



Contenido

[1. Introducción 2](#_heading=h.oarneh9fcc86)

[2. Herramientas Utilizadas 2](#_heading=h.d8csrp7s64n1)

[2.1 Google BigQuery 2](#_heading=h.nlie7gkggdgw)

[2.2 Alteryx Designer Cloud (anteriormente Google Dataprep) 2](#_heading=h.ey58ctrb2ljg)

[2.3 Google Looker Studio 3](#_heading=h.zdiltplzvrs)

[3. Ingesta de Datos 3](#_heading=h.qaxqr8f2gv2z)

[3.1 Fuente de Datos 3](#_heading=h.3ntp3hho4yc)

[3.2 Herramientas Utilizadas 3](#_heading=h.t5fchkc3joih)

[3.3 Proceso de Ingesta 3](#_heading=h.5ndqqqyn7ej3)

[4. Transformación y Limpieza 4](#_heading=h.9hggtafxe4a)

[4.1 Tabla citibike\_trips 4](#_heading=h.kwyrzsfycd8h)

[4.2 Tabla citibike\_stations 5](#_heading=h.9vvdm32zlvsf)

[4.3. Validación Final del Dataset 5](#_heading=h.xocaifwcb2cm)

[5. Enriquecimiento y Unión de Datos 6](#_heading=h.5ff994doz4h)

[6. Visualización y Análisis de Insights 6](#_heading=h.a1xnzuaq3uwe)

[6.1 Viajes por Tipo de Usuario 7](#_heading=h.fz4dzd35yb0d)

[6.2 Viajes por Mes 7](#_heading=h.wflz25kb4b1m)

[6.3 Promedio de Duración por Género 7](#_heading=h.h75qvmnkjwwe)

[6.4 Estaciones Más Usadas 8](#_heading=h.gfuocpp3hhvi)

[6.5 Bicicletas Más Usadas 8](#_heading=h.abfxsb51c2o0)

[7. Conclusión 9](#_heading=h.go84wemeklzf)

# 1. Introducción

En el contexto de la asignatura de Big Data, se desarrolló un proyecto práctico cuyo objetivo fue aplicar los conceptos de ingesta, limpieza, almacenamiento y visualización de datos a gran escala. Para ello, se trabajó con el conjunto de datos públicos NYC Citi Bike Trips, específicamente con la tabla citibike\_stations, la cual contiene información sobre las estaciones de bicicletas públicas de Nueva York, incluyendo ubicación, capacidad y disponibilidad.

El propósito principal fue construir un pipeline de procesamiento de datos en un entorno de Big Data, utilizando Google BigQuery como plataforma de almacenamiento y procesamiento, Alteryx Designer Cloud como entorno de limpieza, y Google Looker Studio para la visualización interactiva.

# 2. Herramientas Utilizadas

Durante el desarrollo del proyecto se emplearon herramientas de análisis y visualización de datos pertenecientes al ecosistema Google Cloud, junto con una plataforma de preparación visual de datos. A continuación, se describen las herramientas principales utilizadas y su función dentro del flujo de trabajo:

2.1 Google BigQuery  
BigQuery es el sistema de almacenamiento y análisis de datos en la nube de Google, diseñado para manejar grandes volúmenes de información de forma rápida y escalable.  
En este proyecto se utilizó para:

* Crear un dataset propio (citibike\_data) a partir de datos públicos.
* Ejecutar consultas SQL para extraer, transformar y almacenar datos.
* Crear una vista extendida (citibike\_with\_geo) que incluyó una columna geoespacial usando ST\_GEOGPOINT.

2.2 Alteryx Designer Cloud (anteriormente Google Dataprep)  
Alteryx Designer Cloud es una herramienta de preparación de datos basada en la nube que permite transformar, limpiar y perfilar conjuntos de datos sin necesidad de codificación avanzada.  
Fue utilizada para:

* Conectar el dataset citibike\_stations desde BigQuery.
* Aplicar reglas de limpieza (eliminar duplicados, eliminar columnas innecesarias, forzar tipos de datos).
* Validar visualmente la estructura del dataset.
* Generar el dataset limpio citibike\_cleaned que fue exportado nuevamente a BigQuery.

