Cyclistic Bike-share Analysis

Google Data Analytics Capstone Project

Abstract

The project was carried out as part of the Google Data Analytics Professional Certificate program and was built upon the Divvy-tripdata dataset. The project aims to **maximize the number of annual memberships by converting casual customers (single-ride and day passes) to members (annual membership)** by using PostgreSQL, Microsoft Power BI for Data Wrangling and Exploratory Data Analysis to understand how casual riders and annual members use bikes differently. The analysis was conducted across multiple dimensions: descriptive analysis, time series analysis, and spatial analysis. The results show that casual riders prefer electric bikes, rent mostly in the afternoon and on weekends, and use bikes for longer durations primarily for leisure. In contrast, annual members show no significant preference between electric and classic bikes, rent both in the morning and afternoon on workdays, and use bikes for shorter durations primarily for commuting to work.

1 Introduction

The Cyclistic Bike-Share Analysis case study is a part of the Google Data Analytics Professional Certificate on Coursera [1]. This case study simulates the real-world tasks of a junior data analyst working at Cyclistic, a bike-share company in Chicago. The primary aim of this project is to understand how annual members and casual riders use Cyclistic bikes differently, and to design a marketing strategy that converts casual riders into annual members.

The main goal of this project is to provide data-driven insights to support Cyclistic's marketing strategy aimed at converting casual riders to annual members. The analysis will help to:

- Identify usage patterns between annual members and casual riders.
- Understand what might encourage casual riders to buy annual memberships.
- Explore how marketing strategies can help convert casual riders into annual members.

1.1 Workflow

The workflow is shown in Fig.1.

[1] https://www.coursera.org/professional-certificates/google-data-analytics

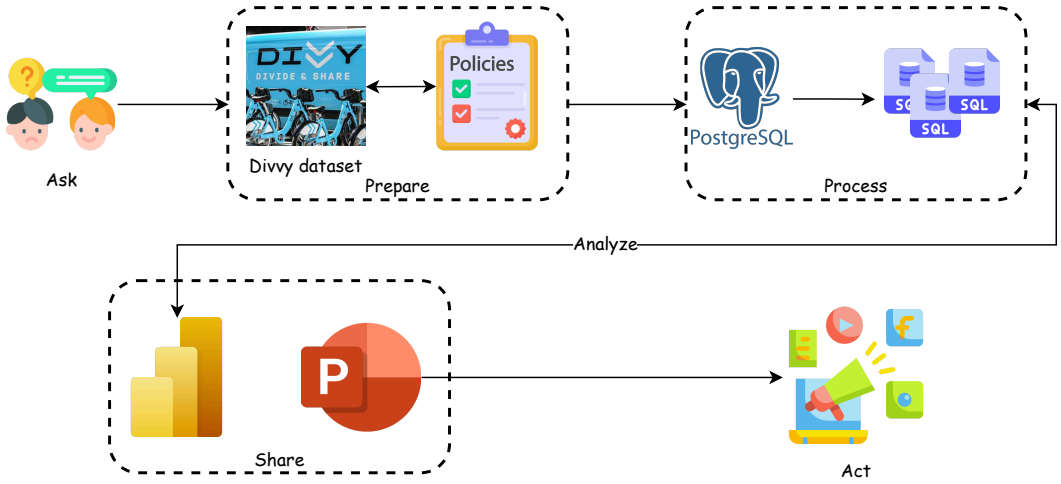

Figure 1: The workflow diagram for analyzing Cyclistic bike usage differences between annual members and casual riders.

Ask In this initial phase, we define the key business question: *How do annual members and casual riders use Cyclistic bikes differently?* This stage involves identifying the business task and considering key stakeholders.

Prepare To answer the business question, we use Cyclistic's historical trip data. This involves gathering, organizing, and understanding the data. Key tasks include:

- Downloading and storing the data appropriately.
- Sorting and filtering the data.
- Assessing the credibility of the data.

Process In the processing phase, we clean and prepare the data for analysis. This involves:

- Checking the data for errors.
- Choosing appropriate tools (in this case, PostgreSQL).
- Transforming the data for effective analysis.
- Documenting the cleaning process.

Analyze This phase involves aggregating, organizing, and analyzing the data to uncover trends and relationships. Key tasks include:

- Conducting descriptive analysis.
- Performing time series analysis to understand trends over time.

- Applying spatial analysis to explore geographic patterns.

Note: If you only want to see the results of our analysis and are not interested in the process or other details, please refer to section 4.

Share After analyzing the data, we create visualizations and share our findings. This involves:

- Creating effective data visualizations.
- Presenting the findings in a clear and accessible manner.
- Ensuring the presentation is tailored to the audience (Cyclistic executive team).

Act In the final step, we interpret the findings and make actionable recommendations based on the analysis. This helps guide the business decisions for Cyclistic's marketing strategy.

1.2 Tools Used

PostgreSQL For this case study, PostgreSQL [2] was used to perform the data extraction, cleaning, and analysis. SQL's powerful querying capabilities allowed for effective manipulation and analysis of large datasets, enabling the discovery of insights and trends critical to answering the business question.

