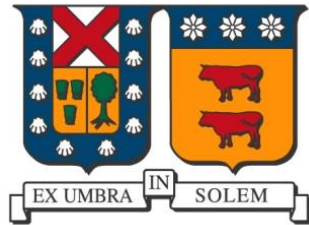


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO - CHILE



“PROYECCIÓN DE DEMANDA GEO-TEMPORAL APLICADA A DELL TECHNOLOGIES LATAM”

ROBERTO ABARCA HENRÍQUEZ

**MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA**

Profesor Guía: Fernando Banz Ramírez
Profesor Correferente: José Luís Martí Lara

Marzo - 2024

DEDICATORIA

Dedicado a mi familia por apoyarme desde siempre a cumplir mis metas, mis logros son suyos.

AGRADECIMIENTOS

A mis padres Jessica y Roberto, por todos estos años, esfuerzo y cariño que me han entregado, les estaré eternamente agradecido por todo lo que han hecho e intentaré retribuirlo día a día.

A mis hermanos y hermanas Javiera, Cristian, Joaquín, Ayelén y Vicente, que a pesar de la distancia y ser pésimos para mantener contacto siempre los sentí cerca.

A Phillipa, que durante toda mi etapa universitaria me ayudó en todos los aspectos posibles, fuiste mi familia, mi amiga y mi compañera y te agradeceré siempre todo lo que significaste para mí.

A mis amigos Bryan, Matías, Diego, Álvaro, Christopher, Cristóbal, Joan, Nicolás, Jorge, Mauricio, con quienes compartí gran parte de mi estadía en la universidad, desde el primer día hasta el último, siempre se los dije, pero qué suerte tuvimos de poder formar este grupo, que la distancia no sea impedimento para continuar con las amistades.

A mis compañeros del colegio y de la vida, con quienes tengo los mejores recuerdos y a pesar de que pasen los años siempre les tendré un lugar en mi vida.

A el profesor Fernando Banz, quien me enseñó el mundo de la inteligencia de negocios y despertó mi cariño por el área, además de siempre ayudarme cuando necesité de su ayuda.

¡A todos ellos, muchísimas gracias!

RESUMEN

Resumen—En este trabajo se diseñó una metodología de limpieza de datos enfocada en inteligencia de negocios para el conjunto de datos de ventas de Dell Technologies LATAM asociadas a proyectos, para luego proponer, construir y evaluar diferentes modelos de predicción de series temporales que nos permitan realizar una proyección de demanda geo-temporal del margen recaudado, además de realizar un trabajo de reportería utilizando *Power BI*. El trabajo se realizó utilizando Python como herramienta de limpieza de datos y creación de modelos de predicción, mientras que el trabajo de reportería se realizó utilizando *Power BI*. Los modelos resultaron exitosos, encontrando ligeras diferencias entre los modelos, pero cada uno con sus beneficios y desventajas, además la limpieza de datos logró que el trabajo de reportería fuese clave. Estos resultados pueden ser de mucha utilidad si se utiliza con el conocimiento del negocio suficiente, pues permitirá mejorar la toma de decisiones y potenciar a Dell Technologies.

Palabras Clave—Inteligencia de Negocios; Series Temporales; Previsión; Limpieza de Datos; Análisis Multidimensional

ABSTRACT

Abstract—*In this work we designed a data cleaning methodology focused on business intelligence for the Dell Technologies LATAM sales data set associated with projects, to then propose, build and evaluate different time series prediction models that allow us to make a geo-temporal demand projection of the margin collected, in addition to performing a reporting work using Power BI. The work was performed using Python as a tool for data cleaning and creation of prediction models, while the reporting work was performed using Power BI. The models were successful, finding slight differences between the models, but each with its benefits and disadvantages, and the data cleansing made the reporting work key. These results can be very useful if used with sufficient knowledge of the business, as it will improve decision making and empower Dell Technologies.*

Keywords—*Business Intelligence; Time Series; Forecasting; Data Cleaning; Multidimensional Analysis.*

GLOSARIO

ADF: Augmented Dickey-Fuller

BI: Business Intelligence

CRISP-DM: Cross-Industry Standard Process for Data Mining

DM: Data Mining

DW: Data Warehouse

EDA: Exploratory data analysis

ETL: Extract, Transform and Load

IOT: Internet of Things

KDE: Kernel Density Estimate

KPI: Key Performance Indicator

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

ML: Machine Learning

MSE: Mean Squared Error

MOLAP: Multi-Dimensional On-Line Analytical Processing

OLAP: On-Line Analytical Processing

OLTP: On-Line Transaction Processing

PDF: Probability Density Function

RMSE: Root Mean Squared Error

ROLAP: Relational On-Line Analytical Processing

STEM: Science, Technology, Engineering and Mathematics

INDICE DE CONTENIDOS

RESUMEN	4
ABSTRACT	4
INDICE DE FIGURAS	8
INDICE DE TABLAS	10
INTRODUCCIÓN	11
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	13
1.1 DESCRIPCIÓN DE LA EMPRESA	13
1.2 PROBLEMA ACTUAL	13
1.3 OBJETIVOS.....	15
1.3.1 OBJETIVO GENERAL	15
1.3.2 OBJETIVOS ESPECÍFICOS	15
1.4 ÁRBOL DEL PROBLEMA	15
CAPÍTULO 2: MARCO CONCEPTUAL	16
2.1 INTELIGENCIA DE NEGOCIOS	16
2.1.1 CONTEXTUALIZACIÓN	16
2.1.2 COMPONENTES DE UN SISTEMA BI	18
2.1.3 BENEFICIOS DE LA INTELIGENCIA DE NEGOCIOS	28
2.1.4 TENDENCIAS	29
2.2 MINERÍA DE DATOS.....	30
2.2.1 INTRODUCCIÓN	30
2.2.2 TAREAS DESCRIPTIVAS	32
2.2.3 TAREAS PREDICTIVAS	33
2.2.4 ANÁLISIS DE SERIES TEMPORALES	34
2.2.5 ANÁLISIS GEOTEMPORAL	35
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN	36
3.1 LIMPIEZA DE DATOS	36

3.1.1	DATASET	36
3.1.2	CARGA DE DATOS	38
3.1.3	LIMPIEZA DE DATOS	41
3.2	ANÁLISIS EXPLORATORIO DE DATOS (EDA)	51
3.3	FORECASTING	64
3.4	REPORTERIA	75
CAPÍTULO 4: VALIDACIÓN DE LA SOLUCIÓN		82
4.1	METODOLOGÍA DE VALIDACIÓN	82
4.2	VALIDACIÓN FORECASTING	83
4.3	INTERFAZ GRÁFICA DE USUARIO	89
4.4	REPORTERÍA EN POWER BI	91
CAPÍTULO 5: CONCLUSIONES		94
5.1	CUMPLIMIENTO OBJETIVO GENERAL	94
5.2	CUMPLIMIENTO OBJETIVOS ESPECÍFICOS	95
5.3	TRABAJO FUTURO	96
REFERENCIAS BIBLIOGRÁFICAS		97

INDICE DE FIGURAS

Figura 1: Árbol del problema. Fuente: Elaboración propia.	15
Figura 2: Datos, información y conocimiento. Fuente: Elaboración propia	17
Figura 3: Proceso BI. Fuente: [Chaudhuri97]	17
Figura 4: Metodología Kimball. Fuente: [Mendoza22]	19
Figura 5: Metodología Inmon. Fuente: [Inmon92]	20
Figura 6: Modelo Estrella Fuente: [Verne22]	26
Figura 7: Modelo Copo de Nieve Fuente: [Verne22]	27
Figura 8: Carga Dataset Fuente: Elaboración Propia	40
Figura 9: Registro dataset Fuente: Elaboración Propia.....	41
Figura 10: Carga de datos en Python Fuente: Elaboración propia	43
Figura 11: Distribución de ventas por año Fuente: Elaboración propia.....	44
Figura 12: Registro agradecimiento Fuente: Elaboración Propia.....	49
Figura 13: Distribución Task Order Date Fuente: Elaboración Propia	52
Figura 14: Distribución Countryname Fuente: Elaboración Propia	53
Figura 15: Distribución Category Fuente: Elaboración Propia	54
Figura 16: Histograma Owner por Project Id Fuente: Elaboración Propia	55
Figura 17: Histograma Owner por Item Quantity Fuente: Elaboración Propia	56
Figura 18: Histograma Owner por Total USD Price Fuente: Elaboración Propia.....	57
Figura 19: Histograma Owner por Total USD Margin Fuente: Elaboración Propia	58
Figura 20: Ventas y margen mensuales Fuente: Elaboración Propia	59
Figura 21: Ventas mensuales por Countryname Fuente: Elaboración Propia	60
Figura 22: Margen mensual por Countryname Fuente: Elaboración Propia	61
Figura 23: Ventas por día de semana Fuente: Elaboración Propia.....	62
Figura 24: Ventas por mes Fuente: Elaboración Propia.....	62
Figura 25: Distribución normal de Total USD Margin y Total USD Price Fuente: Elaboración Propia.....	63
Figura 26: Violin Plot de Total USD Margin y Total USD Price Fuente: Elaboración Propia.....	63

Figura 27: Dickey-Fuller Test Fuente: Elaboración Propia.....	66
Figura 28: Optimización de parámetros Fuente: Elaboración Propia	67
Figura 29: SARIMA Diagnostic Fuente: Elaboración Propia	70
Figura 30: Modelo Prophet resultado de entrenamiento Fuente: Elaboración Propia	71
Figura 31: logLoss NeuralProphet Fuente: Elaboración Propia	73
Figura 32: Diagnóstico Holt-Winters Fuente: Elaboración Propia.....	75
Figura 33: Pasos aplicados tabla Sales Fuente: Elaboración Propia	77
Figura 34: Muestra tabla Sales Fuente: Elaboración Propia	78
Figura 35: Tabla Medidas Principales Fuente: Elaboración Propia.....	78
Figura 36: Tabla Calendar Fuente: Elaboración Propia	80
Figura 37: Modelo Estrella Power BI Fuente: Elaboración Propia.....	81
Figura 38: SARIMA Forecast Fuente: Elaboración Propia.....	84
Figura 39: Prophet Forecast Fuente: Elaboración Propia.....	84
Figura 40: NeuralProphet Forecast Fuente: Elaboración Propia.....	85
Figura 41: Holt-Winters Forecast Fuente: Elaboración Propia	85
Figura 42: GUI Forecast Menú Fuente: Elaboración Propia	90
Figura 43: GUI Forecast ejemplo ejecución Fuente: Elaboración Propia	90
Figura 44: Power BI página 1 Fuente: Elaboración Propia	91
Figura 45: Power BI página 2 Fuente: Elaboración Propia	92
Figura 46: Power BI página 3 Fuente: Elaboración Propia	92
Figura 47: Power BI página 4 Fuente: Elaboración Propia	93

INDICE DE TABLAS

Tabla 1: Tabla comparativa métodos Inmon y Kimball. Fuente: Elaboración Propia.	20
Tabla 2: Tabla comparativa sistemas MOLAP, ROLAP y HOLAP. Fuente: Elaboración Propia.	25
Tabla 3: Columnas LATAM Project Equipment List by Date. Fuente: Elaboración Propia.....	37
Tabla 4: Registro error. Fuente: Elaboración Propia.	39
Tabla 5: Registro SKU Number duplicado. Fuente: Elaboración Propia.	48
Tabla 6: Formato de los campos de salida Fuente: Elaboración Propia.....	50
Tabla 7: Parámetros SARIMA Fuente: Elaboración Propia	69
Tabla 8: Parámetros Prophet Fuente: Elaboración Propia	71
Tabla 9: Parámetros NeuralProphet Fuente: Elaboración Propia	72
Tabla 10: Parámetros Holt-Winters Fuente: Elaboración Propia	74
Tabla 11: Variables tabla Medidas Principales Fuente: Elaboración Propia	79
Tabla 12: Tabla de errores SARIMA Fuente: Elaboración Propia	86
Tabla 13: Tabla de errores Prophet Fuente: Elaboración Propia	87
Tabla 14: Tabla de errores NeuralProphet Fuente: Elaboración Propia	87
Tabla 15: Tabla de errores Holt-Winters Fuente: Elaboración Propia	88
Tabla 16: Resumen de errores Fuente: Elaboración Propia	89

INTRODUCCIÓN

En el contexto empresarial actual, donde las empresas son cada vez más competitivas y eficiente, urge la necesidad de las empresas por mantener su relevancia y asegurar su continuidad en el mercado. En este sentido, la información que se posee previo a la toma de decisiones puede marcar la diferencia ante sus competidores, garantizando el éxito y la sostenibilidad de las organizaciones. Para lograrlo es crucial que las empresas cuenten con la información precisa y oportuna que respalde sus procesos de tomas de decisiones.

Es en este contexto que surge la necesidad de implementar herramientas de *Business Intelligence* (BI) o Inteligencia de Negocios. El BI permite a las organizaciones agregar inteligencia a sus procesos empresariales, facilitando la extracción, análisis y visualización de datos para la toma de decisiones.

En el marco de este trabajo de investigación, se reconoce la importancia estratégica del BI para Dell Technologies en su objetivo de proyectar la demanda de sus productos en el mercado latinoamericano. Se busca aprovechar la gran cantidad de datos disponibles para descubrir conocimientos relevantes que permitan reducir la incertidumbre en la toma de decisiones, anticipar tendencias y optimizar la planificación de la demanda.

A través de la aplicación de técnicas avanzadas de análisis de datos, modelado predictivo y generación de reportes en Power BI, se pretende proporcionar a Dell Technologies una herramienta efectiva para anticipar y gestionar la demanda de sus productos en América Latina, contribuyendo así a su competitividad y éxito en el mercado.

La naturaleza del problema radica en la complejidad de los datos y la necesidad de identificar patrones significativos en la demanda de productos tecnológicos en un contexto geográfico diverso y en constante cambio. La propuesta de solución consiste en aplicar una metodología de limpieza de datos, exploración de los datos y modelado predictivo para obtener proyecciones precisas y útiles para la toma de decisiones estratégicas.

La estructura de la memoria sigue un enfoque sistemático, el cual queda definido en los siguientes capítulos:

- **CAPÍTULO 1 - DEFINICIÓN DEL PROBLEMA:** Se presenta la empresa relacionada con el trabajo, se contextualiza el problema actual y se definen los objetivos generales y específicos.
- **CAPÍTULO 2 - MARCO CONCEPTUAL:** Se definen todos los conceptos del área que permitan a los lectores un mayor entendimiento del trabajo.
- **CAPÍTULO 3 - PROPUESTA DE SOLUCION:** Se realiza la metodología de trabajo, abarcando desde el diseño de la solución hasta el desarrollo de esta, explicando cada paso realizado, desde el comienzo del trabajo hasta el final de este.

- **CAPÍTULO 4 - VALIDACIÓN DE LA SOLUCIÓN:** Se entregan métricas evaluables de la solución entregada, permitiendo definir el cumplimiento de los objetivos planteado en un comienzo.
- **CAPÍTULO 5 – CONCLUSIONES:** Se finaliza el trabajo con la realización de las conclusiones, alcances de este mismo y trabajos futuros por realizar.

CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA

1.1 DESCRIPCIÓN DE LA EMPRESA

Dell Technologies¹ es una empresa estadounidense con sucursales de venta en todo el mundo, líder en el sector de servicios TI y consultorías TI, destacando en las áreas de fabricación, venta y soporte de computadoras, servidores, programas informáticos y otros productos tecnológicos, además ofrece servicios en la nube y servicios de almacenamiento de datos.

Debido a los años de trayectoria que posee la empresa sumada a las buenas decisiones tomadas por su fundador Michael Dell, Dell Technologies ha conseguido posicionarse dentro de las empresas líderes en el área tecnológica.

Dell Technologies nace en el 2016 con la fusión de EMC Corporation (ahora llamada EMC Dell) y la compañía Dell, la cual originó a la mayor compañía tecnológica del mundo del capital privado, se estima que posee 133.000 empleado alrededor del mundo² y en el último año han obtenido un ingreso anual de 101.200 millones de dólares³.

Como se mencionó anteriormente el rubro principal de la empresa es la venta de servicios y productos, por lo cual el obtener información relevante con los procesos de ventas es algo crucial para la organización, sobre todo al poseer numerosos puntos de ventas y sucursales a lo largo del mundo, es bajo esta afirmación que centraremos el problema, la obtención de información para luego analizarla y crear visualizaciones que permitan a la organización una mejor toma de decisiones, o dicho de otra forma, *business intelligence* (BI).

1.2 PROBLEMA ACTUAL

Dell Technologies realiza una variedad de ventas de productos tanto de forma presencial como en línea, y estas transacciones se almacenan en una base de datos que registra todas las operaciones de ventas. Estos datos son fundamentales para llevar a cabo análisis, crear modelos predictivos, comparar estadísticas de diferentes períodos y realizar otras tareas relacionadas.

A pesar de la importancia de estos datos, en la actualidad, Dell Technologies en LATAM se enfrenta a un desafío importante en cuanto al análisis de datos debido a problemas en sus bases de datos de ventas. Estos problemas incluyen errores de separación de datos, columnas que no aportan información relevante, registros relacionados con productos no comprados, como garantías de productos, mensajes de agradecimiento, entre otros. La limpieza y preparación de estos datos es una tarea compleja, especialmente dado el volumen significativo de datos específicos de la empresa. Para abordar este desafío, es necesario aplicar un proceso de *Business Intelligence*.

¹ <https://www.dell.com/es-cl/dt/index.htm>

² <https://es.tradingview.com/symbols/BMV-DELLC/financials-statistics-and-ratios/number-of-employees/>

³ <https://es.investing.com/equities/dell-inc-income-statement>

No obstante, la limpieza y preparación de datos no es el único problema. La principal preocupación de Dell Technologies radica en la falta de información al tomar decisiones estratégicas relacionadas con las ventas de productos en Latinoamérica. La ausencia de análisis de BI tiene consecuencias negativas, como pérdidas financieras, problemas de inventario, pérdida de clientes y otros problemas adicionales. Por lo tanto, la realización de un análisis de proyección geo-temporal puede atenuar estos resultados inesperados, brindando a la empresa una ventaja al permitir la toma de decisiones estratégicas informadas para maximizar el beneficio de las ventas.

Para abordar el problema en profundidad, es importante considerar los procesos de BI necesarios para una solución en la empresa.

La aplicación de un proceso de BI en una empresa no es un proceso simple ni rápido, y es por eso por lo que existen numerosas empresas de consultoría especializadas en BI. En un nivel inicial, existen tres pasos fundamentales para la creación de un *dataset*⁴ utilizable en análisis posteriores. Estos pasos están definidos por las siglas *ETL* (*Extract, Transform, Load*):

1. **Extracción:** Consiste en la recopilación de datos desde diversas fuentes.
2. **Transformación:** Implica la limpieza, estandarización y preparación de los datos para su análisis.
3. **Carga:** Implica la incorporación de los datos limpios y procesados en un sistema o repositorio adecuado.

Este proceso ETL es un proceso bastante largo, según Lohr de *The New York Times* "Los científicos de datos, según entrevistas y estimaciones de expertos, pasan entre el 50 y el 80 por ciento de su tiempo enfangados en la mundana labor de recopilar y preparar datos digitales desordenados, antes de poder explorarlos en busca de pepitas útiles." [Lohr14].

También este apartado es importante debido a que un *dataset* deficiente o con información errónea pueden provocar impactos graves en la generación de reportes, lo cual provoca pérdida de credibilidad y una mala toma de decisiones, provocando pérdidas monetarias, impacto negativo con el negocio y otros problemas.

Además del proceso de ETL, existe el proceso de análisis, el cual se encarga de realizar tareas descriptivas, diagnosticas, predictivas, prescriptivas y cognitivas. Cada una de estas posee un enfoque distinto, pero todas buscan obtener información relevante a partir de los datos obtenidos mediante el proceso de ETL.

⁴ Colección o representación de datos residentes en memoria con un modelo de programación relacional coherente e independientemente del origen de los datos que contiene.

1.3 OBJETIVOS

1.3.1 OBJETIVO GENERAL

Diseñar, construir e implementar un proceso de análisis mediante técnicas de Inteligencia de Negocios aplicado a las ventas de la empresa Dell Technologies en Latinoamérica durante los años 2012 al 2019, con el fin de apoyar la toma de decisiones respecto a las estrategias organizacionales de venta.

1.3.2 OBJETIVOS ESPECÍFICOS

- Analizar e identificar los principales errores que existen en la base de datos de ventas de Dell en Latinoamérica.
- Diseñar un método de limpieza y estandarización de la base de datos.
- Realizar una proyección de demanda geo-temporal de las ventas de Dell en Latinoamérica en periodos futuros.

1.4 ÁRBOL DEL PROBLEMA

El problema descrito anteriormente sumado a los objetivos nombrados queda resumido en la siguiente figura que representa al árbol del problema.

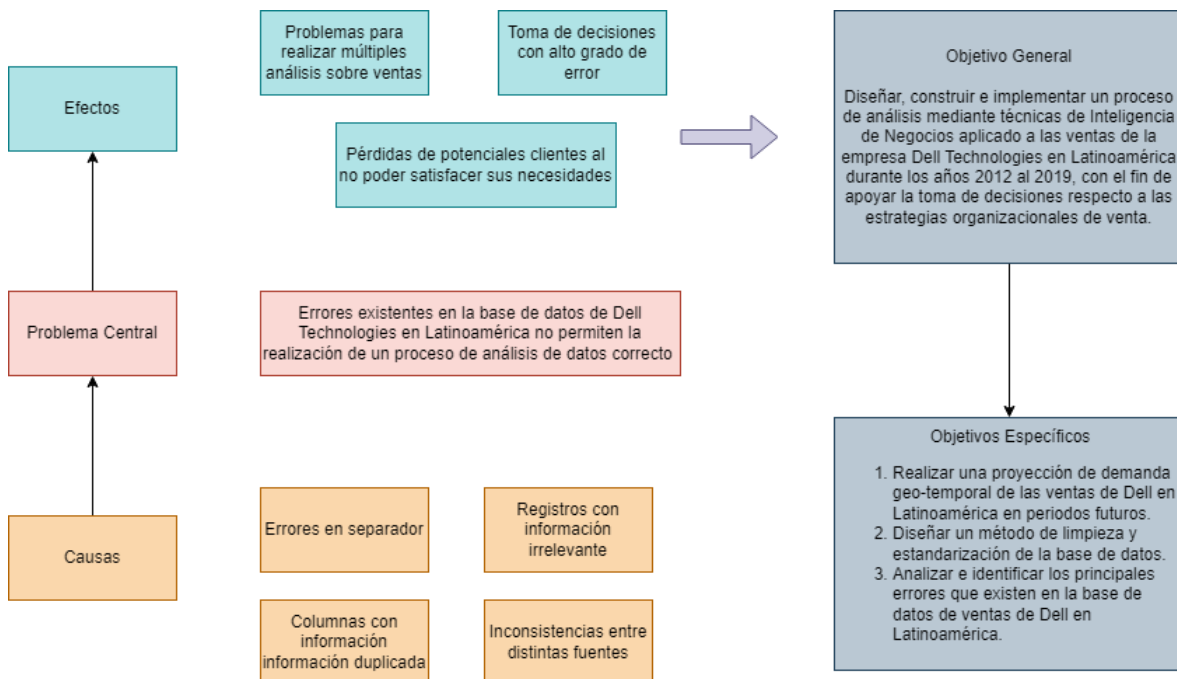


Figura 1: Árbol del problema.

Fuente: Elaboración propia.

CAPÍTULO 2: MARCO CONCEPTUAL

2.1 INTELIGENCIA DE NEGOCIOS

En este apartado analizaremos qué es la Inteligencia de Negocios, su flujo de trabajo, sus beneficios, sus componentes y cualquier concepto importante para la comprensión del desarrollo del trabajo, así como la importancia que tiene en cualquier organización.

2.1.1 CONTEXTUALIZACIÓN

La Inteligencia de Negocios (BI) es una solución tecnológica que permite a las organizaciones gestionar la información necesaria para respaldar la toma de decisiones estratégicas. Comprende un proceso que va desde la extracción de datos de sistemas existentes hasta la explotación de la información a través de herramientas de análisis de datos. Para comprender la BI, es fundamental comprender la diferencia entre dato, información y conocimiento.

- **Dato:** Un dato es una representación simbólica de un atributo o variable, ya sea cuantitativa o cualitativa. Por si solo un dato aislado puede carecer de información relevante.
- **Información:** La información se deriva de la transformación y organización de datos. Se modela para que las organizaciones puedan tomar decisiones informadas. Los sistemas de información actúan como "cajas negras" que toman datos como entrada y producen información como salida.
- **Conocimiento:** El conocimiento es un conjunto de información adquirida a través de la experiencia y el aprendizaje. Es un recurso invaluable para la organización. La BI tiene como objetivo adquirir información a partir de datos para generar conocimiento que respalde la toma de decisiones estratégicas.

Es importante comprender que existe una relación directa entre estos tres conceptos, debido a que para son los datos los elementos que nos permiten obtener la información y es esta información la que nos provee el conocimiento que obtendrá la organización, con este conocimiento se pueden realizar unas mejores tomas de decisiones. La relación de los conceptos se detalla a continuación (Ver Figura 2).

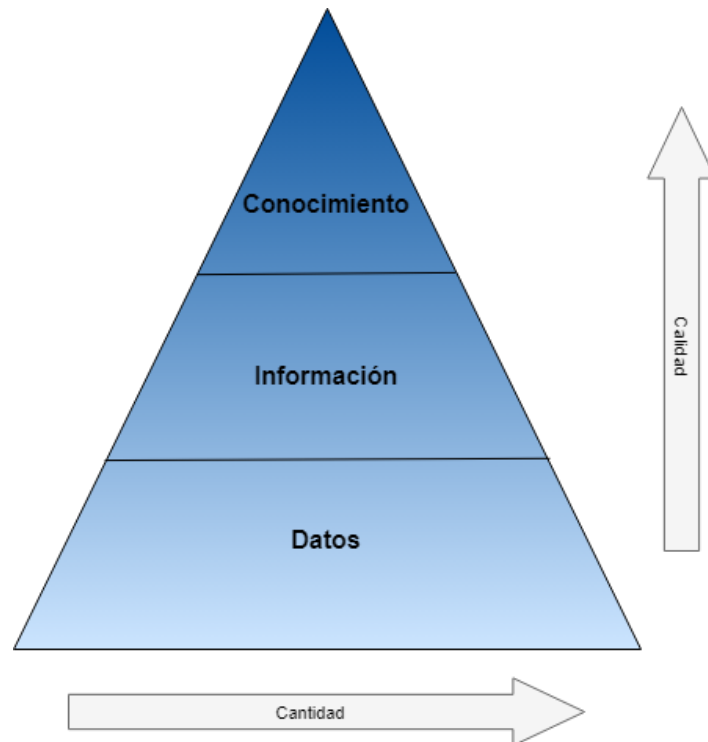


Figura 2: Datos, información y conocimiento.

Fuente: Elaboración propia

El proceso de Inteligencia de Negocios sigue un flujo de trabajo que comienza con la extracción de datos, realizando el proceso de ETL, utilizando almacenes de datos y culmina con la utilización de herramientas de análisis de datos para obtener visualizaciones acordes a las necesidades de la organización. El proceso se detalla a continuación (ver Figura 3).

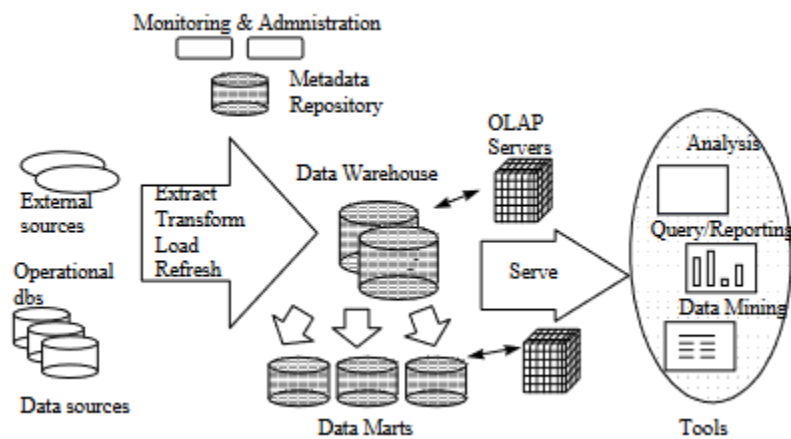


Figura 3: Proceso BI.

Fuente: [Chaudhuri97]

2.1.2 COMPONENTES DE UN SISTEMA BI

En esta sección definiremos los componentes principales que forman parte de un sistema de Inteligencia de Negocios.

2.1.2.1 DATA WAREHOUSE

Un *Data Warehouse* en palabra de William H. Inmon es una “recopilación de datos no volátil, integrada, orientada a un tema y variable en el tiempo, que se utiliza principalmente en la toma de decisiones organizativas.” [Inmon92]. Usando una definición menos formal podemos definir un *Data Warehouse* (DW) como un repositorio donde se almacenan e integran datos históricos de diversas fuentes con el propósito de facilitar el análisis OLAP para respaldar la toma de decisiones. Los *Data Warehouses* deben cumplir con ciertas características fundamentales [Vyas17].

- **Integrado:** Los datos pueden provenir de múltiples fuentes heterogéneas que deben integrarse. Por ejemplo, se pueden agregar datos de diferentes fuentes de la misma empresa en diferentes países para obtener un conjunto de datos integrados.
- **Temporal:** Los datos almacenados pueden ser actuales o históricos, pero es esencial que posean una dimensión temporal que les dé significado. Esto permite realizar análisis temporales, como examinar las ventas de los últimos 6 meses.
- **No Volátil:** El DW se mantiene separado de las bases de datos operativas de la empresa, lo que significa que los datos en el DW no suelen sufrir modificaciones significativas. Mientras que las bases de datos operativas usan operaciones CRUD (Crear, Leer, Actualizar, Eliminar), en un DW solo se realiza la carga y lectura de la información.
- **Orientada a un tema:** Los *Data Warehouses* se centran en un tema específico, ya que todos los datos que se almacenan deben estar relacionados entre sí para permitir un análisis relevante.

2.1.2.2 DATA MART

Un *Data Mart* es un subconjunto de un *Data Warehouse*, diseñado, construido e implementado para satisfacer las necesidades de un área específica de la organización, comúnmente un departamento. Comparten las mismas características que los *Data Warehouses*, pero al tener información más centralizada, permiten análisis y generación de informes separados para cada departamento, optimizando los recursos y controlando mejor la información que manejan.

En relación con los *Data Marts*, existen dos metodologías de creación de *Data Warehouses*. Una se enfoca en la creación del *Data Warehouse* mediante la unión de múltiples *Data Marts*, mientras que la otra implica crear múltiples *Data Marts* a partir de un *Data Warehouse* ya existente. A continuación, se analizarán en detalle ambas metodologías con sus pros y contras.

El primer método es llamado **Método de Kimball** [Forero21], es basado en un enfoque de *Bottom-Up*⁵, según Ralph Kimball la metodología se rige por los siguientes principios:

- Centrarse en el negocio.
- Construir una infraestructura de información adecuada.
- Realizar entregas incrementales significativos.
- Ofrecer una solución completa.

Esta metodología facilita la creación del *Data Warehouse* y permite priorizar la creación de *Data Marts* según las necesidades de la organización. Sin embargo, puede ser más costosa a largo plazo debido a los desafíos de integración. El proceso se detalla a continuación (ver Figura 4).