2.3 Google Looker Studio  
Looker Studio es la herramienta de visualización de datos de Google. Permite crear dashboards dinámicos e interactivos a partir de múltiples fuentes de datos.  
En este proyecto, Looker Studio se utilizó para:

* Conectarse directamente a la vista citibike\_with\_geo de BigQuery.
* Crear gráficos representativos: barras, líneas, mapas geoespaciales y comparativas.
* Aplicar filtros interactivos por región (region\_id) y por fecha (last\_reported).
* Visualizar patrones de uso, disponibilidad y distribución espacial del sistema CitiBike.

Estas tres herramientas actuaron de forma integrada para construir un pipeline completo desde la ingesta de datos hasta la presentación de visualizaciones dinámicas orientadas a la toma de decisiones.

# 3. Ingesta de Datos

## 3.1 Fuente de Datos

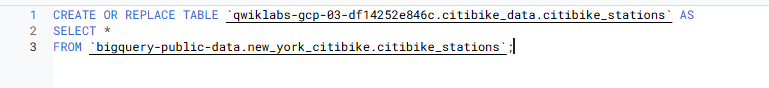
* **Tabla citibike\_trips:** bigquery-public-data.new\_york.citibike\_trips
* **Tabla citibike\_stations:** bigquery-public-data.new\_york.citibike\_stations

## 3.2 Herramientas Utilizadas

* **BigQuery Console:** Para conexión directa al dataset.
* **Google Cloud Platform (GCP):** Como entorno de análisis.
* No se descargaron archivos localmente ni se utilizó almacenamiento en Cloud Storage; se trabajó directamente sobre BigQuery en modalidad **batch**.

## 3.3 Proceso de Ingesta

Para asegurar la disponibilidad de datos dentro del entorno de trabajo personal en Google Cloud, se realizó una copia directa de la tabla pública citibike\_stations al proyecto propio:



**Explicación:**

* CREATE OR REPLACE TABLE: Crea o reemplaza una tabla en el dataset personal.
* qwiklabs-gcp-03-df14252e846c.citibike\_data: Proyecto y dataset de destino.
* SELECT \* FROM: Copia todos los registros desde la tabla pública original.

Este mismo proceso se realizo con la tabla citibike\_trips.

Este paso permitió trabajar con los datos sin modificar la fuente original y aplicar procesos personalizados de limpieza y análisis sobre una versión controlada del dataset.

# 4. Transformación y Limpieza

En esta fase se aplicaron procesos fundamentales para garantizar la calidad, coherencia y utilidad analítica de los datos. Los procesos fueron ejecutados utilizando herramientas de limpieza en entorno BigQuery y plataformas compatibles con flujos tipo “recipe” (receta), facilitando el análisis exploratorio visual y la depuración automática.

## 4.1 Tabla citibike\_trips

**Receta aplicada:**

**Explicación de los pasos:**

1. Eliminación de filas duplicadas.
2. Eliminación de la columna customer\_plan, que se encuentra incompleta y sin utilidad para el análisis solicitado.
3. Filtrado de valores inconsistentes en la columna gender (por ejemplo, 0).
4. Eliminación de filas con valores faltantes en birth\_year.

Descripción de columnas principales:

| **Columna** | **Descripción** |
| --- | --- |
| tripduration | Duración del viaje en segundos. |
| start\_time | Fecha y hora de inicio del viaje. |
| stop\_time | Fecha y hora de término del viaje. |
| start\_station\_id | ID de la estación de inicio. |
| end\_station\_id | ID de la estación de término. |
| bikeid | Identificador de bicicleta. |
| usertype | "Customer" o "Subscriber". |
| birth\_year | Año de nacimiento del usuario. |
| gender | Género del usuario (1: masculino, 2: femenino). |

## 4.2 Tabla citibike\_stations

**Receta aplicada:**

**Explicación de los pasos:**

1. Conversión del tipo de dato short\_name a String.
2. Eliminación de las columnas rental\_methods, eightd\_has\_key\_dispenser y eightd\_has\_available\_keys, al no aportar valor directo al análisis.
3. Eliminación de duplicados, asegurando unicidad en las estaciones.

