



**SISTEMA DE  
RECOMENDACION  
DE RESTAURANTES**

# **CONSULTORA KANGAROO**

## **INFORME**

**Integrantes:**

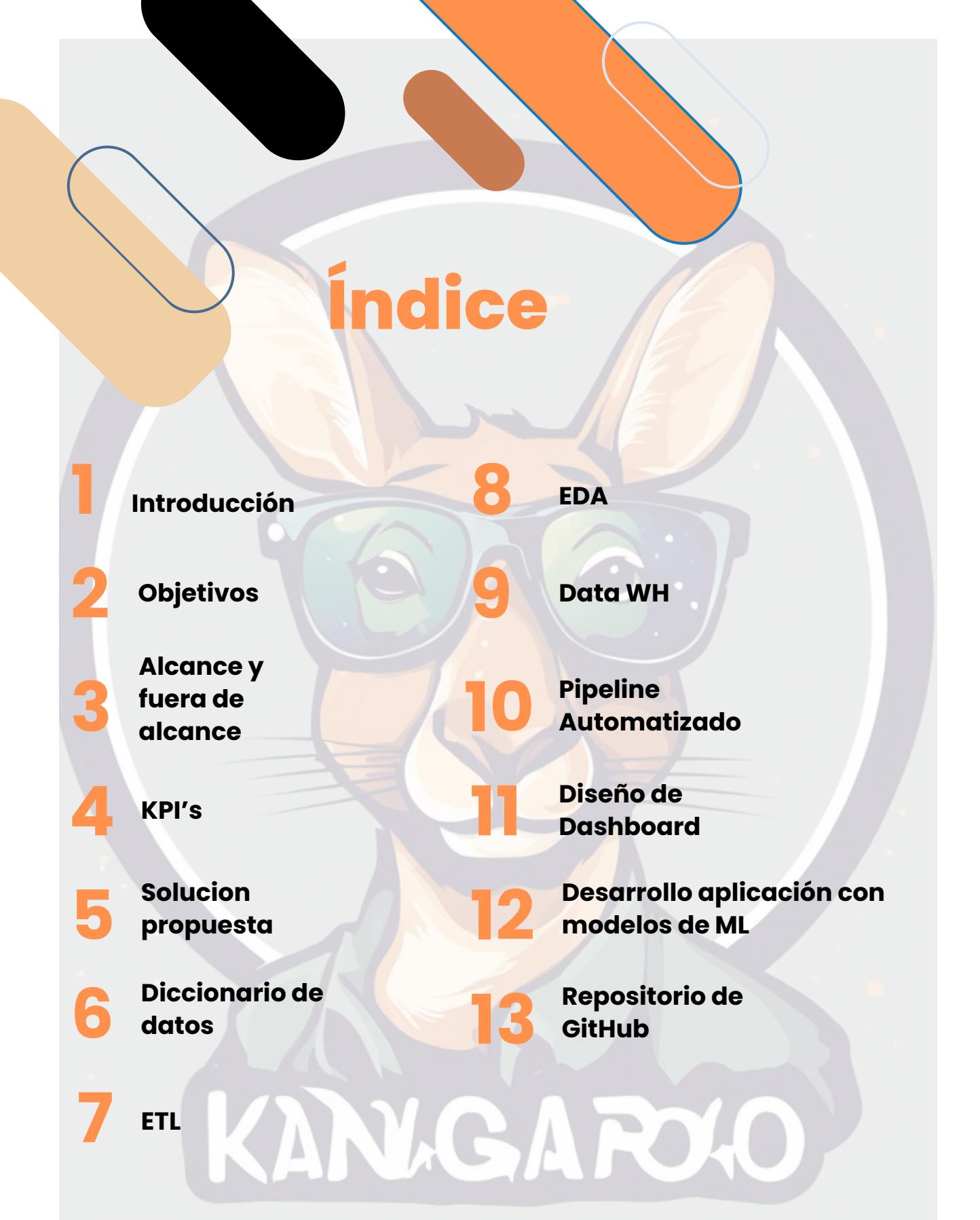
**Fausto Ezquerra**

**Maria Constanza Florio**

**Joaquin Millan**

**Martin Peñas**

**Nicolas Yapur**



# Índice

- 1 Introducción
- 2 Objetivos
- 3 Alcance y fuera de alcance
- 4 KPI's
- 5 Solucion propuesta
- 6 Diccionario de datos
- 7 ETL
- 8 EDA
- 9 Data WH
- 10 Pipeline Automatizado
- 11 Diseño de Dashboard
- 12 Desarrollo aplicación con modelos de ML
- 13 Repositorio de GitHub

# INTRODUCCIÓN

En un mundo en constante movimiento, la experiencia gastronómica juega un papel fundamental en la satisfacción de los viajeros. En este sentido, el Estado de Florida está comprometido en elevar la calidad de la experiencia de sus turistas, y es por eso que ha decidido contratar a Kangaroo, una consultora startup especializada en Business Intelligence.

Kangaroo se ha destacado por su capacidad de transformar datos en soluciones prácticas y personalizadas. En esta ocasión, hemos llevado nuestra experiencia al mundo de la gastronomía turística. Sabemos que cuando estamos de viaje, elegir el lugar perfecto para comer puede ser un desafío abrumador. ¿Dónde ir? ¿Qué opciones se adaptan a nuestras preferencias? ¿Cómo evitar perder tiempo valioso explorando opciones que no son adecuadas?

La respuesta a estos desafíos se llama "Kanguro Viajero", nuestra innovadora aplicación diseñada para simplificar la vida de los viajeros. A través de un análisis de datos exhaustivo y la implementación de un avanzado modelo de recomendación, Kanguro Viajero está diseñada para ofrecerte una experiencia gastronómica personalizada y sin complicaciones.

Imagina simplemente que el turista ingrese sus preferencias y ubicación, y en segundos reciba una lista de restaurantes que se ajusten perfectamente a sus necesidades. Kanguro Viajero le permitirá disfrutar de la deliciosa comida de Florida de manera eficiente y satisfactoria, sin perder tiempo en búsquedas interminables.

# OBJETIVOS

1. Armar un Datalake con todos los datos iniciales.
2. Crear un DataWarehouse que contenga data recopilada, procesada y específica.
3. Generar un pipeline de procesamiento de data automatizado
4. Crear un dashboard interactivo, que integre los datos para poder tener un seguimiento de los KPI's y contar con información valiosa.
5. Crear una APP de recomendación de restaurantes llamada "Kanguro Viajero", basándose en las reseñas de Google, así como en el análisis de los sentimientos de las reviews, en la cual el usuario coloca sus preferencias y su localización y esta devuelve distintas recomendaciones.

# ALCANCE Y FUERA DE ALCANCE

## Alcance

- Recopilación, limpieza y análisis de datos obtenidos de datasets de Google Maps y Yelp para obtener información sobre lugares de interés.
- Desarrollo de modelos de machine learning para analizar y clasificar las reseñas de los usuarios y los lugares en función de sus sentimientos.
- Implementación de un motor de recomendación que sugiere lugares a los usuarios en función de sus preferencias y localización.
- Diseño y desarrollo de un dashboard que permite visualizar el análisis de los datos.
- Puesta en marcha de proyecto en MVP donde se pueda observar la app en funcionamiento

## Fuera de alcance

Este proyecto no abarca recomendación de hoteles u otras atracciones turísticas ni otros estados que no sean Florida.

# KPI's

## 1. Evolución Anual del Rating:

Indicador que mide la evolución del rating promedio del restaurante seleccionado en el año y mes seleccionados respecto al mismo periodo del año pasado.

**Objetivo:** controlar la evolución en ratings de los restaurantes, se quiere estar al menos 5% arriba.

**Temporalidad:** anual/mensual.

## 2. Estado de Rating:

Indicador que mide el rating promedio para el año y mes seleccionados y el restaurante seleccionado y lo compara contra el rating promedio histórico.

**Objetivo:** medir que el rating esté por encima del rating histórico.

**Temporalidad:** anual/mensual.

## 3. Crecimiento de Ratings/Popularidad:

Indicador que mide la cantidad de ratings dados a un restaurante seleccionado en un año y mes seleccionados y lo compara con la cantidad de ratings respecto al mes pasado.

**Objetivo:** medirla evolución instantánea en popularidad de los restaurantes.

**Temporalidad:** mensual.

# KPI's

## 4.Crecimiento de Popularidad Anual:

Indicador que mide la cantidad de ratings dados a un restaurante seleccionado en una fecha seleccionada y lo compara con la cantidad de ratings respecto al mismo periodo del año pasado.

**Objetivo:** medirla evolución sostenida en popularidad de los restaurantes.

**Temporalidad:** mensual.

## 5.Crecimiento en cantidad de Clientes:

Indicador que mide el crecimiento en la cantidad de clientes comparando la cantidad de nuevos clientes en el periodo actual respecto al mismo periodo del pasado año.

**Objetivo:** medir el crecimiento sostenido en cuanto a clientes del restaurante o categoría de restaurantes .

**Temporalidad:** anual/mensual.

## 6.Crecimiento lineal en Clientes:

Indicador que mide el crecimiento lineal en la cantidad de clientes basado en los clientes que dejaron su calificación del restaurante o categoría de restaurantes.

**Objetivo:** medir crecimiento o estancamiento del restaurante o categoría de restaurantes.

**Temporalidad:** anual/mensual.

## **7. Evolución Anual del Rating de Reviews Positivas:**

Indicador que mide la evolución del rating promedio del restaurante seleccionado en la fecha seleccionada (año , mes , día) y compara el año corriente respecto al mismo periodo del año pasado.

**Objetivo:** lograr un incremento del 20% en la evolución review positivas en los restaurantes de forma anual, .

**Temporalidad:** anual.

## **8. Evolución de Anual de Crecimiento de Popularidad(cant. de reviews):**

Indicador que mide el incremento en porcentaje de la cantidad de reviews sobre un restaurante seleccionado en una fecha determinada, (año ,mes , día) y lo compara el año corriente con la cantidad de reviews respecto al año pasado.

**Objetivo:** lograr un crecimiento del 20% en cantidad de reviews anualmente respecto al año anterior.

**Temporalidad:** anual.

## **9. Evaluación Competencia según Ciudad-Polo Gastronómico :**

Indicador que mide la cantidad de restaurants ubicados en la misma ciudad, esto da al usuario la posibilidad de elegir entre restaurants aledaños, y saber si el restaurant se encuentra en una zona turística y competitiva.

