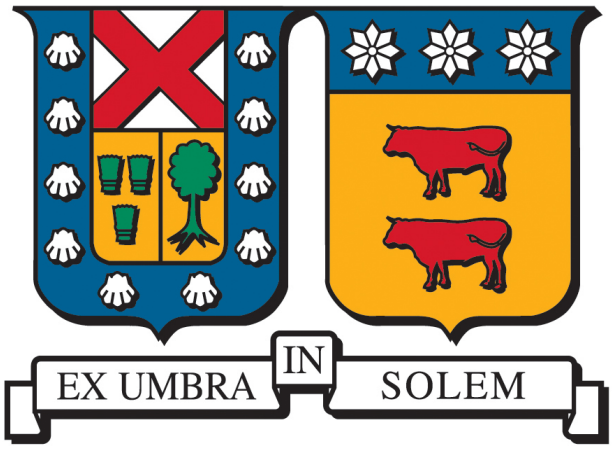
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



"ANÁLISIS DE DATOS, MODELOS PARA LA PREDICCIÓN Y SELECCIÓN DE ATRIBUTOS, APLICADOS A LOS DATOS DE E-RESTÓ"

NICOLÁS MATÍAS BRAVO TORO

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: **RICARDO ÑANCULEF**

PROFESOR CORREFERENTE: **JOSÉ LUIS MARTÍ**

NOVIEMBRE - 2016

Agradecimientos

Primero, agradecer a mi familia que me apoyó en este proceso académico, siempre presente y constante.

A la Fundación Belén Educa, ya que gracias a ellos pude costear mis estudios.

A mis compañeros y amigos, que de una u otra forma también se preocupaban por mí y me ayudaron en las etapas difíciles de este proceso.

A mi jefe de carrera, que confió en mí la oportunidad para continuar con mis estudios.

A mi profesor guía que me guió durante esta última etapa.

Y por último, a todos aquellos que por alguna razón llegaron a leer esto, y gastaron tiempo de su vida en saber en que trabajé durante varios meses de mi vida.

Para todos ustedes, gracias totales.

A mi persona, para demostrarme que puedo cumplir con las metas que me propongo.

Resumen

El propósito de este trabajo es hacer un análisis de los datos de **e-restó** mediante técnicas de inteligencia de negocios y máquinas de aprendizaje, lo que resultó en una serie de KPI (para ser presentados en un *dashboard* de reportes) y un modelo para la predicción de ventas de los negocios (usando el modelo *arima* para series de tiempo).

Palabras Clave: BI, KPI, arima, series de tiempo, Azure ML.

Abstract

The purpose of this work is to do an analysis of e-restó's data using business intelligence and machine learning techniques, wich result in a serie of KPI (to be presented in a report dashboard) and a model for business sales forecasting (using *arima* model for time series).

Keywords: BI, KPI, arima, time series, Azure ML.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Introducción

En el trabajo de Richard Millar Davens (1865), "Cyclopaedia of Commercial and Business Anecdotes [1]" está el primer uso conocido para el término **"Inteligencia de Negocios"**. Él lo usa para describir la forma en que un banquero, Sir Henry Furnese, triunfó: él tenía una comprensión de los problemas políticos, inestabilidades, y el mercado frente a sus competidores. [2]

Es este éxito lo que origina la idea de aplicar técnicas de BI[1] y algoritmos de máquinas de aprendizaje ("machine learning") a los datos de **e-restó**, con el fin de obtener información relevante y así poder asistir en la toma de decisiones a los administradores de estos negocios (restaurantes, bares y cafés).

Se comienza el presente documento presentando la definición del problema junto con los objetivos fijados para el trabajo de memoria, lo que consiste principalmente en determinar las variables de importancia para el análisis de BI y predicciones de ventas de los negocios.

Luego se expone el marco teórico usado para el desarrollo, abarcando los temas de estadística descriptiva, visualización de datos, inteligencia de negocios, series de tiempo y un extracto de la ISO/IEC 25010 para la calidad del software.

Para la presentación de las variables relevantes de un negocio se usaron KPIs, los que fueron obtenidos a través del análisis de los datos y constantes reuniones con algunos de los administradores de locales. Estos KPIs luego serán presentados en un *dashboard* de BI dentro de la misma aplicación e-restó, para así agregar valor a la aplicación.

Para generar las predicciones de ventas se utilizó la plataforma **Microsoft Azure Machine Learning**. Usando la función de series de tiempo *auto.arima* (del módulo de

[1]Se define inteligencia de negocios (BI por sus siglas en inglés "Business Intelligence") como un sistema que combina: recopilación de datos, almacenamiento de datos y gestión del conocimiento; con análisis para evaluar la información competitiva y corporativa para ser presentada a los que planean y toman las decisiones, con el objetivo de mejorar el tiempo y calidad de la entrada en el proceso de toma de decisiones. [3]

R) se obtuvieron resultados con errores bastante bajos para cuatro de los cinco negocios utilizados para las pruebas. El modelo generado será expuesto como un *web service* para ser consumido por la aplicación e-restó.

Se finaliza el desarrollo con comentarios sobre la experiencia usando MS Azure ML, siguiendo como guía la definición de calidad de la ISO/IEC 25010.

A través del desarrollo del trabajo se encontraron también oportunidades para mejorar el entendimiento que tiene e-restó (como proveedores del servicio) de sus clientes, y el uso que estos le dan a la plataforma. Esto a través de un análisis de *missing data* y anomalías en los datos.

Finalmente se presentan las conclusiones del trabajo presentando nuevamente los objetivos y evidencia del cumplimiento estos y razones por las cuales no se pudieron cumplir algunos de los objetivos secundarios.

1. Definición del problema

E-restó[2] es un software online para la gestión de restaurantes, bares y cafés, el cual permite: gestionar el consumo de las mesas, gestionar reservas, generar indicadores de ventas, administrar una base de datos de productos, precios y stock; y control de gastos.

Junto con una trayectoria de más de cinco años y dos mil clientes, se ha generado una gran cantidad de datos proveniente de los negocios que usan el software, por lo que se presentó la posibilidad de usar estos datos históricos para generar conocimiento útil para estos negocios, y así apoyar la continua mejora de éstos.

1.1. Objetivos generales

La idea que motiva este trabajo es aplicar técnicas de aprendizaje automático a los datos de e-restó haciendo uso de la plataforma de Microsoft Azure Machine Learning, lo que dio origen a los siguientes objetivos generales:

- Proporcionar a e-restó una herramienta de BI que ayude a sus clientes[3] a entender mejor su negocio.
- Mejorar el entendimiento que e-restó tiene de sus clientes.

1.2. Objetivos específicos

Después de reuniones con e-restó, se definieron los siguientes objetivos específicos para el presente trabajo:

1.2.1. Objetivo Principal

- Determinar qué variables determinan el éxito de un negocio.

[2] www.e-resto.com

[3] Los clientes son las personas que hacen uso de la aplicación e-restó. Los comensales son las personas que van a los negocios.

- Predecir las ventas de un negocio.

1.2.2. Objetivos Secundarios

- Hacer que el sistema se actualice en tiempo real.
- Generar observaciones en cuanto a: usabilidad, eficiencia y escalabilidad, de la plataforma MS Azure ML.
- Generar un ranking de atributos, por orden de impacto en el negocio.

2. Marco Teórico

En este capítulo se presenta el marco teórico usado en este trabajo haciendo una breve descripción sobre los temas abordados para contextualizar al lector.

Los temas abordados fueron: estadística descriptiva, visualización de datos, inteligencia de negocios, series de tiempo, predicción usando series de tiempo y un extracto de la ISO/IEC 25010.

2.1. Estadística descriptiva

La estadística proporciona métodos para organizar y resumir datos, y de sacar conclusiones basadas en la información contenida en los datos.

Una colección bien definida de objetos constituye a una **población** de interés, y un subconjunto de esta es llamada **una muestra**[4]. Por ejemplo, en una encuesta de aprobación de un alcalde, la población de interés son todos los residentes en la comuna del alcalde, y los residentes encuestados constituyen la muestra.

Las ventajas de elegir una muestra de la población son:

Reducción de costos: Se consumen menos recursos (tiempo, dinero y personas) al recabar información sobre una parte de la población, en vez de recabar la información de todos los individuos.

Viabilidad: En algunos casos es imposible obtener la información de toda la población, por lo que obtener la información de una muestra permite realizar el estudio.

En general existe interés en ciertas características de los objetos en una población: Por ejemplo, la edad de una persona, género, cantidad de ingresos, años que lleva viviendo en la comuna y nota con la cual califica al alcalde. Una característica puede ser categórica como el género, o numérica como la edad.

Se define como **variable** una característica cuyo valor puede cambiar de un objeto a otro en la población. En general las variables se representan con letras minúsculas del alfabeto.

Por ejemplo, para las características anteriores:

- e = edad de la persona
- g = género
- s = ingresos de la persona, en pesos
- a = años que lleva viviendo en la comuna
- x = nota con la cual califica al alcaide

Cuando se tiene solo una característica del objeto, se llaman datos **univariado**, en cambio cuando se tienen dos o más datos sobre el mismo objeto, se llaman datos **multivariado**. [4]

Es posible que un investigador que ha recopilado datos desee resumir y describir características importantes de los mismos. Esto implica utilizar métodos de **estadística descriptiva**. Algunos de ellos son de naturaleza gráfica: histogramas, diagramas de cajas y gráficas de puntos. Otros métodos descriptivos implican el cálculo de medidas numéricas: media, moda, mediana, desviación estándar y coeficientes de correlación. [4]

Los métodos de naturaleza gráfica serán presentados en la parte de visualización de datos 2.2.

La **media muestral** $\overline{x}$ de las observaciones $x_1, x_2, \ldots, x_n$ está dada por:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

La **mediana muestral** $\widetilde{x}$ se obtiene ordenando primero las n observaciones de la más pequeña a la más grande (incluyendo los valores repetidos de modo que cada observación muestral aparezca en la lista ordenada). Entonces:

$$\widetilde{x} = \begin{cases} \text{número en la posición } \left(\frac{n+1}{2}\right) & \text{si } n \text{ es impar} \\ \overline{x} \text{ de los números en las posiciones } \left(\frac{n}{2}\right) \text{ y } \left(\frac{n}{2}-1\right) & \text{si } n \text{ es par} \end{cases} \tag{2}$$

La **moda muestral** es el dato con mayor frecuencia dentro de la muestra, la frecuencia es la cantidad de veces que está repetido un dato en la muestra. [4]

La mediana divide el conjunto de datos en dos partes iguales. Para obtener medidas de ubicación más finas, se podrían dividir los datos en más de dos partes. Tentativamente, los **cuartiles** dividen el conjunto de datos en cuatro partes iguales y las observaciones arriba del tercer cuartil constituyen el cuarto superior del conjunto de datos, el segundo cuartil idéntico a la mediana y el primer cuartil separa el cuarto inferior de los tres cuartos superiores. [4]

Las medidas principales de variabilidad implican las **desviaciones de la media**, $x_1 - \overline{x}, x_2 - \overline{x}, ..., x_n - \overline{x}$. Es decir, las desviaciones de la media se obtiene restando $\overline{x}$ de cada una de las n observaciones muestrales. Una desviación será positiva si la observación es más grande que la media y negativa si la observación es más pequeña que la media. Para evitar que la suma de las diferencias se neutralicen entre ellas (suma de valores positivos y negativos), se consideran las desviaciones al cuadrado $(x_1 - \overline{x})^2, (x_2 - \overline{x})^2, ..., (x_n - \overline{x})^2$. En vez de utilizar la desviación al cuadrado promedio $\sum(x_i - \overline{x})^2/n$, por varias razones se divide la suma de desviaciones al cuadrado entre $n - 1$ en lugar de entre n. [4]

La **varianza muestral**, denotada por s^2 está dada por

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{S_{xx}}{n-1} \tag{3}$$

La **desviación estándar muestral**, denotada por s, es la raíz cuadrada (positiva) de la varianza

$$s = \sqrt{s^2} \tag{4}$$

2.2. Visualización de datos

Las presentaciones con gráficos son a menudo formas muy efectivas de comunicación. Pero también pueden ser formas poco efectivas de comunicación, si se realizan de la forma incorrecta.

Los gráficos proveen un excelente acercamiento para explorar datos y son esenciales para presentar resultados.

Existen principalmente dos tipos de gráficos, los **gráficos para presentación** y los **gráficos para exploración**, sus diferencias radican tanto en forma como en uso.

Los gráficos de presentación son generalmente estáticos, y un sólo gráfico es usado para resumir la información presentada. Estas presentaciones debiesen ser en alta calidad e incluir descripciones detalladas y una explicación completa de las variables mostradas y la forma del gráfico. Los gráficos de presentación deben ser como los teoremas matemáticos; estos no dan pistas de cómo se alcanzó el resultado, pero deben ofrecer una justificación convincente para su conclusión.

Los gráficos para exploración, por otro lado, son usados para buscar resultados. Estos deben ser rápidos e informativos en vez de lentos y precisos. No son diseñados para presentaciones, por lo que no es necesario que contengan etiquetas y título. Un gráfico de presentación debe ser diseñado para que lo vean muchos lectores, en cambio muchos gráficos de exploración son diseñados para apoyar la investigación de un analista.

2.2.1. Literatura

Varios autores han escrito libros sobre como dibujar buenos gráficos de estadísticas, el más conocido es el trabajo de **Edward Tufte**. Sus libros[4] incluyen excelentes ejemplos (y algunos terribles) y describen importantes principios de cómo dibujar buenos gráficos. Tufte critica la decoración inadecuada y la tergiversación de los datos, pero su consejo se limita a representar correctamente los datos. Los libros de **Cleveland** son igualmente valiosos. Y así debe ser. Los estadísticos deben concentrarse en obtener la información estadística básica correcta, y los diseñadores pueden ser consultados para producir una versión final pulida [5].

2.2.2. Elecciones de diseño científico en Visualización de Datos

Trazar una sola variable debiese ser simple. El tipo de variable influenciará el tipo de gráfico elegido. Por ejemplo, histogramas y gráficos de caja son adecuados para variables numéricas continuas, mientras que los gráficos de barras y circulares son apropiados para variables categóricas. En ambos casos es posible elegir otro tipo de gráfico. La transformación y agregación de los datos dependen de la distribución de los datos y el objetivo del gráfico.

Esto es diferente en el caso de gráficos multivariados, donde representar la unión de dos distribuciones categóricas no es tan simple. La decisión principal al momento de hacer un gráfico multivariado es la forma de exhibición, además también es importante la elección de las variables y su orden. Por ejemplo, en un gráfico de dispersión usualmente la variable dependiente se fija en el eje vertical.

Existen gráficos de barra, gráficos de sectores circulares, histogramas, gráficos de puntos, gráficos de caja, gráficos de dispersión, histogramas circulares (*roseplots*), entre otros. Su elección depende del tipo de variable a ser graficada.

Escala

[4]Por ejemplo, The Visual Display of Quantitative Information.

Definir la escala para el eje de una variable categórica es cuestión de elegir el orden de la información. Esto quizás dependa de lo que represente cada variable o su tamaño relativo. En el caso de las variables continuas hay que elegir los criterios de evaluación y las marcas de graduación.

Existe la tentación de elegir la escala desde el valor mínimo al máximo de los datos, pero esto significa que quizás algunos puntos queden en los ejes. A menos que los límites sean definidos por la naturaleza de los datos (Por ejemplo, notas de certámenes: 0 a 100), es una buena práctica extender la escala más allá de los límites de las observaciones.

En la figura 2.1 muestra histogramas para el grosor de las marcas de datos de Hidalgo [6]. Con la configuración por defecto se muestra una distribución sesgada con una posible posición de moda secundaria alrededor de 0,10. En el segundo se especifican criterios de redondeo de las variables y el ancho del intervalo, mostrando aún más evidencia de la posible moda secundaria. El tercero esta hecho de tal forma que cada valor es representado sin redondeo (la precisión de los datos originales es milimétrica). Esto sugiere que la primera moda está compuesta por dos grupos y quizás existen más modas a la derecha. Lo que los datos representan y la forma en la que estos fueron tomados deben ser tomados en cuenta al momento de elegir la escala.

Una cosa es determinar qué escala usar, pero otra es marcar y etiquetar los ejes. Demasiadas etiquetas en los ejes dan una impresión desordenada; pocas etiquetas pueden dificultar la evaluación de los valores y diferencias. (Notar que el objetivo de los gráficos no es proveer un valor exacto, las tablas son mejores para eso.) Marcas entre las etiquetas pocas veces ofrecen un uso práctico.

Clasificar y ordenar

El efecto de una figura puede ser influenciado por diversos factores. Por ejemplo, las variables categóricas pueden ser ordenadas de distinta forma, lo que se traduce en un efecto importante en el gráfico. Un orden alfabético puede parecer apropiado, pero puede ser que otra agrupación (países por continente) sea más relevante. Las categorías

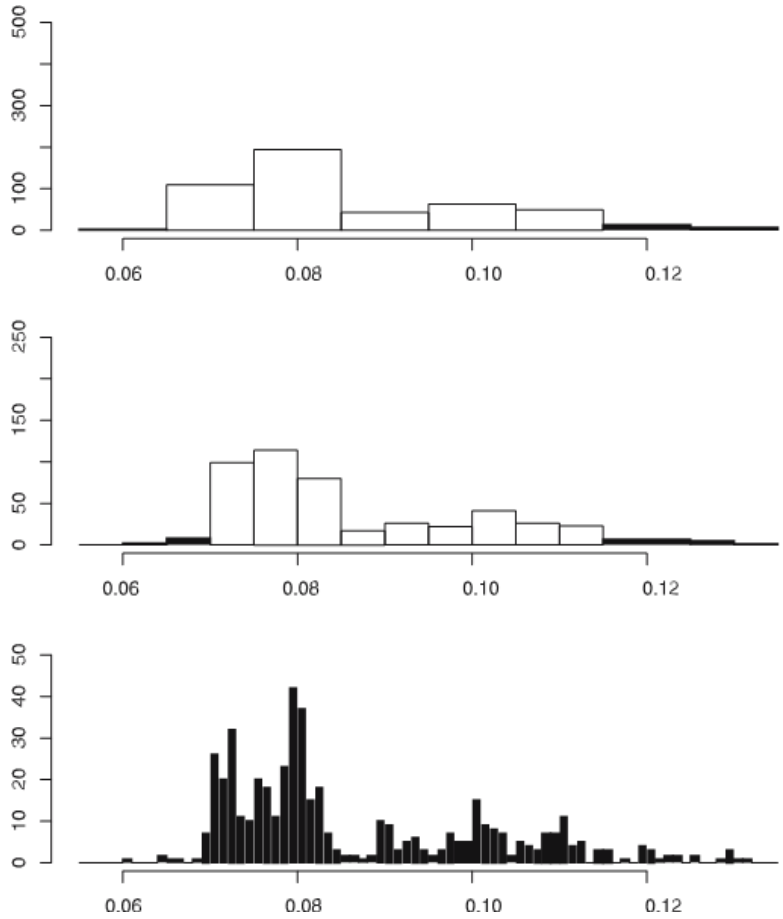


Figura 2.1: Tres histogramas para el grosor de las marcas de datos de Hidalgo, todos con el mismo punto de anclaje pero con diferente ancho de intervalo. La escala horizontal esta alineada y el área total de cada figura es la misma (notar la diferencia de escala en la frecuencia). Fuente: Izenman & Sommer (1988)[6].

pueden ser también ordenadas por el tamaño de una segunda variable. En ambos casos es necesario tener presente lo que se desea diferenciar entre las categorías. [5]

Título, leyendas y comentarios

Idealmente, el título debería explicar completamente el gráfico, incluyendo la fuente de los datos. Las leyendas describen que colores o símbolos representan cada grupo de datos, en este caso se recomienda que la información esté descrita directamente en el gráfico y no separado de éste para que los ojos no estén saltando entre el gráfico y la leyenda.

Los comentarios son usados para destacar una característica particular del gráfico, por ejemplo en eventos en series de tiempo o para atraer la atención a un punto en particular en un gráfico de dispersión. Los gráficos que requieren demasiados comentarios pueden que estén intentando presentar demasiada información en un solo gráfico.