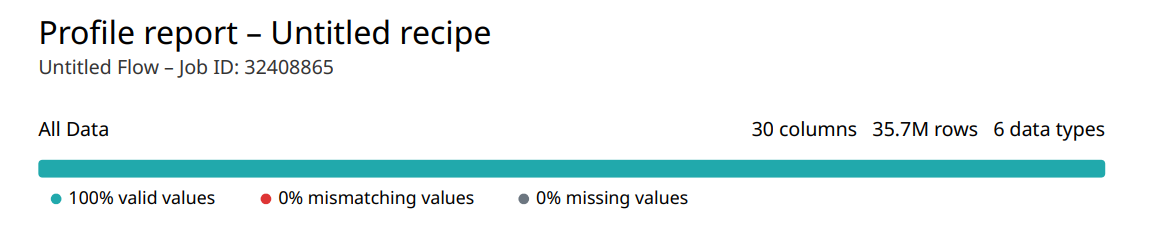
**Descripción de columnas clave:**

| **Columna** | **Descripción** |
| --- | --- |
| station\_id | Identificador único de la estación. |
| name | Nombre de la estación. |
| lat | Latitud geográfica. |
| lon | Longitud geográfica. |
| capacity | Número de bicicletas disponibles en la estación. |

## 4.3. Validación Final del Dataset

Al finalizar el proceso de limpieza y transformación, se generó un perfil de datos que arrojó los siguientes resultados:

**Resultado del perfil de calidad:**

**** Resumen:

* 100% valores válidos
* 0% valores inconsistentes
* 0% valores faltantes

Esto confirma que el dataset se encuentra totalmente limpio, con una estructura de 30 columnas y más de 35.7 millones de registros preparados para análisis posterior y visualización en Looker Studio.

# 5. Enriquecimiento y Unión de Datos

Con el objetivo de potenciar el análisis exploratorio y facilitar la interpretación visual de los datos en Looker Studio, se realizó la **unión entre las tablas citibike\_trips y citibike\_stations**. Esta unión permitió reemplazar los identificadores de estaciones por sus nombres y atributos geográficos (latitud, longitud, capacidad), mejorando la lectura y contextualización de los trayectos.

**Beneficios de esta unión**

* Se incorporó **nombre, coordenadas y capacidad** de las estaciones directamente a cada viaje.
* Permite **crear mapas, filtrar por zonas geográficas o agrupar por estación real**, lo cual no es posible con IDs numéricos.
* Mejora la presentación en dashboards, facilitando la lectura para usuarios no técnicos.

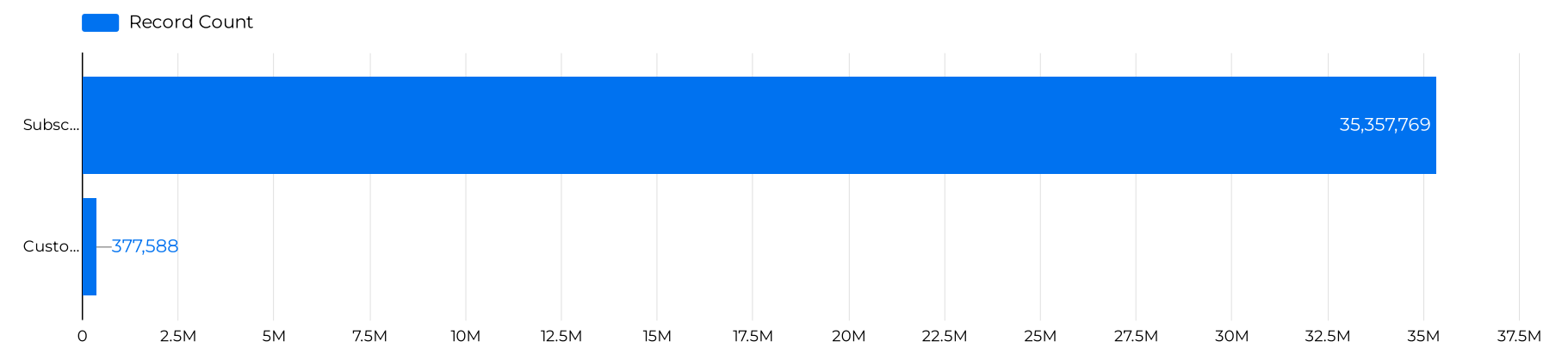
**Esta unión fue especialmente útil para visualizaciones como:**

