Ficha Técnica: Proyecto Análisis de Datos

Titulo: Proyecto Cicloviajes

Objetivo:

- **Realizar un análisis descriptivo de datos.**
- **Calcular métricas de uso de un día promedio.**
    - Número de viajes que se realizan en promedio en un día.
    - Calcular las medidas de dispersión: valor máximo, valor mínimo, promedio y desviación estándar de la duración de un viaje.
- **Calcular métricas históricas.**
    - Calcular el total de viajes realizados.
    - Analizar el crecimiento del número de viajes diarios a lo largo del tiempo.
    - Analizar el total de viajes por usuarios, según género, edad y/o tipo de suscripción.
- **Crear reporte, dashboard con visualizaciones del resumen del análisis.**
- **Presentar conclusiones generales, patrones y recomendaciones al nuevo CEO.**

Equipo
Individual

Herramientas y Tecnologías:

- Google BigQuery
- Google Docs
- Google Slides
- Google Looker Studio

Lenguajes:

- SQL

Insumos:
citi_bike_trips.csv

Diccionario de datos:

- **tripduration:** Duración del viaje (en segundos).
- **stoptime:** Fecha y hora de finalización del viaje.
- **Start_station_id:** Identificador único de la estación donde comenzó el viaje.
- **Start_station_name:** Nombre de la estación donde comenzó el viaje.
- **Start_station_latitude:** Latitud geográfica de la estación donde comenzó el viaje.
- **start_station_longitude:** Longitud geográfica de la estación donde comenzó el viaje. **bikeid:** Identificador único de la bicicleta.
- **usertype:** Tipo de usuario “Subscriber” (suscriptor) o “Customer” (cliente).
- **birth_year:** Año de nacimiento del usuario.

- **gender:** Género del usuario, male (masculino), female (femenino) y unknow (desconocido).
- **customer_plan:** Plan de suscripción del usuario.

Procesamiento y análisis de datos

Conectar/importar datos a otras herramientas

- Se creó el reto-tecnico y el conjunto de datos Dataset en BigQuery.
- Tablas importadas: citi_bike_trip.

Identificar y manejar valores nulos

- Se identifican valores nulos a través de comandos SQL COUNTIF, IS NULL.
- **Birth_year:** se identificaron 4639 valores nulos. Estos datos representan el 9.27% del total, por lo que no serán considerados en el análisis.
- **Customer_plan:** se identificaron 50,000 valores nulos. Esta variable tiene el 100% de datos nulos, por lo que será excluida del análisis.

Identificar y manejar valores duplicados

- Se identifican duplicados a través de comandos SQL COUNT, GROUP BY, HAVING.
- No se encontraron valores duplicados en las variables.

Identificar y manejar datos fuera del alcance del análisis

- Se manejan variables que no son útiles para el análisis a través de comandos SQL SELECT EXCEPT.
- Se excluye la variable customer_plan.
- Al analizar la variable birth_year se observó que hay 32 usuarios con un rango de edad entre 103 - 139 años, por lo que no serán considerados en el análisis.
- Se creó una nueva consulta para excluir los valores nulos de la variable birth_year y excluir la variable customer_plan, quedando un total de 42651 registros.

Crear nuevas variables

- Se creó la variable *tripduration_min* transformando la variable *tripduration* de segundos a minutos.
- Se creó la variable *stopdate* extrayendo la fecha de la variable *stoptime*, usando el comando DATE.
- Se creó la variable *age* para los usuarios, usando los comandos EXTRACT(YEAR FROM CURRENT_DATE()) - birth_year).

Unir tablas

- Se creó una vista con la unión de las consultas anteriores, considerando las variables que se crearon y excluyendo las que no son relevantes para el análisis.

Análisis de Descriptivo de Datos

Métricas de uso de un día promedio

- Con los comandos AVG, MAX, MIN, STDDEV, COUNT, COUNT(DISTINCT) se calculó el número de viajes que se realizan en promedio por día, y las medidas de dispersión: valor máximo, valor mínimo, promedio y desviación estándar de la duración de un viaje.

Row	avg_tripduration_min	max_tripduration	min_tripduration	stddev_tripduration	avg_trips_per_day
1	14.74561575738...	54566.06666666...	1.0	264.4965083840...	48.27369542066...

Métricas históricas

- Con el comando COUNT, se calculó el total de viajes.

Row	total_trips
1	45329

- Con los comandos COUNT, GROUP BY, ORDER BY, se agruparon los viajes realizados en el mismo día para analizar su crecimiento a lo largo del tiempo.