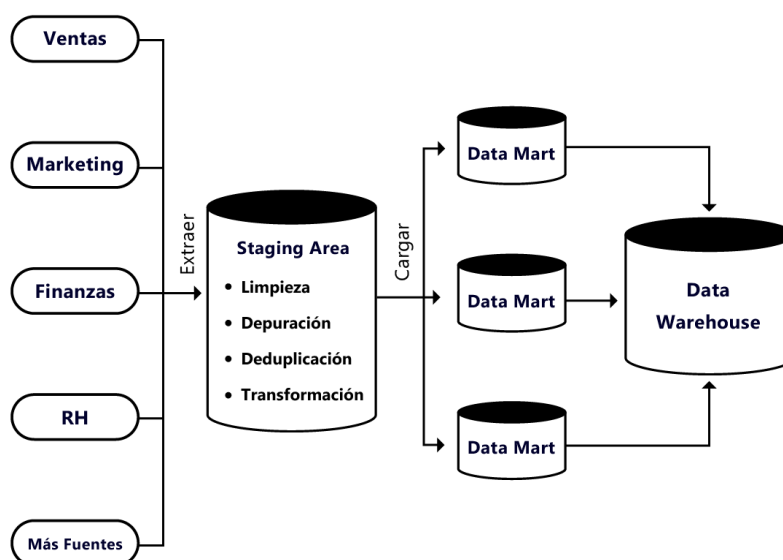


Figura 4: Metodología Kimball.

Fuente: [Mendoza22]

La otra opción es el **Método de Inmon**, presentada por Inmon [Inmon92] y reforzada por E. Malinowski y E. Zimányi en 2008. Basada en un enfoque de *Top-Down*⁶, comienza recopilando los requisitos de los usuarios en las diferentes áreas y niveles de la organización. Luego, se inicia un proceso de diseño y construcción del *Data Warehouse* completo, a partir del cual se seccionan los *Data Marts*. Esta metodología se centra en los datos corporativos, aunque es más costosa y lleva más tiempo. El proceso se detalla a continuación (ver Figura 5).

⁵ Las partes individuales se diseñan con detalle y luego se enlazan para formar componentes más grandes.

⁶ Se formula un resumen del sistema, sin especificar detalles. Cada parte nueva es entonces redefinida, cada vez con mayor detalle, hasta que la especificación completa es lo suficientemente detallada para validar el modelo.

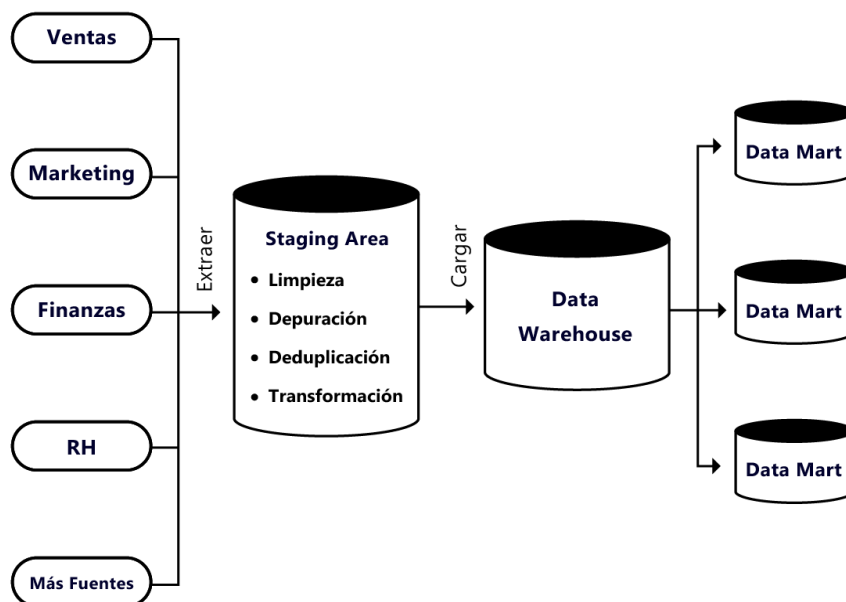


Figura 5: Metodología Inmon.

Fuente: [Inmon92]

A continuación, se puede observar una tabla comparativa entre ambos métodos, donde podrá ver las ventajas y desventajas de cada uno de los métodos en diferentes características.

Tabla 1: Tabla comparativa métodos Inmon y Kimball.

Fuente: Elaboración Propia.

	Método de Inmon	Método de Kimball
Enfoque	Top-Down	Bottom-Up
Inversión	Alta	Baja
Desarrollo	Tiempos de desarrollo largos	Tiempos de desarrollo bajo
Mantenimiento	Simple	Medio-Alto a largo plazo
Tiempo	Mayor tiempo para su inicio	Menor tiempo para la configuración inicial
Calidad	Mayor calidad en el diseño por análisis completo	Menor calidad debido a la unión de cada área
Flexibilidad	Mayor flexibilidad	Menor flexibilidad

2.1.2.3 ETL

La integración de datos se gestiona a través de procesos de extracción, transformación y carga (ETL), que son etapas encargadas de filtrar, transformar y redefinir los datos en cada etapa de su composición. Estas fases son consideradas unas de las más exigentes en los proyectos de BI, ya que consumen entre el 60 y el 80% de la carga de trabajo [Gómez13]. Este trabajo involucra la heterogeneidad y diversidad de fuentes de información, las cuales, como se mencionó anteriormente, pueden existir en varios formatos y deben integrarse mediante procesos ETL:

1. **Extracción** (*Extract*): Esta etapa es responsable del proceso de obtención de datos de diversas fuentes, que pueden estar dispersas en diferentes ubicaciones dentro y fuera de la organización. Además, estos datos pueden presentarse en varios formatos, como bases de datos transaccionales, hojas de cálculo, archivos de texto sin formato, archivos XML, entre otros. En términos generales, esta fase de extracción debe realizarse en momentos en que la organización experimenta baja actividad del sistema operativo, por lo que generalmente se lleva a cabo durante la noche. Esto se hace para minimizar el impacto en las operaciones diarias de la organización.
2. **Transformación** (*Transform*): Dado que la información se origina en diferentes formatos y cada uno de ellos puede tener su propia definición y estructura, es necesario un paso de cuidado de datos para corregir y ajustar los datos según sea requerido. Es la más compleja de las etapas, ya que es donde pueden surgir más problemas, especialmente cuando se utilizan múltiples fuentes durante la extracción de datos. La etapa de transformación se divide en tres subetapas: Limpieza de datos, Formateo de datos y Mezcla de datos.
 - 1) Limpieza de datos: En esta subetapa, se detectan y corrigen anomalías. Se deben buscar las causas de las anomalías y eliminarlas utilizando técnicas de limpieza. Es importante ser cauteloso al eliminar datos anómalos, ya que esto podría sesgar la información, en estos casos hay que comparar los resultados y documentar los criterios a utilizar.
 - 2) Formateo de datos: En esta subetapa, los datos extraídos se formatean para cumplir con los requisitos del *Data Warehouse*. Deben realizarse consideraciones sobre el tipo y el tamaño de los campos. Las transformaciones pueden ser simples o complejas dependiendo de las necesidades del análisis.
 - 3) Mezcla de datos: Aquí se considera la integración, derivación y agregación de datos antes de la carga. Es posible combinar tablas o generar datos derivados o agregados para satisfacer las necesidades de la organización.
3. **Carga** (*Load*): Esta etapa implica la inserción de los datos preparados en el *Data Warehouse*, donde estarán disponibles para su análisis. Es esencial que la carga se realice cuando el *Data Warehouse* no esté en uso y que se utilicen técnicas de carga

que optimicen el tiempo. También se recomienda la programación de cargas automáticas periódicas y la creación de copias de seguridad por si el proceso de carga falla. En esta etapa se pueden agregar tablas con datos pre calculados que permitan una optimización en los tiempos de consulta sobre los requerimientos más frecuentes.

2.1.2.4 LIMPIEZA DE DATOS

La limpieza de datos también conocida como *Data Cleansing*, es un proceso fundamental en la gestión de la inteligencia de negocios que se encarga de identificar y corregir errores, inconsistencia y anomalías en los datos recopilados antes de que se integren en un *Data Warehouse* o un *Data Mart*. Este proceso es esencial para garantizar que los datos utilizados en las decisiones de negocios sean precisos, confiables y coherentes. La limpieza de datos aborda una serie de problemas comunes que pueden surgir en los datos, como valores faltantes, duplicados, errores tipográficos y discrepancias entre fuentes de datos heterogéneas.

La limpieza de datos desempeña un papel crítico en el éxito de cualquier proyecto de inteligencia de negocios. Los datos inexactos pueden llevar a decisiones incorrectas, pérdida de oportunidades de negocio y, en última instancia, un impacto negativo en el rendimiento de la organización. Los beneficios clave de la limpieza de datos incluyen:

- **Precisión en la Toma de Decisiones:** Los datos limpios y precisos proporcionan una base sólida para la toma de decisiones estratégicas. La limpieza de datos ayuda a evitar decisiones basadas en información errónea o incompleta.
- **Eficiencia Operativa:** Los datos limpios mejoran la eficiencia en los procesos empresariales, al eliminar la necesidad de abordar y corregir errores en etapas posteriores del análisis.
- **Reducción de Costos:** La limpieza de datos temprana puede ahorrar costos significativos asociados con la corrección de errores y la pérdida de recursos debido a decisiones erróneas.
- **Errores Reducidos:** La corrección de errores estructurales y tipográficos conduce a una disminución de los errores en los datos, lo que se traduce en una mayor confiabilidad de los resultados.
- **Cumplimiento Normativo:** En muchas industrias, el cumplimiento de regulaciones requiere la precisión y la integridad de los datos. La limpieza de datos contribuye al cumplimiento normativo al garantizar datos confiables.

Aunque no existe un proceso universalmente aplicable para la limpieza de datos, se pueden seguir algunos pasos fundamentales para iniciar el proceso de limpieza de datos. Estos pasos incluyen:

- **Identificación de Problemas:** Esta etapa implica la detección de problemas como valores faltantes, duplicados, valores atípicos, errores tipográficos y discrepancias entre fuentes de datos. El software de limpieza de datos y reglas predefinidas pueden ayudar en este proceso.
- **Evaluación de Datos:** Una vez identificados los problemas, se evalúa su impacto en la calidad de los datos y en las decisiones empresariales. Esto ayuda a priorizar qué problemas deben abordarse primero.
- **Corrección y Estandarización:** En esta fase, se aplican correcciones a los datos problemáticos. Esto puede incluir la eliminación de registros duplicados, la corrección de errores tipográficos y la estandarización de formatos.
- **Validación de Datos:** Los datos corregidos se validan para garantizar que cumplan con los estándares definidos de calidad y precisión. Esto puede incluir la validación de datos utilizando reglas predefinidas.
- **Documentación:** Es importante documentar todo el proceso de limpieza de datos, incluyendo los problemas identificados y las correcciones aplicadas. Esto facilita la trazabilidad y la auditoría.

2.1.2.5 OLAP

El Procesamiento Analítico en Línea (OLAP) es una tecnología de software que permite el análisis de datos empresariales desde múltiples perspectivas. Las organizaciones acumulan datos de diversas fuentes, las cuales son procesadas por la etapa de ETL y almacenadas en un *Data Warehouse* o *Data Mart*. OLAP combina y categoriza estos datos para proporcionar información accionable en el proceso de toma de decisiones estratégicas.

El análisis multidimensional es un componente esencial de OLAP. A diferencia del enfoque unidimensional de las bases de datos relacionales tradicionales, que almacenan datos en tablas, el análisis multidimensional organiza los datos en cubos multidimensionales. Estos cubos contienen información de manera más rica, ya que representan múltiples dimensiones, como tiempo, producto, ubicación, etc. Esto permite una exploración más profunda y flexible de los datos.

Un ejemplo común de análisis multidimensional es el análisis de ventas. Un cubo multidimensional podría representar las ventas en función del tiempo, productos y ubicaciones. Los usuarios pueden navegar a través de estos cubos para descubrir patrones, tendencias y relaciones complejas en los datos.

OLAP desempeña un papel fundamental en el procesamiento y aprovechamiento de la creciente cantidad de información digital. Algunos de los beneficios clave de OLAP son los siguientes:

- **Toma de Decisiones Rápidas:** Permite a las empresas tomar decisiones precisas en un entorno empresarial acelerado. Las consultas analíticas en bases de datos

relacionales pueden ser lentas, pero OLAP calcula previamente e integra datos, lo que acelera la generación de informes.

- **Facilita el Acceso para Usuarios no Técnicos:** Las herramientas OLAP permiten a los usuarios de negocios analizar datos complejos sin necesidad de conocimientos técnicos profundos. Los usuarios pueden crear cálculos analíticos complejos y generar informes sin necesidad de interactuar directamente con las bases de datos.
- **Integración de Datos:** OLAP proporciona una plataforma unificada que permite a las distintas unidades de negocios, como marketing, finanzas y producción, ver el panorama general y tomar decisiones efectivas. Esto facilita la identificación de relaciones y la realización de análisis hipotéticos.

Los sistemas OLAP almacenan datos en una representación multidimensional, lo que significa que se organizan en más de dos dimensiones o categorías. Los elementos claves de la arquitectura OLAP incluyen:

- **Almacenamiento de Datos:** Esta etapa recopila información de diversas fuentes, como bases de datos relacionales, aplicaciones y archivos. La información se procesa y se prepara para el análisis mediante herramientas de Extracción, Transformación y Carga (ETL).
- **Herramientas ETL:** Las herramientas ETL realizan procesos de extracción, transformación y carga, que automatizan la preparación de datos para el análisis. Esta etapa es crucial para estandarizar y convertir datos de diversas fuentes en un formato adecuado para el análisis.
- **Servidor OLAP:** El servidor OLAP es la infraestructura subyacente que impulsa el sistema OLAP. Utiliza herramientas ETL para transformar datos y prepararlos para operaciones OLAP.
- **Base de Datos OLAP:** La base de datos OLAP es independiente y se conecta al almacenamiento de datos. A veces se utiliza para evitar sobrecargar el almacenamiento principal y facilitar la creación de modelos de datos OLAP.
- **Cubos OLAP:** Un cubo de datos es un modelo que representa una matriz multidimensional de información. Los analistas empresariales interactúan con estos cubos para realizar operaciones analíticas.

Un sistema de procesamiento analítico en línea (OLAP) funciona recopilando, organizando, agregando y analizando datos mediante los siguientes pasos:

1. El servidor OLAP recopila los datos desde diversas fuentes, incluyendo bases de datos relacionales y los almacenamientos de los datos.
2. Las herramientas de ETL limpian, agregan, calculan previamente y almacenan los datos en un cubo OLAP según el número de dimensiones especificadas.

- Los analistas empresariales utilizan herramientas OLAP para realizar consultas y generar informes a partir de los datos multidimensionales del cubo OLAP.

Los sistemas OLAP funcionan de tres formas principales:

- **MOLAP:** En MOLAP (Procesamiento Analítico Multidimensional en Línea), los datos son calculados previamente y se almacenan en cubos multidimensionales, lo que permite un análisis rápido.
- **ROLAP:** ROLAP (Procesamiento Analítico Relacional en Línea) permite analizar datos multidimensionales directamente en bases de datos relacionales. Adecuado para datos extensos y detallados, aunque puede tener un rendimiento de consulta más lento.
- **HOLAP:** HOLAP (Procesamiento Analítico Híbrido en Línea) combina características de MOLAP y ROLAP, ofreciendo lo mejor de ambas arquitecturas.

Tabla 2: Tabla comparativa sistemas MOLAP, ROLAP y HOLAP.
Fuente: Elaboración Propia.

Aspecto	ROLAP	MOLAP	HOLAP
Escalabilidad	Escalable para manejar grandes volúmenes de datos.	Menos adecuado para grandes volúmenes de datos.	Escalable y puede manejar grandes volúmenes de datos.
Rendimiento	Rendimiento de consultas puede ser más lento.	Rendimiento rápido en consultas calculadas previamente.	Rendimiento rápido en consultas con datos multidimensionales.
Facilidad de uso	Requiere consultas SQL que pueden ser complejas.	Interfaces intuitivas y fáciles de usar.	Combinación de ROLAP y MOLAP para facilitar el uso.
Modelado	Almacena datos en bases de datos relacionales.	Almacena datos en cubos multidimensionales.	Requiere una infraestructura de almacenamiento de datos bien diseñada.
Soporte para datos en tiempo real	Puede manejar datos actualizados en tiempo real.	No es tan adecuado para datos con cambios frecuentes.	Puede manejar datos actualizados en tiempo real.
Ventajas principales	Flexibilidad, soporte para datos en tiempo real.	Rendimiento rápido en consultas, facilidad de uso.	Combinación de rendimiento y datos detallados.

Desventajas principales	Rendimiento más lento, complejidad en consultas.	Menos adecuado para grandes volúmenes de datos.	Mayor complejidad de implementación y costos.
--------------------------------	--	---	---

El modelado de datos es esencial en OLAP, ya que representa cómo se almacenan los datos. En OLAP, se utilizan esquemas en estrella o copo de nieve para organizar los datos multidimensionales.

- **Esquema en Estrella:** Consiste en una tabla de hechos central y múltiples tablas de dimensiones. Las tablas de hechos contienen valores numéricos relacionados con un proceso empresarial, y las tablas de dimensiones describen atributos.

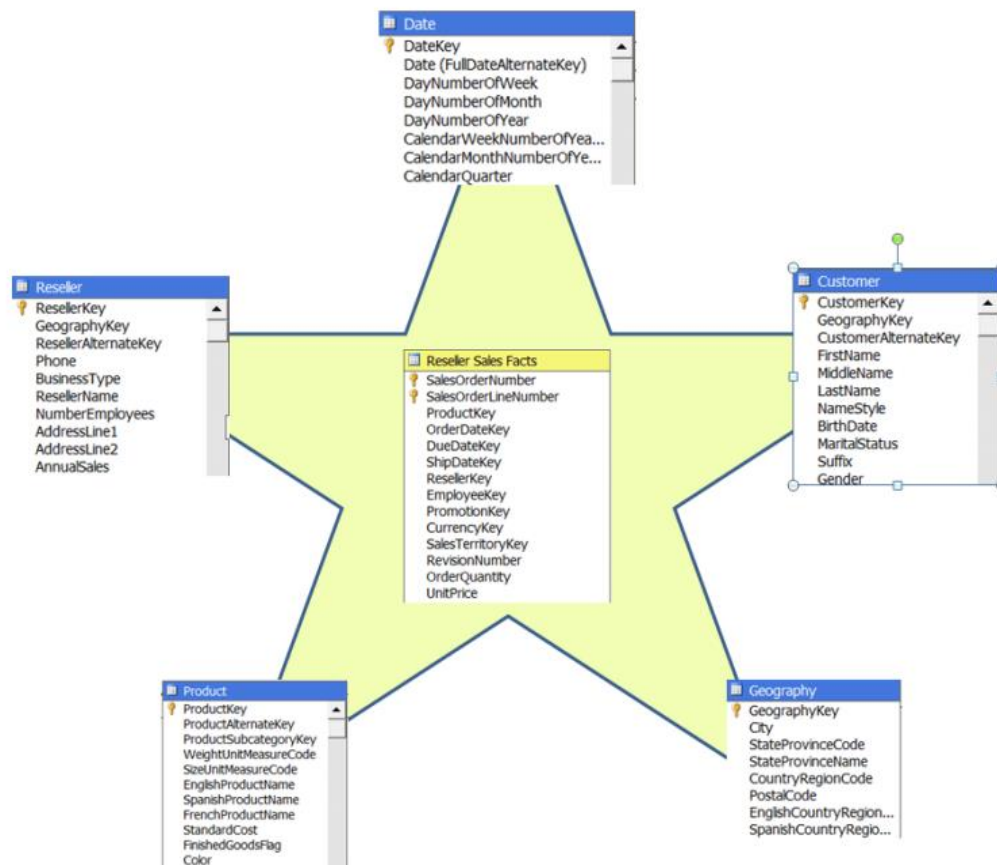


Figura 6: Modelo Estrella

Fuente: [Verne22]

- **Esquema de Copo de Nieve:** Es una extensión del esquema en estrella donde algunas tablas de dimensiones se dividen en tablas secundarias.

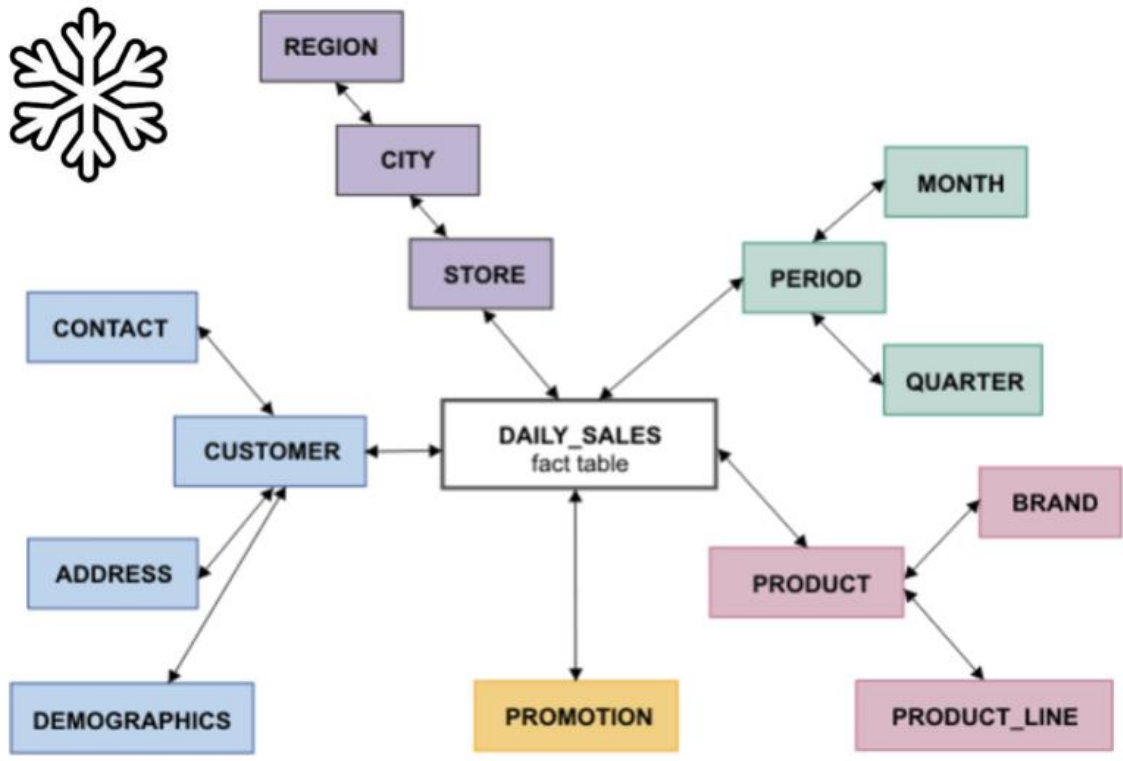


Figura 7: Modelo Copo de Nieve

Fuente: [Verne22]

La principal diferencia de OLAP con OLTP (Procesamiento de Transacciones en Línea) es que OLTP se utiliza para el almacenamiento rápido y confiable de datos transaccionales en una base de datos. OLAP, por otro lado, se enfoca en el análisis multidimensional de datos y no en la gestión de transacciones, esto provoca una relación entre ambos sistemas, ya que se puede conectar una base de datos OLTP a un cubo OLAP para hacer consultas de computación intensivas en los datos históricos.

2.1.2.6 VISUALIZACIONES

La visualización de datos se define como la representación gráfica de información y datos a través de elementos visuales como lo pueden ser gráficos, mapas, líneas temporales, entre otros. En el contexto de *Business Intelligence*, las herramientas y tecnologías de visualización de datos juegan un papel fundamental al permitir la comprensión de tendencias, valores atípicos y patrones en los conjuntos de datos analizados.

El propósito fundamental de las visualizaciones de datos va más allá de la mera presentación; busca contar una historia, eliminando el ruido de los datos y resaltando la información esencial. Todos los campos desde STEM⁷ hasta áreas como gobierno, finanzas, marketing y educación, se benefician de esta comprensión visual de los datos. La capacidad

⁷ STEM por sus siglas en inglés de ciencia, tecnología, ingeniería y matemáticas.

de transmitir puntos claves de manera visual, ya sea a través de un *dashboard* o un informe, es fundamental para aprovechar al máximo la información.

En el ámbito de la visualización de datos, se destacan elementos claves para crear representaciones efectivas. Aspectos como la definición clara de objetivos, elección adecuada de gráficos, la simplicidad en el diseño y el uso intencionados de colores son fundamentales. Estos elementos permiten una comunicación más clara y una interpretación más rápida de la información presentada.

Las ventajas de la visualización de datos son significativas en la actualidad. En un entorno donde las empresas deben lidiar con un volumen masivo de datos, la capacidad de desbloquear información valiosa es esencial. La visualización permite la rápida absorción de grandes cantidades de datos, acelerando el proceso de toma de decisiones e identificando procesos claves, patrones y tendencias cruciales para la supervivencia organizacional.

Algunos beneficios adicionales incluyen la capacidad de revelar rápidamente relaciones entre operaciones y resultados, así como la interactividad que permite a los usuarios profundizar en los datos. Además, la visualización de datos mejora la comunicación entre equipos, mejora la eficiencia en la gestión del tiempo de los empleados y proporciona acceso más amplio a la información en tiempo real, otorgando a las empresas una ventaja competitiva.

2.1.3 BENEFICIOS DE LA INTELIGENCIA DE NEGOCIOS

El uso estratégico de *Business Intelligence* (BI) por parte de las organizaciones conlleva una serie de ventajas fundamentales. En entornos empresariales altamente competitivos, el destacar implica obtener una ventaja distintiva sobre los competidores. Esta distinción se logra mediante la toma de decisiones informadas respaldadas por un sólido conocimiento del negocio, lo cual justifica la creciente adopción de estudios de Inteligencia de Negocios independientemente del tamaño de la empresa.

Los sistemas de BI desplegados en una empresa proporcionan diversos beneficios, entre los cuales se destacan:

- **Optimización de Costos:** La obtención de información relevante a través de BI permite la optimización de procesos y la mejora de la eficiencia operativa a todos los niveles. La unificación de datos en un *Data Warehouse* aumenta la disponibilidad de información precisa, minimizando posibles errores e inconsistencias inherentes a los sistemas OLTP. Además, se agiliza la elaboración de informes de gestión, reduciendo el tiempo y recursos necesarios debido a la facilidad de uso para los usuarios finales.
- **Incremento de Ingresos:** Los sistemas de BI proporcionan conocimientos que permiten una definición más precisa del público objetivo y una segmentación de mercado más acertada. Esta capacidad posibilita la creación de campañas de marketing más específicas para cada grupo de clientes potenciales, lo que se traduce en mejores resultados. Asimismo, al conocer patrones de demanda, se facilita una

distribución más eficaz de las fuerzas de ventas para alcanzar mayores volúmenes de venta.

- **Mejora de la Satisfacción y Fidelización de Clientes:** La optimización de procesos internos en la organización, incluyendo áreas como recursos humanos y operaciones, conlleva una mejora en la atención al cliente. Además, al obtener información sobre las ventas, se pueden tomar decisiones que mantengan los productos siempre disponibles para los clientes, evitando que recurran a los servicios de la competencia.

Dentro de las múltiples aplicaciones de *Business Intelligence* en la industria, se destacan las siguientes áreas principales:

- **Marketing:** Segmentación de clientes para ofrecer promociones específicas basadas en características, intereses y patrones identificados. Análisis de formularios de satisfacción para identificar problemas comunes y solicitudes frecuentes.
- **Ventas:** Identificación de asociaciones entre productos, estrategias promocionales y proyección futura de ventas para una gestión de inventario y fuerzas de ventas más efectiva.
- **Finanzas:** Predicción de tasas de interés y fluctuaciones en mercados de cambio y acciones. Detección de patrones inusuales en transacciones para prevenir fraudes.
- **Banca:** Análisis de cartera y segmentación económica de clientes, identificación de perfiles y beneficios específicos para ciertos clientes.

2.1.4 TENDENCIAS

En el dinámico mundo del *Business Intelligence*, diversas tendencias están dando forma y evolución a esta disciplina fundamental para la toma de decisiones empresariales. Estas tendencias representan cambios significativos en la forma en que las organizaciones interactúan, procesan y utilizan los datos para obtener ventajas competitivas. Entre las principales tendencias actuales se encuentran:

- **Inteligencia Artificial y Machine Learning en BI:** La integración de técnicas de inteligencia artificial y aprendizaje automático está transformando radicalmente el análisis de datos en BI. Esta evolución permite la predicción de patrones, recomendaciones automatizadas y tareas analíticas más avanzadas.
- **Análisis Predictivo y Prescriptivo:** Además del análisis histórico, el análisis predictivo y prescriptivo está ganando terreno. La capacidad de prever eventos futuros y recomendar acciones óptimas en tiempo real se ha vuelto esencial para la toma de decisiones ágiles.
- **BI en la Nube:** La migración hacia soluciones de BI basadas en la nube es una tendencia en ascenso debido a su escalabilidad, accesibilidad y potencia de

procesamiento. Plataformas como *Google Cloud*, *Microsoft Azure* y *Amazon Web Services* están liderando esta transformación.

- **BI Integrado con IoT (Internet de las Cosas):** La convergencia de BI con el Internet de las Cosas está generando grandes volúmenes de datos provenientes de dispositivos conectados. Esta integración se está utilizando para mejorar la gestión de la cadena de suministro, el monitoreo de activos y la optimización de procesos.
- **Autoservicio y Visualización de Datos Avanzada:** La democratización de los datos a través de herramientas de autoservicio está permitiendo que usuarios no técnicos accedan y analicen datos de manera más eficiente. Las herramientas de visualización de datos avanzadas están ofreciendo capacidades interactivas y sofisticadas.
- **Énfasis en la Seguridad y Privacidad de los Datos:** Con la creciente cantidad de datos, la seguridad y privacidad se han convertido en preocupaciones primordiales. Las tendencias se centran en prácticas más robustas de seguridad de datos y cumplimiento de regulaciones.
- **BI Conversacional:** La adopción de interfaces conversacionales, como *chatbots* y asistentes virtuales, para acceder y analizar datos está ganando tracción. Esto permite a los usuarios realizar consultas utilizando lenguaje natural.

2.2 MINERÍA DE DATOS

La minería de datos es un proceso vital para explorar y analizar grandes conjuntos de datos, con el propósito de descubrir patrones, tendencias y relaciones significativas. Esta sección abarcará desde una introducción general hasta las tareas específicas involucradas en este campo.

2.2.1 INTRODUCCIÓN

La Minería de Datos es un pilar esencial en la Inteligencia de Negocios, permitiendo extraer información significativa de conjuntos masivos de datos. En el entorno empresarial actual, donde la generación de datos es exponencial, la habilidad para analizar y comprender esta información es una ventaja competitiva clave.

Dentro del contexto de la Inteligencia de Negocios, la Minería de Datos desvela patrones y relaciones ocultas entre datos. Esta disciplina no solo busca entender el pasado y presente, sino también anticipar y modelar escenarios futuros basados en datos históricos y actuales.

La integración de la Minería de Datos en la Inteligencia de Negocios transforma datos complejos en información estratégica. Esta información se convierte en la base para tomar decisiones informadas, ofreciendo una ventaja competitiva en un mercado global dinámico.

El proceso de Minería de Datos, siguiendo el marco de trabajo flexible de CRISP-DM [Chapman00] (*Cross-Industry Standard Process for Data Mining*), se compone de seis fases interdependientes que permiten a los equipos de datos abordar de manera efectiva problemas empresariales y obtener conocimientos estratégicos.

