



# PROYECTO GOOGLE YELP. (HOTELES)



## Tabla de Contenido

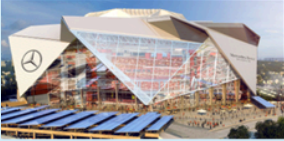


KPIs	<u>1</u>
Alcance del Proyecto	<u>2</u>
EDA	<u>3</u>
Repositorio en GitHub	<u>3</u>
Implementación Stack Tecnológico	<u>3</u>
Metodología de Trabajo	<u>4</u>
Diseño Detallado	<u>4</u>
Fuentes de Información	<u>4</u>
Equipo de Trabajo – Roles y Responsabilidades	<u>4</u>
Cronograma General de Actividades – Diagrama de GANTT	<u>4</u>
Análisis preliminar de calidad de Datos	<u>4</u>

## KPIs

KPI	Descripción	Objetivo	Fórmula	Periodicidad	Comentarios
<b>Puntuación Promedio de Satisfacción en Reseñas</b>	Medir la satisfacción del cliente a través de las reseñas.	Mantener una puntuación de satisfacción de 4.5 estrellas o superior	Promedio de las calificaciones	Anual	Este KPI ayuda a evaluar la salud financiera del hotel.
<b>Densidad de Hoteles por Zona</b>	Número de hoteles por milla cuadrada en áreas clave, lo que indica saturación de mercado.	Mantener la densidad de hoteles por debajo de 1 hotel o 2 max por milla cuadrada en áreas clave	Hoteles / Millas cuadradas	Anual	Este KPI ayuda a tomar decisiones estratégicas sobre la expansión del hotel.
<b>Índice de Competitividad</b>	Comparación de la calificación promedio de un hotel con la calificación promedio de otros hoteles en la misma área.	Superar la calificación promedio de la competencia 0.5 puntos	(Calificación promedio del hotel - Calificación promedio de la competencia)	Trimestral	Este KPI ayuda a medir el posicionamiento del hotel en comparación con la competencia.
<b>Tasa de Crecimiento de Reseñas Anuales</b>	Medir el aumento porcentual año tras año en el número de reseñas para hoteles cercanos a los estadios.	15% de crecimiento anual	$((\text{Reseñas este año} - \text{Reseñas año anterior}) / \text{Reseñas año anterior}) * 100$	Anual	Este KPI ayuda a medir la visibilidad del hotel y la satisfacción del cliente.

## Alcance del Proyecto

Hoteles y negocios que queden alrededor de los estadios donde se jugarán los partidos del mundial de fútbol del año 2026 en Estados Unidos. En la Figura 1 se detallan las características de cada estadio.

			
Estadio: Mercedes-Benz Stadium Ciudad: Atlanta Estado: Georgia Capacidad: 71000	Estadio: AT&T Stadium Ciudad: Dallas Estado: Texas Capacidad: 80000	Estadio: Gillette Stadium Ciudad: Boston Estado: Massachusetts Capacidad: 68756	Estadio: NRG Stadium Ciudad: Houston Estado: Texas Capacidad: 72220
			
Estadio: Arrowhead Stadium Ciudad: Kansas City Estado: Kansas Capacidad: 75416	Estadio: SoFi Stadium Ciudad: Los Angeles Estado: California Capacidad: 70000	Estadio: Hard Rock Stadium Ciudad: Miami Estado: Florida Capacidad: 65326	Estadio: MetLife Stadium Ciudad: New York/New Jersey Estado: New York/New Jersey Capacidad: 82500
			
Estadio: Lincoln Financial Field Ciudad: Philadelphia Estado: Pensilvania Capacidad: 67594	Estadio: Levi's Stadium Ciudad: San Francisco Estado: California Capacidad: 68500	Estadio: Lumen Field Ciudad: Seattle Estado: Washington Capacidad: 72.000	

Por otro lado, y en vista de la necesidad de los inversionistas en los temas de hospedaje, el proyecto se enfocará en los hoteles y sitios dormitorio en general que se encuentren ubicados en las bases de datos; además porque es necesario también recomendar sitios de estadía a los viajeros.

