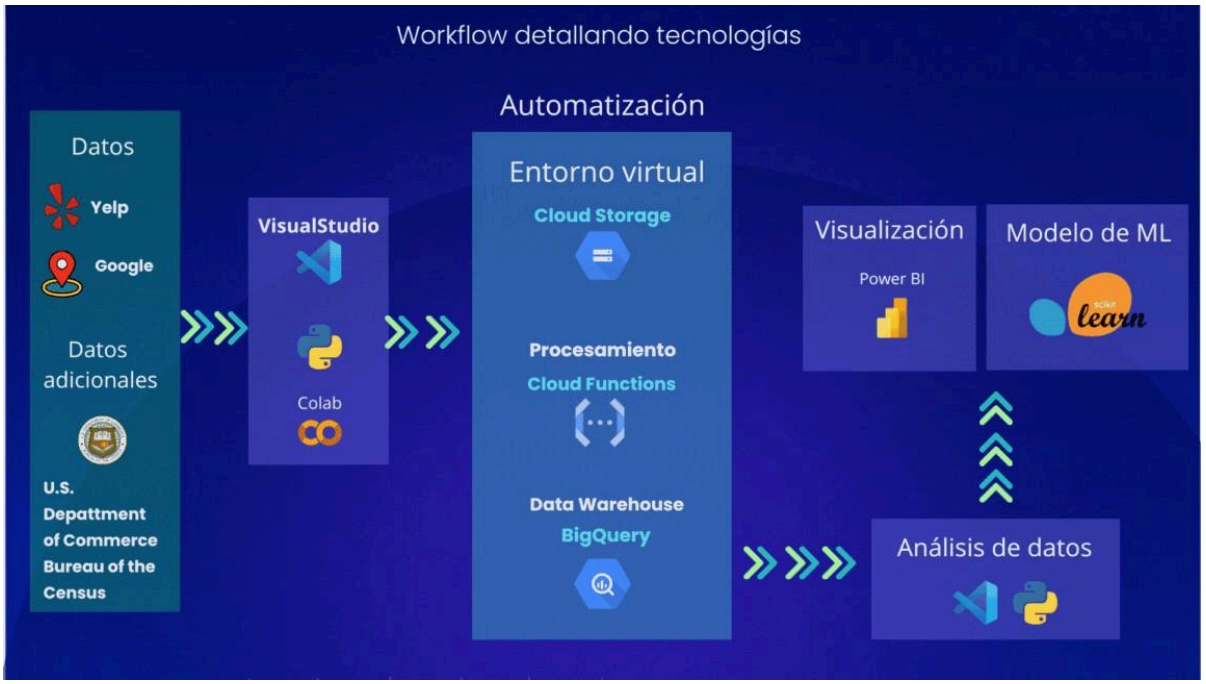
Sprint 2

Información detallada

Avances en el segundo sprint de nuestro proyecto. En esta fase, nos enfocamos en establecer la infraestructura que respalde todo el proceso de análisis y visualización de datos.

Estructura de datos implementada:



Automatización:

- Implementamos un proceso de automatización integral para gestionar la carga, transformación y almacenamiento de datos de manera eficiente.

Pipeline ETL automatizado:

- Nos encontramos desarrollando los pipelines de ETL automatizados utilizando Google Functions, lo que permite realizar el procesamiento de datos de forma ágil y escalable.

Pipelines para alimentar el DW:

- Configuración de pipelines específicos para alimentar nuestro Data Warehouse (DW) en BigQuery con datos limpios y estructurados, listos para su análisis y generación de modelos de Machine Learning.

Data Warehouse:

- Implementamos un Data Warehouse en BigQuery de Google Cloud Platform (GCP), que actúa como el núcleo central para el almacenamiento y la gestión de nuestros datos procesados.

Workflow detallando tecnologías:

- Nuestro workflow se basa en tecnologías de GCP, incluyendo Google Storage para la gestión de datos en crudo, Google Functions para la automatización del ETL, y BigQuery para el almacenamiento y análisis de datos a gran escala.

Es importante destacar que también realizamos ETLs preliminares de los datos originales como parte de nuestra estrategia de preparación y limpieza de datos. Estos procesos preliminares pueden consultarse en detalle en nuestro repositorio de GitHub.