* “Estaciones más usadas”
* “Promedio de viajes por estación”
* “Trayectos más frecuentes” (cuando se replica para estación final)

# 6. Visualización y Análisis de Insights

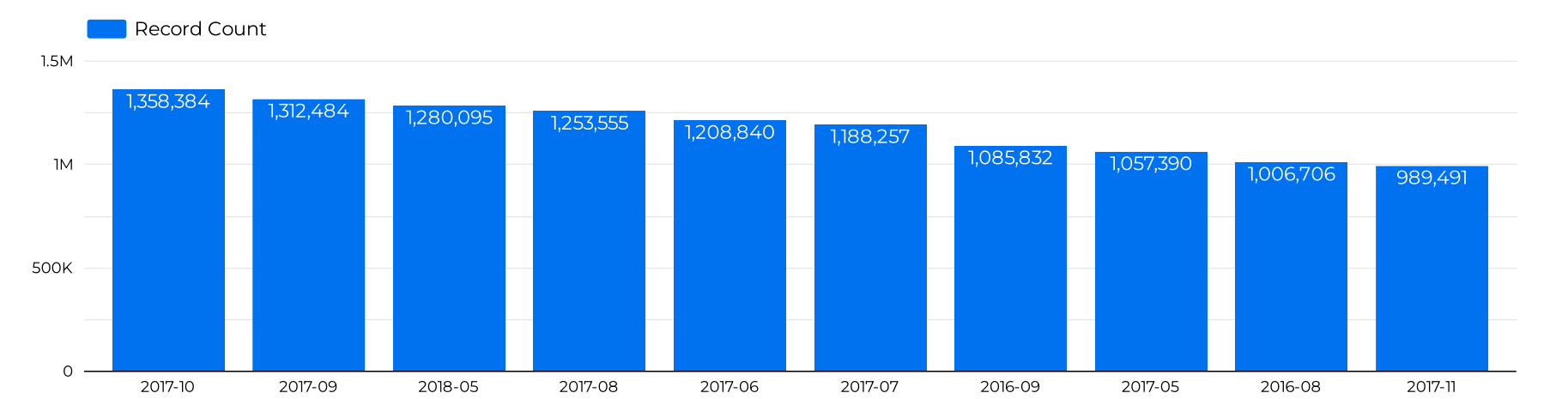
Los siguientes gráficos fueron creados como parte del análisis exploratorio. Todos están extraídos del informe visual en PDF:

## 6.1 Viajes por Tipo de Usuario



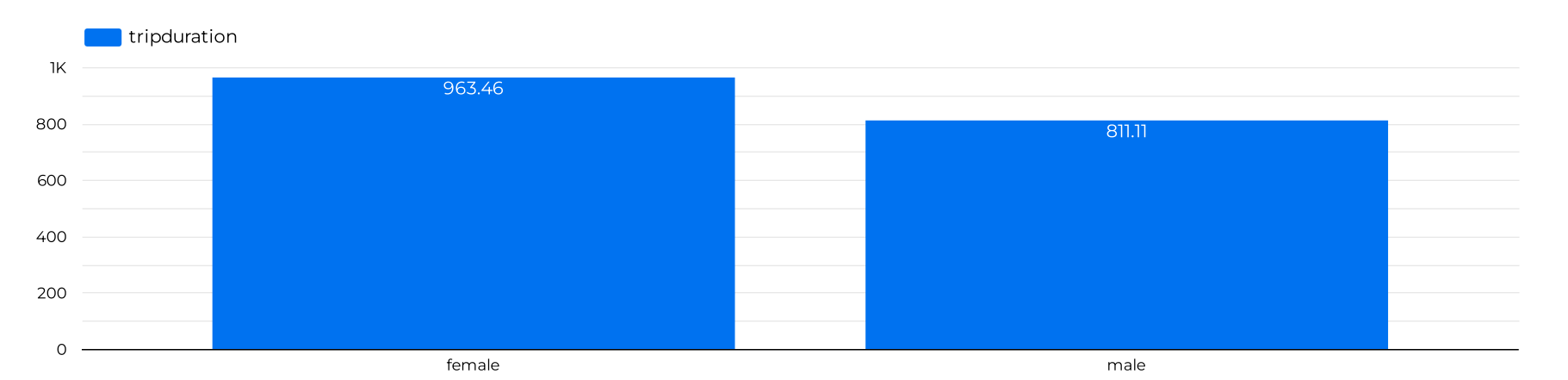
* Mayoría abrumadora de "Subscribers".
* **Insight:** Los usuarios recurrentes dominan el sistema (más de 35M viajes frente a solo 377K de "Customers").

## 6.2 Viajes por Mes



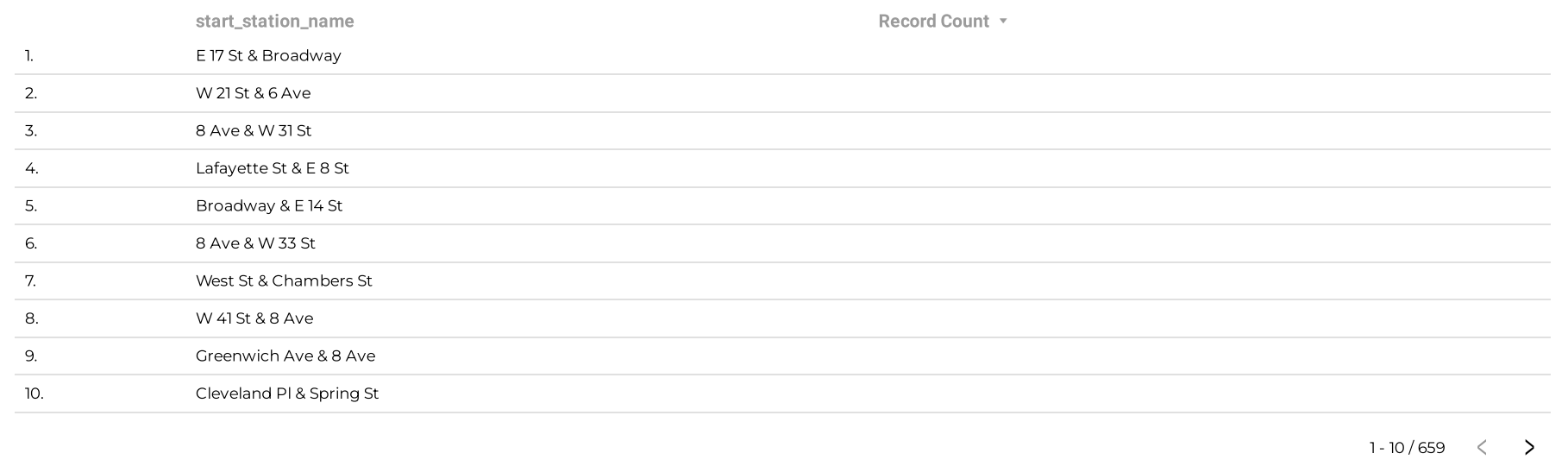
* Peak estacionales entre mayo y septiembre.
* **Insight:** Importante para planificación operativa y mantenimiento preventivo.