**Objetivo:** Brindar al usuario información sobre la zona donde se encuentra el restaurant, si es una zona gastronómica, turística, de competencia.

**Temporalidad:** Actualidad.

# DICCIONARIO DE DATOS

Para realizar este trabajo tenemos dos grandes fuentes de datos, Google y Yelp con información sobre sitios turísticos en Estados Unidos.

Los datos de Google se componen dos carpetas: una llamada 'metadata-sitios' en la cual hay 11 archivos en formato json con información de los negocios en cada uno de los estados, y otra que se llama 'reviews' con 51 carpetas (una por estado) con archivos que contienen las reseñas de los usuarios.

Los archivos de metadata-sitios tienen la misma estructura, es por eso que se puede realizar un cruce entre los mismos, estos cuentan con la siguiente información y con su formato específico:

Nombre del campo	Descripcion	Tipo de dato
name	Nombre del negocio	String
address	Nombre del local y Direccion con codigo postal	String
gmap_id	ID que repesenta un lugar en google maps	String
description	Descripcion breve del negocio	String
latitude	Latitud de la ubicación	Float
longitude	Longitud de la ubicación	Float
category	Categoría a la que pertenece el negocio, pueden ser mas de 1.	String
avg_rating	Promedio de puntaje otorgado al negocio por parte de los opinadores	Float
num_of_reviews	Numero de opiniones dadas al negocio	Int
price	Simbolo de \$ que representa el rango de gasto. Un solo simbolo \$ significa barato, mientras que a mas simbolos de \$, mas caro será.	String
hours	Lista con los dias y horario de aperturas del negocio	String
MISC	Columna diccionario que dentro tiene informacion especifica del negocio, como opciones de comida, informacion en cuanto a la seguridad y salud, opciones del servicio, destacables, famoso por, metodos de pago, amenities, etc	String
state	Estado en el cual se encuentra el negocio. Abierto, cerrado y horarios.	String
relative_results	gmap_id relativo	String
url	Direccion web del sitio en google maps	String

# DICCIONARIO DE DATOS

Por otro lado, los archivos de reviews, de google contienen una amplia cantidad de opiniones realizadas por usuarios, en donde hay una carpeta por estado con diferente cantidad de archivos json dentro de cada una de ellas. Las mismas cuentan con la siguiente estructura:

Nombre del campo	Descripcion	Tipo de dato
user_id	Id del usuario	String
name	Nombre del usuario que realizo la opinion	String
time	Tiempo en timestamp en el que fue realizado la review	Int
rating	Puntaje otorgado por cada usuario, de 1 a 5 estrellas	Int
text	La opinion que dio el usuario al negocio	String
pics	Url con las fotos que sube el usuario a su review	String
resp	Tiempo y respuesta del negocio a la opinion	String
gmap_id	ID que representa el negocio en google maps	String

Dentro de la carpeta de Yelp tenemos 5 archivos que contienen información de distintos negocios. Con respecto al dataset de Yelp, tenemos cinco archivos:

**TIP** → Sugerencias rápidas de los restaurants

**Check-in** → Tiene la fecha de reserva del restaurant.

**Reviews** → Tiene reseñas de los usuarios.

**Users** → Información de usuarios de Yelp.

**Business** → Información de los negocios.

## Tip

Nombre del campo	Descripcion	Tipo de dato
user_id	Id del usuario	String
business_id	Id del negocio en Yelp	String
text	Sugerencia para el restaurant	String
date	Fecha en formato YYYY/MM/DD HH:MM:SS	Date
compliment_count	Cuantos cumplidos totales tiene	Int

## Check in

Nombre del campo	Descripcion	Tipo de dato
business_id	Id del negocio en Yelp	String
fechas	Fechas en las que se efectuaron un check-in	Time

# DICCIONARIO DE DATOS

## Reviews

Nombre del campo	Descripcion	Tipo de dato
review_id	Id de la reseña	String
user_id	Id del usuario	String
business_id	Id del negocio en Yelp	String
stars	Estrellas dadas al negocio	Float
useful	Numeros de votos a los usuarios que encontraron util esta opinion	Int
funny	Numeros de votos a los usuarios que encontraron graciosa esta opinion	Int
cool	Numeros de votos a los usuarios que encontraron genial esta opinion	Int
text	Texto de la reseña	String
date	Fecha en formato YYYY/MM/DD HH:MM:SS	Date

## Users

Nombre del campo	Descripcion	Tipo de dato
user_id	Id del usuario	String
name	Nombre del usuario	String
review_count	Cuenta de las reseñas realizadas	Int
yelping_since	Inicio su cuenta de yelp en formato YYYY/MM/DD HH:MM:SS	Date
useful	Numeros de votos que los usuarios encontraron util las opiniones del usuario	Int
funny	Numeros de votos que los usuarios encontraron graciosas las opiniones del usuario	Int
cool	Numeros de votos que los usuarios encontraron genial las opiniones del usuario	Int
elite	Años en los que el usuario fue elite	Int
friends	Cantidad de amigos del usuario	Int
fans	Cantidad de fans que tiene el usuario	Int
average_stars	Promedio del valor de las reseñas	Float
compliment_hot	Total de cumplidos "hot" recibidos por el usuario	Int
compliment_more	Total de cumplidos varios recibidos por el usuario	Int
compliment_profile	Total de cumplidos por el perfil recibidos por el usuario	Int
compliment_cute	Total de cumplidos "cute" recibidos por el usuario	Int
compliment_list	Total de listas de cumplidos recibidos por el usuario	Int
compliment_note	Total de cumplidos como notas recibidos por el usuario	Int
compliment_plain	Total de cumplidos planos recibidos por el usuario	Int
compliment_cool	Total de cumplidos geniales recibidos por el usuario	Int
compliment_funny	Total de cumplidos graciosos recibidos por el usuario	Int
compliment_writer	Total de cumplidos escritos recibidos por el usuario	Int
compliment_photos	Total de cumplidos en foto recibidos por el usuario	Int

## Business

Nombre del campo	Descripcion	Tipo de dato
business_id	Id del negocio en Yelp	String
name	Nombre del negocio	String
address	Direccion del negocio	String
city	Ciudad a la que pertenece el negocio	String
state	codigo de estado al que pertenece el negocio	String
postal_code	codigo postal del negocio	Int
latitude	Latitud de la ubicación	Float
longitude	Longitud de la ubicación	Float
stars	Promedio de estrellas recibidas	Float
review_count	Cantidad de reseñas en las que fue evaluado el negocio	Int
is_open	Si esta abierto, 0 cerrado	Int
attributes	atributos del negocio como valores	String
categories	lista de categoria a la que pertenece los negocios	String
hours	Horas y dias en las que el negocio opera	String

Estos conjuntos de datos unificados nos permiten crear un diagrama entidad-relación, una representación gráfica que incluye tablas de hechos y dimensiones.

Esto nos brinda la capacidad de visualizar la arquitectura y cómo interactúan nuestros datos en un diagrama.

# ARQUITECTURA DE DATOS

Anteriormente, habíamos diseñado una arquitectura de datos en la nube de Azure que empleaba:

- Azure Blob Storage para el almacenamiento
- Databricks para el procesamiento
- Azure SQL para el almacenamiento de datos.

La automatización de este pipeline de datos estaba siendo gestionada mediante Azure Data Factory.

Sin embargo, debido a una disminución en el volumen de datos que necesitábamos procesar y almacenar, así como a la búsqueda de una mayor eficiencia en costos, decidimos implementar cambios en la infraestructura de nuestro pipeline de datos y optar por cambiar de proveedor de servicios en la nube.

Google Cloud ofrece una alternativa de pipeline más económica y simplificada en comparación con las herramientas más robustas mencionadas anteriormente en Azure.

Nuestro nuevo pipeline estará compuesto por:

- Google Cloud Storage para la ingestión de datos
- Google Cloud Functions para el procesamiento y la automatización del pipeline
- Google BigQuery como almacén de datos procesados y listos para su utilización.

Se eligió Google Cloud Storage porque permite almacenar una cantidad masiva de datos, desde gigabytes hasta exabytes, lo que lo hace ideal para empresas de cualquier tamaño. Además, ofrece una variedad de opciones de precios, incluyendo almacenamiento estándar, Nearline y Coldline, lo que te permite elegir la opción más adecuada para tus necesidades y presupuesto.

Se eligió Cloud functions ya que escala automáticamente según la carga de trabajo. Cuando llegan más datos, se pueden crear nuevas instancias de funciones para manejar la carga adicional, sin necesidad de configuración manual. Otra razón relevante es que se paga por la cantidad de tiempo en la que se ejecutan las funciones y los recursos utilizados durante ese tiempo. Esto puede ser más económico que mantener servidores dedicados o instancias de máquinas virtuales para procesamiento de datos.

Se eligió utilizar BigQuery en un data pipeline ofrece escalabilidad, rendimiento y facilidad de consulta. Permite almacenar grandes volúmenes de datos, realizar consultas SQL rápidas y aprovechar integraciones con herramientas de análisis.

Además, ofrece seguridad, control de acceso y una estructura de precios basada en el uso real, lo que lo convierte en una elección sólida para el almacenamiento y análisis de datos en pipelines eficientes y escalables.

