

# Sprint 3: Data Analytics & Machine Learning Report

## Contenidos

Integrantes, roles y responsabilidades	1
Links asociados	1
Recapitulación	2
Dashboard	4
EDA para ML	11
Sistema de Recomendación de Restaurantes	12

## Integrantes, roles y responsabilidades

ROL	Name	Email	NameAbr	Github
Machine Learning	Diego Osorio	dosoriofc@gmail.com	DO	dosoriofc
Data Engineer	David Marimón	david.neko26@gmail.com	DM	DaAnMaGi
Data Engineer	Salomón Orozco	Salomonorozcojaramillo@gmail.com	SO	SaloLL
Data Analytics	Marcela Correal	mcorreal@gmail.com	MC	MarceCorreal
Data Analytics	Juan Garate	garatejb@gmail.com	JG	Batxa

## Links asociados

### Documentos internos

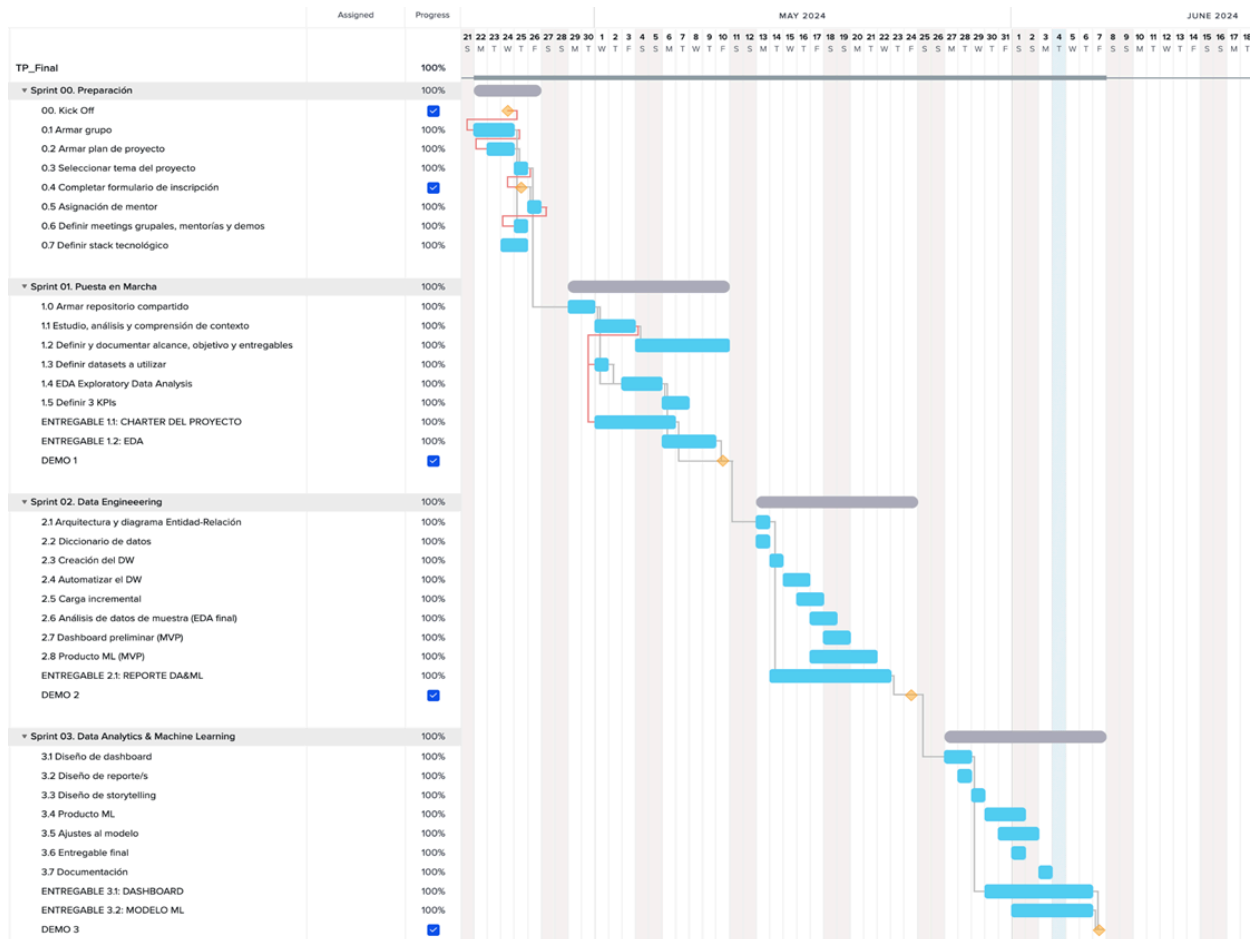
- Repositorio del proyecto [https://github.com/Batxa/DS\\_ProjectFinal.git](https://github.com/Batxa/DS_ProjectFinal.git)
- Follow-up de tareas: <https://trello.com/b/sH9ofad9/tpfinal>
- Cronograma: <https://app.teamgantt.com/projects/gantt?ids=3939144>
- Dashboard:  
[https://lookerstudio.google.com/reporting/d89b639c-7877-4d8b-94d9-312b5c177cd2/page/p\\_6bmjritshd/edit](https://lookerstudio.google.com/reporting/d89b639c-7877-4d8b-94d9-312b5c177cd2/page/p_6bmjritshd/edit)
- Sistema de recomendación: <https://pfyelpml-upkwe29phjyawezyvqffu6s.streamlit.app/>