La idea principal de los gráficos es presentar la información de forma concisa y directa. [5]

Tamaño, marco y relación de aspecto

Los gráficos deben ser lo suficientemente grandes como para presentar la información de forma clara, y no más grande que eso.

Se puede agregar un marco al gráfico, pero como estos agregan tamaño al gráfico y un poco de confusión, solo deberían ser usados para demarcar la separación. Por ejemplo, separar un gráfico del texto.

La relación de aspecto tiene un gran efecto en la percepción del gráfico. Esto se hace más evidente en las series de tiempo. Si se quiere mostrar un cambio gradual, aumente la proporción del eje horizontal respecto al eje vertical. La acción contraria mostrará un cambio más rápido. En la figura 2.2 se muestran dos gráficos de dispersión con los mismos datos, en el primero se aprecia un cambio más rápido, en cambio en el segundo parece ser más gradual.

En un *paper* de 1988 (*The Shape Parameter of a Two-Variable Graph*) [7], Cleveland et al. proponen la idea de que el ángulo promedio en un gráfico de líneas debe ser 45°. Esto se ha llamado *banking to* 45° y ha resultado ser un estándar para determinar la mejor relación de aspecto en un gráfico de líneas.

Color

Un buen uso de los colores puede mejorar y clarificar la presentación de un gráfico. Un uso pobre de estos puede opacar, ensuciar y confundir. Si bien hay una componente de estética fuerte en el uso de colores, usarlos de buena forma para presentar los datos depende esencialmente sobre la función: qué información se intenta transmitir, y cómo los colores mejoran eso.

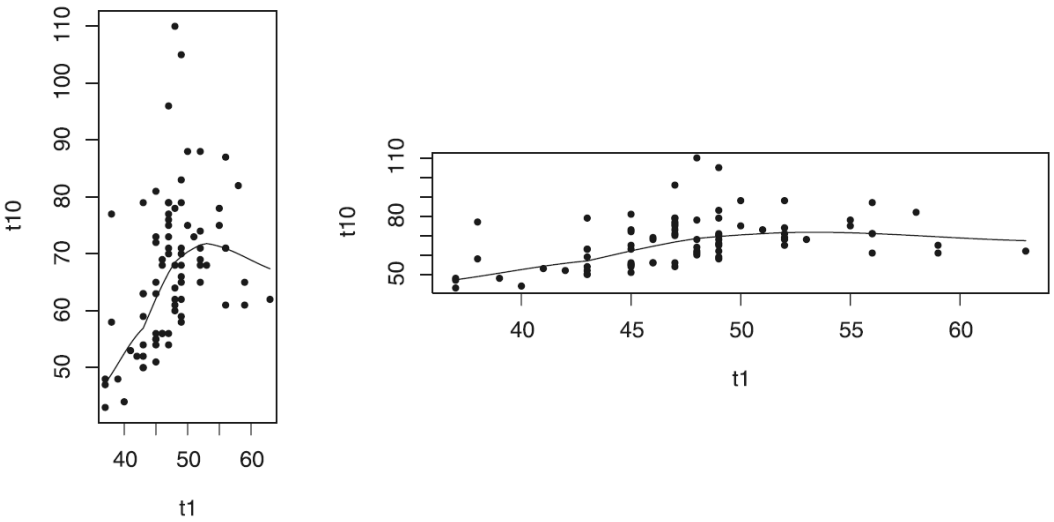


Figura 2.2: Los tiempos de los 80 corredores para la última etapa de una carrera frente a sus tiempos de la primera etapa, con una regresión local (lowess smoother). Fuente: Everitt (1993)[8].

La función más importante de los colores en la presentación de la información es para distinguir los elementos. En la figura 2.3 se usan diferentes colores para distinguir los productos en el gráfico de dispersión, pero también hay que considerar que los colores también incluyen el negro, blanco y tonos de grises. En este caso se usa negro para la leyenda, gris para identificar el área fuera del gráfico y las guías dentro del gráfico, y el blanco para mostrar el área del gráfico.

Un uso útil de los colores es para agrupar los elementos relacionados y dirigir la atención en proporción a la importancia de cada elemento. Por ejemplo, toda la información contextual (guías, ejes, leyendas y bordes) de la figura 2.3 está en tonos de grises, blanco y negro, mientras que los datos están claramente coloreados, lo que atrae la atención a los datos. Los colores de los datos fueron elegidos de tal forma que todos parecen tener igual importancia, y todos son claramente visibles en un fondo blanco. Y las guías gris claro son legibles, pero lo suficientemente similar al fondo como para no interferir con los datos.

Los principios que definen el diseño del color son el contraste y analogía. Los colores en contraste son diferentes, colores análogos son similares. El contraste atrae la

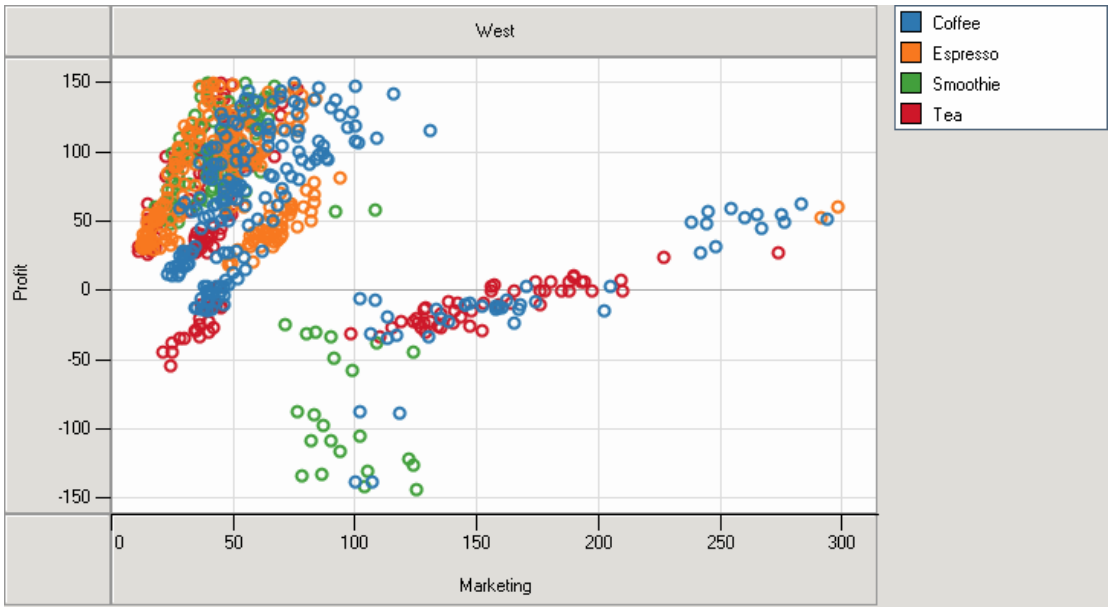


Figura 2.3: Marketing frente a la ganancia de 4 productos distintos. Fuente: Choosing Colors for Data Visualization (2006)[9].

atención, la analogía agrupa.

En el diseño del color se especifican 3 dimensiones: matiz, valor y croma (coloreado). Matiz es el nombre del color, como el rojo, verde o naranjo. Valor es la percepción de luminosidad u oscuridad del color. El coloreado es la intensidad del color. Colores con coloreado alto son vividos o saturados, colores con coloreado bajo son parecidos al gris.

La dimensión de la matiz es circular, usualmente presentado como un círculo de matices (figura 2.4). Existen diferentes círculos de matices, pero todos presentan los colores en el mismo orden. En cualquier círculo de matices, los colores análogos están cercas uno de otro. Matices de contraste se encuentran en el lado opuesto del círculo, aunque una separación pequeña puede resultar suficiente.

La dimensión del valor es visualmente la más importante, el contraste de los valores define tanto la claridad como el poder de atraer la atención. Es fácil ver las variaciones en tonos de grises o de un solo color. Es más difícil comparar el valor de dos colores diferentes. La escala del valor va desde el negro (0) hasta el blanco (100). La figura 2.5

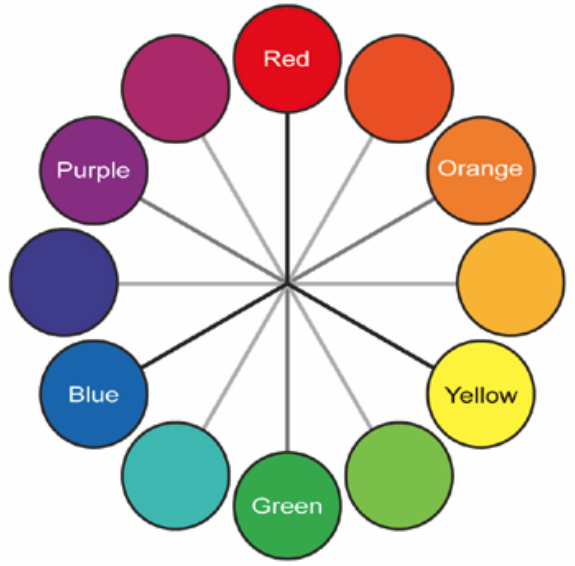


Figura 2.4: Ejemplo de un círculo de matices. Fuente: Choosing Colors for Data Visualization (2006)[9].

muestra las diferencias de graduación de color para demostrar (aproximadamente) el mismo valor de la escala.

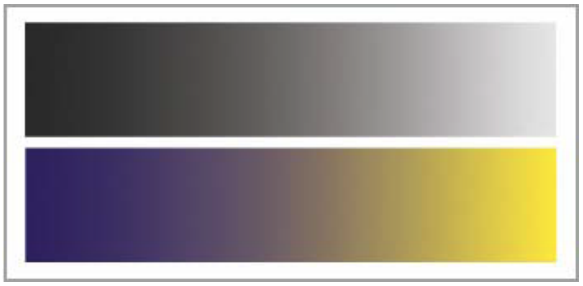


Figura 2.5: Ilustración de graduación de dos colores aproximadamente usando la misma escala. Fuente: Choosing Colors for Data Visualization (2006)[9].

El coloreado indica que tan brillante, saturado, vivido o atractivo es el color. Formalmente, para cualquier color, reducir su coloreado a cero produce un gris con el mismo valor. La figura 2.6 muestra un ejemplo organizado por matiz, tinte, tono y sombreado, para cinco colores distintos.

Para la selección de los colores existen 3 tipos de paletas: divergentes, secuenciales y cualitativas. En la figura 2.7 se muestra un ejemplo para los tipos de paletas.

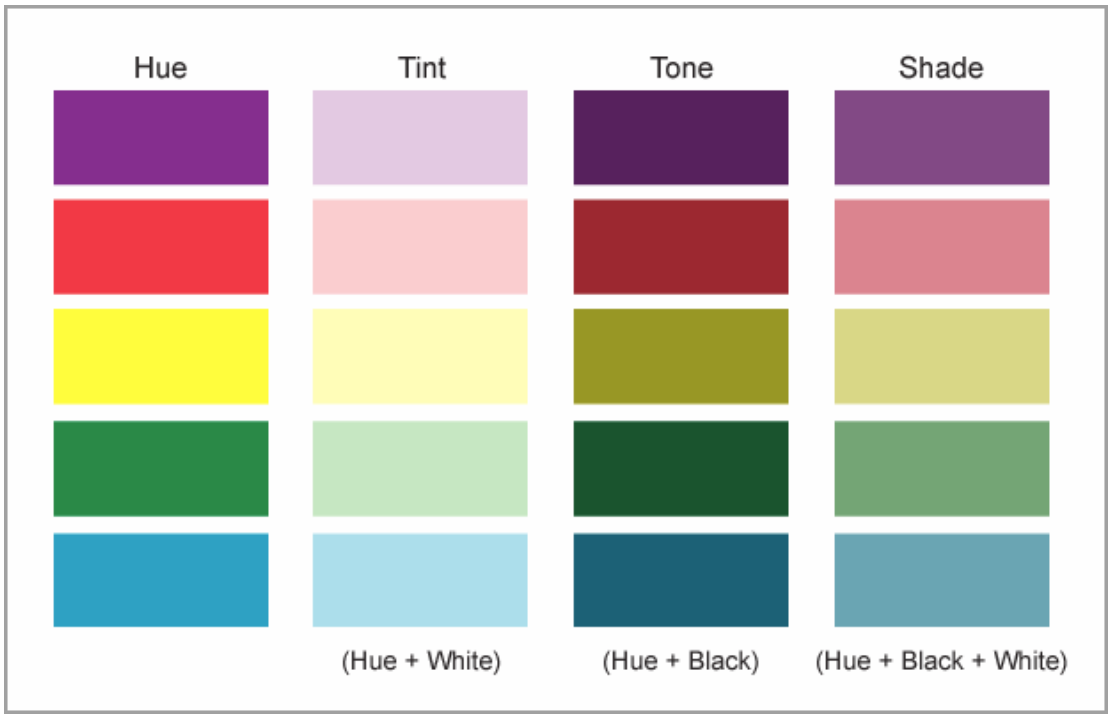


Figura 2.6: Tintes, sombras y tonos de cinco colores distintos. Fuente: Choosing Colors for Data Visualization (2006)[9].

Una paleta divergente es usada para presentar datos que cuente con extremos y un centro neutro. Por ejemplo, la altura geográfica, donde el color central representa el nivel del mar.

Una paleta secuencial son para datos ordenados de un punto a otro.

Y una paleta cualitativa son para datos categóricos o agrupados, sin un orden explícito.

Una buena fuente para paletas de colores es el sitio: www.colorbrewer2.org

2.2.3. Tipos de gráficos

A continuación se presenta una descripción de los los tipos de gráfico más comunes.

Gráfico de barras Un gráfico de barras es una representación gráfica en un eje cartesiano de las frecuencias de una variable cualitativa o discreta.

En uno de los ejes se posicionan las distintas categorías o modalidades de la va-

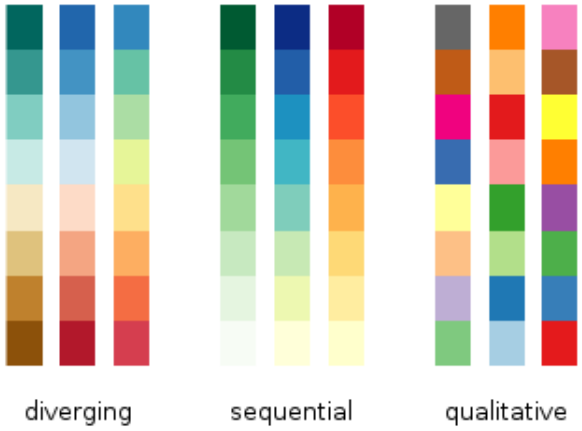


Figura 2.7: Ejemplo de paletas de colores obtenidas de www.colorbrewer2.org.

riable cualitativa o discreta y en el otro el valor o frecuencia de cada categoría en una determinada escala. Ejemplo figura 2.8.

Estos gráficos suelen ser usados para: comparar magnitudes de varias categorías o ver la evolución en el tiempo de una magnitud concreta.

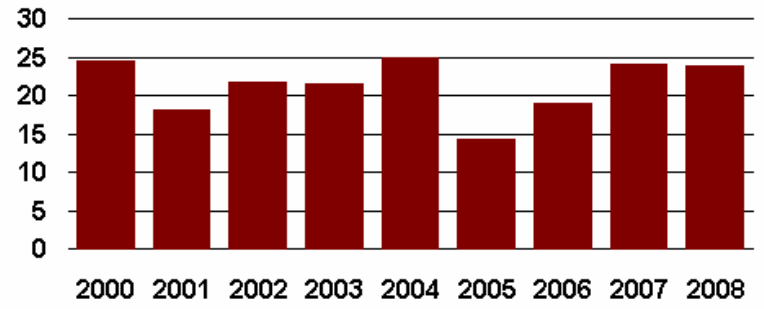


Figura 2.8: Producción de cereales en España, en millones de toneladas [10].

Histograma Un histograma es un tipo especial de gráfico de barras. Éste, a diferencia del gráfico de barras, solo puede ser usado con variables continuas, ya que es usado para representar una distribución de frecuencias en variables continuas. Lo importante es el área de la barra y no solo a su altura, ya que es proporcional al ancho del intervalo y a la frecuencia de este.

En la figura 2.9, se puede ver que las columnas están juntas, y el punto medio es el que da nombre al intervalo.

Estos gráficos suelen ser usados para representar distribuciones de probabilidad, además pueden representar la probabilidad de valores que no fueron medidos, pero que sí se encuentran contenidos en alguno de los intervalos.

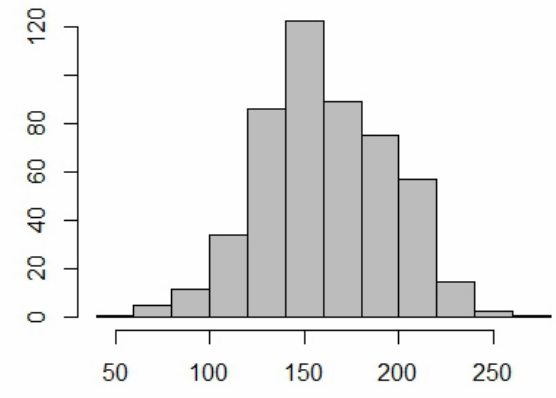


Figura 2.9: Histograma.

Gráfico de líneas Un gráfico de líneas es una representación gráfica en un eje cartesiano de la relación que existe entre dos variables reflejando con claridad los cambios producidos.

En cada eje se representa cada una de las variables cuya relación se quiere observar. Se presenta un ejemplo en la figura 2.10.

Estos gráficos suelen ser usados para presentar tendencias temporales, en el eje horizontal se posiciona la variable temporal y en el eje vertical se introduce la escala de la variable cuya variación en el tiempo se desea ver. Permite la comparación de varias variables.

Gráfico de sectores Un gráfico de sectores es una representación circular de las frecuencias relativas de una variable cualitativa o discreta que permite, de una manera sencilla y rápida, su comparación.

El círculo representa la totalidad que se desea observar y cada porción, llamadas

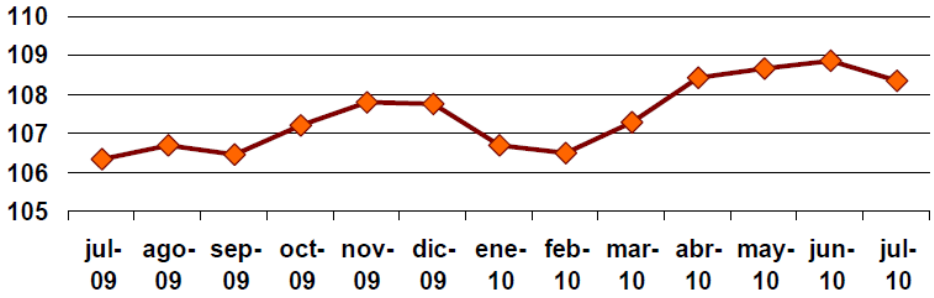


Figura 2.10: IPC 2006 [10].

sectores, representa la proporción de cada categoría de la variable respecto el total. Suele expresarse en porcentajes. Ejemplo figura 2.11.

Para obtener el ángulo de cada sector, primero se calcula la frecuencia relativa de cada categoría, y luego esta se multiplica por 360°. Para obtener el porcentaje se multiplica la frecuencia relativa por 100 %.

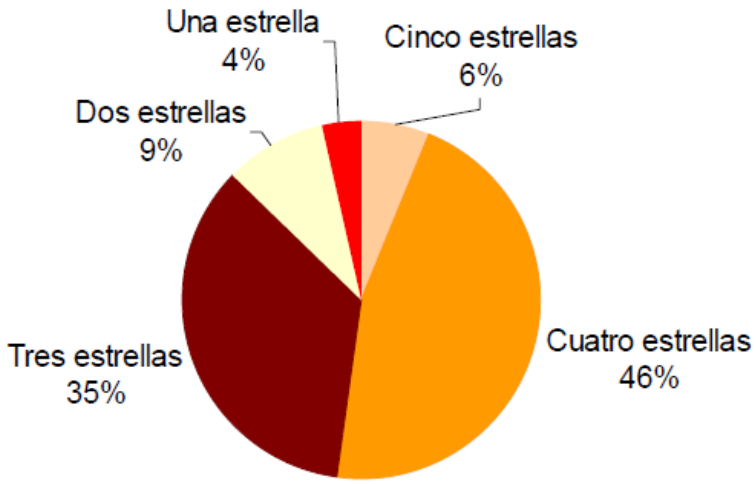


Figura 2.11: Viajeros hospedados en hoteles españoles por categoría del establecimiento (2009) [10].