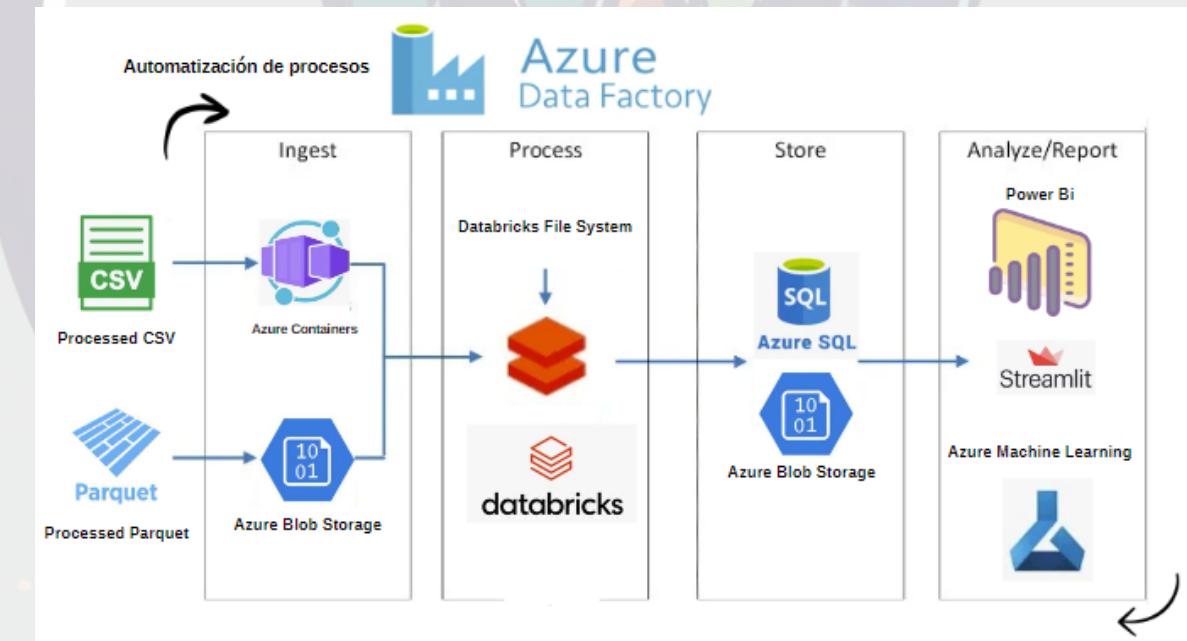
# ARQUITECTURA DE DATOS

Las herramientas utilizadas para la visualización/análisis y la creación de KPIs, así como el desarrollo de modelos de Machine Learning, seguirán siendo Power BI y Streamlit, por lo que no se realizaron modificaciones en esta parte del pipeline.

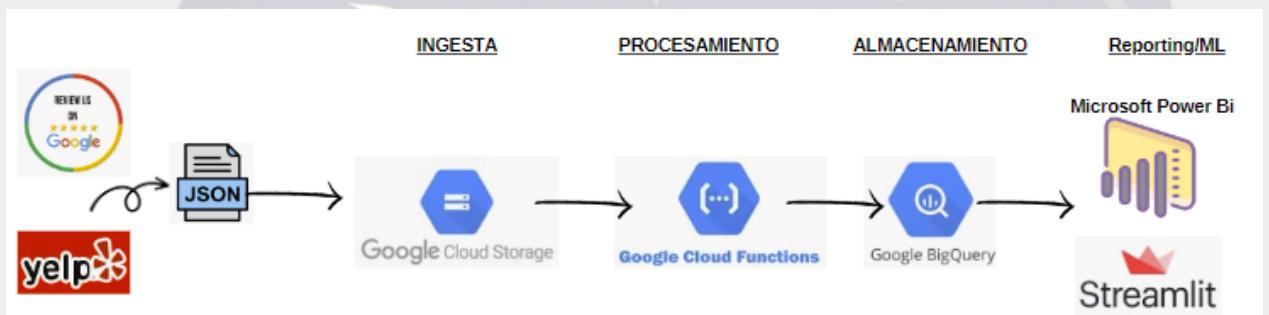
Decidimos usar Power BI para construir KPIs y dashboards es beneficioso debido a su facilidad de uso, capacidad de conectar múltiples fuentes de datos, visualización efectiva, actualización en tiempo real y opciones de colaboración, lo que permite tomar decisiones basadas en datos de manera eficaz y accesible.

Por su parte decidimos usar Streamlit para implementar un modelo de machine learning ya que se integra de manera fluida con bibliotecas de machine learning como TensorFlow y scikit-learn, lo que facilita la presentación y demostración de nuestros modelos de manera efectiva.

## Imagen del Pipeline Propuesto Anteriormente:



## Imagen del Pipeline Final a utilizar en el Proyecto:



# ETL

## (EXTRACCIÓN - TRANSFORMACIÓN - CARGA)

Nuestro pipeline de ETL, alojado en Google Cloud Platform, fue diseñado para garantizar la eficiencia y calidad de los datos. El proceso se inició en la primera fase del proyecto, cuyos detalles están documentados en el informe de esa etapa y en nuestro repositorio de GitHub.

Durante la segunda fase, implementamos dos funciones, construidas a través de Google Cloud Functions, para manejar datos provenientes de archivos JSON. Estos archivos, en su estado crudo, son primero ingeridos en un 'bucket' de Google Cloud Storage. Tras una primera fase de tratamiento, se trasladan a un segundo 'bucket' donde son transformados en archivos Parquet. Esta transformación no sólo reduce su peso y uso de recursos, sino que también los optimiza, haciéndolos más eficientes para el proceso subsiguiente.

Estos archivos Parquet son sometidos a un segundo proceso mediante Google Cloud Functions, en el cual se lleva a cabo un desanidamiento y una meticulosa distribución de los datos en tablas de consulta, respetando un riguroso modelo de entidad-relación. Esta estructura es esencial para garantizar la accesibilidad y la coherencia de la información.

Nuestro objetivo primordial es asegurar la integridad y pertinencia de la información. Por ello, aplicamos múltiples filtros para:

1. Excluir registros fuera del estado de Florida.
2. Eliminar datos considerados "vacíos" o irrelevantes.
3. Centrarnos exclusivamente en el nicho de "restaurants".

La elección de BigQuery no es casualidad. Nos permite transformar eficientemente grandes volúmenes de información en tablas estructuradas y coherentes. La robustez de esta herramienta asegura a nuestros clientes un acceso a datos de máxima calidad, constantemente actualizados y disponibles 24/7. Con la infraestructura que ofrecemos, nuestros usuarios tienen la libertad de realizar consultas detalladas y visualizar los resultados en múltiples formatos, ya sea mapas, gráficos o tablas, siempre respaldados por la precisión y confiabilidad de nuestras fuentes y el poder de Google Cloud Platform.

KANGAROO

# EDA (ANALISIS DE DATOS EXPLORATORIO)

## EDA GENERAL

El análisis exploratorio de los datos que tenemos lo realizamos en una notebook de Python. En el equipo utilizamos tanto Jupyter Notebook como Visual Studio Code.

El desarrollo del mismo se encuentra en el repositorio de Github, donde se podrá observar los gráficos y los análisis que realizamos buscando qué y cómo utilizar los datos para la puesta a punto en producción de la app.

## GOOGLE/REVIEWS

De todas las carpetas de reseñas del dataset de google, contamos con reseñas en todos los estados de Estados Unidos, en algunos con mayor y con menor cantidad de reseñas, desde las 324.725 de Vermont a las 2.85 M de reseñas en Florida. Totalizando unas 89.045.193 cantidad de reseñas, emitidas por 11.218.423 usuarios.

A su vez, podemos ver que las reseñas comprenden períodos que van desde 1990 a 2021, pero que la gran mayoría de los estados del dataset arrancan por los años 2007/8 y se repite un patrón muy marcado en todos los estados, que en 2018 se genera un crecimiento exponencial en la cantidad de reseñas que los usuarios otorgan, manteniendo una alta cantidad de datos a partir de 2018 en todos los estados en comparación a períodos anteriores, siendo muy pocas las de reseñas anteriores a 2018.

Estas reseñas de los estados tenían 8 columnas, user\_id, name, time, rating, text, pics, resp, gmap\_id (descritas en el diccionario de datos), 3 columnas repiten un patrón que es común a casi todos los estados, que son los nulos y su relación con respecto al total del estado. En cuanto a las "pics", si bien los estados se mantienen con porcentajes altos y parecidos entre ellos, el promedio es de un 97,2%. Con respecto a las "resp" que ocurre lo mismo que "pics", en promedio tienen un total de 87,9% de datos nulos. Por ese alto porcentaje de nulos y datos que son irrelevantes, decidimos eliminar esas columnas.

Se destaca en las nubes de palabras realizadas a todos los estados por separado, la cantidad de texto vinculado a lo que es la comida, el servicio, la calidad del lugar, y elogios. En base a nuestro cliente de Florida, vemos que el dataset dado es acorde ya que posee la mayor cantidad de reseñas del estado, la segunda mayor cantidad de usuarios, la menor cantidad de nulos y la gran mayoría de sus reseñas vienen con texto.

KANGAROO

# EDA

## Analisis de estados

Estado	Cantidad de reseñas	Usuarios	Años	Comentarios	% de nulos pcts	% de nulos resp	% de nulos texto	Reseñas con texto
Alabama	1.800.000	171.379	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,0%	88,5%	46,3%	966.600
Alaska	512.515	20.022	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	96,0%	91,5%	42,8%	293.107
Arizona	2.100.000	332.285	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,0%	83,9%	39,6%	1.267.770
Arkansas	2.400.000	99.053	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,0%	89,4%	46,0%	1.295.760
California	2.700.000	973.518	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	96,3%	90,9%	43,4%	1.529.010
Colorado	2.400.000	284.109	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	96,8%	82,8%	38,4%	1.479.600
Connecticut	2.700.000	93.804	Review mas antigua 2004. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 2004. Se destaca mucho palabras de elogio al servicio por sobre todo, y se destaca en menor cantidad las palabras relacionadas a la comida	97,4%	90,4%	47,2%	1.424.520
Delaware	905.537	34.261	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,9%	89,2%	45,1%	497.230
District of Columbia	564.783	26.021	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	94,8%	92,6%	47,9%	294.478
Florida	2.850.000	903.683	Review mas antigua 2003. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 2003. Se destaca mucho palabras de elogio al servicio por sobre todo, y se destaca en menor cantidad las palabras relacionadas a la comida	96,3%	84,0%	37,9%	1.770.420
Georgia	1.950.000	407.118	Review mas antigua 2004. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 2004. Se destaca mucho palabras de elogio al servicio por sobre todo, y se destaca en menor cantidad las palabras relacionadas a la comida	97,3%	85,2%	41,1%	1.149.525
Hawaii	1.504.347	64.336	Review mas antigua 2002. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 2002. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	92,8%	91,2%	43,3%	852.664
Idaho	2.085.487	72.929	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,6%	85,9%	43,0%	1.189.770
Illinois	2.100.000	401.282	Review mas antigua 2001. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 2001. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	97,2%	87,9%	43,4%	1.187.760
Indiana	2.100.000	249.331	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,9%	87,5%	44,8%	1.158.570
Iowa	2.677.684	94.907	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,2%	90,0%	48,4%	1.382.488
Kansas	1.950.000	106.129	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,0%	88,0%	46,2%	1.049.490
Kentucky	1.650.000	144.963	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,0%	88,2%	45,9%	893.145
Louisiana	1.500.000	138.469	Review mas antigua 2001. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 2001. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	97,6%	89,3%	47,4%	789.150
Maine	1.123.881	41.435	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,1%	91,5%	45,4%	614.201
Maryland	2.400.000	194.773	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	97,4%	87,3%	44,7%	1.326.480
Massachusetts	2.400.000	195.753	Review mas antigua 1999. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1999. Se destaca mucho palabras de elogio al lugar, la comida y al servicio	97,1%	91,0%	47,0%	1.272.000
Michigan	2.250.000	384.538	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,6%	86,5%	42,1%	1.303.875
Minnesota	1.800.000	181.043	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,3%	86,8%	43,3%	1.021.320