# Recapitulación

En el primer y segundo sprint del proyecto, se cumplieron con los entregables propuestos y organizados en Trello, a saber:

- Repositorio compartido
- EDA
- Project Charter
- DER
- Diccionario de Datos
- Nube en GCP
- DWH

Este Sprint, tal como se observa en el Gantt de Trello, se centró, en el desarrollo del dashboard, la última parte del desarrollo de modelo de ML, y en la automatización de la carga de datos.



En esta fase se cumplió con la metodología Scrum sugerida al inicio del proyecto.

El equipo se reunió los martes y los jueves de las 6 semanas del módulo y con el tutor Emilio Santander los días Lunes, Miércoles y Viernes. Al finalizar cada sprint se sostuvieron encuentros con la Project Owner del proyecto.

A continuación se describen los resultados finales del proyecto

# Dashboard

Este dashboard pretende cumplir con las expectativas del Proyecto y responder a los objetivos trazados al inicio, consignados en el Project Charter.

Está dirigido a propietarios y directivos de restaurantes y bares, que contratan a la consultora SMART CHOICE ANALYTICS, interesados en obtener información de calidad sobre el sector y analizar información que ayude a la definición de proyectos y mejoras en los establecimientos.

Enlace Dashboard Dashboard:

[https://lookerstudio.google.com/reporting/d89b639c-7877-4d8b-94d9-312b5c177cd2/page/p\\_u9byqk4shd/edit](https://lookerstudio.google.com/reporting/d89b639c-7877-4d8b-94d9-312b5c177cd2/page/p_u9byqk4shd/edit)

En este caso, la información se analiza en su completitud, como un demo y oferta de inicio hacia el cliente, quien podrá seleccionar los filtros que se consideren relevantes dependiendo del caso particular.

La página introductoria del Dashboard, se presenta como un resumen de la información procesada y algunos datos condensados:



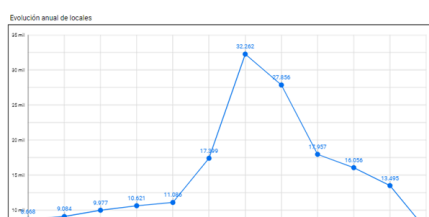
En la parte superior se puede observar el resultado de los KPIs sugeridos al inicio del proyecto, que se explicarán más adelante.

En el cuerpo a mano izquierda se presenta al cliente el tamaño de las muestras de datos limpios utilizados.

En este caso, el número de registros en las bases de datos sugeridas: datos sugeridos: sites, users, checkins y reviews.

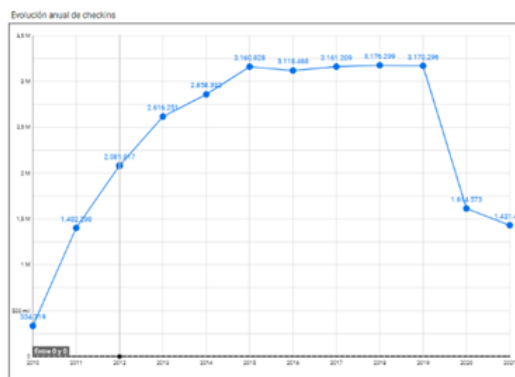
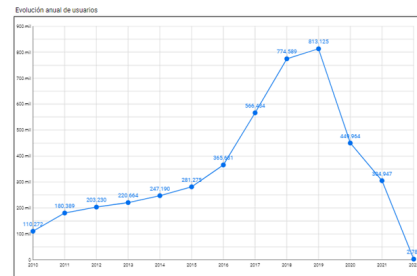
En la parte derecha del cuerpo, una gráfica que compara el progreso de los checkins y los reviews, evidenciando que desde el 2011 se cuenta con la información de las referencias de los usuarios que aumentan significativamente año a año, y que hasta el año 2019 refiere que, alrededor del 30% de los visitantes, incluyen sus reseña en la aplicación.

