CYCLISTIC BIKE-SHARE CASE STUDY: UNDERSTANDING USER BEHAVIOR

As part of my Google Data Analytics Certification, I conducted an in-depth analysis of Cyclistic's bike-share data to understand how casual riders and annual members use bikes differently.

Using BigQuery for SQL-based analysis, I explored ride duration trends, peak usage times, and station popularity.

My findings showed that casual riders take longer rides, prefer weekends, and use bikes near tourist areas, while members ride frequently for shorter durations, likely for commuting.

Based on these insights, I developed three key business recommendations: launching a Weekend Warrior membership plan, targeted promotions at high-traffic stations, and incentives for long rides.

This case study demonstrates my ability to analyze large datasets, extract meaningful insights, and translate findings into business strategies.

1. ASK - A Clear Statement Of The Business Task

What is the problem i am trying to solve?

The problem is understanding how **annual members** and **casual riders** use Cyclistic bikes differently. This insight will help Cyclistic design targeted marketing strategies to convert casual riders into annual members, thereby increasing the profitability and long-term growth of the company.

How can your insights drive business decisions?

By analyzing the historical bike usage data, we can uncover patterns and behaviors specific to annual members and casual riders, such as:

- **Ride duration**: Do casual riders take longer or shorter rides compared to annual members?
- **Time of usage**: Are casual riders more active on weekends, while members ride during weekdays?
- **Popular locations**: What docking stations or routes are preferred by casual riders versus annual members?

These insights can inform Cyclistic's marketing team on:

1. **How to position annual memberships** as a more appealing option to casual riders.
2. **What types of promotions or campaigns** would resonate most effectively with casual riders.
3. **How to leverage digital media platforms** to maximize conversions from casual riders to members.

*--- The primary goal is to analyze how **annual members** and **casual riders** use Cyclistic bikes differently to support the company's objective of converting casual riders into annual members. By examining historical trip data, we aim to identify key patterns in ride duration, time of usage, and popular locations for each group. These insights will drive data-informed marketing strategies, helping Cyclistic design targeted campaigns to increase annual memberships, which are more profitable and critical for the company's long-term*

growth. This analysis will also guide the use of digital media to effectively influence casual riders to become loyal members. ---

2. PREPARE - A Description Of All Data Sources Used

Where is your data located?

- The data is stored in Google Cloud Storage as 12 CSV files, each representing monthly trip data for Cyclistic in 2024.
- The files have been imported into Google BigQuery for querying, analysis, and processing.

How is the data organized?

Table Structure: Each file (or its data) is organized in tabular format with the following columns:

- *ride_id*: Unique identifier for each ride.
- *rideable_type*: Type of bike used (e.g., docked bike, electric bike).
- *started_at* & *ended_at*: Timestamp of the trip start and end times.
- *start_station_name* & *end_station_name*: Names of the stations where the trip began and ended.
- *start_lat*, *start_lng*, *end_lat*, *end_lng*: Geographic coordinates of the stations.
- *member_casual*: User type (either "member" or "casual").

Integration: The data from all 12 files was combined into a single table in BigQuery for ease of analysis.

Are there issues with bias or credibility in this data? Does your data ROCCC?

ROCCC Analysis:

- **Reliable**: The data has been made available by Motivate International Inc., making it a credible source.
- **Original**: The data is first-party, collected directly from Cyclistic's bike-sharing platform.
- **Comprehensive**: Covers an entire year (2024), ensuring robust temporal coverage for analyzing trends.
- **Current**: Data from 2024 is recent and relevant for strategic planning.
- **Cited**: As Cyclistic's data, it is directly linked to the company's operations.

Potential Bias:

Geographic Bias:

- The data may overrepresent specific areas or stations where Cyclistic operates, potentially skewing insights for expansion opportunities.

Behavioral Bias:

- The dataset includes only Cyclistic's existing riders, which may not generalize to potential new customers.

Missing Demographics:

- The data lacks user-specific demographics (e.g., age, income), which could provide deeper insights into rider behavior.

How are you addressing licensing, privacy, security, and accessibility?

Licensing:

- Ensure the data usage adheres to Cyclistic's policies and is limited to internal analysis.

Privacy:

- The dataset does not contain **personally identifiable information (PII)**, making privacy concerns minimal. However, geographic patterns (e.g., popular routes) should not be exposed without aggregation.