Gráfico de puntos Un gráfico de puntos es utilizado para ilustrar un **número reducido de datos**, mostrando cada uno de los elementos de un conjunto de datos numéricos por encima de una recta numérica (horizontal).

Este tipo de gráfico es usado principalmente para: localizar los datos y ver la dispersión o variabilidad de los datos.

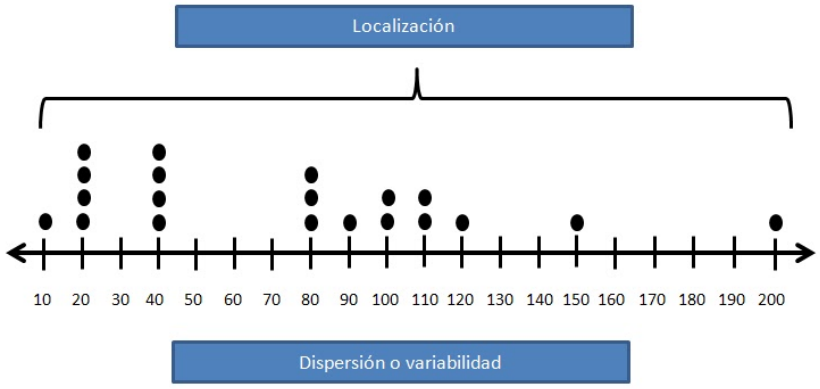


Figura 2.12: Gráfico de puntos.

Gráfico de caja (*Boxplot*) Un gráfico de caja, como su nombre lo indica, es una representación de los datos en una caja, el que es usado para representar varias características prominentes de un conjunto de datos: el centro, la dispersión, el grado y naturaleza de cualquier alejamiento de la simetría y la identificación de las observaciones "extremas o apartadas" inusualmente alejadas del cuerpo principal de los datos. [4]

La caja, o cuerpo principal, contiene el 50 % de los datos, la que se encuentra delimitada por los cuartiles 1 ($Q1$) y 3 ($Q3$), y dentro de esta se identifica la mediana, correspondiente al cuartil 2 ($Q2$). En los extremos de la caja se trazan los "bigotes" que van desde la caja a los datos más pequeños y más grandes. El largo de cada bigote va desde la caja hasta el dato más grande (por un lado, y pequeño por el otro) contenido en $1,5$ veces el rango inter cuartil (IQR), siendo este rango la diferencia entre $Q3 - Q1$. Cualquier valor fuera de estos bigotes se marca como dato "outlier". Ejemplo figura 2.13

Gráfico de dispersión Un gráfico de dispersión representa la relación entre dos variables.

Se asigna cada variable a un eje, generalmente se asigna al eje horizontal la variable independiente y en el vertical la variable dependiente. Dentro del área del

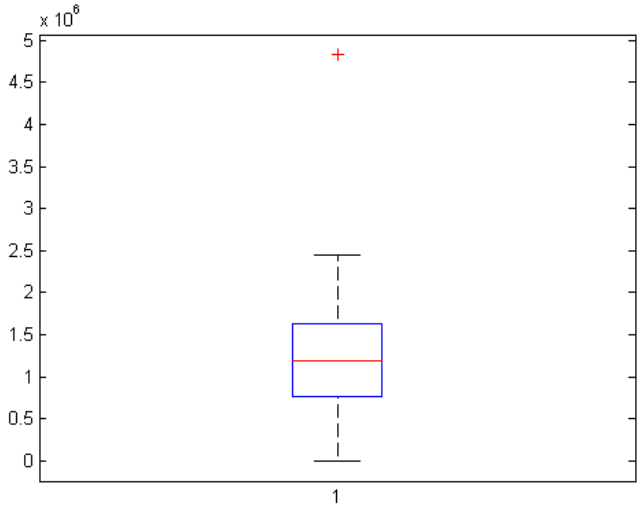


Figura 2.13: Ventas de un negocio con 1 outlier.

gráfico se representan las duplas. En la figura 2.14 se muestra la corriente como variable independiente y el voltaje como variable dependiente, se puede ver que la relación es lineal, siguiendo la Ley de Ohm: $V = IR$.

Estos gráficos son usados usualmente para representar la relación entre dos variables (hasta 3 si se genera un cubo en tres dimensiones), también se aplican métodos de interpolación para predecir un valor dentro del intervalo de mediciones, y extrapolación para predecir un valor fuera del intervalo de mediciones.

Histograma circular (*roseplot*) Al igual que un histograma normal, un histograma circular sirve para representar variables con un comportamiento "circular". i.e. Los grados de una brújula poseen un con comportamiento circular, entonces si se quiere representar de mejor forma la dirección del viento registrada esta debe ser en un círculo, y su velocidad promedio diaria sería el tamaño de la barra, como lo muestra la figura 2.15.

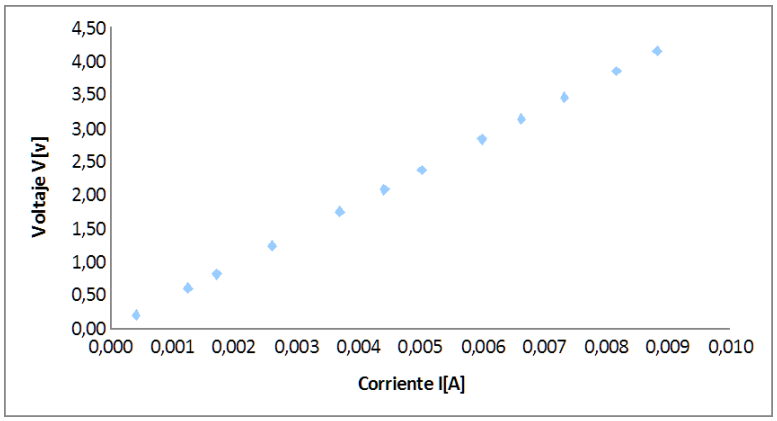


Figura 2.14: Diferencia de voltaje [V] medido en una resistencia al aplicar corriente en el circuito.

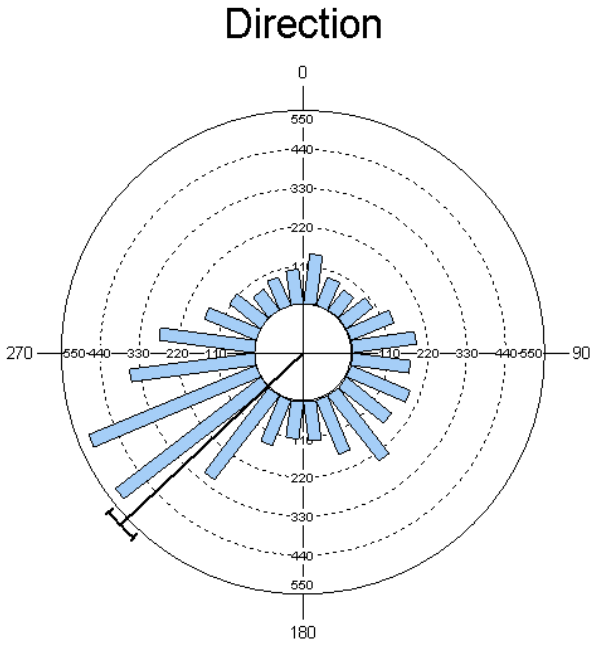


Figura 2.15: Dirección y velocidad del viento promedio diaria en un año.

2.3. Inteligencia de negocios

En el día de hoy y a lo largo de la historia es posible encontrar varias definiciones de Inteligencia de negocios (desde ahora en adelante BI, por sus siglas en inglés *Business Intelligence*). Todas estas definiciones apuntan a la misma intención: “BI es un conjunto de teorías, metodologías, procesos, arquitecturas, y tecnologías que transforman datos

brutos en información útil y significativa para los procesos de negocios" [11].

En el libro *Fundamentals of Business Intelligence* [11] se destacan las siguientes características y desafíos en BI:

Tareas: La tarea principal de BI es proveer apoyo a la toma de decisiones para metas específicas definidas en un contexto de las actividades del negocio en diferentes dominios, tomando en cuenta la estructura organizacional e institucional.

Actualmente se puede encontrar una comprensión bien estructurada de la lógica de negocio en casi todas las áreas. Esta nueva comprensión también ha conducido a una perspectiva de conceptos orientada a los procesos, las que consideran el flujo de trabajo y la minería de procesos en BI. Otro aspecto son las estructuras organizacionales como la organización descentralizada que quieren aplicar el apoyo de decisiones dentro de su entorno, y por lo tanto, las ideas de inteligencia colectiva y *crowdsourcing* son aplicadas en BI.

Base: El apoyo a la toma de decisiones se basa principalmente en información empírica basada en los datos. Además de los antecedentes empíricos, BI usa distintos tipos de teorías y conocimiento para la generación de la información.

Además de los *data warehouse* tradicionales, hay que tomar en cuenta los datos en la Web. Estos datos generalmente no son bien estructurados, sino que son semi-estructurados como el texto. La necesidad de integrar diferentes datos de forma útil y coherente para el apoyo de toma de decisiones ha generado modelos para vincular los datos en BI. En relación con estos nuevos datos, el enfoque de los métodos analíticos ha sido ampliado y han surgido nuevas herramientas, tales como la minería de imágenes, minería de texto, minería de opiniones, o análisis de las redes sociales.

Realización: El apoyo a la toma de decisiones tiene que ser construido en un sistema usando las capacidades actuales de las tecnologías de información y comunicación (TIC).

La arquitectura de software actual permite la construcción de sistemas de BI más interesantes. Desde la perspectiva del usuario, el software como servicio (*Sofware as a Service*, SaaS) constituye un desarrollo importante para estos sistemas. Desde un punto de vista computacional, se tiene que tratar con una gran cantidad de datos complejos. Es más, la computación en la nube y la distribuida son con conceptos que abren nuevas oportunidades para la aplicación de sistemas de BI (como la oportunidad aprovechada para el desarrollo de este trabajo).

Entrega: Un sistema de BI tiene que entregar la información en el momento correcto y a la gente correcta de una forma adecuada.

Los dispositivos móviles ofrecen una nueva dimensión para la entrega de la información en tiempo real a los usuarios. Además, hay que tomar en cuenta la calidad de la información entregada en tiempo real significa un desafío para el desarrollo de los sistemas de BI.

Para entender mejor la conección entre los modelos de negocios y BI desde un punto de vista administrativo, se definen 4 escenarios que conectan BI con el contexto del negocio, tomando en cuenta la siguiente definición de modelo de negocio: "*Un modelo de negocios refleja la estrategia de una empresa para generar valor*" [11].

BI separado de la gestión estratégica: En este caso el interés principal de BI es lograr objetivos al corto plazo en una sección de la organización. Típicamente los resultados de la aplicación de BI son reportes estandarizados para una parte específica del negocio.

BI apoyando el monitoreo de desempeño estratégico: Una aplicación de este tipo de BI es motivado por objetivos globales y formulado de acuerdo a estos objetivos. El monitoreo del desempeño es hecho mediante objetivos medibles. Un *data warehouse* permite una visión unificada del negocio, usualmente esto es un prerrequisito para la aplicación de este escenario.

BI en la retroalimentación sobre la formulación de estrategias: Esta aplicación va un paso más allá de la estrategia anterior, y apunta a la evaluación del desempeño usando métodos analíticos. En el mejor caso, esta aplicación puede ser usada para optimizar una estrategia. El resultado típico de este escenario puede ser un cuadro de mando integral (*balanced scorecard*).

BI como un recurso estratégico: Esta estrategia usa la información generada por BI no solo para la optimización, sino que también como una fuente esencial para la definición de la estrategia a un nivel administrativo. Por ejemplo marketing basado en los clientes o el desarrollo de un procedimiento estándar para el tratamiento de un paciente.

2.3.1. Objetivos de BI

Los primeros objetivos de aplicaciones de BI son de análisis. Estos objetivos van desde la adquisición de la información de algunos aspectos del proceso de negocios, a través de la mejora del rendimiento del proceso hasta el entender la participación del proceso para lograr los objetivos estratégicos.

Una forma de formular los objetivos está basada en los llamados "indicadores claves de rendimiento", KPI por sus siglas en inglés *key performance indicators*. Los KPI permiten medir el rendimiento en cualquier aspecto del negocio. Por ejemplo, en un contexto educacional, un KPI puede ser la taza de reprobación de los estudiantes. Existen KPI cuantitativos y no cuantitativos (estos últimos son más difíciles de medir).

Un KPI une las actividades del negocio a los objetivos (estratégicos) mediante la definición de un indicador medible. Estos pueden estar relacionados con algún proceso del negocio o al negocio completo. Es posible distinguir diferentes categorías de indicadores: cualitativos y cuantitativos; principales y retrasados; de entrada, de proceso y de salida; prácticos, direccionales y accionables; y financieros.

Cualitativos y cuantitativos: Es la clasificación más fácil de distinguir. Los indica-

dores cualitativos no pueden ser presentados con números a diferencia de los cuantitativos que sí pueden ser presentados con números.

Principales y retrasados (*Leading and lagging*): Estos indicadores se diferencian en que el indicador principal predice el resultado de un proceso, y el indicador retrasado presenta el éxito o falla *post hoc*.

De entrada, de proceso y de salida: Estos indicadores están relacionados con los recursos, eficiencia y resultados del proceso. Un indicador de entrada refleja los recursos a ser consumidos por un proceso, el indicador de proceso refleja la eficiencia de una tarea dentro del proceso, y el indicador de salida que representan los resultados del proceso.

Prácticos, direccionales y accionables: Estos indicadores son usados generalmente para confirmar algo que se cree o se sabe respecto a un proceso. Los indicadores prácticos son aquellos KPIs que revelan datos sobre un proceso existente. Por otra parte un indicador direccional especifica si el negocio está mejorando o empeorando. Y un indicador accionable es aquel que uno puede cambiar (Por ejemplo, aumentar la propaganda en las redes sociales para cambiar el indicador accionable de ventas mensuales).

Financieros: Son los más conocidos y está relacionado con el rendimiento del negocio desde un punto de vista financiero, por ejemplo, márgenes de ganancias.

Otra forma de formular los objetivos son los llamados "objetivos de análisis". Esta forma está basada en tres tipos de objetivos de análisis: descriptivo, predictivo y de entendimiento. El objetivo descriptivo generalmente se representa con reportes y es usado para lograr los otros dos objetivos, se relaciona generalmente con el denominado *unsupervised learning*[5]. El segundo objetivo es el análisis predictivo, que ya es poco más

[5]Método de aprendizaje automático donde se ajusta un modelo a observaciones con datos sin etiquetar.

ambicioso que el descriptivo, y se relaciona generalmente con el denominado *supervised learning*[6]. Y tercer objetivo es el entendimiento, los que ayudan a los participantes a entender su proceso de negocio.

El análisis descriptivo está relacionado con la estadística descriptiva, la que busca representar un resumen de las instancias dentro del proceso de negocio. Se pueden resumir tres objetivos: reportes con resumen de las instancias para tomar decisiones (por ejemplo, un *dashboard*), segmentar las instancias de acuerdo a similitud entre ellas creando grupos que las representen (por ejemplo, un algoritmo de *Clustering*), y detectar patrones de comportamiento en un determinado proceso que permita identificar aspectos relevantes relacionados a dicho proceso (por ejemplo, detectar que la cantidad de fallas en el uso de determinado vehículo se hacen más frecuente a medida que pasa el tiempo).

El análisis predictivo, como su nombre lo indica predice el comportamiento de un proceso en el futuro. Se identifican dos tipos de predicciones: la regresión que busca encontrar una función para predecir el resultado (generalmente algún KPI) a partir de un número de variables (factores que influyen en el resultado), y la clasificación que busca asignar una nueva instancia a algún grupo de instancias (instancias segmentadas en el análisis descriptivo).

En el objetivo de entendimiento, se busca que los participantes entiendan su proceso de negocio. Se identifican 2 objetivos: identificar las reglas que determinan la relación entre los eventos y los procesos, e investigación del desempeño de las instancias respecto a la conformidad definida en el proceso de negocios.

[6]Método de aprendizaje automático donde se ajusta un modelo a observaciones previamente etiquetadas.

2.4. Series de tiempo

Una serie de tiempo[7] es un conjunto de observaciones x_t, cada una de las cuales es registrada en un tiempo específico t [13]. Se dice que $t \in \tau$, donde el tiempo τ corresponde generalmente a un conjunto discreto y equidistante.

En el libro de Brockwell et al. [14] se define el modelo de series de tiempo como:

Definición 2.1. *Un modelo de series de tiempo para una secuencia de datos observados $\{x_t\}$ es una especificación de la distribución conjunta (o posiblemente las medias y covarianzas) de una secuencia de variables aleatorias $\{X_t\}$ de los cuales $\{x_t\}$ es postulado a ser una realización. [14]*

El principal objetivo de una X_t es su análisis para hacer pronóstico. Algunos ejemplos donde se puede utilizar series de tiempo:

- Proyecciones del empleo y desempleo.
- Beneficios netos mensuales de un ente bancaria.
- Índices de precio.
- Número de habitantes por año.
- Temperatura media mensual.

El análisis clásico de las series temporales se basa en la suposición de que los valores que toma la variable de observación es la consecuencia de tres componentes, cuya actuación conjunta da como resultado los valores medidos, estos componentes son [15]:

Componente tendencia: Se puede definir como el cambio a largo plazo que se produce en la relación al nivel medio, o el cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo.

[7] Existen libros completos dedicados al análisis de las series de tiempo y métodos de predicciones usando éstas. Para entender las series de tiempo en más detalle se recomienda leer el trabajo de **Box-Jenkins** *Time Series Analysis, Forecasting and Control* [12].

Componente estacional: Muchas series temporales presentan cierta periodicidad o dicho de otro modo, variación de cierto período (semestral, mensual, etc.). Por ejemplo, las "Ventas al Detalle en Puerto Rico" aumentan por los meses de noviembre y diciembre por las festividades navideñas. Estos efectos son fáciles de entender y se puede medir explícitamente o incluso se pueden eliminar de la serie de datos (a este último se le llama desestacionalización de la serie).

Componente aleatoria: Esta componente no responde a ningún patrón de comportamiento, sino que es el resultado de factores fortuitos o aleatorios que inciden de forma aislada en una serie de tiempo.

De estos tres componentes los primeros son componentes determinísticos, mientras que la última es aleatoria. Así se puede denotar la serie de tiempo como

$$X_t = T_t + E_t + I_t, \tag{5}$$

donde T_t es la tendencia, E_t es la componente estacional e I_t es la componente aleatoria.

Las series de tiempo se pueden clasificar en [15]:

Estacionarias: Una serie es estrictamente estacionaria[8] cuando su distribución de probabilidad es estable a lo largo del tiempo, esto implica que la media y varianza son constantes en el tiempo. Esto se refleja gráficamente en que los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también permanece constante en el tiempo.

No estacionarias: Son series en las cuales la tendencia y/o variabilidad cambian en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante.

[8]Para este trabajo se considera que es una serie débilmente estacionaria, ver definición 2.3.

En la realidad, es posible observar una serie de tiempo una cantidad limitada de veces, y en ese caso la secuencia de valores aleatorios subyacente $(X_1, X_2, \ldots, X_n)$ es solo un vector de n dimensiones. Sin embargo, a menudo es conveniente permitir que el número de observaciones sea infinito. En ese caso $\{X_t, t = 1, 2, \ldots\}$ es llamado un proceso estocástico [13]. Un proceso estocástico puede ser descrito como una secuencia de datos que evolucionan en el tiempo.

A grandes rasgos, un proceso estocástico es estacionario si sus propiedades estadísticas no cambian con el tiempo. Esto implica que su media, varianza y covarianza (en diferentes momentos) permanecen iguales sin importar el momento en el que se midan; es decir, son invariantes en el tiempo.

