Análisis de los registros efectivos de la empresa Wif ID empleando ciencia de datos

Roberth Castaño García.
Noviembre 26 de 2019.

Universidad Tecnológica de Pereira.
Facultad de Ingenierías.
Ingeniería de Sistemas y Computación.

Presentación del proyecto

El presente proyecto, describe el proceso de análisis de datos aplicado a los registros efectivos contenidos en la base de datos del establecimiento <<El Barista Slow Coffee>> perteneciente a la empresa Wif ID. El proceso involucra la obtención, validación, limpieza y el análisis de los usuarios y los dispositivos conectados a la red WiFi.

También se muestra el proceso de extracción, limpieza, transformación en instalación tal y como lo plantea la metodología conocida como E.T.L. Para transformar, visualizar y analizar los datos se muestra como se puede usar Pandas como biblioteca de análisis estadístico y iPython como infraestructura de computación interactiva. En última instancia, se muestran los resultados de los análisis obtenidos sobre los días de la semana y las horas en los que es más probable que ocurra un registro efectivo.

Planteamiento del problema

La satisfacción de los clientes es una de las principales preocupaciones de los gerentes y dueños de los establecimientos públicos. Si un administrador con visión conoce bien a sus clientes puede ejecutar procesos de marketing para lograr fidelización y adquisición de nuevos clientes. Entre las estrategias de fidelización empleadas por las empresas se encuentra el ofrecer servicios adicionales, tales como el acceso a Internet mediante redes inalámbricas Wifi.

La conexión a la red inalámbrica es un potencial activo el cual es asimilado por parte de los administradores asimilado comúnmente como un gasto mensual que, aunque sea necesario, no reporta beneficios directos a la empresa.

Wif ID ofrece un servicio de recolección de datos que aprovecha la red inalámbrica para capturar información de los gustos y preferencias de sus usuarios. Para lograr esto, Wif ID utiliza un dispositivo llamado WifBox que hace las veces de *hombre en la mitad* gestionando el acceso a la red inalámbrica al mismo tiempo que captura los datos de los usuarios y dispositivos conectados. Esto le permite a los dueños de los establecimientos conocer su público objetivo y de este modo, se pueden planificar estrategias de fidelización y marketing que sean atractivas para el sector específico en el que se encuentran sus clientes.

Esto es lo que lleva haciendo Wif ID desde el año 2017, un proceso de análisis del público objetivo de establecimientos tales como: bares, restaurantes y cafeterías de la ciudad de Pereira. Sin embargo, hasta la fecha en la que se comenzó este proyecto (20 de octubre del a), el análisis de la información se había limitado a la descripción de su público objetivo. Al hablar con los fundadores de Wif ID, se nota una clara intención de querer llegar a que el sistema genere análisis predictivos poderosos, sin embargo, para que un análisis predictivo sea confiable se necesita que el análisis descriptivo también lo sea; esto involucra que el proceso de análisis descriptivo se realice bajo una metodología que ayude a filtrar todos aquellos posibles valores erróneos que estén provocando ruido dentro del conjunto de datos.

¿Cómo podría Wif ID mejorar su proceso para garantizar un correcto análisis descriptivo?

Objetivos

Objetivo principal

Analizar los registros efectivos y otros factores relativos al comportamiento de los clientes de bares, restaurantes y cafeterías que son clientes de WiF ID.

Objetivos Específicos

1. Extraer la información desde las bases de datos de Wif ID.
2. Transformar las bases de datos a un formato de texto plano.
3. Cargar los archivos en Jupyter.
4. Validar, limpiar y transformar los datos con respecto al conjunto de reglas establecidas.
5. Hacer un análisis descriptivo de los días de la semana y los horas en las cuales se dan más los registros efectivos.
6. Hacer un análisis de los valores esperados para los días de la semana y las horas en los cuales se dan los registros efectivos.

Justificación

Dado que se busca garantizar que la capacidad de análisis descriptivo sobre la información que se encuentra en la base de datos del sistema Wif ID está libre de ruido; se presenta la propuesta de estudiar el proceso que usa la ciencia de datos para así, aplicarlo, en un primer análisis descriptivo. El análisis se desarrollaría sobre una pequeña porción de los datos con el fín de acercar la metodología propuesta a su posible integración al flujo de trabajo dentro de la plataforma.

Al implementar un proceso robusto de análisis de datos, los errores causados por: valores inexistentes, datos corruptos, valores atípicos, distribuciones asimétricas y relaciones débiles entre variables; serían atenuados considerablemente. Esto mejoraría, en un futuro en el que se implemente el análisis inferencial, la calidad de las interpretaciones sobre la base de que existiría un proceso sólido de extracción, limpieza, transformación y visualización de la información.

Alcance y delimitación

El proyecto comprende desde el proceso de extracción, transformación y carga de los datos y concluye en el análisis descriptivo de las horas y los días más probables en los que ocurran los registros efectivos en la base de datos del establecimiento conocido como <<El Barista Slow Coffee>>. Se agrega una visualización del *valor esperado* por trimestre y por año de las horas y los días en los que es más probable que ocurra un registro efectivo.

Marco de Referencia

Marco Conceptual

API: Interfaz de Programación de Aplicaciones por sus siglas en inglés. Es una capa de abstracción que expone las interfaces que un programador necesita para interactuar con una biblioteca de funciones.

Activo de información: Según la norma *ISO 27001*, los activos de información hacen referencia a cualquier componente lógico o abstracto que sustenta uno o más negocios de una unidad o área de negocios.

Base de Datos: Es un conjunto de datos pertenecientes a un mismo contexto y almacenado sistemáticamente para su uso posterior.

Business Intelligence: Se denomina Inteligencia de Negocios al conjunto de estrategias y tecnologías usadas por las empresas para el análisis de la información de un negocio. Proporciona vistas históricas, actuales y predictivas de las operaciones del negocio.

Character Varying Values: Un campo de caracteres variables es un tipo de dato que contiene cualquier tipo de dato: Números, caracteres, espacios o puntuación. Sirve para almacenar temporalmente datos de cualquier tipo o para almacenar cadenas de texto.

Conocimiento: Hechos o información adquiridos a través de la experiencia, la comprensión teórica, práctica de un asunto referente a la realidad.

Dataframe: Es una estructura de datos de dos dimensiones dispuesta en forma de tabla con ejes coordenados etiquetados como filas y columnas. Puede ser pensado como un diccionario que contiene un conjunto de elementos Serie enlazados a una llave columna.

Datos: Es una representación simbólica (numérica, alfabética, etc) de un atributo de una variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y

entidades.

Datos Estructurados: Es la información clasificada que se suele encontrar en las bases de datos. Suelen estar correctamente organizados, clasificados y listos para ser accedidos.

Datos no Estructurados: Son datos binarios que no contienen una estructura interna identificable. Entre ellas pueden estar: Correos electrónicos, archivos de texto enriquecido, imagenes, videos, publicaciones en medios sociales.

Depuración de Datos: Es el proceso de alterar los datos en un almacenamiento para asegurarse de que son exactos y correctos. También es conocido como *Limpieza de datos*.

Esperanza: En estadística, la esperanza de una variable aleatoria X es el número E[X] que formaliza la idea de valor medio de un fenómeno aleatorio.

Cuando la variable aleatoria es discreta, la esperanza es igual a la suma de la probabilidad de cada posible suceso aleatorio multiplicado por el valor de dicho proceso.

Por ejemplo, el valor esperado cuando se lanza un dado equilibrado de 6 caras es de 3,5. El cálculo se realiza de la siguiente manera:

$$E(X) = 1\cdot\tfrac{1}{6} + 2\cdot\tfrac{1}{6} + 3\cdot\tfrac{1}{6} + 4\cdot\tfrac{1}{6} + 5\cdot\tfrac{1}{6} + 6\cdot\tfrac{1}{6} = \tfrac{1+2+3+4+5+6}{6} = 3.5$$

Hombre en el medio: Es un tipo de arquitectura, que se basa regularmente en técnicas de hacking, en la que se ubica un dispositivo en medio de un emisor y un receptor para capturar información.

Información: Es un conjunto organizado de datos que constituyen un mensaje que cambia el estado del conocimiento del sujeto o del sistema que recibe dicho mensaje.

IPython: Es un ambiente de programación enfocado a la computación interactiva y a la exploración. Se apoya en un Shell que cuenta con las funcionalidades propias de un editor de código profesional y permite ejecutar porciones de código por separado.

MAC address: Es un identificador único asignado a un dispositivo para ser usado como una dirección única de red. Generalmente los fabricante queman las direcciones MAC físicamente en el dispositivo para asegurar la unicidad de la misma.

NaN: Es un identificador usado por la biblioteca Pandas que denota un valor faltante o que no está en un formato reconocible.