Security:

- The data is stored in Google BigQuery, which uses **Google Cloud's robust encryption and access control mechanisms**.
- Access to the dataset is limited to authorized users via **IAM roles** (e.g., BigQuery Viewer, Editor).

Accessibility:

- BigQuery ensures that data is easily accessible to stakeholders for querying and analysis.
- Analysis results can be exported to Google Sheets for easier sharing.

How did you verify the data's integrity?

Steps Taken:

Schema Validation:

- Ensured consistent column names, formats, and data types across all 12 files (e.g., TIMESTAMP for started_at and ended_at).

Missing Data Check: Used SQL queries in BigQuery to identify missing or blank fields:

```
SELECT COUNT(*) AS missing_count, column_name
FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`
WHERE column_name IS NULL
GROUP BY column_name;
```

Duplicate Removal: Verified and removed duplicate records using:

```
SELECT DISTINCT *
FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`;
```

Data Quality Validation:

- Verified that trip durations (*ended_at - started_at*) were positive and within reasonable limits.
- Ensured station names and IDs matched known Cyclistic stations.

How does it help you answer your question?

The data helps answer the question of **how to convert casual riders into members** by providing:

1. **Behavioral Insights**:

- Identifying trip patterns (e.g., duration, start and end locations) for casual riders compared to members.

2. **Temporal Trends**:

- Examining seasonal and daily usage patterns to identify peak times for casual riders.

3. **Operational Insights**:

- Pinpointing popular stations and routes used by casual riders.

4. **Engagement Metrics**:

- Highlighting differences in trip durations and frequency between members and casual users.

These insights can inform marketing strategies (e.g., targeting specific locations or times) and operational improvements (e.g., better bike availability).

Are there any problems with the data?

Potential Problems:

1. **Missing Data:**
   - Some rows may lack station names, IDs, or geographic coordinates.
   - Missing values can limit the granularity of the analysis.
2. **Outliers:**

- Unusual trip durations (e.g., rides lasting a few seconds or several days) may distort metrics.

3. Station Information:

- Some station names or IDs may be inconsistent (e.g., typos, duplicates), requiring cleaning.

4. Limited Demographics:

- The data does not include user demographics, limiting insights into customer profiles.

Description of All Data Sources Used

Data Source: Cyclistic Trip Data (2024)

Description:

- 12 CSV files, each representing monthly trip data for Cyclistic's bike-sharing service in 2024.
- Includes details about trip duration, bike type, start and end stations, geographic coordinates, and user type

Location:

- Initially stored in Google Cloud Storage.
- Imported into BigQuery for analysis.

Fields:

- *ride_id*: Unique identifier for each trip.
- rideable*_type*: Type of bike used (e.g., docked, electric).
- *started_at* and *ended_at*: Start and end timestamps of trips.
- *start_station_name*, *start_station_id*, *end_station_name, end_station_id*: Station names and is's where trips began and ended.
- *start_lat, start_lng, end_lat, end_lng:* Geographic coordinates of the stations.
- *member_casual*: User type (member or casual).

---

3. PROCESS - Documentation Of Any Cleaning Or Manipulation Of Data

1. What tools are you choosing and why?
Google BigQuery
Why BigQuery?

- Handles Large Data: BigQuery is designed for processing and analyzing large datasets like my 12 Cyclistic CSV files.
- SQL Queries: Enables powerful data manipulation and analysis using SQL.
- Scalability: Easily scales as the dataset grows.
- Accessibility: Data is cloud-based and accessible from anywhere.
- Integration: Integrates seamlessly with tools like Google Sheets, Google Data Studio, and other visualization platforms for downstream analysis.