La empresa de inversiones Adubai Invest, grupo empresarial árabe con sede principal en New York e inversiones en varias ciudades de Estados Unidos y Europa está auscultando la posibilidad de realizar inversiones en hotelería en las ciudades de Estados Unidos sede del mundial de fútbol 2026. Para este propósito contrató la empresa Opportunity Hunters, especialista en buscar oportunidades de negocio mediante el análisis de datos tomados de redes sociales y páginas donde los diferentes usuarios recomiendan sitios de compra.

En este sentido, el alcance del proyecto está dado en los siguientes parámetros:

- **Lugares objetivo:** 10 estadios donde se jugarán los partidos del mundial.
- **Ciudades objetivo:** 10 ciudades donde están ubicados los estadios respectivos.
- **Sitios de interés para el análisis:** todos los puntos referenciados en la metadata de Google y Yelp con ubicación geográfica 'latitude' y 'longitude' con distancia menor a 30 kilómetros alrededor del estadio respectivo, y se agregarán algunas zonas de interés turístico de acuerdo a la ciudad.

Esta distancia objetivo la determinamos teniendo en cuenta que en Estados Unidos los sitios en general quedan muy distantes y seguramente los visitantes buscarán aquellos lugares de estadía más cercanos al estadio.

Por otro lado, y en vista de la necesidad de los inversionistas en los temas de hospedaje, el proyecto se enfocará en los hoteles y sitios dormitorio en general que se encuentren ubicados en las bases de datos; además porque es necesario también recomendar sitios de estadía a los viajeros.

En la Tabla 1, se detallan los estadios, la ciudad donde está ubicada, la capacidad en personas y el estado.

Ciudad	Estado	Estadio	Capacidad
Atlanta	Georgia	Atlanta: Mercedes-Benz Stadium (71,000)	71.000
Boston	Massachuts	Boston: Gillette Stadium (68,756)	68.756
Dallas	Texas	Dallas: AT&T Stadium (80,000)	80.000
Huston	Texas	Houston: NRG Stadum (72,220)	72.220
Kansas City	Kansas	Kansas City: Arrowhead Stadium (76,416)	76.416
Los Angeles	California	Los Angeles: SoFi Stadium (70,000)	70.000
Miami	Florida	Miami: Hard Rock Stadium (65,326)	65.326
New York/New Jersey	New York/New Jersey	New York/New Jersey: MetLife Stadium (82,500)	82.500
Philadelphia	Pensilvania	Philadelphia: Lincoln Financial Field (67,594)	67.594
San Francisco	California	San Francisco Bay Area: Levi's Stadium (68,500)	68.500
Seattle	Washinton	Seattle: Lumen Field (72,000)	72.000

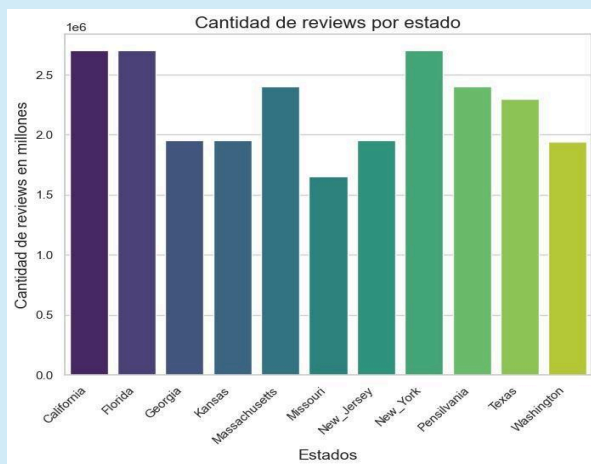
## EDA

En un primer análisis exploratorio de los datos, se pudo observar que los datos provienen de dos fuentes principales: la plataforma de reseñas Yelp y Google Maps, ambas utilizadas para recopilar información sobre comercios en Estados Unidos.