Row	stopdate	daily_trips
1	2015-04-01	4
2	2015-04-02	6
3	2015-04-03	4
4	2015-04-04	7
5	2015-04-05	1
6	2015-04-06	9
7	2015-04-07	3
8	2015-04-08	5
9	2015-04-09	6
10	2015-04-10	5
11	2015-04-11	3
12	2015-04-12	4
13	2015-04-13	5
14	2015-04-14	6
15	2015-04-15	9

- Con los COUNT, GROUP BY, ORDER BY se hace un conteo de los viajes por usuario por género, edad y tipo de suscripción

Row	gender	age_group	usertype	total_trips
1	male	22	Subscriber	1
2	male	23	Subscriber	16
3	male	23	Customer	2
4	female	23	Subscriber	4
5	female	24	Subscriber	8
6	male	24	Subscriber	32
7	female	24	Customer	2
8	male	24	Customer	1
9	unknown	24	Subscriber	1
10	female	25	Subscriber	29
11	male	25	Subscriber	63

Análisis Exploratorio de Datos

Agrupar Datos

- Se importaron y conectaron los datos desde BigQuery a Looker Studio.
- Se crearon score cards:
    - Total de viajes: 45,329.
    - Suscriptor: 43,591.
    - Cliente: 1,738.
    - Bicicletas: 14,607.
    - Estación inicio: 130.
    - Estación fin: 835.
- Métricas de uso de un día:
    - Viajes promedio: 48.3.
    - Viaje promedio (min): 14.7.
    - Min viaje (min): 1.
    - Máx viaje (min): 54.6K.
    - STD viaje (min): 264.5.

Visualizar Variables

- Se creó gráfica circular , barras, y bivariable para visualizar los datos en looker studio.

Conclusiones

- **Suscriptores:** El 96% de los viajes provienen de suscriptores (43,591 de un total de 45,329). Esto sugiere que la mayor parte del uso del servicio proviene de usuarios recurrentes que valoran la accesibilidad y conveniencia.

- **Género:** El 74.9% de los usuarios identificados son hombres, mostrando una gran disparidad en el uso del servicio entre géneros. Esto podría sugerir que hay oportunidades de atraer a más mujeres al servicio.
- **Tendencias de Viajes a lo Largo del Tiempo:** La gráfica de crecimiento muestra un aumento en el número de viajes diarios y bicicletas hasta 2017, pero luego se observa una caída en los años siguientes. Es crucial entender las causas detrás de esta disminución, como cambios en la demanda o problemas operativos.
- **Duración Promedio de Viajes:** El tiempo promedio de viaje es de 48.3 minutos, con una desviación estándar significativa de 264.5 minutos. Esto sugiere que hay una alta variabilidad en la duración de los viajes.
- **Distribución por Género en los Viajes:** Los viajes realizados por hombres (33,944) son considerablemente más altos que los de mujeres (10,254), reafirmando el patrón de uso en función del género.
- **Rango de Edad de Usuarios Activos:** Los usuarios más activos están en el rango de 31-40 años, con un total de 16,415 viajes diarios, seguidos por el grupo de 41-50 años con 11,453 viajes. Los grupos de edades mayores presentan una participación decreciente, siendo menos activos los usuarios de más de 70 años.

Patrones Descubiertos:

- **Preferencia por el Servicio de Suscripción:** La mayoría de los usuarios opta por la suscripción, lo que muestra que la estrategia de fidelización ha sido efectiva. Sin embargo, también revela que el segmento de clientes ocasionales es pequeño y podría haber oportunidades para captarlos.
- **Segmentación por Edad y Género:** Los hombres de 31-50 años son el grupo más activo, lo que sugiere un perfil demográfico claro para los usuarios frecuentes. Es importante considerar estrategias de expansión hacia otros grupos de edad o género para equilibrar la base de usuarios.
- **Desaceleración en el Crecimiento:** Aunque el servicio experimentó un crecimiento hasta 2017, la caída posterior indica que algo cambió. Es crucial explorar factores como la competencia, cambios en la población, o la experiencia del usuario.

Recomendaciones

- **Diversificación de la Base de Usuarios:** Se deberían implementar campañas de marketing dirigidas a mujeres y usuarios de diferentes rangos de edad para aumentar la inclusión y el uso del servicio en estos segmentos.
- **Investigación de la Disminución en el Número de Viajes:** Realizar un análisis más profundo para entender por qué la cantidad de viajes ha disminuido en los últimos años. Esto podría involucrar estudios de mercado, encuestas a usuarios, y análisis de la competencia para ajustar la estrategia de negocio.
- **Optimización de Planes de Suscripción:** Dado que la mayoría de los usuarios son suscriptores, se podría explorar la introducción de planes personalizados o descuentos adicionales para aumentar la retención y captar nuevos usuarios.

- **Mejorar la Infraestructura:** Evaluar la distribución de estaciones y su capacidad con base en la demanda real para identificar áreas de mejora o expansión en la infraestructura.
- **Enfoque en Usuarios Jóvenes:** Dado que los usuarios entre 31 y 50 años son los más activos, se podrían desarrollar programas o incentivos dirigidos específicamente a este grupo para seguir potenciando su lealtad y aumentar el uso en otros grupos de edad.