NaT: Es un identificador que denota la falta de un valor tipo Timestamp.

Portales cautivos: Es un programa o máquina de una red informática que vigila el tráfico HTTP y fuerza a los usuarios a pasar por una página especial si estos quieren navegar en Internet de forma normal. Es una página web, también conocida como una página de login que el usuario ve antes de acceder a la red (generalmente una red WiFi pública).

PosgreSQL: Es un potente sistema de base de datos relacional de objetos de código abierto que usa y amplía el lenguaje SQL combinado con muchas características que almacenan y escalan de manera segura las cargas de trabajo de datos más complicadas.

Proxy: Servidor, programa o dispositivo, que hace de intermediario, en las peticiones de recursos que realiza un cliente a otro servidor.

REST: Representational State Transfer. Transferencia de representación de de estado.

RV: Random Value: Su abreviatura(RV) da comienzo a la casilla “asunto”del encabezado de los mensajes de correo electrónico que ha sido reenviados.

Red Neuronal: Es un modelo computacional vagamente inspirado en el comportamiento, observado en su homólogo biológico, consiste en un conjunto de unidades llamadas neuronas artificiales, conectadas entre sí para transmitir señales.

Registro Efectivo: Representa un objeto único de datos, implícitamente estructurados en una tabla.

Regresión Lineal: Es un modelo matemático usado para aproximar la relación de dependencia entre una variable independiente, xi y un término aleatorio ε.

Repositorio: Es un espacio centralizado, donde se almacena, organiza, mantiene y difunde información digital, habitualmente archivos informáticos.

Router: Es un dispositivo hardware, que permite la interconexión de ordenadores en red. Es un dispositivo que opera en capa tres de nivel 3, así permite que varias redes u ordenadores se conecten entre sí y por ejemplo compartan una misma conexión de Internet.

Semántica de un conjunto de datos: Es un conjunto de actividades desarrolladas en el seno de World Wide web consortium, con tendencia a la creación de tecnologías para publicar datos legibles, por aplicaciones informáticas.

Shell: Intérprete de comandos.

Timestamp: El sistema Timestamp, es un número que se refiere a la cantidad de segundos transcurridos, desde la 00:00:00 UTC del 1 de enero de 1970 a Diciembre 1 de 2010

WiFi: Es una tecnología de comunicación inalámbrica que permite conectar a Internet equipos electrónicos como, computadoras portátiles, etc mediante el uso de radiofrecuencia o infrarrojos para la transmisión de la información.

Marco de Antecedentes

En el año 2017, en el paper titulado como: "Restaurant's Feedback Analysis System using Sentimental Analysis and Data Mining Techniques" escrito por: Atharva Patil, Nishita S. Upadhyay, Karan Bheda, Rupali Sawant se usaron técnicas de análisis de datos en redes sociales para hacer análisis de sentimientos. En este trabajo, se ha propuesto un método para el análisis sentimental de la retroalimentación de los estudiantes utilizando algoritmos de aprendizaje automático como Support Vector Machine, Multinomial Naïve Bayes Classifier y Random Forest. También se realiza un análisis comparativo entre estas técnicas de aprendizaje automático. Los resultados experimentales sugieren que el Multinomial Naïve Bayes Classifier es más preciso que otros métodos.

¿Por qué los bares y restaurantes comenzaron a ofrecer wifi gratuitamente a sus clientes?.

Wif ID es una empresa del sector de las tecnologías de la información especializada en Marketing Wi-Fi. Se encuentra ubicada en el sector CDV (Centro De Desarrollo Vecinal) Barrio San Luis de la ciudad de Pereira y es apoyada activamente por la incubadora de empresas de tecnología conocida como Parquesoft Risaralda, ubicada en el mismo sector.

Wif ID ofrece servicios de captura y análisis de datos, los cuáles, hasta la fecha del 21 de octubre del año 2018, recolectaban datos valiéndose de un dispositivo llamado WifBox. Un WifBox es, en esencia, un proxy basado en tecnología Raspberry Pi 2 ordenador de

placa reducida, cuya función es establecer un enlace entre el Router que ofrece la conexión a Internet y los usuarios que desean acceder a dicha conexión.

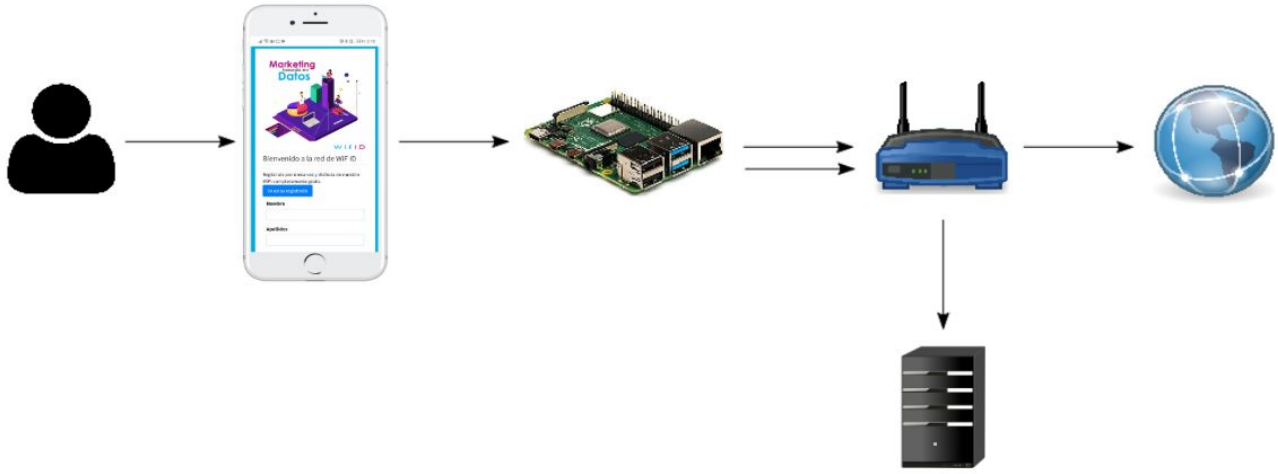

Figura 1. Despliegue del sistema WifBox. Adaptado de Wikimedia Commons.

Para otorgar acceso a la red, los WifBox despliegan un portal cautivo, con el fin de recolectar los datos, en los dispositivos móviles de sus usuarios. El portal permite el acceso gratuito a Internet a los usuarios que se registren con sus datos personales y que además, respondan alguna de las preguntas programadas por el dueño del establecimiento. Las respuestas se almacenan en el centro de datos de Wif ID para un posterior análisis que se entrega por medio de un informe, con una periodicidad mensual.

Wif ID ha tenido éxito, entregando informes nutridos, con la información previamente generada por la plataforma; sin embargo parte del análisis, hasta la fecha de inicio de este proyecto, se hacía de forma manual y el análisis se veía limitado por sus propios estadísticos generados automáticamente.

Marco teórico

Ciencia de Datos

La Ciencia de Datos es un campo multidisciplinario que usa procesos de estadística y aprendizaje de máquinas para obtener un mejor entendimiento de un conjunto datos y así modelar su comportamiento. Dichos datos pueden ser estructurados o no-estructurados.

Un científico de datos se define como un profesional en computación, matemática, estadística y experticia en el tema a ser investigado. En el siguiente diagrama de Venn, Drew Conway define las tres habilidades que debe poseer un científico de datos.