En el costado izquierdo, se pueden encontrar enlaces que direccionan a tableros particulares, que muestran el análisis de la información a saber:



**Sites:** Este enlace refiere la evolución anual de la cantidad de locales, lo que indica que tienen cierta tendencia a la baja que se ratifica en el 2019 con la llegada de la pandemia. Esta información ubicará al cliente en cuanto al tamaño general del sector,

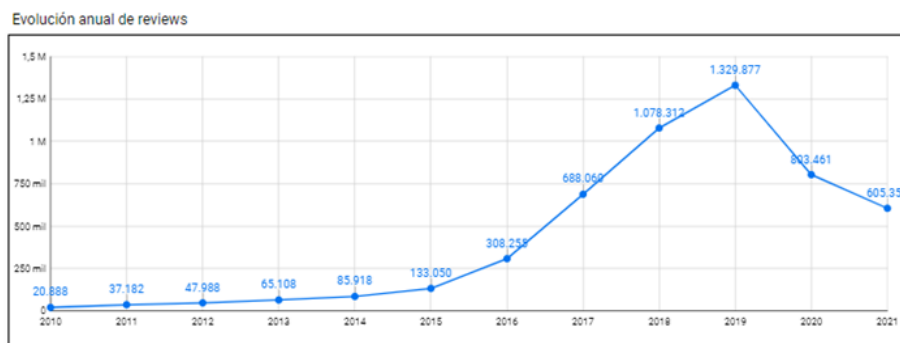
**Users:** Esta gráfica de cantidad de usuarios cuenta con la misma tendencia de la cantidad de sitios, lo que puede indicar, además del análisis anterior, que los datos se encuentran bien ajustados y corresponden a la realidad. Será importante para el cliente entender el tamaño de la demanda y uso de los productos o servicio que la empresa provee



**Checks:** En este caso, se observa cómo terminando la pandemia se inicia la recuperación del sector mediante el insípido aumento de las visitas a los locales.

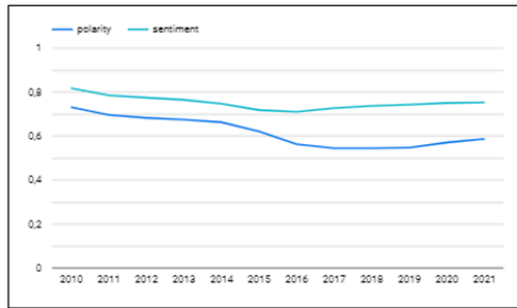
Esto demuestra la relación de los datos con los hechos reales, actividad que siempre estará contemplada en los escenarios de análisis con los clientes

**Reviews:** En esta sección se analiza no sólo la evolución de la cantidad de reviews que lógicamente es la proporcional que la cantidad de visitas, como se evidencia en el gráfico,

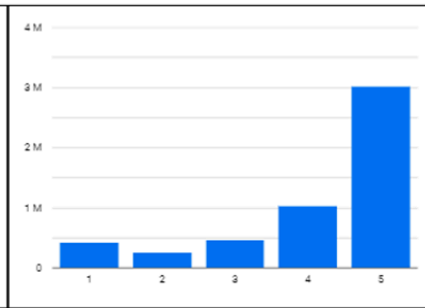


sino que se analiza la evolución del desempeño de los locales, utilizando para este indicador el la polaridad y el sentimiento que devuelve la función "Vader" que se utilizó en Python en el

Evolución anual de performance



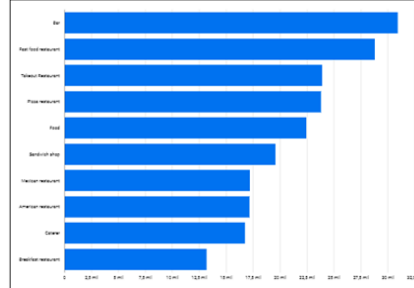
Distribución del rating



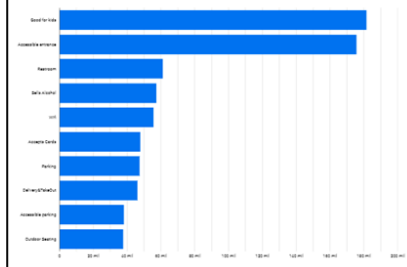
marco del desarrollo del EDA, y analizando las calificaciones que indica el usuario y los comentarios que realiza sobre la experiencia vivida en sus visitas.

**Categories y Attributes:** Se contempla en este análisis las categorías y atributos mejor rankeadas de la muestra, lo que le indicaría a un nuevo inversionista luces de qué tipo de lugares y atributos podría incluir en su plan de negocio con altas expectativas de éxito.