Definición 2.2. *Sea $\{X_t, t \in \mathbb{Z}\}$ una serie de tiempo estacionaria. La función de autocovarianza (ACVF) de $\{X_t\}$ es*

$$\gamma_X(h) = Cov(X_{t+h}, X_t). \tag{6}$$

La función de autocorrelación (ACF) es

$$\rho_X(h) \equiv \frac{\gamma_x(h)}{\gamma_X(0)}. \tag{7}$$

La **autocorrelación** mide la correlación entre dos variables separadas por k periodos, por lo tanto, otra forma de escribirla es:

$$\rho_k = corr(X_j, X_{j-k}) = \frac{cov(X_j, X_{j-k})}{\sqrt{V(X_j)}\sqrt{V(X_{j-k})}}. \tag{8}$$

La función de autocorrelación simple tiene las siguientes propiedades:

- $\rho_0 = 1$
- $-1 \leq \rho_j \leq 1$
- Simetría $\rho_j = \rho_{-j}$

Definición 2.3. *Una serie de tiempo $\{X_t, t \in \mathbb{Z}\}$ se dice que es débilmente estacionaria si cumple*

(i) $Var(X_t) < \infty$ *para todo* $t \in \mathbb{Z}$,

(ii) $\mu_X(t) = \mu$ *para todo* $t \in \mathbb{Z}$,

(iii) $\gamma_X(r, s) = \gamma_X(r + t, s + t)$ *para todo* $r, s, t \in \mathbb{Z}$.

(iii) implica que $\gamma_X(r, s)$ es una función de $r - s$, y es conveniente definirla como

$$\gamma_X(h) \equiv \gamma_X(h, 0). \tag{9}$$

El valor h es referido como el "*lag* (retraso)".

Definición 2.4. *Un proceso $\{X_t, t \in \mathbb{Z}\}$ se dice que es un ruido blanco (white noise) con una media μ y varianza σ^2, como*

$$\{X_t\} \sim WN(\mu, \sigma^2), \tag{10}$$

si se cumple (11) y (12)

$$EX_t = \mu, \tag{11}$$

$$\gamma(h) = \begin{cases} \sigma^2 & si\ h = 0 \\ 0 & si\ h \neq 0 \end{cases}. \tag{12}$$

Un ruido blanco es un caso simple de los procesos estocásticos, donde los valores son independientes e idénticamente distribuidos[9] a lo largo del tiempo con media cero e igual varianza, se denota por ε_t [15].

$$\varepsilon_t \sim N(0, \sigma^2) \quad cov(\varepsilon_{t_i}, \varepsilon_{t_j}) = 0 \quad \forall t_i \neq t_j. \tag{13}$$

[9]Típicamente se usa una distribución normal $N(\mu, \sigma^2)$, pero no siempre es así.

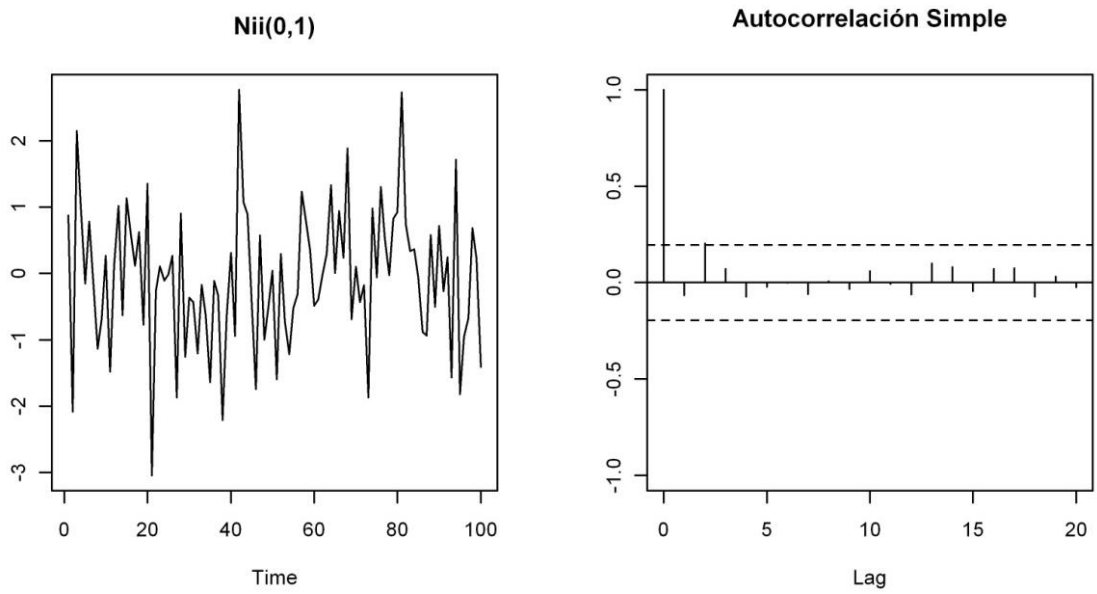


Figura 2.16: La grafica muestra un ruido blanco con media cero y varianza constante e igual a uno. Fuente: Introducción a Series de Tiempo [15].

Un camino aleatorio (*Random Walk*) o camino al azar es un proceso estocástico X_t, donde la primera diferencia de este proceso estocástico es un ruido blanco, esto es $\nabla X_t = \varepsilon_t$.

2.4.1. Procesos lineales estacionarios

Procesos autoregresivos *AR(P)*

Los modelos autoregresivos se basan en la idea de que el valor actual de la serie, X_t, puede explicarse en función de p valores pasados $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, donde p determina el número de rezagos necesarios para pronosticar un valor actual [15].

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + +\phi_p X_{t-p} + \varepsilon_t, \tag{14}$$

donde ε_t es un proceso de ruido blanco. Expresado de manera compacta como:

$$\phi_p(L) X_t = \varepsilon_t. \tag{15}$$

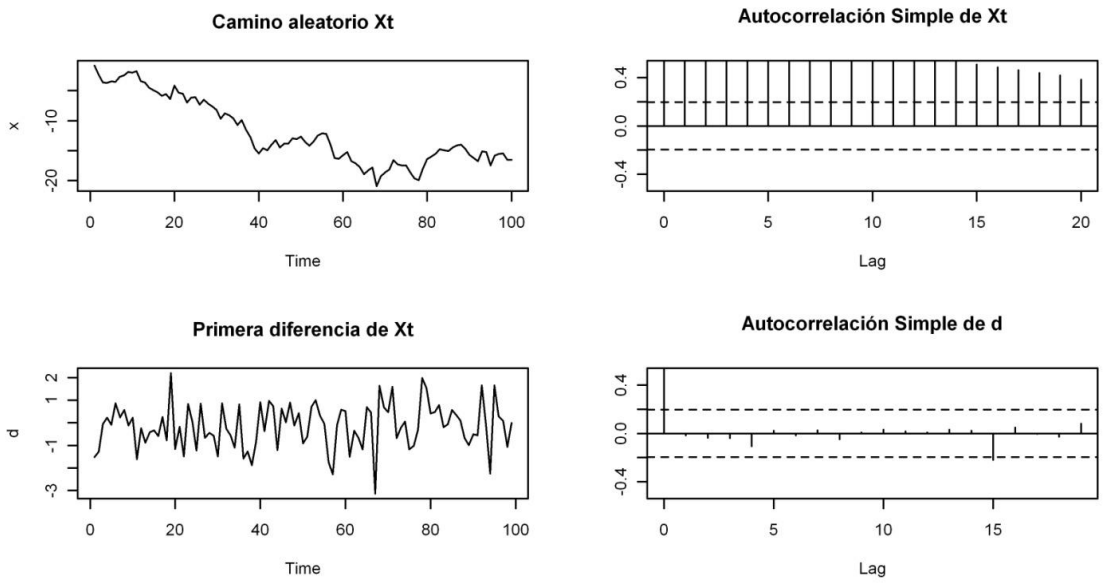


Figura 2.17: La figura muestra el proceso estocástico X_t y su autocorrelación simple, y la primera diferencia del proceso estocástico y su gráfica de autocorrelación. Fuente: Introducción a Series de Tiempo [15].

En los procesos *AR(1)* la variable X_t está determinado únicamente por el valor pasado, esto es $X_t - 1$.

$$X_t = \phi_1 X_{t-1} + \varepsilon_t, \tag{16}$$

donde ε_t es un proceso de ruido blanco con media 0 y varianza constante σ^2, ϕ es el parámetro del modelo.

El modelo es estacionario si $|\phi_1| < 1$, esta constituye una condición necesaria y suficiente para que el proceso *AR(1)* sea estacionario. Para este proceso la función de autocorrelación (ACF) estará dada por $\rho(k) = \phi_1^k$ por lo que se apreciará un decaimiento exponencial en el gráfico de autocorrelación. En el caso general *AR(p)* el proceso es estacionario si el polinomio característico $\phi(z) = 0$ tiene por raíces complejas, números z tales que $|z| > 1$.

Se pueden resumir las condiciones de estacionalidad como:

Modelo *AR(1)*: $X_t = \phi X_{t-1} + \varepsilon_t$, entonces $(1 - \phi L)X_t = \varepsilon_t$

Polinomio autoregresivo: $\phi_1(L) = 1 - \phi L$, las raíces de $1 - \phi L = 0$ son:

$$L = \frac{1}{\phi}. \tag{17}$$

La condición de estacionalidad del modelo *AR(1)* es:

$$|L| = \left|\frac{1}{\phi}\right| > 1, \tag{18}$$

entonces

$$|\phi| < 1. \tag{19}$$

Modelo *AR(2)*: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$, entonces $(1 - \phi_1 L - \phi_2 L^2)X_t = \varepsilon_t$

Polinomio autoregresivo: $\phi_2(L) = 1 - \phi_1 L - \phi_2 L^2$, las raíces de $1 - \phi_1 L - \phi_2 L^2 = 0$ son:

$$L_1, L_2 = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}. \tag{20}$$

La condición de estacionalidad del modelo *AR(2)* es:

$$|L_1| = \left|\frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}\right| > 1, \tag{21}$$

y

$$|L_2| = \left|\frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}\right| > 1. \tag{22}$$

Procesos de medias móviles *MA(q)*

Modelo "determinados por una fuente externa". Estos modelos suponen linealidad, el valor actual de la serie, X_t, está influenciado por los valores de la fuente externa.

El modelo de medias móviles de orden q está dado por:

$$X_t = \theta_0 - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} - \varepsilon_t, \tag{23}$$

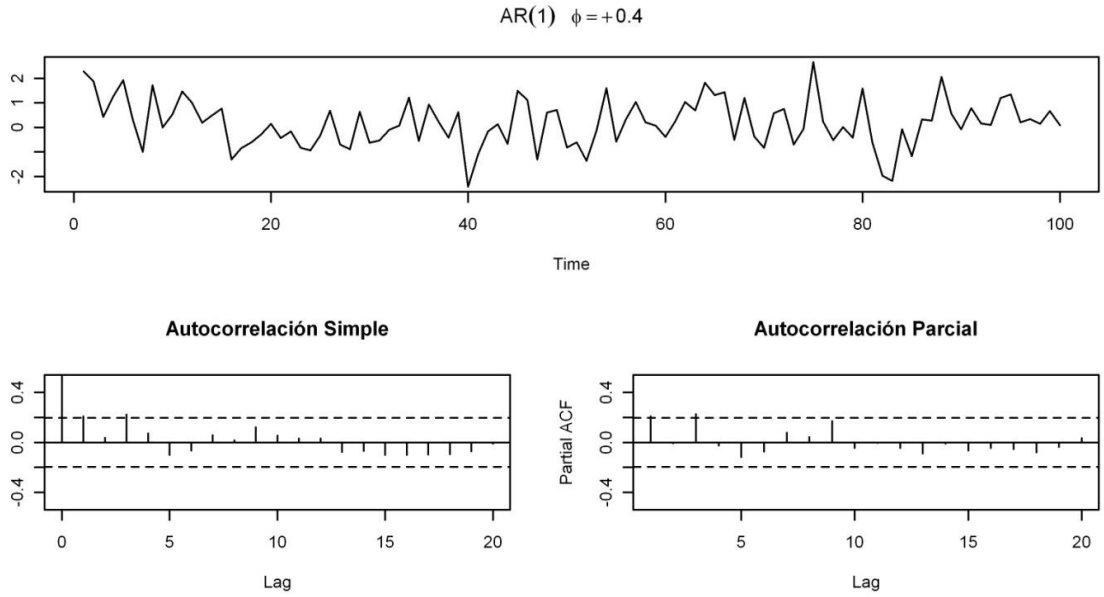


Figura 2.18: La figura muestra la simulación de un proceso autoregresivo de orden uno, esto es $X_t = 0{,}4X_{t-1} + \varepsilon_t$, con sus respectivas gráficas de autocorrelación simple y parcial. Fuente: Introducción a Series de Tiempo [15].

donde ε_t es un proceso de ruido blanco. Expresado de manera compacta como:

$$X_t = \theta_q(L)\varepsilon_t. \tag{24}$$

El modelo de medias móviles *MA(1)* determina el valor de X_t en función de la innovación actual y su primer "lag", esto es:

$$X_t = \varepsilon_t - \theta\varepsilon_{t-1}. \tag{25}$$

Expresado en función del polinomio operador de "lag" es:

$$X_t = (1 - \theta)\varepsilon_t, \tag{26}$$

$$X_t = \theta_1(L)\varepsilon_t, \tag{27}$$

donde ε_t es un proceso de ruido blanco y θ es el parámetro.

Se pueden resumir las condiciones de invertibilidad como:

Modelo *MA(1)*: $X_t = \varepsilon_t - \theta\varepsilon_{t-1}$, entonces $X_t = (1 - \theta L)\varepsilon_t$.

El polinomio de medias móviles está dado por $\theta_1(L) = 1 - \theta L$.Para encontrar las raíces del polinomio se tiene que resolver la ecuación $1 - \theta L = 0$, entonces $L = \frac{1}{\theta}$.

La condición de invertibilidad para un modelo *MA(1)* está dado por: $|L| = \left|\frac{1}{\theta}\right| > 1$, esto es $|\theta| < 1$.

Modelo *MA(2)*: $X_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2}$, entonces $X_t = (1 - \theta_1 L - \theta_2 L^2)\varepsilon_t$.

El polinomio de medias móviles está dado por $\theta_2(L) = 1 - \theta_1 L - \theta_2 L^2$. Para encontrar las raíces del polinomio se tiene que resolver la ecuación $1 - \theta_1 L - \theta_2 L^2 = 0$, las raíces son:

$$L_1, L_2 = \frac{\theta_1 \mp \sqrt{\theta_1^2 + 4\theta_2}}{-2\theta_2}. \tag{28}$$

Las condiciones de invertibilidad para el modelo *MA(2)* están dados por:

$$|L_1| = \left|\frac{\theta_1 + \sqrt{\theta_1^2 + 4\theta_2}}{-2\theta_2}\right| > 1, \tag{29}$$

y

$$|L_2| = \left|\frac{\theta_1 - \sqrt{\theta_1^2 + 4\theta_2}}{-2\theta_2}\right| > 1. \tag{30}$$

Los procesos *MA* se suelen denominar procesos de memoria corta, mientras que a los *AR* se les denomina procesos de memoria larga.

Proceso autoregresivo de medias móviles *ARMA(p,q)*

Cuando una serie de tiempo tiene características de *AR* y de *MA* a la vez[10], se

[10]Es importante destacar que estas series de tiempo pueden ser presentada como proceso *AR(p)* o *MA(q)*, pero al hacer esto el parámetro p o q del proceso tiende a ser mucho mayor que a los parámetros p **y** q del proceso *ARMA(p,q)*. Por ejemplo, una serie puede ser presentada como un proceso *AR(100)* o como uno *ARMA(2,1)*.

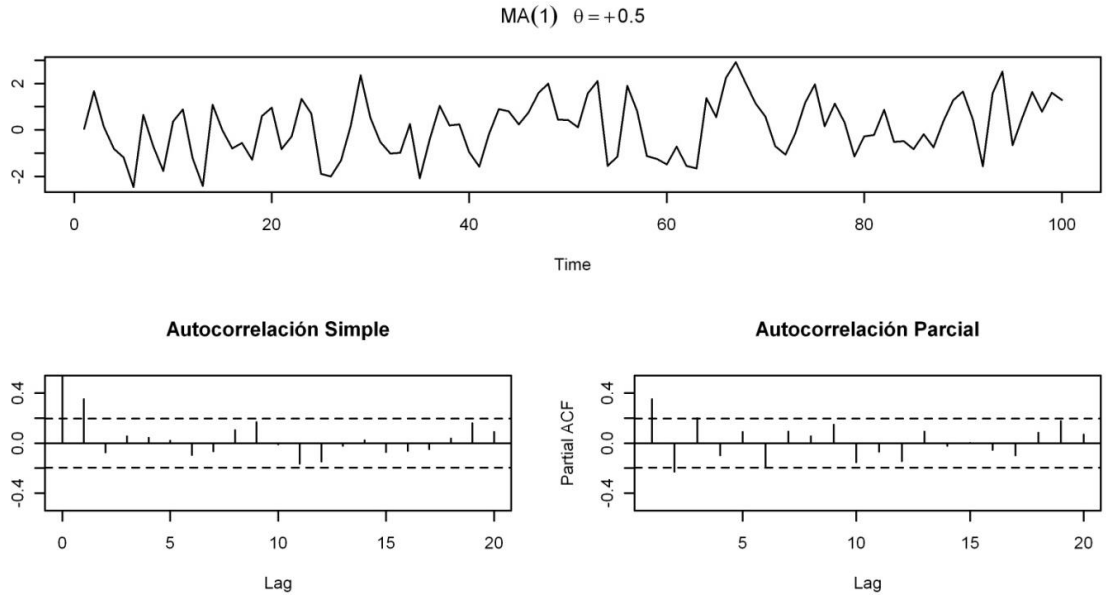


Figura 2.19: La figura muestra un proceso de medias móviles de orden uno, esto es $X_t = \varepsilon_t + 0{,}5\varepsilon_{t-1}$ con sus respectivas gráficas de autocorrelación simple y parcial. Fuente: Introducción a Series de Tiempo [15].

dice que es un proceso *ARMA* con parámetros p y q, dónde estos son los términos autoregresivos y de media móvil, respectivamente.

$$X_t = c + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{31}$$

donde ε_t es un proceso de ruido blanco, y $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ son los parámetros del modelo.

Para un proceso *ARMA(p,q)* una condición de estacionalidad es la misma que para un proceso *AR(p)*, del mismo modo una condición de invertibilidad es la misma que para el proceso *MA(q)*.

Los modelos *ARMA(p,q)* siempre va a compartir las características del modelo *AR(p)* y *MA(q)*, esto es porque contiene a ambas estructuras a la vez. El modelo *ARMA(p,q)* tiene media cero, varianza constante y finita y una función de autocorrelación infinita. La función de autocorrelación es infinita decreciendo rápidamente hacia cero

Modelo *ARMA(1,1)*: Consideremos el modelo *ARMA(p,q)*, donde X_t se determina en función de su pasado hasta el primer "lag", la innovación contemporánea y el pasado de la innovación hasta el "lag" 1.

$$X_t = \phi X_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}, \tag{32}$$

donde ε_t sigue un proceso de ruido blanco, ϕ y θ son los parámetros del modelo.