1. **Comprensión del negocio:** Los científicos de datos inician este proceso identificando los objetivos y el alcance del proyecto, colaborando estrechamente con las partes interesadas para comprender los problemas a abordar, las restricciones del proyecto y el impacto potencial de las soluciones propuestas. Esta etapa establece la base para definir los objetivos de la minería de datos y determinar los recursos necesarios.
2. **Comprensión de los datos:** Tras la identificación de los objetivos empresariales, se inicia un análisis preliminar de los datos disponibles. Se recopilan conjuntos de datos de diversas fuentes, se realiza un informe descriptivo que detalla los tipos y cantidades de datos, y se evalúan los requisitos de hardware y software para el procesamiento de datos. Una vez aprobado el plan, se procede a explorar y verificar los datos, utilizando técnicas estadísticas básicas para evaluar su calidad y seleccionar un conjunto final para la siguiente fase.
3. **Preparación de los datos:** Esta fase ocupa la mayor parte del tiempo debido a que el software de minería de datos requiere datos de alta calidad. Implica procesos como la limpieza, integración y formateo de datos para asegurar su idoneidad y precisión en el modelado.
4. **Modelado de datos:** En esta etapa, los datos preparados se introducen en el software de minería de datos, donde se aplican diversas técnicas y herramientas para analizar los resultados. Los científicos de datos tienen la flexibilidad de elegir entre múltiples métodos de minería de datos y, posteriormente, evaluar la calidad de los resultados a través de pruebas rigurosas. Pueden entrenar modelos de *machine learning* con conjuntos de datos conocidos y utilizar estos modelos para analizar conjuntos de datos desconocidos, ajustando y refinando iterativamente el proceso hasta obtener resultados satisfactorios.
5. **Evaluación:** Una vez que se han creado los modelos, los científicos de datos miden su eficacia en relación con los objetivos empresariales. Comparten los resultados con los analistas de negocio, quienes proporcionan retroalimentación. Esta etapa es crucial para validar si los modelos cumplen con los objetivos establecidos inicialmente. Basándose en los comentarios recibidos, los científicos de datos pueden ajustar los modelos, reconsiderar los objetivos empresariales o revisar los datos nuevamente para obtener un análisis más preciso.
6. **Implementación:** Durante esta fase, los modelos creados se ponen en práctica para generar inteligencia empresarial. Los científicos de datos planifican y supervisan la implementación, instruyendo a otras partes interesadas sobre el uso y las funciones del modelo. Además, se realiza un seguimiento continuo y se mantiene la aplicación de la minería de datos. Los analistas empresariales utilizan esta aplicación para crear informes, compartir resultados con los clientes y mejorar los procesos empresariales.

Las técnicas de minería de datos se basan en varios campos de aprendizaje que se entrelazan, como el análisis estadístico, el *machine learning* y las matemáticas. Además, se pueden realizar diferentes tipos de tareas, las cuales se pueden separar en dos grandes grupos como lo son las tareas descriptivas y las tareas predictiva.

2.2.2 TAREAS DESCRIPTIVAS

Las tareas descriptivas representan una fase fundamental dentro de la Minería de Datos y constituyen una de las dos grandes secciones de este proceso. Su objetivo principal radica en explorar y descubrir patrones, características y relaciones inherentes a los datos, permitiendo una descripción precisa que brinda un entendimiento tanto del comportamiento histórico como del actual.

El proceso de las tareas descriptivas es el siguiente:

1. **Exploración de Datos:** Esta etapa se inicia con la exploración visual de los datos mediante gráficos, tablas y resúmenes estadísticos. Esto posibilita la identificación de patrones, distribuciones, relaciones y valores atípicos, facilitando un análisis exhaustivo posterior.
2. **Análisis Estadístico:** Se emplean métodos estadísticos para comprender la distribución de los datos. Se calculan medidas de tendencia central como la media, mediana, desviación estándar, etc. Además, se evalúa la correlación entre variables para comprender posibles relaciones.
3. **Métodos y Técnicas:** Se aplican métodos y técnicas pertinentes al contexto del problema. Algunas de las más comunes incluyen:
 - a. **Correlaciones y Dependencias:** Centrado en atributos numéricos, nos permite saber si dos o más atributos numéricos están correlacionados linealmente o de algún otro modo.
 - i. **Asociación:** Se utiliza para determinar acontecimientos que ocurren en común en cierto conjunto determinado de datos, nos permite ver como las diferentes columnas pueden estar relacionadas.
 - ii. **Agrupamiento:** También conocido como segmentación o *clustering*, consiste en obtener grupos o conjuntos entre la muestra, de tal manera que los elementos asignados a un mismo grupo sean similares bajo ciertas características.
 - b. **Técnicas de Visualización:** La representación gráfica de datos es esencial para comprender la información de manera intuitiva. Se utilizan herramientas como mapas de calor, histogramas, gráficos de dispersión, entre otros.

- c. **Detección de Anomalías:** Esta técnica es útil para detectar comportamientos anómalos como fraudes, fallas, intrusos, etc. Es importante percatarse de la existencia de estos para que los resultados que obtengamos sean valiosos.

2.2.3 TAREAS PREDICTIVAS

Las tareas predictivas son una parte esencial de la minería de datos y se centran en utilizar datos históricos para anticipar tendencias, patrones o resultados futuros. Estas tareas permiten la construcción de modelos predictivos que estiman o clasifican datos desconocidos basándose en la información histórica.

El proceso de las tareas predictivas consta de varios pasos fundamentales:

1. **Selección de Atributos Relevantes:** Inicia la creación de un modelo predictivo identificando y seleccionando las variables más significativas y pertinentes para el análisis. Este proceso se basa en comprender qué atributos tienen una mayor influencia en los resultados previstos.
2. **Elección del Modelo Predictivo:** Seleccionar el modelo adecuado es crucial y requiere comprender el contexto del problema y las expectativas de resultados. Algunas técnicas comunes incluyen:
 - a. **Clasificación:** El objetivo es encontrar una función lambda que nos permita obtener un resultado único a partir de un input correspondiente a los datos de entrada. Los modelos de clasificación pueden estar basados en diferentes técnicas como lo son los árboles de decisión, métodos bayesianos, métodos basados en casos y vecindad, redes neuronales, entre otros.
 - b. **Categorización:** En este caso no buscamos un valor de salida como resultado, sino más bien una correspondencia, la función en este caso puede asignar varias categorías a un mismo input.
 - c. **Regresión Lineal:** Basado en una regresión estadística, las muestras son correspondencias entre dos conjuntos de tipo numérico, el primer conjunto corresponde a los datos independientes y el dato a predecir, conocido también como variable dependiente. Esta técnica nos permite obtener una función matemática que nos entregue el valor de la variable dependiente dado que le entregamos las variables independientes.
 - d. **Regresión Logística:** Similar a la regresión lineal es basado en regresión estadística, la diferencia consta en el resultado obtenido, pues la regresión logística nos entrega una probabilidad, de 0 a 1, la cual nos responderá si el output es verdadero o falso.
 - e. **Pronósticos:** Modelo estadístico que buscan obtener información futura, a diferencia de los modelos visto con anterioridad este se basa en series temporales, por lo cual es importante que los datos que queremos predecir posean la dimensión temporal para utilizar este modelo.

3. **Entrenamiento del Modelo:** Se lleva a cabo el entrenamiento del modelo escogido, para esto se selecciona una parte del conjunto de datos donde se le entregan datos históricos junto con su valor conocido, a esto se le llaman datos etiquetados, es con esta información que el modelo podrá predecir los datos nuevos entregados. Un paso importante en esta sección es el ajuste de los parámetros del modelo, ya que esto nos permitirá una personalización de los resultados obtenidos, como se mencionó anteriormente cada problema es diferente al anterior, por lo cual es importante utilizar diferentes parámetros para encontrar el mejor modelo posible.
4. **Validación del Modelo:** Tras el entrenamiento, se evalúa la eficacia del modelo utilizando datos de prueba independientes. Este proceso se enfoca en asegurar que el modelo tenga la capacidad de generalizar y predecir con precisión datos nuevos y no vistos previamente, verificando su capacidad para trabajar con información no utilizada en su fase de entrenamiento.
5. **Predicción y Evaluación:** Finalmente, se aplican los modelos a datos desconocidos para realizar predicciones reales. Se evalúa la precisión del modelo al comparar estas predicciones con los resultados reales, permitiendo ajustar y refinar el modelo según sea necesario, con un enfoque más práctico en la aplicación efectiva del modelo en situaciones del mundo real.

2.2.4 ANÁLISIS DE SERIES TEMPORALES

Una serie temporal es un conjunto de observaciones que son obtenidas midiendo una variable única de manera constante a través del tiempo. Es importante que las mediciones sean regulares y equidistantes cronológicamente para que posea un valor el ser estudiadas. La forma más sencilla de analizar una serie temporal es mediante un gráfico ya que nos permite observar diferentes características de nuestros datos como lo pueden ser la dispersión, los valores atípicos, frecuencia, tendencia, estacionalidad, etc.

El principal objetivo de un análisis de series temporales es obtener un modelo y pronóstico de eventos futuros utilizando la información histórica, las características esenciales para lograr esto son las tendencias, estacionalidad y ciclos recurrentes de los datos históricos. Si los datos no son estacionarios entonces no podremos realizar un análisis correcto.

Algunos de los modelos más utilizados son los siguientes:

- **Modelo AR (Auto-Regresivo):** Este modelo asume que el valor de la serie temporal en un momento dado está relacionado con los valores anteriores. La idea principal es predecir el valor actual utilizando una combinación lineal de sus valores pasados. En un modelo AR(p), se utiliza la información de los últimos 'p' períodos para hacer la predicción.
- **Modelo MA (Media Móvil):** En contraste con AR, este modelo predice el valor actual basándose en el error residual de las predicciones anteriores. En un modelo MA(q), se utiliza la información de los últimos 'q' errores para hacer la predicción.

- **Modelo ARMA (Auto Regresivo Media Móvil):** Es una combinación de los modelos AR y MA. $ARMA(p,q)$ utiliza tanto la información de los valores pasados como los errores residuales para hacer la predicción.
- **Modelo ARIMA (Auto Regresivo Integrado Media Móvil):** Este modelo incorpora la diferenciación para hacer estacionaria una serie temporal que no lo es. $ARIMA(p,d,q)$ incluye los componentes AR, MA y una etapa de diferenciación de orden 'd' para convertir la serie en estacionaria.
- **SARIMA (Seasonal ARIMA):** Asume patrones estacionales en los datos. Incorpora los componentes de ARIMA junto con componentes estacionales. $SARIMA(p,d,q)(P,D,Q)s$ incluye la estacionalidad (P,D,Q) con un periodo 's'.
- **SARIMAX:** Es una extensión de SARIMA que incorpora covariables exógenas. Además de los componentes de SARIMA, permite integrar variables externas para mejorar las predicciones.

2.2.5 ANÁLISIS GEOTEMPORAL

El análisis geo-temporal es una técnica que puede resultar muy útil según el contexto del problema, permite comprender como evolucionan los datos en el tiempo y el espacio. Esta integración nos permite que las organizaciones obtengan una visión más completa y detallada de las tendencias de los datos que se quieren analizar.

En el ámbito de BI, la visualización de datos ventas mediante análisis geoespaciales y temporales revela patrones de compra en diferentes ubicaciones a lo largo del tiempo. Esto facilita la identificación de áreas de alto rendimiento y oportunidades de crecimiento, así como la optimización de la logística y la gestión de inventario.

Al aplicar análisis geo-temporales a los datos de ventas, es posible prever la demanda futura en diversas ubicaciones, lo que resulta clave para la planificación de inventario y el diseño de estrategias de marketing personalizadas para cada región.

La combinación de series temporales y datos geoespaciales proporciona una visión detallada de cómo las tendencias de ventas evolucionan en el tiempo y el espacio. Esto permite identificar patrones de compra específicos por región y comprender cómo cambian con el tiempo, lo que resulta fundamental para la predicción de la demanda y la toma de decisiones informadas.

CAPÍTULO 3: PROPUESTA DE SOLUCIÓN

En este capítulo, se presenta una metodología diseñada para cumplir con los objetivos previamente establecidos. El trabajo se enfoca en dos etapas principalmente, siendo la limpieza de los datos la primera de estas. El enfoque del trabajo se divide en dos etapas clave, siendo la primera la limpieza de los datos. El *dataset* en cuestión corresponde a las ventas de insumos tecnológicos asociados a proyectos de Dell Technologies en Latinoamérica durante el periodo 2012-2019. Para alcanzar este propósito, se lleva a cabo un análisis exploratorio de los datos, extrayendo contexto, significado de las columnas, posibles valores, entre otros. Posteriormente, se realiza una limpieza de los datos para preprocesarlos, permitiendo así su utilización en análisis subsiguientes.

La segunda etapa del proyecto se centra en el aprovechamiento de los datos para obtener información relevante. Se llevará a cabo una proyección geo-temporal en múltiples dimensiones de las ventas de la empresa en Latinoamérica. Esta proyección tiene como objetivo ser utilizada en periodos futuros para mejorar la toma de decisiones en las estrategias organizacionales del área de ventas, aprovechando los valores proporcionados por el modelo.

Para concluir, se generará un informe en Power BI que facilite la resolución de tareas descriptivas y el análisis de diferentes KPIs. Esto permitirá un análisis más detallado de los procesos clave de la empresa, mejorando así la toma de decisiones.

Con estos enfoques, se busca abordar el problema de los errores existentes en la base de datos de ventas, permitiendo un análisis en múltiples áreas y solucionando problemas como pérdidas económicas, errores en la toma de decisiones, disminución de clientes, entre otros. Esto contribuirá a que Dell Technologies continúe manteniéndose como un referente en su área de disciplina.

3.1 LIMPIEZA DE DATOS

Para iniciar la tarea de limpieza de datos, es crucial seleccionar la herramienta adecuada. Existen múltiples herramientas y software para realizar el proceso de ETL, abarcando la etapa de "Transformación", donde se personalizan los datos. En este caso, se opta por el lenguaje de programación Python debido a su versatilidad y las numerosas librerías que agilizan y personalizan cada etapa de limpieza.

3.1.1 DATASET

Comencemos analizando cada columna del *dataset* para comprender el contexto de los datos y trazar una ruta de limpieza que cumpla con nuestros objetivos.

El archivo por analizar es un documento CSV (*Comma Separated Values*) llamado "LATAM Project Equipment List by Date". Consta de 734,712 registros y 23 columnas. La información detallada sobre estas columnas se presenta en la siguiente tabla:

Tabla 3: Columnas LATAM Project Equipment List by Date.

Fuente: Elaboración Propia.

Nombre	Tipo	Descripción
Countryname	VARCHAR	Nombre del país.
Owner	VARCHAR	Usuario del empleado a cargo del proyecto.
Plannerfirstname	VARCHAR	Nombre del empleado a cargo del proyecto.
Plannersurname	VARCHAR	Apellido del empleado a cargo del proyecto.
Project Id	INTEGER	ID del proyecto.
Project Name	VARCHAR	Nombre del proyecto.
Proj Created Dt	DATE	Fecha de creación del proyecto.
Customer Number	INTEGER	Número de cliente.
Project State	VARCHAR	Estado del proyecto.
Projectworkflowstatename	VARCHAR	Nombre estado flujo de trabajo del proyecto.
Order Num	INTEGER	Número de orden.
Task Order Date	DATE	Fecha del pedido.
DOMs Order Status	VARCHAR	Estado DOMs de la orden.
SKU Number	VARCHAR	Código del producto.
SKU Description	VARCHAR	Descripción del producto.
Accounting Class	VARCHAR	Categoría contable del producto.
Item Quantity	INTEGER	Cantidad de productos en la orden.
SKU USD Unit Price	FLOAT	Precio del producto por unidad en dólares estadounidense.

Total USD Price	FLOAT	Precio final del pedido en dólares estadounidense.
Total USD Margin	FLOAT	Beneficio neto total del pedido en dólares estadounidense.
Total Order Price	FLOAT	Precio final del pedido en moneda local.
Total Order Margin	FLOAT	Beneficio neto total del pedido en moneda local.
SKU Local Sell Price	FLOAT	Precio del producto por unidad en moneda local.

Cada registro del *dataset* corresponde a la venta de un producto asociado a una boleta, la cual se identifica con la columna **“Order Num”**, por lo cual una misma boleta puede contener múltiples registros, si se compra un mismo producto más de una vez en la misma boleta este queda reflejado en la columna **“Item Quantity”**, de esta manera corresponde a un único registro.

3.1.2 CARGA DE DATOS

El primer paso en la preparación y limpieza de datos es la carga del *dataset* en nuestra herramienta de trabajo, que en este caso es Python. No obstante, durante este proceso, se identifica un error debido a la presencia de registros con más campos de los especificados anteriormente. Para observar con mayor profundidad por qué ocurre esto se revisa el archivo como texto plano, donde se puede observar el formato de este, también se verifica con Excel como son separados los campos, aquí se puede observar que existen registros donde el campo **SKU Description** es seccionado derivando en un desplazamiento hacia la derecha en el resto de los campos.

La razón del desplazamiento se debe a que cada campo en el CSV es rodeado por comillas dobles y debido a que en el campo **SKU Description** existen algunos registros donde se incluyen dimensiones expresadas en pulgadas, generando confusiones en la estructura del CSV. Por ejemplo, registros como *"Chassis with up to 8, 3.5" Hot Plug Hard Drives, Tower Configuration"* provocan que "Tower Configuration" sea considerado el siguiente campo debido a las comillas dobles que cierran el *string* después de 3.5".

La razón subyacente a este comportamiento es que, en los CSV, las comas se utilizan como separadores, pero si están dentro de un *string* rodeado por comillas dobles, se consideran parte del campo y no como separadores. En situaciones como *"Dell Limited Hardware Warranty Plus Service, Initial Year"*, las comillas dobles encierran el *string* completo, incluidas las comas internas. Sin embargo, en el caso que analizamos al comienzo donde el

valor del campo es *"Chassis with up to 8, 3.5" Hot Plug Hard Drives, Tower Configuration"* las comillas dobles del **3.5"** cierran el *string* y la coma que viene luego de *Hard Drives* funciona como separador de valores.

Una vez entendido el problema se debe solucionar como trabajar con los separadores o modificar el archivo, de forma que podamos cargar el *dataset* en Python sin que existan problemas.

Podemos notar que *"Chassis with up to 8, 3.5" Hot Plug Hard Drives, Tower Configuration"* debiese corresponder a un único campo, esto debido a que luego de las comillas dobles que siguen a 3.5 no prosigue una coma seguida de otra comilla doble, es decir, si el campo fuese *"Chassis with up to 8, 3.5","Hot Plug Hard Drives, Tower Configuration"*, se podría considerar como dos campos diferentes.

Utilizaremos esta lógica para separar los campos, de esta manera usaremos como separador de valores la cadena `' ',' '` es decir, una coma encerrada en comillas dobles. Esta elección permite evitar los problemas en el campo **SKU Description**. Sin embargo, surge un nuevo inconveniente ya que el primer campo posee como valor **"CAMPO_1"**, y el último campo posee como valor **"CAMPO_N"**, es decir, el primer y último carácter de cada fila del CSV corresponde a `"` y `"`, respectivamente. Para solucionar este problema deberemos quitar el primer y último carácter de cada fila del CSV.

Además de este problema de carga tenemos unos registros con casos donde dos filas del CSV corresponden al mismo registro. A continuación, podemos observar un caso en el cual ocurre:

Tabla 4: Registro error.
Fuente: Elaboración Propia.

Fila	Valor
511828	"PERU","Danisa_Saldivia","Danisa","Saldivia","367767","IBM DEL PERU S.A.C-AB6X3M01","2018-06-11 17:26:21.0","108001693","COMPLETED","Project Closure Completed","571531789","2018-06-08 18:33:24.0","IN","321-BCSM","Chassis with up to 8 x 2.5" SAS/SATA Hard Drives
511829	for 2CPU Configuration","Bse HW Dell Syst","6","23.32","139.92","-592.866","139.92","-592.866","78.82"

Para solucionar esto debemos verificar si la línea que estamos leyendo posee una comilla doble al final, ya que esto significaría que el registro está completo, en el caso de que no sea así deberemos guardar la información y concatenarla con la línea que continúa. Con esto en mente el código que permite cargar los datos de manera correcta en Python se puede observar en la Figura 8.

```
1 file = open('LATAM Project Equipment List by Date.csv',
2             encoding="utf8")
3 file2 = open('TESTING.csv', 'w',
4             encoding="utf8")
5
6 merged_line = ''
7
8 for line in file:
9     if line.strip().endswith(''):
10         if merged_line: # Se concatena con el anterior
11             line = merged_line + ' ' + line.strip()[:-1] + '\n'
12             merged_line = ''
13         else: # Línea normal
14             line = line.strip()[1:-1] + '\n'
15             file2.write(line)
16         else:
17             # Línea que debe seguir
18             merged_line = line.strip()[1:]
19 file.close()
20 file2.close()
21
22 df = pd.read_csv('TESTING.csv',
23                 index_col=None,
24                 keep_default_na=True,
25                 on_bad_lines="error",
26                 sep=';',
27                 engine="python",
28                 encoding="utf8")
```

Figura 8: Carga Dataset

Fuente: Elaboración Propia

Para verificar si la carga se realizó de manera exitosa procederemos a verificar el registro que poseía problemas con el campo **SKU Description**, en la Figura 9 se puede observar que la carga se realiza de manera correcta y las modificaciones resultaron exitosas debido a que todas las columnas muestran sus valores correctos. Se incluyen los valores de las otras columnas para facilitar el análisis de los datos contenidos en el *dataset*.

Un ejemplo de registro para ver como se cargó el dataset

```
df = frames['dell']
df.iloc[64]
```

[13] Python

Countryname	ARGENTINA
Owner	Mun_Yee_Chow
Plannerfirstname	Mun Yee
Plannersurname	Chow
Project Id	200086.0
Project Name	LEVEL3-638101791
Proj Created Dt	2016-07-29 17:16:55.0
Customer Number	100205402
Project State	DRAFT
Projectworkflowstatename	Customer Declined
Order Num	524669199
Task Order Date	2016-04-15 13:14:06.0
DOMs Order Status	IN
SKU Number	350-BBEL
SKU Description	Chassis with up to 8, 2.5" Hard Drives, Softwa...
Accounting Class	Bse HW Dell Syst
Item Quantity	2.0
SKU USD Unit Price	0.0
Total USD Price	0.0
Total USD Margin	-304.158
Total Order Price	0.0
Total Order Margin	-304.158
SKU Local Sell Price	0.0
Name: 64, dtype: object	

Figura 9: Registro dataset

Fuente: Elaboración Propia

3.1.3 LIMPIEZA DE DATOS

Ahora que hemos realizado la carga tenemos que realizar la limpieza y estandarización de los datos, el primer paso que realizaremos será eliminar las columnas que posean información personal, datos sensibles y aquellas que no entreguen una información adicional al conjunto de datos.

La información personal corresponde a los nombres de clientes y empleados que se encuentran en los registros. Dentro de las columnas que contienen información personal tenemos las siguientes:

- **“Plannerfirstname”**: Nombre del empleado de Dell Technologies a cargo del proyecto.
- **“Plannersurname”**: Apellido del empleado de Dell Technologies a cargo del proyecto.
- **“Project Name”**: Nombre del cliente que está a cargo del proyecto.

Dentro de las columnas que no entregan información adicional definiremos aquellas que están asociadas al valor de las ventas en su respectiva moneda local, esto debido a que existen columnas con la misma información, pero en dólares, lo cual permite que el análisis sea global, dentro de las columnas que entran en este tipo están las siguientes:

- **“Total Order Price”**: Precio total de la venta en la moneda local.
- **“Total Order Margin”**: Ganancia total de la venta en la moneda local.
- **“SKU Local Sell Price”**: Precio unitario del producto en la moneda local.

Una vez hemos eliminado las columnas que no utilizaremos procederemos a modificar ciertos errores que podemos encontrar en los diferentes campos que posee el dataset, partiremos por los campos relacionados a fechas, como lo son **“Task Order Date”** y **“Proj Created Dt”**.

El primer paso al trabajar con campos del tipo fecha corresponde a modificar el tipo de dato en el *dataframe*, para realizar esto utilizaremos el comando **to_datetime()** que ofrece la librería pandas, una vez realizado esto podemos verificar si existen campos con errores en las fechas. En el caso de **“Proj Created Dt”**, no posee ningún error con nulos o valores de fechas erróneos, por lo cual no es necesario realizar un cambio, por otra parte **“Task Order Date”** posee dos problemas, el primero es que posee valores erróneos como lo es “2029-07-29 08:19:20”, el cual posee como año de venta el 2029, lo que no tiene lógica y probablemente sea un error al ingresar el dato, además posee 4 registros donde el valor es nulo.

Ver registros con errores en la fecha

```
# Supongamos que tu dataframe se llama df y la columna de fecha es 'fecha_columna'
df['Task Order Date'] = pd.to_datetime(df['Task Order Date'], errors='coerce')
df['Proj Created Dt'] = pd.to_datetime(df['Proj Created Dt'], errors='coerce')

# Crea una máscara booleana para identificar las filas con valores NaT
mask_nat = df['Task Order Date'].isna()

# Imprime o realiza otras operaciones con las filas que contienen NaT
df[mask_nat]
```

✓ 0.0s

	Countryname	Owner	Project Id	Proj Created Dt	Customer Number	Project State	Projectworkflowstatename	Order Num	Task Order Date
429392	BRAZIL	rogerio_garcia	242729	2017-01-06 16:21:10	110601177	COMPLETED	Project Closure Completed	543067441	NaT
429393	BRAZIL	rogerio_garcia	242729	2017-01-06 16:21:10	110601177	COMPLETED	Project Closure Completed	543207294	NaT
617409	BRAZIL	luiz_christofoletti	219719	2016-10-14 22:33:05	108974211	COMPLETED	Project Closure Completed	542784087	NaT
617410	BRAZIL	luiz_christofoletti	219719	2016-10-14 22:33:05	108974211	COMPLETED	Project Closure Completed	542784087	NaT

```
print('Min date from dataset: %s' % df['Task Order Date'].min().date())
print('Max date from dataset: %s' % df['Task Order Date'].max().date())
```

✓ 0.0s

Min date from dataset: 2012-03-09
Max date from dataset: 2029-07-29

Figura 10: Carga de datos en Python

Fuente: Elaboración propia

Para solucionar esto se verifica a qué proyecto están relacionadas las ventas de 2029, encontrando que todos corresponden al proyecto con valor **“Project Id” = 207373**, mientras que los valores nulos corresponden a dos proyectos diferentes. Analizando las ventas relacionadas al proyecto 207373 se pueden observar que las compras fueron realizadas todas en el año 2016, por lo cual podemos asumir que las ventas del 2029 corresponden a un error al ingresar los datos y debiesen corresponder al año 2016, este cambio se realiza buscando las ventas del 2029 del proyecto 2016 y modificando su campo a “2016-07-29 08:19:20”. Una vez hecha esta modificación se puede observar que los rangos de venta corresponden desde “2012-03-09” hasta “2019-06-28”.

La distribución de las fechas de venta también es algo que se debe considerar, debido a que nuestro trabajo consiste en realizar un *forecasting* de las ventas debemos poseer registros de ventas constantes en el tiempo, y como podemos observar en la figura 11 los años 2012 y 2013 poseen muy pocos registros, lo que podría afectar de manera negativa a nuestro análisis, es por esto por lo que los datos a utilizar corresponderán desde el 2014 hasta el 2019.

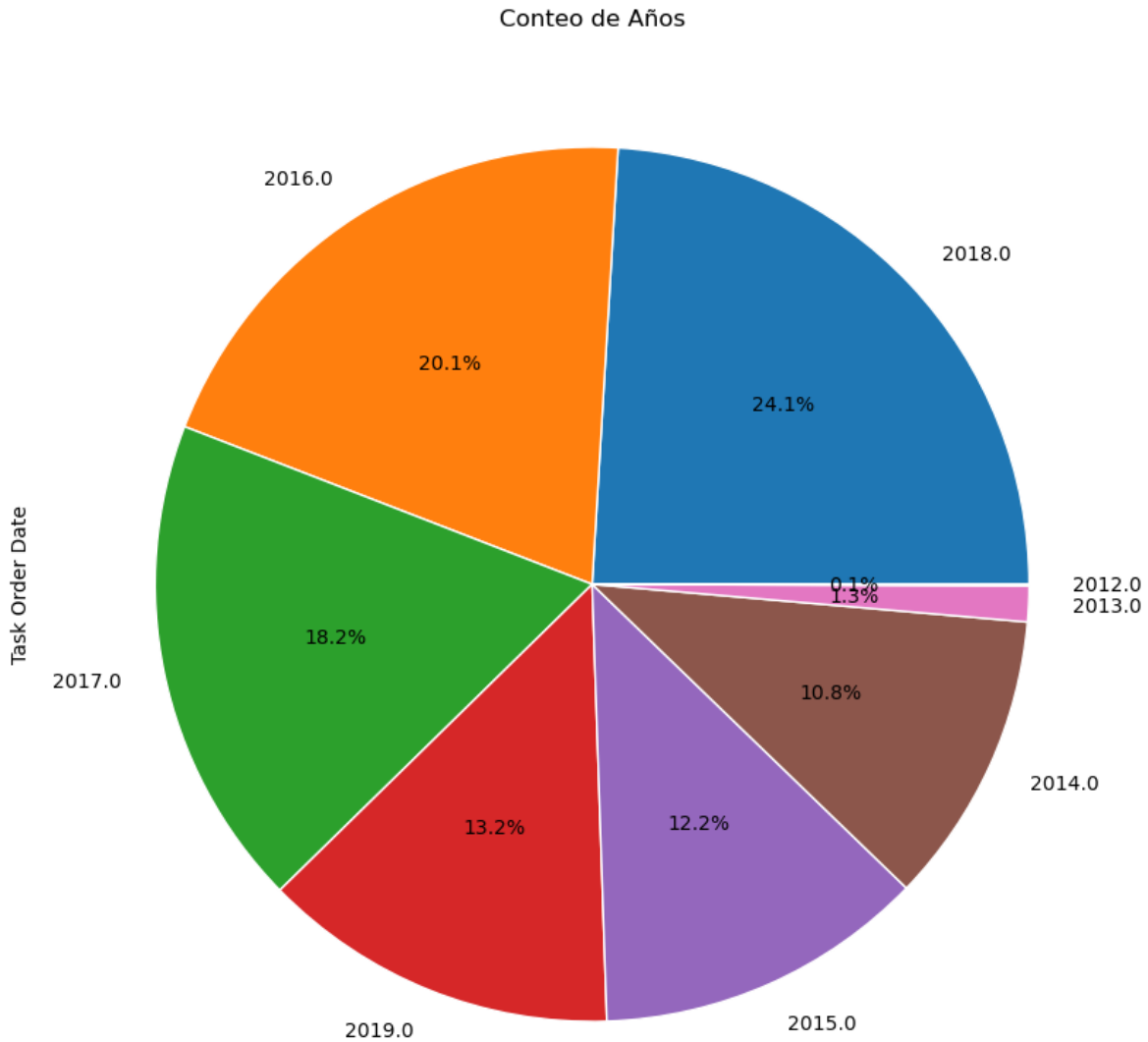


Figura 11: Distribución de ventas por año

Fuente: Elaboración propia

Por otra parte, aquellos registros que no poseían valores en el campo de fecha de venta, se puede observar un rango de fechas donde se realizaron ventas para el proyecto asociado, sin embargo, como no sabemos con exactitud el valor de la venta y debido al bajo número de registros que poseen errores estos serán eliminados del *dataset*.