Top 10 de categorías



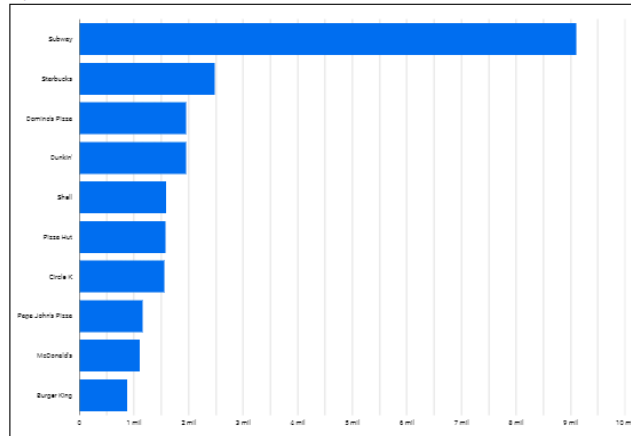
Top 10 de atributos (servicios auxiliares)



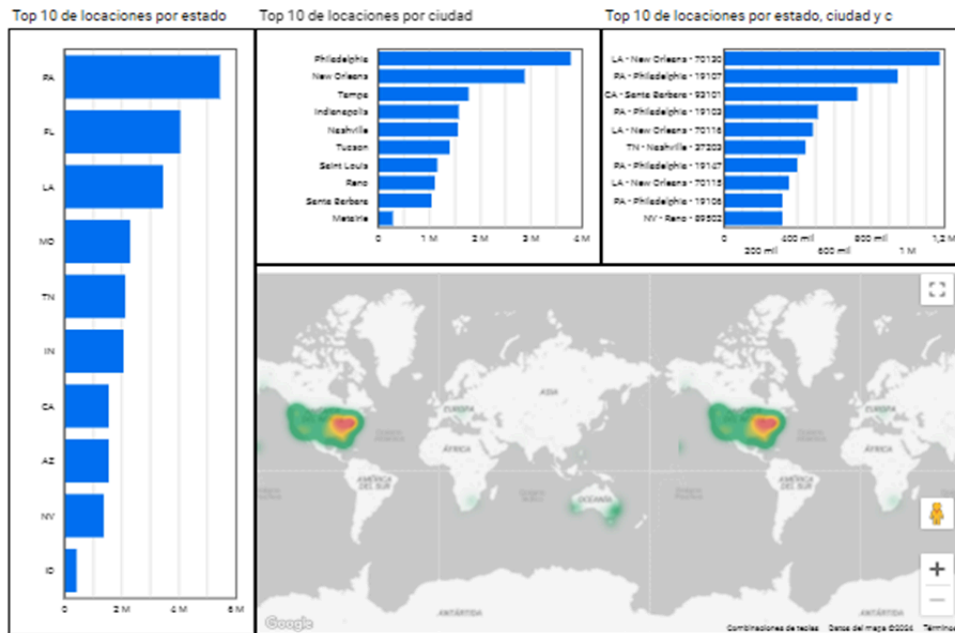
En este caso, se recomendaría al cliente tener en sus establecimientos actividades para niños, parqueaderos y facilidades de ingreso, así como servicios de baño y venta de bebidas alcohólicas.

**Brands:** El análisis de las marcas en el marco de la consultoría, será relevante en clientes que cuentan con una marca importante en el mercado, como en este caso, Subway, Starbucks y Domino's, ya que será posible comparar y dimensionar su presencia en el mercado por los estados de interés, número de locales y hasta el sentimiento de los reviews en cada uno de ellos.

Top 10 de marcas



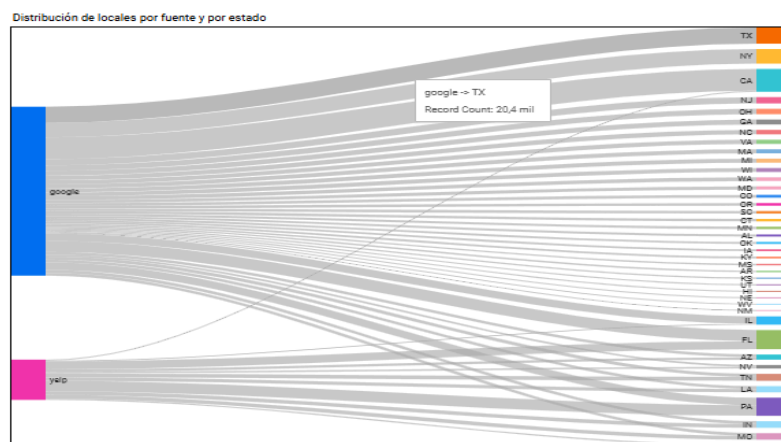
**Locations:** En esta sección se encuentra el análisis de las ubicaciones con mayor cantidad de establecimientos de interés, indicando al inversionista lugares atractivos para ubicar nuevos negocios. Con la información adquirida se podrían analizar estados como Philadelphia y New Orleans.



En este análisis se presenta el top 10 de los lugares de interés, sin embargo, las herramientas permiten visualizar el número de ítems que facilite el análisis particular al cliente.

