

Bootcamp Data Analysis-Montella Yanina

1. Título del Proyecto:	2
2. Objetivo General:	2
El objetivo de este proyecto es realizar un análisis del sistema de bicicletas compartidas de Austin, Texas, para identificar patrones de uso, determinar factores clave que influyen en la demanda y la oferta de bicicletas, y proponer recomendaciones basadas en datos para mejorar la eficiencia operativa y la satisfacción del usuario.	2
2.1. Justificación de la Elección del Proyecto	2
3. Dataset a Utilizar:	2
4. Objetivos Específicos del Análisis:	6
Plan de métricas	6
5. Limpieza y preparación de los datos.	7
Diagrama del diseño del datamart	8
Script para la creación del datamart	9
Algunas de las modificaciones realizadas en Power BI	11
6. Herramientas Utilizadas:	12
7. Interpretación y Presentación de Resultados	13
Pestaña General	13
Pestaña Estaciones	14
Pestaña Usuarios	15
Rutas Frecuentes	17
CONCLUSIONES GENERALES	18

Proyecto de Data Analysis: Sistema de Bicicletas Compartidas de Austin

1. Título del Proyecto:

Análisis del Uso del Sistema Austin Bike Share para la Optimización de Operaciones y la Experiencia del Usuario.

2. Objetivo General:

El objetivo de este proyecto es realizar un análisis del sistema de bicicletas compartidas de Austin, Texas, para identificar patrones de uso, determinar factores clave que influyen en la demanda y la oferta de bicicletas, y proponer recomendaciones basadas en datos para mejorar la eficiencia operativa y la satisfacción del usuario.

2.1. Justificación de la Elección del Proyecto

Los sistemas de bicicletas compartidas son una tendencia creciente en ciudades de todo el mundo, incluida Argentina. Analizar un sistema ya establecido como el de Austin nos permite comprender desafíos y oportunidades que son escalables y aplicables a contextos urbanos similares, incluso a nivel local. Las lecciones aprendidas sobre optimización, logística y comportamiento del usuario son directamente transferibles.

Así mismo el dataset ofrece un escenario real y útil para aplicar habilidades de análisis de datos ya que la combinación de información de viajes (bikeshare_trips) y estaciones (bikeshare_stations) permite un análisis desde varias dimensiones.

3. Dataset a Utilizar:

bigquery-public-data.austin_bikeshare

- **Fuente:** Datos obtenidos de una plataforma bigQuery base: austin_bikeshare.
- **Origen:** SQL.

La misma contiene 2 tablas:

1-bikeshare_stations (información sobre las estaciones)

2-bikeshare_trips (información sobre los viajes)

Descripción breve de sus campos:

bikeshare_stations. Tamaño: 101 registros.	
station_id:	clave única por estación
name	nombre de la estación
name: status:	Indica el estado operativo actual de la estación. Los valores existentes son "active" (activa), "closed" (cerrada),
location	Indican la latitud y longitud geográfica de la estación.
address	Dirección de la estación
alternate_name:	nombre alternativo o secundario para la estación (no siempre existe).
city_asset_number	número de identificación o un código de activo que la Ciudad de Austin.
property_type	<p>Tipo de terreno o espacio donde está ubicada la estación de bicicletas, específicamente en relación con las regulaciones de estacionamiento o el uso del suelo.</p> <p>Valores posibles:</p> <p>nonmetered_parking: La estación está ubicada en un área de estacionamiento no tarifado (es decir, gratuito o sin parquímetro). Esto podría ser una calle residencial sin restricciones de estacionamiento, un estacionamiento público sin costo, o un área designada donde no se cobra por dejar el vehículo.</p> <p>paid_parking: La estación se encuentra en un área de estacionamiento tarifado (con parquímetro o en un estacionamiento pago).</p> <p>parkland: La estación está ubicada dentro de un parque o un área verde designada.</p> <p>sidewalk: La estación está ubicada en una acera o vereda. Es una ubicación muy común para las estaciones de bicicletas compartidas en áreas urbanas, integrada directamente en el espacio peatonal.</p> <p>undetermined_parking: El tipo de estacionamiento o la regulación de la propiedad en la que se encuentra la estación no ha sido determinado o no está claramente clasificado en las otras categorías. Puede ser un valor por defecto o para ubicaciones que no encajan.</p>