Una vez hemos trabajado con los campos relacionados a las fechas verificaremos si existen otros campos que posean problemas con los valores que posee.

Antes de comenzar el análisis de los errores de cada columna, hay que analizar las diferencias entre eliminar un registro completo y eliminar el valor de la columna, recordemos que este trabajo está enfocado en realizar un análisis geo-temporal de la

demanda, el cual se puede realizar en múltiples dimensiones adicionales, bajo esta primicia, las columnas fundamentales para nuestro trabajo son aquellas temporales que nos definen cuando se realizó la venta y las columnas de hechos, especialmente el margen obtenido, si otras columnas como lo pueden ser el número del cliente, el producto, el país, entre otras, poseen un valor erróneo se puede eliminar el valor de la columna y que el registro siga siendo valioso, en estos casos simplemente eliminaremos el valor de la columna. Sin embargo, en caso de que el error se encuentre en la columna temporal o en las de ventas y no se pueda realizar alguna técnica que nos permita obtener el dato, en ese caso el registro puede ser eliminado sin perder información importante.

Con las columnas **“Project State”**, **“DOMs Order Status”**, **“Project Id”**, **“Item Quantity”** no poseen errores en sus valores, por lo cual no hay realizar ninguna modificación en estos campos.

Por el lado del campo **“Countryname”**, podemos encontrar valores nulos, en específico 652 registros no poseen el país asociado, dado que los **“Owner”** pueden estar asociados a diferentes países y que el resto de ventas del mismo **“Project Id”** también poseen como valor de **“Countryname”** nulo, no podemos completar el campo con el valor verdadero, y esto afectará al momento de realizar el análisis enfocado en los países, de todas maneras la cantidad de registros con problemas en el campo es mínimo respecto al total de registros, y esto afectará netamente a un valor más certero al analizar las ventas por países, por lo cual podemos dejar estos registros sin problemas.

De la misma manera que existen valores nulos en **“Countryname”**, también se pueden encontrar valores nulos en el campo **“Owner”**, en este caso la cantidad de registros es mayor con 15.961 casos. Los registros pertenecen a múltiples países y proyectos, similar al caso anterior no tenemos forma de conocer a que Owner le pertenecen las ventas, pero no es un gran problema y podemos realizar el trabajo con estos datos faltantes.

Para la columna **“Customer Number”**, se encuentran registros con un tipo de dato no numérico, para trabajar con estos casos se observó la cantidad de registros que no corresponden a dígitos obteniéndose 778 casos donde el valor del campo no era numérico, siendo 4 valores únicos que puede tomar, estos valores son ['AMER_EDT_FY16Q4', '105796108-8585', '3696-107660172', '11-14136094'], para solucionar esto se modificarán los valores de estos registros y se le asignará un valor nulo, el total de registros con estos problemas son 778.

Para el campo **“Projectworkflowstatename”** se pueden obtener 36.112 registros que poseen valor nulo, la importancia del campo se deriva en los análisis de flujo de trabajo, donde se puede observar la cantidad de proyectos que fueron abandonados, donde existen cuellos de botella y otros análisis. En este caso la cantidad de registros que poseen problemas es mayor a cualquier otra columna, sin embargo, no afecta al trabajo principal de este trabajo, por lo cual estos registros serán mantenidos para el análisis.

Con la columna **“Orden Num”**, se realiza un trabajo similar al de **“Customer Number”**, se obtienen los registros donde el campo no es numérico, obteniendo 4.303 registros, aquellos

valores que no son numéricos son reemplazados por un valor nulo, esto puede afectar para ver una orden en específico, pero para análisis en grandes cantidades no genera problemas.

En el caso del campo **“SKU Number”** podemos observar que existen 3 registros con valor nulo, estos serán eliminados ya que no entrega ninguna información útil, son registros donde los campos:

- “SKU Number”
- “SKU Description”
- “Accounting Class”
- “Item Quantity”
- “SKU USD Unit Price”
- “Total USD Price”
- “Total USD Margin”

Poseen un valor nulo, además de esto existen ciertos registros que no respetan el formato del campo. Los registros de este campo poseen la siguiente estructura: “XXX-YYYY”, donde “X” es cualquier dígito e “Y” es cualquier carácter alfanumérico, un ejemplo de valor de este campo lo es “340-AKPJ” o “341-9096”, sin embargo, existen unos valores con un formato diferente, como lo es “A9918009”. Para trabajar con estos datos se realizará una transformación para estandarizar los valores. En este caso se eliminará el primer carácter, y se insertará un guion luego de los primeros tres dígitos, logrando la siguiente transformación “A9918009” – “991-8009”, la cual respeta el formato del resto de los registros.

En el caso de **“Accounting Class”** se observa que existen demasiadas categorías que pueden poseer los productos vendidos, siendo un total de 60 posibles valores, esto provoca que realizar análisis respecto a el campo se obtenga una baja muestra de valores. Debido a esta se generará un nuevo campo que categorice el producto el cual estará basado en el valor que posee el registro en el campo “Accounting Class”, el nombre del nuevo campo es **“Category”** y sus posibles valores serán “Hardware”, “Software & License”, “Warranty”, “Support & Services”, “Information”, “Other”.

Estas categorías serán un nivel superior respecto a las asignadas en “Accounting Class”, la forma en que se distribuyeron los valores se realizó viendo los productos que pertenecían a cada uno de los valores del campo y asignándole manualmente una categoría a cada valor, finalmente la relación entre el nuevo campo llamado **“Category”** y el campo desde el que se obtuvo la información “Accounting Class” es la siguiente:

Category	Accounting Class
Hardware	"Bse HW Dell Syst", "Base EMC Purple", "SnP DB HW CLIENT", "SnP DB HW ENTERPRISE", "SnP DB HW DISPLAY", "Freight", "COIP VMWare"
Software & License	"Base SW", "VMWare OEM", "SnP Kit ESG SW VL Amort", "SVC PRM SW SUP", "Base SW Gross_Net Review", "Base Software", "VMWare OEM PROD", "SnP SW VL Amort"
Warranty	"Base Syst WAR nRA", "SVC UpsWAR NBD", "SVC UpsWAR SAME DAY", "SVC UpsWAR nRA", "Base SW Gross_Net Review", "Base SW Amort", "SVC APS UpsWAR nRA"
Support & Services	"SVC Msc SUP", "SVC PESS", "SVC EDT CNSLT", "SVC EIS SVCs", "SVC CFI", "SFTSUPPORT", "SVC TCH SUP", "SVC Apps/BPO", "SVC SW SUP", "SVC CMPLT CARE", "SVC Msc SUP", "EMC SW", "SVC DOC", "EMC STRG SW", "COIP VMW SW SUP", "SVC DPS CNSLTG", "SVC OTH CNSLTG", "SVC APS Msc SUP", "SVC APS Msc SUP", "SVC ARS Fees", "SVC GTS SVC"
Information	"Information", "SVC CLNT INSTLN", "Bse HW CFI Fact", "SVC OSpt SVCs", "SFTTRAINING", "SVC DES TRNG", "COIP SERVER HW SUP", "SnP HW CFI NnFact"
Other	Todas aquellas no poseen "Accounting Class" o que el valor del campo no sea ninguno de los anteriormente nombrados.

Para el campo **"SKU Description"** encontramos que un valor de **"SKU Number"** podía tener más de algún valor en el campo **"SKU Description"**, esto se puede deber a que diferentes países utilizan SKU's distintos en el mismo producto, errores al momento de realizar la carga, cambios en la descripción del producto con los años, entre otros. La finalidad en este campo es obtener una relación de uno a uno entre los campos **"SKU Number"** y **"SKU Description"**, para realizar esto se obtendrán todos los valores que posean dos destinos, para ambas direcciones.

Partimos con los SKU's que poseen dos o más descripciones, se obtendrán todas las combinaciones SKU Number – SKU Description y todos los SKU Number que se encuentren repetidos serán aquellos que posean dos o más SKU Description, esta información se almacenará en un archivo CSV donde además añadiremos las veces que aparece esa combinación en nuestro conjunto de datos, el número de países que posee esta combinación y la última fecha en el que se utilizó. Utilizaremos esta información para

realizar la actualización del valor del campo, la cantidad de SKU Number con más de una descripción son 3.201, un ejemplo de la información almacenada es la siguiente:

Tabla 5: Registro SKU Number duplicado.

Fuente: Elaboración Propia.

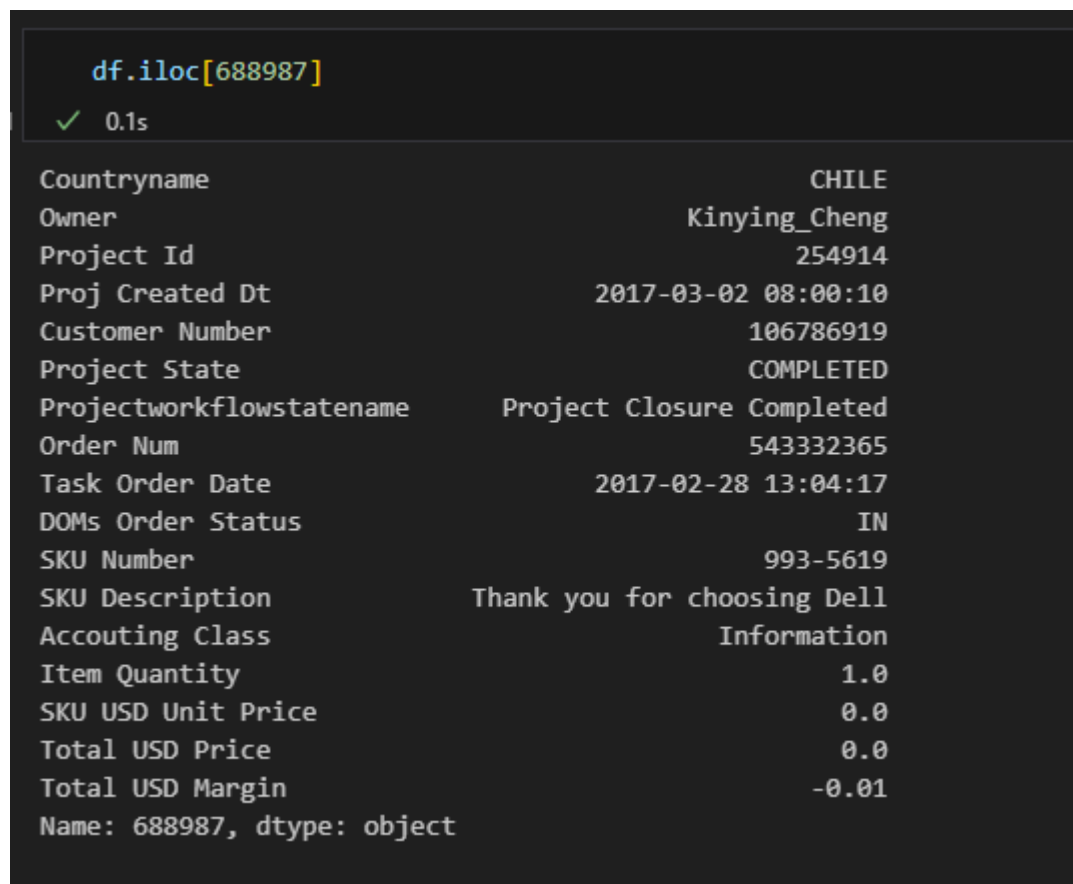
SKU Number	SKU Description	TotalCount	NumCountries	MaxDate
330-6397	PDU PWR Cord,L6-20P to C19,20A,250V,3.7m	46	1	2014-07-07 23:00:00
330-6397	PDU Power Cord L6-20P to C19, 20A, 250V, 3.7m	2	1	2014-09-29 13:16:37

Con esta información podemos tomar decisiones sobre qué descripción conservaremos, hay tres factores a tomar en cuenta, la cantidad de veces que ha aparecido la combinación, el número de países y la última fecha. El algoritmo que utilizaremos será ver si existe una gran diferencia entre la cantidad de veces que se ha utilizado la descripción, si la cantidad de veces es por lo menos tres veces mayor será esta la que usaremos, aplicando el mismo criterio luego con el número de países, finalmente si los valores son parecidos se decidirá con el valor más nuevo. Una vez realizado el cambio podremos modificar las descripciones de aquellos SKU Number que poseían más de dos descripciones. Ahora se realizará algo similar, pero en sentido inverso, si una descripción posee dos o más SKU Number hablamos del mismo producto con dos códigos diferentes, por lo cual realizaremos un proceso similar al anterior con uno que otro detalle, principalmente el hecho de considerar dos descripciones iguales en el caso de que una posea algunos caracteres en mayúsculas y la otra descripción no. Se encontraron 1.994 descripciones que poseen más de un SKU Number. Al finalizar esta etapa el resultado es un conjunto de datos cuyos campos “SKU Number” y “SKU Description” poseen una relación 1 a 1.

Con las últimas columnas, relacionadas al precio de venta del producto tenemos las columnas “SKU USD Unit Price”, “Total USD Price” y “Total USD Margin”, existen 11 registros que no poseen información en ninguna de las columnas mencionadas y 14 registros sin información en la columna Total USD Margin, es decir, hay tres registros que poseen información en las otras columnas, pero no en Total USD Margin. Sobre los primeros once registros se verificará si los productos vendidos han sido vendidos en otras ocasiones, y obtendremos la media de la columna “SKU USD Unit Price”. De esta manera podremos completar la información faltante con un valor cercano al real. Una vez hemos completado el precio unitario del producto, podremos completar el campo “Total USD Price”, ya que poseemos el precio unitario y las unidades vendidas, por lo cual, al multiplicar ambos valores podemos obtener el valor del campo “Total USD Price”. De manera similar a las anteriores completaremos el campo “Total USD Margin”, para esto se obtendrá el promedio

del campo “Total USD Margin” dividido por “Item Quantity”, obteniendo una media del margen por unidad, esto se utilizará para completar el valor del campo “Total USD Margin” en aquellos registros donde el valor sea nulo, el valor del campo estará conformado por la media del margen por unidad multiplicado por la cantidad de unidades vendidas.

Continuando con la etapa de limpieza de datos, se eliminaron los registros que no aportaban valor alguno, estos registros corresponden a información de agradecimiento hacia el cliente por preferir a Dell Technologies. Un ejemplo de registro de los comentados es el siguiente.



```
df.iloc[688987]
```

✓ 0.1s

Countryname	CHILE
Owner	Kinying_Cheng
Project Id	254914
Proj Created Dt	2017-03-02 08:00:10
Customer Number	106786919
Project State	COMPLETED
Projectworkflowstatename	Project Closure Completed
Order Num	543332365
Task Order Date	2017-02-28 13:04:17
DOMs Order Status	IN
SKU Number	993-5619
SKU Description	Thank you for choosing Dell
Accounting Class	Information
Item Quantity	1.0
SKU USD Unit Price	0.0
Total USD Price	0.0
Total USD Margin	-0.01
Name: 688987, dtype: object	

Figura 12: Registro agradecimiento

Fuente: Elaboración Propia

El “SKU Description” del registro de ejemplo corresponde a “Thank you choosing Dell”, y además no agrega un valor de venta, por lo cual solo es un gasto de memoria, todos los registros de este tipo serán eliminados. También se modificaron los tipos de datos del *dataset* de modo que coincidieran con el contexto de cada campo. Las columnas y sus tipos son los siguientes:

Tabla 6: Formato de los campos de salida

Fuente: Elaboración Propia

Número	Nombre	Tipo
1	Countryname	STRING
2	Owner	STRING
3	Project Id	INT64
4	Proj Created Dt	DATE
5	Customer Number	INTEGER
6	Project State	STRING
7	Projectworkflowstatename	STRING
8	Order Num	INT64
9	Task Order Date	DATE
10	DOMs Order Status	STRING
11	SKU Number	STRING
12	SKU Description	STRING
13	Category	STRING
14	Accouting Class	STRING
15	Item Quantity	INT64
16	SKU USD Unit Price	FLOAT
17	Total USD Price	FLOAT
18	Total USD Margin	FLOAT

Para finalizar la etapa de limpieza de datos procederemos a eliminar los *outliers*, para esto se eliminarán los extremos donde el campo “Total USD Margin” correspondan al 1% más bajo y al 1% más alto. El conjunto de datos trabajado durante esta etapa será almacenado

en un nuevo CSV, que corresponderá a el *dataset* limpio, serán estos datos los que utilizaremos para las siguientes etapas del trabajo.

Al comienzo del proceso de limpieza contábamos con 735.363 registros y 22 columnas, al finalizar la etapa poseemos 691.875 registros y 18 columnas.

3.2 ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Una vez que hemos realizado la limpieza los datos podemos comenzar a trabajar con estos, antes de realizar las tareas principales que hemos descrito debemos observar de qué manera se comportan los datos, realizando un análisis exploratorio de los datos, conocido también como EDA. Esto nos permitirá conocer de mejor manera la naturaleza de los datos y trazar un camino a seguir para cumplir nuestros objetivos.

Comenzaremos analizando de qué manera se distribuyen ciertos campos categóricos del *dataset*, también analizaremos las ventas a lo largo de los años, para identificar ciertos patrones. En el caso de la distribución registros por año podemos observar que existe una uniformidad en los valores tal como se puede observar en la Figura 13, esto nos es útil ya que un *forecast* necesita que los periodos de ventas sean equilibrados para que la predicción de ventas posea una base sólida de todos los años. Los años que más registros poseen son el 2018 con un 24.4% de los valores, seguido por 2016 con un 20.4%, los registros más bajos corresponden a 2014, 2015 y 2019 con 11%, 12.4% y 13.4% respectivamente. Esto tiene una explicación sencilla los años 2014 y 2015 poseen menos valores al ser años más antiguos, mientras que 2019 posee una baja cantidad ya que el *dataset* posee datos hasta el 28 de junio de 2019, por lo cual mientras que el resto de los años poseen datos de todo el año, en 2019 solo acumuló datos de la primera mitad del año.

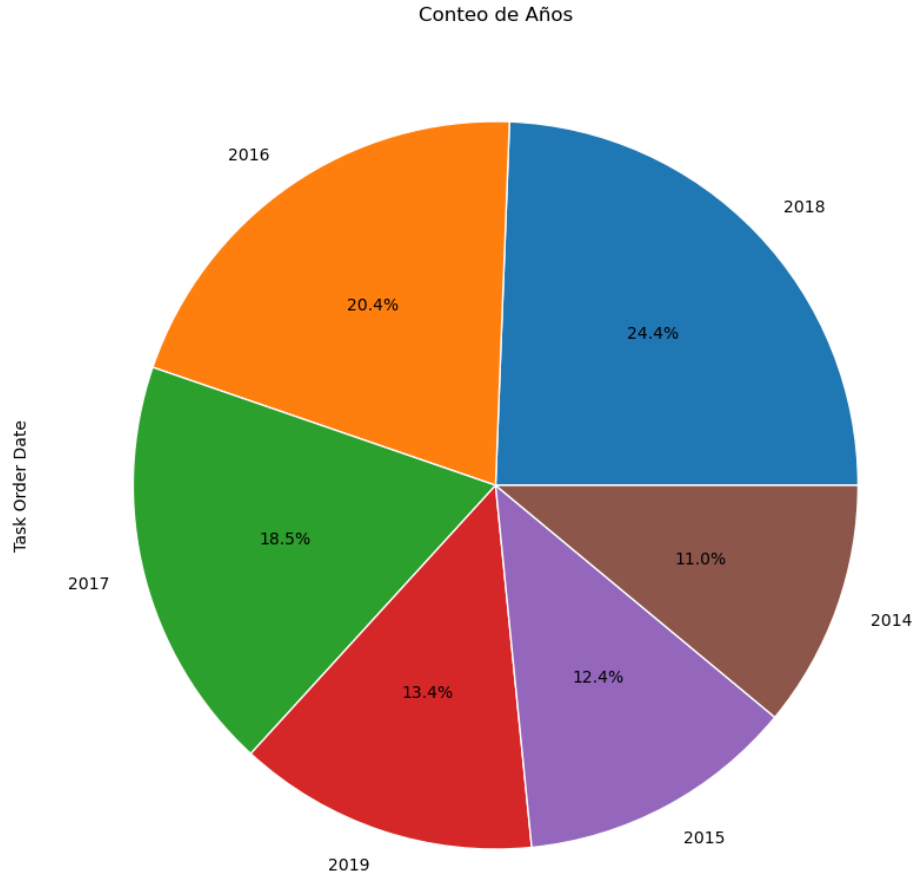


Figura 13: Distribución Task Order Date

Fuente: Elaboración Propia

En el caso del campo “Countryname”, correspondiente al país en donde se realizó la transacción podemos observar que Brasil, México, Chile, Colombia y Argentina son los países donde se producen mayores transacciones, el resto de los países que fueron agrupados el valor “Otros” poseen un valor menor al 3%. Esto nos puede ser de utilidad para realizar el *forecast* considerando los países más importantes para la empresa. El gráfico se puede observar en la Figura 14.

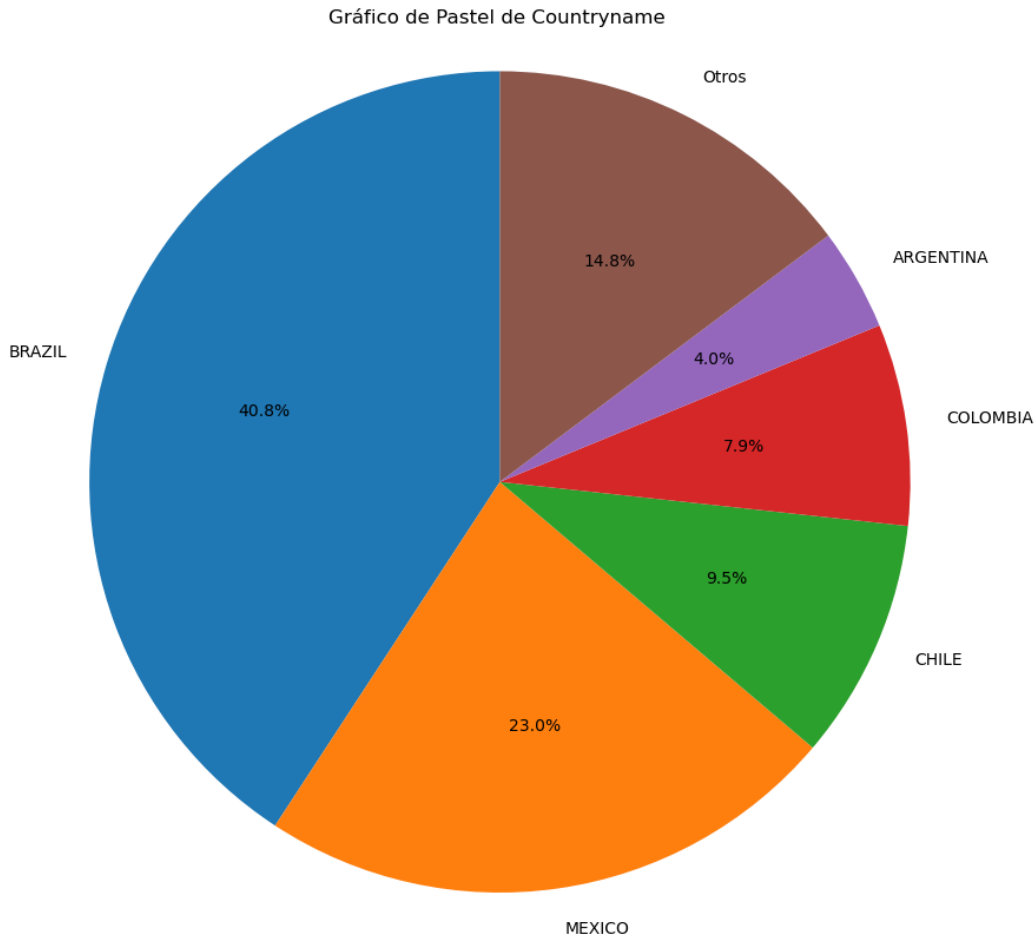


Figura 14: Distribución Countryname

Fuente: Elaboración Propia

De manera análoga a las mencionadas con anterioridad, ahora veremos cuales son las categorías que aparecen en un mayor número de transacciones en el *dataset*, recordemos que este campo “Category” fue creado durante la etapa de transformación y es un campo que agrupa los valores de “Accounting Class”. Según lo que se puede observar en la Figura 15, la categoría principal del *dataset* corresponde a Hardware con una amplia mayoría, seguido de “Support & Services”. Así como mencionamos anteriormente esto nos puede ser de utilidad para realizar un *forecast* más específico.

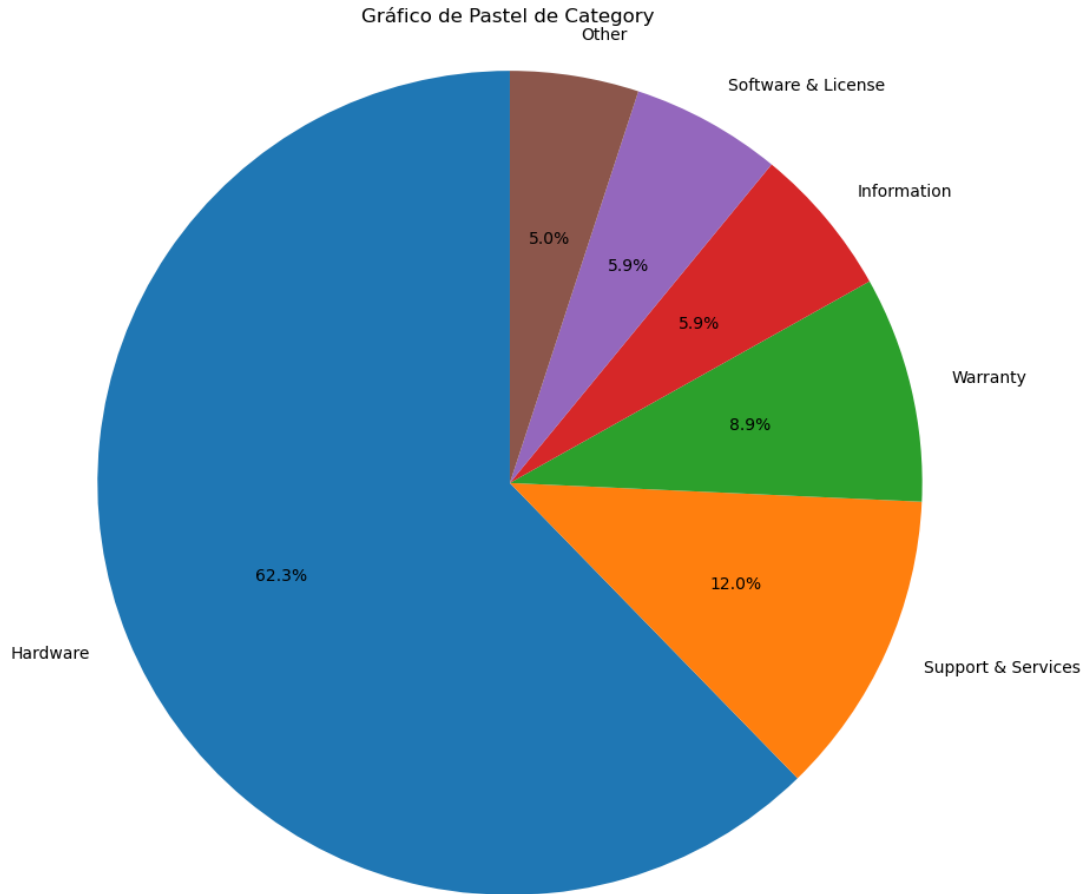


Figura 15: Distribución Category
Fuente: Elaboración Propia

Ahora realizaremos un análisis exploratorio sobre el desempeño de los “Owner” o propietarios sobre las métricas de ventas, las que se considerarán serán los proyectos liderados, las unidades vendidas, las ventas realizadas, y el margen obtenido. Para representar esto debido al gran número de propietarios que existen se mostrará un top 10 de estos en cada una de las métricas, utilizando un histograma para observar también el valor que alcanza.

Comenzaremos con los propietarios que han liderado una mayor cantidad de proyectos. Existe un gran número de propietarios y a su vez también un gran número de proyectos asociados a Dell Technologies, podemos observar del gráfico de la figura 16 que *Lucas_Azevedo*, que corresponde al propietario con una mayor cantidad de proyectos liderados, posee una cantidad cercana a los 600 proyectos, mientras que *fernando_banz* cierra el top 10 con una cifra cercana a los 450 proyectos.

La métrica de proyectos liderados es relevante para la empresa y puede mostrar una primera mirada al desempeño de los propietarios, ya sea por experiencia poseída, por años de trabajo o la capacidad organizativa de manejar múltiples proyectos de forma simultánea,

lo que permitiría focalizar el trabajo en los propietarios más importantes que puedan manejar los proyectos más importantes.

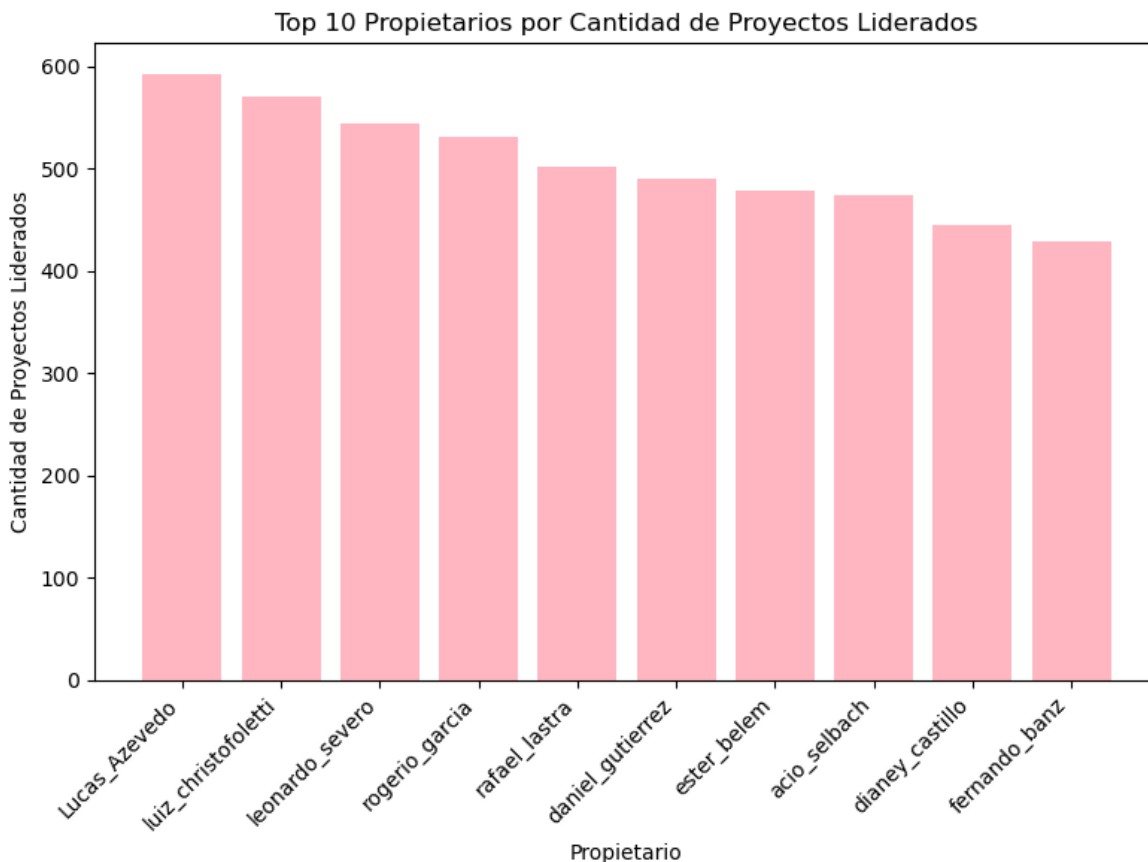


Figura 16: Histograma Owner por Project Id

Fuente: Elaboración Propia

Por otra parte, podemos analizar las unidades vendidas por los propietarios, se puede observar del gráfico de la figura 17 que el propietario *fernando_menegoli* ha vendido más de 160.000 unidades, obteniendo una amplia diferencia frente al resto. Mientras tanto, el resto de los propietarios con más unidades vendidas poseen valores que oscilan entre las 60.000 y 100.000 unidades vendidas. Si bien esta métrica por sí sola no nos muestra la eficacia que posee un propietario si debiese ser una métrica para tomar en cuenta, ya que el hecho de realizar un mayor número de ventas puedes derivar a un mayor número de ganancias para la empresa.