Documentación y validación de datos:

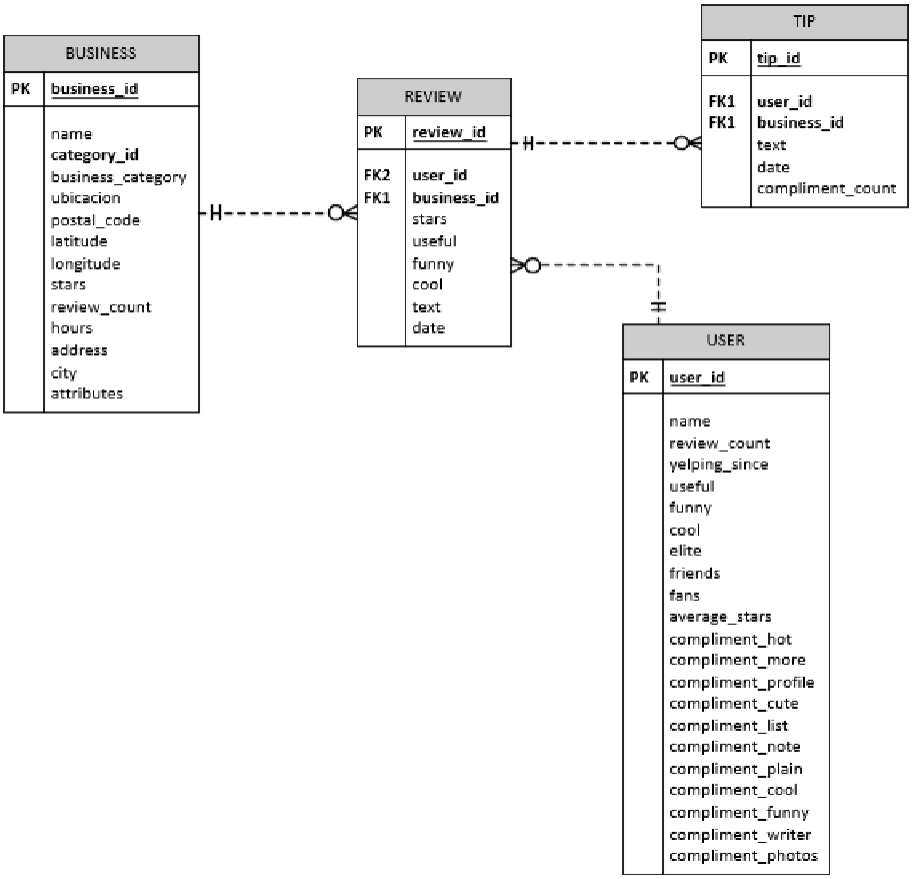
ETL Completo:

- Hemos realizado el proceso completo de Extracción, Transformación y Carga (ETL) de los datasets relevantes para nuestro proyecto. Esto incluye los datos originales de Yelp y Google Maps, así como los datasets adicionales del Censo de Estados Unidos que hemos integrado para enriquecer nuestro análisis. Estos documentos pueden consultarse en el siguiente link de Github.

Modelo ER:

- Se ha desarrollado un Modelo Entidad-Relación (ER) detallado para representar la estructura y las relaciones de las tablas en nuestra base de datos. Este modelo es fundamental para comprender la organización de los datos y las interconexiones entre las entidades clave.

Modelo Entidad-Relación YELP



El modelo de datos consta de cuatro tablas: business, review, user, y tip. A continuación, una descripción detallada de cada tabla y sus relaciones:

1. BUSINESS

COLUMNA	DESCRIPCIÓN	TIPO DE DATO
business_id (PK)	Identificación única	OBJECT (str)
category_id(FK)	Id de categoría	OBJECT (str)
name	Nombre	OBJECT (str)
business_category	Nombre de la categoría	OBJECT (str)
ubicacion	Identifica si pertenece a Florida	BOLEANO
postal_code	codigo postal (ubicación)	FLOAT
latitude	Coordenada de latitud.	FLOAT
longitude	Coordenada de longitud.	FLOAT
stars	Calificación promedio de 1 - 5	FLOAT
review_count	Número de reseñas	INTEGER

hours	Hora a la que se realizó esa calificación.	DATETIME64
address	Dirección	OBJECT (str)
city	cuidad (ubicación)	OBJECT (str)
attributes	Diccionario con los atributos que ofrece	OBJECT (str)

2. REVIEW

COLUMNA	DESCRIPCIÓN	TIPO DE DATO
review_id (PK)	Id de la reseña	OBJECT (str)
user_id (FK)	id del usuario que realizó la reseña	OBJECT (str)
business_id (FK)	id de la empresa a la que se le realizó la reseña	OBJECT (str)
stars	valor de calificación del 1-5	FLOAT
useful	número de veces que fue calificada como 'useful' por otros usuarios	INTEGER
funny	número de veces que fue calificada como 'funny' por otros usuarios	INTEGER
cool	número de veces que fue calificada como 'cool' por otros usuarios	INTEGER
text	El texto de la reseña.	OBJECT (str)
date	Fecha de la reseña	DATETIME64