number_of_docks	Representa el número total de "docks" o soportes disponibles en una estación específica donde los usuarios pueden recoger o dejar una bicicleta. Es el espacio físico que tiene la estación para almacenar bicicletas.
power_type	<p>Indica la fuente de energía utilizada para alimentar la estación en sí, especialmente para funciones como la carga de bicicletas eléctricas o los sistemas de bloqueo/desbloqueo.</p> <p>Valores posibles:</p> <p>solar:: La estación utiliza energía solar para su funcionamiento. Esto es común en sistemas de bicicletas compartidas, donde los paneles solares pueden cargar las baterías de las bicicletas eléctricas y/o alimentar los sistemas de la estación (pantallas, lectores de tarjetas, etc.).</p> <p>non-metered: La estación está conectada a la red eléctrica tradicional, pero su consumo de energía no se mide individualmente o no se factura por un medidor específico para esa estación. Esto podría significar que la electricidad se obtiene de una fuente compartida o que está incluida en un acuerdo general de servicios públicos para una propiedad más grande.</p>
footprint_length	Representa la medida de la longitud de la superficie que la estación de bicicletas abarca en el suelo. Esto incluiría el espacio ocupado por los anclajes de las bicicletas (docks), la terminal de pago (kiosk), y cualquier otra infraestructura fija de la estación dispuesta en una línea. Unidad de Medida: Se mide en pies (feet).
footprint_width	Indica el ancho de la huella física que la estación ocupa en el suelo. Complementa a footprint_length para describir las dimensiones rectangulares o aproximadas del área de la estación.
notes	Indica una observación o descripción de la estación.
council_district	Indica el distrito del consejo municipal de la ciudad de Austin al que pertenece cada estación de bicicletas. Austin está dividida en 10 distritos geográficos, cada uno representado por un miembro del consejo de la ciudad
image	este campo tiene valor null para todos los registros.
modified_date	fecha y hora de modificación del registro

bikeshare_trips. Tamaño: 2,271,152 registros.	
trip_id:	clave única de cada viaje
subscriber_type	Clasifica al usuario que realizó el viaje. Según la suscripción que tiene al servicio.
bike_id:	Id único de la bicicleta usada en el viaje.
bike_type	Indica el tipo de bicicleta. Posibles valores: classic(clásica), electric(eléctrica).
start_time	La fecha y hora exacta en que comenzó el viaje
start_station_id	El identificador numérico de la estación donde comenzó el viaje. Este ID se corresponde con el <i>station_id</i> en la tabla <i>bikeshare_stations</i> .
start_station_name	Nombre de la estación donde comenzó el viaje.
end_station_id	El identificador numérico de la estación donde finalizó el viaje. Este ID se corresponde con el <i>station_id</i> en la tabla <i>bikeshare_stations</i> .
end_station_name	Nombre de la estación donde finalizó el viaje.
duration_minutes	Indica la duración total del viaje, expresada en minutos

4. Objetivos Específicos del Análisis:

- **Uso General:**
 - Evaluar la evolución y duración promedio de los viajes a lo largo del tiempo (meses, años).
 - Analizar horas pico y días de la semana de mayor y menor actividad
- **Estaciones:**
 - Estaciones de origen y destino más frecuentes
 - Analizar si existen estaciones que con frecuencia se quedan sin bicicletas (déficit de oferta) o con demasiadas bicicletas (exceso de oferta)
 - Analizar el flujo neto de bicicletas entre las estaciones (entradas vs. salidas) y su relación con los atributos de la estación (*property_type*, *power_type*)
- **Usuarios:**

- Analizar diferencias en los patrones de uso (duración del viaje, estaciones, horarios) entre los distintos tipos de usuarios (subscriber_type).
- Proporción de suscriptores o usuarios ocasionales y su evolución.
- **Rutas Frecuentes:**
 - Analizar cuáles son las rutas más frecuentes (estación de origen a estación de destino)

Plan de métricas

Tipo	Ord	Nombre	Definición de Negocio	Puntos de Vista	Como calculamos	Formula
Métricas de Uso General	1	Total de Viajes	Número total de viajes registrados en el sistema.	Tiempo, Estación, Tipo de Usuario	Conteo de todas las transacciones de viaje.	COUNT(trip_id)
	2	Duración Promedio del Viaje	El tiempo promedio que los usuarios emplean en un viaje.	Tiempo, tipo de Usuario	Suma total de la duración de todos los viajes dividida por el total de	AVG(duration_minutes)
	3	Viajes por Hora/Día	Conteo de viajes por hora del día y día de la semana para identificar patrones de actividad.	Tiempo	Agrupando los viajes por hora y día de la semana.	COUNT(trip_id) GROUP BY HOUR(start_time), DAYOFWEEK(start_time)
Métricas de Estaciones	4	Estaciones Activas	Cantidad de estaciones activas	Estación	Conteo de viajes agrupados por start_station_id	COUNT(trip_id) GROUP BY start_station_id WHERE status = "active"
	5	Cantidad de viajes por tipo de estación (electrica /clásica)	Cantidad de viajes según ubicación geográfica	bike_type, Estación		Count(trip) GROUP BY bike_type
	6	Cantidad de viajes por tipo la propiedad de la estación (parking/non parking)	Cantidad de viajes según ubicación geográfica	property_type, Estación		Count(trip) GROUP BY property_type
	7	Estaciones mas frecuentes según ubicación	Cantidad de viajes según ubicación geográfica	Ubicación, Estación		Count(trip) GROUP BY start_station_id
	8	Estaciones Cerradas	Cantidad de estaciones cerradas	Estación	Conteo de viajes agrupados por start_station_id	COUNT(trip_id) GROUP BY start_station_id WHERE status = "close"
Métricas de Usuarios	9	Proporción de los distintos Suscriptores	Distribución porcentual de los viajes entre los distintos tipos de subscriber_type.	Usuario, Tiempo	Conteo de viajes por subscriber_type dividido por el total de viajes.	COUNT(trip_id) GROUP BY subscriber_type
	10	Duración Promedio por Tipo de Usuario	Duración promedio del viaje según el tipo de suscriptor.	Usuario, Tiempo		AVG(duration_minutes) GROUP BY subscriber_type
	11	Cantidad minima de viajes por usuario	Cantidad minima de viajes por tipo de usuario	Usuario, Tiempo		MIN(trip_id) GROUP BY subscriber_type
	12	Cantidad max de viajes por usuario	Cantidad minima de viajes por tipo de usuario	Usuario, Tiempo		MAX(trip_id) GROUP BY subscriber_type
Métricas de Rutas	13	Estaciones Más Frecuentes (Origen)	Las estaciones desde donde se inician más viajes.	Estación, Tiempo	Conteo de viajes agrupados por start_station_id	COUNT(trip_id) GROUP BY start_station_id
	14	Estaciones Más Frecuentes (Destino)	Las estaciones donde se finalizan más viajes.	Estación, Tiempo	Conteo de viajes agrupados por end_station_id	COUNT(trip_id) GROUP BY end_station_id

