Title : How does a bike-share navigate speedy success?

Name: Prateek Sharma

Date: 10 September 2024

Business Task: How do annual members and casual riders use Cyclistic bikes differently?

Objective: It's about a bike-share company in Chicago, this project aims to analyse and recognize the usage patterns of casual riders versus annual members. The primary goal is to uncover significant differences in how these two groups use Cyclistic bikes. The insights gained from this analysis will be used to design a new marketing strategy with the objective of converting casual riders into annual members.

Cyclistic: *A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike.*

Data Source: Cyclistic bike-share data has been used. This dataset contains comprehensive records of bike trips taken by users, including trip start and end times, bike IDs, ride durations, and user types (casual or annual) and is sourced from Cyclistic's internal tracking system.

This is public dataset that you can use to explore how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit you from using riders' personally identifiable information(PII).

Data Preparation: Here, data spanning six months, from November 2023 to April 2024, was utilized. Upon acquiring the data from the source, I organized it into a structured folder hierarchy to manage the files effectively. Initially, I attempted to open the data files in Google Sheets; however, due to the large size of each file—approximately 3 to 5 lakh rows and 13 columns per month—it was not feasible.

To efficiently process and analyze the data, I opted to use SQL. The data was exported to Google BigQuery, where each month's dataset was imported into separate tables within the same dataset. This approach facilitated more robust

data processing and allowed for effective handling of the extensive volume of data.

Data Preparation and Cleaning :

1) Verification of Data Types -
   Initially, I verified the data types of all variables by inspecting the schema of the dataset to ensure that each column had the appropriate data type. This step is crucial for maintaining data integrity and ensuring accurate processing.
2) Combining Datasets –
   All six months of data were combined into a single dataset to facilitate comprehensive analysis. The dataset was aggregated using the following SQL code:

SQL CODE CHUNK ```

-- Combine all 6 months dataset into 1 single dataset by creating a new table named as combined table

CREATE TABLE `ornate-apricot-435209-g7.tripdata.combined_tripdata` AS (

SELECT * FROM `ornate-apricot-435209-g7.tripdata.202311_tripdata`

UNION ALL

SELECT * FROM `ornate-apricot-435209-g7.tripdata.202312_tripdata`

UNION ALL

SELECT * FROM `ornate-apricot-435209-g7.tripdata.202401_trip`

UNION ALL

SELECT * FROM `ornate-apricot-435209-g7.tripdata.202402_tripdata`

UNION ALL

SELECT * FROM `ornate-apricot-435209-g7.tripdata.202403_tripdata`

UNION ALL

SELECT * FROM `ornate-apricot-435209-g7.tripdata.202404_tripdata`

); ```

3) Data Quality Checks-
   Various quality checks were performed to identify null values and inconsistencies:

SQL CODE CHUNK ```

```
-- For checking varieties in start_station_name

SELECT DISTINCT start_station_name

FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`;


-- For checking varieties in start_station_id

SELECT DISTINCT start_station_id

FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`;


-- For checking varieties in end_station_name

SELECT DISTINCT end_station_name

FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`;


-- For checking varieties in end_station_id

SELECT DISTINCT end_station_id

FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`;


-- For checking member types

SELECT DISTINCT member_casual

FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`;


-- For checking bike types

SELECT DISTINCT rideable_type

FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`;
```

```
-- Checking longitude values
SELECT start_lng
FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`
WHERE CAST(start_lng AS STRING) NOT LIKE "-%";

SELECT end_lng
FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`
WHERE CAST(end_lng AS STRING) NOT LIKE "-%";
```

4) Data Cleaning-
   A new table was created by removing all null values and duplicates, and adding additional columns for ride length, day of week, and start hour. The SQL code used is:

SQL CODE CHUNK ```

```
-- Creating a cleaned data table
CREATE TABLE `ornate-apricot-435209-g7.tripdata.clean_tridpdata` AS (
 SELECT *
 FROM (
  -- Creating a new table by removing all null values
  WITH combined_tripdata_with_no_null AS (
   SELECT *
   FROM `ornate-apricot-435209-g7.tripdata.combined_tripdata`
   WHERE start_station_id IS NOT NULL
    AND start_station_name IS NOT NULL
    AND end_station_name IS NOT NULL
```

```
    AND end_station_id IS NOT NULL
  ),

  -- Creating a new common Table Expression (CTE) using previous CTE without any duplicate and adding new column of ride_length_minutes
  -- Also add new columns of weekdays and start hour for ride
  combined_tripdata_without_duplicates AS (
    SELECT *,
      TIMESTAMP_DIFF(ended_at, started_at, MINUTE) AS ride_length_minute,
      ROW_NUMBER() OVER (PARTITION BY ride_id ORDER BY started_at) AS row_num,
      EXTRACT(DAYOFWEEK FROM started_at) AS day_of_week,
      EXTRACT(HOUR FROM started_at) AS start_hour
    FROM combined_tripdata_with_no_null
  )

  SELECT *
  FROM combined_tripdata_without_duplicates
  WHERE row_num = 1
    AND ride_length_minute > 1
    AND ride_length_minute < 1440
 )
); ```
```

5) Verification-
   After cleaning, I verified the data to ensure it met the required conditions and to check for any remaining anomalies:

SQL CODE CHUNK ```

```
-- Checking for ride length less than 1 minute
SELECT *
FROM `ornate-apricot-435209-g7.tripdata.clean_tridpdata`
WHERE ride_length_minute < 1;

-- Checking for ride length greater than 1440 minutes (more than 24 hours)
SELECT *
FROM `ornate-apricot-435209-g7.tripdata.clean_tridpdata`
WHERE ride_length_minute > 1440;

-- Checking for missing ride length values
SELECT *
FROM `ornate-apricot-435209-g7.tripdata.clean_tridpdata`
WHERE ride_length_minute IS NULL;

-- Checking distinct values in day_of_week
SELECT DISTINCT day_of_week
FROM `ornate-apricot-435209-g7.tripdata.clean_tridpdata`;

-- Checking distinct values in start_hour
SELECT DISTINCT start_hour
FROM `ornate-apricot-435209-g7.tripdata.clean_tridpdata`;
```

```
```

Data Analysis:

1. Average Ride Length by Member Type -

Query: ```

SELECT member_casual,

  AVG(ride_length_minute) as Avg_ride_length

FROM `ornate-apricot-435209-g7.tripdata.clean_tripdata`

GROUP BY member_casual;

```

Summary:

a) This analysis calculates the average ride length (in minutes) for casual and annual members.

b) Provides insight into how long each type of member typically uses the bike per ride. This can help determine whether casual riders tend to take shorter or longer rides compared to annual members.

2. Ride Frequency by Member Type -

Query: ```

SELECT member_casual,

  COUNT(*) as num_rides

FROM `ornate-apricot-435209-g7.tripdata.clean_tripdata`

GROUP BY member_casual;

```

Summary:

a) This analysis counts the number of rides taken by casual and annual members.

b) Indicates the total usage frequency for each type of member. Helps in understanding the overall engagement of casual vs. annual members with the bike-share system.

3. Hourly Usage Patterns by Member Type -

Query: ```

SELECT member_casual,

start_hour,

count(*) as num_rides

FROM `ornate-apricot-435209-g7.tripdata.clean_tripdata`

GROUP BY member_casual, start_hour

ORDER BY member_casual, start_hour;

```

Summary:

a) This analysis shows the number of rides taken at each hour of the day for both casual and annual members.

b) Reveals peak usage times and hourly patterns for each member type. This can be useful for identifying high-traffic periods and optimizing bike availability.

4. Weekly Usage Patterns by Member Type -

Query: ```

SELECT member_casual,

day_of_week,

count(*) as num_rides

FROM `ornate-apricot-435209-g7.tripdata.clean_tripdata`

GROUP BY member_casual, day_of_week

ORDER BY member_casual, day_of_week;

```

Summary:

a) This analysis examines the frequency of bike usage across different days of the week for each member type.

b) Highlights which days see higher or lower bike usage and helps to identify weekly trends in usage.