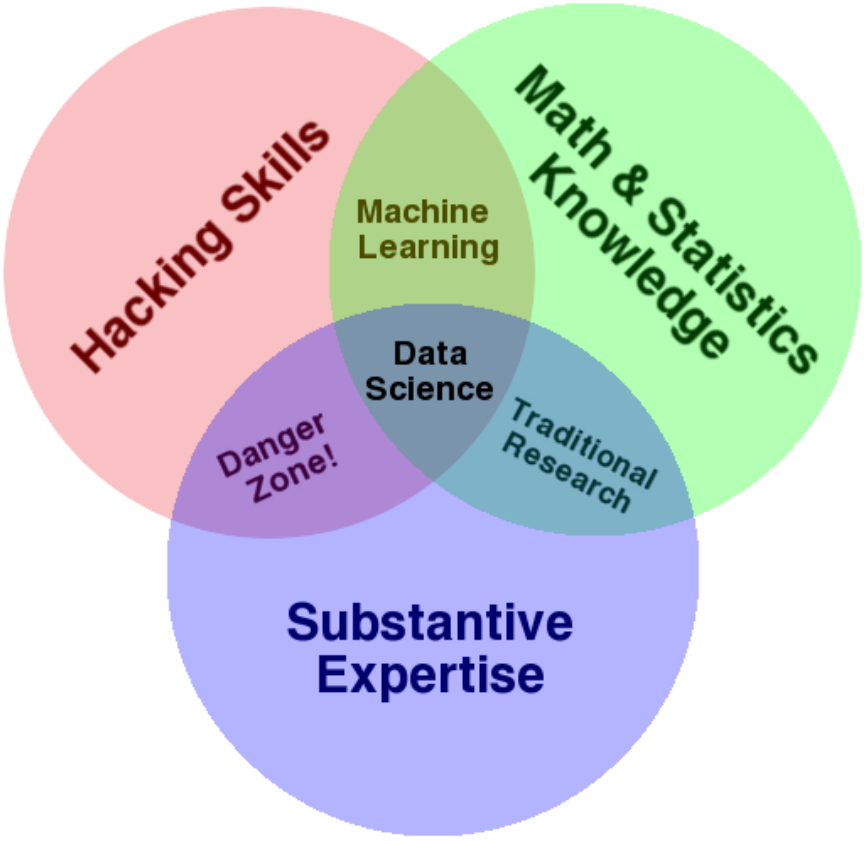

Figura 3. Conway, D. (2010). *The Data Science Venn Diagram*. Fuente: https://images.squarespace-cdn.com/content/v1/5150aec6e4b0e340ec52710a/1364352051365-HZAS3CLBF7ABLE3F5OBY/ke17ZwdGBToddI8pDm48kB2M2-8_3EzuSSXvzQBRsa1Zw-zPPgdn4jUwVcJE1ZvWQUxwkmyExglNqGp0IvTJZUJFbgE-7XRK3dMEBRBhUpxPe_8B-x4gq2tfVez1FwLYYZXud0o-3jV-FAs7tmkMHY-a7GzQZKbHRGZboWC-fOc/Data_Science_VD.png?format=750w [Recuperado el 26 Nov. 2019].

El método que emplea la ciencia de datos es el siguiente.

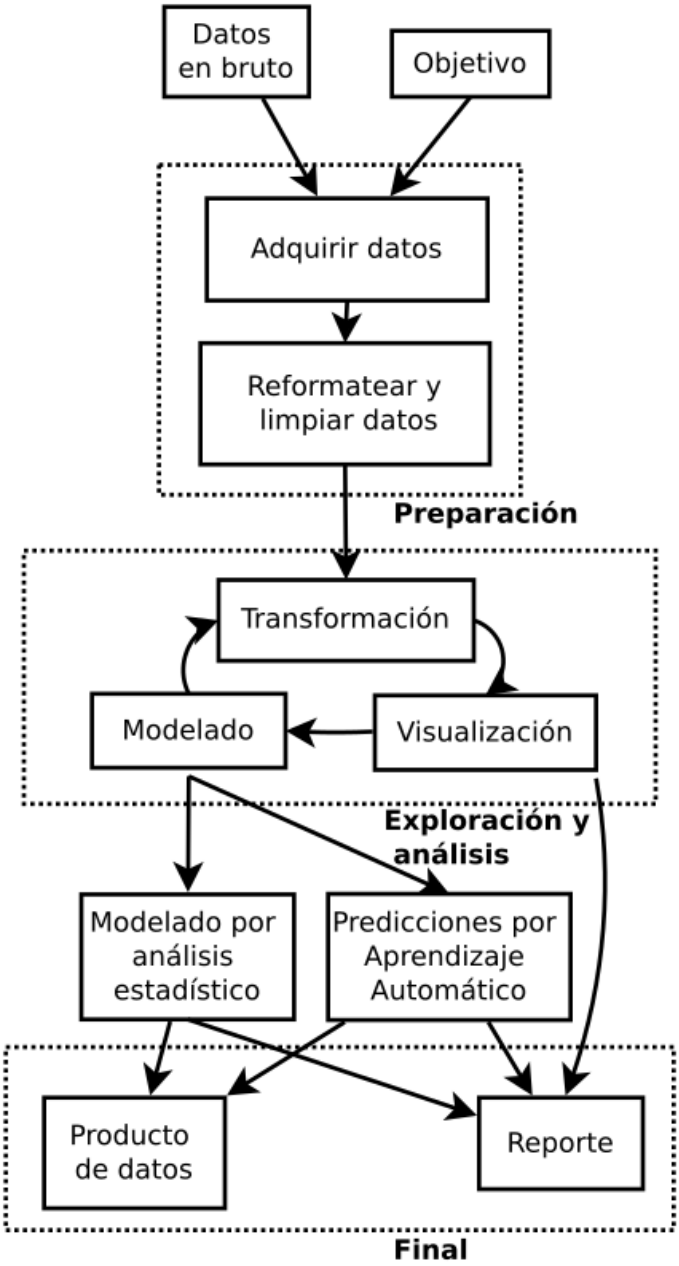

Figura 4. Flujo de trabajo en Ciencia de Datos. Recurso propio.

Objetivo: El objetivo se declara por medio de una pregunta inicial; la cuál es el problema que el analista intenta resolver.

Datos en bruto (Raw data): Es el conjunto de datos en su formato y medio de almacenamiento original.

Fase de preparación: En esta fase el científico de datos deberá primero, adquirir los datos y luego reformatearlos de forma que puedan ser sometibles a computación.

Adquirir Datos: Este es el proceso de adquirir los datos que van a ser analizados. Los

datos pueden venir desde distintas fuentes como:

- Datos que provienen de repositorios públicos en línea.
- Datos que pueden ser leídos en tiempo real por medio de una API.
- Datos que pueden ser adquiridos en tiempo real, por medio de un dispositivo físico.
- Datos que son ingresados manualmente por un humano, a una hoja de cálculo.

Reformatear y Limpiar Datos: En esta fase se deben convertir los datos en bruto a un formato que pueda ser leído por el ecosistema de aplicaciones usado para el análisis de datos. El formato más empleado en los ecosistemas de Ciencia de Datos es el CSV (v*alores separados por comas* por sus siglas en inglés) ya que es una formato fácilmente legible. Así mismo, se busca modificar los datos de una manera en que la semántica del conjunto de datos, encaje con la manera en que ha sido almacenada. Los datos estarán organizados, cuándo cada columna represente correctamente a una variable y cada fila represente a una observación. Al organizar los datos la estructura permitirá enfocarse más en hacer preguntas adecuadas en lugar de desperdiciar tiempo en intentar manipular las tablas.

Exploración y análisis: En esta fase se realiza un proceso denominado Análisis Exploratorio de Datos (EDA por sus siglas en inglés). Este proceso busca hacer inferencias sobre una población por medio del análisis gráfico y estadístico de los datos. Esto permitirá luego poder modelar su comportamiento bajo una regresión lineal o una predicción arrojada por una red neuronal. El Análisis Exploratorio de Datos se divide en tres sub-procesos: *Transformación*, *Visualización* y *Modelado*.

Transformación: Es el proceso de manipular la estructura de los datos, con respecto a la pregunta formulada, de forma que se obtenga un conjunto de tablas y campos adecuados para ser explorados. En esta fase también se buscan anomalías en los datos, se limpian registros vacíos y se cambian valores registrados de forma incorrecta por valores que tengan sentido según el dominio del problema.

Visualización: Es el proceso de observar gráficamente el comportamiento de los datos. Generalmente se hace graficando un conjunto cruzado de variables y dependiendo de la cantidad y el tipo de variables, se elige el tipo de gráfica a usar.

Modelado: Es el proceso sobre el cual se construye un modelo matemático que logre predecir el comportamiento de un conjunto de datos. El modelado se puede realizar por medio de métodos estadísticos, ya sean descriptivos o inferenciales, o por sometimiento de los datos al entrenamiento de una red neuronal.

La fase de exploración es iterativa; por lo que un hallazgo puede dar lugar a reformular

la pregunta inicial. Lo importante de esta fase es entrar a familiarizarse poco a poco con los datos, notar como van surgiendo tendencias, ver si dichas tendencias son estadísticamente consistentes y, por último, hallar un modelo que me permita predecir nuevos valores.

Modelado por análisis estadístico: Sucede cuando, en la fase de modelado, se utiliza un método de modelado estadístico como por ejemplo: una regresión lineal.

Predicciones por aprendizaje automático: Sucede cuando, en la fase de modelado, se usa un conjunto de datos para entrenar una red neuronal y esta retorna un comportamiento predictivo que luego se puede probar con otro conjunto de datos del mismo tipo.

Fase final: En esta fase se pueden dar dos tipos de resultados: El Producto de datos o un Reporte final.

Producto de datos: Un producto de datos es un activo de información envuelto en analítica que proporciona valor a sus clientes. Es decir, es un producto que facilita un objetivo final mediante el uso de datos. Un producto de datos se compone tanto de datos como de procesos analíticos.

Reporte: Un reporte es una presentación organizada de los resultados a un público objetivo. Dicha presentación se apoya en gráficos y tablas para así dar una descripción detallada del proceso con el cual se llegó a la conclusión.