5. Limpieza y preparación de los datos.

Los datos se obtienen de BigQuery, de la base de datos `bigrquery-public-data.austin_bikeshare` y se someten a un proceso ETL para ser cargados con el diseño del datamart en BigQuery, en la base de datos `dataanalisis-461715`.

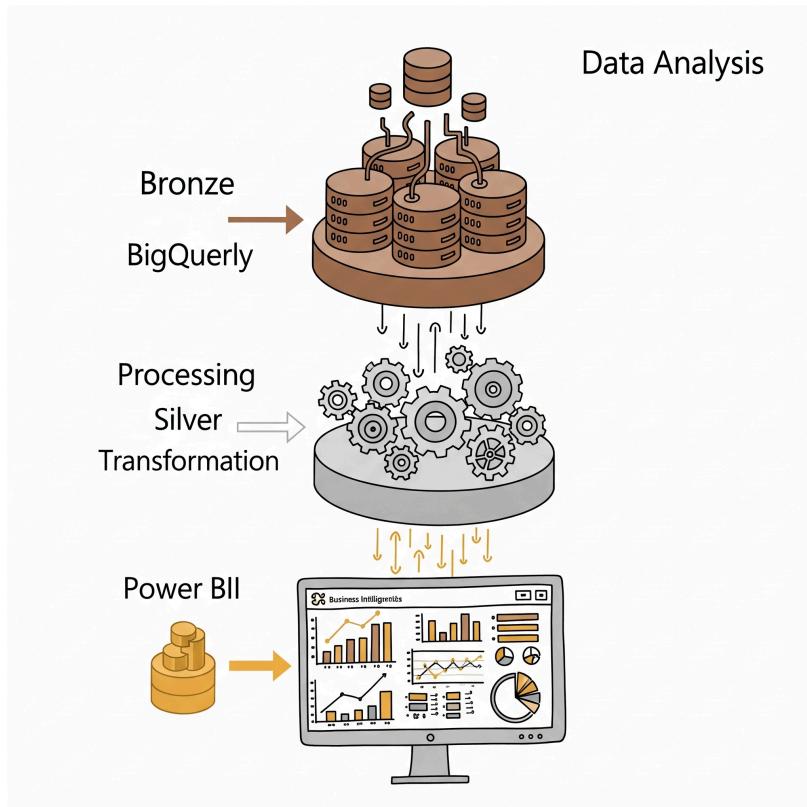
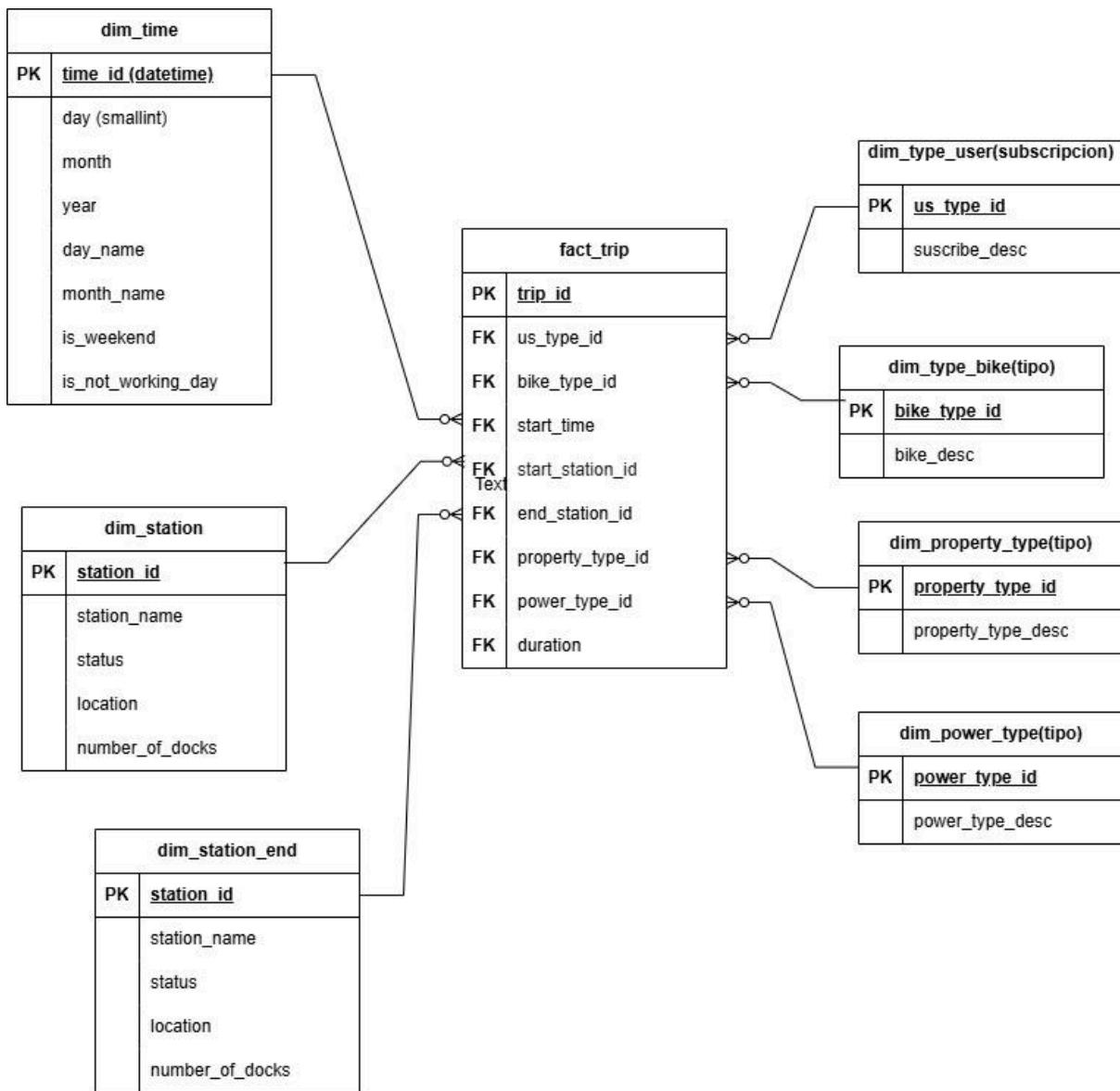