Para comprobar la estacionalidad del modelo se calculan las raíces del polinomio autoregresivo:

$1 - \phi L = 0$, entonces $|L| = \left|\frac{1}{\phi}\right|$ esto es $|\phi| < 1$

Para comprobar la condición de invertibilidad del modelo se calculan las raíces del polinomio de media móviles:

$1 - \theta L = 0$, entonces $|L| = \left|\frac{1}{\theta}\right|$ esto es $|\theta| < 1$

Características de un proceso *ARMA(1,1)* estacionario:

Media:

$$E(X_t) = E(\phi X_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}) = \phi E(X_{t-1}), \tag{33}$$

$$E(X_t) = 0. \tag{34}$$

Autocovarianzas:

$$\gamma_k = \begin{cases} \gamma_0 = \frac{(1+\theta^2-2\theta\phi)\sigma^2}{1-\phi^2} & k = 0 \\ \gamma_1 = \phi\gamma_0 - \theta\sigma^2 & k = 1 \cdot \\ \gamma_k = \phi\gamma_{k-1} & k > 1 \end{cases} \tag{35}$$

Autocorrelación:

$$\rho_k = \begin{cases} \rho_1 = \phi - \frac{\theta\sigma^2}{\gamma_0} & k = 1 \\ \\ \rho_1 = \phi\rho_{k-1} & k > 1 \end{cases} . \tag{36}$$

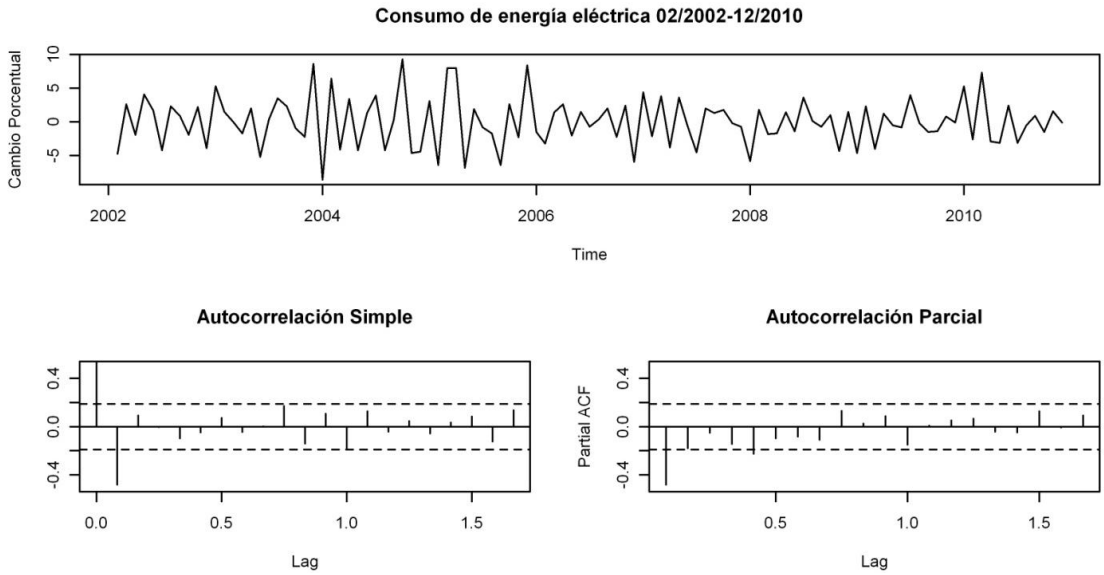


Figura 2.20: La figura muestra la serie del cambio porcentual mes a mes de la serie desestacionalizada del consumo de energía eléctrica en Puerto Rico con sus respectivas graficas de autocorrelación simple y parcial. Fuente: Introducción a Series de Tiempo [15].

2.4.2. Procesos lineales no estacionarios

Proceso Autoregresivo Integrado y de Media Móvil *ARIMA(p,d,q)*

Un proceso *ARIMA(p,d,q)* integra las funciones del modelo *ARMA(p,q)* con la posibilidad de que haya un cambio en la media, varianza y covarianza a lo largo del tiempo[11]. Por consiguiente se debe diferenciar (cálculo diferencial) d veces para hacerla estacionaria y luego aplicarla a esta serie diferenciada un modelo *ARMA(p,q)*, se dice que la serie original es *ARIMA(p,d,q)*, es decir, una serie de tiempo autoregresiva integrada de media móvil. Donde p denota el número de términos autoregresivos, d

[11] Siempre y cuando la serie pueda hacerse estacionaria luego de d diferenciaciones.

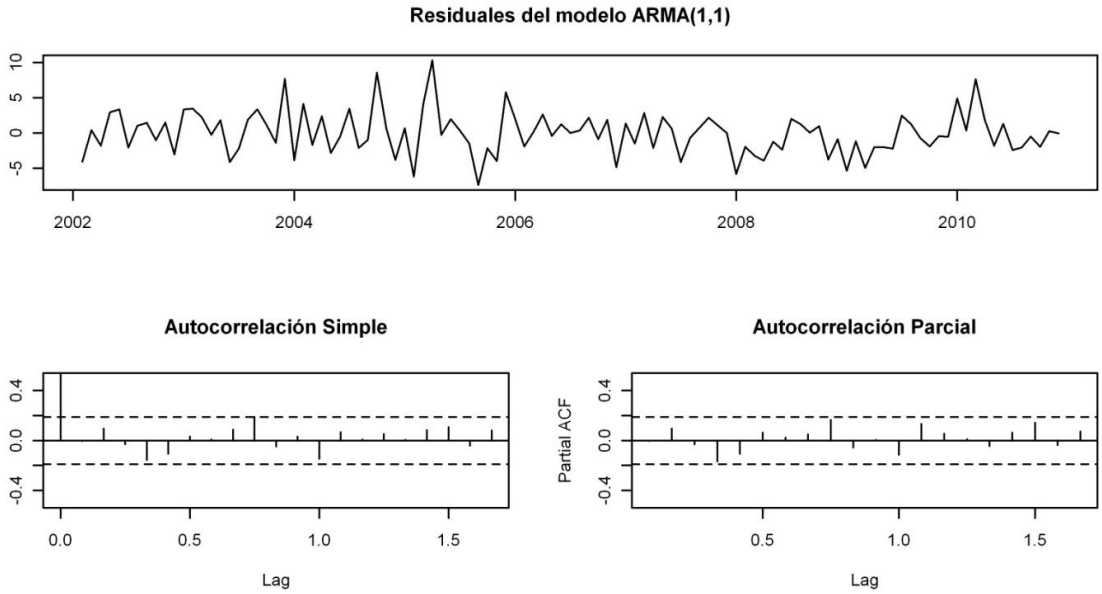


Figura 2.21: La figura muestra los residuales del modelo ARMA(1,1) y las respectivas autocorrelaciones para la serie consumo de energía eléctrica de Puerto Rico. Fuente: Introducción a Series de Tiempo [15].

el número de veces que la serie debe ser diferenciada para hacerla estacionaria y q el número de términos de la media móvil invertible [15].

Su expresión algebraica es:

$$X_t^d = c + \phi_1 X_{t-1}^d + ... + \phi_p X_{t-p}^d + \theta_1 \varepsilon_{t-1}^d + ... + \theta_q \varepsilon_{t-q}^d + \varepsilon_t^d. \tag{37}$$

Expresado en forma del polinomio operador de retardos el modelo *ARIMA(p,d,q)* es:

$$\phi(L)(1-L)^d X_t = c + \theta(L)\varepsilon_t, \tag{38}$$

donde X_t^d es la serie de las diferencias de orden d, ε_t^d es un proceso de ruido blanco, y $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ son los parámetros del modelo.

La construcción de los modelos *ARIMA(p,d,q)* se lleva de manera iterativa mediante un proceso en el que se puede distinguir cuatro etapas [15]:

Identificación Utilizando los datos ordenados cronológicamente se intentará sugerir un modelo *ARIMA(p,d,q)* que merezca la pena ser investigada. El objetivo es determinar los valores p, d y q que sean apropiados para reproducir la serie de tiempo. En esta etapa es posible identificar más de un modelo candidato que pueda describir la serie.

Estimación Considerando el modelo apropiado para la serie de tiempo se realiza inferencia sobre los parámetros.

Validación Se realizan contraste de diagnóstico para validar si el modelo seleccionado se ajusta a los datos. Si no es así, escoger el próximo modelo candidato (lo más simple es variar p y/o q en ± 1) y repetir los pasos anteriores.

Predicción Una vez seleccionado el mejor modelo candidato *ARIMA(p,d,q)* se pueden hacer pronósticos en términos probabilísticos de los valores futuros.

2.4.3. Función *auto.arima* de R

La función *auto.arima* de R usa una variación del algoritmo de *Hyndman y Khandakar* [16], el que combina un **test de raíz unitaria** (*unit root test*), minimización del criterio **AIC** (*Akaike's Information Criterion*) y **MLE** (*Maxium Likelihood Estimation*) para obtener el modelo *ARIMA*.

Unit root test: Es una prueba para saber si una serie de tiempo es estacionaria o no estacionaria con una raíz unitaria. Si el valor de una de las raíces de la ecuación característica es 1, se dice que es una raíz unitaria. Si el valor de las demás raíces es menor a 1, entonces se dice que la primera diferencia del proceso es estacionaria. En caso contrario será necesario aplicar la diferencia múltiples veces para hacer que el proceso sea estacionario.

AIC: El criterio de información de Akaike mide la calidad del modelo estadístico para un set dado de datos. Dado un conjunto de modelos, AIC estima la calidad de

cada modelo respecto a los demás modelos.

$$AIC = L(\hat{\theta}, \hat{x}_0) + 2q, \tag{39}$$

donde q es el número de parámetros en θ más el número de estados fijos en x_0, $\hat{\theta}$ y $\hat{x}_0$ son las estimaciones de θ y x_0, y L es la función de verosimilitud.

MLE: La estimación de Máxima Verosimilitud es un método para estimar los parámetros de un modelo estadístico, seleccionando aquellos valores que maximizan la función de verosimilitud. O sea, aumenta la coincidencia entre el modelo seleccionado y los datos observados, maximizando la probabilidad de que al elegir un valor aleatorio de la muestra esta quede dentro de la distribución resultante.

El procedimiento de la función *auto.arima* puede ser resumido como:

1. El número de diferencias d es determinado usando el test $KPPS$ (test de raíz unitaria elegido).

2. Luego de diferenciar los datos d veces, se estiman los valores p y q usando AIC. En vez de probar todas las combinaciones posibles de p y q, el algoritmo usa una búsqueda escalonada para recorrer el espacio de búsqueda.

   a) Se selecciona el modelo con menor AIC de los siguientes 4:
   ARIMA(2,d,2)
   ARIMA(0,d,0)
   ARIMA(1,d,0)
   ARIMA(0,d,1)
   Si $d = 0$ entonces la constante c es incluida, de otra forma $c = 0$.

   b) Se consideran las siguientes variaciones del modelo seleccionado:

      1) Variar p y/o q del modelo en ± 1;

      2) Incluir o excluir c del modelo.

Se selecciona el mejor modelo considerando también las variaciones.

c) Repetir el paso 2(b) hasta que no se encuentre un AIC menor.

2.5. ISO/IEC 25010:2011 SQuaRE

El estándar de calidad ISO/IEC 25010 SQuaRE (*Software Product Quality Requirements and Evaluation*) define:

- Un modelo de calidad de software compuesto de 8 características, cada una subdividida en subcaracterísticas, las que pueden ser medidas internamente o externamente.

  Estas características son:

  Funcionalidad: Completitud funcional, corrección funcional y pertinencia funcional.

  Eficiencia: Comportamiento temporal, consumo de recursos y capacidad.

  Compatibilidad: Coexistencia e interoperabilidad.

  Usabilidad: Inteligibilidad, aprendizaje, operabilidad, protección frente a errores de usuario, estética y accesibilidad.

  Fiabilidad: Madurez, disponibilidad, tolerancia a fallos y capacidad de recuperación.

  Seguridad: Confidencialidad, integridad, autenticidad y responsabilidad.

  Mantenibilidad: Modularidad, reusabilidad, analizabilidad, capacidad de ser modificado y capacidad de ser probado.

  Portabilidad: Adaptabilidad, facilidad de instalación y capacidad de ser reemplazado.

- Un modelo de calidad para el uso del sistema compuesto de 5 características, cada una subdividida en subcaracterísticas, las que pueden ser medidas cuando

el software es usado en una situación real.

Estas características son:

Eficacia: Eficacia.

Eficiencia: Eficiencia.

Satisfacción: Utilidad, confianza, placer y comodidad.

Seguridad: Riesgo económico, riesgo para la salud y seguridad, y riesgo para el medio ambiente.

Usabilidad: Aprendizaje, flexibilidad y accesibilidad.

3. Desarrollo y resultados

Como se planteó en un comienzo, el objetivo del presente trabajo es obtener información útil para los clientes de e-restó, a partir de los datos históricos de éstos. En este capítulo se presentan los resultados, estamos tomando como guía de *Business Intelligence* los objetivos tres objetivos de análisis: descriptivo, predictivo y de entendimiento.

Primero se presentan como antecedentes el esquema de la base de datos y la cantidad de datos con los que se realizó el análisis, destacando que el modelo presentado es genérico para cualquier negocio de gastronomía, sin incluir aquellos datos sensibles tanto para e-restó como para los clientes.

Luego se presentan los KPI para la generación de reportes y gráficos, con el fin de generar un *dashboard* de BI, considerando como público objetivo los clientes de e-restó. Junto a esto se propone un "layout" para el *dashboard*.

Para finalizar el análisis descriptivo se hace un análisis de los datos faltantes (*missing data*) junto con un análisis de anomalías, para que la gente de e-restó pueda entender mejor cuál es el uso que sus clientes le están dando a la aplicación.

El análisis predictivo busca determinar las ventas de un negocio en el futuro, mediante el pronóstico de series de tiempo usando el modelo lineal ARIMA.

3.1. Análisis descriptivo

3.1.1. Antecedentes

A continuación se presenta el estado inicial de los datos. Esto comprende una breve descripción de la base de datos y el estado de las tablas utilizadas en el presente trabajo.

Esquema de datos

Para el desarrollo del presente, se utilizaron 4 tablas de la base de datos: tabla de Productos, Ventas, Adiciones y Gastos.

A continuación se presenta una breve descripción de cada tabla y las variables relevantes para el análisis.

- La tabla de Productos contiene la información correspondiente a los productos ofrecidos por el negocio.

  Id identificador del registro. Tipo: único y numérico.

  Nombre nombre del producto. Tipo: texto.

  Precio precio de venta por unidad. Tipo: numérico.

  Costo costo de producción por unidad. Tipo: numérico.

  Stock unidades del producto en inventario. Tipo: numérico.

  Categoría tipo de producto. Tipo: numérico.

- La tabla de Ventas contiene la información de la venta hecha en una mesa.

  Id identificador del registro. Tipo: único y numérico.

  Fecha fecha de la venta. Tipo: fecha.

  Personas cantidad de personas en la mesa. Tipo: numérico.

  Creado a las fecha y hora del comienzo de la venta. Tipo: fecha con hora.

  Cerrado a las fecha y hora del término de la venta. Tipo: fecha con hora.

  Mesa identificador de la mesa en la que se hace la venta: Tipo: numérico.

  Camarero identificador del camarero que atiende la venta. Tipo: numérico.

- La tabla de Adiciones contiene la información de cada adición asociada a una venta.

  Id identificador del registro. Tipo: único y numérico.

  Creado a las fecha y hora de la orden del producto. Tipo: fecha con hora.

  Producto identificador del producto que se ordenó. Tipo: numérico.

Precio precio por unidad del producto. Tipo: numérico.

Cantidad cantidad de unidades del producto ordenado. Tipo: numérico.

Venta identificador de la venta asociada a la adición. Tipo: numérico.

- La tabla de Gastos contiene la información de los gastos producidos en el negocio.

Id identificador del registro. Tipo: único y numérico.

Fecha fecha del gasto. Tipo: fecha.

Monto monto del gasto. Tipo: numérico.

Categoría categoría del gasto. Tipo: numérico.

La relación entre las tablas está representada en la figura 3.1, donde cada venta tiene asociada varias adiciones, las cuales a su vez están asociadas a un producto. En otras palabras, una venta asocia todos los productos comprados en una mesa.

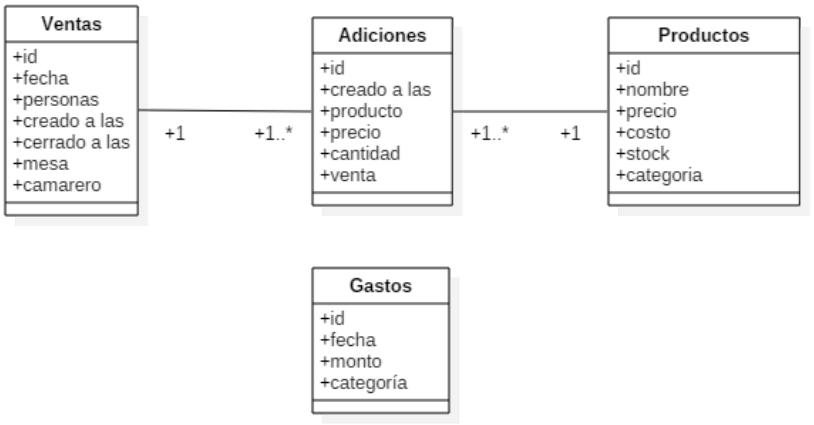


Figura 3.1: Modelo UML que representa la relación existente entre las tablas usadas.

La tabla de Gastos no posee asociación directa con las otras tres tablas, pero si es usada al momento de hacer la cuadratura entre los gastos y los ingresos del negocio.

Para mantener los datos de los negocios separados uno de los otros cada negocio es creado como un esquema en la base de datos, así es posible que cada cliente tenga sus propias tablas compartiendo el modelo de datos.

Por seguridad de los datos y derechos de autor no se presenta el esquema completo de la base de datos, como también se han modificado el nombre de las variables con el fin de presentarlo de la forma más genérica posible.

Estado de los datos

Se utilizaron como muestra los datos de 5 negocios distintos: A, B, C, D y E, todos pertenecientes a un mismo país. Estos negocios fueron seleccionados al azar entre otros 600 negocios del mismo país. Si bien esta no es una muestra representativa, lo que se busca es generar indicadores que puedan ser aplicables a todos los negocios, por lo que estos 5 negocios son suficientes para el análisis de las variables.

En la tabla 3.1 se encuentra anotada la cantidad de registros que posee cada tabla y el mes de inscripción en la plataforma[12].

	Productos	**Ventas**	**Adiciones**	**Gastos**	**Cliente desde**
Negocio A	1.504	37.957	327.475	6.765	Dic-2013
Negocio B	59	112.246	152.842	7.952	Feb-2012
Negocio C	249	18.791	135.367	25	Abr-2014
Negocio D	419	128.363	233.494	5.270	Jun-2012
Negocio E	136	15.734	43.654	4.009	Dic-2013

Tabla 3.1: Cantidad de registros por tabla y fecha de inscripción del cliente.

Es importante destacar que existe la libertad para que cada negocio use los módulos que estime conveniente, por lo que existen columnas en las tablas que se encuentran completa o parcialmente vacías. Esto se tratará más adelante como un KPI práctico, ya que puede revelar información sobre el uso que le dan los clientes a la aplicación.

[12]Se considera desde el segundo mes, ya que el primer mes es de pruebas y aprendizaje.

3.1.2. Variables de éxito de un negocio

Haciendo un catastro de la cantidad de datos con la se contaba de cada usuario, se determinó que el uso que estos le dan a la herramienta son para satisfacer las necesidades de sus negocios, las cuales varían mucho entre los negocios, por lo que se descartó la idea de generar un modelo "genérico" para la clasificación y ranking de variables, y se optó por generar un *dashboard* de BI para presentar las variables más relevantes (y con las que se contaban más datos) para la toma de decisiones de los administradores de los negocios (cliente).

Además, se hizo un análisis a nivel de datos (capítulo 3.1.4) para que e-restó mejore el entendimiento de las necesidades de sus clientes, y de esta forma dar más valor a su aplicación.

3.1.3. KPIs

Para poder determinar los KPI a usar, se hicieron varias reuniones, en las cuales se realizó la siguiente pregunta a los clientes: ¿Qué es lo que ellos quieren de sus negocios? Hubo diversas respuestas que apuntaban a lo mismo, **hacer que sus negocios crecieran**. Tomando este objetivo en cuenta, se definieron variables que pudiesen ser controladas por los clientes y afectaran en el crecimiento de sus negocios, tomando como fuente de datos la misma base de datos de e-restó. Con estas variables se definieron los KPI presentados a continuación.

Estos KPI son todos del tipo cuantitativo, ya que las variables utilizadas son todas numéricas, y la calidad del valor de cada una depende del negocio.

Como se cuenta con datos históricos, se intenta sacar el máximo provecho de esto, presentando en la mayor parte de los casos la variable temporal, la que puede ser agrupada por día, semana, mes o año, dependiendo del nivel de granularidad que se quiera mostrar (de acuerdo a lo presentado en el capítulo 2.2).