Marketing WiFi

El márketing WiFi consiste en la creación de una zona de cobertura basada en tecnología inalámbrica. El acceso a dicha red WiFi requiere que el usuario ingrese su información personal, ya sea de forma manual, por medio de un cuestionario o, de manera indirecta y automatizada, a través de una solicitud de acceso a sus redes sociales o por medio del acceso a la información proveniente de su dispositivo.

El objetivo del Márketing WiFi es encontrar el perfil de los usuarios que acuden a un establecimiento, de esta forma, quien diseña una estrategia de márketing, puede apoyarse en las métricas obtenidas y así obtener segmentaciones con respecto a variables demográficas o conocer mejor el comportamiento que los usuarios de su red están teniendo dentro del negocio.

ETL

E.T.L. (Extraer, transformar y cargar por sus siglas en inglés) es el proceso que se usa para recolectar datos que vienen de múltiples fuentes, transformar dichos datos a un formato estándar para que la información pueda ser analizada y almacenar los datos obtenidos en otro sistema.

ETL se desarrolla en tres pasos.

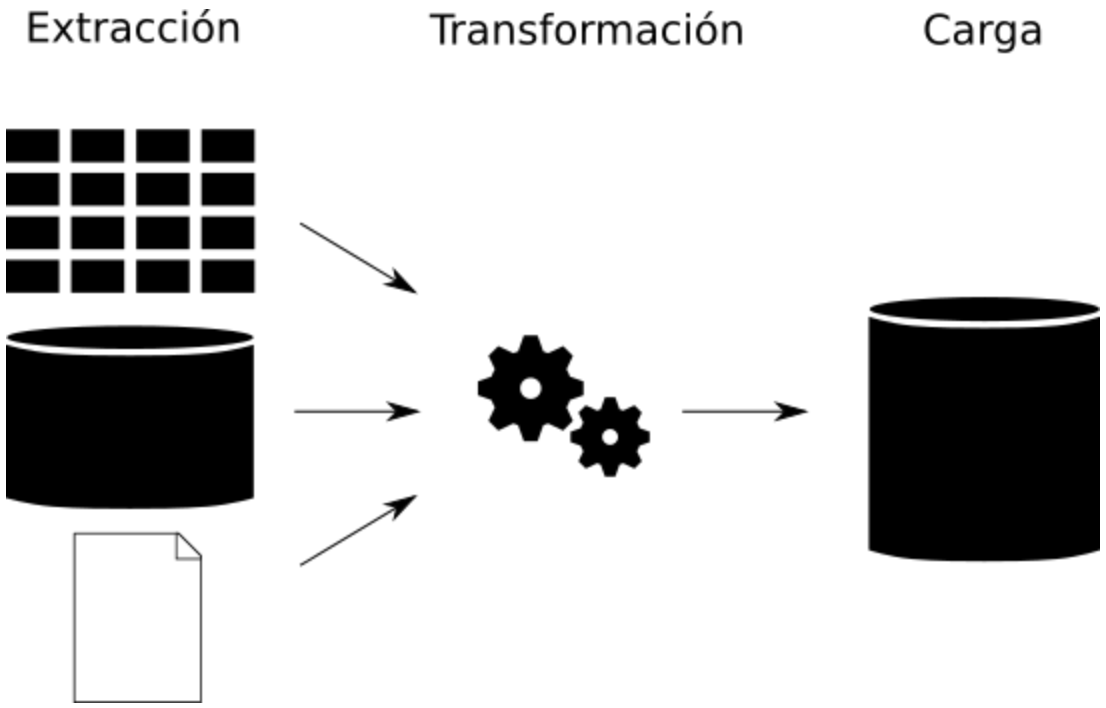

Figura 5. Proceso de la ETL. Recurso propio.

Extracción: En este primer paso se extraen datos desde distintas fuentes. Cada fuente puede almacena sus datos usando formatos que suelen ser incompatibles con los formatos de otras fuentes. Generalmente, entre los medios desde los que se extraen los datos se encuentran: las bases de datos relacionales, los archivos de texto plano, las hojas de cálculo; sin embargo, los datos pueden provenir de otras fuentes como fuentes de información en vivo o servicios de API's REST

Transformación: Una vez se hayan extraído los datos y se hayan exportados a un formato adecuado para que puedan ser manipulados y analizados, lo siguiente será limpiar los datos. La limpieza es un proceso que, en la teoría, es comúnmente ignorado sin embargo, es un proceso importante en sí mismo ya que ayuda a estandarizar la

información bajo unas reglas de unificación de datos que aseguren la calidad de los dichos datos. Algunas operaciones básicas de limpieza son:

- Unificar los identificadores. (ejemplo: identificadores de género como 'Mujer'/'Femenino'/'F' se convierten todas en 'Femenino' y 'Hombre'/'Masculino'/'H' se convierten en 'Masculino')
- Convertir los valores nulos en un formato estandarizado como 'NA'.
- Convertir los números telefónicos a un formato estándar.
- Estandarizar direcciones a un estilo estándar de nomenclatura.
- Confrontar y validar direcciones unas contra otras para hallar errores.

Lo que sigue entonces es transformar dichos datos resultantes para obtener unos resultados esperados empleando operaciones de filtrado, ordenamiento, agregación, combinación, generación de datos calculados, entre otros.

Carga: El último paso es almacenar los datos resultantes en la ubicación destino, generalmente es otra base de datos para uso en producción como las que existen en las bodegas de datos.

Cálculo de la Esperanza

El valor esperado o la media de una variable aleatoria continua **x** está definido por la integral:

$$E\{x\} = \int_{-\infty}^{\infty} xf(x)dx \quad \textbf{(1)}$$

Suponga que x toma los valores xi con probabilidad pi en este caso.

Si:

$$f(x) = \frac{dF(x)}{dx}$$

Siendo $F(x)$ la función de densidad o la función de frecuencia de RV x.

Si RV x es de tipo discreto tomando los valores x_i con probabilidades p_i. Entonces.

$$f(x) = \sum_i p_i \delta(x - x_i) \qquad p_i = P\{x = x_i\} \quad \textbf{(2)}$$

Para tipos de valores discretos la integral **(1)** puede escribirse como una sumatoria.

Insertando **(2)** en la integral **(1) y** usando la identidad

$$\int_{-\infty}^{\infty} x\delta(x - x_i)dx = x_i$$

Obtenemos:

$$E\{x\} = \sum_i p_i x_i \qquad p_i = P\{x = x_i\}$$

Metodología

Se usó un computador portatil Lenovo G40-70 con 8 Giga Bytes de RAM ejecutando Arch Linux con un escritorio i3.

Se aplicó el Proceso E.T.L. para la exportación, transformación y carga de los datos:

Proceso de Exportación:

Se copiaron los archivos de respaldo de los clientes desde servidor de WIF-ID al computador a través ssh. Las bases de datos tomadas fueron las copias de respaldo del día domingo 21 de octubre del 2018 de los siguientes establecimientos:

- El Barista Slow Coffee

Proceso de Transformación:

Las bases de datos fueron descomprimidas desde su formato original .gz y fueron cargadas a PostgreSQL para luego exportadas a un archivo de texto plano .csv usando la bandera para arropar con comillas todos los valores Timestamp, Macaddr y Character Varying.

Proceso de Carga:

Usando Jupyter Notebook, se cargaron las tablas *dispositivos* y *users* correspondientes a cada establecimiento en Jupyter.

Segunda iteración del proceso de Transformación:

Se hizo una limpieza de las tablas *dispositivos* y *users* con respecto a las siguientes reglas:

1. Las fechas deben estar almacenadas en formato datetime.
2. Los *id_user* deben estar almacenados en el formato de cadena de caracteres.
3. Los *macaddr* deben estar almacenados en el formato de cadena de caracteres.
4. Las fechas de modificación deben ser mayores o iguales a las fechas de creación.

Se hizo la limpieza de la tabla *dispositivos*:

- Se validaron las fechas created y modified para probar que ninguno de los datos

sea un *NaT*.
- Se renombraron las columnas *created* y *modified* a *d_created* y *d_modified* respectivamente para evitar solapamiento de datos a la hora de combinar las tablas.
- Se validó *id_user* de manera que ningún dato sea *NaN*.
- Se transformó el id_user en un entero para eliminar el punto flotante del formato y luego convertirlo en una cadena para que no se den operaciones de sumas.
- Se transformó macaddr en una cadena para posterior manipulación.
- Se eliminaron las columnas: *id*, *asignado*, *ultimo_visto*, *fabricante*, *ssid*, *primer_lugar_visto*, *routermac*, *potencia*; ya que son datos que no sirven al estudio de los registros efectivos.
- Se validó si las fechas en *d_modified* son posteriores a las de *d_created*, puesto que las modificaciones se deberían hacer luego de la creación de los registros.