Diagrama estrella del diseño del datamart



Script para la creación del datamart

```
CREATE TABLE `dataanalisis-461715.bike_share.dim_station` AS
SELECT
station_id, name, status, location, COALESCE(number_of_docks, 0) as
colenumber_of_docks
FROM `bigquery-public-data.austin_bikeshare.bikeshare_stations` AS s
WHERE station_id IN( select min(station_id) )
    FROM `bigquery-public-data.austin_bikeshare.bikeshare_stations` GROUP BY name)
UNION ALL
SELECT 0 as station_id,
'Sin Especificar' as name,
'0' as status,
NULL as location,
0 as colenumber_of_docks
```

Se reemplazan los valores nulos por 0 y se agrega como nombre 'No especificado'. Del análisis surge que las estaciones que tienen null son aquellas que tienen estado="closed" (cerrada). Se toman el id mínimo porque en algunos casos el nombre estaba duplicado.

```
CREATE TABLE `dataanalisis-461715.bike_share.dim_user_type` AS
WITH distinct_Susbscriber AS (
    SELECT DISTINCT subscriber_type
    FROM `bigquery-public-data.austin_bikeshare.bikeshare_trips`
)
SELECT
    ROW_NUMBER() OVER(ORDER BY subscriber_type) AS us_type_id, -- Añadimos un ORDER
    BY para que el ROW_NUMBER sea determinístico
    COALESCE(subscriber_type, 'No especificada') AS us_type
FROM
    distinct_Susbscriber;
```

Como el campo subscriber_type era tipo varchar y lo necesitaba para crear una dimensión en base a este dato, con lo cual decidí crear una clave numérica única en la dimensión. La misma reemplazará al texto de la tabla bikeshare_trips en la creación de la fact_trip. Este trabajo se realizó para la dim_type_bike, dim_user_type, dim_power_type y dim_property_type.

```

CREATE TABLE `dataanalisis-461715.bike_share.dim_type_bike` AS
WITH distinct_Bike AS (
    SELECT DISTINCT bike_type
    FROM `bigquery-public-data.austin_bikeshare.bikeshare_trips`
)
SELECT
    ROW_NUMBER() OVER(ORDER BY bike_type) AS byke_type_id,
    COALESCE(bike_type, 'No especificado') AS byke_type
FROM
    distinct_Bike;

```

```

CREATE TABLE `dataanalisis-461715.bike_share.dim_property_type` AS
WITH distinct_Property AS (
    SELECT DISTINCT property_type as property_type
    FROM `bigquery-public-data.austin_bikeshare.bikeshare_stations`
)
SELECT
    ROW_NUMBER() OVER(ORDER BY property_type) AS property_type_id,
    COALESCE(property_type, 'sin propiedad') AS property_type
FROM
    distinct_Property;

```

```

CREATE TABLE `dataanalisis-461715.bike_share.dim_power_type` AS
WITH distinct_Power AS (
    SELECT DISTINCT power_type as power_type
    FROM `bigquery-public-data.austin_bikeshare.bikeshare_stations`
)
SELECT
    ROW_NUMBER() OVER(ORDER BY power_type) AS power_type_id,
    COALESCE(power_type, 'sin clasificar') as power_type
FROM
    distinct_Power;