3. TIP

COLUMNA	DESCRIPCIÓN	TIPO DE DATO
id_tip (PK)	Id del tip	INTEGER
user_id (FK)	Id del usuario	OBJECT (str)
business_id (FK)	Id de la empresa	OBJECT (str)
text	El texto de la reseña.	OBJECT (str)
date	Fecha de la reseña	DATETIME64
compliment_count	Cantidad de elogios o cumplidos recibidos en todas las reseñas	INTEGER

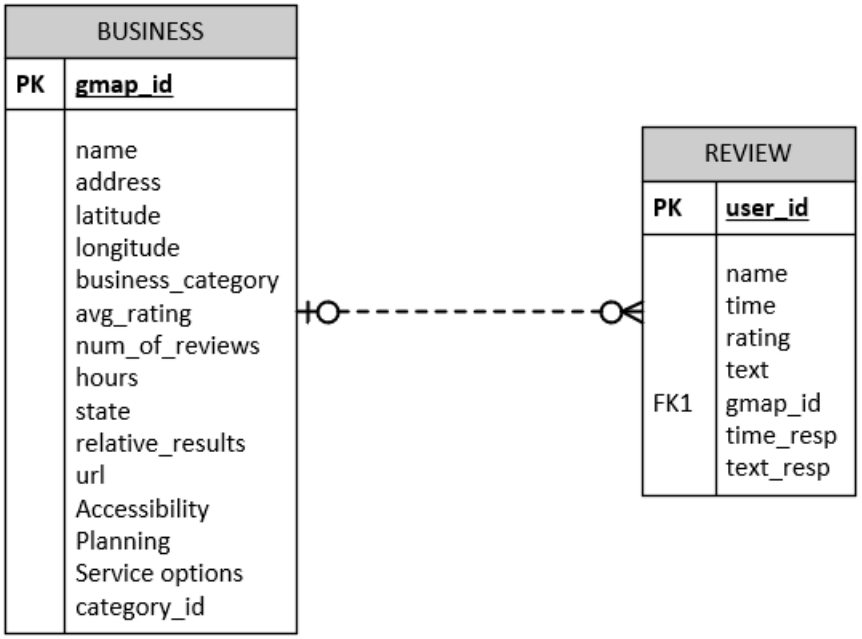
4. USER

COLUMNA	DESCRIPCIÓN	TIPO DE DATO
user_id (PK)	Id del usuario	OBJECT (str)
name	Nombre	OBJECT (str)
review_count	El número total de reseñas	INTEGER
yelping_since	Fecha en la que el usuario se unió a Yelp formato AAAA-MM-DD HH:MM:SS	DATETIME64
useful	Cantidad total de votos " useful " que el usuario ha recibido	INTEGER
funny	Cantidad total de votos " funny " que el usuario ha recibido	INTEGER
cool	cantidad total de votos " cool " que el usuario ha recibido	INTEGER
elite	Usuario ha sido miembro de Yelp Elite	OBJECT (str)
friends	User_id de amigos del usuario	OBJECT (str)
fans	Cantidad de seguidores que el usuario	INTEGER
average_stars	Promedio de estrellas que el usuario ha dado en sus reseñas	FLOAT
compliment_hot	Cantidad de cumplidos que el usuario ha recibido de acuerdo a cada item.	INTEGER
compliment_more		INTEGER
compliment_profile		INTEGER
compliment_cute		INTEGER
compliment_list		INTEGER
compliment_note		INTEGER
compliment_plain		INTEGER
compliment_cool		INTEGER
compliment_funny		INTEGER
compliment_writer		INTEGER
compliment_photos		INTEGER

Relaciones

- **BUSINESS A REVIEW**: Relación uno a muchos. Un negocio puede tener muchas reseñas.
- **USER A REVIEW**: Relación uno a muchos. Un usuario puede hacer muchas reseñas.
- **BUSINESS A TIP**: Relación uno a muchos. Un negocio puede recibir muchos tips.
- **USER A TIP**: Relación uno a muchos. Un usuario puede dejar muchos tips.