KANGAROO

# EDA

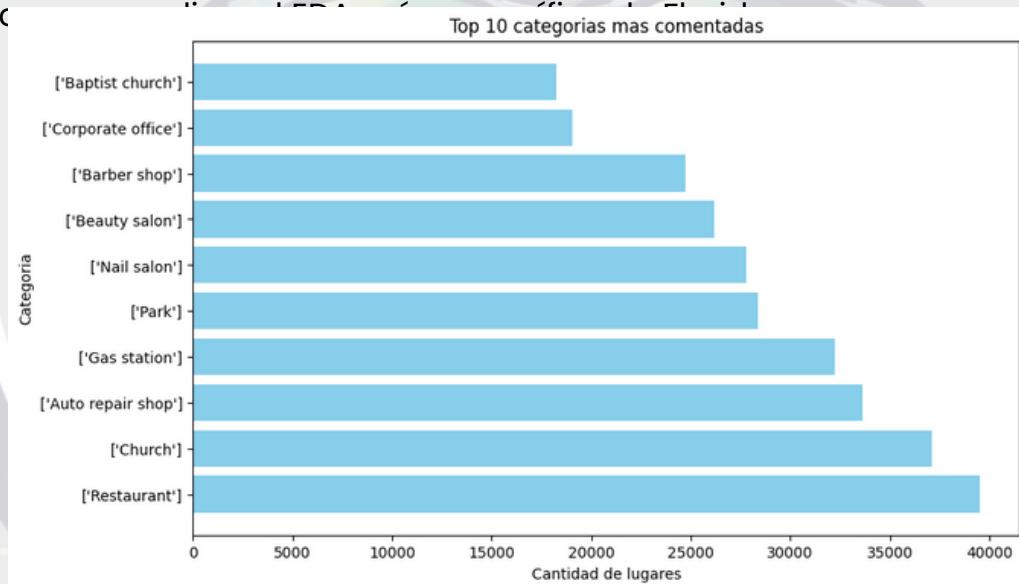
Estado	Cantidad de reseñas	Usuarios	Años	Comentarios	% de nulos pics	% de nulos resp	% de nulos texto	Reseñas con texto
Mississippi	1.971.181	73.386	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,3%	90,5%	48,0%	1.025.211
Missouri	1.650.000	251.135	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,6%	86,0%	42,3%	952.710
Montana	950.370	34.896	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,8%	88,5%	44,1%	531.637
Nebraska	1.817.866	61.675	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,1%	88,5%	47,9%	947.835
Nevada	1.800.000	148.846	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	96,3%	87,5%	40,8%	1.064.880
New Hampshire	1.296.603	49.024	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,5%	89,8%	45,1%	712.094
New Jersey	1.950.000	276.687	Review mas antigua 1999. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1999. Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,1%	88,7%	44,3%	1.085.760
New Mexico	1.650.000	88.388	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,7%	87,8%	43,7%	929.610
New York	2.700.000	567.559	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar , la comida y al servicio	96,2%	90,2%	43,1%	1.536.300
North Carolina	2.250.000	398.660	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,3%	85,1%	41,8%	1.310.400
North Dakota	563.693	21.685	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	98,4%	85,3%	49,0%	287.709
Ohio	1.950.000	412.445	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,7%	85,9%	43,9%	1.094.925
Oklahoma	1.650.000	161.182	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,9%	86,7%	43,3%	935.220
Oregon	2.250.000	203.349	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar , la comida y al servicio	96,6%	88,0%	39,9%	1.352.925
Pennsylvania	2.400.000	402.953	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,6%	87,5%	43,4%	1.358.640
Rhode Island	890.006	32.897	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,5%	91,4%	47,8%	464.227
South Carolina	2.100.000	219.015	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho elogio al servicio y en menor cantidad de reseñas a la comida	97,5%	86,6%	43,4%	1.189.020
South Dakota	673.048	26.037	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,7%	87,6%	48,4%	347.091
Tennessee	1.800.000	285.247	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,6%	85,9%	43,3%	1.020.960
Texas	2.296.824	882.471	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar y al servicio	97,0%	83,7%	40,0%	1.377.635
Utah	1.500.000	159.245	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio al lugar , la comida y al servicio	96,7%	83,2%	37,6%	936.000
Vermont	324.725	13.851	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio a la comida y al servicio	97,0%	91,7%	45,1%	178.177
Virginia	1.050.000	258.253	Review mas antigua 1990. 2018 y años posteriores gran cantidad de reseñas	La review mas antigua es de 1990. Se destaca mucho palabras de elogio al lugar , la comida y al servicio	97,3%	86,9%	42,5%	604.170
Washington	1.942.020	317.932	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio a la comida y al servicio	96,2%	88,5%	40,9%	1.147.928
West Virginia	1.080.333	41.688	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio a la comida y al servicio	96,3%	84,0%	37,9%	671.103
Wisconsin	1.686.482	196.575	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio a la comida y al servicio	97,6%	87,3%	44,4%	937.347
Wyoming	427.808	18.201	2018 y años posteriores gran cantidad de reseñas	Se destaca mucho palabras de elogio a la comida y al servicio	97,5%	89,7%	45,2%	234.653
<b>TOTAL</b>	<b>89.045.193</b>	<b>11.218.423</b>			<b>97,2%</b>	<b>87,9%</b>	<b>43,9%</b>	<b>50.243.101</b>

KANGAROO

# EDA

## GOOGLE/Metadata-sitios

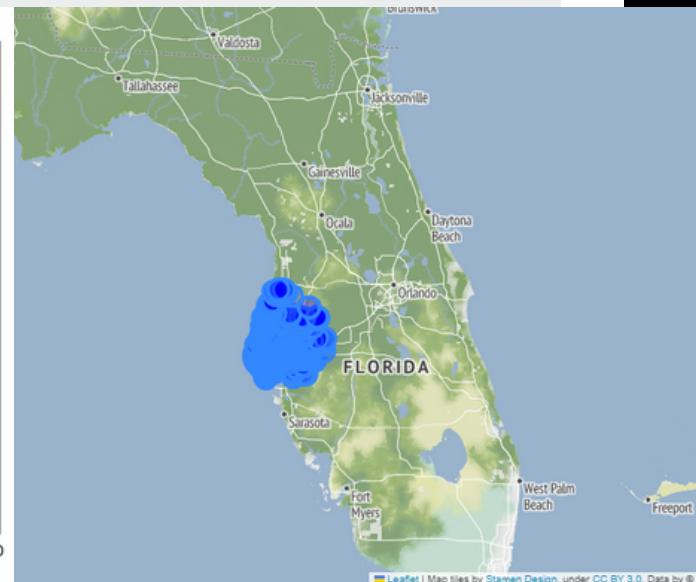
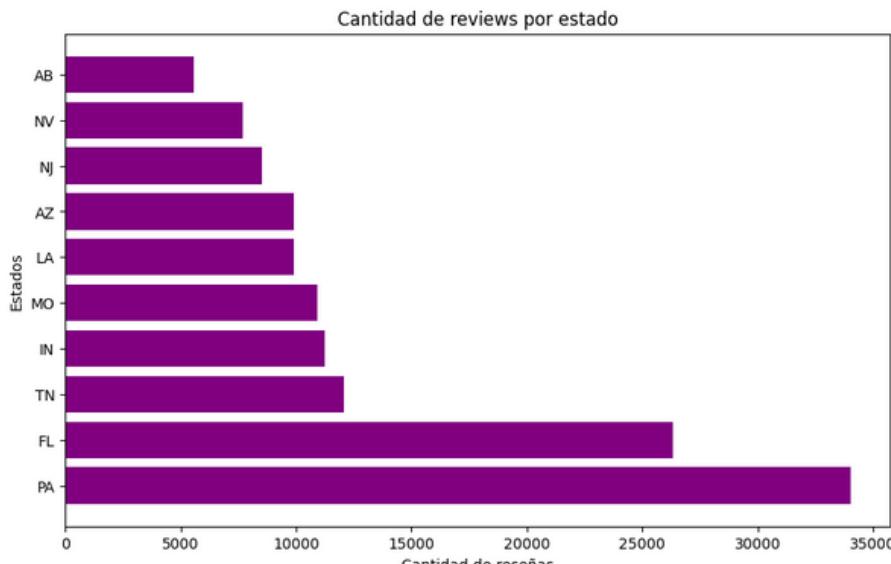
Dentro de la carpeta Google también tenemos una carpeta llamada 'metadata-sitios' que contiene 11 archivos json que contienen información sobre sitios turísticos, localización, reseñas y puntuación de todos los estados de EEUU, para realizar el EDA unimos esos archivos. Descartamos las columnas que contienen más de 80% de valores nulos 'price' y 'description'. Nos quedan aproximadamente 3 millones de valores. Podemos ver que los sitios con mayor cantidad de reseñas van desde parques de diversiones, iglesias, hasta shoppings y restaurantes. Al mirar las categorías con mayor cantidad de sitios con reviews realizadas, la primera es Restaurant. Luego de ese archivo filtramos por categoría restaurant y por latitud y longitud los valores comprendidos para el estado de Florida y lo unimos con los archivos de la carpeta reviews del mismo estado, en un archivo parquet que luego utilizamos.