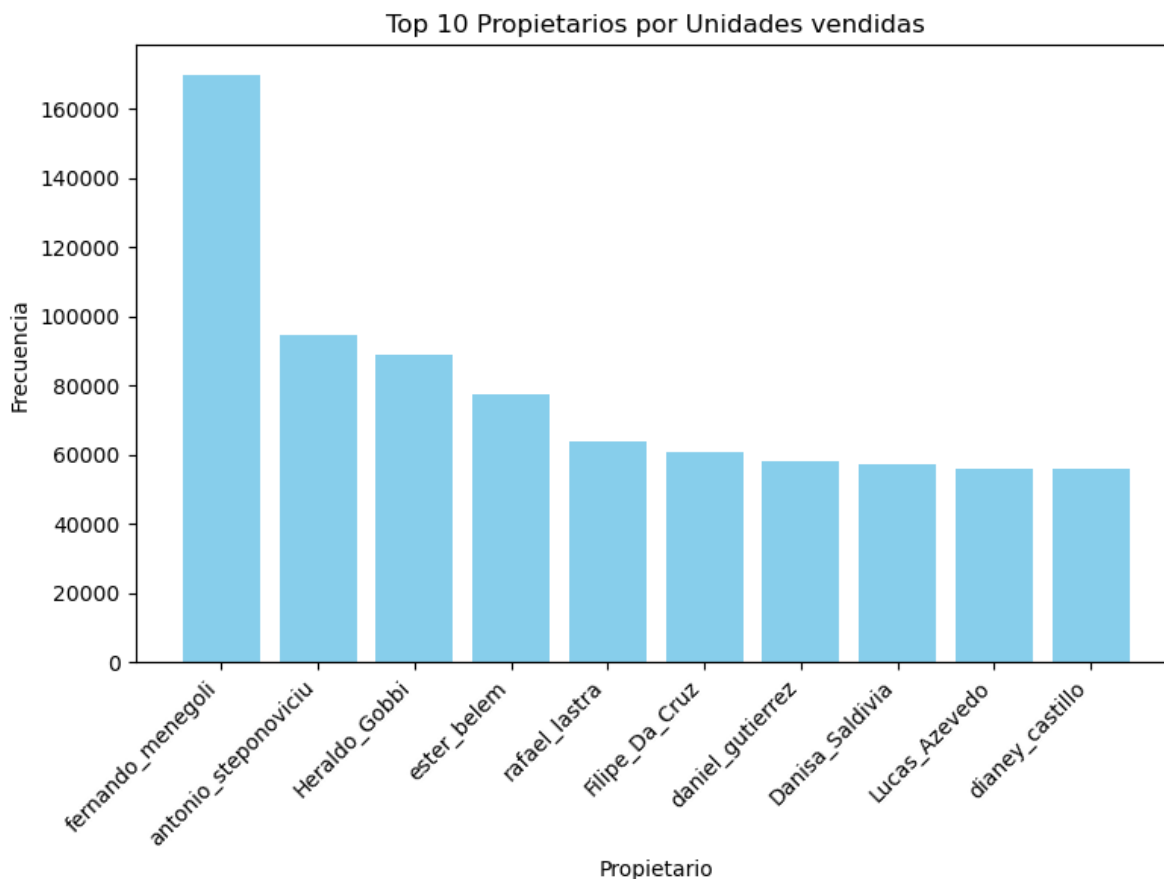


Figura 17: Histograma Owner por Item Quantity

Fuente: Elaboración Propia

Como mencionamos antes la siguiente métrica que se analizará será los propietarios que realizaron las mayores ventas, cabe destacar que las ventas en este contexto es el precio final del producto y no es la ganancia que obtuvo la empresa por esta venta, de todas maneras este gráfico tiene una mayor importancia que el anterior, ya que un propietario puede vender muchos productos bajos que no generan tanto impacto como otros productos, en este caso podemos observar que la propietaria *ester_belen* ha sido quien ha realizado una mayor venta durante el periodo que abarcamos en el *dataset*, por otra parte *leonardo_severo* cierra el top con uno valor cercano. Algo que podemos notar a diferencia del gráfico pasado es que las cifras de ventas de cada propietario son muy cercanas, y también que a pesar de existir propietarios con muchas unidades vendidas no representan esos números en los montos de venta.

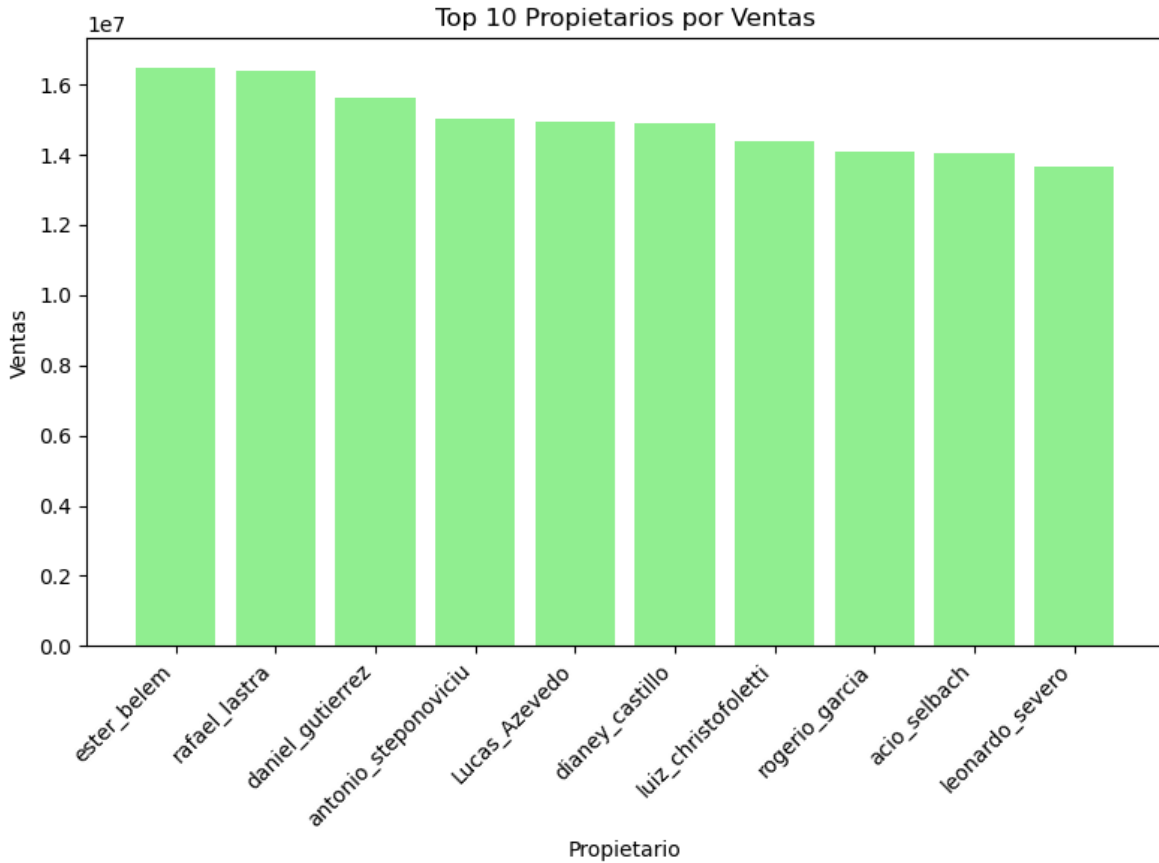


Figura 18: Histograma Owner por Total USD Price
Fuente: Elaboración Propia

Finalmente queda por analizar la métrica más importante para una empresa, la cual corresponde al margen de ganancia. Las empresas buscan generar un beneficio de su trabajo y este se ve reflejado en el margen que estas generan, los propietarios que posean un mayor margen acumulado son aquellos que son más valiosos para la empresa, aunque también se pueden fijar en cómo ha crecido el margen de los propietarios durante los años, debido a que, si un propietario poseía un margen anual bajo, pero ha ido en crecimiento puede representar una pieza muy importante para la empresa en un futuro.

Los propietarios que han generado un mayor margen son los que podemos observar en la figura 19, *rafael_lastra*, es el propietario que ha generado un mayor margen, aunque de manera similar a la figura 18, no existen grandes diferencias entre el resto.

Podemos observar que el propietario *rafael_lastra* formó parte del top 10 propietarios en cada una de las métricas, lo cual representa que es un propietario valioso para Dell Technologies.

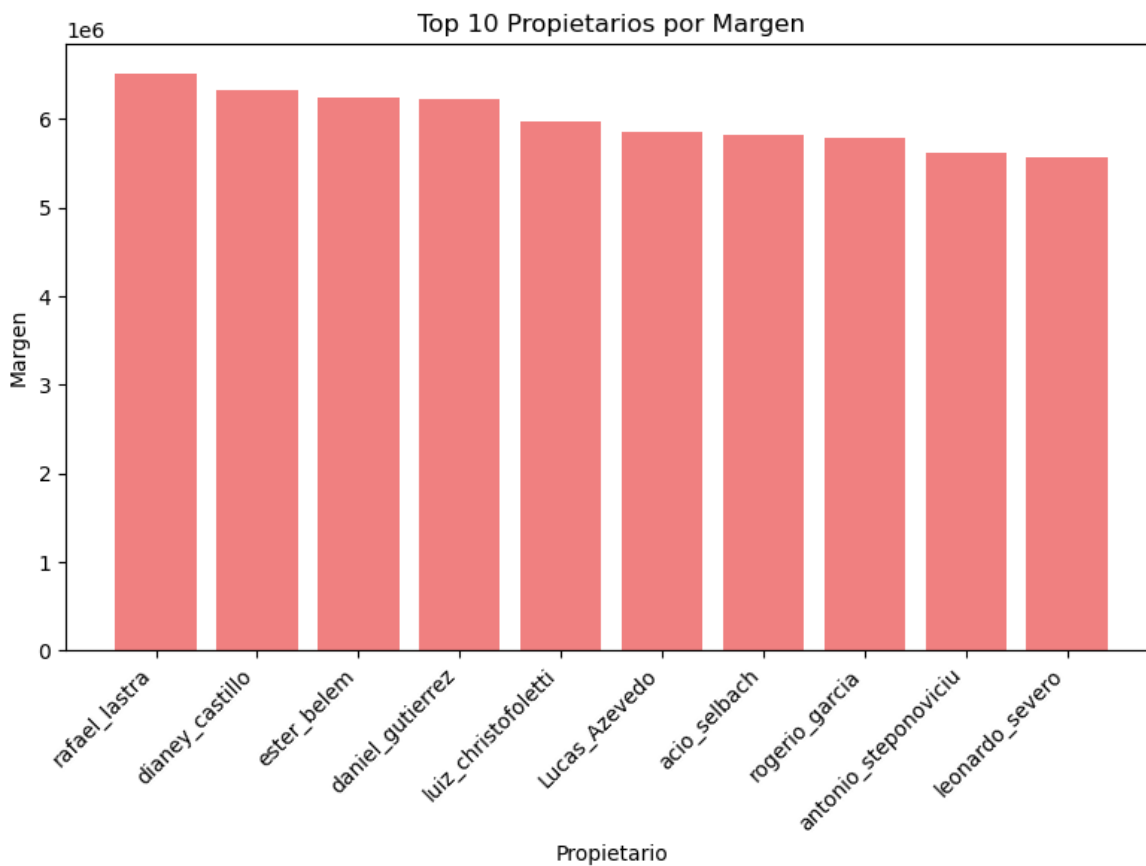


Figura 19: Histograma Owner por Total USD Margin
Fuente: Elaboración Propia

Nos moveremos ahora a las variables numéricas de nuestro dataset, considerando en mayor parte a los campos “Total USD Price” y “Total USD Margin”, podemos obtener información como tendencia de ventas, distribución de estas, estadísticas básicas, etc.

En primera instancia realizaremos un gráfico con las ventas mensuales y el margen mensual, para observar como han ido variando las ventas a través de los años y si se puede observar ciertas tendencias en las ventas, esto queda representado en la Figura 20, donde podemos ver que ambos valores están relacionados, pues una mayor cantidad de ventas refleja una mayor cantidad de margen, por otro lado se puede observar cierta estacionalidad a excepción de ciertos meses donde las cifras se escapan de lo esperado.

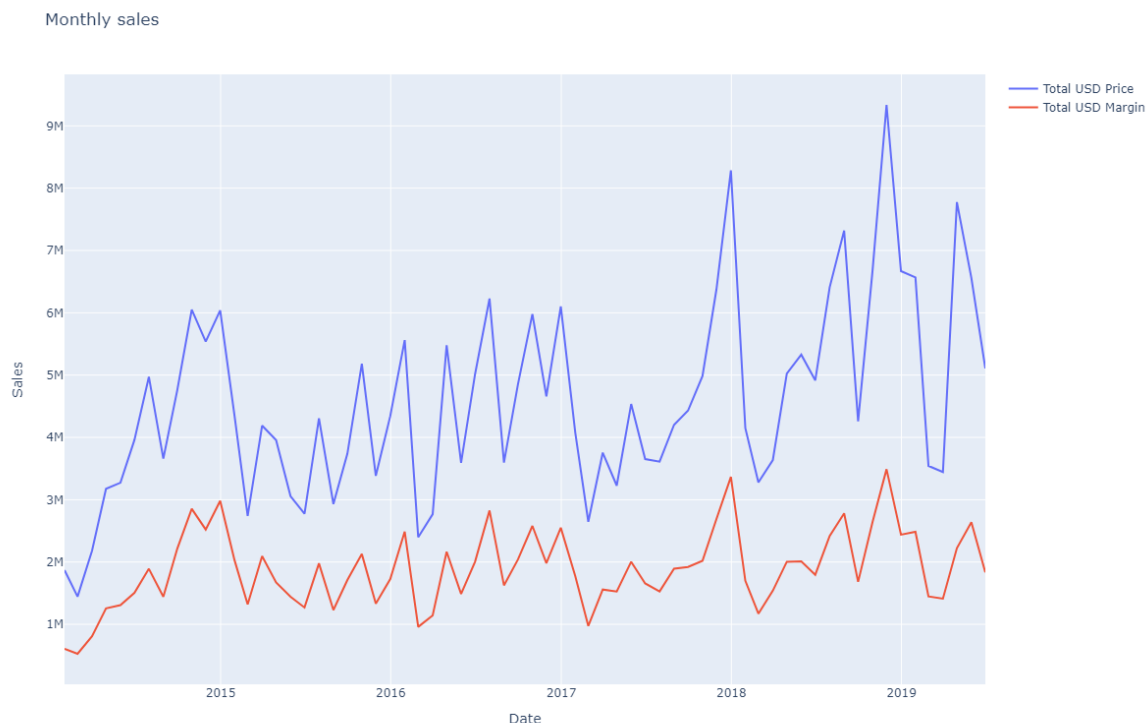


Figura 20: Ventas y margen mensuales

Fuente: Elaboración Propia

Ahora analizaremos las cifras de ventas y margen mensuales para cada uno de los países, podemos observar que existen 42 países en nuestra base de datos, cada uno de estos países serán identificados con una línea en nuestro gráfico, el eje y serán las ventas y el eje x representará la fecha del valor, la fecha corresponde las ventas mensuales en el respectivo año. En la Figura 21 podemos observar las ventas mensuales mientras que en la Figura 22 se presenta el margen mensual.

En ambos gráficos es Brasil el país con mayor cantidad de venta y margen, esto debido al gran número de registros que se posee en el *dataset* correspondiente a ese país, México es el siguiente país con un gran volumen de ventas, con un número de ventas similar al del resto de países que posean un número de ventas menores al tres por ciento.

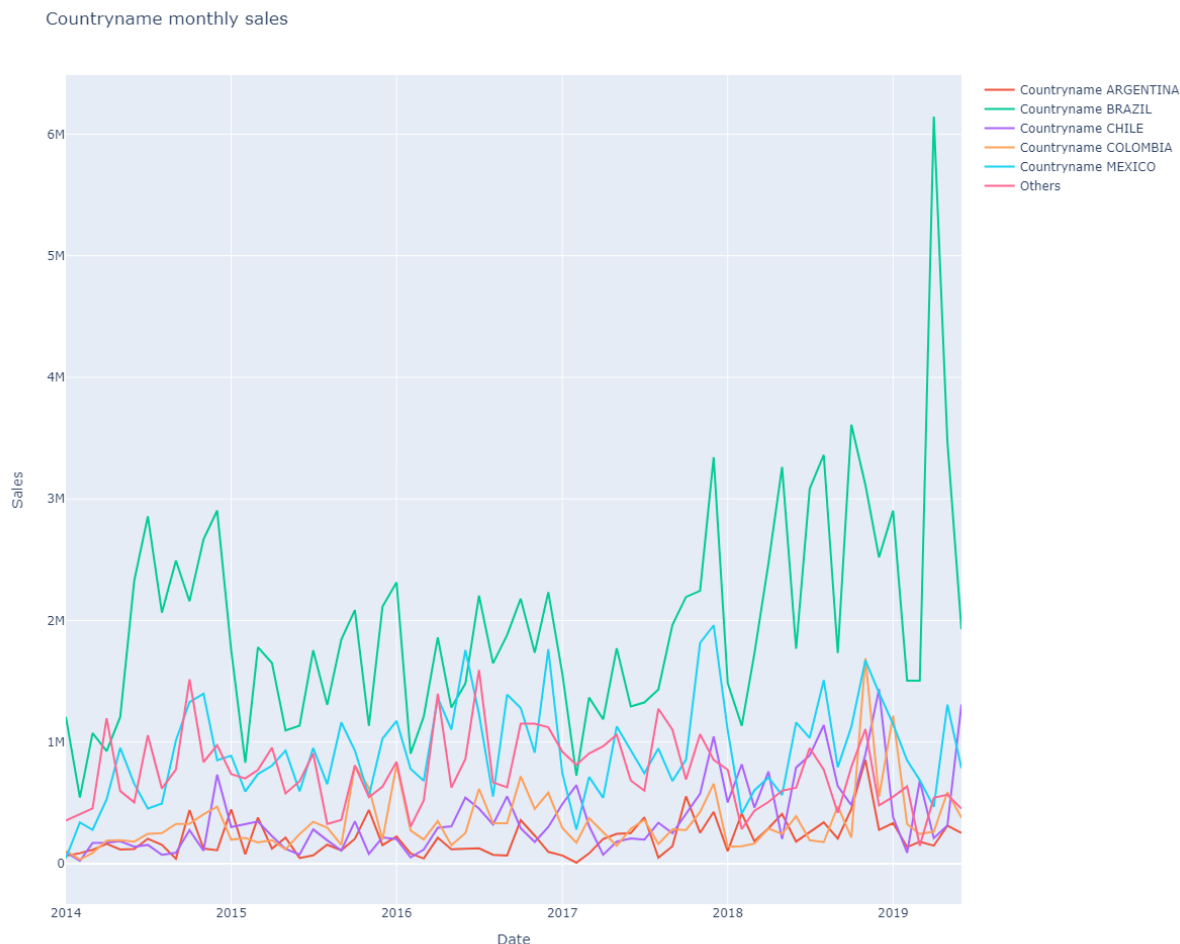


Figura 21: Ventas mensuales por Countryname

Fuente: Elaboración Propia

Mientras tanto con el margen de manera análoga a lo visto anteriormente se pueden observar valores relacionados al margen de las ventas, sin embargo, podemos observar que casos donde el margen es cercano a cero o incluso en ocasiones es negativo, como lo es en el caso de Argentina durante el periodo de febrero 2017, donde existe un valor muy cercano a cero.

La razón por la cual el margen puede poseer un valor negativo puede deberse a diferentes motivos, uno es que, debido a descuentos, el precio final del producto es menor a el costo asociado a la fabricación de este, también se puede deber a productos que han sido regresados lo que provoca un margen negativo en la venta.

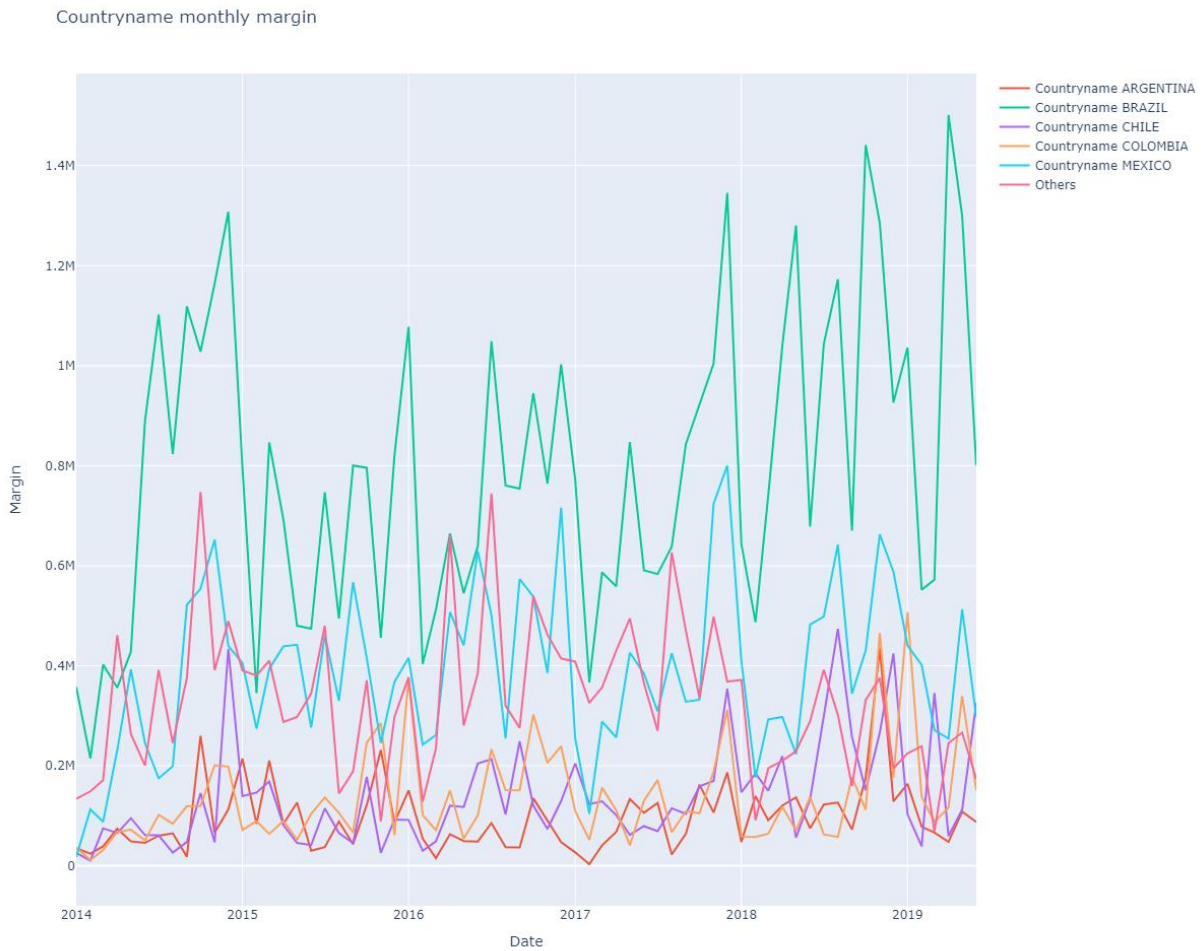


Figura 22: Margen mensual por Countryname
Fuente: Elaboración Propia

Siguiendo con nuestras exploraciones sobre las ventas nos centraremos en analizar las ventas y margen con respecto a el día que se realizó la compra, analizaremos en dos etapas, primero con el día de semana en que se realizó la compra y, en segundo lugar, el mes correspondiente.

En la Figura 23 podemos observar que las ventas los fines de semana es mínima, esto se puede explicar debido a que muchas empresas no trabajan los fines de semana y debido a que las ventas del *dataset* son enfocadas en proyectos es normal que las ventas disminuyan los fines de semana. Por otro lado, podemos observar que desde el lunes hasta el viernes las ventas crecen paulatinamente, los viernes son cruciales para la empresa debido al gran número de ventas que realizan este día.

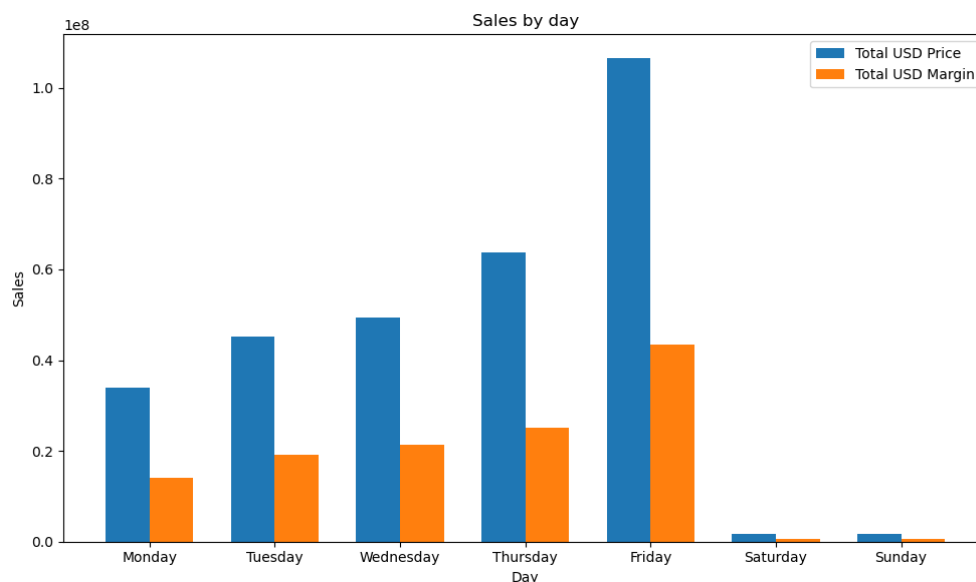


Figura 23: Ventas por día de semana

Fuente: Elaboración Propia

Por otro lado en la Figura 24 se observa que la mayoría de las ventas son generadas en los meses de mitad de año (abril-julio) y finalizando el año (octubre-enero). Esta diferencia puede deberse a que el año fiscal de Dell Technologies comienza en febrero, esto implica que muchas áreas y equipos deben completar objetivos durante este último mes de enero, también ocurren muchas vacaciones durante el mes de febrero, y así se puede desglosar más información. Esto nos es útil para que los recursos de la empresa puedan ser utilizados con mayor estrategia.

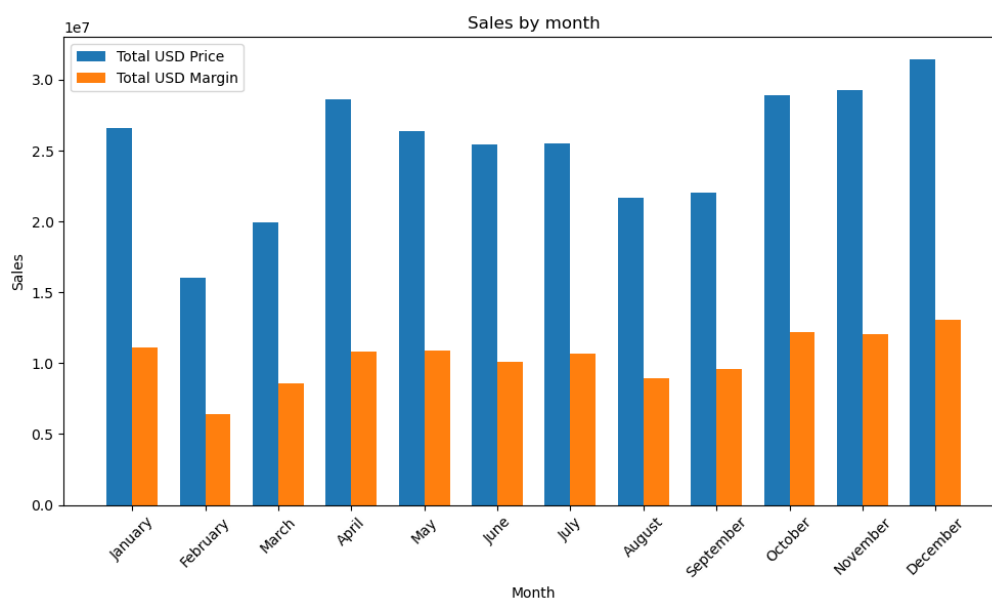


Figura 24: Ventas por mes

Fuente: Elaboración Propia

Mientras tanto la distribución de los campos “Total USD Price” y “Total USD Margin” se pueden observar en la Figura 25, donde se grafican mediante una distribución normal los valores de ambas columnas, podemos notar que mientras el margen presenta valores más uniformes entre su media, mínimo y máximo el campo “Total USD Price” representa una distribución con pocos valores inferior a la media.

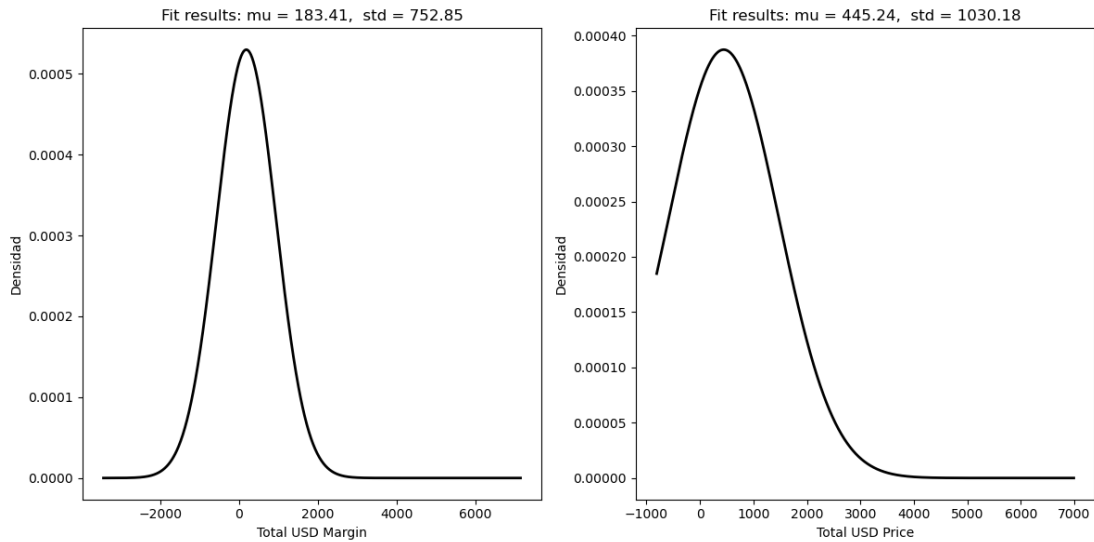


Figura 25: Distribución normal de Total USD Margin y Total USD Price

Fuente: Elaboración Propia

De manera análoga en la Figura 26 podemos ver la distribución con un *Violin Plot*, representando la densidad de los valores que toman los campos.