2. Have you ensured your data's integrity?
Steps to Ensure Data Integrity:

1. **Schema Validation**:
   - Confirmed column headers and data types are consistent across the 12 CSV files during upload to BigQuery.
   - Ensured critical columns (*ride_id*, *started_at*, *ended_at*, *member_casual*) are properly typed.
2. **Data Type Checks:**
   - Verified started_at and ended_at columns are stored as TIMESTAMP.
   - Confirmed numeric columns like ride_length_seconds (before deletion) and station IDs are properly typed.
3. **Handling Missing Data:**
   - Identified missing or null values using SQL queries like:

```
SELECT COUNT(*) AS missing_count, 'start_station_name' AS column_name
FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`;
WHERE start_station_name IS NULL
UNION ALL
SELECT COUNT(*) AS missing_count, 'end_station_name' AS column_name
FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`;
WHERE end_station_name IS NULL;
```

   - Replaced missing station names with "Unknown."
4. **Duplicate Removal:**
   - Checked and removed duplicate rows using:

```
SELECT DISTINCT *
FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`;
```

5. **Consistency Across Files:**
   - Unified column headers and data formats across all files before combining them in BigQuery.

3. What steps have you taken to ensure that your data is clean?
Combining Files:

- Loaded the 12 monthly CSV files into BigQuery as a single table for consistent analysis.

Column Renaming:
- Standardized column names (e.g., *member_casual* → *user_type*).

Created New Columns:
- Added *ride_length* (HH:MM:SS format) for better readability.
- Added *day_of_week* to facilitate temporal analysis.

Outlier Detection:
- Filtered out invalid rides (e.g., negative or zero-duration trips):

```
DELETE FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`;
WHERE TIMESTAMP_DIFF(ended_at, started_at, SECOND) <= 0;
```

Duplicate Removal:
- Removed duplicate rows using DISTINCT.

Missing Data:
- Replaced missing station names with "Unknown."

4. How can you verify that your data is clean and ready for analysis?

Column Consistency:
- Ensured all columns are correctly named, formatted, and typed (TIMESTAMP, STRING, INTEGER).

Value Checks:
- Verified ride_length values align with started_at and ended_at timestamps:

```
SELECT COUNT(*)
FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`;
WHERE ride_length <> FORMAT_TIMESTAMP('%H:%M:%S',
TIMESTAMP_SECONDS(TIMESTAMP_DIFF(ended_at, started_at, SECOND)));
```

- Confirmed no invalid or outlier rows remain.

Data Completeness:
- Checked for null or missing values using COUNT(*) queries.

Sampling:
- Manually inspected a random sample of rows to confirm accuracy:

```
SELECT *
FROM `cyclistic-case-study-450611.cyclistic_data_2024.table`;
ORDER BY RAND()
LIMIT 10;
```

5. Have you documented your cleaning process so you can review and share those results?

Below is a template for documenting my cleaning and manipulation process.

Documentation of Data Cleaning and Manipulation

Source Data

- **Data Location**: Google BigQuery.
- **Source Files**: 12 monthly CSV files for Cyclistic trip data (2024).

Cleaning and Manipulation Steps

1. **File Combination**:
   - All 12 CSV files were uploaded to Google Cloud Storage and imported into BigQuery as a single table.
2. **Column Standardization**:
   - Unified column names across all files.
   - Added two new columns:
     - *ride_length* (formatted as HH:MM:SS).
     - *day_of_week* (numeric, 1 = Sunday, 7 = Saturday).
3. **Duplicate Removal**:
   - Removed duplicate rows using the DISTINCT keyword.
4. **Missing Data**:
   - Replaced missing station names (*start_station_name*, *end_station_name*) with "*Unknown*"
   - Checked for null values in other critical fields (e.g., *started_at*, *ended_at*).
5. **Outlier Removal**:
   - Filtered out rides with:
     - Negative or zero durations.
     - Unrealistic durations exceeding 24 hours.
6. **Validation**:
   - Verified timestamps and trip durations align.
   - Randomly sampled rows for manual inspection.

Resulting Data

Table Name: `*cyclistic-case-study-450611.cyclistic_data_2024.enhanced_table*`

Columns:

- Key Columns: *ride_id*, *started_at*, *ended_at*, *user_type*.
- Derived Columns: *ride_length* (HH:MM:SS), *day_of_week*.

4. ANALYZE

1. How should you organize your data to perform analysis on it?

- combined 12 months of data into a single table in BigQuery, making analysis easier;
- included key fields (ride_id, timestamps, station names, user type);
- added useful new columns like *ride_length* and *day_of_week* for deeper insights;
- group data by time intervals: summarized data by month to identify seasonal trends in bike usage;
- segment data by user type (casual vs. member).