Se hizo la limpieza de la tabla *users*:

- Se validaron las fechas created y modified para probar que ninguno de los datos sea un NaT.
- Se renombraron las columnas created y modified a u_created y u_modified respectivamente para evitar solapamiento de datos a la hora de combinar las tablas.
- Se transformó el id_user en un entero para eliminar el punto flotante del formato y luego convertirlo en una cadena para que no se den operaciones de sumas.
- Se transformó macaddr en una cadena para posterior manipulación.
- Se transformó y validó la columna fecha_nacimiento de manera que se eliminaran las fechas no existentes.
- Se eliminaron las columnas: nombre, apellidos, presente, password, username, control_ausencia, estado, ultimo_visto, ultimo_ingreso, celular, cedula, coins; ya que son datos que no sirven al estudio de los registros efectivos.
- Se filtraron los usuarios por *rol*, ya que la columna *reg* determina en esta tabla que un usuario se encuentre efectivamente registrado.
- Se validó el que los datos de la columna *email* no fueran nulos, ya que el que exista un correo es otro indicador de que hubo un registro exitoso.
- Se convirtieron los caracteres de la columna *email* a minúscula, esta era cndición necesaria para comprobar luego si existían correos duplicados.
- Se eliminaron los registros que contenían correos duplicados y se dejaron solamente los registros que tenían la fecha *u_created* más antiguo.

Transformación:

En este paso se combinaron las tablas *dispositivos* y *users* por medio de las columnas

id_user y *macaddr* para crear la tabla *reg_efectivos*.

- Se ordenaron los registros efectivos por fecha de creación de usuarios *u_created*, ya que son los usuarios los que importan a la hora de hacer el análisis de los registros efectivos.
- Se validó que la fecha de creación de los usuarios, *u_created* debe ser mayor a la fecha de creación de los dispositivos *d_created*, pues los usuarios deben haberse creado luego de la captura de su dispositivo.
- Se creó la columna id con los identificadores de la tabla para reg_efectivos para usarlos en el cálculo de valores hallados por groupby, esto ayuda a reubicar dichos valores calculados de nuevo en los índices adecuados.
- Se creó la columna u_created_dayofweek que contiene el día de la semana en el que ocurre un registro efectivo.
- Se creó otra columna llamada u_created_dayofweek_expected_value calculando la esperanza del día de la semana en el que es más probable que ocurra un registro efectivo.
- Se creó la columna u_created_hour que contiene la hora en el que ocurre un registro efectivo.
- Se creó otra columna llamada u_created_hour_expected_value calculando la esperanza de la hora en la que es más probable que ocurra un registro efectivo.
- Se creó una columna u_created_date con la fecha la cual resulta útil a la hora de mostrar información en las gráficas.
- Se creó una columna u_created_dayname con el día de la semana en formato de cadena de caracteres la cual resulta útil a la hora de mostrar el nombre del día de una manera más legible.
- Se creó una columna u_created_quarter que obtiene el trimestre actual el cuál es útil para graficar el comportamiento de los días de la semana y las horas separadas en una misma gráfica
- Se creó una columna ano_fecha_nacimiento que solo contiene el componente del año de la fecha de nacimiento que será útil.

Eliminación de los valores atípicos en u_created_hour y u_created_dayofweek:

- Apoyados en el gráfico de cajas y bigotes se identifican valores atípicos.
- Se eliminaron los valores típicos de u_created_hour y u_created_dayofweek usando el método del rango intercuartílico.
- Se creó la columna u_created_dayofquarter para almacenar el día relativo al trimestre.
- Se calculó la esperanza por trimestre para las horas y los días de la semana, los valores fueron almacenados en u_created_hour_expected_value_by_quarter y u_created_dayofweek_expected_value_by_quarter respectivamente.

Visualización:

Se construyeron las siguientes gráficas:

- Un histograma para los días de la semana en los que ocurren los registros efectivos.
- Un diagrama de línea para valor esperado de los días de la semana en los que ocurren registros efectivos.
- Un diagrama de líneas para el valor esperado por trimestre de los días de la semana en los que ocurren registros efectivos.
- Un histograma para las horas del día en las que ocurren registros efectivos.
- Un diagrama de línea para valor esperado de las horas a lo largo del tiempo en las que ocurren registros efectivos.
- Un diagrama de líneas para el valor esperado por trimestre de las horas en las que ocurren registros efectivos.
- Un histograma para el año de nacimiento de las personas que se registran

Análisis

A continuación, se muestra el análisis desarrollado sobre la base de datos del establecimiento <<El Barista Slow Coffee>> cargado en el dataframe *reg_efectivos*.

En una primera visualización de las distribuciones de las columnas ‘u_created_hour’ y ‘u_created_dayofweek’:

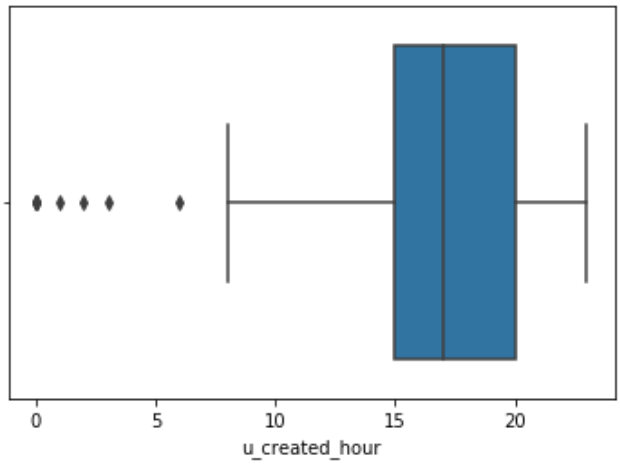

Figura 1. Distribución de horas. Recurso propio.

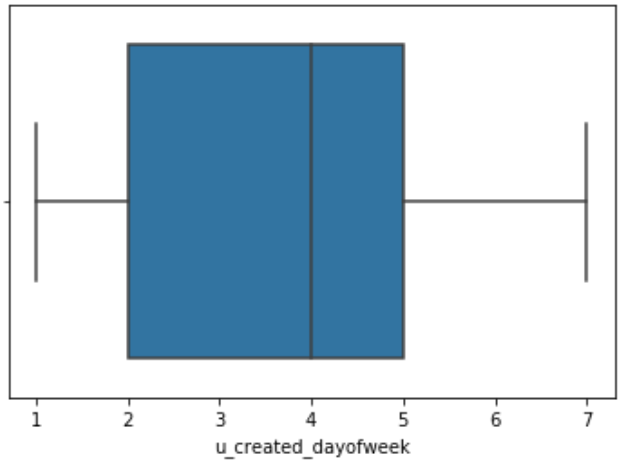

Figura 2. Distribución de los días de la semana. Recurso Propio.

Se evidencian datos atípicos en el diagrama de Distribución de horas antes de ser pasados por algún proceso de limpieza.

A continuación se muestra el diagrama de cajas y bigotes una vez se ejecuta el proceso de eliminación de datos atípicos utilizando el método del rango intercuartílico IQR.

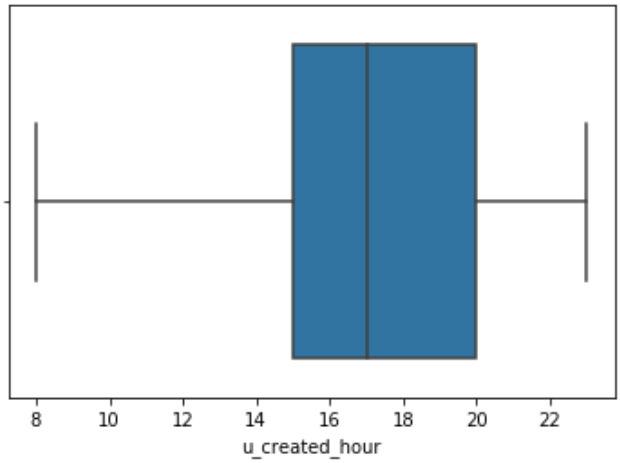

Figura 3. Distribución de horas con datos atípicos eliminados. Recurso propio.

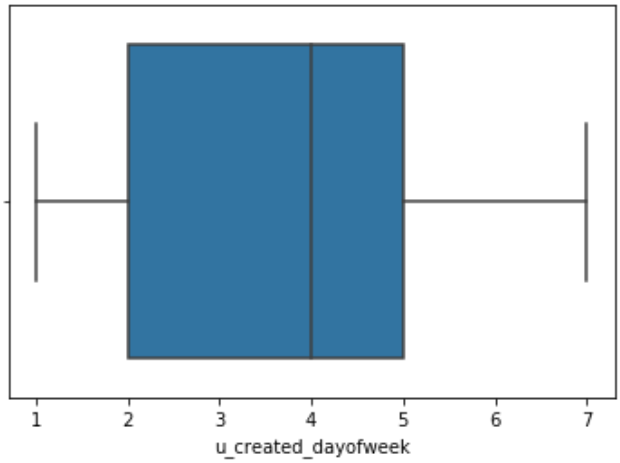

Figura 4. Distribución de los días de la semana datos atípicos eliminados. Recurso propio.