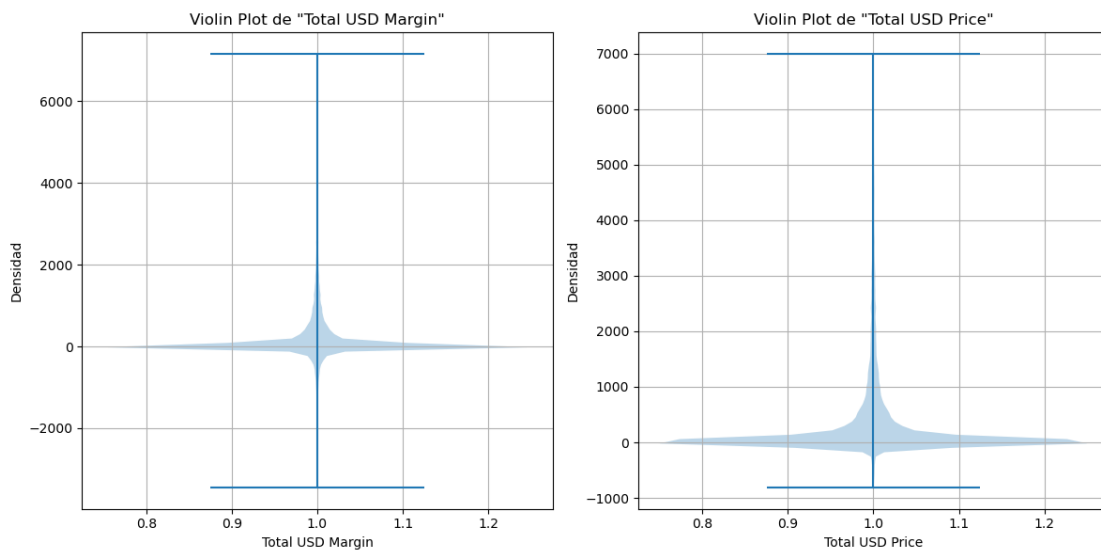


Figura 26: Violin Plot de Total USD Margin y Total USD Price

Fuente: Elaboración Propia

Ambos gráficos en conjunto nos permiten entender de mejor manera la forma en que ambos campos se distribuyen.

Con esto hemos finalizado la etapa de exploración de datos, y podemos proseguir con las siguientes etapas de la metodología.

3.3 FORECASTING

Siguiendo con la metodología de trabajo comenzaremos con la etapa del *forecasting* o pronóstico, lo que nos permitirá prever valores futuros basados en datos históricos y tendencias observadas.

A continuación, se explorarán diversas técnicas y enfoques utilizados en el *forecasting*, desde modelos estadísticos clásicos hasta técnicas más avanzadas basadas en *machine learning* y redes neuronales.

Entre las técnicas que se utilizarán se encuentran:

- **SARIMA**
- **Prophet**
- **NeuralProphet**
- **Holt-Winters**

Comenzaremos por verificar la estacionalidad de nuestro *dataset* para comenzar con el modelamiento de las técnicas de *forecasting*.

El primer paso es transformar nuestro *dataset* a un formato que sea sencillo de leer para el modelo, esto corresponde a un formato de serie temporal, donde una columna sea la variable de tiempo y la otra la variable a predecir, para esto cambiamos los nombres de las columnas "Task Order Date" y "Total USD Margin", a "ds" e "y" respectivamente. Luego de esto podemos agrupar los registros que contengan el mismo valor en la columna "ds" y sumar los valores correspondientes de la columna "y", esto permitirá obtener un *dataframe* con las ventas diarias.

Como sabemos que existen días donde no se registraron ventas procederemos a completar el *dataframe*, asignándole a los valores de "ds" que no se encuentran un valor de "y" igual a 0. Con esto podremos realizar técnicas de *forecasting* sin problemas.

Nuestro *dataframe* consta de 2.005 registros, los cuales corresponden a todas las fechas desde el 2014-01-01 hasta el 2019-06-28 con su respectivo valor del margen diario.

Como realizaremos el *forecasting* respecto a las ventas mensuales necesitamos agrupar estos datos de manera mensual, para esto agrupamos los valores del *dataframe* de manera mensual y designamos el valor del campo "y" como la suma de la agrupación de registros, de esta manera el *dataframe* contiene como valor "ds" el último día de cada mes, y como valor de "y" la suma del campo "Total USD Margin".

Un paso importante para la creación de modelos de *forecasting* corresponde a la separación de los datos en dos grupos, el primer grupo correspondiente a los datos de entrenamiento serán aquellos datos con los cuales crearemos y entrenaremos nuestros modelos de *forecasting*, mientras que el segundo segmento corresponde a los datos de prueba, con estos datos verificaremos que tan acertado es nuestro modelo, en este caso designamos los últimos 12 meses a nuestro grupo de datos de prueba. Dicho de otra manera, utilizaremos la información hasta un año antes de la que poseemos para crear el modelo y comprobaremos con los resultados del último año para ver qué tan acertado fue nuestro modelo.

Una vez realizada la transformación y segmentación de nuestro *dataframe* estamos listos para comenzar a diseñar y elaborar nuestros modelos, el primer paso corresponderá a verificar si nuestro *dataframe* es estacionario. En el contexto del *forecasting* la estacionariedad se logra cuando los valores estadísticos como la media y la varianza se mantienen constantes en el tiempo, el hecho de que una serie temporal sea estacionario es de utilidad por múltiples factores.

- **Modelado más simple:** Los modelos de *forecasting*, como SARIMA o Holt-Winters, asumen que la serie temporal es estacionaria o puede transformarse en una serie estacionaria. Esto simplifica significativamente el proceso de modelado, ya que los patrones y tendencias en los datos pueden ser más fáciles de identificar y capturar.
- **Estabilidad de las propiedades estadísticas:** Cuando una serie temporal es estacionaria, sus propiedades estadísticas, como la media y la varianza, son estables a lo largo del tiempo. Esto significa que los modelos entrenados en datos pasados pueden aplicarse de manera más confiable a datos futuros, ya que se espera que las relaciones entre las variables se mantengan consistentes.
- **Facilita la identificación de patrones:** La estacionariedad facilita la identificación de patrones y tendencias en los datos. Los modelos de *forecasting* buscan capturar estos patrones para hacer predicciones precisas sobre el futuro. Si la serie temporal no es estacionaria, puede ser difícil discernir entre los efectos temporales y los cambios reales en el comportamiento de los datos.
- **Mayor precisión en las predicciones:** Los modelos entrenados en datos estacionarios tienden a producir predicciones más precisas y confiables. Esto se debe a que los patrones identificados en los datos históricos son más propensos a persistir en el futuro si la serie temporal es estacionaria.

Para comprobar la estacionariedad de una serie temporal se pueden utilizar diversos métodos como:

- **Visualizar Gráficos:** Puedes revisar un gráfico de la serie temporal y analizar si se detectan tendencias o patrones a lo largo de ciertos periodos de tiempo.

- **Resúmenes Estadísticos:** Se pueden analizar resúmenes estadísticos de los datos para cierto tiempo y revisar si existen diferencias notorias entre los periodos.
- **Test Estadísticos:** Puedes utilizar test estadísticos para comprobar si la estacionariedad de una serie temporal.

Dentro de las pruebas estadísticas más comunes para comprobar si una serie temporal es estacionaria nos encontramos con el “Augmented Dickey-Fuller test”, conocido también como ADF. Esta prueba se basa en una hipótesis nula en la cual la serie temporal posee una raíz unitaria, esto implica que la serie tiene una tendencia que persiste en el tiempo y no se revierte con el paso de este, en otras palabras, los cambios en la serie no se amortiguan a medida que transcurre el tiempo, esto deriva en una serie temporal no estacionaria.

Los resultados de la prueba nos entregan el valor p (p -value), el cual es una medida que indica la probabilidad de obtener resultados al menos tan extremos como los observados en la muestra, al suponer la hipótesis nula como verdadera. En el contexto del ADF test, si el p -value es menor a el umbral, generalmente con un valor de 0.05, entonces se rechaza la hipótesis nula, lo que sugiere que la serie temporal es estacionaria. Si el p -value es mayor que el umbral, no hay suficiente evidencia para rechazar la hipótesis nula, lo que indica que la serie temporal no es estacionaria.

Al realizar el ADF en nuestro *dataframe* se obtuvieron los resultados que se observan en la Figura 27.

```
from statsmodels.tsa.stattools import adfuller

#Augmented Dickey-Fuller test:
print('Results of Dickey Fuller Test:')
dfctest = adfuller(group_df['y'], autolag='AIC')

dfoutput = pd.Series(dfctest[0:4], index=['Test Statistic', 'p-value', '#Lags Used', 'Number of Observations Used'])
for key,value in dfctest[4].items():
    dfoutput['Critical Value (%s)'%key] = value

print(dfoutput)
```

✓ 0.0s

```
Results of Dickey Fuller Test:
Test Statistic      -4.238673
p-value             0.000566
#Lags Used          3.000000
Number of Observations Used  62.000000
Critical Value (1%)  -3.540523
Critical Value (5%)  -2.909427
Critical Value (10%) -2.592314
dtype: float64
```

Figura 27: Dickey-Fuller Test

Fuente: Elaboración Propia

Como se puede observar el p -value posee un valor de 0.000566 lo cual es menor a nuestro umbral 0.05, por lo cual podemos concluir que nuestra serie temporal es estacionaria y por consiguiente los resultados de nuestro *forecasting* tendrá mayor precisión.

Una vez finalizado con el ADF test, nos queda comenzar con el modelado de las técnicas de *forecasting*, comenzaremos por la técnica SARIMA, recordemos que el entrenamiento del modelo se realizará dejando a parte los últimos 12 meses, ya que estos serán utilizados para validar los resultados del modelo.

El primer paso para obtener el modelo esperado con los mejores resultados posibles es el definir los hiperparámetros que utilizaremos en el modelo, los hiperparámetros son aquellos valores que podemos configurar dentro de la creación del parámetro, esto permite una personalización en la forma en que trabaja el modelo. Una técnica para encontrar los mejores valores de los parámetros consiste en crear una grilla con los parámetros a modificar y una serie de valores a considerar. Luego se ejecuta el modelo con cada posible combinación de parámetros y se valora la calidad del resultado, luego de crear todas las combinaciones dadas guardamos los parámetros que nos entregaron el mejor valor, serán estos parámetros los que utilizaremos en nuestro modelo final, esto se debe realizar para cada modelo. Un ejemplo del código que nos permite encontrar los mejores parámetros se puede observar en la Figura 28, en este código encontramos los mejores parámetros para el modelo Holt Winters.

```

1  from statsmodels.tsa.holtwinters import ExponentialSmoothing
2  from sklearn.model_selection import ParameterGrid
3
4  def find_best_holt_params(param_grid):
5      best_params = None
6      best_score = float("inf")
7
8      for params in ParameterGrid(param_grid):
9          model = ExponentialSmoothing(train_prophet['y'], **params)
10         model_fit = model.fit()
11
12         # Realiza pronósticos para el mismo número de pasos que el conjunto de prueba
13         forecast = model_fit.forecast(steps=len(test_prophet))
14         hwes_data = prophet_df.copy()
15         hwes_data['forecast'] = forecast
16
17         y_true = test_prophet['y'].values
18         y_pred = hwes_data['forecast'].tail(len(test_prophet)).values
19
20         score = mean_absolute_percentage_error(y_true, y_pred)
21
22         if score < best_score:
23             best_score = score
24             best_params = params
25
26     return best_params, best_score
27
28 param_grid = {
29     'trend': ['add', 'mul'],
30     'damped_trend': [True, False],
31     'seasonal': ['add', 'mul'],
32     'seasonal_periods': [12]
33 }
34
35 best_params, best_score = find_best_holt_params(param_grid)
36 print("Mejores hiperparámetros:", best_params)
37 print("Mejor puntuación (MAPE):", best_score)

```

Figura 28: Optimización de parámetros

Fuente: Elaboración Propia

Tal como se obtuvieron los parámetros para el modelo Holt Winters se realizó el mismo proceso con los otros modelos, una vez obtenidos los mejores parámetros se creó el modelo con los resultados obtenidos.

El primer modelo obtenido es SARIMA, el cual corresponde a un modelo estadístico utilizado para el análisis y la predicción de series temporales. Combina componentes autoregresivos (AR), de media móvil (MA) e integrados (I) para modelar la tendencia, la estacionalidad y la aleatoriedad en los datos temporales. La "S" en SARIMA indica que el modelo puede manejar datos estacionales. Específicamente, SARIMA se caracteriza por los parámetros (p, d, q) para la parte no estacional del modelo y (P, D, Q, s) para la parte estacional, donde p, d y q son los órdenes de los componentes AR, I y MA respectivamente, y P, D, Q y s son los órdenes de los componentes AR, I, MA y el período estacional. El significado de los parámetros son los siguientes:

- **p (Autoregressive order):** Representa el orden del componente autoregresivo, que captura la relación lineal entre la variable de interés y sus valores pasados en el tiempo.
- **d (Integrated order):** Indica el orden de diferenciación necesario para hacer estacionaria la serie temporal. La diferenciación implica tomar la diferencia entre los valores observados en diferentes puntos de tiempo.
- **q (Moving Average order):** Define el orden del componente de media móvil, que modela el efecto de los errores pasados en la serie temporal actual.
- **P (Seasonal Autoregressive order):** Es similar al parámetro p pero para la parte estacional del modelo.
- **D (Seasonal Integrated order):** Similar a d pero para la parte estacional del modelo.
- **Q (Seasonal Moving Average order):** Similar a q pero para la parte estacional del modelo.
- **s (Seasonal Period):** Representa la periodicidad estacional de la serie temporal, es decir, el número de períodos de tiempo en un ciclo estacional.

Al realizar la optimización de parámetros sobre nuestros datos se obtuvieron los siguientes valores:

Tabla 7: Parámetros SARIMA

Fuente: Elaboración Propia

Parámetro	Valor
p	3
q	1
d	0
P	1
Q	1
D	0
S	12

En la Figura 29 se observan gráficos relacionados al modelo SARIMA.

El primero corresponde a el residuo estandarizado, lo cual nos muestra el residuo de la predicción dividido en su error estándar, valores cercanos al cero implican un forecast certero.

El segundo es un gráfico con tres componentes dentro, el primero es histograma, el segundo corresponde a kernel density estimation (KDE), el cual representa una manera no parametrica de estimar la función de probabilidad de densidad (PDF) de una variable aleatoria.

En tercer lugar se encuentra el gráfico Normal Q-Q que nos muestra la distribución de los datos, si los puntos se encuentran cerca de la línea roja entonces la distribución corrponde a una distribución normal.

Finalmente el correlograma el cual se utiliza para identificar patrones de correlación entre los valores de una serie temporal y sus rezagos (valores pasados). Se espera que los valores se encuentren dentro de la banda de confianza para asumir que los valores no actuales no dependen de los anteriores.

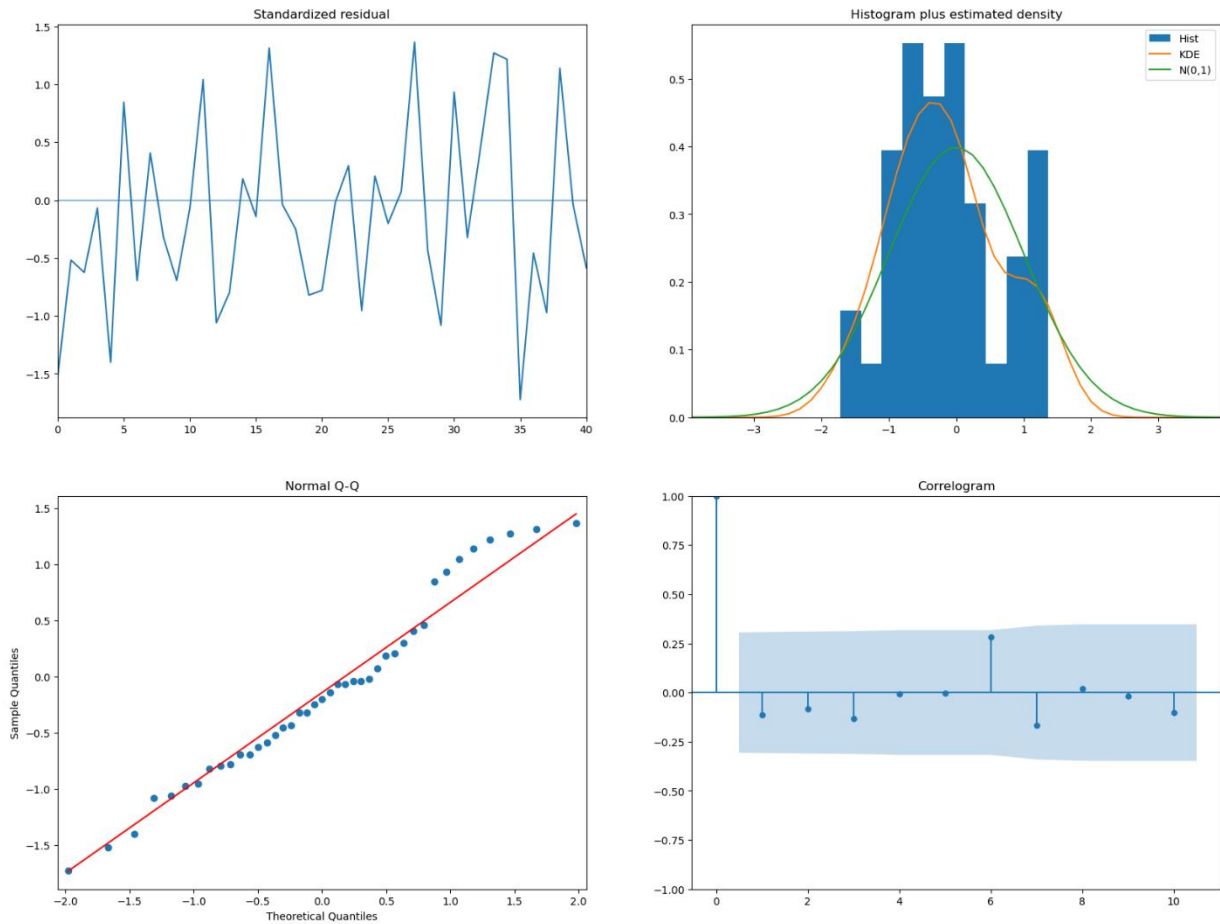


Figura 29: SARIMA Diagnostic
Fuente: Elaboración Propia

Con esta información ya hemos creado el modelo y analizado sus resultados de manera general, continuaremos con los siguientes modelos.

El modelo que continúa en la lista es Prophet, el cual es un modelo de pronóstico desarrollado por Facebook que se utiliza para predecir series temporales con estacionalidad. Es especialmente útil para conjuntos de datos con tendencias no lineales y múltiples fuentes de incertidumbre. Prophet se basa en un modelo aditivo que descompone la serie temporal en componentes de tendencia, estacionalidad y días festivos. Utiliza un enfoque de ajuste de curvas flexible que puede manejar cambios bruscos en los datos y proporciona intervalos de confianza para las predicciones.

Al realizar la optimización de parámetros en el modelo de Prophet se obtuvieron resultados los siguientes valores de los parámetros:

Tabla 8: Parámetros Prophet

Fuente: Elaboración Propia

Parámetro	Valor
changepoint_prior_scale	0.8
seasonality_mode	multiplicative
seasonality_prior_scale	0.05

Con estos parámetros se configura el modelo, el resultado del modelo se puede observar en la Figura 30, la línea azul corresponde a los valores predichos, la sombra corresponde a la banda de confianza que posee el modelo, mientras que los puntos negros corresponden a los valores reales de las ventas. Se puede observar que los valores normalmente se encuentran dentro del intervalo de confianza del modelo, lo cual es una buena noticia.

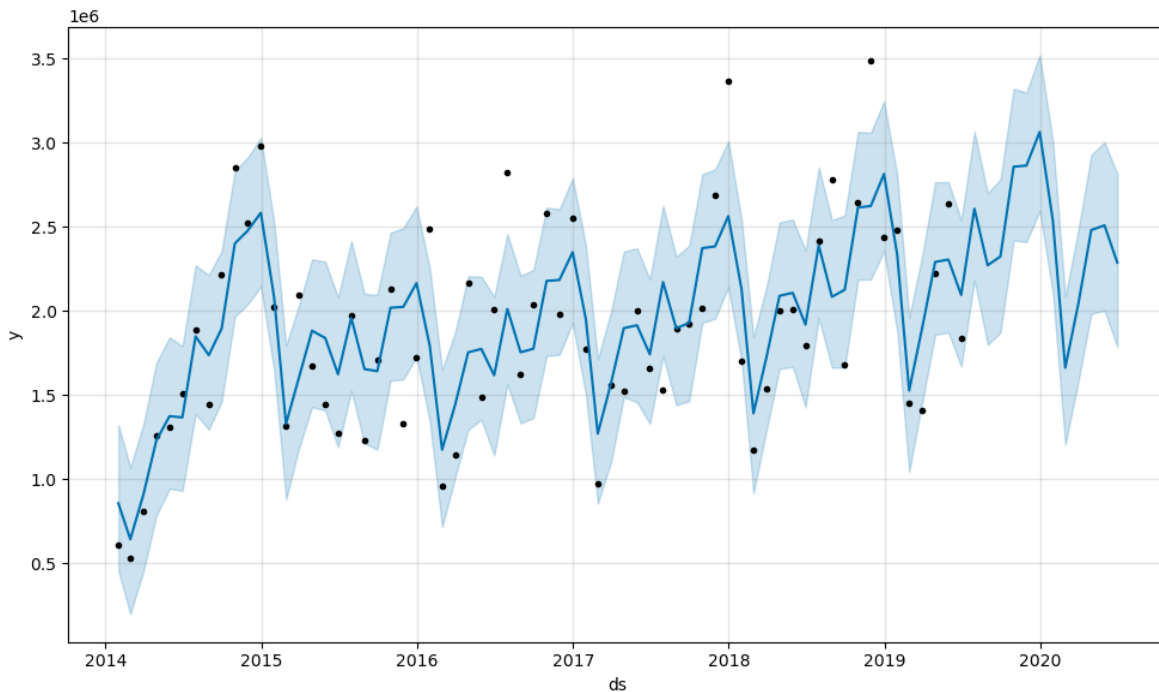


Figura 30: Modelo Prophet resultado de entrenamiento

Fuente: Elaboración Propia

El siguiente modelo que utilizaremos corresponde a NeuralProphet. NeuralProphet es una extensión de Prophet que utiliza redes neuronales para modelar las tendencias y la estacionalidad en los datos de series temporales. A diferencia de Prophet, que utiliza un enfoque aditivo, NeuralProphet emplea una red neuronal recurrente (RNN) para capturar patrones complejos y no lineales en los datos temporales. Esto le permite manejar de manera más efectiva relaciones temporales complejas y capturar características más sutiles en los datos.

Al realizar la optimización de parámetros con el modelo NeuralProphet se obtuvieron como resultado de los parámetros los valores:

Tabla 9: Parámetros NeuralProphet

Fuente: Elaboración Propia

Parámetro	Valor
epochs	700
learning_rate	0.3
seasonality_mode	additive

En el contexto del aprendizaje automático y el entrenamiento de modelos, "Log-SmoothL1Loss" es una función de pérdida utilizada para medir la discrepancia entre las predicciones del modelo y los valores reales en un problema de regresión. La "SmoothL1Loss" es una versión suavizada de la función de pérdida L1 (también conocida como pérdida de valor absoluto) que se utiliza para reducir el impacto de los valores atípicos en el cálculo de la pérdida.

Esto lo hace especialmente útil en problemas de regresión donde los datos pueden contener valores atípicos o extremos que pueden afectar negativamente la eficacia de la pérdida cuadrática tradicional.

Por otra parte, los "*epochs*" son simplemente iteraciones completas a través del conjunto de datos de entrenamiento durante el proceso de entrenamiento de un modelo de aprendizaje automático. En cada *epoch*, el modelo ajusta sus parámetros para minimizar la función de pérdida, lo que generalmente implica actualizar los pesos de las conexiones en una red neuronal utilizando algoritmos de optimización como el descenso de gradiente. Cuantas más *epochs* se realizan, más oportunidades tiene el modelo de aprender patrones complejos en los datos y mejorar su rendimiento predictivo.

Cuando se entrena un modelo de aprendizaje automático, como una red neuronal, el objetivo es minimizar esta función de pérdida. Por lo tanto, una disminución en el valor de la "Log-SmoothL1Loss" a medida que aumentan los *epochs* indica que el modelo está mejorando su capacidad para hacer predicciones precisas en el conjunto de datos de entrenamiento a lo largo del tiempo. Esto es lo que observamos en la Figura 31, donde a medida que los *epochs* aumentan el *Log-SmoothL1Loss* comienza a disminuir y a estabilizarse.

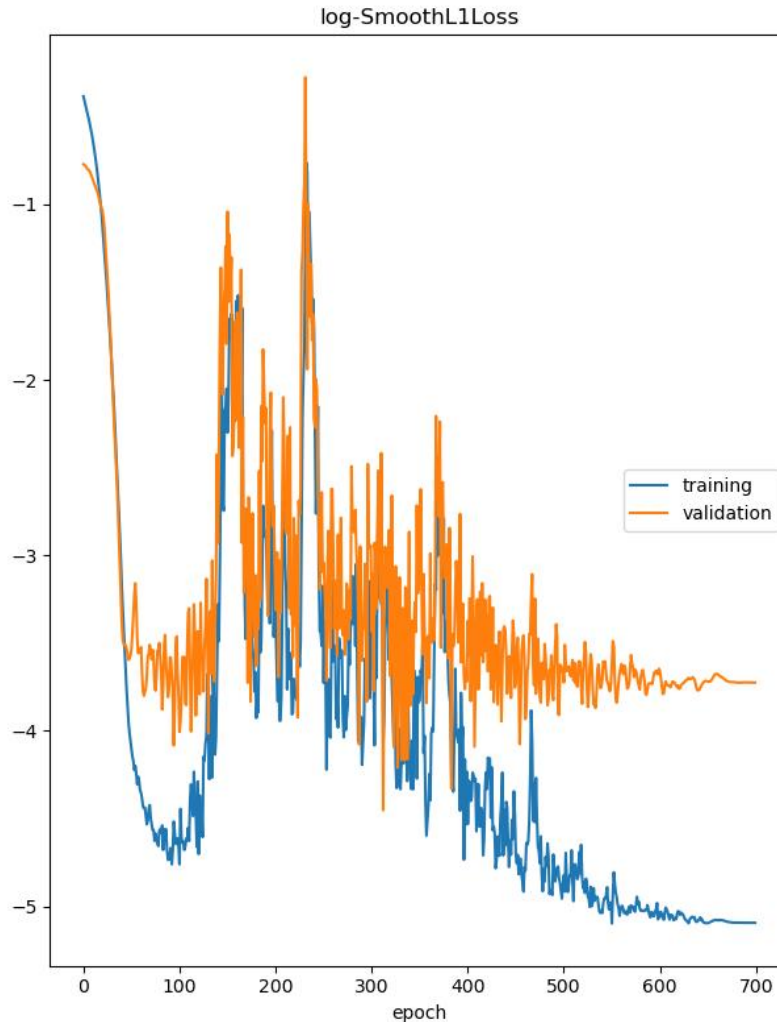


Figura 31: logLoss NeuralProphet

Fuente: Elaboración Propia

Por último, tenemos el modelo Holt-Winters, el cual es un método clásico de suavizado exponencial utilizado para predecir series temporales con tendencia y estacionalidad. Se compone de tres componentes principales: suavizado exponencial para la nivelación, suavizado exponencial doble para la tendencia y suavizado exponencial triple para la estacionalidad. Holt-Winters calcula de forma iterativa los niveles, tendencias y factores de estacionalidad para realizar predicciones futuras. Es especialmente útil para datos que exhiben patrones estacionales claros y cambios en la tendencia a lo largo del tiempo.

Al realizar la optimización de parámetros en el modelo Holt-Winters, se obtuvieron como resultado los valores:

Tabla 10: Parámetros Holt-Winters
Fuente: Elaboración Propia

Parámetro	Valor
damped_trend	True
seasonal	add
seasonal_periods	12
trend	add

Con estos valores se configura el modelo, una vez configurado el modelo podemos observar ciertas métricas en la Figura 32, donde se puede observar la correlación de los residuos, la correlación parcial de los residuos y los residuos.

Como mencionamos anteriormente la correlación de los residuos es una medida de la relación lineal entre los residuos en diferentes momentos en el tiempo. Si los residuos muestran autocorrelación significativa en ciertos rezagos, puede indicar que el modelo no está capturando toda la información temporal en los datos. En este caso podemos observar que la correlación se mantiene dentro de las bandas de confianza.

Por otro lado, la autocorrelación parcial de los residuos mide la correlación entre los residuos en dos momentos en el tiempo, controlando el efecto de los rezagos intermedios. Es útil para identificar la dependencia temporal directa entre los residuos, una vez que se ha eliminado la influencia de los rezagos intermedios. En este gráfico observamos como los valores pueden en ocasiones obtener valores muy altos y escapar de la banda de confianza.

Similar a lo visto con SARIMA tenemos en nuestro gráfico los residuos estandarizados, esto debido a que como manejamos valores muy altos en las predicciones, los errores también son muy altos, para obtener valores más simples de interpretar utilizamos los residuos estandarizados, en este caso se observan valores cercanos al modelo SARIMA, lo cual está dentro de lo esperado y nos indica que el modelo se ha modelado correctamente.