Estos datos incluyen detalles sobre la ubicación de los establecimientos, su categoría, puntajes promedio, estado de apertura, así como información relacionada con los usuarios, como las reseñas que han realizado y la cantidad de votos que han recibido.

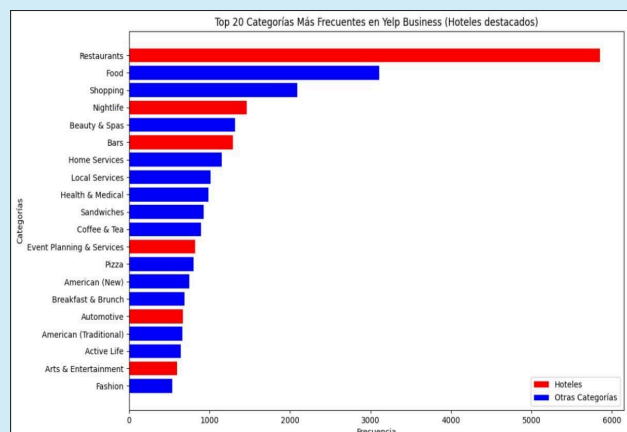
Se puede observar que en el dataset de Google Maps contiene los datos de los negocios de los 51 estados y en Yelp solo de 15 estados, por esa razón será necesario reducir la cantidad de estados a analizar. Además, se puede observar que será necesario complementar el análisis con conjuntos de datos adicionales que proporcionen información relevante para el estudio, como los valores de las acciones de las empresas o datos geográficos específicos de los locales, dado que algunas columnas de los diferentes datasets contienen hasta un 90% de valores faltantes, datos nulos, datos duplicados, entre otros.

En este contexto, se utilizarán técnicas de procesamiento del lenguaje natural (NLP) para analizar y extraer información relevante de las reseñas recopiladas.



```
df_1.info()
✓ 0.1s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 275001 entries, 0 to 275000
Data columns (total 15 columns):
#   Column              Non-Null Count  Dtype
---  -
0   name                 274994 non-null object
1   address              264939 non-null object
2   gmap_id              275001 non-null object
3   description          13155 non-null  object
4   latitude             275001 non-null float64
5   longitude            275001 non-null float64
6   category             272740 non-null object
7   avg_rating           275001 non-null float64
8   num_of_reviews       275001 non-null int64
9   price               13450 non-null  object
10  hours               192448 non-null object
11  MISC                194972 non-null object
12  state               195523 non-null object
13  relative_results    238771 non-null object
14  url                 275001 non-null object
dtypes: float64(3), int64(1), object(11)
memory usage: 31.5+ MB
```



## Repositorio en GitHub

**Enlace:** (<https://github.com/patricio-martinez-cintas/Proyecto-google-yelp>)

## Implementación del Stack Tecnológico

Para el desarrollo del proyecto, se han elegido cuidadosamente una serie de herramientas tecnológicas que permitirán un procesamiento eficiente y escalable de los datos. A continuación, se presentan las herramientas seleccionadas del Stack Tecnológico:

**Python:** Utilizado para el procesamiento y análisis de datos

**Pandas y Numpy:** Librerías para el manejo tabular de datos.

**Matplotlib y Seaborn:** Herramientas para visualización y análisis exploratorio de datos.

**NLTK (Natural Language Toolkit):** Librería para procesamiento de lenguaje natural, incluyendo análisis de sentimiento a las reseñas hechas por los usuarios de ambas plataformas.

**boto3:** Librería para interactuar con servicios de AWS como Amazon S3.

**Docker:** Sistema para gestionar contenedores, simplificando la manipulación de datos.

**Amazon S3 AWS:** Almacenamiento escalable y duradero para datos a través de internet.

**GitHub:** Plataforma de control de versiones para colaboración en equipo.

**Google Cloud Storage:** Ideal para datasets escalables y optimizados en la nube.

**Google Drive:** Utilizado para sincronización y acceso a archivos desde cualquier lugar.

**Looker Studio:** Es una herramienta utilizada para la visualización de datos a través de paneles o informes.

**Streamlit:** Modelo de Machine Learning





## Metodología de Trabajo

### Enfoque del Proyecto

El equipo del proyecto final se ha embarcado en una simulación de un ambiente laboral para fortalecer competencias técnicas y suaves, enfatizando la sinergia y la respuesta a los requerimientos de un Product Owner.