Modelo Entidad-Relación GOOGLE MAPS



El modelo de datos consta de dos tablas: business, review. A continuación, una descripción detallada de cada tabla y sus relaciones:

1. BUSINESS

COLUMNA	DESCRIPCIÓN	TIPO DE DATOS
gmap_id (PK)	ID de Google Maps.	OBJECT (str)
Category_id (FK)	Ide de categoria	OBJECT (str)
address	Dirección.	OBJECT (str)
latitude	Coordenada de latitud.	FLOAT
longitude	Coordenada de longitud.	FLOAT
business_category	Nombre de la categoría	OBJECT (str)
avg_rating	Calificación promedio.	FLOAT
num_of_reviews	Número de reseñas.	INTEGER
hours	Horario de atención. (apertura y cierre)	DATATIME64
state	Estado actual al subir la información (abierto, cerrado, etc.).	OBJECT (str)
relative_results	Id de google maps de otros negocios relacionados	OBJECT (str)
url	URL de Google Maps.	OBJECT (str)

Accessibility	Información adicional: Accesibilidad para personas con discapacidad. (accesible para sillas de ruedas)	OBJECT (str)
Planning	Información adicional: Planificación requerida para visitar. (Visita rápida)	OBJECT (str)
Service options	Información adicional: Opciones de servicio disponibles. (recojo tienda, autoservicio, compra tienda)	OBJECT (str)
name	Nombre.	OBJECT (str)

