Google Data Analytics Professional Certificate - Capstone Project
Author: Ashraf Un Noor
Date: 2025/04/20

Cyclistic Q1 2020 Data Analysis Report

How Member types differ in their usage of Cyclistic

1. Business Task Statement

The business task is to analyze historical trip data to determine how annual members and casual riders use Cyclistic bikes differently. This analysis will focus on identifying key patterns and behaviors—such as differences in ride duration, frequency, day and time of use—between the two user groups. The insights gained will help the marketing team better understand casual riders and inform strategies to convert them into annual members.

2. Description of Data Source

The dataset used for this analysis is the **Cyclistic Q1 2020 trip data**, obtained from the official Divvy Bike Share system's open data portal. The specific file used is titled **"Divvy_Trips_2020_Q1.csv"** and is publicly accessible at https://divvy-tripdata.s3.amazonaws.com/index.html.

This dataset contains **426,887 rows** of unique bike trip records and includes the following fields:

- **ride_id**: Unique identifier for each bike trip
- **rideable_type**: Type of bike used (*Only one type of bike was found in the data*)
- **started_at**: Start date and time of the trip
- **ended_at**: End date and time of the trip
- **start_station_name** and **start_station_id**: Name and ID of the trip's starting station
- **end_station_name** and **end_station_id**: Name and ID of the trip's ending station
- **start_lat** and **start_lng**: Latitude and longitude of the starting location
- **end_lat** and **end_lng**: Latitude and longitude of the ending location
- **member_casual**: User type (either "member" for annual subscribers or "casual" for non-subscribers)

This dataset provides the necessary information to analyze and compare the behavior of casual riders and annual members, helping to uncover usage trends and patterns.

3. Data Cleaning and Manipulation

BigQuery

Given the size and complexity of the dataset, the raw data was first imported into **Google BigQuery** to facilitate efficient exploration and analysis.

During the initial review, several data quality issues were identified:

- **Invalid timestamps**: Many records showed an `ended_at` time earlier than the `started_at` time, indicating clear errors. *(Appendix 1)*
- **Zero or extremely short durations**: A number of rides had the same `started_at` and `ended_at` times, or a duration of just a few seconds. These likely represent

failed attempts to unlock a bike or aborted rides. *(Appendix 1)*

- **Unusually long durations**: Some casual riders kept bikes for over **500 to 1,000+ minutes**, suggesting issues such as bikes not being docked properly or usage outside normal patterns. *(Appendix 1)*

To systematically identify outliers and better understand the distribution of ride durations, a new column called `ride_duration_min`*(Appendix 1)* was added to represent ride duration in minutes. This enabled the calculation of key metrics such as averages, medians, and percentiles, and supported more accurate comparative analysis between casual riders and annual members. A **Percentile analysis** was conducted using SQL. The key results from this analysis showed:

- The **median ride time (50th percentile)** was *9 minutes* for **Members** and *22 minutes* for **Casual riders**. *(Appendix 2)*
- The **99th percentile** of rides lasted up to 85 minutes, with a small portion exceeding this threshold. *(Appendix 2)*

To eliminate noise and potential data errors, records were filtered as follows:

- **Members**: Retained rides between the **10th(7 minutes) and 80th percentiles(41 minutes).** *(Appendix 2)*
- **Casual riders**: Retained rides between the **5th(5 minutes) and 95th percentiles(14 minutes).** *(Appendix 2)*

Afterwards a new Table was generated through BigQuery reducing the dataset from **426,887 rows** to **292,916 rows**, resulting in a more reliable and cleaner dataset for analysis. *(Appendix 3)*

Sample Size Selection for Analysis

To make the analysis more manageable without sacrificing statistical validity, a **sample size calculator** was used:

- **Population size**: 292,916

- **Confidence level**: 99%

- **Margin of error**: ~2.03%

This yielded a recommended sample size of **4,000**. The resulting sample was then exported to **Google Sheets** for further analysis and visualization. *(Appendix 4)*

Google Sheets

To further enrich the dataset for visualization, a new column was created in **Google Sheets** to extract the **day of the week** from each ride's `started_at` timestamp. This was done using the `WEEKDAY()` function, followed by formatting using the `TEXT()` functions to label the days (e.g., Monday, Tuesday). This transformation was essential for identifying trends by day and enabled more intuitive **visualizations in Tableau**.

4. Brief Summary

The analysis examined behavioral differences between annual and casual riders using the Cyclistic Q1 2020 dataset. After addressing data quality issues—including negative ride durations, implausibly short rides, and extreme outliers—the dataset was trimmed to include rides within the 5th to 95th percentile range, resulting in a refined dataset of approximately 293,000 entries. A statistically valid random sample of 4,000 rides was then extracted for focused analysis.