Ahora se puede apreciar que los valores atípicos en el diagrama de Distribución de horas ya no están.

Análisis hecho a los días de la semana:

A continuación se muestra el histograma de frecuencias de los registros efectivos por día de la semana.

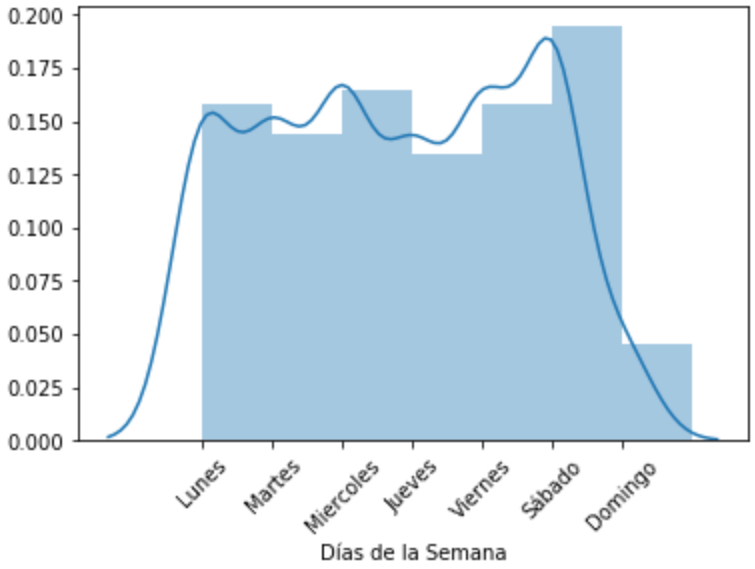

Figura 5. Histograma de frecuencias de los registros efectivos por días de la semana. Recurso Propio.

El tamaño de la muestra es

N = 1182

Visualmente se evidencia que la distribución no es normal y para este caso la media aritmética no ofrecerá una buena aproximación al valor central. Se deben analizar las medidas de tendencia central y compararlas entre sí para obtener la orientación de la asimetría detectada visualmente.

- Moda = 6
- Mediana = 4
- μ = 3.757191

Como μ es menor que la Mediana y como la Mediana es menor que la Moda; se dice que la distribución tiene asimetría negativa. Al revisar de nuevo el gráfico de cajas y bigotes se confirma que la mayor cantidad de datos están concentrados entre los días martes y viernes.

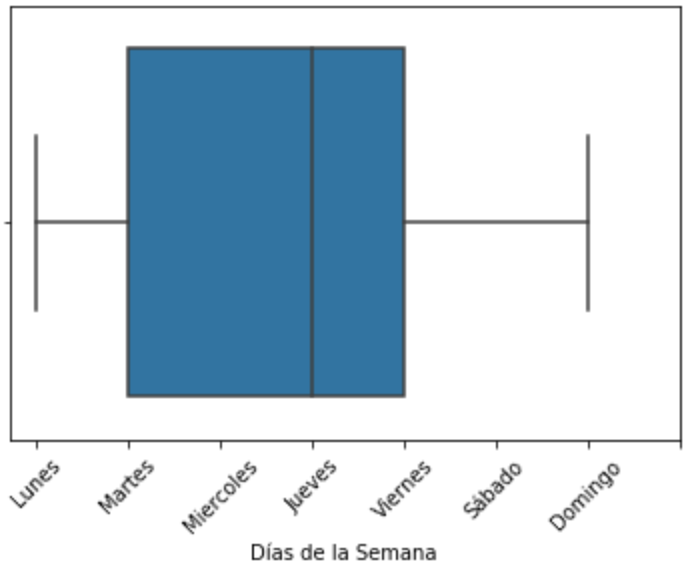

Figura 6. Diagrama de cajas y bigotes de los registros efectivos por días de la semana. Recurso propio.

Esto significa que, aunque la moda indique que el sábado sea el día con mayor cantidad de registros efectivos, existe una mayor cantidad de ocurrencias entre los días martes y viernes. Al calcular la esperanza de los días de la semana en los que ocurren los registros efectivos se puede percibir una clara tendencia a ubicarse entre los días miércoles y jueves.

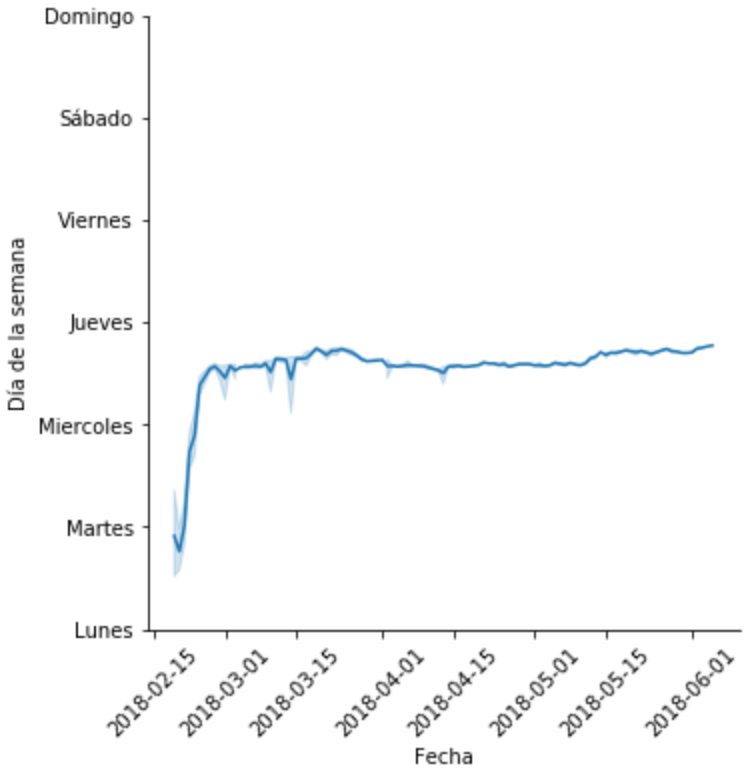

Figura 7. Valor esperado de los registros efectivos para los días de la semana. Recurso propio.

Al dividir el tiempo total por trimestres se logra percibir una tendencia que se hace fuerte y que tiende a ubicar el día de la semana entre el miércoles y el jueves.

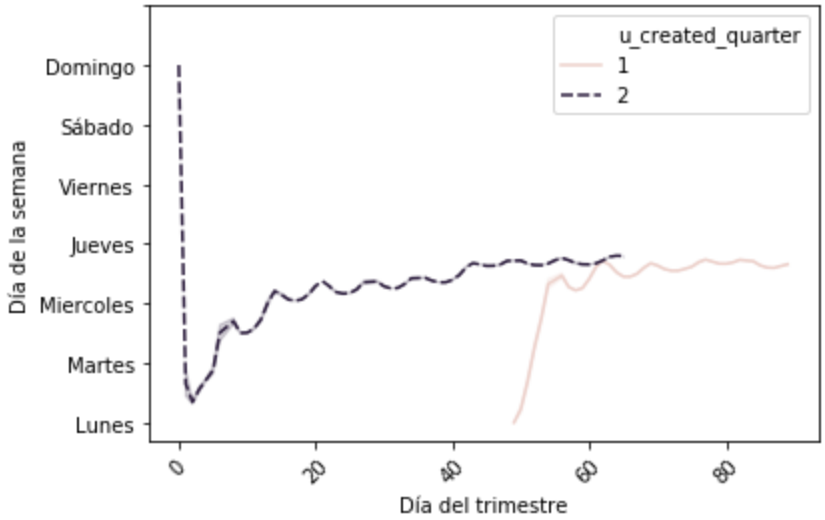

Figura 8. Valor esperado por trimestre de los registros efectivos para los días de la semana. Recurso propio.