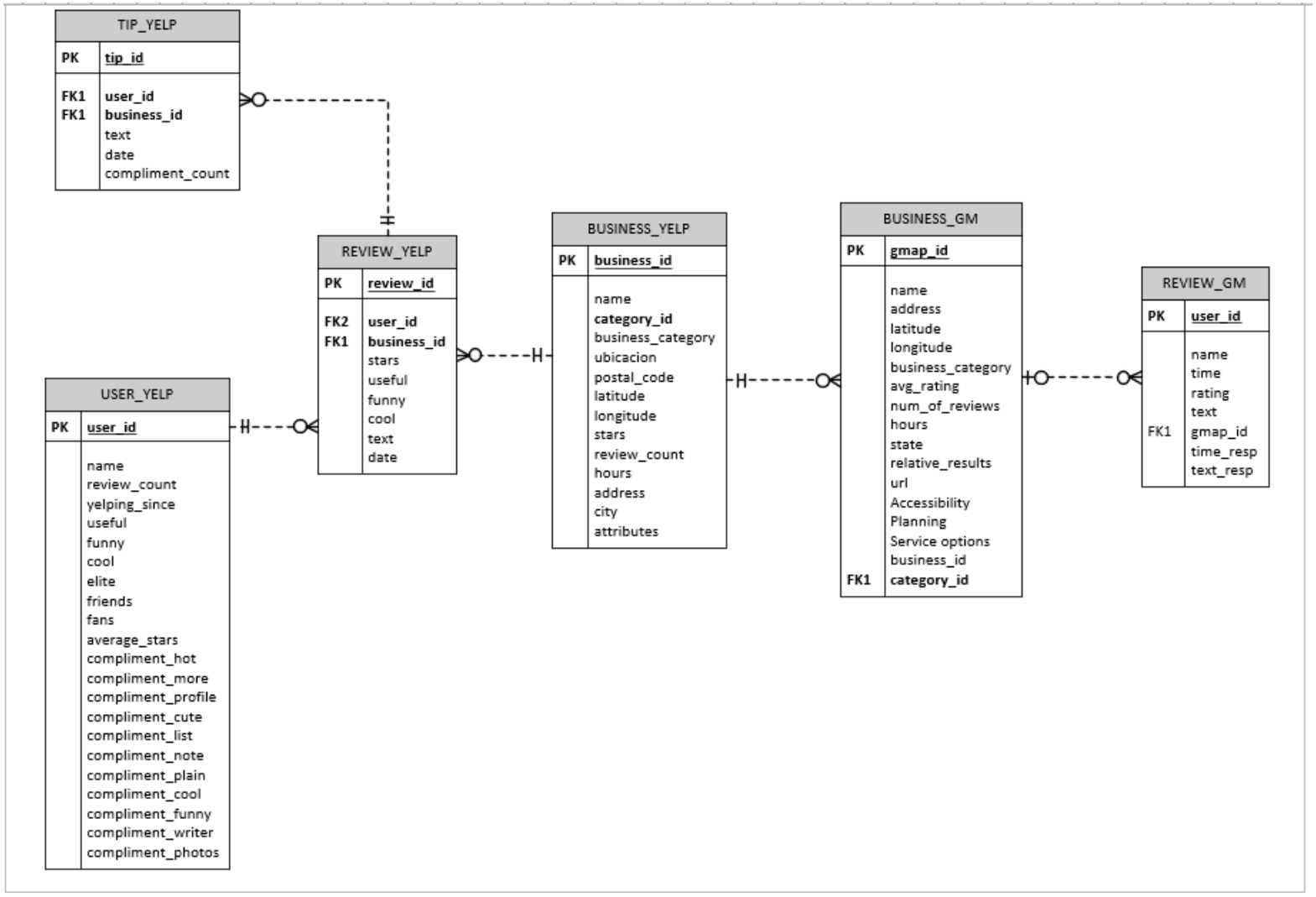
2. REVIEWS

COLUMNA	DESCRIPCIÓN	TIPO DE DATOS
gmap_id (FK)	Identificación de la ubicación en Google Maps.	OBJECT (str)
user_id	Id de usuario	OBJECT (str)
name	Nombre del usuario.	OBJECT (str)
time	Tiempo en formato AAAA-MM-DD HH-MM-SS.	DATATIME64
rating	Calificación otorgada.	INTEGER
text	El texto de la reseña.	OBJECT (str)
time_resp	Tiempo en formato AAAA-MM-DD HH-MM-SS, de la respuesta de la empresa hacia la reseña.	DATATIME64
text_resp	El texto de la respuesta de la empresa hacia la reseña.	OBJECT (str)

Relaciones

- **BUSINESS A REVIEW**: Relación uno a muchos. Un negocio puede tener muchas reseñas.

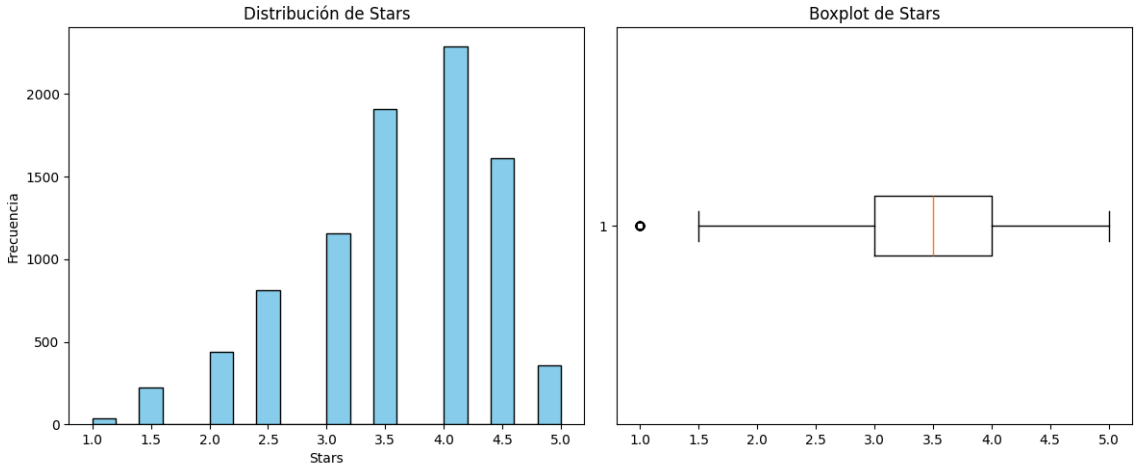
DIAGRAMA ER YELP – GOOGLE MAPS



Análisis de Dispersión de Datos YELP

Análisis de la dispersión de los datos en las columnas stars y review_count de la tabla BUSINESS de YELP. El objetivo principal es comprender la distribución y la presencia de outliers en estas dos variables clave.