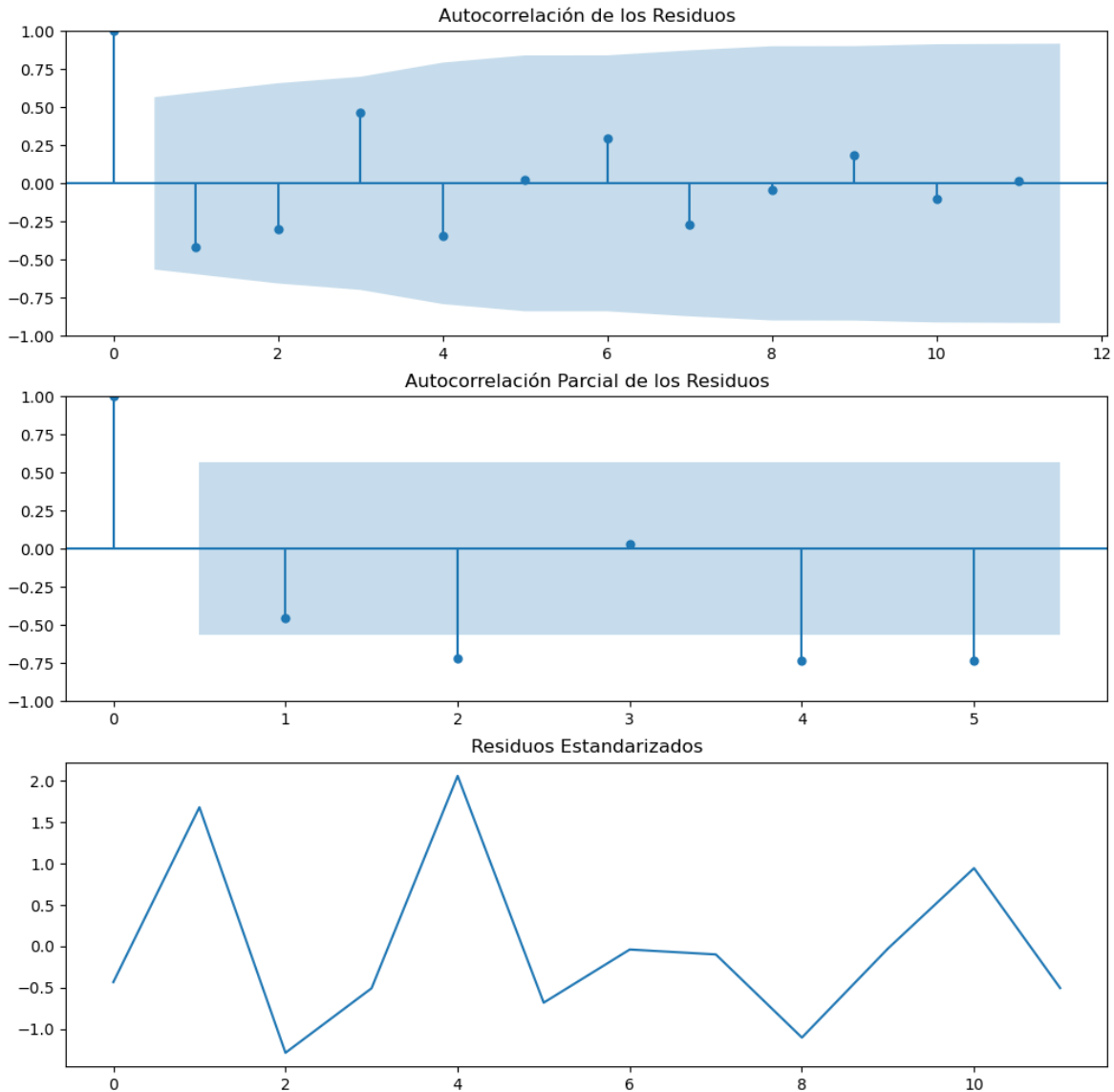


Figura 32: Diagnóstico Holt-Winters

Fuente: Elaboración Propia

3.4 REPORTERIA

Como hemos mencionado anteriormente, uno de los beneficios de realizar una limpieza de datos correcta es que nos permite trabajar de manera más certera en todo lo relacionado a la inteligencia de negocios, dentro de los trabajos más útiles corresponde a la realización de sistemas de reportería, donde podremos crear reportes donde se puedan analizar diferentes métricas, revisar KPI's, realizar gráficos, etc.

Para trabajar con reportería utilizaremos Power BI, la cual es una potente herramienta de análisis y visualización de datos desarrollada por Microsoft la cual permite realizar el traspaso de datos a información significativa y visualmente atractiva. Con Power BI podemos acceder a los datos desde diferentes fuentes, desde bases de datos hasta hojas

de calculo en excel, también se puede realizar conexiones con los servicios de la nube, dado nuestro trabajo previo se cargará el archivo de excel que ha sido trabajado en la etapa de limpieza. Una vez realizada la carga de nuestro archivo podemos realizar algunos pasos antes de comenzar a trabajar con este. Unos de los fundamentos que hemos hablado corresponde al modelo estrella, actualmente poseemos toda la información en nuestro archivo de excel, pero esto no es lo más óptimo para realizar trabajos de análisis multidimensionales, para esto deberemos seccionar los datos en múltiples tablas que puedan relacionarse con la estrategia del modelo estrella.

Para iniciar con este procedimiento se deben definir cuales serán las tablas de hecho y las tablas de dimensión. En este caso la tabla de hecho corresponde a aquellos valores numéricos que buscamos analizar, mientras que las tablas de dimensiones corresponden a los valores descriptivos por los cuales buscamos filtrar la información.

Las tablas de dimensión corresponderán a “Countries”, “Projects”, “Products”, “Customers” y “Owners”. Estas almacenarán la información relacionada a estos campos.

- **Countries:** Tabla de dimensión que almacena la información de los países. Posee el campo “Countryname”.
- **Projects:** Tabla de dimensión que almacena la información de los proyectos. Posee los campos “Proj Created Dt”, “Project Id”, “Project State”, “Projectworkflowstatename”.
- **Products:** Tabla de dimensión que almacena la información de los productos. Posee los campos “Accouting Class”, “Category”, “SKU Description”, “SKU Number”. Además se realizó una jerarquía entre los campos “Category”, “Accouting Class” y “SKU Number”.
- **Customers:** Tabla de dimensión que almacena la información de los clientes. Posee el campo “Customer Number”.
- **Owners:** Tabla de dimensión que almacena la información de los propietarios. Posee el campo “Owner”.

La tabla de hechos corresponderá a “Sales”, la cual constará con los campos “Countryname”, “Customer Number”, “Owner”, “Project Id” y “SKU Number” que actúan como llaves foráneas que permiten relacionar la tabla de hecho con las de dimensión, también posee un campo que actúa como identificador de la tabla que corresponde a “Order Num” y finalmente están los campos numéricos que nos interesa analizar, siendo “Item Quantity”, “SKU USD Unit Price”, “Total USD Margin” y “Total USD Price”. Finalmente también posee el campo “Task Order Date”, que representa la fecha de la compra. La separación y configuración de cada tabla se puede observar en la Figura 33.

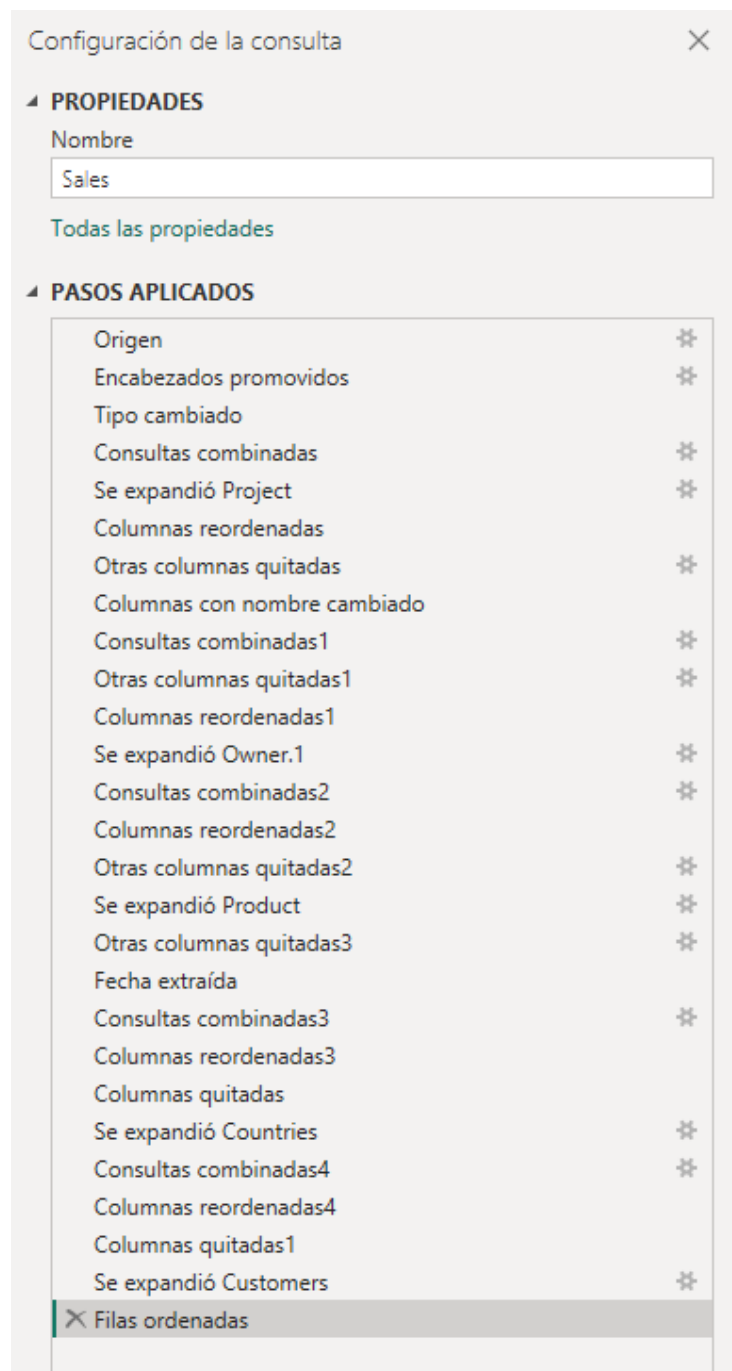


Figura 33: Pasos aplicados tabla Sales

Fuente: Elaboración Propia

El resultado final de la tabla de hechos “Sales” se puede observar en la Figura 34, esto correspondería a el resultado de las transformaciones comentadas recientemente.

Country	Order Name	Project ID	Customer Number	Order Date	SKU Number	Item Quantity	Unit Price	Total Price	Total Margin
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	310-ACDU	4	455.06	1820.24	1819.09
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	402-ACDU	4	22.06	88.24	75.12
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	591-88CH	4	0	0	-1056.148
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	997-3274	4	200	800	687.2
HONDURAS	L_Jim	381806	94920502	20/07/2016	997-3274	2	500	1000	853.2
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	997-3241	4	0	0	0
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	955-0418	4	799	3196	2943.4
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	332-1286	4	0	0	0
HONDURAS	L_Jim	381806	94920502	20/07/2016	332-1286	2	0	0	0
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	911-0418	4	0	0	-0.04
HONDURAS	L_Jim	381806	94920502	20/07/2016	911-0418	2	0	0	-0.02
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	996-0349	4	539	2076	1176
HONDURAS	L_Jim	381806	94920502	20/07/2016	996-0349	2	539	1078	588
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	340-44X8	4	0	0	-86.238
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	330-88CO	4	17.85	71.4	26.012
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	330-88CO	4	17.85	71.4	26
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	374-88WJ	4	0	0	-4.266
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	340-88CW	4	135.37	541.08	368.28
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	406-88EB	8	379.75	3038	3056.72
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	340-88BW	4	26.77	107.08	-30.96
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	385-88HD	4	167.89	671.56	643.96
HONDURAS	L_Jim	381806	94920502	20/07/2016	385-88HD	2	159	318	218
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	634-88WJ	4	35.31	141.22	141.12
HONDURAS	L_Jim	381806	94920502	20/07/2016	634-88WJ	2	35.36	67.12	67.12
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	350-88EN	4	0	0	471.864
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	350-88EJ	4	17.49	69.96	28.888
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	384-88BL	4	0	0	0
HONDURAS	L_Jim	381806	94920502	20/07/2016	384-88BL	2	0	0	0
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	780-88BK	4	0	0	0
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	405-44EG	4	213.79	855.16	233.404
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	338-88PI	4	795.55	3182.2	-2531.8
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	374-88BK	4	0	0	0
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	370-48VW	32	263.75	8440	896.96
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	370-48UP	4	0	0	0
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	370-48AP	4	0	0	0
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	400-44XK	8	117.42	939.36	348.04
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	631-44UG	4	0	0	-8.18
HONDURAS	Mun_Yee_Chow	279479	94920502	20/05/2016	319-1391	4	35.31	141.22	141.12

Figura 34: Muestra tabla Sales
Fuente: Elaboración Propia

Una vez creadas nuestras tablas nos quedan unos pasos para finalizar el modelo, partiendo por la creación de una tabla con medidas calculadas, esta nos será de utilidad para definir valores calculados relacionados a nuestros campos, logrando una mayor organización y facilidad en la creación de reportes. Las medidas corresponden a las que se observan en la Figura 35, y la función que las define se puede observar en la Tabla 11.

Medidas Principales
☐ Año Actual
☐ Año Anterior
☐ Porcentaje de margen
☐ Suma de margen
☐ Suma Unidades
☐ Variacion Ventas
☐ Ventas LYTD
☐ Ventas totales
☐ Ventas YTD

Figura 35: Tabla Medidas Principales
Fuente: Elaboración Propia

Tabla 11: Variables tabla Medidas Principales

Fuente: Elaboración Propia

Medida	Función DAX
Año Actual	<code>MAX('Calendar'[Date].[Año])</code>
Año Anterior	<code>'Medidas Principales'[Año Actual]-1</code>
Porcentaje de margen	<code>DIVIDE(SUM('Sales'[Total USD Margin]), SUM('Sales'[Total USD Price]))</code>
Suma de margen	<code>SUM(Sales[Total USD Margin])</code>
Suma Unidades	<code>SUM(Sales[Item Quantity])</code>
Variacion Ventas	<code>(([Ventas YTD]-[Ventas LYTD])/[Ventas LYTD])</code>
Ventas LYTD	<code>VAR X = [Año Anterior] VAR Y = [Año Actual] RETURN CALCULATE([Ventas totales], 'Calendar'[Date].[Año]=X, CALCULATETABLE(VALUES('Calendar'[Date].[Mes]), 'Calendar'[Date].[Año]=Y))</code>
Ventas totales	<code>SUM(Sales[Total USD Price])</code>
Ventas YTD	<code>VAR X = [Año Actual] RETURN CALCULATE([Ventas totales], 'Calendar'[Date].[Año]=X)</code>

La siguiente tabla que se debe realizar es la tabla calendario, esto es una práctica muy común en el ámbito de la inteligencia de negocios, crear una tabla de calendario nos permite filtrar por año, mes y día, también nos permite relacionar esta tabla con todos los campos de tipo fecha, de manera que se puedan realizar análisis históricos y tendencias, agrupación y filtrado de fechas, integración con visualizaciones temporales, etc.

La función que nos permite crear la tabla es:

Calendar = `CALENDAR(MIN(Sales[Task Order Date]), MAX(Sales[Task Order Date]))`

El resultado de la tabla se puede observar en la Figura 36.

1 2 Calendar = CALENDAR(MIN(Sales[Task Order Date]), MAX(Sales[Task Order Date]))				
Date	Year	Month	Day	
01/01/2014	2014	1	1	
02/01/2014	2014	1	2	
03/01/2014	2014	1	3	
04/01/2014	2014	1	4	
05/01/2014	2014	1	5	
06/01/2014	2014	1	6	
07/01/2014	2014	1	7	
08/01/2014	2014	1	8	
09/01/2014	2014	1	9	
10/01/2014	2014	1	10	
11/01/2014	2014	1	11	
12/01/2014	2014	1	12	
13/01/2014	2014	1	13	
14/01/2014	2014	1	14	
15/01/2014	2014	1	15	
16/01/2014	2014	1	16	
17/01/2014	2014	1	17	
18/01/2014	2014	1	18	
19/01/2014	2014	1	19	
20/01/2014	2014	1	20	
21/01/2014	2014	1	21	
22/01/2014	2014	1	22	
23/01/2014	2014	1	23	
24/01/2014	2014	1	24	
25/01/2014	2014	1	25	
26/01/2014	2014	1	26	
27/01/2014	2014	1	27	
28/01/2014	2014	1	28	
29/01/2014	2014	1	29	
30/01/2014	2014	1	30	
31/01/2014	2014	1	31	
01/02/2014	2014	2	1	

Figura 36: Tabla Calendar

Fuente: Elaboración Propia

Una vez que hemos realizado los pasos anteriores hemos obtenido múltiples tablas las cuales nos permiten construir nuestro modelo de estrella, recordemos que el modelo de estrella posee nuestra tabla de hechos con los campos de valores numéricos relacionados al proceso de ventas, la tabla de hecho se relaciona con múltiples tablas de dimensión las cuales describen los atributos por los que se puede filtrar los datos numéricos.

Como podemos observar en la Figura 37, nuestra tabla de hecho corresponde a “Sales”, mientras que “Calendar”, “Countries”, “Projects”, “Owners”, “Products” y “Customers” corresponden a las tablas de dimensión. La tabla de “Medidas Principales” no se asocia como el resto con nuestra tabla de hechos, sin embargo, los valores de sus campos dependen del resto de tablas. Podemos notar que la relación de las tablas de dimensión con la de hechos es de uno a muchos, debido a que estas tablas representan la descripción de los atributos de esta. Además de esto podemos notar los campos que posee cada una de las tablas y también podemos observar cómo ciertos campos de la tabla de hechos fueron ocultados, esto es debido a que corresponden a las llaves foráneas de la tabla de hechos y podemos acceder a estos campos desde las otras tablas.

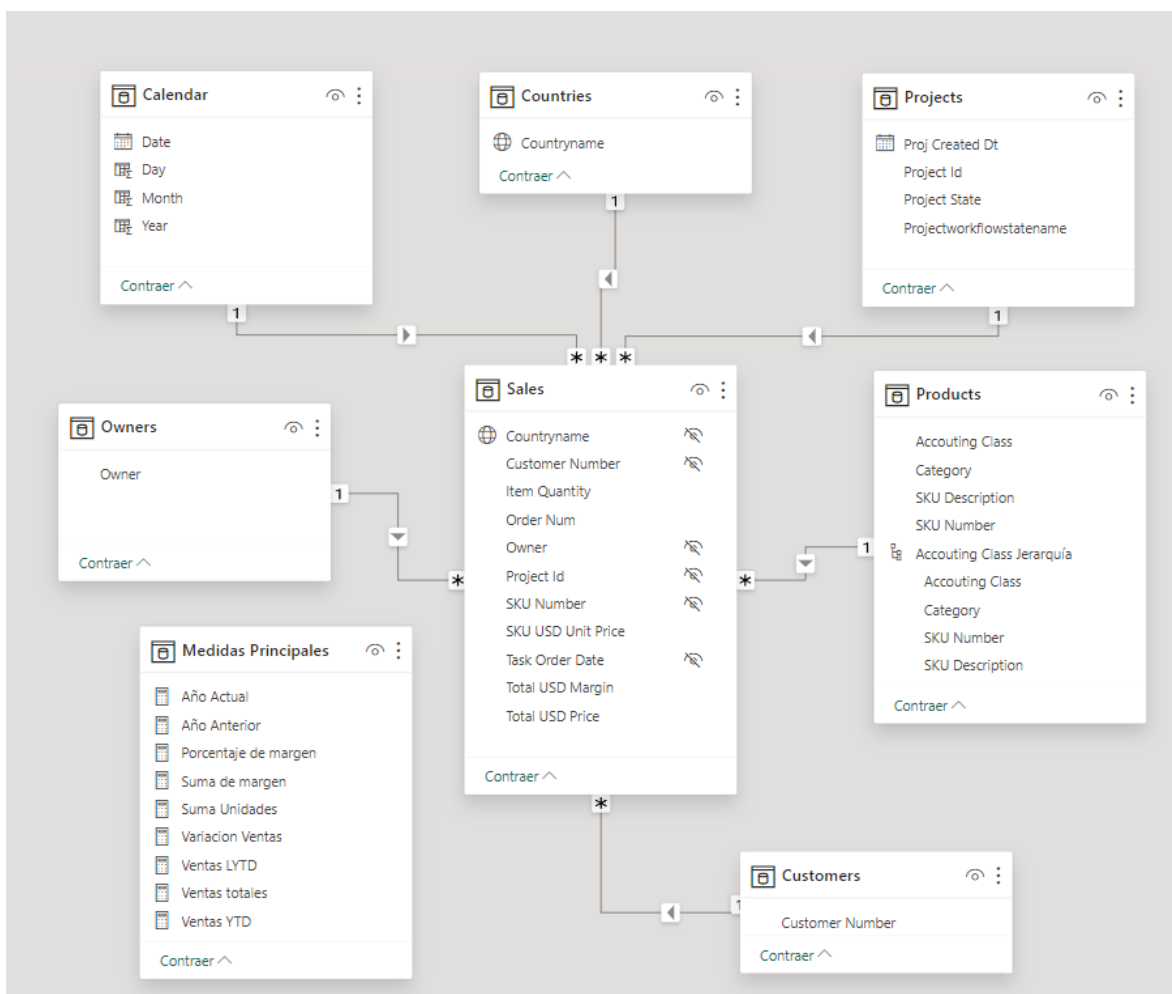


Figura 37: Modelo Estrella Power BI

Fuente: Elaboración Propia

Con esto realizado podemos trabajar con nuestro modelo estrella construido.

CAPÍTULO 4: VALIDACIÓN DE LA SOLUCIÓN

4.1 METODOLOGÍA DE VALIDACIÓN

La etapa de validación de la solución es un aspecto clave en nuestro trabajo, pues nos permite demostrar que la solución propuesta es válida para el entorno donde fue planteada. En nuestro trabajo realizamos tres etapas principales, la primera corresponde a la limpieza de datos, donde solucionamos los problemas de carga y realizamos una transformación en el CSV, los registros y las columnas, de modo que nuestro trabajo sea de mayor calidad. En segundo lugar, generamos los modelos predictivos que nos permitirán conocer la demanda que tendrá Dell Technologies en un futuro, permitiendo un mejor manejo de recursos y una toma de decisiones con más información. Finalmente, la generación de reportes en Power BI, que le permitirá a Dell Technologies la observación de métricas e indicadores de manera multidimensional permitiendo cubrir la información necesaria para la compañía.

Para comprobar que nuestro trabajo ha sido realizado con éxito analizaremos en profundidad los resultados de los modelos realizados, los cuales corresponden a SARIMA, Prophet, NeuralProphet y Holt-Winters.

Estos cuatro modelos se evaluarán en distintos contextos tomando en consideración dos dimensiones, el país donde se realizaron las ventas, donde consideraremos tres casos, el primer caso corresponde a las ventas de Chile, en segundo lugar, las ventas de Brasil y por último las ventas del resto de los países. La otra dimensión que utilizaremos será el año que buscamos predecir, si utilizamos el año 2018 se generará una predicción de las ventas de este año utilizando los datos hasta el año 2017 para la creación de nuestro modelo.

Para cuantificar los resultados de los modelos utilizaremos cuatro métricas de error, las cuales corresponden:

- **MSE (Mean Squared Error):** Mide el promedio de los cuadrados de los errores entre los valores predichos por el modelo y los valores reales. La fórmula es la siguiente:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **RMSE (Root Mean Squared Error):** El RMSE es simplemente la raíz cuadrada del MSE. Esto se hace para devolver la métrica a la misma escala que los datos originales, lo que hace que sea más fácil de interpretar. La fórmula es la siguiente:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- **MAE (Mean Absolute Error):** Es la media de las diferencias absolutas entre los valores predichos y los valores reales. Es decir, toma el valor absoluto de la diferencia entre cada valor predicho y el valor real, y luego calcula el promedio de estas diferencias. La fórmula es la siguiente:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- **MAPE (Mean Absolute Percentage Error):** El MAPE es similar al MAE, pero en lugar de calcular el error absoluto como una diferencia entre los valores reales y predichos, calcula el error absoluto como un porcentaje del valor real. La fórmula es:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

Donde:

- y_i son los valores reales
- $\hat{y}_i$ son los valores predichos
- n son el número de observaciones

Para esto se generará una tabla para cada modelo de predicción, que contendrá todos los errores mencionados evaluando cada contexto en específico, además se obtendrá un promedio de cada uno de los errores para cada modelo en particular.

4.2 VALIDACIÓN FORECASTING

Comenzaremos la validación de las predicciones por observar a grandes rasgos los resultados de cada modelo en específico, comenzando por la predicción de SARIMA.

En la Figura 38 podemos observar la predicción realizada por SARIMA, el gráfico muestra tres líneas, la primera es la azul que corresponde a los datos con los que se realizó el entrenamiento del modelo, recordemos que corresponden a los registros de hasta un año antes de la fecha máximo de los datos.

Por otro lado, podemos observar en la línea roja los valores que se utilizarán como validación del modelo y son los registros que intentamos predecir, estos corresponden a los datos del último año del *dataset*.

Finalmente, la línea verde corresponde a la predicción realizada por nuestro modelo, esta posee la misma cantidad de registros que nuestros datos de validación y corresponden a los valores predichos por el modelo, estos serán comparados y generarán un error el cual

servirá como métrica de desempeño del modelo y nos permitirá encontrar el mejor modelo para nuestro contexto.

Como se puede observar la línea de predicción realizada por el modelo SARIMA no se aleja de los valores reales, salvo el comienzo donde se puede observar que la predicción es que las ventas bajarán sin embargo suben, a partir de acá el modelo responde de manera satisfactoria generando un buen resultado en general.

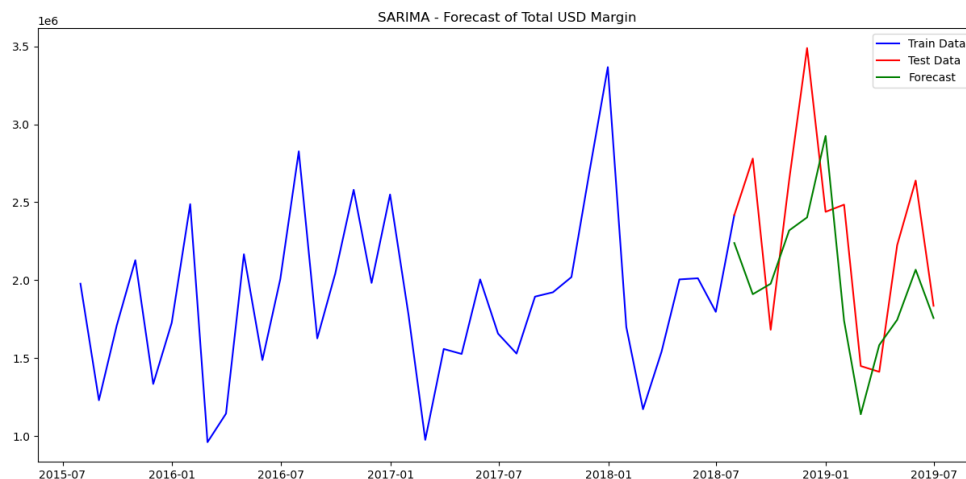


Figura 38: SARIMA Forecast

Fuente: Elaboración Propia

En la Figura 39 podemos analizar la predicción del modelo Prophet, esta de manera similar a el modelo SARIMA falla en el primer mes predicho pero se repone en los siguientes y finaliza con una predicción bastante acertada de las ventas de los últimos meses.

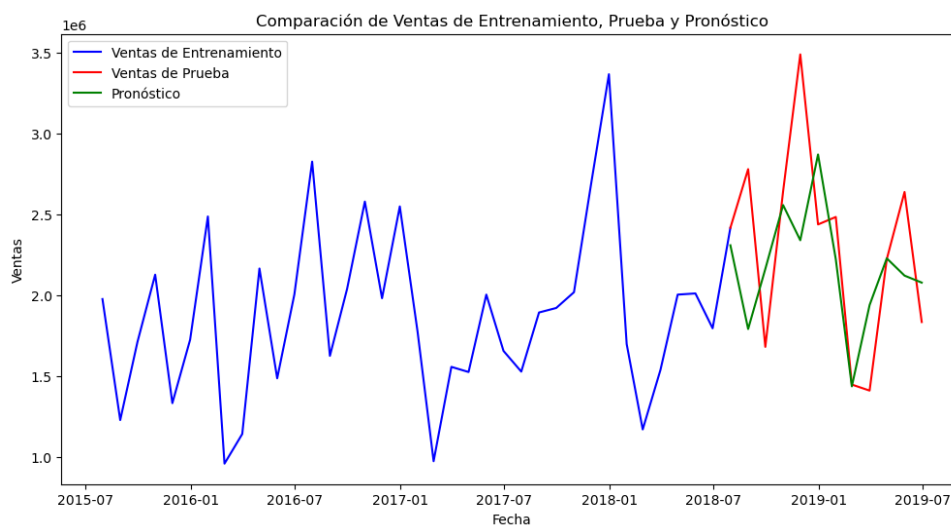


Figura 39: Prophet Forecast

Fuente: Elaboración Propia

En la Figura 40 podemos observar el caso de NeuralProphet, este es muy similar a el modelo Prophet y falla en lo mismo que los otros vistos que es el mes de inicio, luego de esto el modelo es preciso en sus resultados generando una predicción certera.

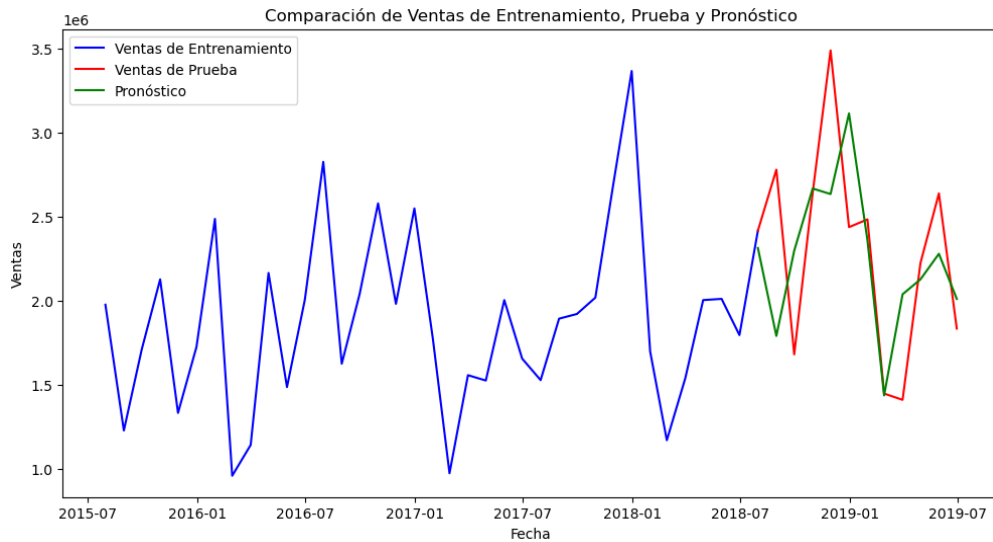


Figura 40: NeuralProphet Forecast

Fuente: Elaboración Propia

Finalmente, en la Figura 41, se puede observar el análisis de Holt-Winters, siguiendo la tónica de los anteriores modelos falla en un comienzo acertando en ciertos meses de manera muy certera, pero en general es el modelo que ha generado mayores errores de predicción al menos para este periodo de tiempo.

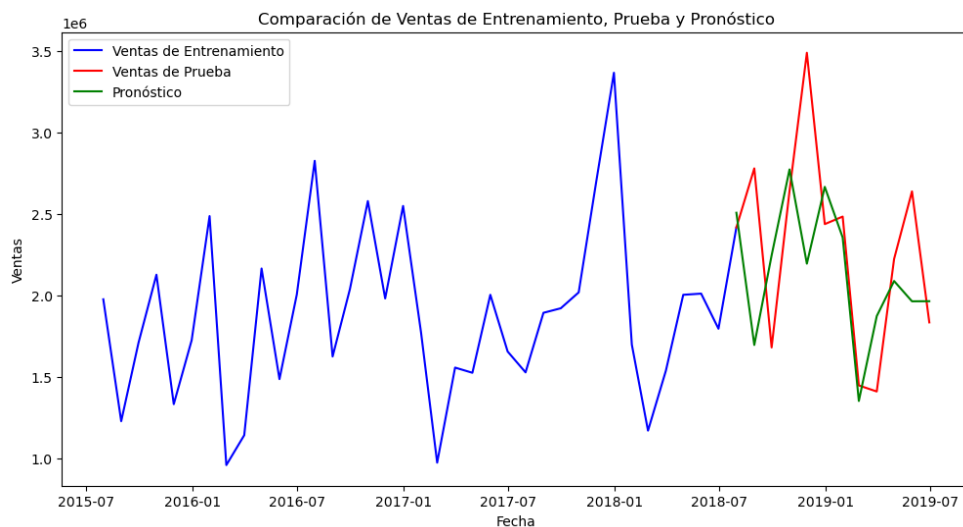


Figura 41: Holt-Winters Forecast

Fuente: Elaboración Propia

Comenzaremos el análisis de los modelos para cada uno de los contextos mencionados anteriormente, la Tabla 12 muestra los errores del modelo SARIMA, se puede observar un error constante en cada uno de los contextos, mostrando una consistencia en la generación de resultados, en general los resultados son aceptables y permite trabajar desde una base.