## 6.3 Promedio de Duración por Género



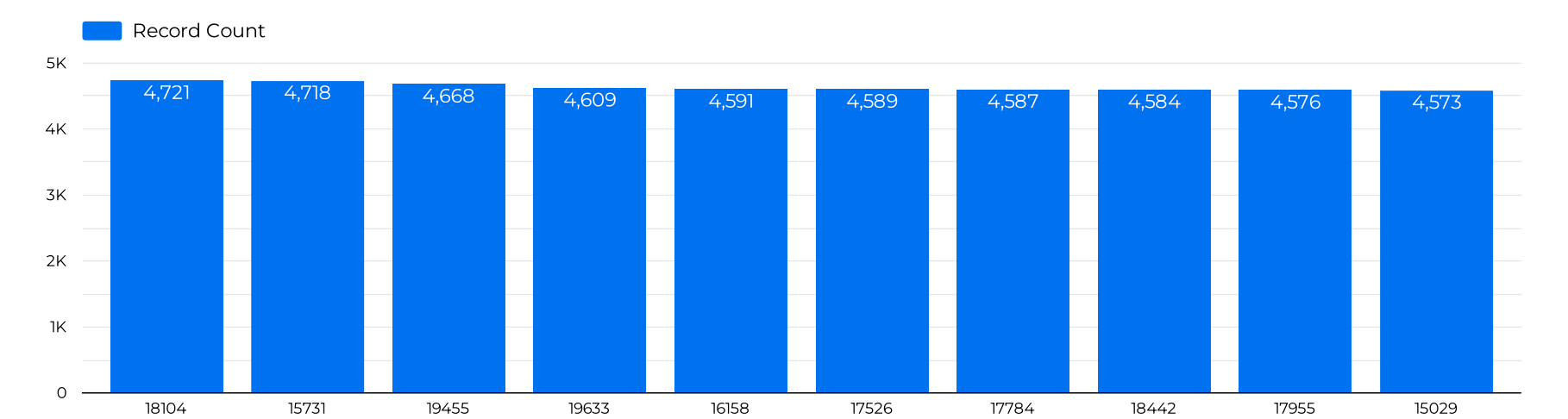
* Las mujeres tienden a tener viajes más largos.
* **Insight:** Posible personalización de servicios.

## 6.4 Estaciones Más Usadas



* *E 17 St & Broadway* es la más activa.
* **Insight:** Punto estratégico para expansión o campañas.

## 6.5 Bicicletas Más Usadas



* Algunas bicicletas tienen más de 4.500 viajes.
* **Insight:** Mantenimiento preventivo y ciclo de vida útil.

# 7. Conclusión

El desarrollo de este proyecto permitió aplicar de forma integral los conceptos fundamentales del procesamiento de grandes volúmenes de datos en entornos Big Data. A través del uso de BigQuery, se logró ejecutar un flujo completo que abarcó desde la **ingesta directa desde fuentes públicas**, pasando por una **transformación y limpieza rigurosa**, hasta la **unión de datos clave** para enriquecer el análisis visual.

Los resultados de la limpieza fueron validados mediante herramientas que confirmaron la integridad y calidad del dataset final, con un 100% de valores válidos en más de **35 millones de registros**. Esta validación garantiza la confiabilidad de los datos utilizados en las visualizaciones.

La **unión entre los viajes y la información detallada de las estaciones** permitió generar análisis más ricos y visualmente interpretables, facilitando la comprensión de patrones de uso, rutas más frecuentes y estaciones críticas en el sistema de CitiBike.

Finalmente, los dashboards elaborados en Looker Studio reflejan insights clave sobre el comportamiento de los usuarios, su distribución por género, estacionalidad en los viajes y uso de infraestructura. Esto demuestra no solo habilidades técnicas en manipulación de datos, sino también la capacidad de traducir información compleja en decisiones operativas y estratégicas.

Este proyecto representa un ejemplo exitoso del uso de tecnologías de Big Data en un entorno real, respondiendo a las necesidades del negocio con una solución robusta, validada y visualmente efectiva.