Volviendo al histograma también se denota una trivialidad, y es que el día con menor actividad es el domingo

Análisis hecho a las horas:

Al graficar las horas en un histograma se muestra una distribución que pareciera ser bimodal con un valle alrededor de las 12 horas:

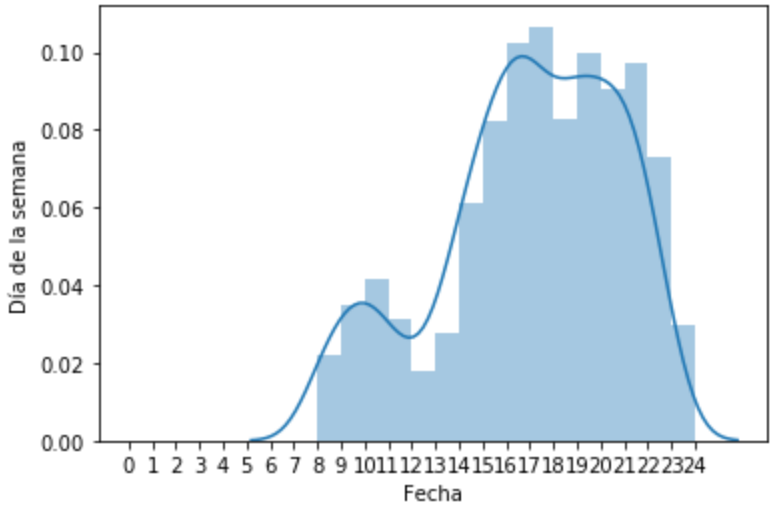

Figura 9. Histograma de frecuencias de los registros efectivos por horas. Recurso propio.

También se observa que el rango de actividad de los registros efectivos se encuentra entre las 8 y las 24 horas. Un análisis con el gráfico de cajas y bigotes, como en el caso

de los días de la semana, puede ayudar a construir una visión al interior de los datos dentro de la distribución.

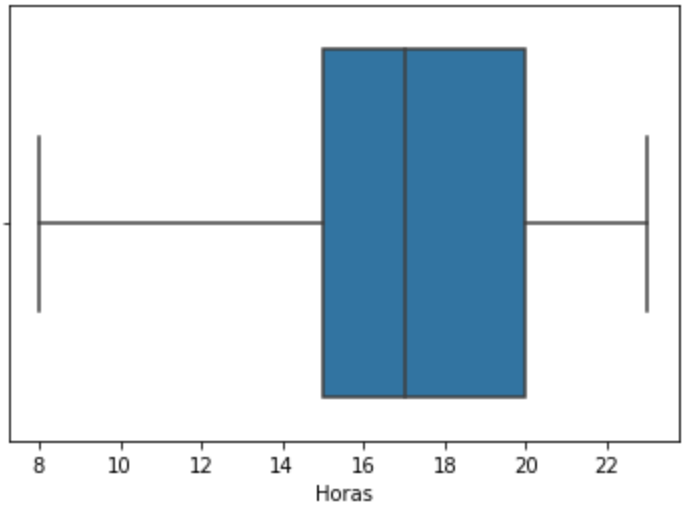

Figura 10. Diagrama de cajas y bigotes de los registros efectivos por horas. Recurso propio.

Se puede observar que efectivamente el rango intercuartílico se ubica entre las 15 y las 20 horas, lo que señala una importante actividad de registros efectivos en esa franja horaria. Para obtener el sentido de la asimetría se calculan las medidas de tendencia central:

- Moda = 17
- Mediana = 17
- $\mu =$ 16.8815

Tanto la moda y la mediana son iguales y la media se ubica apenas por debajo de este valor; así que se puede decir que el comportamiento de los datos es prácticamente el de una distribución normal. Al tomar la población como una distribución normal se pueden calcular las medidas de variación:

- $\sigma = 3.8427$

No es el objetivo de este proyecto adelantar un análisis más allá del descriptivo; pero cabe mencionar que la desviación estándar ofrece una idea de cuán alejados se suelen encontrar los datos alrededor de la media y esto es vital para hacer análisis inferencial en un trabajo posterior.

A continuación se muestra un gráfico con el valor esperado:

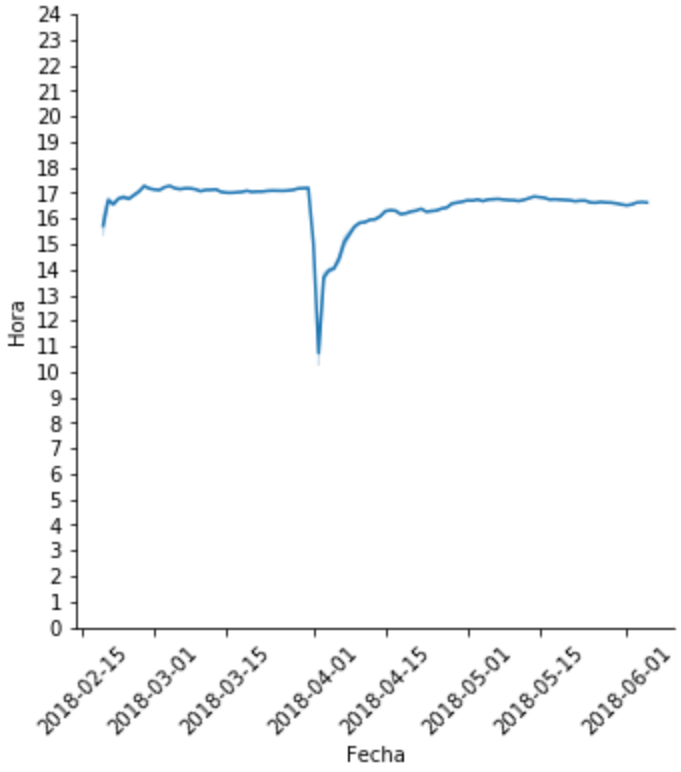

Figura 11. Valor esperado de los registros efectivos para las horas. Recurso propio.

Se logra evidenciar que, aunque los registros efectivos sufren un cambio brusco a principios de abril del 2018, la tendencia es a que la actividad se sostenga alrededor de las 16 horas. Si se divide de nuevo el tiempo por trimestres, una fuerte tendencia a estabilizar la hora de mayor actividad de registros efectivos entre las 16 y las 17 horas.

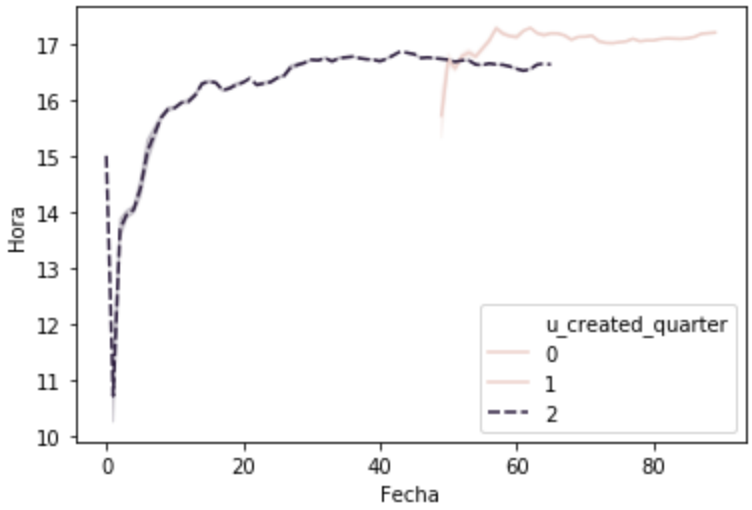

Figura 12. Valor esperado por trimestre de los registros efectivos para las horas. Recurso propio.

Como extra,y como una curiosidad, se aprovecha el dato de la fecha de nacimiento para ver entre qué rango de edades se encuentran las personas que son más proclives a

registrarse para acceder gratuitamente a la red WiFi del establecimiento de El Barista.

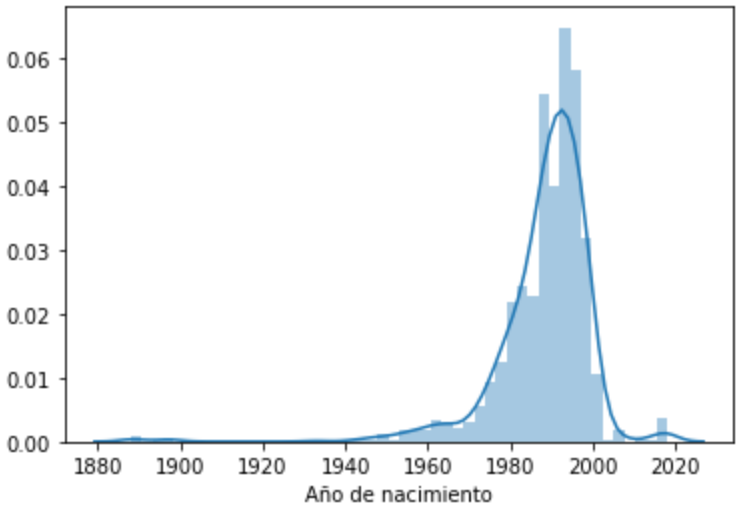

Figura 13. Histograma del año de nacimiento de los usuarios. Recurso propio.