```

```

CREATE TABLE `dataanalisis-461715.bike_share.fact_trip` AS
SELECT trip_id, COALESCE(us_type_id,0) as us_type_id, COALESCE(bike_type_id,0)as
bike_type_id, start_time,
COALESCE(COALESCE(start_station_id, start_station_id2),0) as start_station_id,
COALESCE(end_station_id , '0') as end_station_id , COALESCE(property_type_id,0) as
property_type_id,
COALESCE(power_type_id,0) as power_type_id, duration
FROM (

```

```

SELECT trip_id,
  (SELECT COALESCE(us_type_id,0) FROM
`dataanalisis-461715.bike_share.dim_user_type` AS u WHERE
u.us_type=t.subscriber_type) as us_type_id,
  (SELECT COALESCE(byke_type_id,0) FROM
`dataanalisis-461715.bike_share.dim_type_bike` AS b WHERE b.byke_type=t.bike_type)
as bike_type_id,
t.start_time,
  (SELECT CASE WHEN ss.name IN('Main Office', 'Repair Shop') THEN 1001 ELSE
COALESCE(station_id, 0) END
  FROM `dataanalisis-461715.bike_share.dim_station`as ss where
TRIM(ss.name)=TRIM(t.start_station_name)) as start_station_id,
  (SELECT CASE WHEN ss.name IN('Main Office', 'Repair Shop') THEN 1001 ELSE
COALESCE(station_id, 0) END
  FROM `bigquery-public-data.austin_bikeshare.bikeshare_stations`as ss where
ss.station_id=t.start_station_id) as start_station_id2,
  COALESCE(t.end_station_id , '0') as end_station_id,
  (SELECT COALESCE(property_type_id,0) FROM
`bigquery-public-data.austin_bikeshare.bikeshare_stations` AS s
  INNER JOIN `dataanalisis-461715.bike_share.dim_property_type` AS p ON
p.property_type=s.property_type
  WHERE t.start_station_id =s.station_id) as property_type_id ,
  (SELECT COALESCE(power_type_id,0) FROM
`bigquery-public-data.austin_bikeshare.bikeshare_stations` AS s
  INNER JOIN `dataanalisis-461715.bike_share.dim_power_type` AS w ON
s.power_type=w.power_type
  WHERE t.start_station_id =s.station_id) as power_type_id ,
  t.duration_minutes as duration
  FROM `bigquery-public-data.austin_bikeshare.bikeshare_trips` AS t
)

```

En todos los casos se reemplazan las claves null por 0 en la fact.

Para el caso de la columna start_station_id se toma el join por name o por station_id para completar los datos ya que en algunos casos el station_id no era suficiente porque figuraba como nulo(no así el campo name).

Algunas de las modificaciones realizadas en Power BI

1-Se quitaron los valores duplicados para la tabla dim_station ya que desde Big Query inserté un registro con clave =0 y nombre 'No especificado' para aquellos viajes que no tenían identificada la estación(null) y resultó que la tabla station ya tenía un registro con station_id =0;

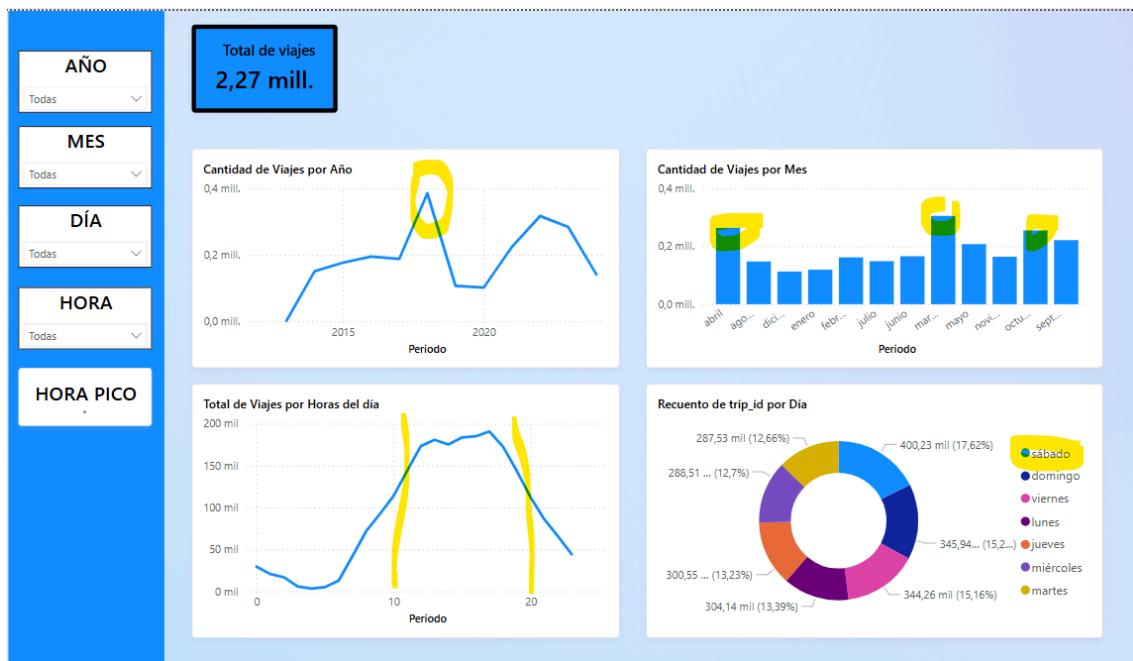
- 2-En la fact_trip: se crea una columna calculada que solo extraiga el valor date(ya que tengo un campo datetime) para poder relacionarlo con la dimensión tiempo.
- 3-En la fact_trip: se crea una columna calculada de minuto del día para poder relacionarla con la dim_hora. Esto me permite realizar un análisis sobre las horas de uso más frecuentes, si son horas pico, etc.
- 4-Se duplica la tabla dim_station en Power Query en otra llamada dim_station_End para las estaciones de llegada. Es la forma estándar y más efectiva que encontré para resolver el problema manejar dimensiones de rol de juego en Power BI ya que mi fact_trip tiene 2 campos con station_id(star y end).
- 5-Se crearon varias medidas que fueron utilizadas mayormente para la creación de las tarjetas del dashboard entre ellas: de estaciones cerradas y estaciones activas