5. Monthly Usage Trends by Member Type -

Query: ```

```
SELECT member_casual,
  EXTRACT(MONTH FROM started_at) as month,
  EXTRACT(YEAR FROM started_at) as year,
  COUNT(*) AS num_rides
FROM `ornate-apricot-435209-g7.tripdata.clean_tripdata`
GROUP BY member_casual, month, year
ORDER BY member_casual, month, year;
```

```

Summary:

a) This analysis tracks the number of rides per month for each type of member.

b) Provides insights into monthly trends and seasonality of bike usage. Helps in understanding how usage patterns change over time and identifying any seasonal effects.

## Supporting Visuals And Key Findings:

1. Average Ride Duration (Bar Chart)-
    - This chart clearly shows the difference in ride duration between casual riders and annual members, with casual riders riding nearly twice as long.
    - Casual riders have significantly longer ride durations compared to annual members.
    - The average ride length for casual riders is **20.58 minutes**, while for members, it's **11.43 minutes**. This suggests that casual riders use
```

the bikes for leisure or infrequent long trips, while annual members use them for shorter, more frequent trips, potentially for commuting.

2. Ride Frequency (Bar Chart)-

    - The stark contrast in ride counts emphasizes the regularity of member usage compared to the occasional rides by casual users.
    - Annual members account for a much higher number of rides (906,942 rides) compared to casual riders (312,466 rides).
    - This difference indicates that members are more likely to use the service regularly, which aligns with their subscription-based usage model, while casual riders are more occasional users.

3. Monthly Usage Trends (Line Chart)-

    - This visual depicts the seasonal fluctuations in ride frequency, with a steady decline for both groups during the winter and a gradual increase towards April.
    - Casual riders show a steep decline in usage during the winter months, with rides dropping from **70,254 in November 2023** to **17,108 in January 2024**.
    - Annual members also show a drop in winter usage, though less drastic than casual riders. This indicates that both groups ride less in colder months, but members continue to use the service more consistently than casual riders during this time.

4. Bike Usage Patterns by Time of Day (Heat Map)-

    - The color gradient shows the intensity of bike usage across different hours, highlighting when each group is most active during the day.
    - Heatmap shows that both casual and annual members have peak usage during late morning and afternoon hours (between 10 AM to 6 PM).
    - Annual members show significant usage during early mornings and late evenings as well, which could indicate commuting patterns, while casual riders are more concentrated in midday hours.

5. Bike Usage Patterns by Day of Week (Heat Map)-
    - This heat map visualizes how casual riders tend to ride more on weekends, while annual members have consistent usage across the week.
    - Casual riders are most active on **weekends**, while annual members have more consistent usage throughout the week, with a slight peak around **Wednesday and Friday**.
    - This supports the idea that casual riders use the service for leisure, while members are likely using it for regular weekday activities like commuting.

Top 3 Recommendations Based On Analysis:

1. Targeted Seasonal Promotions for Casual Riders:

- Insight: Casual riders show a significant drop in usage during the winter months (November to February).
- Recommendation: Offer seasonal discounts or limited-time winter promotions to encourage casual riders to continue using Cyclistic during the off-peak season. This could include discounted membership offers or reduced ride costs for longer rides, incentivizing casual riders to upgrade to an annual membership.

2. Weekday Commuter Incentives:

- Insight: Annual members have consistent weekday usage, likely for commuting, while casual riders tend to use the service more on weekends.
- Recommendation: Create weekday-specific promotions aimed at casual riders, such as a "commuter package" that allows for discounted rides during rush hours. Highlight the benefits of becoming a member for frequent commuters, such as unlimited rides, priority access, and more savings over time.

3. Customized Membership Plans Based on Usage Patterns:

- Insight: Casual riders have longer average ride durations compared to members, suggesting a preference for leisure or longer trips.

- Recommendation: Introduce a "flexible membership plan" tailored for casual riders who prefer longer, less frequent rides. This plan could offer features like monthly memberships, pay-per-use options with added benefits, or a hybrid model with discounts for long rides. Promote this as a stepping stone towards full annual membership.