## YELP

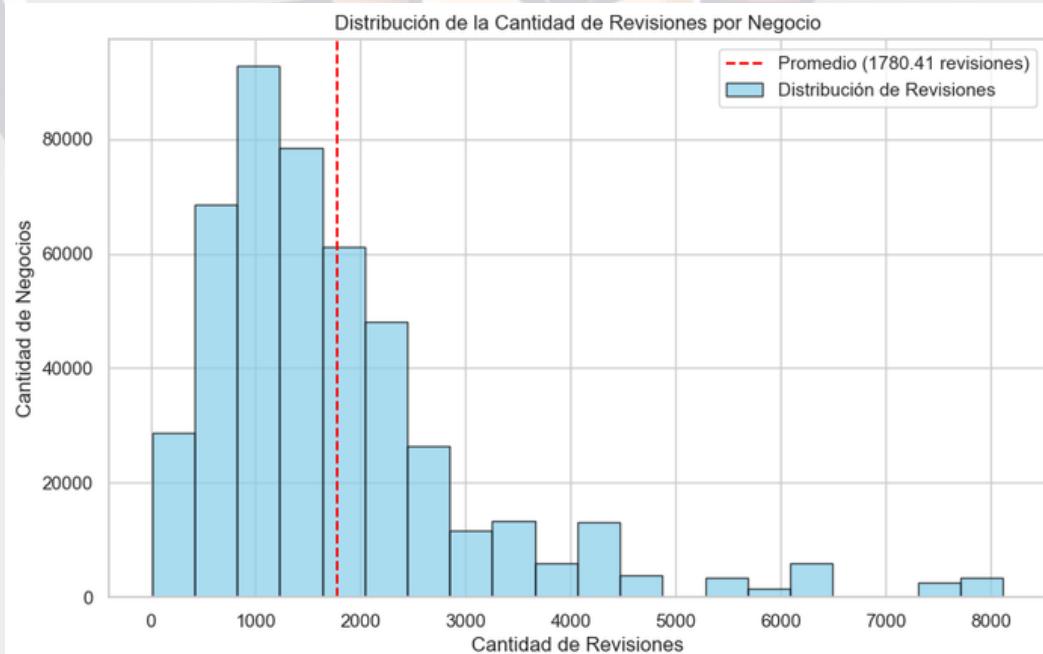
Otra fuente de datos que tenemos es la de YELP. Para analizar los datos de YELP, unimos los 5 archivos comprendidos en la carpeta y como conclusión general, la mayoría de las reviews pertenecen al estado de Florida cuando filtramos los datos por la columna 'state'. Nos quedan aproximadamente 26 mil registros. Luego filtramos por categoría restauran teniendo cerca de 8 mil registros. Pero al afinar el filtrado por geolocalización utilizando longevidad y latitud y un mapa, nos damos cuenta que los lugares que realmente están en Florida son 1577, y todos en un radio de aprox 30km a la ciudad de Tampa. Por este motivo decidimos desestimar los valores que aporta el dataset de YELP, ya que no termina siendo un aporte significativo para nuestra muestra.

# EDA



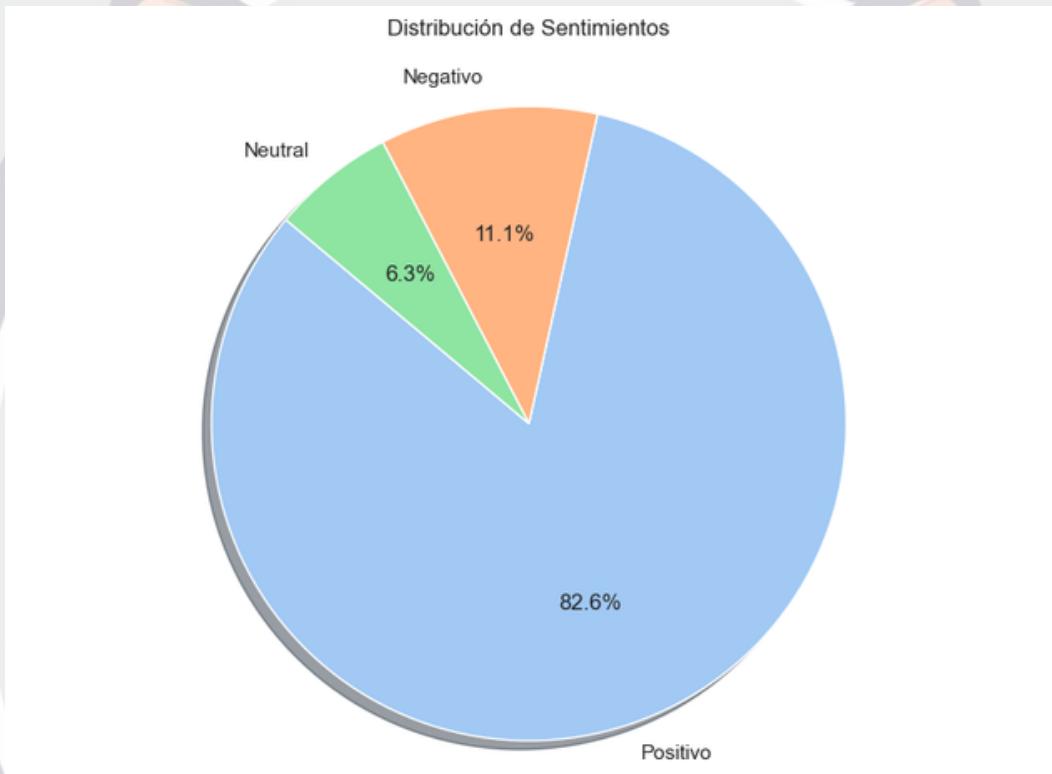
## EDA FLORIDA

El estado de Florida tras el ETL, quedó en unas 563227 reseñas. Analizamos la cantidad de reviews en relación a su distribución en la cantidad de negocios, y nos da que en promedio tenemos 1780 reseñas por negocio, con un máximo de 8000 reviews.

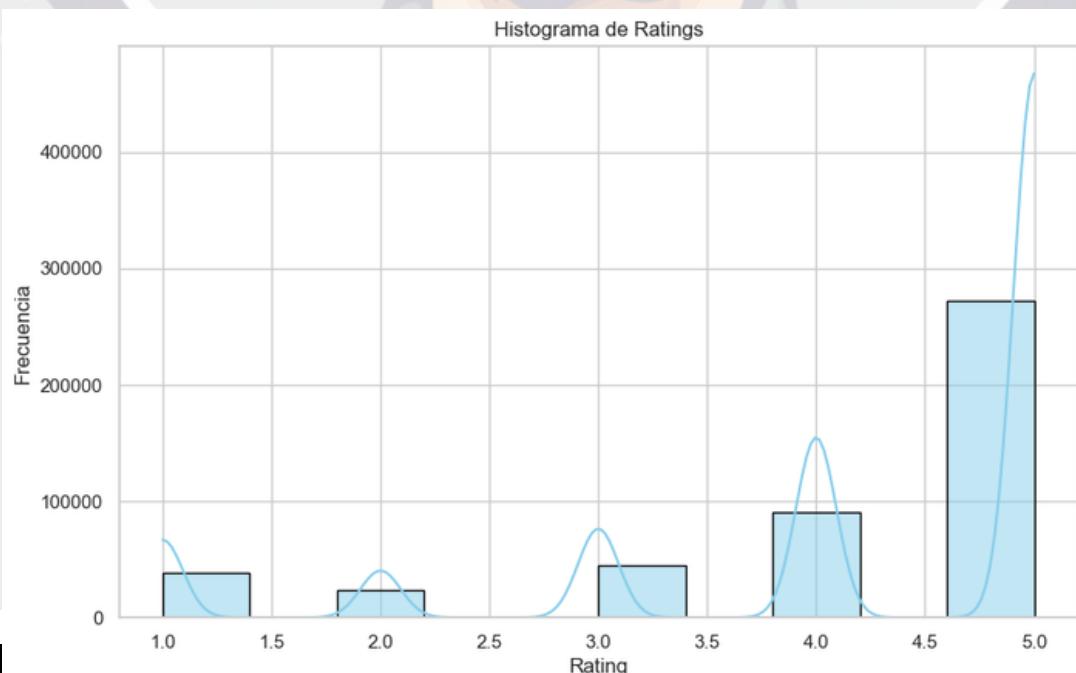


## EDA

Además, teniendo en cuenta el texto de las reseñas, pudimos analizar con TextBlob, en busca de patrones, extraer información y realizar análisis de sentimientos, que según la polaridad definimos tres categorías: si es mayor a 0 se considera positivo, si es igual a 0 neutral y si es menor a 0 negativo.

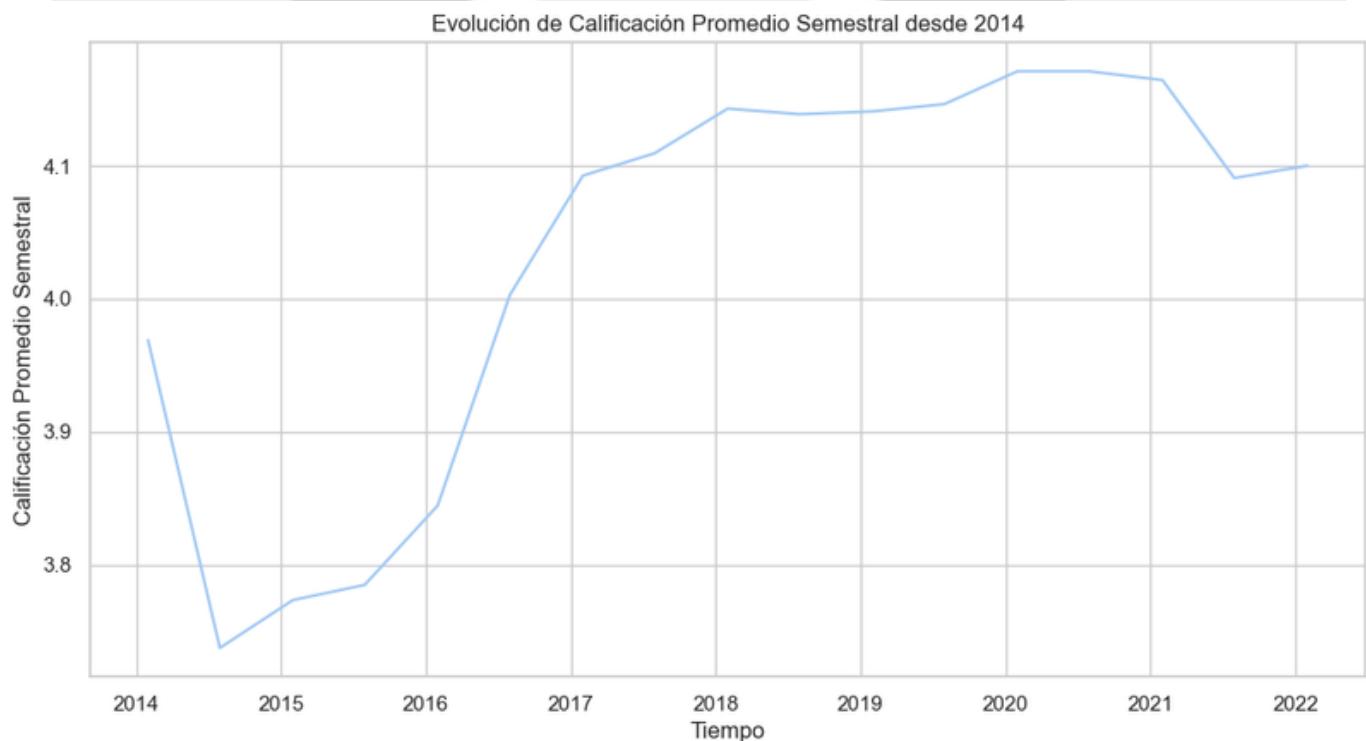


Analizamos la cantidad de reseñas y la distribución de frecuencias de rating de los restaurantes, obteniendo como conclusión que la gran mayoría se encuentran por encima de 3.