6. Herramientas Utilizadas:

- BigQuery: Para almacenamiento, consulta de datos (SQL), análisis exploratorio de datos y proceso para la transformación de la base de datos original al datamart (ETL).
- Power BI: Para limpieza final de los datos y visualización y creación de dashboards .
- Github para la creación del repositorio y publicación del proyecto.

7. Interpretación y Presentación de Resultados

Pestaña General



- Evolución y duración promedio de los viajes a lo largo del tiempo (meses, años):

Se observa un incremento de los viajes en 2018, esto puede deberse a eventos importantes en la ciudad o factores externos. Austin es conocida por albergar grandes eventos como SXSW (South by Southwest), el Gran Premio de Fórmula 1, o festivales de música. Puede que en 2018, que tuvo un número particularmente alto, haya tenido lugar este tipo de eventos, y el servicio de bicicletas compartidas se promocionó o utilizó como medio de transporte alternativo.

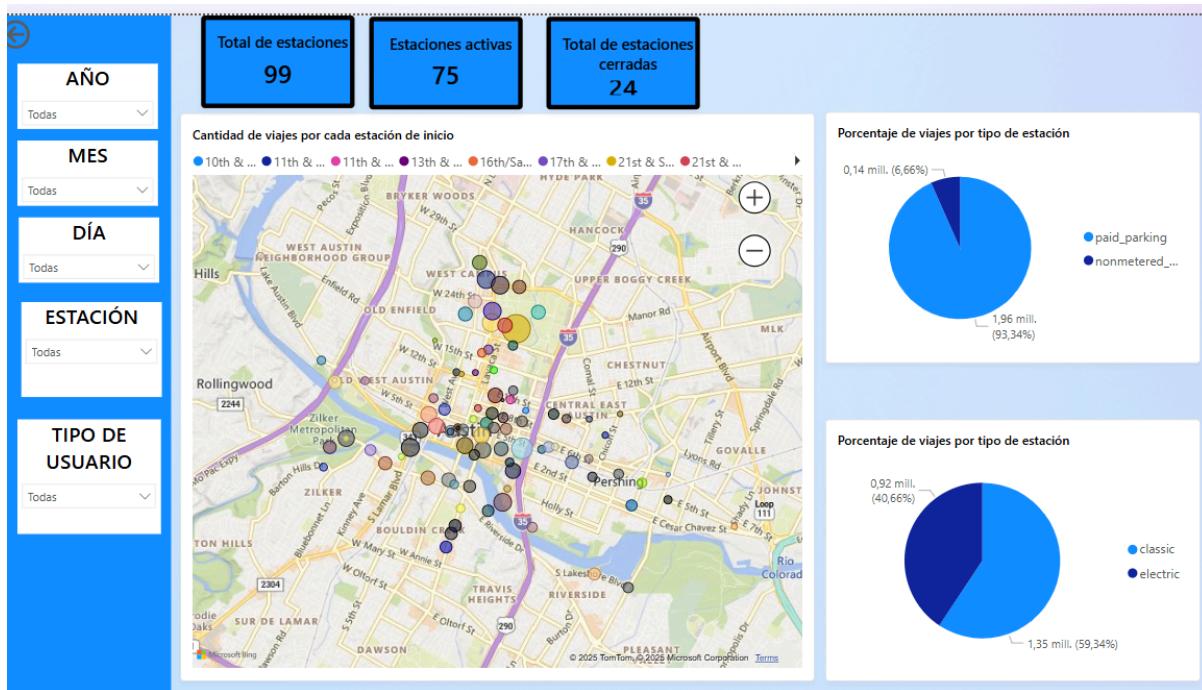
Pero debido al consumo lineal que se observa en ese año a través de todos los meses me inclinaría más por que se deba a la madurez del mercado.

Para 2018, el servicio podría haber alcanzado un punto de "madurez" donde ya era bien conocido por la población local y los visitantes, pero aún no había enfrentado la saturación o la competencia de otros modos de transporte que podrían surgir en años posteriores (como patinetes eléctricos).