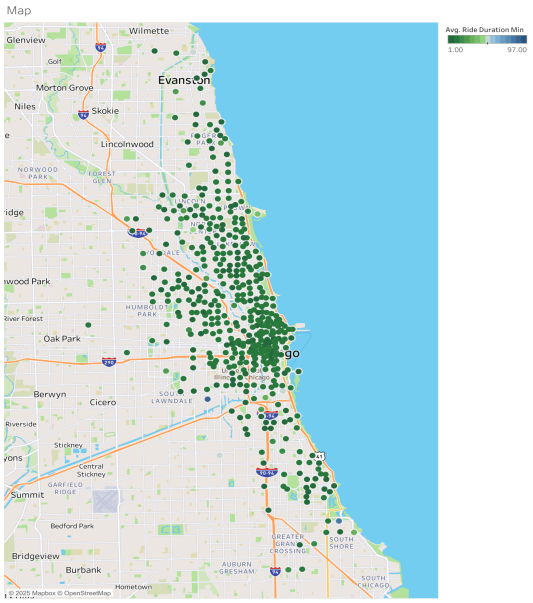
The results revealed that **annual members predominantly ride on weekdays**, suggesting usage for routine, day-to-day activities such as commuting. In contrast, **casual riders are more active on weekends** and, despite lower ride counts, demonstrate **longer ride durations**. This pattern suggests that casual riders are more likely engaged in leisure activities such as sightseeing or recreational use.

These insights highlight clear behavioral distinctions that can inform Cyclistic's targeted marketing strategies and membership growth initiatives.

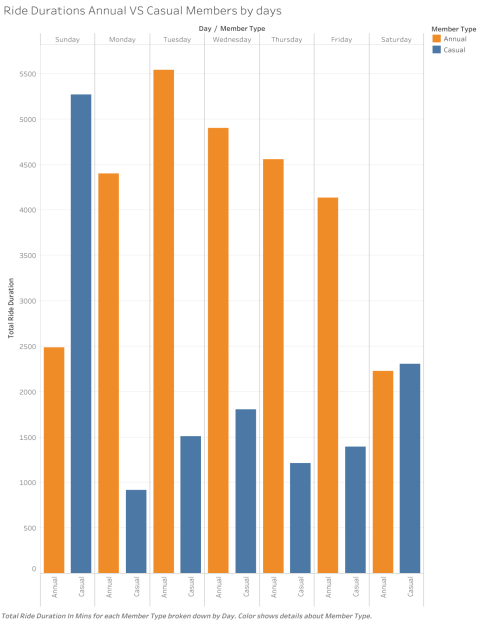
5. Data Visualization in Tableau ******Members*** *were referred to as* ***Annual*** *from here onwards.*

To uncover trends in user behavior and ride patterns, four Tableau visualizations were created using the cleaned and sampled dataset.

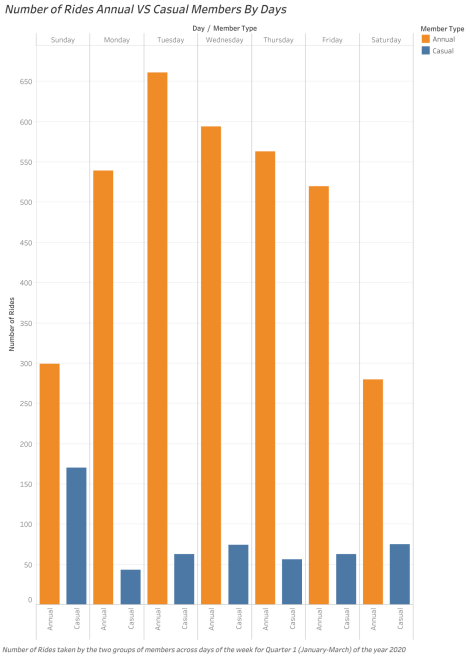
- ***Interactive Geographical Map:*** This map of Chicago enables clicking on individual stations to view the station's name, latitude/longitude, average ride duration from that station, and whether annual or casual members are the more frequent users. It helps visually identify which areas are more active among different member types with filters.



- ***Ride Duration by Member Type and Day of the Week:*** A color-coded bar chart compares total ride duration across each weekday, segmented by member type. The visualization clearly shows that while annual members dominate weekday usage, casual riders take longer rides on weekends.



- ***Number of Rides by Day and Member Type:*** This color-coded bar chart compares the total number of rides taken by annual and casual users across weekdays and weekends. While annual members consistently take more rides during the week, casual ridership increases significantly over the weekend. Combined with the previous visualization, this reinforces the idea that casual members make fewer rides but with longer durations—highlighting recreational usage.



- ***Rides and Durations by Hour of Day:*** This visualization shows both the number of rides and total ride duration by hour of day for each user group. Casual riders show increased activity during midday and afternoon hours and sustain longer ride durations compared to annual members. This further supports the insight that casual riders are using the service more for sightseeing or leisure rather than short, time-constrained trips.

Access the full Tableau dashboard [here] for the interactive map and full-sized visualizations.