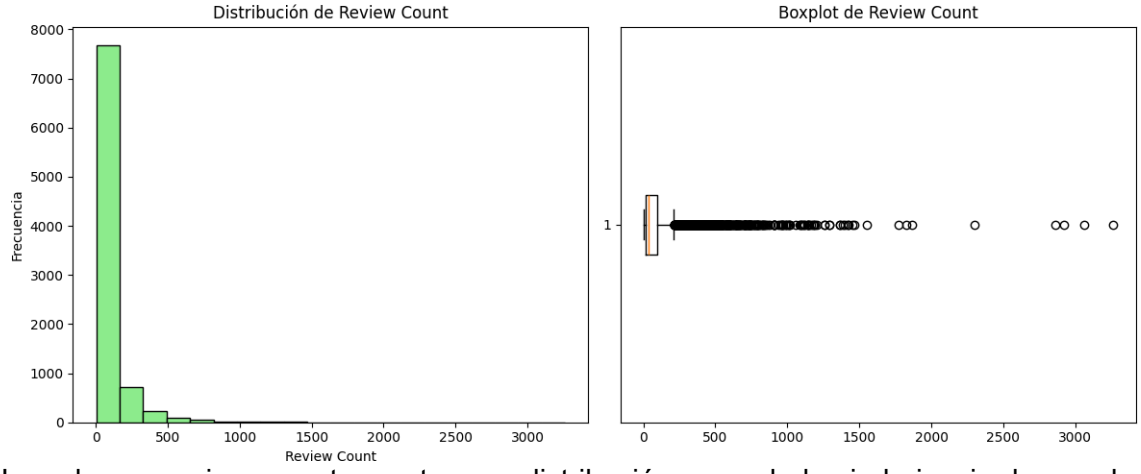
Distribución de Stars



La columna stars muestra una distribución aproximadamente normal, con la mayoría de los valores concentrados a un sesgo a la derecha.

El histograma de stars indica que la mayoría de las empresas tienen una calificación de estrellas en torno a 4 puntos.

Distribución de Review Count



La columna review_count muestra una distribución sesgada hacia la izquierda, con la mayoría de las empresas teniendo un número relativamente uniforme de reseñas.

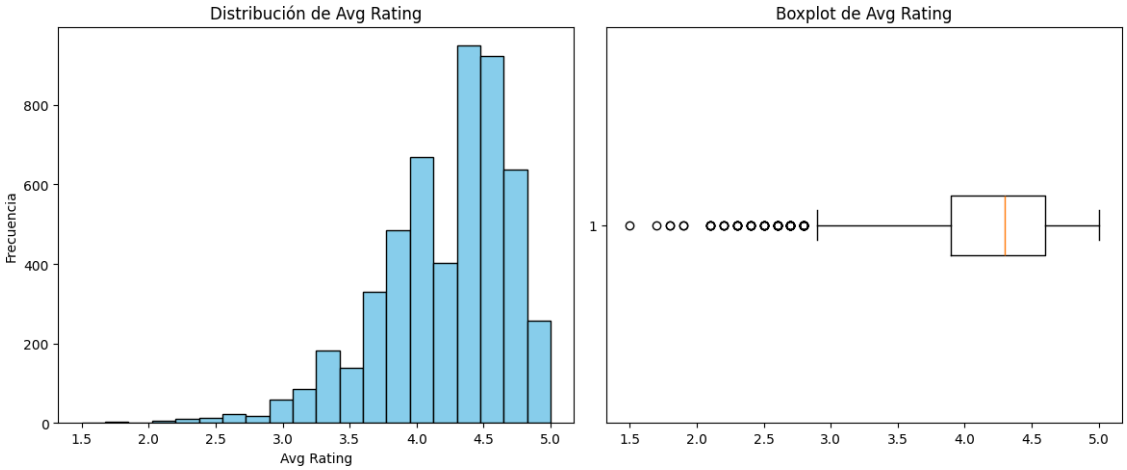
El histograma de review_count muestra una concentración de empresas con un número de reseñas entre 0 y 500, con una disminución grande en la frecuencia a medida que aumenta el número de reseñas.

El boxplot indica la presencia de outliers en la distribución de review_count, lo que sugiere la existencia de algunas empresas con un número excepcionalmente alto de reseñas.

Análisis de Dispersión de Datos GOOGLE MAPS

Análisis de la dispersión de los datos en las columnas Avg Rating y Num of Reviews de la tabla BUSINESS de GOOGLE MAPS. El objetivo principal es comprender la distribución y la presencia de outliers en estas dos variables clave.