## EDA

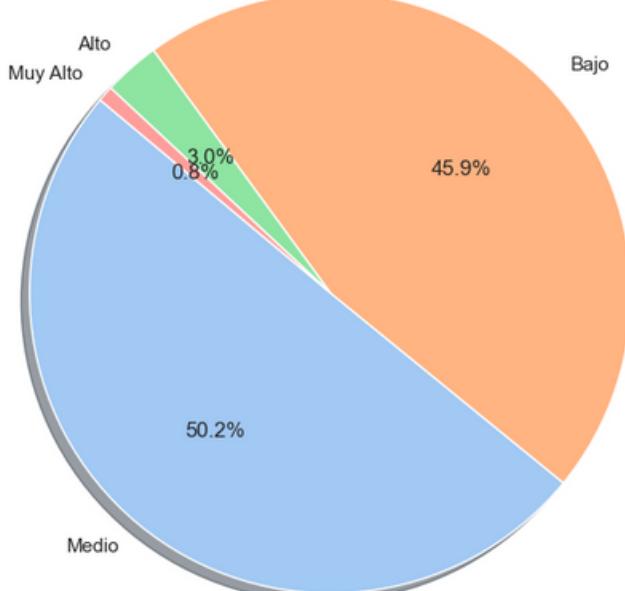
Podemos ver que la cantidad de reseñas aumenta a lo largo del tiempo, creciendo exponencialmente en 201 y a su vez, acompañado como vemos en el siguiente gráfico de aumento en la calificación promedio de los restaurantes.



Dentro de la cantidad de restaurantes que contaban con precio, realizamos una categorización de la siguiente forma:

- \$ → Bajo
- \$\$ → Medio
- \$\$\$ → Alto
- \$\$\$ → Muy alto

Distribución de Restaurantes por Categorías de Precios



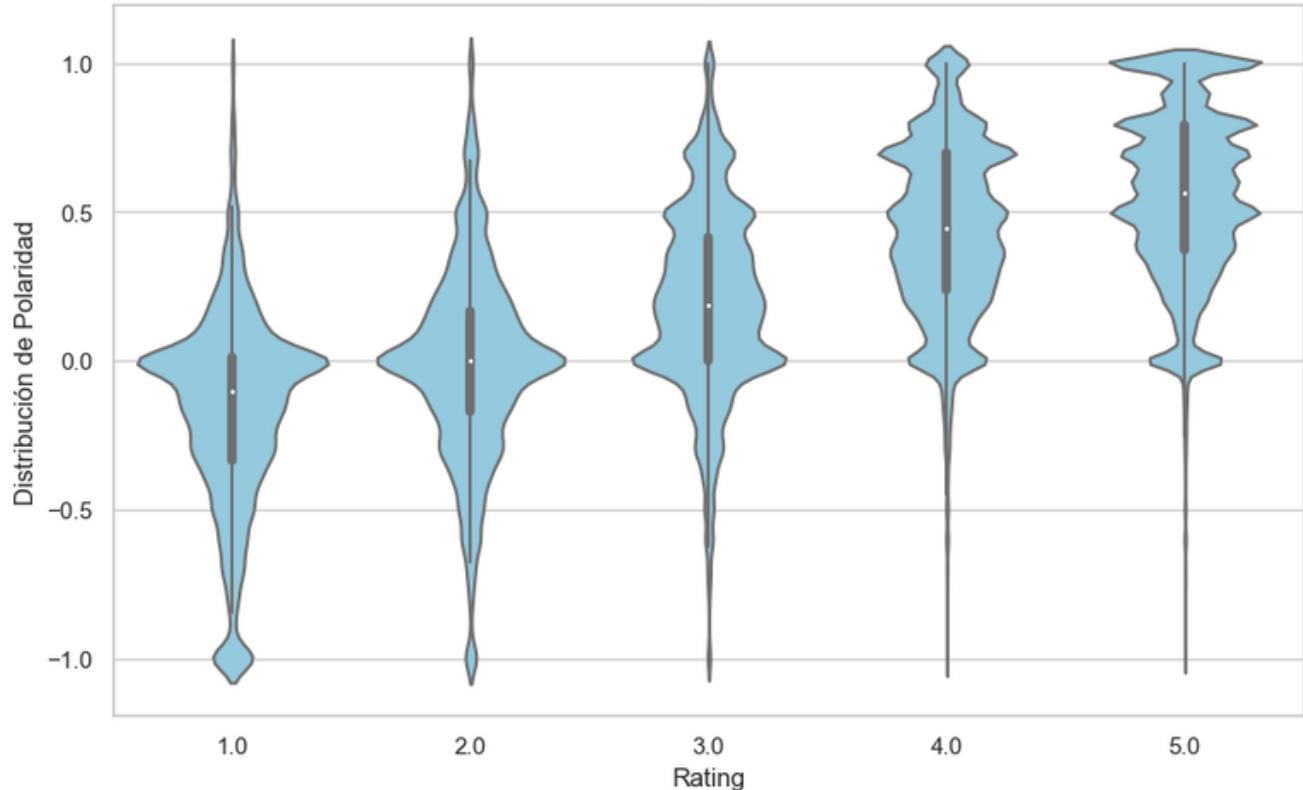
## EDA

Teniendo en cuenta el top 6 de Restaurantes con mayor promedio en las calificaciones, podemos ver una evolución de las mismas a lo largo del tiempo, con una variación similar entre las mismas.

En el siguiente gráfico mostramos la relación que existe entre los sentimientos (polaridad numérica), entre -1 (negativo) y 1 (positivo) y las calificaciones brindadas (reviews). A priori podemos decir que existe una relación lineal.

Distribución de Ratings de acuerdo al Sentimiento - Polaridad

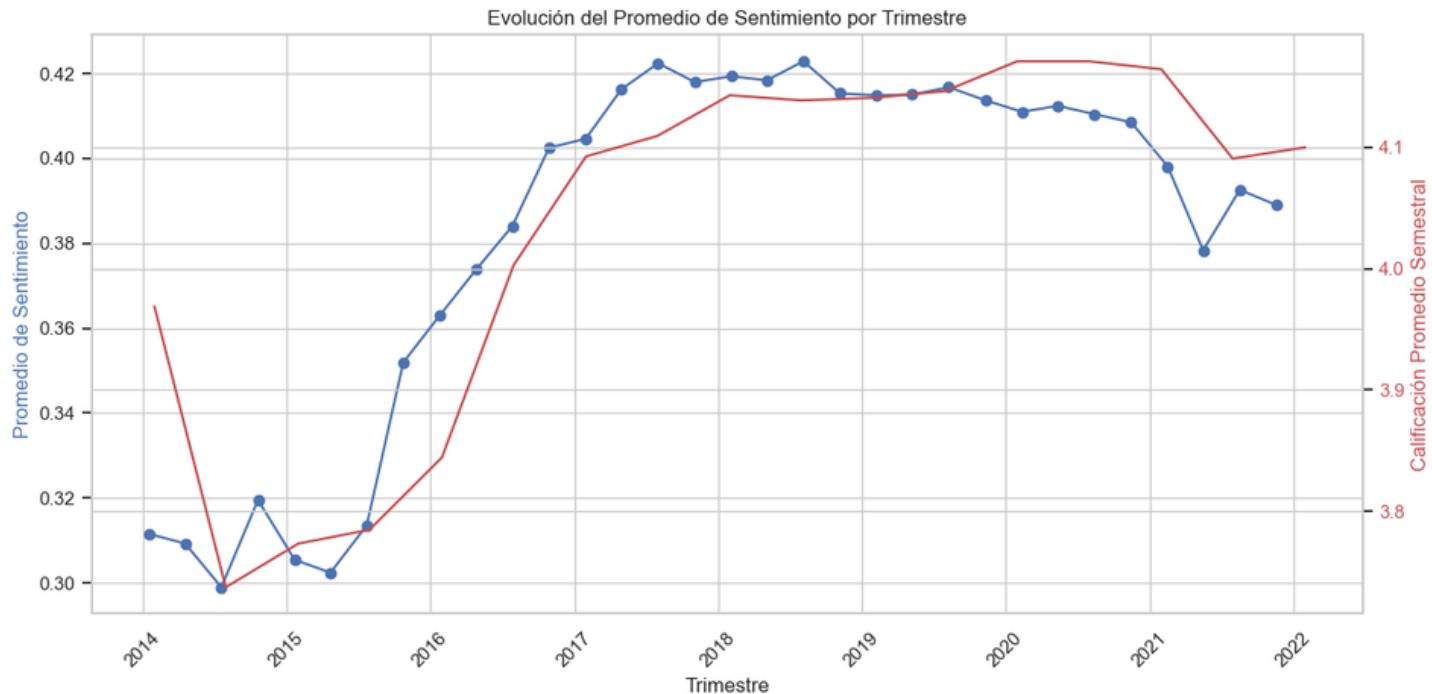
Sentimiento -1 Negativo // 1 Positivo



Observamos cierta relación entre el análisis de sentimiento a través del tiempo y la puntuación de rating otorgada a los usuarios.

