# **Google Case Study :**

# **How Can a Bike-Sharing Company Achieve Rapid Success ?**

**EXECUTIVE SUMMARY**

This analysis explores data collected by Divvy, a **bike-sharing company** in Chicago, focusing on user behavior and ride characteristics to identify trends that could **enhance membership conversion rates**. By examining the differences between casual riders and members, this study aims to provide actionable insights for increasing subscriptions.

The analysis follows a structured framework consisting of six key steps :

1. [**Ask :** Define the Questions](#ASK)
2. [**Prepare :** Gather and Structure Data](#PREPARE)
3. [**Process :** Clean and Prepare Data](#PROCESS)
4. [**Analyze :** Data Analysis](#ANALYZE)
5. [**Share :** Visualizations and Results](#SHARE)
6. [**Act : Strategic Recommendations**](#ACT)

**SCENARIO**

In the context of Cyclistic's marketing team, the goal was to understand how casual cyclists and annual members utilize the bike-sharing service differently. Identifying these usage patterns is essential for designing effective marketing strategies to **convert casual users into annual subscribers.**

1. **ASK : Define the Questions**

In this phase, we establish key questions to guide our analysis, ensuring we address critical issues that inform Cyclistic's marketing strategy.

**GUIDING QUESTIONS :**

* What are the key differences in ride behavior between casual users and members?
* How do ride durations vary across different days of the week?
* What strategies can be implemented to convert casual users into members?

1. **PREPARE : Gather and Structure Data**

**DATA SOURCES :**

The historical ride data was sourced from Divvy [here](https://divvy-tripdata.s3.amazonaws.com/index.html). The dataset includes 12 months of data, spanning from March 2023 to February 2024, with files named in the format YYYYMM-divvy-tripdata.csv.

**DATA STRUCTURE :**

Each monthly file contains 13 columns with detailed information about each ride, including:

* **ride\_id:** Unique identifier for each ride.
* **rideable\_type:** Type of bike utilized.
* **started\_at:** Timestamp for ride commencement.
* **ended\_at:** Timestamp for ride conclusion.
* **start\_station\_name:** Departure station name.
* **start\_station\_id:** Unique identifier for the departure station.
* **end\_station\_name:** Arrival station name.
* **end\_station\_id:** Unique identifier for the arrival station.
* **start\_lat:** Latitude of the departure station.
* **start\_lng:** Longitude of the departure station.
* **end\_lat:** Latitude of the arrival station.
* **end\_lng:** Longitude of the arrival station.
* **member\_casual:** Classification of user type, either member or casual.

**DATA INTEGRITY CHECKS :**

 To verify data integrity, each monthly file was examined using Excel, focusing on the following aspects:

* **Ride ID:**Each ride has a unique identifier.
* **Rideable Type:**Three types of bikes are present (electric\_bike, classic, docked). The term "docked" was previously referred to as "classic" in archived data.
* **Timestamps:**The started\_at and ended\_at columns are formatted as dd/mm/yyyy hh:mm.
* **Station Names and IDs:**Missing values were noted for start\_station\_name and start\_station\_id, particularly for electric bikes that do not require docking. Similar patterns of missing data were observed for end\_station\_name and end\_station\_id.
* **Latitude and Longitude:**Values for start\_lat and start\_lng are consistent with the expected ranges for Chicago. However, some end\_lat and end\_lng values were recorded as 0.0, indicating potential data entry errors or the use of test stations.
* **User Type:**The member\_casual column correctly categorizes users as either members or casual riders.

While some missing or anomalous data points were identified, the overall integrity of the dataset is satisfactory. Any issues will be addressed during the data cleaning phase.

**DATA CREDIBILITY AND BIAS CONSIDERATIONS :**

The dataset is evaluated against the ROCCC framework:

* **R: Reliable** - free from bias, comprehensive, and accurate.
* **O: Original** - directly sourced from the company.
* **C: Cited** - data obtained from Divvy, provided by Motivate International Inc.
* **C: Comprehensive** - detailed ride information segmented by user type.
* **C: Current** - includes data from the last 12 months.

**COMPLIANCE WITH PRIVACY, SECURITY, AND ACCESSIBILITY STANDARDS :**

The dataset is anonymized to ensure user privacy and complies with personal data protection regulations. Licensing agreements facilitate legal access to the data, while ownership remains with the city of Chicago. The data is regularly updated on a monthly basis and is made publicly available in CSV format through official channels. Additionally, the dataset is filtered to exclude irrelevant rides, such as those conducted by maintenance staff or rides lasting less than 60 seconds, further mitigating potential bias and ensuring the integrity of the analysis.

**3. PROCESS : Clean and Prepare Data**

SQL is employed for data processing, which involve identifying and addressing missing or erroneous values, as well as ensuring consistency across different variables. We import the 12 CSV files into the database, resulting in individual tables for each month.

For detailed methodologies on data cleaning and preparation, refer to the following code repository :

[Data Cleaning and Preparation Code.](https://github.com/ambrb/Divvy-Bike-Analysis/blob/main/Cleaning_Divvy_Analysis.sql)

**4. ANALYZE : Data Analysis**

In this phase, we use **R** and we applied the following key techniques to analyze user behavior and generate insights :

**Descriptive and Comparative Analysis:**

* Calculated **summary metrics** (mean, median, quantiles) to understand ride length distributions.
* **Mann-Whitney U, Wilcoxon**, and**Kruskal-Wallis tests**to identify significant differences in ride lengths between members and casual users, especially across weekdays and weekends.

**Categorical and Correlation Analysis :**

* **Chi-square Tests**explored associations between user type (member vs. casual) and bike type (classic vs. electric), as well as bike choice across weekdays vs. weekends among casual users.
* **Correlation matrix** to identify key factors affecting user type.
* **Spearman Correlation**showed minimal influence of time of day on ride length.

**Regression Models:**

* **Linear Regression**examined the relationship between ride length and factors like bike type and ride time among casual users.
* **Logistic Regression**modeled the likelihood of different ride patterns between members and casual users, including tendencies for longer, weekend rides.

**Clustering :**

* **K-means Clustering** segmented users into behavioral groups based on ride length, bike type, and day of use.

**Visualization and Spatial Analysis:**

* **Density Plots** and **heatmaps** illustrated ride length distributions and peak usage by user type.
* **Spatial heatmaps**to highlight high-usage areas for different user types across the city.

To explore the analysis conducted, please visit :

[Data Analysis R Code](https://github.com/ambrb/Divvy-Bike-Analysis/blob/main/Divvy_bike_analysis.Rmd)

**5. SHARE : Visualizations and Results**

**6. ACT : Strategic Recommendations**

These 2 sections present the analysis results, the visualizations designed to engage stakeholders and the strategic recommendations. To view the document containing the insights and results, please see :

[Tableau Vizualisation](https://public.tableau.com/app/profile/ambre.b.gu./viz/DIVVYDATAANALYSIS/DIVVYBIKEUSERS)