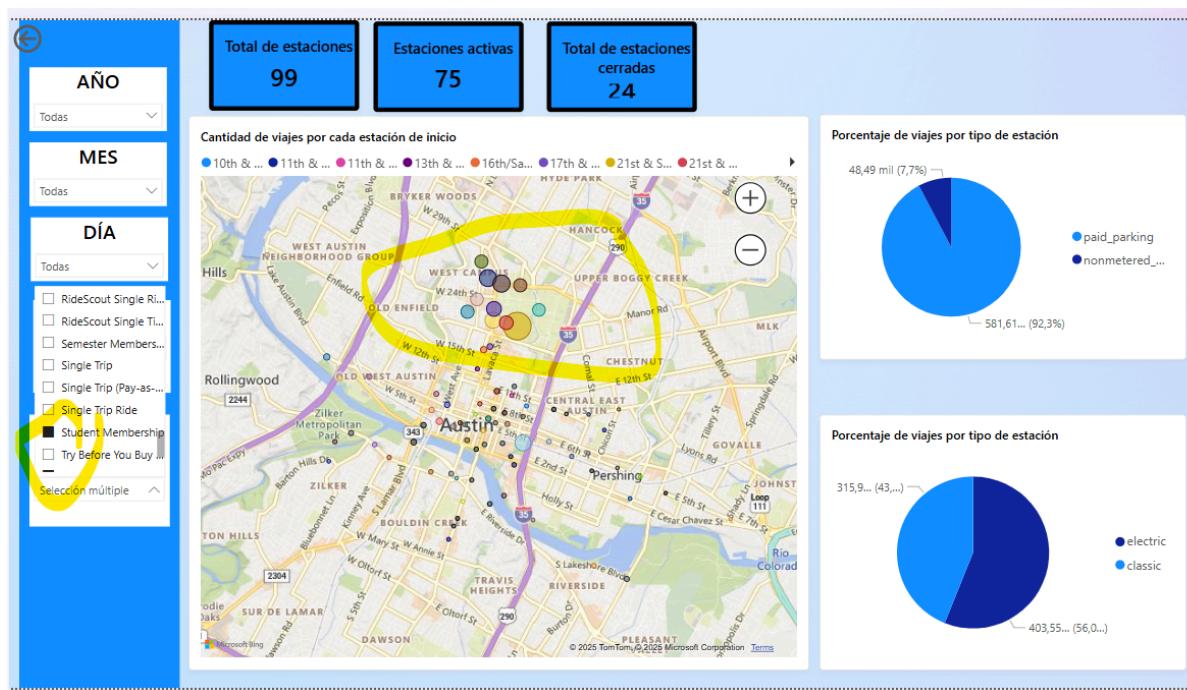
En general se destacan en todos los años los meses de marzo, abril y octubre como los de mayor consumo.

- Horas Pico: Se identifican horarios pico por la tarde (aproximadamente entre las 12 y las 17) lo que indica que el uso se da en la franja horaria de la tarde. Esto quizás pueda reflejar que el uso de bicicletas sea más utilizada como fines recreativos y de paseo y no como medio de transporte en sí, ya que además la mayor frecuencia se da los días sábados.

Pestaña Estaciones



Visualización filtrando por usuarios UT Student Membership y UT Student Membership



- Se puede observar que las estaciones con mayor uso son las que se encuentran en el centro de la ciudad, lo cual puede ratificar el uso de las bicicletas como medio de paseo. Pero el foco más importante de consumo se encuentra en la zona cercana al West Campus que es un vecindario conocido por su alta densidad de población

estudiantil, lo cual puede indicar una segmentación dentro de las clasificación del público que usa este servicio.

- Las estaciones clásicas superan ampliamente en consumo a las estaciones eléctricas.
- La prevalencia de paid_parking refuerza la idea de que los usuarios están utilizando las bicicletas principalmente para desplazamientos de corta a media distancia dentro de estas zonas céntricas o bien conectadas. Esto incluye viajes al trabajo, reuniones, visitas a tiendas, restaurantes o atracciones. Se aleja de un modelo donde el uso es predominantemente recreativo en zonas residenciales o parques sin costo de estacionamiento.
-

Pestaña Usuarios



Filtrada por año 2018



- subscriber_type: Generalmente se encuentran tres categorías principales:
 - 1-UT Student Membership (Membresía Estudiantil de la Universidad de Texas) que es una membresía estudiantil altamente específica y a menudo más subsidiada, diseñada exclusivamente para los estudiantes, profesores y personal de la Universidad de Texas en Austin (UT Austin).
 - 2-Student Membership (Membresía Estudiantil General)
 - 3-La suscripción "local365" es una membresía anual diseñada para **residentes de Austin** que buscan un acceso frecuente y asequible al sistema de bicicletas compartidas. Generalmente, ofrece: tarifa plana anual, viajes diarios incluidos y costo adicional por tiempo extra

La combinación de usuarios "local365" y estudiantes como los principales usuarios del servicio demuestra que tanto estudiantes como residentes locales utilizan las bicicletas compartidas como un medio de transporte habitual para moverse por la ciudad. Esto sugiere que el sistema es efectivo para viajes cortos y frecuentes, como ir al trabajo, a la universidad, hacer recados o conectar con el transporte público. Lo que destaca la importancia del servicio como una opción de transporte diaria y accesible para los residentes y la comunidad universitaria.