KANGAROO

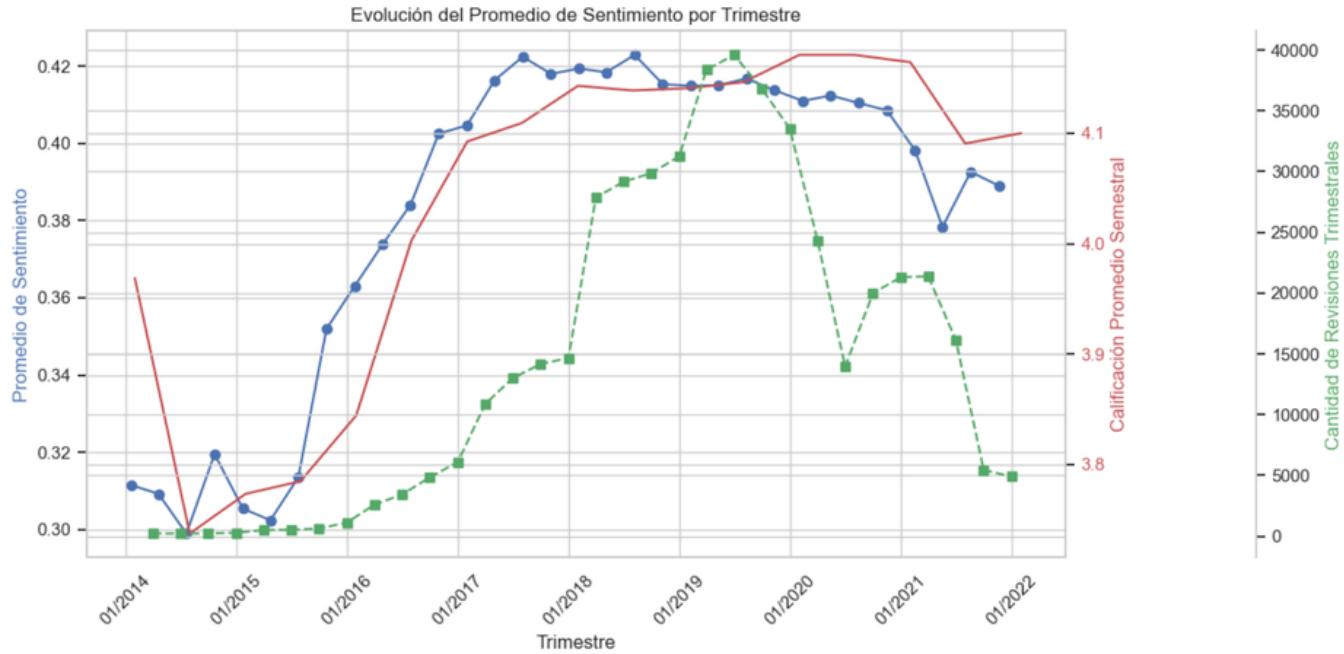
# EDA



Continuando con el análisis anterior, agregamos una nueva variable en la comparación, siendo: (término de tiempo trimestral):

- **Calificación:** es el promedio de calificación brindada por el usuario (estrellas, valoración).
- **Cantidad de Revisiones:** es la cantidad de revisiones que dieron dentro del período
- **Promedio Sentimiento:** utilizando de la librería TextBlob conocemos la polaridad de las reseñas en texto, siendo la escala entre -1 para negativo y 1 para positivo.

# EDA



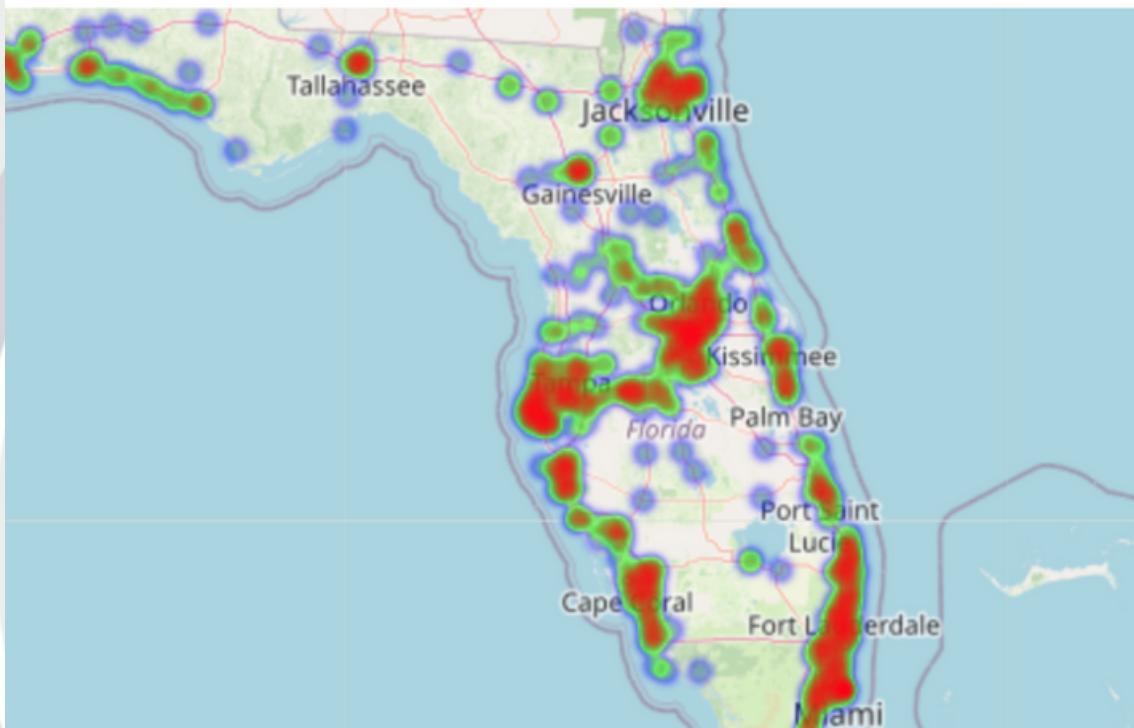
Del gráfico anterior podemos concluir que existe relación directa entre lo que un usuario informa y describe en texto y la puntuación que le dan en su experiencia. Nos parece importante haber encontrado dicha relación, lo que nos permite poder manejar los datos de una manera cercana sobre ambas variables. Se observa que existe una pequeña disminución en las calificaciones a partir del 2021, lo que también se ve agregado en la cantidad de revisiones que los usuarios otorgaron

A modo gráfico, compartimos aquí la acumulación de revisiones que los usuarios han brindado dentro del estado de Florida y el mapa con la cantidad de restaurantes. Podemos ver que se acumulan dentro de las ciudades, en relación al tamaño y cantidad de restaurantes que existen. Existe una relación lineal entre las variables restaurantes y reseñas.

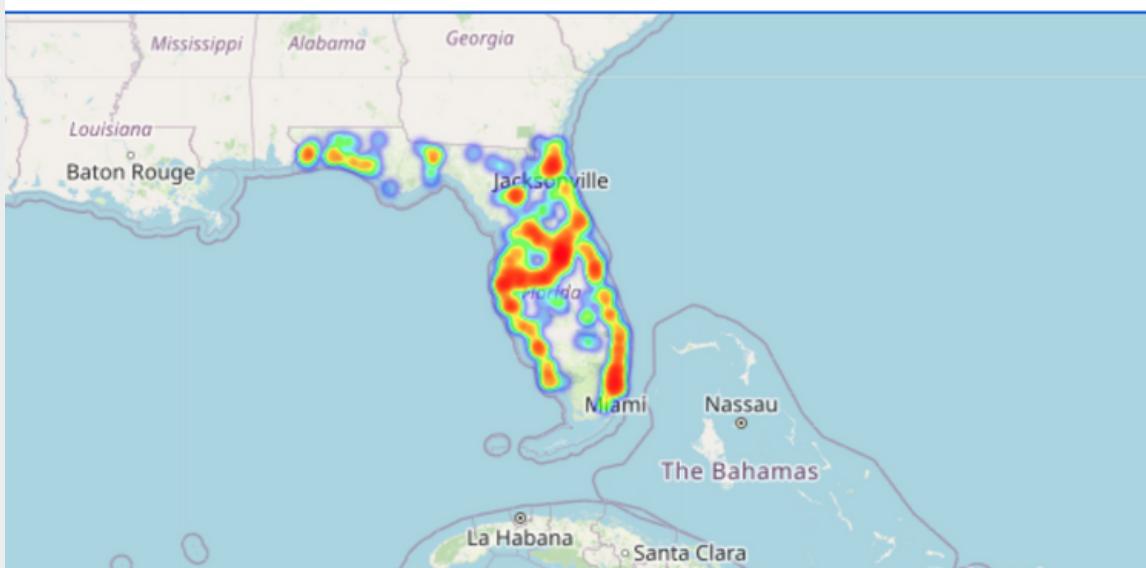
KANGAROO

# EDA

## Acumulación de Restaurantes



## Acumulación de Reviews



# CREACIÓN DEL DATAWAREHOUSE (Diagrama Entidad – Relación)

Una vez que ya contamos con un pipeline automatizado para la ingesta de datos a través de Google cloud en el cual un archivo es ingresado al flujo y procesado de datos, mediante un proceso de ETL que garantiza su disponibilización en un dato de útil y de calidad.