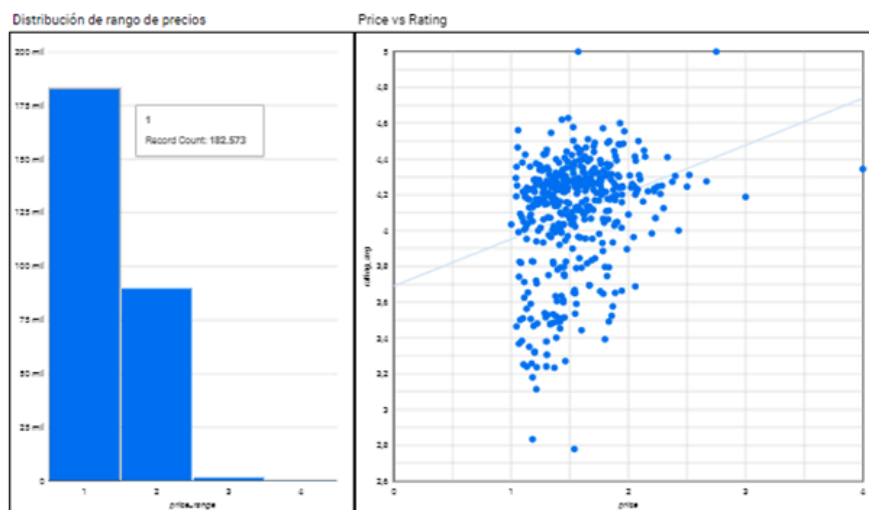
### Datasource-Locations:

Para el análisis de la distribución de locales y estados teniendo en cuenta la fuente de la información, que en este caso fue Yelp y Google, se utilizó la gráfica lo que indica que sistema de información es más utilizado en cada uno de los estados. Esta gráfica entrega información que ayudaría a analizar al cliente en qué sistema de información puede encontrar datos relevantes para el análisis de sus locales. Es muy útil en los casos que el acceso a diferentes bases de datos cuentan con costo monetario.



Será importante ofrecer al cliente la consecución de datos de diferentes fuentes. Este gráfico ayuda en el análisis de la calidad y cantidad de registros de cada fuente.

**Prices:** El análisis de de precios, que se ofrecerá en las consultorías, ayuda a entender los datos referentes a la relación de precio y calificación de los locales. Es un análisis que sin importar el sector es ineludible realizar, pues el precio de los bienes estará determinado no sólo por el costeo si no por la disposición de los clientes a pagar el nivel de precios, y eso lo dicta el análisis de los datos. Para este análisis suele elegirse una gráfica de dispersión que ofrece al cliente la relación del rating con el nivel de precios. Para este caso, de los 4 niveles que se establecieron, resulta que los establecimientos con precios en el segundo cuartil cuentan con la mayor frecuencia de ratings altos.



También se presenta un gráfico de barras que analiza la cantidad de locales en cada franja de precios, arrojando que la mayoría de los locales se encuentran en la franja de precios más baja, y pudiendo concluir también, que la diferencia entre cada una de las franjas de precio es bastante pronunciada.

## KPIs

De los servicios más útiles que prestará la consultoría realizada por SMART CHOICE ANALYTICS es la definición, cálculo y análisis de los indicadores de gestión, ya que siempre se diseñarán con la metodología SMART; serán específicos, medibles, alcanzables y con un tiempo definido.

Adicional, se presentarán al cliente estos KPI en formatos fáciles de entender para hacer seguimiento continuo sobre los indicadores señalados.

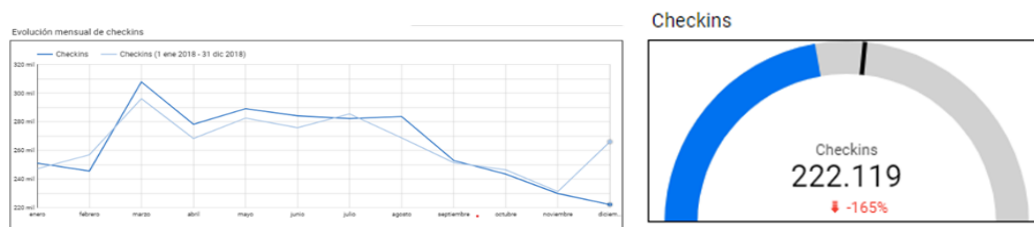


Para este caso en particular, los KPI definidos son:

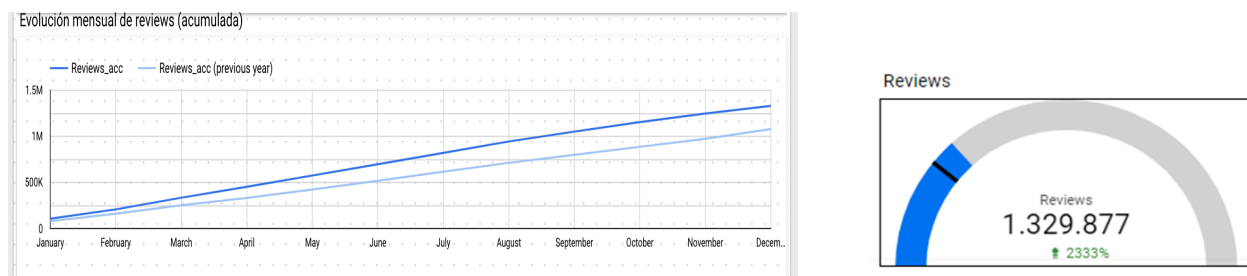
**KPI1 - Checkins:** Aumentar en 5% los checkins respecto al mismo período del año anterior.