SQL query:

```
SELECT
  EXTRACT(MONTH FROM started_at) AS month,
  member_casual,
  COUNT(ride_id) AS ride_count,
  AVG(TIMESTAMP_DIFF(ended_at, started_at, MINUTE)) AS avg_ride_duration
FROM `cyclistic-case-study-450611.cyclistic_data_2024.enhanced_table`
GROUP BY month, member_casual
ORDER BY month, member_casual;
```

2. Has your data been properly formatted?

- checked for null values and replaced missing station names with "*Unknown*";
- removed duplicates and eliminated invalid rides (negative or unrealistic ride durations);
- verified that timestamps were properly formatted;
- investigate station name inconsistencies;
- check for outlier ride durations.

SQL query:

```
SELECT ride_id, member_casual, TIMESTAMP_DIFF(ended_at, started_at, MINUTE) AS ride_duration
FROM  `cyclistic-case-study-450611.cyclistic_data_2024.enhanced_table`
WHERE TIMESTAMP_DIFF(ended_at, started_at, MINUTE) > 180  -- More than 3 hours
ORDER BY ride_duration DESC;
```

3. What trends or relationships did you find in the data?

- outlined key behavioral insights to extract (ride duration, usage patterns, station popularity);
- checked for weekly patterns by adding *day_of_week*;
- analyze ride duration trends:
    - Do casual riders take longer rides compared to members?
    - How do ride durations change by time of day (morning vs. evening)?

SQL query:

```
SELECT
  EXTRACT(HOUR FROM started_at) AS hour_of_day,
  member_casual,
  AVG(TIMESTAMP_DIFF(ended_at, started_at, MINUTE)) AS avg_ride_duration
FROM `cyclistic-case-study-450611.cyclistic_data_2024.enhanced_table`
GROUP BY hour_of_day, member_casual
ORDER BY hour_of_day, member_casual;
```

- top stations for casual vs. member riders:
    - What are the most used start and end stations for each rider type?

SQL query:

```sql
SELECT
  start_station_name,
  member_casual,
  COUNT(ride_id) AS ride_count
FROM `cyclistic-case-study-450611.cyclistic_data_2024.enhanced_table`
GROUP BY start_station_name, member_casual
ORDER BY ride_count DESC
LIMIT 10;
```

4. How will these insights help answer your business questions?

1. Monthly Ride Trends

- Casual riders take significantly longer trips than members.
- Ride volume is higher for members across all months.
- January: Members: 120,413 rides, Casual: 24,460 rides.
- March: Members: 176,001 rides, Casual: 47,163 rides.
- Casual riders' average trip duration is around 20-25 minutes, while members' average trip duration is 12-13 minutes.

Insights:

- Casual riders may use bikes more for recreational purposes, while members use them for commuting.
- Marketing should focus on converting casual users into members by promoting commuter-friendly benefits like reduced pricing for frequent rides.

2. Ride Duration Distribution

- The dataset contains 21,324 ride records with their respective durations.
- Ride durations range widely, with casual riders showing longer durations.

Insights:

- Identifying the peak ride duration ranges can help in designing targeted pricing models.
- If many casual rides are longer than 30 minutes, promoting monthly passes may be an effective conversion strategy.

3. Ride Usage by Hour of the Day

- Casual riders take longer trips during all hours of the day.
- Members' ride duration stays consistent throughout the day (~12 minutes), whereas casual rides increase in length late at night (25-27 minutes at 1-2 AM).

Insights:

- Members likely use bikes for regular commuting (work/school), leading to shorter, structured trips.

- Casual riders may use bikes more for leisurely rides, tourism, or late-night activities.

4. Most Popular Start Stations

Top stations for casual riders:

- Streeter Dr & Grand Ave (51,050 rides)
- DuSable Lake Shore Dr & Monroe St (34,107 rides)

Top stations for members:

- Kingsbury St & Kinzie St (29,522 rides)

Insights:

- Casual riders' top stations are likely near tourist attractions or recreational areas.
- Members' top stations may be near business districts or residential areas.

Marketing Action:

- Place membership promotions at stations heavily used by casual riders.
- Target commuters near member-frequented stations with ads on quicker, cheaper, and more convenient trips.

Cyclistic Bike-Share Case Study - Data Analysis Summary

Objective

The goal of this analysis is to identify patterns in Cyclistic's bike usage for casual riders vs. annual members. These insights will help convert casual riders into annual members through data-driven marketing strategies.