### Estructura y Roles

Se establecieron roles dentro del equipo, asignando responsabilidades principales que abarcan desde la ingeniería de datos hasta la analítica, asegurando que el proyecto reciba una atención detallada en cada fase crítica.

### Ciclo de Revisión y Mejora

Un Head Mentor supervisa diariamente, ofreciendo apoyo general y asesoramiento técnico, aunque la ejecución de tareas recae en la autonomía del equipo.

### Metodología Ágil y Herramientas de Planificación

Adoptando Scrum, el equipo trabaja en sprints, utilizando Jira para gestionar tareas y un diagrama de Gantt para la planificación. Esto permite un seguimiento detallado y la adaptación a cambios durante el proyecto.

### Hitos y Entregables

Los hitos se establecen como indicadores clave de progreso, mientras que los entregables son productos específicos resultantes de las tareas realizadas.

### Documentación Continua

La documentación meticulosa es una práctica central, manteniendo un registro detallado de cada acción y decisión.

### Planificación y Herramientas

El equipo desarrolla un cronograma detallado y selecciona un stack tecnológico que permita abordar eficientemente la limpieza y transformación de datos.

### Adelanto y Retroalimentación

Se anticipan tareas de sprints futuros para obtener retroalimentación temprana del PO y mantener el alineamiento con los objetivos.

## Sprint del Proyecto

- ★ **Sprint 1:** Se enfoca en el análisis inicial y la propuesta de cómo abordar el proyecto, estableciendo un entendimiento claro de la situación y objetivos.
- ★ **Sprint 2:** Se trabaja en la infraestructura del proyecto con un enfoque en ETL, preparando el terreno para análisis y visualizaciones de datos.
- ★ **Sprint 3:** Culmina con el desarrollo de Dashboard interactivos y modelos de ML, integrando análisis y datos procesados para proporcionar insights valiosos.

En cada sprint, la iteración, la colaboración y la flexibilidad son fundamentales, guiando al equipo hacia la entrega de soluciones innovadoras de data.

## Diseño Detallado

Para desarrollar un diseño detallado del proyecto utilizando exclusivamente Google Cloud, se puede seguir la siguiente estructura:

### Ingesta de Datos

- ★ Utilizar Google Cloud Storage (GCS) para almacenar los datasets en formatos Parquet y JSON.
- ★ GCS es escalable y ofrece una capa de uso gratuito que es adecuada para el almacenamiento de grandes conjuntos de datos.

### Procesamiento y Análisis

- ★ Emplear Google Cloud Dataproc para procesar y analizar grandes volúmenes de datos.
- ★ Dataproc es un servicio de Hadoop y Spark gestionado que permite procesar datos directamente en GCS.
- ★ Para análisis interactivos y ejecución de queries a gran escala, se puede utilizar Google BigQuery.

### Despliegue y Orquestación

- ★ Google Cloud Functions y Google Cloud Composer (basado en Apache Airflow) son adecuados para automatizar flujos de trabajo y desplegar aplicaciones.
- ★ Estos servicios permiten una orquestación eficiente y están integrados con el resto de servicios de Google Cloud.

### Visualización de Datos

- ★ Google Data Studio o Looker (si está disponible en la capa gratuita) pueden ser utilizados para crear dashboards interactivos que se conecten directamente a BigQuery o a otras fuentes de datos en GCP.

### Ciclo de Vida del Dato

- ★ Recolección: Automatización de la ingesta de datos mediante servicios de Google Cloud como Pub/Sub o directamente desde GCS.
- ★ Limpieza y Preparación: Herramientas como Google Cloud Dataprep para limpiar y transformar datos antes del análisis.