Distribución de Avg Rating

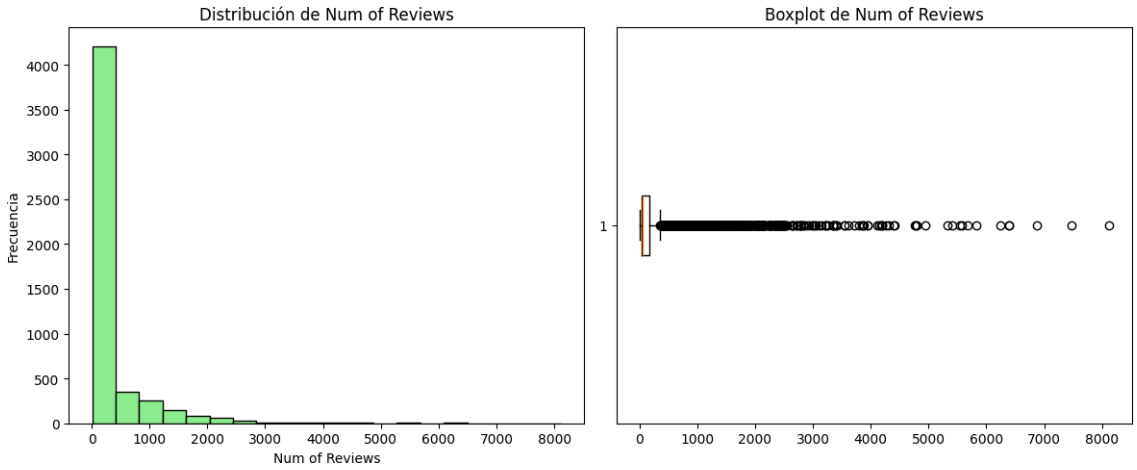


La columna Avg Rating muestra una distribución aproximadamente normal, con la mayoría de los valores concentrados a un sesgo a la derecha y poca presencia de datos mínimos

El histograma de Avg Rating indica que la mayoría de las empresas tienen una calificación de estrellas en torno a 4 y 5 puntos.

El boxplot revela la presencia de algunos outliers en la distribución de Avg Rating, lo que sugiere la existencia de algunas empresas con calificaciones de estrellas inusuales.

Distribución de Num of Reviews

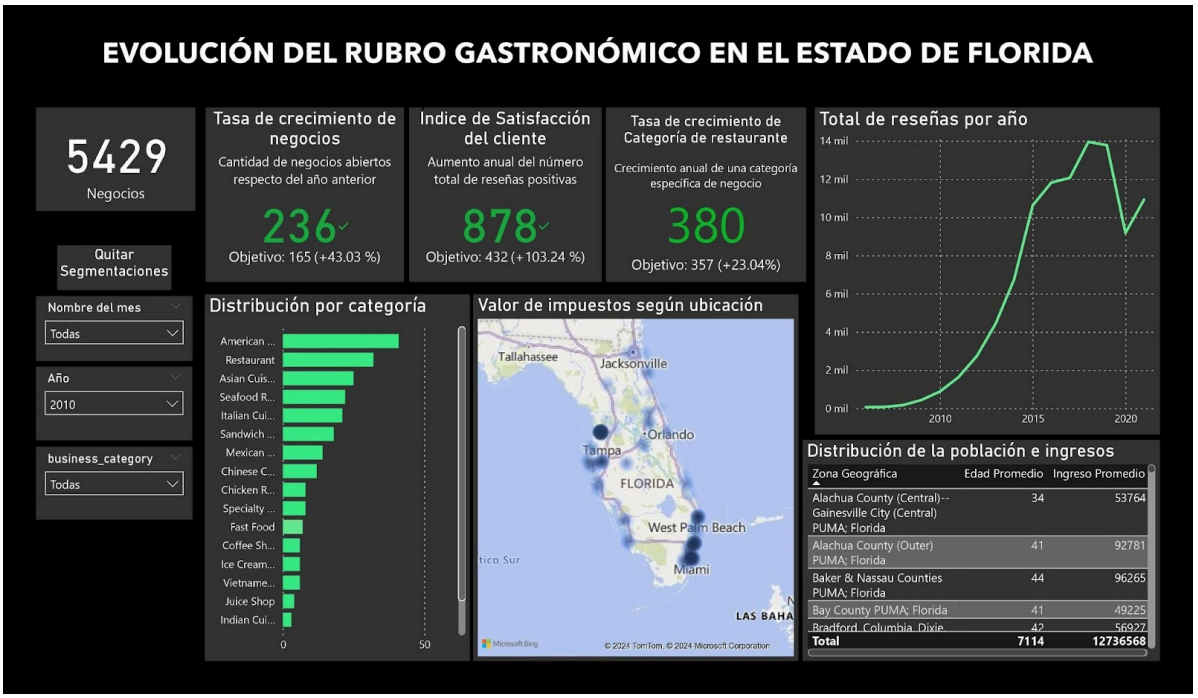


La columna Num of Reviews muestra una distribución sesgada hacia la izquierda, con la mayoría de las empresas teniendo un número relativamente uniforme de reseñas.

El histograma de Num of Reviews muestra una concentración de empresas con un número de reseñas entre 0 y 3000, con una disminución grande en la frecuencia a medida que aumenta el número de reseñas.

El boxplot indica la presencia de outliers en la distribución de Num of Reviews, lo que sugiere la existencia de algunas empresas con un número excepcionalmente alto de reseñas.