6. Recommendations

1. ***Launch an Annual Weekend Membership***
   Given the increased activity of casual riders on weekends, Cyclistic should introduce a specialized "Annual Weekend Membership." This membership would cater to those who primarily ride during weekends, providing them with a cost-effective option that encourages conversion from casual riders to annual members. It could offer exclusive weekend perks such as discounts, free minutes, or extended ride durations.
2. ***Promote Annual Membership Benefits at High-Traffic, Casual-Dominant Stations***
   Stations with higher casual rider traffic should be prioritized for targeted promotional efforts. Strategically placed ads, both physical (e.g., posters or digital kiosks) and digital (e.g., in-app pop-ups), can help raise awareness of the advantages of an annual membership. This promotion should highlight the benefits, including savings on frequent rides and unlimited access during peak hours, directly at the stations where casual riders are more likely to see them.
3. ***Revise Pricing Structure for Midday and Afternoon Rides***
   To further attract casual riders and encourage more frequent usage, Cyclistic should consider adjusting its pricing for midday and afternoon rides. Offering discounts or lower rates during off-peak hours would entice casual riders to utilize the service more frequently, potentially increasing the likelihood of them upgrading to an annual membership for better value and convenience.
4. ***Offer Ride Duration-Based Incentives***
   Since casual riders often take longer trips, implementing loyalty rewards into the Annual program (e.g., points per minute ridden or discounted rates after a certain ride duration) could incentivize them to consider membership for better cost efficiency.

These recommendations align with the data-driven insights from the analysis and could help Cyclistic increase its annual membership conversion rates.

Appendix

SQL Queries

1.

```
SELECT
 TIMESTAMP_DIFF(ended_at, started_at, MINUTE) AS ride_duration_min
FROM `cyclistic-457220.cyclistic_data.cyclistic_q1_2025`
WHERE ended_at > started_at;
```

2.

```
WITH rides AS (
 SELECT
   member_casual,
   TIMESTAMP_DIFF(ended_at, started_at, MINUTE) AS ride_duration_mins
 FROM `cyclistic-457220.cyclistic_data.filtered_by_percentile`
 WHERE ended_at > started_at
)
SELECT DISTINCT
 member_casual,
 PERCENTILE_CONT(ride_duration_mins, 0.01) OVER (PARTITION BY member_casual) AS
p01,
 PERCENTILE_CONT(ride_duration_mins, 0.05) OVER (PARTITION BY member_casual) AS
p05,
 PERCENTILE_CONT(ride_duration_mins, 0.10) OVER (PARTITION BY member_casual) AS
p10,
```

```
 PERCENTILE_CONT(ride_duration_mins, 0.25) OVER (PARTITION BY member_casual) AS
p25,

 PERCENTILE_CONT(ride_duration_mins, 0.50) OVER (PARTITION BY member_casual) AS
p50,

 PERCENTILE_CONT(ride_duration_mins, 0.75) OVER (PARTITION BY member_casual) AS
p75,

 PERCENTILE_CONT(ride_duration_mins, 0.80) OVER (PARTITION BY member_casual) AS
p80,

 PERCENTILE_CONT(ride_duration_mins, 0.90) OVER (PARTITION BY member_casual) AS
p90,

 PERCENTILE_CONT(ride_duration_mins, 0.95) OVER (PARTITION BY member_casual) AS
p95,

 PERCENTILE_CONT(ride_duration_mins, 0.99) OVER (PARTITION BY member_casual) AS
p99

FROM rides
```

3.

```
CREATE TABLE `cyclistic-457220.cyclistic_data.filtered_by_percentile` AS

WITH ride_duration_data AS (

 SELECT *,

   TIMESTAMP_DIFF(ended_at, started_at, MINUTE) AS ride_duration_min

 FROM `cyclistic-457220.cyclistic_data.cyclistic_q1_2025`

),

percentiles AS (

 SELECT

   member_casual,
```

```
    APPROX_QUANTILES(ride_duration_min, 100)[OFFSET(5)] AS p05,
    APPROX_QUANTILES(ride_duration_min, 100)[OFFSET(10)] AS p10,
    APPROX_QUANTILES(ride_duration_min, 100)[OFFSET(80)] AS p80,
    APPROX_QUANTILES(ride_duration_min, 100)[OFFSET(95)] AS p95
  FROM ride_duration_data
  GROUP BY member_casual
)
SELECT r.*
FROM ride_duration_data r
JOIN percentiles p
  ON r.member_casual = p.member_casual
WHERE (
  (r.member_casual = 'member' AND r.ride_duration_min > p.p10 AND
r.ride_duration_min <= p.p80)
  OR
  (r.member_casual = 'casual' AND r.ride_duration_min > p.p05 AND
r.ride_duration_min <= p.p95)
);
```

4.

```
SELECT *
FROM `cyclistic-457220.cyclistic_data.filtered_by_percentile`
ORDER BY rand() limit 4000
```