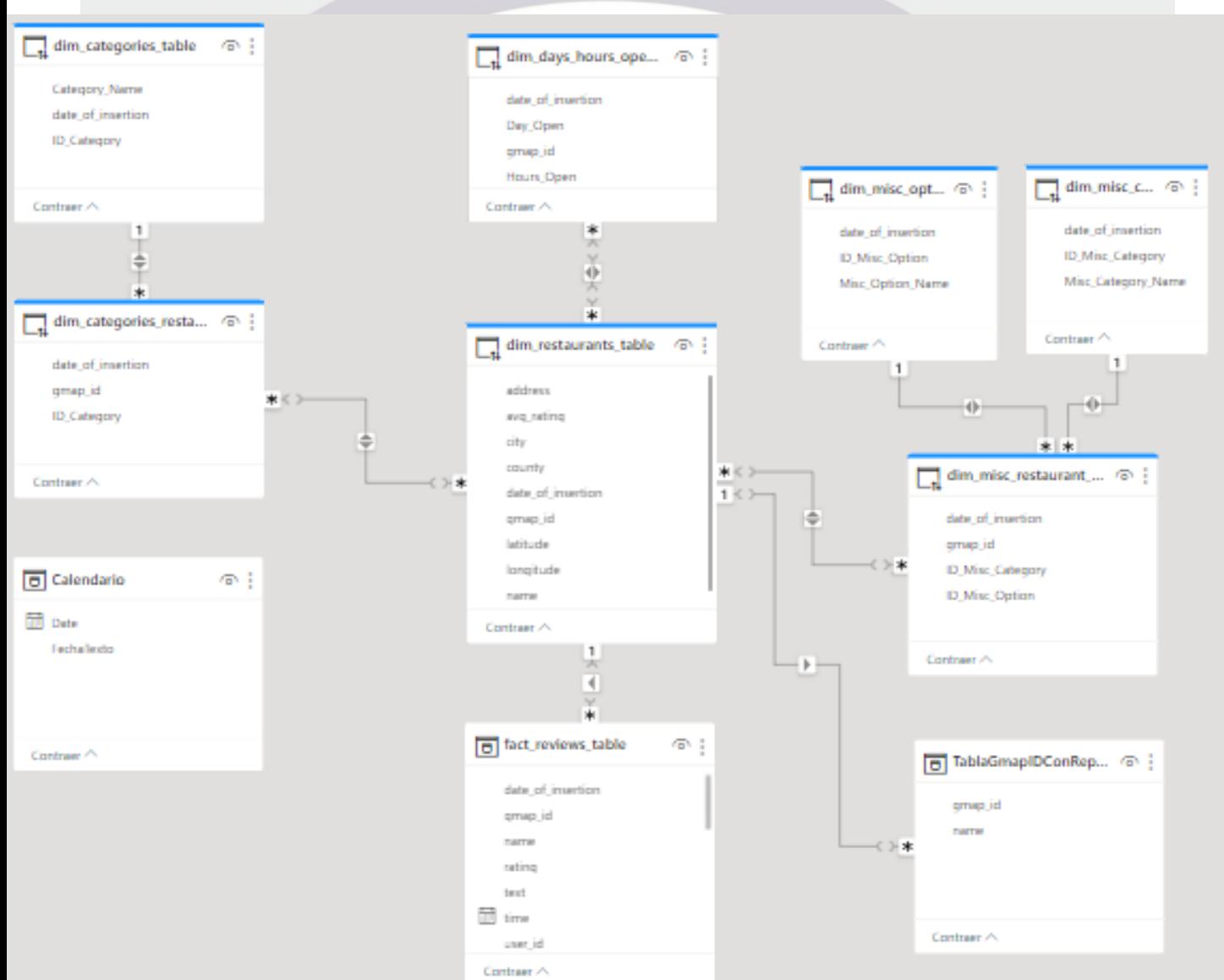
Los datos procesados se almacenan en un DataWarehouse y, al mismo tiempo, se ponen a disposición en el motor de datos SQL de BigQuery.

Luego, se realizó la conexión con Power BI para poder desarrollar nuestro Dashboard interactivo, con el siguiente esquema:

Las tablas dentro de este Dashboard, todas cuentan con un campo “date\_of\_insertion”, qué es la fecha en la cual estos datos fueron ingresados post procesado. Las tablas son las siguientes:

- fact\_reviews\_table → Tabla de hechos de las reviews que realizan los usuarios.
- dim\_restaurants\_table → Tabla de dimensiones con información de los restaurantes.
- dim\_categories\_restaurant\_table → Tabla intermedia con id del restaurante y id de la categoría del restaurante.
- dim\_categories\_table → Tabla de dimensiones con el id de la categoría y su nombre asociado.
- dim\_days\_hours\_open\_restaurant\_table → Tabla con los días y horarios en los cuales el restaurant se encuentra abierto.
- dim\_misc\_restaurant\_category\_options\_table → Tabla intermedia con los gmap\_id, el id de la categoría miscelánea y el id de la opción miscelánea
- dim\_misc\_options\_table → Tabla de dimensiones con el id de la opción miscelánea y su nombre asociado.
- dim\_misc\_categories\_table → Tabla de dimensiones con el id de la categoría miscelánea y su nombre asociado.
- Calendario → Tabla de calendario auto generada contemplando las fechas de las reviews

# DATAWAREHOUSE



KANGAROO

# PIPELINE AUTOMATIZADO

Hemos logrado la automatización de nuestro Data Warehouse y la carga incremental utilizando la potencia de Google Cloud. Iniciamos configurando dos buckets en Cloud Storage: "data\_lake\_kanguro" y "data\_procesada\_kanguro". En el primero, subimos nuestros datos manualmente, mientras que en el segundo implementamos Cloud Functions para simplificar el proceso.

## **Creamos dos funciones:**

- ETL\_bucket\_a\_bucket\_limpio: Esta función se activa automáticamente cada vez que se añaden nuevos archivos a "data\_lake\_kanguro". Su tarea principal es ejecutar nuestro proceso ETL para procesar los datos recién cargados y, a continuación, trasladarlos al bucket "data\_procesada\_kanguro".
- De\_bucket\_limpio\_a\_BigQuery: La segunda función entra en acción cuando se carga un archivo en el bucket "data\_procesada\_kanguro". Su propósito es tomar los datos procesados y cargarlos de manera eficiente en BigQuery, asegurando que nuestra base de datos esté siempre actualizada con la información más reciente.

Esta automatización garantiza que nuestros datos fluyan desde su origen en "data\_lake\_kanguro" hasta BigQuery, lo que simplifica la gestión de nuestro Data Warehouse y permite un análisis en tiempo real de los datos procesados. Además, al eliminar la intervención manual, hemos mejorado significativamente la eficiencia de nuestro proceso de datos.

# DASHBOARD



## MACHINE LEARNING

### Modelo de recomendación basado en ubicación y tendencia

Nuestro sistema de recomendación es una solución integral que utiliza la ubicación geográfica del usuario para proporcionar recomendaciones de restaurantes altamente personalizadas. No nos limitamos simplemente a ofrecer opciones basadas en la proximidad; también destacamos los restaurantes en tendencia.

El núcleo de nuestro sistema radica en la capacidad de considerar la ubicación actual del usuario, lo que nos permite presentar restaurantes cercanos que son convenientes y accesibles. Pero no nos detenemos ahí. Además, evaluamos las tendencias culinarias y gastronómicas en tiempo real para destacar aquellos restaurantes que están ganando popularidad y generando interés. Ello se obtiene a través del análisis de los textos de las reseñas y su evolución a través del tiempo. Luego, con tecnología de análisis de tiempo e inferencias sobre el comportamiento de este ratio en el futuro, logramos seleccionar aquellos restaurantes que vienen dando que hablar.

Este enfoque combina la conveniencia de encontrar restaurantes cercanos con la emoción de descubrir lugares de moda. Así permitiremos al usuario poder seleccionar los mejores lugares para poder disfrutar de una experiencia única, junto al tercer modelo presentado a continuación.

KANGAROO

# MACHINE LEARNING

## Modelo de recomendación basado en contenido

Nuestro modelo de recomendación tiene como objetivo proporcionar recomendaciones de restaurantes similares basadas en múltiples dimensiones clave, incluyendo categorías, atmósfera, ofertas y aspectos destacados. Para lograr este objetivo, hemos implementado un enfoque de recomendación basado en contenido que aprovecha técnicas avanzadas de procesamiento de datos.

Al considerar estas dimensiones, nuestro modelo es capaz de sugerir restaurantes que se ajusten no solo al tipo de cocina deseado, sino también a la experiencia general que el usuario busca, incluyendo factores como el ambiente, las opciones culinarias y los aspectos más destacados del restaurante. Sin embargo, para hacer esto de manera efectiva, utilizamos dos técnicas clave: TF-IDF (Term Frequency-Inverse Document Frequency) y la similitud del coseno.

El TF-IDF es una técnica de procesamiento de lenguaje natural que nos permite evaluar la importancia de las palabras clave en los perfiles de los restaurantes. Este método asigna un peso a cada palabra en función de cuán a menudo aparece en el perfil de un restaurante en particular y cuán rara es en el conjunto completo de perfiles de restaurantes. Esto nos permite identificar características distintivas y relevantes de cada restaurante, como su tipo de cocina, atmósfera y aspectos destacados.

La similitud del coseno, por su parte, es una métrica fundamental que utilizamos para medir cuán similares son dos perfiles de restaurantes. Esto nos permite comparar cuán estrechamente se alinean los restaurantes con las preferencias de un usuario en términos de categorías, atmósfera y otras características clave. Cuanto más cercano a 1 sea el valor del coseno, más parecidos son los perfiles, lo que indica que el restaurante coincide estrechamente con las preferencias del usuario.

En resumen, al combinar TF-IDF y la similitud del coseno, nuestro enfoque de recomendación basado en contenido es capaz de evaluar y comparar las características de los restaurantes con las preferencias de los usuarios. Esto nos permite generar recomendaciones personalizadas y precisas que mejoran significativamente la experiencia de descubrimiento gastronómico de nuestros usuarios, permitiéndoles explorar una amplia gama de opciones que se ajustan de manera más precisa a sus preferencias individuales.

# MACHINE LEARNING

## Modelo GPT-3 (Generative Pre-trained Transformer 3)

Es el modelo de machine learning que utiliza Chat GPT, específicamente en GPT-3.5. GPT-3.5 es una versión de la serie GPT desarrollada por OpenAI. Estos modelos son modelos de lenguaje generativo pre-entrenados en una amplia variedad de texto en línea y pueden generar texto coherente y contextualmente relevante en respuesta a preguntas o solicitudes de los usuarios.

GPT-3.5 es una versión mejorada y más avanzada de GPT-3, con un conocimiento y capacidad de generación de lenguaje más profundos.

Nuestro modelo incorpora mediante una API este modelo de openAI utilizando la información que se obtiene al ingresar la dirección de los usuarios.

## REPOSITORIO GITHUB

Creamos el siguiente repositorio de Github para trabajar en equipo y poder estar conectados: [https://github.com/Constanzafl/Proyecto\\_Final](https://github.com/Constanzafl/Proyecto_Final)

Y además para visualizar bien el trabajo y ordenarnos lo asociamos a Git Kraken

KANGAROO

# EQUIPO KANGURO



**Martin Peñas**

*Data Science*

[✉](#) [in](#)



**Constanza Florio**

*Data Engineer*

[✉](#) [in](#)



**Joaquin Millan**

*Data Analyst*

[✉](#) [in](#)



**Fausto Ezquerra**

*Data Analyst*

[✉](#)



**Nicolas Yapur**

*Data Engineer*

[✉](#) [in](#)

KANGURO