Proof of Concept de Dashboard



El dashboard diseñado presenta información sobre la evolución del sector gastronómico en el estado de Florida. Los datos se muestran en un formato que facilita la comprensión de las tendencias y patrones clave.

Secciones principales

- Segmentaciones: Año, mes y categoría de negocio.
- Análisis:

Total de negocios: indica el total de negocios gastronómicos a lo largo de los años. Esto se puede segmentar por año, mes y categoría.

KPIs: se tienen los indicadores en la parte principal del tablero para tener un seguimiento constante.

Total de reseñas por año: indica la evolución de la cantidad de reseñas realizadas a los largo del tiempo.

Distribución por categoría: muestra jerárquicamente cuáles son las categorías que más negocios tienen, esto puede segmentarse también por año o por mes.

Valor de impuestos según ubicación: a partir de un mapa de calor indica qué zona geográfica tiene impuestos más altos.

Distribución de la población según ingresos: indica el promedio de edad por zona geográfica y

el ingreso promedio por persona, por zona.

KPIs:

- KPI de Tasa de Crecimiento de Negocios

-Número de nuevos negocios restaurantes abiertos en el año actual y en el año anterior.
-Objetivo: Lograr una Tasa de Crecimiento del 10% de crecimiento anual.
Este KPI muestra la evolución anual de la Tasa de Crecimiento para analizar las tendencias de crecimiento de negocios restaurantes en el tiempo.

- KPI de Índice de Satisfacción del Cliente

-Aumento anual del número total de reseñas
-Objetivo: Lograr un aumento del 10% respecto del año anterior.
Este KPI te ayudaría a medir el crecimiento estacional de los negocios gastronómicos asesorados, estableciendo un objetivo claro de incremento en las visitas durante las temporadas altas y permitiendo evaluar su desempeño de manera periódica.

- KPI de Índice de Crecimiento estacional

-Medir el crecimiento anual de visitas durante la temporada baja (junio a agosto)
-Objetivo: Lograr un crecimiento de un 5% por año.
Este KPI permite identificar las tendencias de una categoría de interés para proporcionar información estratégica a nuestros clientes.

- KPI de Cobertura Competitiva

-Análisis trimestral de la aparición de competidores directos.
-Objetivo: identificar a los competidores directos de nuestro cliente.

Reporte de costos estimados:

Servicio	Parámetro	Uso mensual estimado	Costo unitario (USD)	Costos Estimados (USD)
Google BigQuery	Almacenamiento (GB)	500	0.02	10
Google BigQuery	Consultas (TB)	10	5	50
Google Cloud Functions	Invocaciones (millones)	10	3.2	32
Google Cloud Functions	Tiempo de computación: (GB-sec)	1000000	0.0000025	2.5
Google Cloud Functions	Red (GB)	50	5.4	270
Google Cloud Storage	Almacenamiento (GB)	100	0.026	2.6
Google Cloud Storage	Salida de red (GB)	50	0	0
Costo mensual total				367.1
Costo anual total				4405.2

Presupuesto Proyecto de Desarrollo y Despliegue en Google Cloud Platform

Resumen: El presente informe detalla los costos estimados para el desarrollo y despliegue de un proyecto en Google Cloud Platform (GCP), incluyendo servicios de almacenamiento, consultas y ejecuciones, redes, servicios adicionales, integración con Streamlit y posibles APIs externas, así como la mano de obra asociada.

Costos Estimados:

- **Servicios GCP:** Se estima un costo total mensual de $348.00, incluyendo almacenamiento, consultas, ejecuciones, redes y servicios adicionales.
- **Integración con Streamlit y APIs:** Se prevé un costo adicional variable para la integración con Streamlit y las APIs de Yelp y Google Maps, dependiendo del uso.
- **Mano de Obra:** Se estima un costo mensual de $12,600 para el desarrollo, despliegue, mantenimiento y soporte técnico del proyecto.

Total, Presupuesto Mensual: Se estima un presupuesto total mensual de $348.00 para los Servicios GCP. Además, un tentativo Mano de Obra de $12,600. Para un total de $12,948, sin incluir los costos variables asociados con la integración de Streamlit y las APIs externas.

**Nota: Los costos estimados son puramente ilustrativos y están sujetos a variaciones basadas en el uso real de los servicios y las tarifas vigentes.*

Para obtener un informe más detallado, consulta el documento "Google Cloud Platform (GCP) Gastos" en GitHub o sigue este enlace.