Este KPI evalúa la tasa de variación porcentual entre el 2018 y el mismo período de 2019.

Para el desarrollo de este KPI se presenta una tabla donde se comparan la cantidad de visitas mes a mes del año 2018 con el resultado del año 2019, el cálculo del objetivo y el indicador que resulta disminución en 16.5%



**KPI2- Reviews:** Aumentar en un 20% las reviews acumuladas con respecto al mismo período del año anterior



Este indicador busca asegurarla cantidad adecuada de datos de los clientes para tener la capacidad de analizar puntos de mejora.

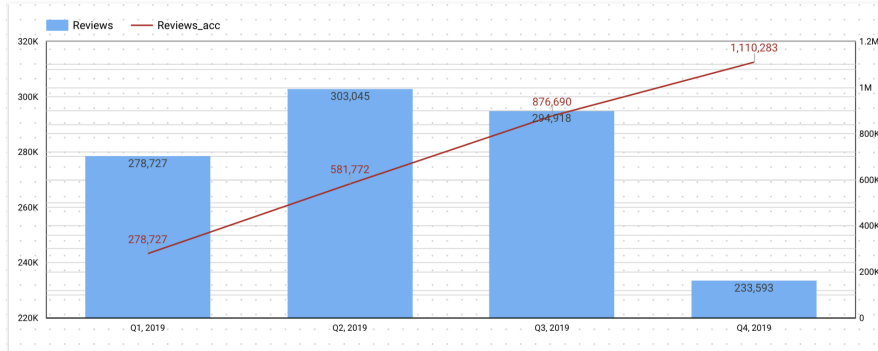
Este KPI evalúa la tasa de variación porcentual de las reviews acumuladas entre los periodos 2018 y 2019. En el dashboard se puede ver el cálculo del objetivo, la tabla de datos y el resultado del indicado en formato velocímetro que ofrece al cliente lecturas rápidas y claras en tiempo real.

**KPI3 – Positive Reviews:** Superar el millón de reviews acumuladas positivas en el último quarter.

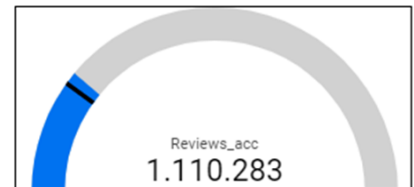
Además de presentar la tabla con la cantidad de reviews acumuladas por quarter, se presenta el cálculo del objetivo y la gráfica de líneas que indica una tendencia positiva y finalmente el velocímetro con el resultado final del KPI superado.

En esta ocasión el KPI evalúa el valor absoluto de las reviews acumuladas y establece un objetivo arbitrario.

Evolución mensual de reviews



Positive Reviews



## EDA para ML

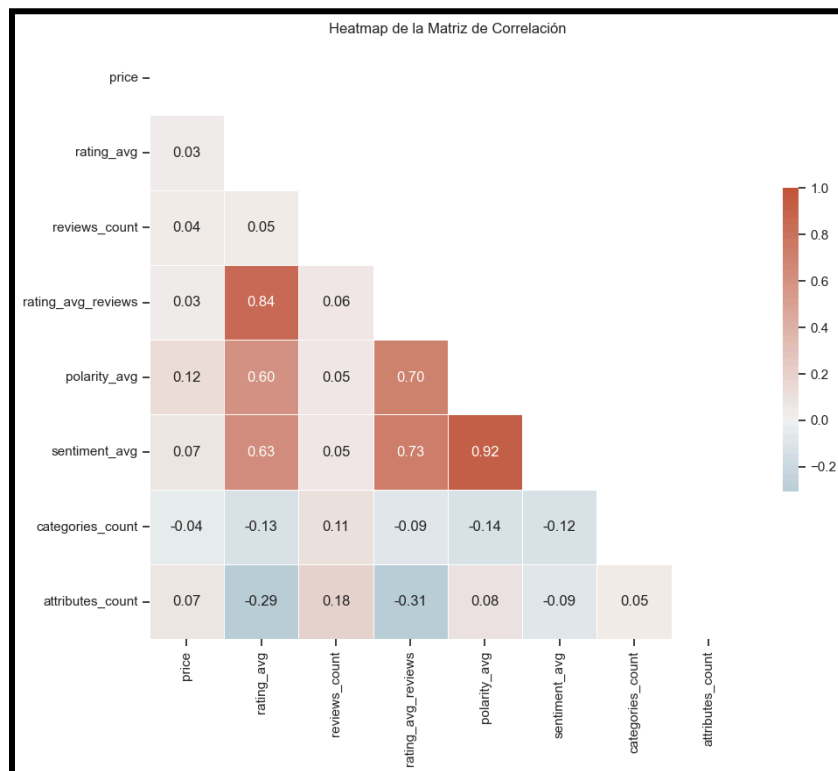
Se unen los principales datasets:

- Restaurants
- Reviews
- Categories
- Attributes

Obteniendo los siguientes campos:

- State
- City
- Price
- rating\_Avg
- Reviews\_count
- Rating\_avg\_reviews
- Polarity\_avg
- Sentiment\_avg
- Categories\_count
- Attributes\_count

Los resultados de la matriz de correlaciones se muestran a continuación:



# Sistema de Recomendación de Restaurantes

Se desarrolló un sistema de recomendación de restaurantes que se basa en la comparación de las reseñas de un usuario con las de los otros usuarios registrados en el sistema y, mediante técnicas de Machine Learning, determina cuáles son los usuarios con gustos más parecidos, y en base a esta similitud le recomienda uno (o más restaurantes) de cualquier categoría o de sólo una categoría especificada por el usuario.

## Herramientas de desarrollo

El algoritmo de recomendación se desarrolló en Python utilizando la biblioteca open-source de **Machine Learning Scikit-learn**, y principalmente hace uso de dos de sus funciones *TfidfVectorizer* y *Cosine\_similarity*. Adicionalmente, se utilizaron las bibliotecas de Python **NLTK/SentimentIntensityAnalyzer** para el análisis de sentimientos de las reseñas y **fuzzywuzzy** para la homologación de las categorías, y finalmente **Streamlit** para el desarrollo de la interfaz web interactiva.

### 1. **Scikit-learn**

- a. **TfidfVectorizer**: se utilizó para el procesamiento de lenguaje natural (NLP) para transformar el texto de las reseñas en vectores numéricos que fueron utilizadas en el algoritmo de similitud.
- b. **Cosine\_similarity**: se utilizó para calcular la similitud (mediante el algoritmo de la similitud del coseno) entre todos los vectores numéricos que representan las reseñas de los usuarios.

### 2. **NLTK (Natural Language Toolkit)**

- a. **SentimentIntensityAnalyzer**: se utilizó para evaluar el tono emocional de las reseñas y obtener una puntuación de sentimiento que refleja la positividad, negatividad, neutralidad del sentimiento expresado en las reseñas.

### 3. **Fuzzywuzzy**: se utilizó para comparar las categorías de ambos set de datos, Google y Yelp, y obtener una puntuación de similitud que va del 0% al 100%, donde una puntuación del 100% indica que las cadenas son idénticas; esto se hizo para generar un listado reducido y estandarizado de categorías

### 4. **Streamlit/Github**: se utilizó para crear interfaz web interactiva que permite el ingreso de los datos y las selecciones de los usuarios y mostrar el resultado del sistema de recomendación

## Datos de Entrada:

El sistema permite al usuario ingresar y seleccionar los siguientes parámetros para pedir la recomendación:

1. Identificador único del usuario en la base de datos
2. El número de recomendaciones que desea - Disponible: de 1 al 10
3. El estado donde desea la recomendación - Disponible: todos los estados de Estados Unidos
4. La categoría de restaurantes en la que desea la recomendación - Disponible: todas las categorías y la opción All (recomienda sin discriminar la categoría)