Tabla 12: Tabla de errores SARIMA

Fuente: Elaboración Propia

Modelo	País	Año	MSE	RMSE	MAE	MAPE
SARIMA	Chile	2016	4.28e+11	6.54e+05	5.67e+05	0.295
		2017	3.78e+11	6.15e+05	5.00e+05	0.257
		2018	4.95e+11	7.04e+05	5.61e+05	0.260
		2019	4.60e+11	6.78e+05	5.60e+05	0.226
	Brasil	2016	4.27e+11	6.54e+05	5.67e+05	0.295
		2017	3.78e+11	6.15e+05	5.00e+05	0.257
		2018	5.04e+11	7.10e+05	5.69e+05	0.263
		2019	4.60e+11	6.78e+05	5.60e+05	0.226
	Otros	2016	4.27e+11	6.54e+05	5.67e+05	0.295
		2017	3.78e+11	6.15e+05	5.00e+05	0.257
		2018	5.04e+11	7.10e+05	5.69e+05	0.263
		2019	4.60e+11	6.78e+05	5.60e+05	0.226
Promedio			4.42e+11	6.64e+05	5.48e+05	0.260

La Tabla 13 corresponde a los errores que posee el modelo Prophet, en este caso podemos observar que los errores en ciertos periodos son más elevados que el resto, en el año 2017 el error porcentual es cercano a 0.4, lo cual implica una predicción bastante errónea, sin embargo, el error para el año 2019 disminuye significativamente. Prueba de que el modelo con un mayor número de datos puede comenzar a ser más certero que otros modelos. El uso de este modelo puede ser de utilidad cuando se posee un número alto de información para generar el modelo.

Tabla 13: Tabla de errores Prophet

Fuente: Elaboración Propia

Modelo	País	Año	MSE	RMSE	MAE	MAPE
Prophet	Chile	2016	4.20e+11	6.48e+05	5.61e+05	0.290
		2017	6.03e+11	7.77e+05	6.46e+05	0.393
		2018	4.78e+11	6.91e+05	6.23e+05	0.316
		2019	2.83e+11	5.32e+05	4.00e+05	0.170
	Brasil	2016	4.18e+11	6.46e+05	5.60e+05	0.290
		2017	6.03e+11	7.77e+05	6.46e+05	0.393
		2018	4.57e+11	6.76e+05	6.15e+05	0.310
		2019	2.84e+11	5.33e+05	4.01e+05	0.170
	Otros	2017	4.21e+11	6.49e+05	5.62e+05	0.290
		2017	6.03e+11	7.77e+05	6.46e+05	0.393
		2018	4.56e+11	6.75e+05	6.15e+05	0.310
		2019	2.84e+11	5.33e+05	4.01e+05	0.170
Promedio			4.42e+11	6.59e+05	5.56e+05	0.291

Por otra parte, en la Tabla 14 podemos observar los errores que generó el modelo NeuralProphet, este similar a Prophet posee periodos con errores muy grandes, en el caso del año 2018 posee errores en la métrica MAPE superiores a 0.4, lo cual es una predicción bastante errónea, sin embargo, en el resto de los periodos posee un buen desempeño, logrando en promedio errores más bajos que los conseguidos en el modelo Prophet.

Tabla 14: Tabla de errores NeuralProphet

Fuente: Elaboración Propia

Modelo	País	Año	MSE	RMSE	MAE	MAPE
NeuralProphet	Chile	2016	3.39e+11	5.82e+05	5.07e+05	0.278
		2017	4.03e+11	6.35e+05	5.06e+05	0.252
		2018	1.04e+12	1.02e+06	9.05e+05	0.480

		2019	2.64e+11	5.14e+05	3.88e+05	0.172
	Brasil	2016	3.48e+11	5.90e+05	5.14e+05	0.281
		2017	4.06e+11	6.38e+05	4.97e+05	0.241
		2018	7.53e+11	8.68e+05	7.61e+05	0.403
		2019	2.55e+11	5.05e+05	3.88e+05	0.171
	Otros	2017	3.36e+11	5.79e+05	5.04e+05	0.277
		2017	3.65e+11	6.04e+05	4.76e+05	0.237
		2018	9.97e+11	9.98e+05	8.92e+05	0.471
		2019	2.63e+11	5.13e+05	3.87e+05	0.172
Promedio			4.80e+11	6.70e+05	5.60e+05	0.286

Finalmente, en la Tabla 15 podemos observar los errores asociados a el modelo Holt-Winters, en este modelo se posee errores muy altos en un principio los cuales disminuyen drásticamente en un final, si bien posee un error en promedio mayor al del resto de los modelos es un error aceptable para predicciones de ventas.

Tabla 15: Tabla de errores Holt-Winters

Fuente: Elaboración Propia

Modelo	País	Año	MSE	RMSE	MAE	MAPE
Holt-Winters	Chile	2016	1.30e+12	1.14e+06	1.03e+06	0.528
		2017	2.57e+11	5.07e+05	4.61e+05	0.251
		2018	3.95e+11	6.28e+05	5.82e+05	0.288
		2019	3.31e+11,	5.75e+05	4.18e+05	0.176
	Brasil	2016	1.32e+12	1.15e+06	1.04e+06	0.534
		2017	2.46e+11	4.95e+05	4.33e+05	0.239
		2018	3.85e+11	6.20e+05	5.74e+05	0.284
		2019	3.31e+11	5.75e+05	4.18e+05	0.176
	Otros	2017	1.32e+12	1.15e+06	1.04e+06	0.534
		2017	2.57e+11	5.07e+05	4.61e+05	0.251

		2018	3.85e+11	6.20e+05	5.74e+05	0.284
		2019	3.31e+11	5.75e+05	4.18e+05	0.176
Promedio			5.71e+11	7.12e+05	6.21e+05	0.310

Una vez analizado cada modelo en profundidad podemos observar en la Tabla 16 un resumen de los errores. Los valores marcados corresponden a los valores más bajo de cada error. Aquí podemos observar que el modelo SARIMA posee los errores más bajos además de como vimos anteriormente los más persistentes en los diferentes contextos.

Por otro lado, Prophet y NeuralProphet poseen errores similares y como bien vimos anteriormente los errores para el último periodo eran bastante bajos en estos dos modelos.

Para el modelo Holt-Winters se obtuvieron errores mayores, sin embargo, una gran parte de esto fueron los errores de los periodos más antiguos, esto puede deberse a la falta de datos para crear un buen modelo de predicción.

En resumen, podemos concluir que si bien cada modelo ofrece una solución diferente y el mejor modelo depende ciertamente del contexto, en este caso el modelo SARIMA corresponde a el modelo que nos ofrece un mejor resultado para la predicción de ventas de Dell Technologies.

Tabla 16: Resumen de errores

Fuente: Elaboración Propia

Modelo	MSE	RMSE	MAE	MAPE
SARIMA	4.42e+11	6.64e+05	5.48e+05	0.260
Prophet	4.42e+11	6.59e+05	5.56e+05	0.291
NeuralProphet	4.80e+11	6.70e+05	5.60e+05	0.286
Holt-Winters	5.71e+11	7.12e+05	6.21e+05	0.310

4.3 INTERFAZ GRÁFICA DE USUARIO

Como todo el trabajo se realizó por medio de un archivo en formato Jupyter Notebook y por comodidad y sabiendo que no todas las personas se manejan con la programación se realizó un software que permita la realización de la proyección de demanda geo-temporal, donde se permite filtrar por diferentes campos.

Como se observa en la Figura 42 podemos filtrar por país, categoría y fecha, además se puede seleccionar el modelo con el cual queremos predecir nuestros resultados.

Figura 42: GUI Forecast Menú

Fuente: Elaboración Propia

Una vez seleccionados los campos y el modelo podemos hacer clic en “Submit” y esto permitirá la creación, evaluación y muestra del modelo y los resultados, agregando además de el gráfico y las bandas de confianza los valores en específico en el sector de abajo.

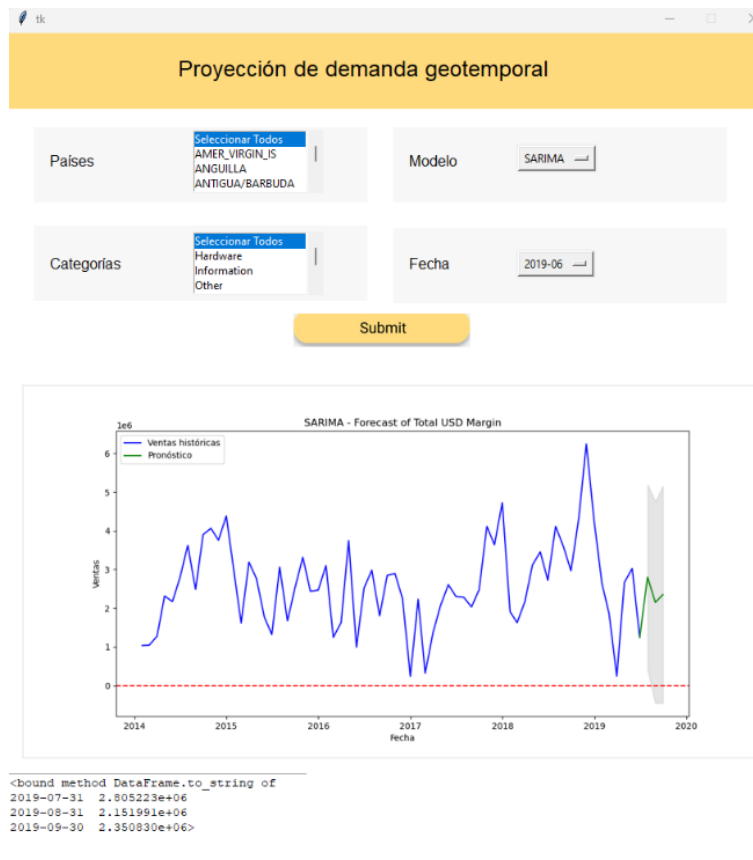


Figura 43: GUI Forecast ejemplo ejecución

Fuente: Elaboración Propia

Esto permitirá realizar predicciones de las ventas mediante un programa sencillo de utilizar.

4.4 REPORTERÍA EN POWER BI

Finalmente, se ha realizado el trabajo de reportería en Power BI, donde se ha confeccionado un reporte de cuatro páginas, donde se incluyen elementos claves para cubrir la información esencial que posee el proyecto.

La primera página de Power BI se puede observar en la Figura 44, en esta poseemos un filtro con los años para enfocarnos en los datos anuales que poseemos, a la derecha se observan indicadores claves que nos permiten saber de manera general la situación actual de la empresa. Por otra parte, y más debajo se encuentran dos gráficos, el primero corresponde a un gráfico de ventas y margen para cada país en específico, el cual va asociado a una línea de tiempo, esta se puede reproducir y muestra los cambios de los valores a través del tiempo, por otra parte, se puede observar las ventas totales y el margen para ciertos días, meses o años en específico, esto nos permite reconocer de manera rápida cuales son los días y meses claves para la compañía.

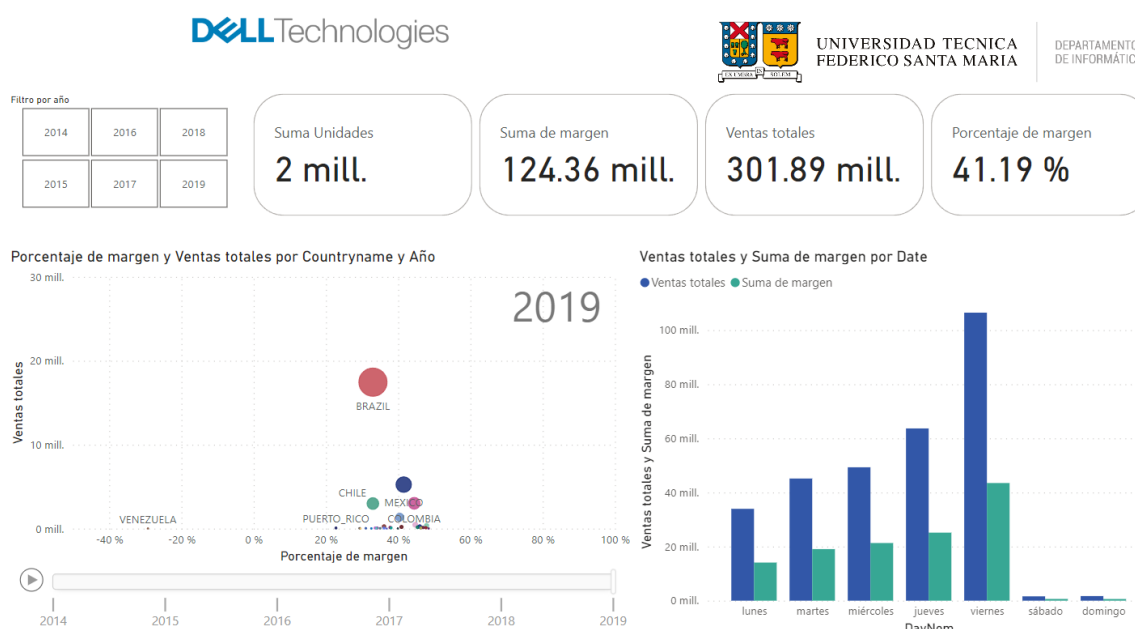


Figura 44: Power BI página 1

Fuente: Elaboración Propia

Por otra parte en la Figura 45 nos encontramos con la página dos de nuestro reporte en Power BI, esta página está enfocada en una comparativa anual frente al año anterior respecto a las ventas. Aquí podemos filtrar anual y mensualmente las ventas, además de esto se observan indicadores con las ventas anuales del año seleccionado y el año anterior y un porcentaje de las variaciones de las ventas. También se muestran los valores en forma de gráficos para observar de manera más amigable la comparativa y si queremos datos más concretos tenemos la tabla con las ventas por países en ambos años y la variación para cada uno de ellos.

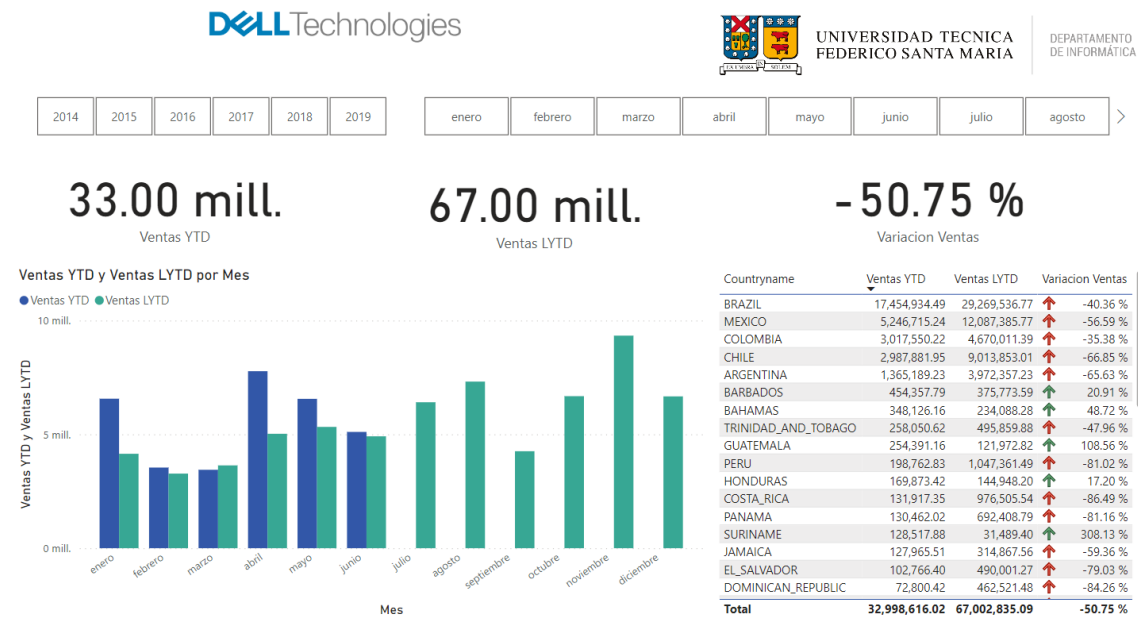


Figura 45: Power BI página 2
Fuente: Elaboración Propia

En la Figura 46 se observa la página 3 del reporte, esta página nos permite obtener información más gráfica de las ventas en diferentes dimensiones, tenemos filtros anuales y mensuales y gráficos que nos permiten observar el aporte del margen de cada propietario al margen total recaudado, también las ventas por país y por categoría.

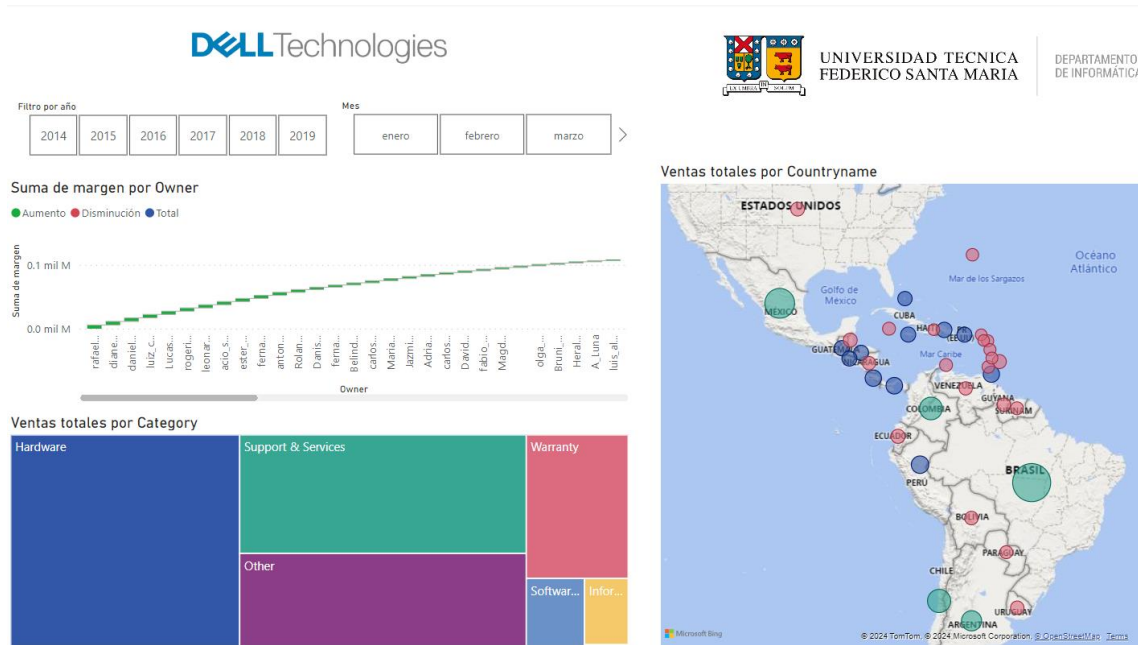


Figura 46: Power BI página 3
Fuente: Elaboración Propia

Finalmente llegamos a la última página de nuestro reporte en Power BI, la cual corresponde a la muestra de ventas, margen y unidades vendidas a través del tiempo, las cuales se pueden filtrar por propietario, categoría, país y además seleccionar el intervalo de fechas a considerar.

Los tres gráficos poseen además una línea de previsión, el cual cumple la labor de realizar también una proyección de los valores, a esta se le asigna una duración de 12 meses, omitiendo los últimos 12 meses y asignándole una estacionariedad de 12 meses. Esto permite obtener resultados del mismo tipo que los vistos con anterioridad.

En este caso los valores obtenidos para la predicción de la suma del margen mensual fueron los siguientes.

- MSE: $3.68e+11$
- RMSE: $6.06e+05$
- MAE: $4.09e+05$
- MAPE: 0.168

Considerando todos los datos que se poseen, este error es menor que el del resto de los modelos lo que nos muestra que la alternativa de realizar las previsiones con Power BI es altamente efectiva, siempre y cuando el *dataset* esté bien trabajado.

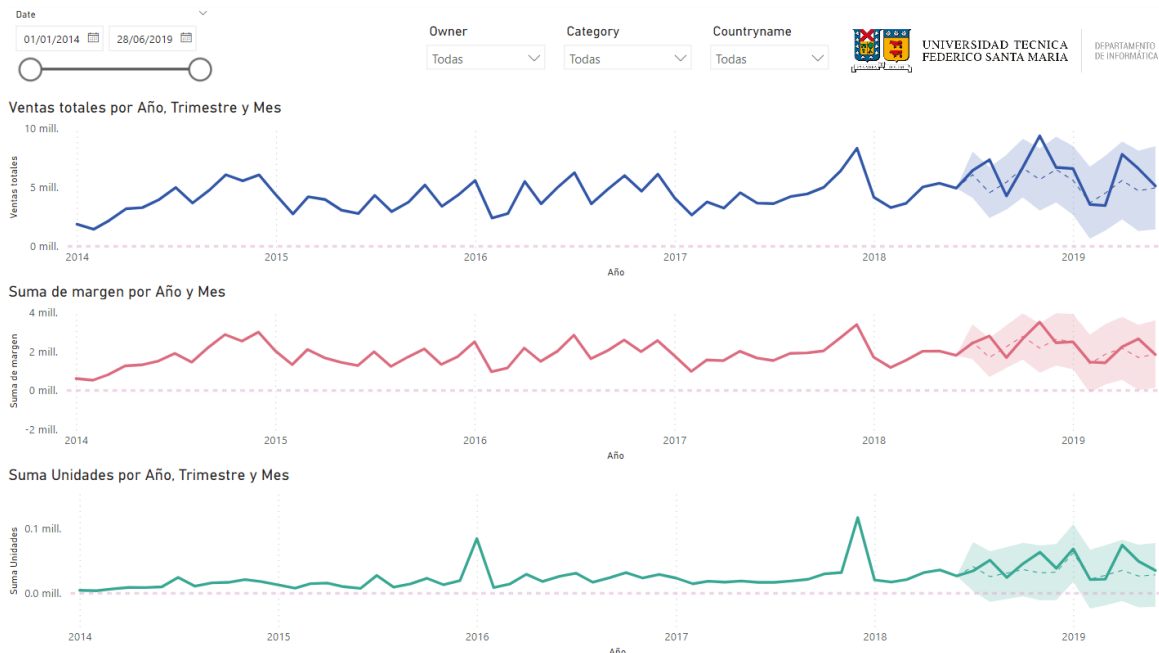


Figura 47: Power BI página 4

Fuente: Elaboración Propia

CAPÍTULO 5: CONCLUSIONES

En este capítulo sintetizaremos a grandes rasgos múltiples aspectos del trabajo realizado, considerando los alcances y limitaciones de la propuesta de la solución, estableciendo de forma simple y directa los resultados, analizando el cumplimiento de los objetivos previamente formulados, identificando las contribuciones, aplicaciones y aporte del trabajo realizado para Dell Technologies finalizando con recomendaciones para futuras mejoras a el trabajo realizado.

En primer lugar, es importante destacar los alcances y limitaciones del trabajo realizado. Se lograron resultados significativos en la limpieza de los datos, así como en la implementación de modelos predictivos para la proyección de la demanda y el trabajo de reportería realizado en Power BI. Sin embargo, en un futuro puede que el *dataset* de ventas comience a desarrollar nuevos errores las cuales demandarán un trabajo de limpieza que pueda abarcar los problemas futuros, por lo cual este proceso si bien es efectivo para el *dataset* actual, puede que en un futuro necesite de arreglos. Además, ocurre lo mismo con los modelos de predicción, a medida que el *dataset* se actualice y almacene un volumen más grande de datos estos modelos deberán actualizarse para que puedan seguir siendo precisos. Finalmente, los resultados de los modelos pueden no ser certeros en un futuro debido a cambios en el comportamiento de la sociedad, este *dataset* contiene información hasta mediados de 2019, sin embargo, durante ese periodo de año y el actual ocurrió una pandemia que afectó de manera significativa las ventas de las empresas, debido a esto puede que los resultados actuales sean difíciles de predecir, a menos que se realice una actualización a los modelos con la información más actualizada.

En cuanto a los resultados obtenidos, se evidenció la efectividad de la propuesta de solución generada durante el trabajo, obteniendo un conjunto de datos de calidad a diferencia de la poseída en un comienzo, permitiendo una generación de proyecciones precisas y útiles y unos reportes de análisis multidimensional que potencia la toma de decisiones estratégicas. Los modelos predictivos desarrollados demostraron ser capaces de anticipar tendencias y variaciones en la demanda, permitiendo a Dell Technologies planificar de manera más eficiente sus operaciones y recursos, mientras que el sistema de reportes permite resolver consultas oportunas de las métricas más relevantes con un enfoque multidimensional.

En términos de contribuciones y aplicaciones, este trabajo ofrece una herramienta valiosa para la gestión de la demanda en el mercado latinoamericano, con potenciales aplicaciones en otras áreas y contextos empresariales. La implementación de la propuesta de solución puede tener un impacto significativo en la competitividad y el éxito de la empresa, al permitir una mejor anticipación y gestión de la demanda de sus productos.

5.1 CUMPLIMIENTO OBJETIVO GENERAL

El objetivo general definido en el Capítulo 1 es:

Diseñar, construir e implementar un proceso de análisis mediante técnicas de Inteligencia de Negocios aplicado a las ventas de la empresa Dell Technologies en Latinoamérica

durante los años 2012 al 2019, con el fin de apoyar la toma de decisiones respecto a las estrategias organizacionales de venta.

El objetivo fue cumplido según lo planeado, a lo largo del trabajo del documento se ha ido trabajando en las áreas que hemos definido en un comienzo, se diseñó una metodología centrada en análisis multidimensional enfocado en la inteligencia de negocios que permita resolver diferentes necesidades de la empresa.

Para esto se realizó la etapa de ETL, un análisis de datos estadísticos, la creación de modelos de predicción geo-temporal en las ventas de la empresa, reportes enfocados en el análisis multidimensional y además se implementó un software que permita el uso de los modelos predictivos de manera amigable para los usuarios.

En resumen, se logró diseñar, construir e implementar el proceso de análisis en las ventas de Dell Technologies, lo que permitirá apoyar la toma de decisiones respecto a las estrategias organizacionales de venta.

5.2 CUMPLIMIENTO OBJETIVOS ESPECÍFICOS

Los objetivos específicos definidos en el Capítulo 1 fueron:

- **Analizar e identificar los principales errores que existen en la base de datos de ventas de Dell en Latinoamérica.**

Durante la etapa de propuesta de solución hemos identificado los errores que poseían los registros y las columnas del *dataset*, como también el mismo archivo CSV con la manera de almacenar la información, estos errores se repiten y se repetirán en múltiples registros por lo cual es importante definir metodologías de limpieza que permitan atacar estos errores.

Este objetivo ha sido completado al analizar e identificar los errores principales de los datos, permitiendo diseñar un método de limpieza que aumente la calidad de estos.

- **Diseñar un método de limpieza y estandarización de la base de datos.**

Un objetivo que va muy ligado a el anterior, debido a que el método de limpieza y estandarización de los datos debe ir enfocada en los errores encontrados en un comienzo. Toda la metodología se puede encontrar en el Capítulo 3, donde se realiza la propuesta de solución, aquí se ha documentado los errores que posee el conjunto de datos y de qué manera de debe atacar cada error, de manera de poseer un conjunto de datos de calidad.

Este objetivo fue completado debido a que hemos obtenido un conjunto de datos de calidad, que permite el uso de este en múltiples áreas obteniendo buenos resultados, esto permite no tan solo realizar los trabajos que hemos hecho como las proyecciones de ventas o los reportes, sino que permite realizar cualquier trabajo futuro que requiera que el conjunto de datos se encuentre correcto.

- **Realizar una proyección de demanda geo-temporal de las ventas de Dell en Latinoamérica en periodos futuros.**

Corresponde al objetivo principal de nuestro trabajo, donde con el conjunto de datos previamente analizado y corregido se realice una proyección de la demanda geo-temporal de las ventas de Dell en Latinoamérica.

Este objetivo fue cumplido y los resultados han sido bastante certeros para la complejidad de la proyección, se han evaluado cuatro modelos además de la opción entregada por Power BI, e inclusive se ha realizado un software que permita utilizar los modelos.

5.3 TRABAJO FUTURO

En cuanto a recomendaciones y futuras líneas de investigación, se sugiere continuar explorando nuevas técnicas y enfoques para mejorar la precisión y robustez de los modelos predictivos desarrollados e incluso la mejora del software que permite realizar las proyecciones de demanda. Además, se recomienda realizar análisis más detallados sobre el impacto económico y estratégico de la implementación de la propuesta de solución en Dell Technologies, así como explorar posibles expansiones o adaptaciones de esta en otros mercados o sectores. También se recomienda evaluar la usabilidad de los reportes en Power BI, permitiendo que los reportes generados sean de utilidad y con facilidad de uso para aquellos que lo utilizarán.

En resumen, este trabajo representa un paso significativo en la dirección de mejorar la capacidad de Dell Technologies para anticipar y gestionar la demanda de sus productos en el mercado latinoamericano. A través de la aplicación de técnicas avanzadas de análisis de datos y modelado predictivo, se ha proporcionado a la empresa una herramienta poderosa para la toma de decisiones estratégicas, contribuyendo así a su competitividad y éxito en el mercado global.

REFERENCIAS BIBLIOGRÁFICAS

- [Chapman00] Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide.
- [Chaudhuri97] Chaudhuri, S. y Dayal, U. (1997). An overview of data warehousing and OLAP technology. SIGMOD Rec., 26, 65-74.
- [Forero21] Forero, D. A. y Sánchez J. A. (2021). Introducción a la inteligencia de negocios basada en la metodología Kimball. Tecnol.Investig. Academia TIA, 9, 5-17.
- [Gómez13] Gómez, A. (2013). Inteligencia de negocios, una ventaja competitiva para las organizaciones. Revista CIENCIA Y TECNOLOGÍA, 8(22), 85-96. Recuperado a partir de <https://revistas.unitru.edu.pe/index.php/PGM/article/view/193>
- [Inmon92] Inmon, W.H. (1992). Building the Data Warehouse. Nueva York, Estados Unidos: John Wiley & Sons
- [Lohr14] Lohr, S. (2014). *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*. The New York Times. Recuperado a partir de <https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>
- [Malinowski08] Malinowski, E. y Zimányi, E. (2008). Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications. Colorado, Estados Unidos:Springer.
- [Mendoza22] Mendoza, A. (2022). Metodologías de Data Warehouse. Gravatar. Recuperado de <https://gravitar.biz/Datawarehouse/metodologias-data-warehouse/>
- [Verne22] Verne Academy. (2022). ¿Qué es Data Warehouse y Data Marts? Diferencias y diseño del modelo. Verne academy. Recuperado a partir de <https://verneacademy.com/blog/articulos-data/que-es-data-warehouse-data-marts-diferencias/>
- [Vyas17] Vyas, S. y Vaishnav, P. (2017). A comparative study of various ETL process and their testing techniques in data warehouse. Journal of Statistics and Management Systems, 20, 753 - 763.