A continuación, se listan los KPIs generados, presentando una breve descripción de

estos y el modo en que se presentarán:

Ventas El objetivo de este KPI es presentar un historial de las ventas totales desde una fecha predeterminada.

Este es un KPI financiero que muestra las ventas agrupadas desde un determinado mes, esta agrupación puede ser diaria, semanal, mensual o anual.

La forma en que se presenta este KPI es un gráfico de líneas, el que puede ser presentado solo o junto al KPI "Gastos". El nivel de agregación con el que se presenta por defecto este KPI es mensual y contando los últimos 24 meses. Esta cantidad de datos fue elegida para poder dilucidar lo que ocurrirá con las ventas el siguiente mes, considerando el comportamiento de los últimos 2 años, y también para mantener el tamaño del eje temporal. En la figura 5.1 se muestran gráficos de ejemplo para las ventas de los negocios A y E, teniendo en cuenta que la idea es presentar de una forma simple y clara la información.

Gastos El objetivo de este KPI es presentar un historial de los gastos totales desde una fecha predeterminada.

Este es un KPI financiero que muestra los gastos agrupados desde un determinado mes, esta agrupación puede ser diaria, semanal, mensual o anual.

Se usó el mismo criterio del KPI de "Ventas" para presentar este KPI. En la figura 5.2 se muestran gráficos de ejemplo para los gastos de los negocios A y E, esta vez con un color más cercano al rojo, para contrastar con el color de las ventas.

Como opción en el *dashboard* se podrán mostrar ambas variables en el mismo gráfico, para poder comparar ambas variables en la misma escala, lo que queda plasmado en la figura 5.3, donde se contrastan las ventas y gastos del negocio A.

Ventas por mesa El objetivo de este KPI es presentar ranking de las mesas que más venden en un determinado mes.

Al igual que los KPI de ventas y gastos, este es un KPI financiero que muestra las ventas en un determinado mes, pero a un nivel más granular, agregando la

dimensión de las mesas.

El propósito de este KPI es generar un ranking "top *n*" de las mesas en un determinado mes (siendo *n* un valor parametrizable), en orden descendiente, el que es presentado en forma de tabla para mostrar precisión en los valores de las ventas de cada mesa. Como ejemplo se encuentra la tabla 5.1, donde se muestra un Top 10 de las mesas con más ventas en el mes de enero del 2015.

Una de las conclusiones que se busca de este KPI es determinar cuáles serán las mesas con más ventas en el siguiente mes, tomando como referencia el ranking del mismo mes del año pasado, suponiendo que la posición de estas no ha cambiado en el tiempo.

Mesas Se definen 3 KPI para tener datos más detallados de cada mesa.

Este es un KPI práctico, que revela datos sobre el proceso existente.

El objetivo de estos es presentar información detallada en cuanto al uso de las mesas, ya sea en tiempo de uso, cantidad de veces que se usa (rotación) y monto promedio que vende por hora.

Tiempo de uso diario: El objetivo de este KPI es presentar el tiempo promedio de uso de las mesas durante un día.

Cantidad de usos diarios: El objetivo de este KPI es presentar la cantidad de veces que se usa una mesa durante un día.

Producción por hora: El objetivo de este KPI es presentar las ventas promedio por hora de una mesa en un intervalo de tiempo.

Los resultados de estos KPI son presentados en forma de tabla (ejemplo tabla 5.2) para mostrar todo el detalle de estos valores. Se agrega una columna que muestra la producción promedio por día, siendo esta el producto de la producción por hora y las horas de uso de la mesa.

Unidades vendidas por producto de cierta categoría El objetivo de este KPI es presentar la cantidad de unidades que se venden de los productos de cierta categoría

en un determinado mes.

Este es un KPI salida, que revela datos sobre el resultado de un proceso existente, tomando en cuenta como resultado la cantidad de ventas del negocio.

Se presenta en un gráfico de sectores (*pie chart*) para poder destacar y mostrar el porcentaje de ventas de cada producto respecto al total de esa categoría. En la figura 5.4, se muestra un gráfico de sectores correspondiente a la categoría de Cervezas del negocio A, en el mes de enero del 2015. Los productos se presentan en orden decreciente para facilitar la asociación entre el producto y el sector en el gráfico.

Otra forma de presentar este KPI es en una tabla para tener más detalle de los datos.

Monto obtenido por ventas por producto de cierta categoría El objetivo de este KPI es presentar el monto obtenido por las ventas de los productos de cierta categoría en un determinado mes.

Este es un KPI salida, que revela datos sobre el resultado de un proceso existente, tomando en cuenta como resultado el monto de ventas del negocio.

Al igual que en el KPI anterior, se presenta un gráfico de sectores, pero esta vez tomando en cuenta el monto vendido y no la cantidad. El fin de esta diferencia es poder notar si existe alguna diferencia entre la cantidad que se vende y el monto que se vende, esto queda más claro en la figura 5.5, donde se nota que el producto "1890 1Lt" esta vez ocupa el 17 % del gráfico y "Bajo Cero 1Lt" un 16 %, notar que los productos ya no estan ordenados de forma decreciente.

Otra forma de presentar este KPI es en una tabla para tener más detalle de los datos.

Monto obtenido por ventas de los productos por categoría El objetivo de este KPI es presentar el monto obtenido por las ventas de los productos agrupados por categoría.

Este es un KPI salida, que revela datos sobre el resultado de un proceso existente,

tomando como resultado el monto de ventas del negocio.

Se presenta en un gráfico de sectores al igual que en los KPI anteriores, pero esta vez mostrando el monto vendido por categorías, para así poder diferenciar cuál es la categoría que más vende. Esta vez solo se presenta el monto vendido, ya que la cantidad de unidades vendidas siempre será mayor en aquellos productos que son de consumo frecuente, tales como los bebestibles. En la figura 5.6 se muestran las ventas de las distintas categorías del negocio A (como ejemplo se muestra la ID y no los nombres).

Otra forma de presentar este KPI es en una tabla para tener más detalle de los datos.

Camareros Se definen 2 KPI para medir la eficiencia de cada camarero, tomando en cuenta tanto la cantidad de ventas como el monto vendido en un día.

Este es un KPI de proceso, que revela la eficiencia del uso del tiempo de cada camarero.

Cantidad de ventas diaria de los camarero El objetivo de este KPI es presentar la cantidad de ventas diaria promedio de los meseros desde una fecha predeterminada.

Monto obtenido por ventas diarias de los camareros El objetivo de este KPI es presentar el monto diario promedio generado por los camareros desde una fecha predeterminada.

Ambos KPI se presentan en una tabla para mostrar el detalle de los datos obtenidos. Se habilita la opción para ordenar la tabla por cualquiera de las columnas, ya sea para generar un ranking o para buscar el desempeño de un camarero determinado. En la tabla 5.3 se muestra el resultado para los camareros del negocio C desde enero del 2015, ordenados por ID del camarero (No se muestra el nombre del camarero en el ejemplo).

3.1.4. *Missing Data*

A partir del análisis de los antecedentes se descubrió que la base de datos no se utilizaba completamente por los clientes, generando como primera impresión que los procesos de negocios usados por los clientes no eran los mismos procesos planteados por e-restó en su aplicación.

Para mejorar el entendimiento del uso que los clientes le dan a la aplicación, se generaron consultas de base de datos para obtener la cantidad de datos faltantes por columnas de cada tabla, siendo estos datos faltantes datos en blanco o valores que no corresponden a la realidad en determinada variable. Por ejemplo, en la tabla 3.2 se hace una revisión de la tabla de productos para los 5 clientes, donde se muestra el porcentaje de datos “faltantes”, siendo 100 el caso en que no hay datos en esa columna. En este caso se puede ver que el Stock para cada producto es cero o menor a cero (se cuentan los valores negativos ya que un stock menor a cero no es real").

Con estos datos, e-restó puede detectar aquellos negocios que no usan todas las funciones de su aplicación y así generar una capacitación focalizada a estos negocios, además de también obtener información sobre los procesos de negocios de sus clientes y así mejorar la experiencia de usuario de la aplicación de e-restó.

3.1.5. Anomalías

A partir del KPI de “Ventas desde un determinado momento” y del análisis predictivo (capítulo 3.2), se notó que existían anomalías (*outliers*) en los datos de ventas (y gastos), por lo que surgió la necesidad de poder identificarlos para su revisión, tanto para el cliente como para e-restó.

Para la detección de las anomalías, se probaron 3 opciones: criterio del *BoxPlot* para la detección de *outliers*, detectar el 1 % de los datos (0.5 % menor y 0.5 % mayor) y la función de probabilidad que representara de mejor forma los datos y considerar como anómalos aquellos con menos de x % de probabilidad de aparición.

Table "productos"					
Columna	**Negocio A**	**Negocio B**	**Negocio C**	**Negocio D**	**Negocio E**
ID	0	0	0	0	0
Var1	99	0	0	0	0
Nombre	0	0	0	0	0
Precio	0	0	0	0	0
Var2	0	0	0	0	0
Costo	40	100	50	80	50
Stock	100	100	100	100	100
Var3	0	0	0	0	0
Var4	0	0	0	0	0
Var5	0	100	100	100	100
Categoria	0	0	0	0	0
Var6	93	100	99	92	92
Var7	0	0	100	0	0
Var8	31	10	0	8	8
Var9	100	100	100	100	100
Var10	0	100	0	100	100
Var11	0	100	100	100	100
Var12	100	100	100	100	100
Var13	0	100	0	0	0
Var14	0	0	0	0	0

Tabla 3.2: Porcentaje de datos faltantes por cada columna de la tabla de productos para los 5 negocios.

Los resultados esperados y obtenidos por las distintas opciones fueron los siguientes:

Criterio del *BoxPlot*: En este caso, se pudo detectar fácilmente aquellos valores que superaban el tercer cuartil en $1{,}5 * IQR$, tanto para ventas como para gastos.
La desventaja encontrada en esta opción fue que la cantidad de anomalías cambiaba dependiendo del nivel de "agregación" que se usaba, lo mismo que ocurría al momento de elegir una escala en la figura 2.1, mientras mayor agregación, menor cantidad de anomalías encontradas.

Detectar el 1 %: En este caso, se pudo detectar el 0,5 % de los valores de cada extremo (mayor y menor), tanto para las ventas como para los gastos. Teniendo menor consumo de recursos que la primera opción.
La desventaja encontrada fue que no se discrimina entre anomalías y datos normales. Además se requería que hubiese una cantidad mínima de 200 datos (después de la agregación) para poder detectar el primer y último dato. Aún después de cumplir esa condición, no se aseguraba que estos fuesen los únicos datos anómalos.

Función de probabilidad: En este caso, no hubo forma de implementar a nivel de base de datos una función para encontrar una función de probabilidad que representara los datos, por lo que se requería de un sistema externo a la base de datos, aumentando así el costo computacional de esta opción.

Finalmente para la detección de estos datos se usó el criterio del *BoxPlot* dado su costo de implementación y consumo de recursos, además de ser un criterio bastante usado en la literatura para la detección de datos *outliers*.

3.2. Análisis predictivo

Para la generación de las predicciones, se usó la plataforma de Microsoft Azure Machine Learning, en donde se optó por usar un módulo de R para la predicción mediante

la función *auto.arima*.

La razón por la que se optó por este módulo, y no un modelo prefabricado de Azure ML, es que se necesita poder generar un modelo de predicción para cada negocio, dado que se esta haciendo uso de series de tiempo. Un modelo genérico para las predicciones (considerando todos los negocios) requiere de más variables para distinguir el tipo de negocio, y por ende un modelo más complejo que éste, además de la necesidad de tener que entrenar el modelo periódicamente (y no *on-demand*) dada la gran cantidad de datos existentes.

3.2.1. Resultados función *auto.arima*

Para las pruebas se cargaron los datos de ventas de los 5 negocios en Azure ML, luego, en un experimento se probaron los 5 set de datos por separados para comparar los resultados.

En la figura 3.2 se encuentra el experimento hecho en Azure ML, donde primero se muestra el set de datos (negocio A). Luego se usa un módulo de transformación SQL para manejar el nivel de agregación de los datos (ya sea diario, semanal, mensual o anual). Después se usa Python para la detección y manejo de anomalías, donde estas son llevadas dentro del rango del límite superior del *BoxPlot*[13]. Finalmente se usa el módulo de R para generar las predicciones mediante *auto.arima*.

Éste último módulo, genera 2 resultados, el nodo de la izquierda retorna las predicciones y el nodo de la derecha el gráfico auto generado por R.

Usando los datos del negocio D (el cual contiene dos anomalías destacables), se generaron 2 predicciones, la primera sin hacer uso del manejo de anomalías y la segunda haciendo que las anomalías queden dentro de los límites del *BoxPlot* (figura 3.3).

En el gráfico de la izquierda de la figura 3.3 los *outliers* hacen que la serie de tiempo parezca una recta en el eje $y = 0$, además el intervalo de confianza que se

[13] Actualmente solo se considera el límite superior porque todos los datos son positivos, y el límite inferior del *boxplot* se encontró siempre en los negativos. Es posible, de ser necesario, homologar el criterio para el límite inferior.

Figura 3.2: Experimento en Microsoft Azure Machine Learning.

genera alrededor de las predicciones varía desde $y > -1 * 10^7$ y $y < 1 * 10^7$, el cual considera valores negativos. En cambio en el gráfico de la derecha, se puede apreciar que la serie de tiempo no es lineal y que el intervalo de confianza alrededor de las predicciones no es tan amplio como en el primer caso, haciendo notar que la función *auto.arima* es sensible a los *outliers*. También se puede ver diferencia en la elección del modelo, siendo el primero un modelo ARIMA(0,0,0) con media cero y el segundo un modelo ARIMA(1,1,0) con media distinta de cero.

Para medir la calidad de los resultados, se usan las siguientes métricas:

Mean Error (ME) Promedio de los errores de predicción (considerando como error la diferencia entre el valor predicho y el valor real) del set de prueba.

Root Mean Squared Error (RMSE) Es la raíz cuadrada del promedio de los errores al cuadrado del set de pruebas.

Mean Absolute Error (MAE) Es el promedio del valor absoluto de los errores.

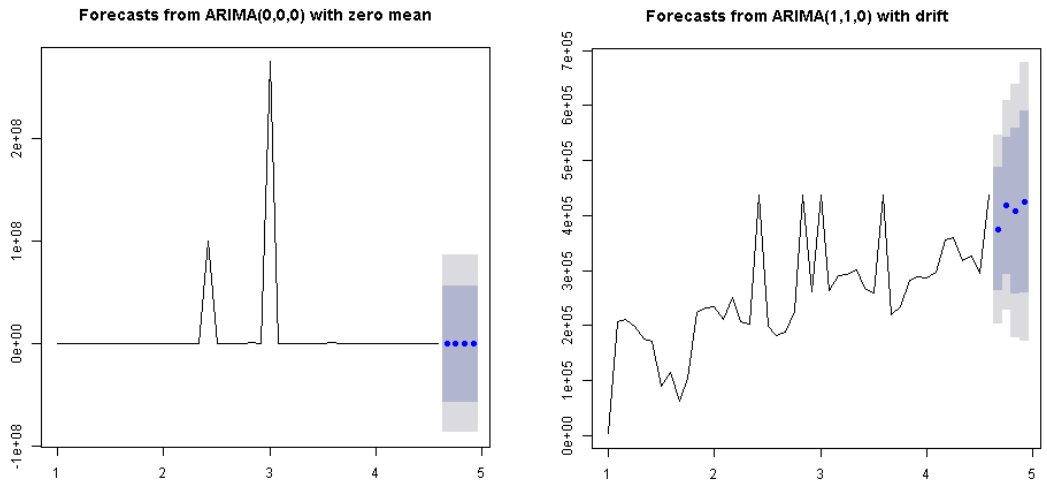


Figura 3.3: Predicciones de ventas usando auto.arima en R para el negocio D, en el gráfico de la izquierda sin manejo de anomalías y a la derecha usando el criterio del BoxPlot.

Mean Percentage Error (MPE) Es el promedio porcentual de los errores.

Mean Absolute Percentage Error (MAPE) Es el promedio porcentual del valor absoluto de los errores.

Se aplicaron estas métricas para las predicciones aplicadas en los 5 negocios, los resultados de esto se encuentran en la tabla 3.3.

	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Negocio A	-63918.61	142199.12	124372.39	4.61 %	10.91 %
Negocio B	-347210.98	580521.36	356327.28	57.85 %	61.85 %
Negocio C	-27714.51	306253.63	267231.37	-11.55 %	37.26 %
Negocio D	-197066.17	432399.11	264771.74	10.09 %	32.21 %
Negocio E	-271204.73	550363.40	289283.85	17.24 %	34.97 %

Tabla 3.3: Métricas de error de predicción.

Si bien los valores obtenidos para el ME, RMSE y MAE son “grandes”, estos en comparación a los valores de las variables son menores en orden, lo que se nota en

los valores de MPE y MAPE, los que corresponden a porcentaje de error. Solo en un negocio se obtuvo un MPE y MAPE mayor al 50 %.

Con estos valores es posible afirmar que la función *auto.arima* genera predicciones aceptables con un error menor al 50 % para la mayoría de los casos.

3.2.2. Web Service

Es posible utilizar el experimento generado en Azure ML como un Web Service, el cual puede ser consumido por cualquier aplicación externa que cuente con la llave API (*API key*) del experimento.

En la figura 3.4 se puede ver que ahora el origen de los datos proviene de un *Web service input*, el cual debe ser configurado para recibir los datos en un formato determinado, al igual que el *Web service output* debe ser configurado para retornar los valores de la predicción.

Figura 3.4: Web Service en Microsoft Azure Machine Learning.

Con este *Web Service* implementado es posible para la aplicación de e-restó enviar los datos de ventas o gastos de un cliente y recibir como respuesta la predicción hecha por el modelo, con un tiempo de respuesta *near real time*.

Un ejemplo del resultado obtenido se puede puede presentar como gráfico en la figura 3.5, donde se predijeron las ventas del negocio A para los meses de febrero, marzo, abril y mayo de 2016. También es posible apreciar que la predicción sigue la tendencia de crecimiento que llevan las ventas del negocio.

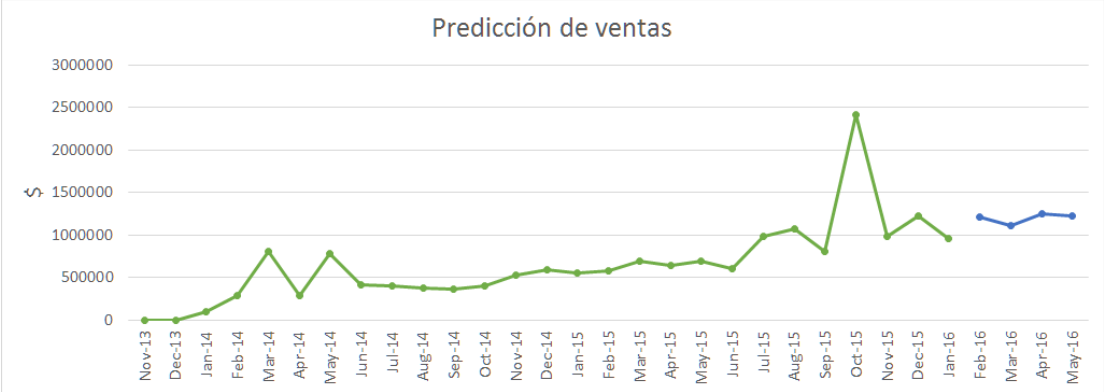


Figura 3.5: En verde se muestra el historial de ventas del negocio A y en azul las predicciones obtenidas de Azure.

3.3. Propuesta de *Layout* para el *Dashboard*

Para presentar los datos a los clientes de e-restó (dueños de los locales), se plantea el uso de un *dashboard* para presentar los KPI definidos anteriormente.

Siguiendo el diseño actual de la aplicación de e-restó (figura 5.7), se propone generar una pantalla principal con un resumen de los KPI más importantes y simples de entender, lo que abarca los KPI de:

- Gráfico de "Ventas desde un determinado momento" con predicciones, con la opción de cambiarlo o combinarlo con el gráfico de "Gastos desde un determinado momento", considerando los últimos 24 meses.