- ★ **Análisis y Modelado:** Uso de BigQuery ML para modelado directamente en el almacén de datos, y AI Platform para modelos de ML más complejos.
- ★ **Visualización y Reporte:** Data Studio para presentar resultados y KPIs.
- ★ **Monitoreo y Mantenimiento:** Google Cloud Monitoring y Google Cloud Logging para el seguimiento del rendimiento de las aplicaciones y los costos asociados.

## Equipo de Trabajo – Roles y Responsabilidades

Dado que el Mundial de Fútbol en 2026 se efectuará en varias de las ciudades más importantes de los Estados Unidos, este proyecto es crucial para determinar dónde emplazar nuevos sitios de hospedaje, alojamiento y otros negocios relacionados con el turismo. En este orden de ideas, Opportunity Hunters DC está comprometida en brindar soluciones efectivas y estratégicas para su cliente en la industria de la hotelería y el ocio.

El objetivo es realizar un análisis exhaustivo del mercado estadounidense, centrándose en las opiniones de los usuarios en Yelp y Google Maps sobre hoteles y otros negocios relacionados con el turismo y el ocio.

Los siguientes serán los roles clave que desempeñarán los miembros del equipo de Opportunity Hunters:

### **Data Engineers (Ingenieros de Datos):**

- **Patricio Martínez Cintas**
- **María Gabriela Pacheco Franco**
  - ★ Los Data Engineers son los arquitectos detrás de la infraestructura de datos. Su tarea principal es diseñar y mantener sistemas eficientes para la recopilación, almacenamiento y procesamiento de datos.
  - ★ Construir los Data Pipeline para extraer, transformar y cargar (ETL) la información de Yelp y Google Maps.
  - ★ Garantizar que los datos estén limpios, disponibles y listos para su análisis por parte de los Data Scientists y el Data Analyst.

### **Data Scientists (Científicos de Datos):**

- **Hector Fabio Ocampo Gaviria**
- **Luis Mary Esmeralda Gaince Pereira**
  - ★ Los Data Scientists son los responsables del análisis. Su tarea principal es utilizar técnicas avanzadas de aprendizaje automático y análisis de sentimientos para extraer información valiosa de las reseñas de usuarios.
  - ★ Identificar patrones, tendencias y oportunidades de crecimiento en los rubros de negocios relacionados con la hotelería y el turismo.