Con respecto a la duración de los viajes se puede observar que quienes realizan los viajes más largos son los "Single Trip" (Viaje Único) que es el modelo de uso más básico y casual del sistema de bicicletas compartidas. La misma no implica una membresía de largo plazo ni un compromiso. Esto revela una dinámica muy diferente en comparación con los suscriptores habituales.

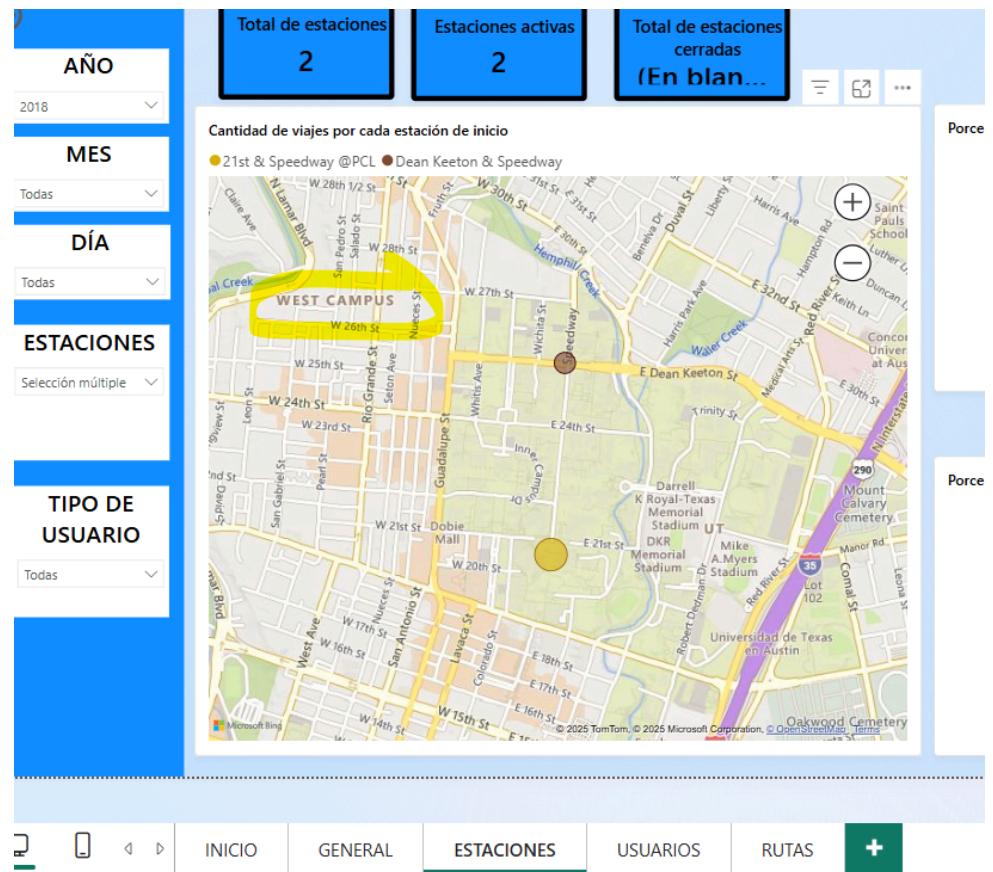
De aquí se puede deducir que los viajes largos por parte de usuarios de "Single Trip" podrían ser un indicador de que están utilizando las bicicletas con fines recreativos, turísticos o de exploración. Es menos probable que sean viajes diarios de transporte, que suelen ser más cortos y directos. Recordando que estos últimos son los más frecuentes, y que en el caso de los estudiantes y residente (usuarios habituales) tienden a realizar viajes más cortos y frecuentes, con patrones más consistentes relacionados con el transporte diario

Rutas Frecuentes

Filtrado por año 2018.



Observando las 2 estaciones principales en el mapa



- La identificación de las combinaciones start_station_id y end_station_id más comunes revelará que los viaje más populares dentro de la ciudad son los que se realizan en la cercanía de West Campus que como mencionamos anteriormente es

un vecindario conocido por su alta densidad de población estudiantil. Lo cual refuerza la teoría de la importancia del público estudiantil en este sistema.

CONCLUSIONES GENERALES

Este análisis podría ser útil para modelar el comportamiento de las estaciones en función de su ubicación y características. Entender las diferencias del uso de los distintos usuarios permite adaptar estrategias de marketing (promover suscripciones para commuters, pases diarios para turistas) y optimizar la disponibilidad de bicicletas según el perfil de usuario predominante en ciertas áreas o momentos.

Conocer las rutas más frecuentes puede aportar información útil tanto para la planificación urbana y de infraestructura, como para la creación de carriles bici dedicados, la optimización de la ubicación de nuevas estaciones, o la evaluación de la demanda de bicicletas en esos trayectos específicos. También puede servir para la publicidad y promoción de rutas turísticas.

El análisis de datos en este proyecto permite tomar decisiones informadas y estratégicas, mejorar la experiencia del usuario, desarrollar y expandir el servicio en forma planificada y fundamentada.

En resumen, el análisis de datos convierte el "ruido" de la información bruta en conocimiento accionable, que es esencial para que el sistema de bicicletas compartidas no sólo opere, sino que prospere, sirva eficazmente a la comunidad y se adapte a las cambiantes necesidades de movilidad urbana.