### Datos de Salida:

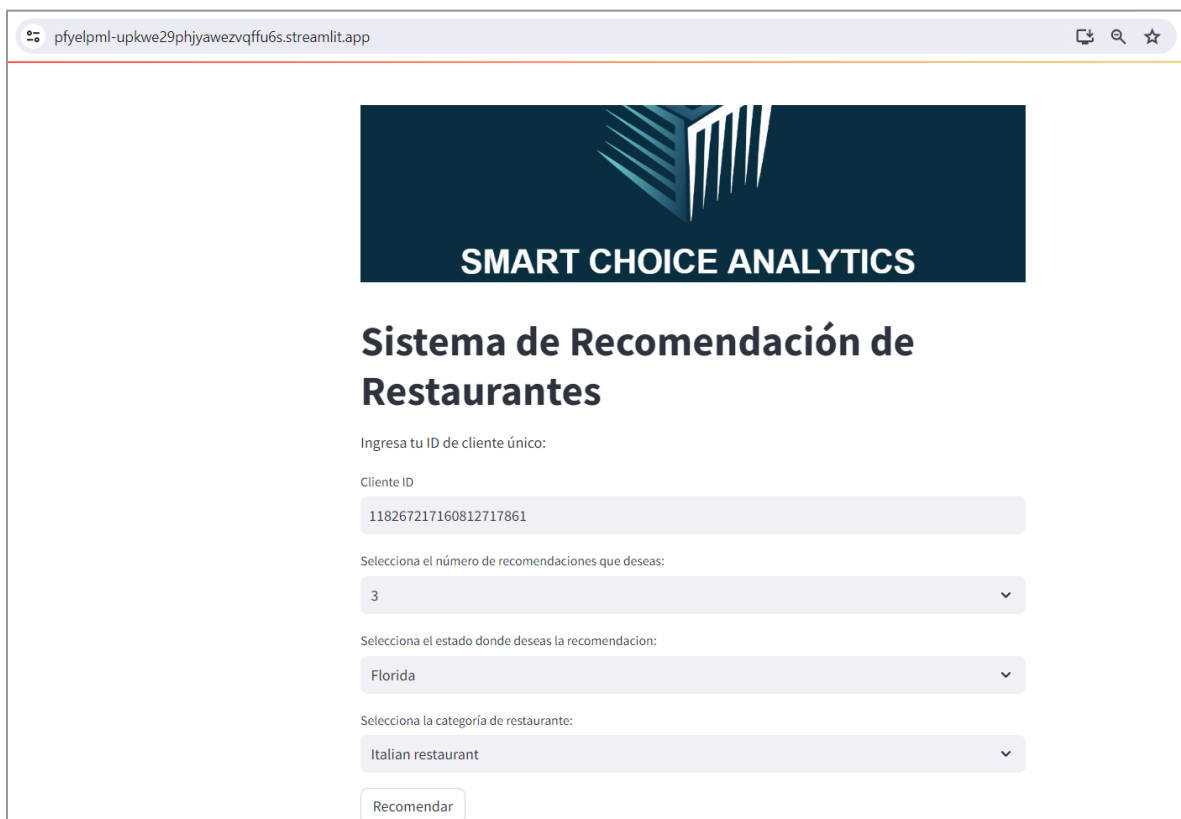
1. Nombre(s) de restaurantes recomendados y para cada uno muestra:
  - a. La categoría a la cual pertenece el restaurante
  - b. Una reseña - imprime a manera de muestra la reseña que obtiene el mayor puntaje positivo obtenido con un algoritmo de análisis de sentimiento
  - c. Rating - el puntaje otorgado al restaurante por el cliente que emitió la reseña mostrada

### Link de Ingreso al Sistema de Recomendación:

<https://pfyelpml-upkwe29phjyavezvqffu6s.streamlit.app/>

### Caso de Uso

- Valores de entrada:
  - cliente\_id = '118267217160812717861'
  - número de recomendaciones: 3
  - estado: Florida
  - categoría: Italian restaurant



The screenshot shows a web application interface for "SMART CHOICE ANALYTICS". The title is "Sistema de Recomendación de Restaurantes". Below the title, there is a form with the following fields:

- Ingresa tu ID de cliente único:** A text input field containing the value "118267217160812717861".
- Selecciona el número de recomendaciones que desees:** A dropdown menu with the value "3" selected.
- Selecciona el estado donde desees la recomendación:** A dropdown menu with the value "Florida" selected.
- Selecciona la categoría de restaurante:** A dropdown menu with the value "Italian restaurant" selected.

At the bottom of the form is a button labeled "Recomendar".

- Resultados Obtenidos:

Restaurantes recomendados:

1. Rossini Italian Bistro:

Categoria: Italian restaurant

Muestra de Reseña: The food was delicious and the decor was clean and inviting. I especially enjoyed the two appetizers we had; the meatballs and calamari. Rating: 5

2. Beccofino Italian Bistro:


Categoria: Italian restaurant




Muestra de Reseña: Not much from the outside but inside you feel like you've stepped into Italy. Every bite prepared with fresh ingredients is a pleasure for the palate. If you appreciated excellent authentic Italian food this is your restaurant. Make a reservation, it gets busy quickly! Rating: 5

3. Benito's Pizza & Pastabilities:

Categoria: Italian restaurant

Muestra de Reseña: Cute and cozy inside. Great customer service. Benito is a sweet heart. First time customer and def going to add to the list of local places to go. Rating: 4

 pfyelpml-upkwe29phjyawezvqffu6s.streamlit.app



Recomendar

Restaurantes recomendados:

1. Rossini Italian Bistro:

Categoría: Italian restaurant

Muestra de Reseña: The food was delicious and the decor was clean and inviting. I especially enjoyed the two appetizers we had; the meatballs and calamari. Rating: 5

2. Beccofino Italian Bistro:

Categoría: Italian restaurant

Muestra de Reseña: Not much from the outside but inside you feel like you've stepped into Italy. Every bite prepared with fresh ingredients is a pleasure for the palate. If you appreciated excellent authentic Italian food this is your restaurant. Make a reservation, it gets busy quickly! Rating: 5

3. Benito's Pizza & Pastabilities:

Categoría: Italian restaurant

Muestra de Reseña: Cute and cozy inside. Great customer service. Benito is a sweet heart. First time customer and def going to add to the list of local places to go. Rating: 4