Key Insights from Data Analysis

1.Monthly Ride Trends

Findings:

- Members consistently have more rides per month than casual users.
- Casual riders' trips are significantly longer (20-25 minutes) compared to members' average trip duration (12-13 minutes).
- Casual ridership increases in warmer months (potentially due to tourism and leisure use).

Insights:

- Encourage casual riders to buy annual memberships by promoting discounts on longer rides.
- Seasonal promotions (e.g., summer discounts) may help convert high-usage casual riders.

2. Ride Duration Distribution

Findings:

- Casual riders have a wider range of ride durations, with many trips exceeding 30 minutes.
- Members tend to have shorter and more consistent ride durations.

Actionable Insights:

- A time-based pricing strategy (e.g., free first 30 minutes for members) could encourage casual users to join.
- Offer membership perks like unlimited rides for trips under 30 minutes.

3. Ride Usage by Hour of the Day

Findings:

- Members' ride durations remain steady throughout the day (~12 minutes).
- Casual riders take longer trips late at night (averaging 25-27 minutes between 1-2 AM).

Actionable Insights:

- Target casual riders with late-night ride passes or promotions for evening memberships.
- Highlight the convenience of shorter trips for commuters to attract casual users.

4. Most Popular Start Stations

Findings:

- Casual riders favor stations near tourist hotspots (e.g., *Streeter Dr & Grand Ave*).
- Members prefer stations closer to business or residential areas.

Actionable Insights:

- Advertise memberships at casual hotspots (e.g., station-based promotions).
- Focus on commuter benefits at member-frequented stations.

Business Recommendations

- Offer "Weekend Warrior" memberships for casual users who ride mostly on weekends.
- Provide discounts for long trips (e.g., free extra minutes for members).
- Place digital ads at top casual rider stations to increase membership sign-ups.
- Emphasize convenience for commuters (shorter, cheaper, faster rides).

---

5. SHARE

Cyclistic_Case_Study_Presentation.ppt
Monthly_Ride_Trends.png
Ride_Duration_Comparisation.png
Ride_Duration_By_Hour.png
Top_5_Stations.png

6. ACT

1. Introduce a "Weekend Warrior" Membership Plan

Insight:
-Casual riders take longer rides and primarily use bikes on weekends.
-Members take shorter, more frequent trips, indicating a commuter-based pattern.

Recommendation:
- Launch a Weekend Warrior membership plan tailored for high-usage casual riders who ride mainly on weekends
- Offer discounted weekend passes to encourage sign-ups.
- Provide incentives such as unlimited 45-minute weekend rides for a fixed price.

Expected Impact:
- Converts high-usage casual riders into members.
- Creates an entry-level membership tier that casual riders find attractive.

2. Target Casual Riders with Promotions at Popular Stations

Insight:
-Casual riders frequently start trips from tourist-heavy stations (e.g., *Streeter Dr & Grand Ave*).
-Members use stations closer to business and residential areas, showing commuter behavior.

Recommendation:
- Place digital ads and station-based promotions at popular casual rider stations.
- Use in-app pop-ups and QR codes at bike stations to advertise membership benefits.
- Highlight cost savings (e.g., "Save 30% per ride with a membership").

Expected Impact:
- Increased membership sign-ups from frequent tourist and recreational users.
- Helps casual riders see the financial benefits of membership vs. per-ride payments.

3. Offer Incentives for Longer Rides & Commuters

Insight:

-Casual riders take longer rides (20-25 min) vs. members (12-13 min).

-Members have consistent usage throughout the day, likely for commuting.

Recommendation:

- Offer "First 30 Minutes Free" for New Members to encourage sign-ups.
- Introduce loyalty-based discounts for casual riders who frequently take long trips.
- Promote monthly commuter plans with lower rates for weekday riders.

Expected Impact:

- Encourages longer trip casual riders to subscribe.
- Captures daily commuters looking for affordable alternatives.

Next Steps for Implementation

-Update Cyclistic's marketing strategy to focus on high-usage casual riders.
-Launch a trial campaign for the Weekend Warrior Plan and measure conversions.
-Monitor engagement at casual-dominated stations and adjust promotional efforts accordingly.
-Expand analysis with additional data, such as user demographics or surveys, to refine strategies further.