- Tabla “top 10 de Ventas por mesa desde un determinado momento” considerando los últimos 3 meses.
- Tabla “top 10 de productos más vendidos de cierta categoría”, correspondiente al mes en curso, con las opciones de elegir una categoría y cambiar el mes.
- Tabla “top 10 camareros”, con opción para ordenar por cantidad de ventas o monto promedio, considerando los últimos 2 meses.

En la figura 3.6 queda reflejada la propuesta inicial para presentar un resumen de los datos.

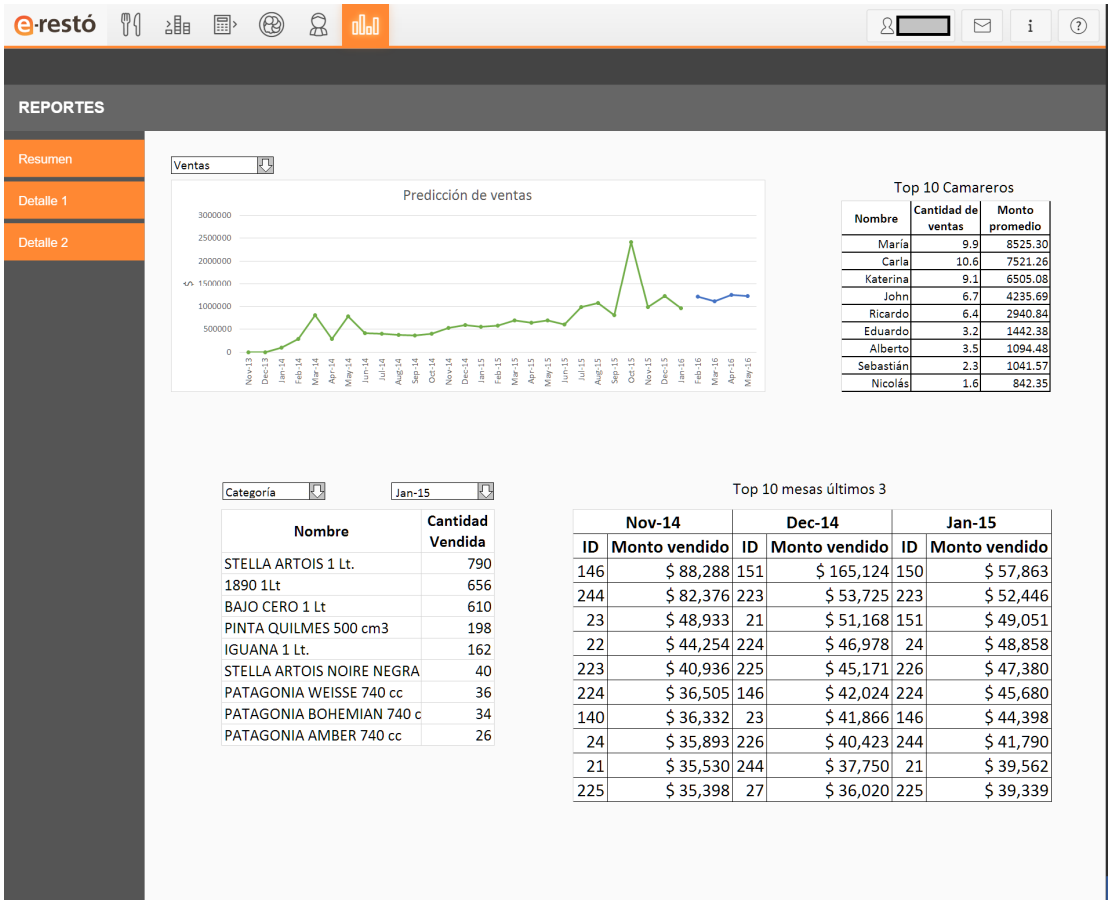


Figura 3.6: Propuesta de layout para la pantalla de resumen.

Además de la pantalla principal, se propone agregar 2 pantallas más, para el detalle de los productos y mesas, dejando en estos sus KPIs correspondientes. Ejemplos figuras

5.8 y 5.9.

Es importante destacar que esta presentación es la primera versión propuesta, la que irá cambiando para mejorar la experiencia del usuario, así como también se irán agregando KPIs a medida que los usuarios los vayan sugiriendo.

3.4. Experiencia con Microsoft Azure Machine Learning

Se hace un análisis de la experiencia del uso de Microsoft Azure Machine Learning, siguiendo las métricas de la ISO 25010. Lo siguiente son comentarios en base a la experiencia personal y bajo ninguna circunstancia es con el objetivo de hacer un *benchmarking*. Se utiliza la ISO para ordenar y contextualizar los factores.

Eficacia: En cuanto a la eficacia, en el momento del desarrollo del presente trabajo, se contó con todas las funcionalidades esperadas, esto incluye los módulos de *machine learning* implementados en Azure ML como también los módulos de R y Python para hacer implementaciones personalizadas a las necesidades del usuario.

Los resultados concuerdan con lo esperado, esto es en otras palabras que se obtienen los mismos resultados de un módulo de R en la Azure ML y un programa en R ejecutado localmente.

Para la interacción con otras plataformas, Azure ML cuenta con la posibilidad de publicar el modelo generado como un Web Service, donde uno puede configurar las variables que uno desea que el modelo reciba y envíe como respuesta. Para la fuente de datos, la plataforma los puede extraer como archivo en formato csv desde una URL o desde una base de datos en Azure (además de los archivos que uno puede cargar en la plataforma).

No se tuvieron problemas con la seguridad de los datos o accesos no autorizados al experimento.

Eficiencia: Una característica importante de destacar de la plataforma es que para aho-

rrar tiempo de ejecución en los experimentos, solo se ejecutan aquellos módulos que han sido modificados o reciben como entrada la información de un módulo modificado. Por ejemplo, si se cuenta con los módulos A->B->C->D->E, donde A es la fuente de datos y E es el último módulo, cuando se ejecuta el experimento por primera vez se "compilarán" todos los módulos, ahora si se modifica el módulo C y luego se ejecuta el algoritmo solo se compilarán los módulos C, D y E, ahorrando el tiempo de compilado de los primeros 2 módulos.

Satisfacción: La herramienta resultó ser bastante útil para el desarrollo del trabajo, ya que se contaba con una gran cantidad de herramientas en un solo lugar, sin la necesidad de tener que hacer instalaciones en la máquina local, resultando ser una característica bastante cómoda para hacer las pruebas.

Seguridad: No es posible hacer comentarios sobre este punto, ya que durante el desarrollo del trabajo no se encontraron fallas en la plataforma y esta se encontraba disponible cada vez que se hacían pruebas.

Usabilidad: En lo que se refiere a la usabilidad, la plataforma puede parecer un poco abrumadora para un usuario que se está iniciando en el mundo de las máquinas de aprendizaje por la cantidad de funciones que tiene, pero se vuelve más simple de usar a medida que uno va aprendiendo sobre los algoritmos y sus funciones.

El hecho que los modelos puedan ser creados a modo de "*Drag and Drop*" acelera la curva de aprendizaje del uso de la plataforma por la facilidad para poner a prueba los modelos sin necesidad de crear un código completo (*end to end*), con esto uno comienza a visualizar los resultados de una forma más rápida.

Es posible personalizar los algoritmos prefabricados a través de los parámetros de cada uno, aumentando así el control sobre el algoritmo.

Otro factor importante que influye en el aprendizaje del uso de la plataforma es la cantidad de ejemplos genéricos que existe en la galería *Cortana Intelligence Gallery*, con estos ejemplos uno se puede hacer una idea rápidamente de los

módulos que hay que usar para hacer un análisis, como por ejemplo, un análisis léxico.

Otro punto importante a destacar es que durante el desarrollo del trabajo se hizo uso del contacto a través del chat con un experto de Azure ML. El motivo de este contacto fue por unas dudas técnicas y se obtuvo una respuesta en tiempo real por parte de un experto, el cual asistió hasta que todas las dudas fueron aclaradas.

Al momento de comenzar el trabajo, la única forma de entrenar el algoritmo era ejecutando el experimento a través de la plataforma. Luego se habilitó la opción para exponer un *web service* y entrenar el algoritmo sin necesidad de acceder a la plataforma. Lo que da a cuentas de que existe un desarrollo constante de esta por parte de Microsoft.

4. Conclusiones

El propósito principal del presente trabajo fue asistir en la toma de decisiones a los administradores de los locales que usan la aplicación **e-restó**, lo que fue logrado a través de la definición de KPIs para presentar las variables que pudiesen contribuir a tomar estas decisiones. Estos KPI fueron pensados y diseñados a partir de varias reuniones con algunos de los administradores de locales para que cumplieran sus necesidades y fuesen útiles para ellos.

Se sugirió el diseño de un *dashboard* para la presentación de estos KPI en la aplicación de e-restó, siguiendo el marco teórico para la visualización de datos. Este diseño es una primera versión que se espera que vaya mejorando a través de la retroalimentación obtenida de los clientes, como también la generación de nuevos KPI para mejorar la experiencia del usuario.

Este *dashboard* se complementa con la predicción generada en la plataforma de Microsoft Azure Machine Learning, la cual arrojó errores menores al 20 % para cuatro de los cinco negocios usados como prueba.

A medida que se analizaban los datos, también se descubrieron cosas interesantes sobre los clientes de e-restó, estos no hacían uso de la totalidad de las funciones de la aplicación, por lo que surgió la necesidad por parte de e-restó de identificar cuáles son las funciones menos usadas y ajustar éstas para que se adapten a las reglas de negocios de sus clientes. Esto fue hecho a través del análisis de *missing data* y anomalías, donde por un lado se obtienen los datos faltantes en la base de datos y por otro lado anomalías producto de fallas de la aplicación o uso inadecuado de ésta.

Como primera medida e-restó identifica a los clientes con menor uso y analiza sus casos para descubrir si la razón es producto de la incoherencia entre las funciones ofrecidas y el proceso de negocio del cliente, o por falta de capacitación del cliente en cuestión.

Uno de los objetivos secundarios del trabajo fue generar observaciones en cuanto

a la experiencia del uso de la plataforma de MS Azure ML. Para dar orden a estas observaciones, se usó como guía el modelo de calidad de uso del sistema de la ISO/IEC 25010. De la experiencia se destaca principalmente la facilidad para aprender a usar la herramienta y la cantidad de ejemplos que existen en la galería *Cortana Intelligence Gallery*.

Dado el tipo de datos y la cantidad de datos faltantes en la base de datos, no fue posible generar un ranking de atributos, por lo que se optó por generar un trabajo amplio que abarcaba BI, visualización de datos y predicciones con series de tiempo. A partir de esto, tampoco fue necesario hacer que el sistema se actualizara en tiempo real, y no habría sido posible hacerlo ya que la plataforma no soporta aprendizaje en línea, solo soporta aquellos que son entrenados una vez para luego ser consumidos.

Finalmente, como trabajo a futuro queda complementar el presente trabajo con la retroalimentación obtenida de los clientes, usar otros algoritmos de predicción de series de tiempo y esperar a que el sistema madure para reconsiderar la clasificación de variables.

Para e-restó se generó y aprovechó la oportunidad para aumentar el valor de su aplicación y se probó que esta aún tiene mucho potencial para ser utilizada por sus clientes.

5. Anexo

5.1. Figuras

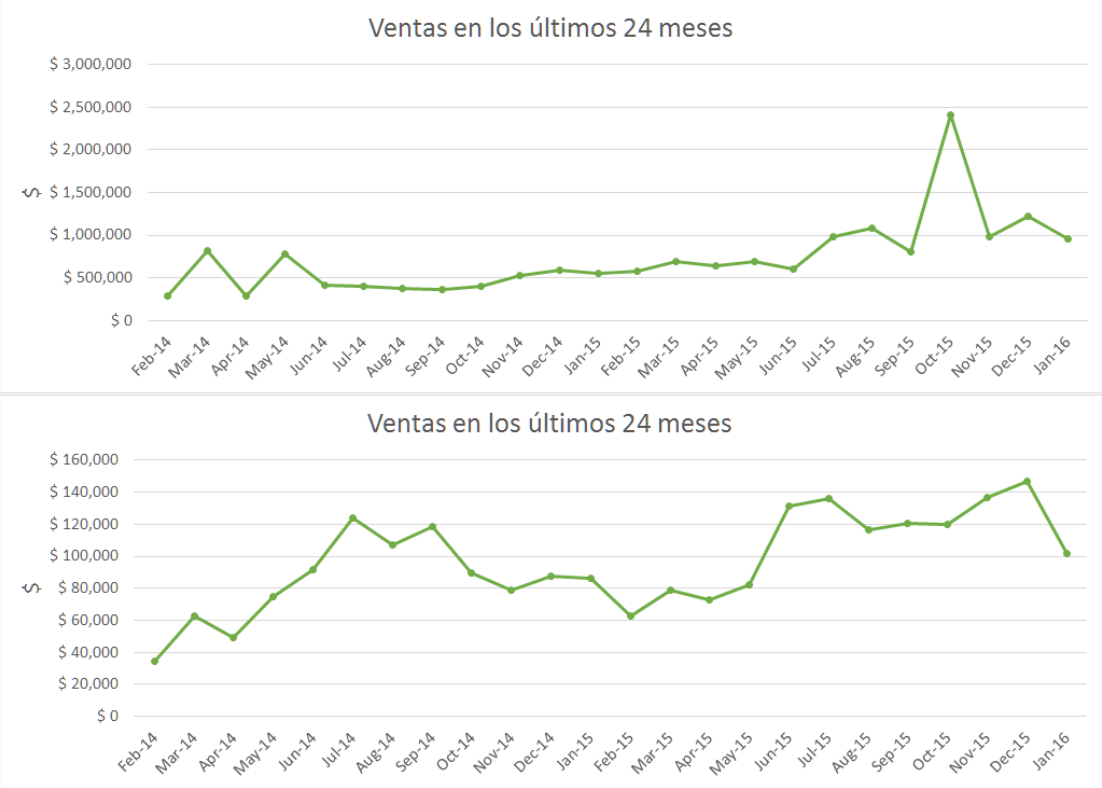


Figura 5.1: El primer gráfico de líneas muestra el historial de ventas del negocio A, y el segundo el historial de ventas del negocio E, desde Febrero del 2014 a Enero del 2016.

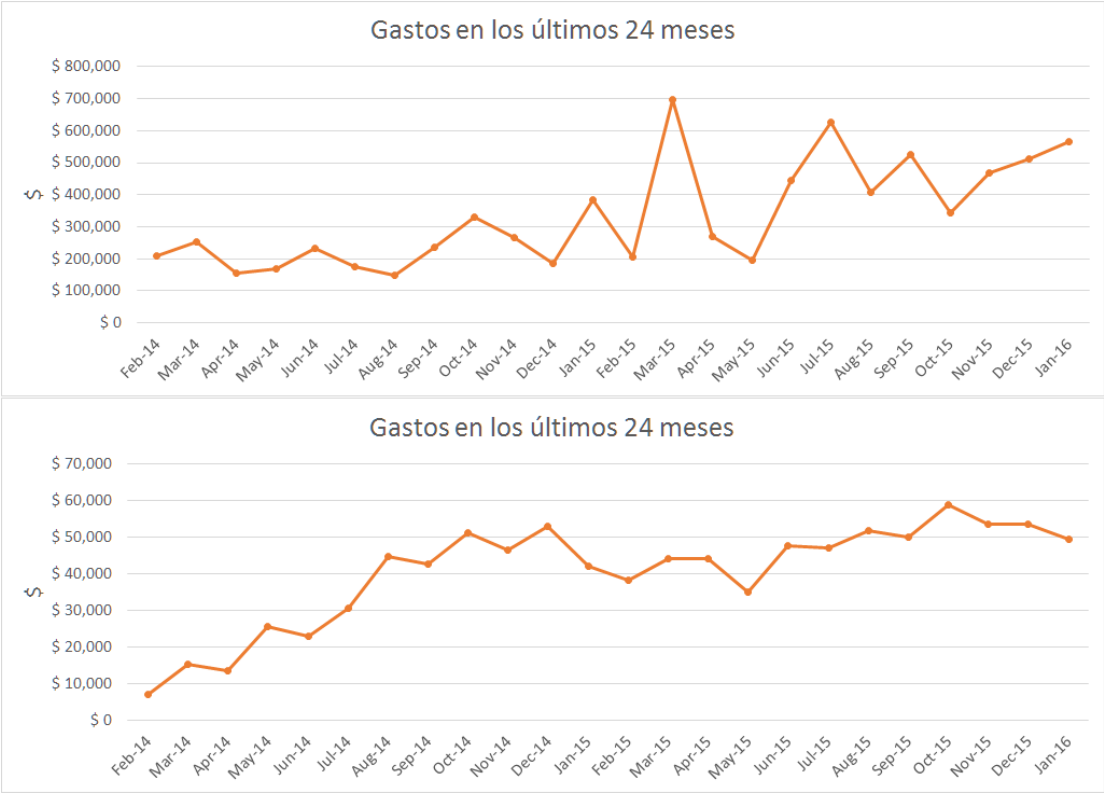


Figura 5.2: El primer gráfico de líneas muestra el historial de gastos del negocio A, y el segundo el historial de gastos del negocio E, desde Febrero del 2014 a Enero del 2016.

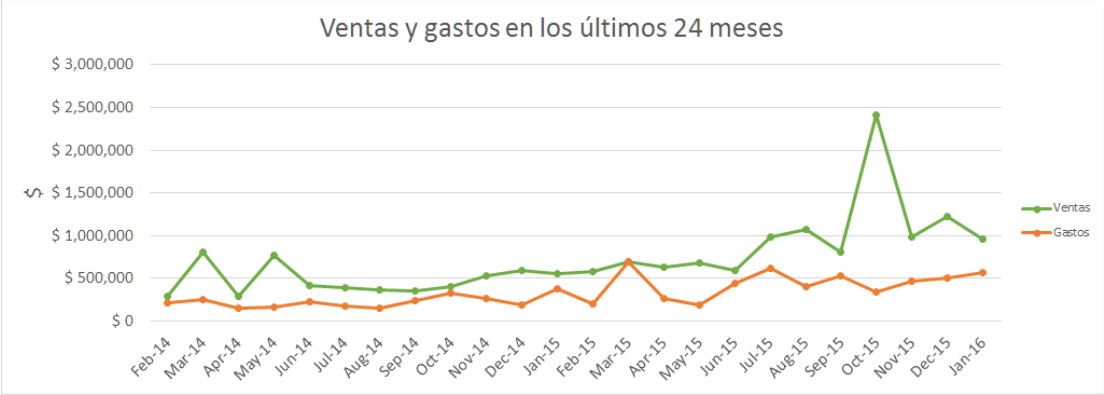


Figura 5.3: En verde se encuentra el historial de ventas y en naranjo el historial de gastos del negocio A, desde Febrero del 2014 a Enero del 2016.