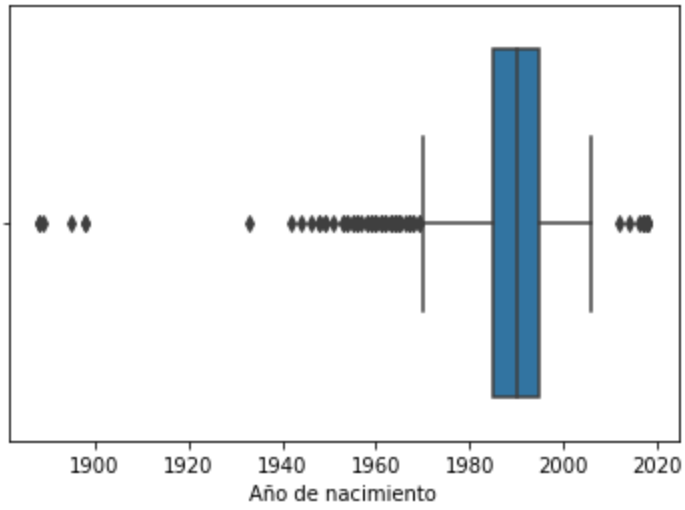

Figura 14. Diagrama de cajas y bigotes para el año de nacimiento de los usuarios. Recurso propio.

Un vistazo superficial muestra que:

- Existen datos atípicos sin tratar entre los años 1880 y 1900, probablemente sean errores de escritura a la hora del registro.
- También se registran algunos datos atípicos cerca al 2020, posiblemente debido a las mismas razones del punto anterior.
- Hay una mayor cantidad de registros efectivos hechos por personas cuyo año de nacimiento se encuentra entre 1985 y el año 2000.

Al ejecutar describe() en la columna *ano_fecha_nacimiento* se obtienen los estadísticos:

- Moda = 1992
- Mediana = 1990
- μ = 1988.2
- IQR=[1985,1995]
- Max: 2018
- Min: 1888
- Std: 12.06

Esta información da soporte a la lectura al vuelo que se hizo sobre las gráficas, las personas dentro del rango de 1985 y 1995 son más proclives a registrarse para conectarse a Internet dentro de *El Barista Slow Coffee*. Existen datos atípicos que hacen que la desviación estándar sea elevada y por ese motivo, puede hacer menos confiable la información mostrada en la distribución. Un paso a seguir sería el de eliminar los datos atípicos y medir el valor esperado a lo largo del tiempo como primera medida, pero no es el objetivo de este proyecto extenderse hasta ese punto.

Conclusiones

- Las bases de datos suelen venir con información errada y es importante limpiar dicha información antes de entrar en la fase de análisis.
- Los problemas en los datos suelen ocurrir por motivos de:
    - Fallos de en el ingreso de los datos por parte de los usuarios.
    - Registros que no tienen información.
    - Fechas de modificación anteriores a las fechas de creación de los registros.
- Los datos atípicos se pueden detectar usando el diagrama de cajas y bigotes.
- Los datos atípicos pueden ocasionar que las distribuciones se muestren asimétricas.
- La esperanza es una medida de tendencia central que entrega información útil a la hora de hallar comportamientos a largo plazo.
- Un comportamiento que se muestra fuerte puede ser validado si se hace un análisis trimestral por medio de los valores esperados.
- El peso de los datos se encuentran entre el rango intercuartílico y puede ser aón más importante ver ese rango a ver el pico de una distribución.

Trabajo Futuro

El siguiente paso será ampliar este proceso a otras columnas de la tabla *reg_efectivos* o que se encuentren en otras tablas de la misma base de datos. Este es, simplemente, un primer paso de un proceso de descripción de datos que necesita hacerse detalladamente.

Una vez terminado este proceso, se podrá comenzar con el análisis inferencial que se hace sobre el estudio de las medidas de dispersión encontradas en las distribuciones. esto permitirá detectar relaciones de covarianza entre pares de datos.

Yendo mucho más allá, se podrá hacer un análisis predictivo aplicando procesos de regresión lineal e incluso aplicando algoritmos de aprendizaje automático.

Anexos

Tabla 1
reg_efectivos

macaddr	d_created	d_modified	id_user	u_modified
cc:2d:b7:15:d9:0f	2018-02-19 16:51:49	2018-02-19 17:14:10	4017	2018-02-19 16:52:19
ac:ed:5c:8b:93:9a	2018-02-19 15:26:07	2018-02-19 17:40:06	4011	2018-02-19 15:33:51
2c:33:61:92:e1:9f	2018-02-19 15:35:43	2018-02-19 16:30:07	4012	2018-02-19 15:36:28
cc:61:e5:4c:23:1a	2018-02-19 21:25:28	2018-02-19 23:19:31	4032	2018-02-19 21:32:02
2c:0e:3d:bb:70:54	2018-02-19 19:57:24	2018-02-19 20:26:09	4027	2018-02-19 19:58:35

u_created	rol	email
2018-02-19 16:52:19	reg	[redacted]@outlook.es
2018-02-19 15:33:51	reg	[redacted]@hotmail.com
2018-02-19 15:36:28	reg	[redacted]@hotmail.com
2018-02-19 21:32:02	reg	[redacted]@gmail.com
2018-02-19 19:58:35	reg	[redacted]@gmail.com

fecha_nacimiento	u_created_dayofweek	u_created_dayofweek_expected_value
1999-09-09 00:00:00	1	1.0
1984-03-16 00:00:00	1	1.0
1987-10-08 00:00:00	1	1.0
1946-11-18 00:00:00	1	1.0
1980-04-21 00:00:00	1	1.0

u_created_hour	u_created_hour_expected_value	u_created_date	created_dayofyear
16	16.0	2018-02-19	50
15	15.5	2018-02-19	50
15	15.333333333333332	2018-02-19	50
21	16.75	2018-02-19	50
19	17.2	2018-02-19	50

u_created_dayname	u_created_quarter	ano_fecha_nacimiento	u_created_dayofquarter
Monday	1	1999	49
Monday	1	1984	49
Monday	1	1987	49
Monday	1	1946	49
Monday	1	1980	49

u_created_hour_expected_value_by_quarter	u_created_dayofweek_expected_value_by_quarter
16.0	1.0
15.5	1.0
15.333333333333332	1.0
16.75	1.0
17.2	1.0

Nota. Recurso propio.

Tabla 2
Tabla dispositivos

macaddr	d_created	d_modified	id_user
34:e6:ad:64:1c:61	2017-10-20 22:24:35	2017-10-21 20:32:19	0
f0:c8:50:0b:2f:f1	2017-10-11 21:50:35	2017-10-11 23:39:48	0
a4:ba:76:7a:cd:74	2017-10-12 21:12:03	2017-10-13 23:50:06	0
7c:b1:5d:ab:5c:bc	2017-10-21 06:24:18	2017-10-21 20:32:21	0
e4:58:b8:b5:1b:a3	2017-10-21 14:58:27	2017-10-21 20:32:27	0

Nota. Recurso propio.

Tabla 3
Tabla users

id_user	u_modified	macaddr	u_created
11023	2018-10-08 11:51:10	14:dd:a9:a8:94:7b	2018-09-25 19:10:25
11025	2018-09-25 20:01:10	f4:8e:92:7f:b7:bb	2018-09-25 19:40:12
11075	2018-10-13 21:13:26	78:ca:39:af:64:20	2018-10-01 12:06:39
11084	2018-10-05 18:51:35	74:1b:b2:5a:46:c4	2018-10-05 18:37:35
11085	2018-10-06 14:15:10	7c:1c:68:e7:92:b4	2018-10-06 12:19:44

rol	email	fecha_nacimiento
reg	steven.mejia@wif-id.co	1985-08-13 00:00:00
reg	stefania.infantesanchez@gmail.com	1990-05-19 00:00:00
reg	gerencia@ekisushi.co	1989-02-14 00:00:00
reg	johajuanes23@hotmail.com	2018-10-05 00:00:00
reg	luisalbertoyepes11@gmail.com	2002-04-25 00:00:00

Nota. Recurso propio.

Bibliografía

[1] Gour, V., Sarangdevot, S. S., Tanwar, G. S., & Sharma, A. (2010). Improve performance of extract, transform and load (ETL) in data warehouse. *International Journal on Computer Science and Engineering*, *2*(3), 786-789.

[2] Papoulis, A. (1984). Expected value; dispersion; moments.

[3] Guo, P. (2013). Data science workflow: Overview and challenges. *Communications of the ACM*.

[4] O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

[5] Skewness and the Mean, Median, and Mode - Introductory Statistics - OpenStax. (2019). Retrieved 27 November 2019, from https://openstax.org/books/introductory-statistics/pages/2-6-skewness-and-the-mean-median-and-mode