**Laboratoria Bootcamp**

**Reto Técnico**

**Betsabé Morales Castro**

**Agosto 26, 2024**

**Ficha Técnica**

**Título del Proyecto: Citi Bike Trips**

**Objetivo:**

Realizar una exploración y análisis de los datos proporcionados con el fin de presentar a la nueva CEO de Citi Bike nuestra información mas relevante en cuanto a usuarios.

**Listado de herramientas y tecnologías utilizadas en el proyecto:**

1) Workspace Google (Hojas de cálculo, Presentaciones, Gemini y Documentos).

2) Chat GPT

3) Big Query (SQL)

4) Looker Studio

Para el proyecto de Citi Bike Trips contamos con la [hoja de datos](https://drive.google.com/file/d/1ywodvLvsoaREi5NZ3fG6l28nTYqKHkR-/view?usp=drive_link) que nos proporcionó la empresa.

**Procesamiento y análisis.**

Para el proceso y preparación de la base de datos realizamos los siguientes pasos:

1. Se revisó la tabla de datos proporcionada para conocer las variables y comenzar a ver que tratamiento se le iba a dar.
2. Conectar/importar datos a herramienta Big Query.
3. Identificar y manejar valores nulos, se encontraron 4,639 nulos en la columna birth\_year.

La query utilizada es la siguiente:

SELECT

COUNTIF(tripduration IS NULL) AS tripduration\_null\_count,

COUNTIF(stoptime IS NULL) AS stoptime\_null\_count,

COUNTIF(start\_station\_id IS NULL) AS start\_station\_id\_null\_count,

COUNTIF(start\_station\_name IS NULL) AS start\_station\_name\_null\_count,

COUNTIF(start\_station\_latitude IS NULL) AS start\_station\_latitude\_null\_count,

COUNTIF(start\_station\_longitude IS NULL) AS start\_station\_longitude\_null\_count,

COUNTIF(end\_station\_id IS NULL) AS end\_station\_id\_null\_count,

COUNTIF(end\_station\_name IS NULL) AS end\_station\_name\_null\_count,

COUNTIF(end\_station\_latitude IS NULL) AS end\_station\_latitude\_null\_count,

COUNTIF(end\_station\_longitude IS NULL) AS end\_station\_longitude\_null\_count,

COUNTIF(bikeid IS NULL) AS bikeid\_null\_count,

COUNTIF(usertype IS NULL) AS usertype\_null\_count,

COUNTIF(birth\_year IS NULL) AS birth\_year\_null\_count,

COUNTIF(gender IS NULL) AS gender\_null\_count,

COUNTIF(customer\_plan IS NULL) AS customer\_plan\_null\_count

FROM

`citi\_bike\_trips.citi\_bike\_trips`;

1. Identificar y manejar valores duplicados, no se identificaron duplicados en la tablas ya que no existe un id de usuario.
2. Identificar y manejar datos fuera del alcance del análisis, no eliminé ningún dato, lo que realicé fué su acotación por medio de winsorización para que no altere los cálculos de las estadísticas. No se tomará en cuenta la variable género para el análisis por considerarlo discriminatorio.
3. Se crearon nuevas variables como age y se obtuvo la edad promedio de los usuarios sin contar a los nulos, para posteriormente utilizar esa edad promedio de 45.78 años en los valores nulos. Lo realicé con la siguiente query

SELECT

AVG(age) AS average\_age

FROM

`citi\_bike\_trips.view\_birth\_year\_edad`

WHERE

age IS NOT NULL;

1. Se hizo una nueva view en donde también se sustituyeron los nulos de la edad por la edad promedio, se acotó el máximo de edad a 80 años para manejar los valores atípicos, la variable tripduration estaba en segundos y se convirtió a minutos. De igual forma se dividió la columna stoptime utilizando el comando date en stoptime\_date y en stoptime\_time y se creó la variable start\_time con la siguiente query:

SELECT

-- Especificar las columnas que deseas incluir

birth\_year,

tripduration,

stoptime,

start\_station\_id,

start\_station\_name,

start\_station\_latitude,

start\_station\_longitude,

end\_station\_id,

end\_station\_name,

end\_station\_latitude,

end\_station\_longitude,

bikeid,

usertype,

gender,

-- Crear columna age con valores nulos reemplazados por 46 y limitar el máximo a 80 años

CASE

WHEN birth\_year IS NOT NULL THEN

LEAST(EXTRACT(YEAR FROM CURRENT\_DATE) - birth\_year, 80)

ELSE 46

END AS age,

-- Convertir tripduration de segundos a minutos

tripduration / 60 AS tripduration\_minutes,

-- Separar stoptime en fecha y hora

DATE(stoptime) AS stoptime\_date,

TIME(stoptime) AS stoptime\_time,

-- Crear columna start\_time con la resta de stoptime y tripduration

TIME(TIMESTAMP\_SUB(stoptime, INTERVAL tripduration SECOND)) AS start\_time

FROM

`citi\_bike\_trips.citi\_bike\_trips`;

Se realizó un análisis exploratorio con los siguientes pasos:

1. Agrupar datos según variables categóricas
2. Visualizar las variables categóricas
3. Aplicar medidas de tendencia central
4. Visualizar distribución
5. Aplicar medidas de dispersión
6. Visualizar el comportamiento de los datos a lo largo del tiempo
7. Calcular cuartiles, deciles o percentiles
8. Calcular correlación entre variables

Se aplicó técnica de análisis

1. Aplicar segmentación por tipo de usuario
2. Se realizó un análisis descriptivo

**Distribución de Edad -** Se analizó la distribución de la edad de los usuarios usando gráfico de densidad para visualizar cómo se distribuyen las edades.

**Duración de los Viajes -** Se examinó la duración de los viajes (tripduration). Se Utilicé estadística descriptiva como la media, mediana, máximo y mínimo. **Uso por Género:**

* Compara la duración del viaje, número de viajes y otros aspectos entre géneros. Puedes usar gráficos de barras o boxplots para estas comparaciones.

Resumì la información en un dashboard de Looker Studio

1. Representar los datos en una tabla o scorecard
2. Representar datos a través de gráficos simples
3. Representar datos a través de gráficos o visuales avanzados
4. Aplicar opciones de filtro para manejo e interacción

Para presentar resultados se realizó lo siguiente:

1. Seleccioné gráficos con información relevante
2. Creamos una presentación en GitHub

**Enlaces de interés:**

**Looker Studio -** [**Dashboard**](https://lookerstudio.google.com/reporting/b0f60433-15c2-49ba-ab1b-e0835d977e90)