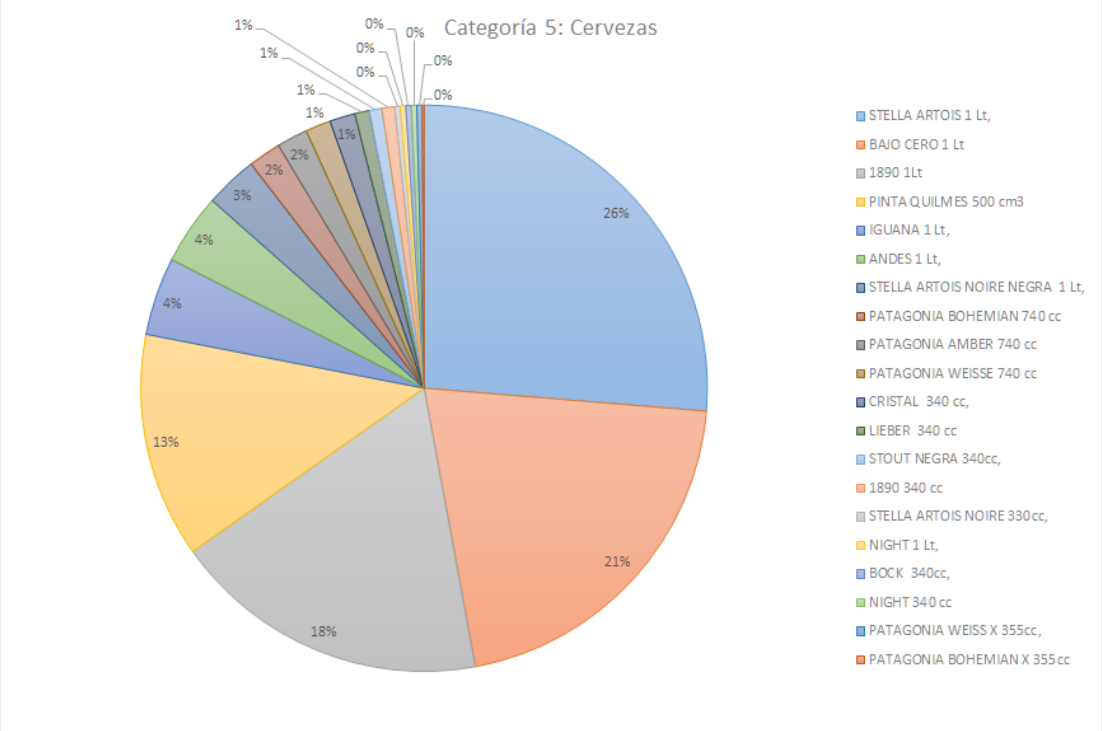


Figura 5.4: Cada sector representa el porcentaje de ventas de un producto sobre las ventas totales correspondiente a la categoría de cervezas del negocio A, en enero del 2015.

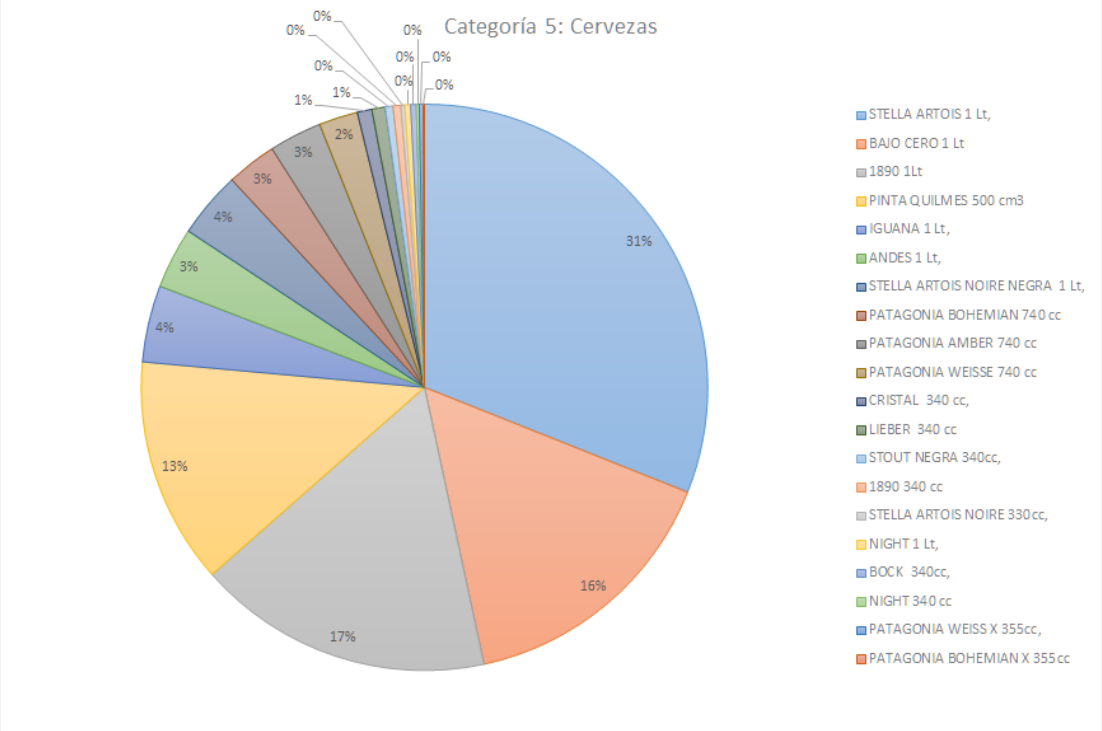


Figura 5.5: Cada sector representa el porcentaje del monto ventas de un producto sobre el monto total correspondiente a la categoría de cervezas del negocio A, en enero del 2015.

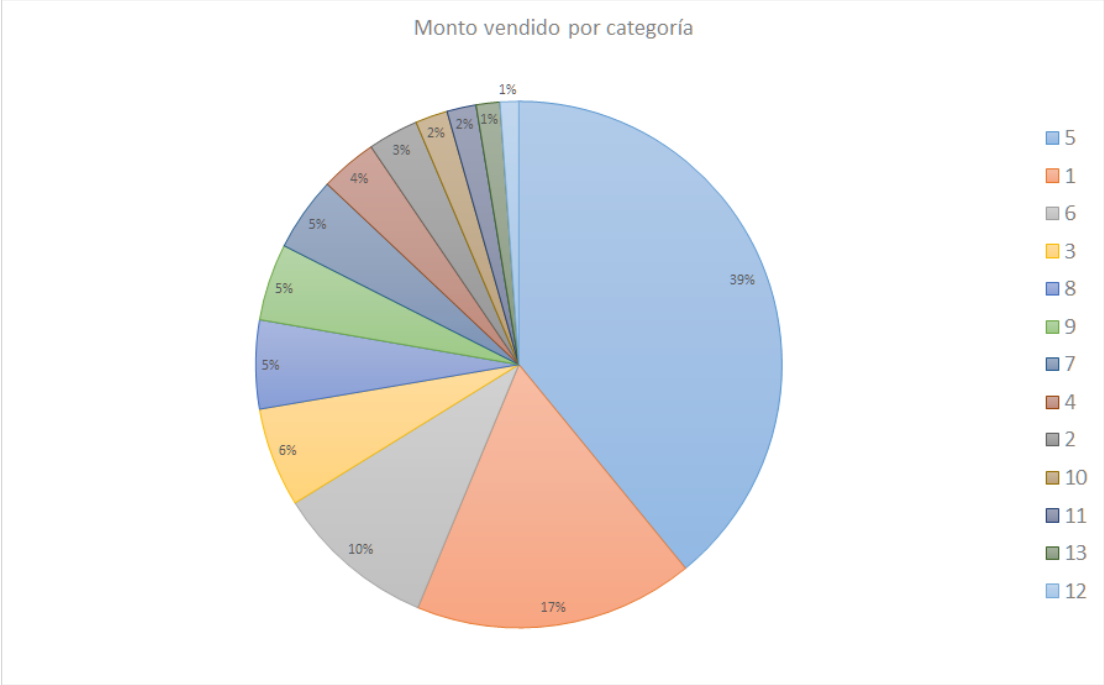


Figura 5.6: Cada sector representa el porcentaje del monto ventas de una categoría sobre el monto total correspondiente al negocio A, en enero del 2015.

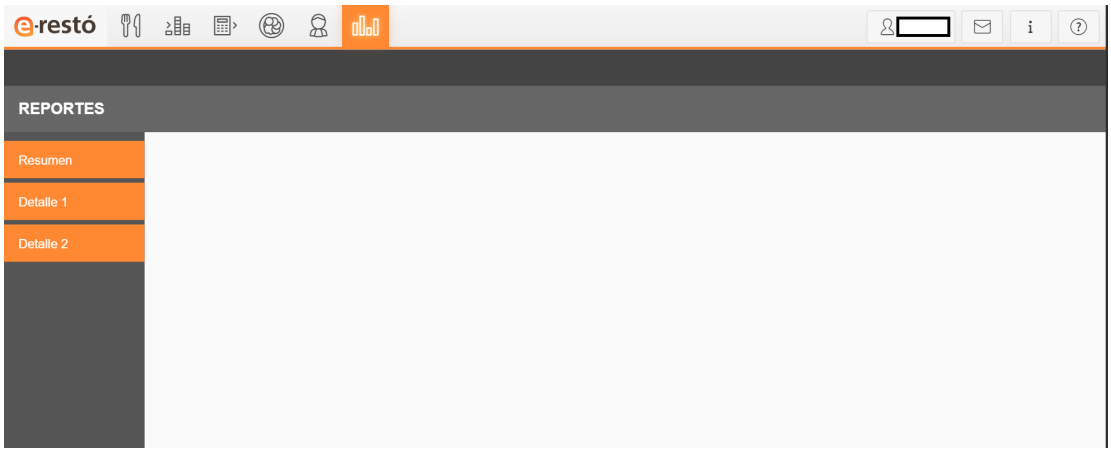


Figura 5.7: Diseño actual de la aplicación de e-restó para la sección de reportes.

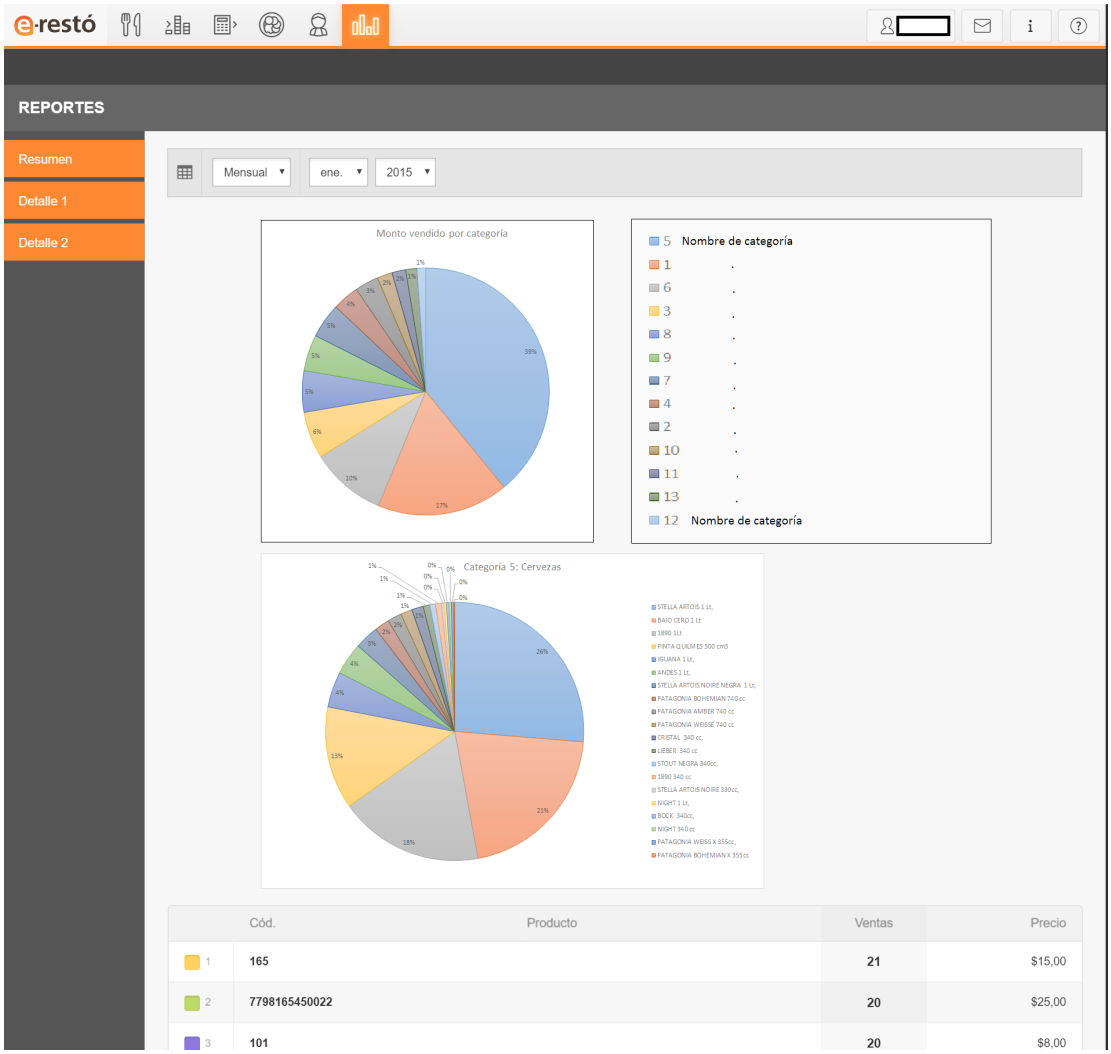
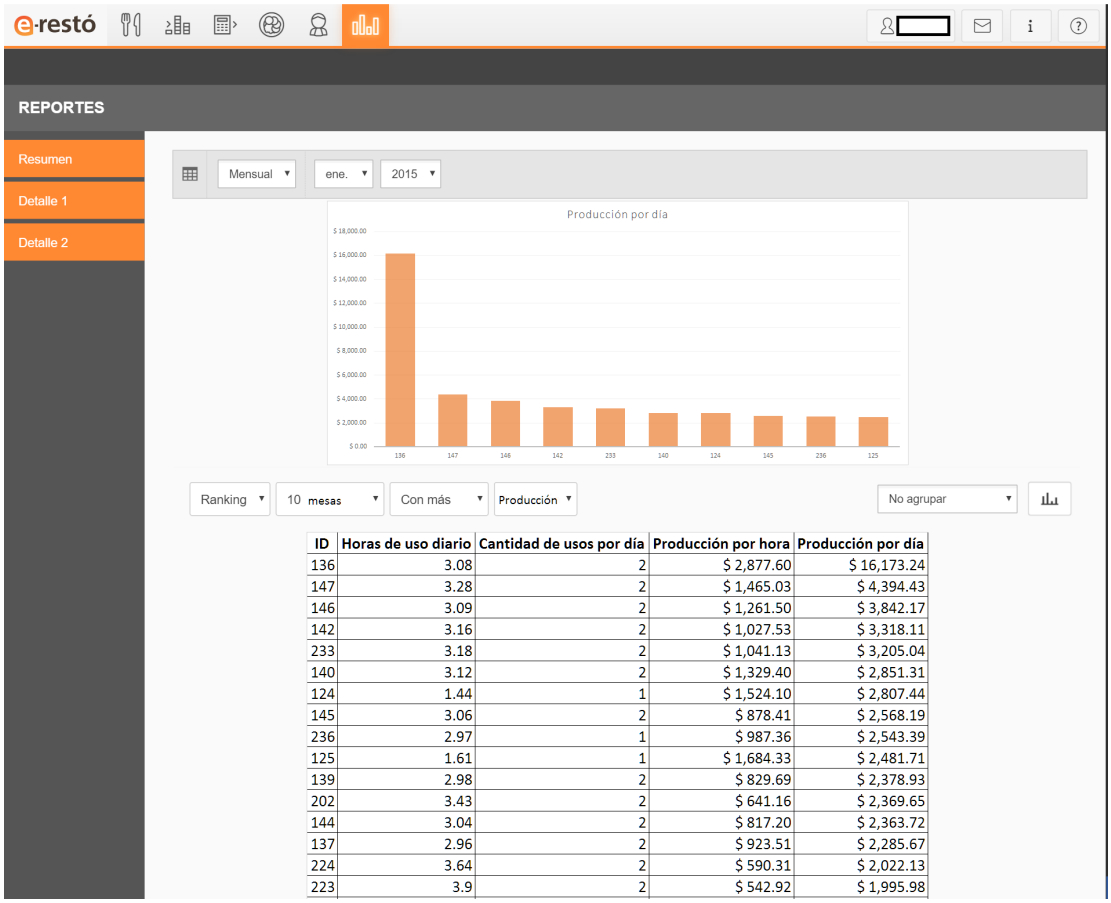


Figura 5.8: Propuesta de layout para la pantalla de detalles de productos.



ID	Horas de uso diario	Cantidad de usos por día	Producción por hora	Producción por día
136	3.08	2	$ 2,877.60	$ 16,173.24
147	3.28	2	$ 1,465.03	$ 4,394.43
146	3.09	2	$ 1,261.50	$ 3,842.17
142	3.16	2	$ 1,027.53	$ 3,318.11
233	3.18	2	$ 1,041.13	$ 3,205.04
140	3.12	2	$ 1,329.40	$ 2,851.31
124	1.44	1	$ 1,524.10	$ 2,807.44
145	3.06	2	$ 878.41	$ 2,568.19
236	2.97	1	$ 987.36	$ 2,543.39
125	1.61	1	$ 1,684.33	$ 2,481.71
139	2.98	2	$ 829.69	$ 2,378.93
202	3.43	2	$ 641.16	$ 2,369.65
144	3.04	2	$ 817.20	$ 2,363.72
137	2.96	2	$ 923.51	$ 2,285.67
224	3.64	2	$ 590.31	$ 2,022.13
223	3.9	2	$ 542.92	$ 1,995.98

Figura 5.9: Propuesta de layout para la pantalla de detalle de mesas.

5.2. Tablas

Jan-15	
ID	**Monto vendido**
150	$ 57,863
223	$ 52,446
151	$ 49,051
24	$ 48,858
226	$ 47,380
224	$ 45,680
146	$ 44,398
244	$ 41,790
21	$ 39,562
225	$ 39,339

Tabla 5.1: Top 10 mesas negocio A

ID	Horas de uso diario	Cantidad de usos por día	Producción por hora	Producción por día
21	3.67	2	$ 550.94	$ 1,799.98
22	3.69	2	$ 504.33	$ 1,734.72
23	3.74	2	$ 460.73	$ 1,664.17
24	3.76	2	$ 475.53	$ 1,696.95
25	3	2	$ 471.35	$ 1,287.12
26	3.17	2	$ 499.71	$ 1,390.35
27	3.23	2	$ 390.87	$ 1,196.53
...	...	...	...	...

Tabla 5.2: KPIs Mesas

ID Camarero	Cantidad promedio diario	Monto promedio diario
1	1.56	842.35
2	9.89	8525.30
3	10.64	7521.26
4	6.67	4235.69
5	3.47	1094.48
6	9.10	6505.08
8	3.19	1442.38
9	2.29	1041.57
...	...	...

Tabla 5.3: KPI Camareros

Referencias Bibliográficas

[1] Richard Miller Devens. *Cyclopædia of commercial and business anecdotes*, volume 1. New York, London, D. Appleton and company, 1865. Microfilm. Ann Arbor, Mich. : University Microfilms International, 1978. – 1 microfilm reel ; 35 mm. – (American culture series. Economics collection ; reel 12:15).

[2] Justin Heinze. A brief history of business intelligence. `https://www.betterbuys.com/bi/history-of-business-intelligence/`. (Consulta: 22 Marzo 2016).

[3] Solomon Negash and Paul Gray. *Business intelligence*. Springer, 2008.

[4] J. L. Devore. *Probabilidad y estadística para Ingenierías y ciencias*. Seventh edition, 2008.

[5] Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. *Handbook of data visualization*. Springer Science & Business Media, 2007.

[6] Alan J Izenman and Charles J Sommer. Philatelic mixtures and multimodal densities. *Journal of the American Statistical association*, 83(404):941–953, 1988.

[7] Robert McGill William S. Cleveland, Marylyn E. McGill. The shape parameter of a two-variable graph. *Journal of the American Statistical Association*, 83(402):289–300, 1988.

[8] Morven Leese Daniel Stahl Brian S. Everitt, Sabine Landau. *Cluster Analysis*. Wiley, 2011.

[9] Maureen Stone. Choosing colors for data visualization. *Business Intelligence Network*, 2, 2006.

[10] Insituto Nacional de Estadística. www.ine.es/explica/docs/pasos_tipos_graficos.pdf. `http://www.ine.es/explica/docs/pasos_tipos_graficos.pdf`. (Consulta: 21 Marzo 2016).

[11] Wilfried Grossmann and Stefanie Rinderle-Ma. *Fundamentals of Business Intelligence*. Springer, 2015.

[12] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

[13] Jan Grandell. *Time series analysis*. Lecture notes, KTH Stockholm, 1998.

[14] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer Science & Business Media, 2006.

[15] John Villavicencio. Introducción a series de tiempo. `http://www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?fileticket=4_BxecUaZmg%3D`. (Accessed on 05/28/2016).

[16] Rob Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(1):1–22, 2008.