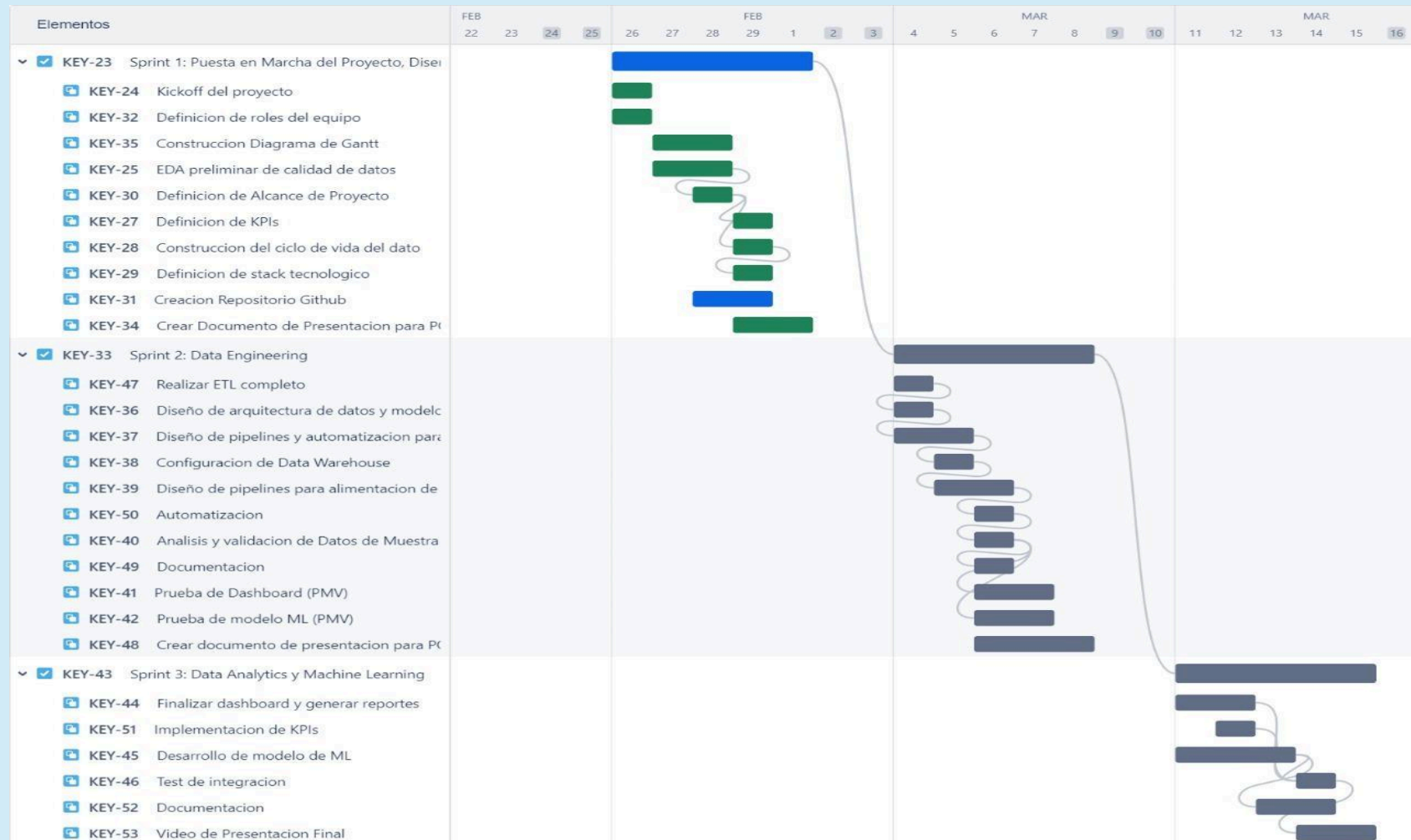
- ★ Su trabajo ayudará a predecir qué sectores experimentarán un crecimiento significativo y cuáles podrían decaer.

#### **Data Analyst (Analista de Datos):**

- **Facundo José Cuervo**

- ★ El Data Analyst será el narrador de la historia detrás de los datos. Su tarea principal es analizar las opiniones de los usuarios y crear visualizaciones claras y concisas.
- ★ Utilizar análisis de sentimientos para comprender las experiencias de los usuarios en Yelp y Google Maps.
- ★ Además, desarrollar un sistema de recomendación de hospedaje y alojamiento basado en las preferencias y experiencias previas de los usuarios.

## Cronograma General de Actividades – Diagrama GANTT



## **Análisis Preliminar de la Calidad de los Datos**

El análisis preliminar de la calidad de los datos es crucial para garantizar que los datos sean confiables y adecuados para su uso. En consecuencia se puede resumir lo siguiente:

- ★ Existencia de valores atípicos, datos faltantes o inconsistencias.
- ★ Algunos datos no están en el formato correcto.
- ★ Existencia de duplicados o registros repetidos.
- ★ Se realizó el cálculo de las estadísticas básicas, como la media, la mediana y la desviación estándar para comprender la distribución de los datos.
- ★ Se observaron los valores máximos y mínimos para detectar posibles errores y valores extremos.
- ★ Validación de los dominios específicos (calificaciones están en el rango correcto de 1 a 5).
- ★ Los datos siguen un patrón lógico.
- ★ Hay inconsistencias entre diferentes columnas o atributos.
- ★ Se decide eliminar registros con datos faltantes o imputar valores donde sea necesario.
- ★ Se corrigen errores obvios, como errores tipográficos o entradas incorrectas.
- ★ Se compararon los datos con fuentes externas para detectar discrepancias o errores.