Microsoft Power BI Microsoft Power BI [3] was utilized to create interactive and comprehensive data visualizations. Its robust features allowed for the effective presentation of findings, making the data insights accessible and understandable for stakeholders.

Microsoft PowerPoint Microsoft PowerPoint [4] was employed to compile and present the results of the analysis. Through clear and professional presentations, the key findings and recommendations were communicated effectively to the Cyclistic executive team.

2 Dataset

The Divvy dataset [5] is a comprehensive collection of data pertaining to Chicago's bike-sharing system, Divvy. This dataset includes detailed information on bike trips, such as trip duration, start and end times, start and end stations, and the bike IDs. Additionally, it provides data on user demographics, including user types (member or casual), and occasionally age and gender. The dataset is valuable for analyzing urban mobility patterns, understanding bike usage trends, and improving the efficiency and planning of bike-sharing systems. Researchers, data analysts, and

[2] https://www.postgresql.org/
[3] https://www.microsoft.com/en-us/power-platform/products/power-bi
[4] https://www.microsoft.com/en-us/microsoft-365/powerpoint
[5] https://divvybikes.com/system-data

city planners can leverage this dataset to gain insights into sustainable transportation and urban infrastructure development.

Table 1 provides a Codebook detailing the dataset, which includes 13 variables and 5,667,717 observations. This Codebook serves as a comprehensive reference guide for the data's variables, their definitions, and associated values. Utilizing the Codebook enhances data transparency, clarity, and understanding. This is useful for the Data Wrangling and Exploratory Data Analysis stages.

Based on the Codebook, it is evident that some columns should be adjusted to their appropriate data types. For example, `"rideable_type"` could be transformed into a categorical variable, and `"start_at"` could be converted to a date-time format. Assigning the correct data types can help reduce memory usage and simplify the data wrangling and analysis stages. Additionally, several columns contain multiple null values. We need to address these null values and determine the best approach for handling them. There are no duplicate values in the dataset.

Table 1: Divvy Dataset Codebook.

Variable	**Descriptions**	**Type**	**Missing(%)**
ride_id	the unique identifier for each trip	unique key	-
rideable_type	the type of rented vehicle, include electric_bike, classic_bike, docked_bike	categorical	-
started_at	the time when the rental period began	datetime	-
ended_at	the time when the rental period ended	datetime	-
started_station_name	the name of the station where the ride was initiated.	text	14.70
start_staion_id	the unique identifier for the station where the ride began	numeric	14.70
end_station_name	the name of the station where the bike was returned after the rental.	text	15.75
end_station_id	the unique identifier for the station where the bike was returned.	numeric	15.75
start_lat	the latitude coordinate of the starting point of the ride.	float	-
start_lng	the longitude coordinate of the starting point of the ride.	float	-
end_lat	the latitude coordinate of the ending point where the bike was returned.	float	0.10
end_lng	the longitude coordinate of the ending point where the bike was returned.	float	0.10
member_casual	the type of customer, include casual and member	categorical	-

3 Data Wrangling

The data wrangling process encompassed several critical steps to prepare the dataset for subsequent analysis. These steps are outlined as follows:

- Creating the Cyclistic bike table in PostgreSQL based on the Codebook schema.
- Cleaning data to ensure consistency and removing unnecessary columns.
- Creating time metrics for time series analytical capabilities.
- Saving the cleaned dataset.

Data cleaning involved several steps to maintain the dataset's relevance and consistency. Initially, unnecessary columns were removed to optimize memory usage. Specifically, the `"start_station_id"` and `"end_station_id"` columns were eliminated as they were not pertinent to the analysis. Additionally, maintenance stations were removed from the `"start_station_name"` and `"end_station_name"` columns. For classic bikes, trips with null values in both stations (begin, end) were excluded, as these bikes must be rented and returned at stations. In contrast, for electric bikes, which can be locked at any location, null values in the station columns were replaced with `'On bike lock'`.

The `"rideable_type"` column, which originally included `docked_bike`, `classic_bike`, and `electric_bike`, was updated to merge `docked_bike` with `classic_bike`, reflecting the outdated nature of the former term. Columns were then converted to their respective data types to ensure accurate handling and analysis. Specifically, the `"rideable_type"` and `"member_casual"` columns were converted to categorical data types, while the `"start_at"` and `"ended_at"` columns were changed to date-time data types. These conversions facilitated precise data manipulation and time-based calculations.

Subsequently, Time metrics were created to enhance the time series analytical capabilities of the dataset. A column named `ride_time_minutes` was introduced to capture the duration of each trip by subtracting the `"ended_at"` timestamp from the `"start_at"` timestamp. Additionally, to provide further insights, the hour, day of the week, and month were also extracted. These derived features are essential for generating various analytical charts, which will be discussed in section 4. Further refinement included dropping invalid trip durations. Trips with durations less than 0 minutes or exceeding 1440 minutes were filtered out to focus on trips within a one-day duration. Finally, the cleaned dataset was thoroughly reviewed for accuracy and completeness. It was saved in Parquet format to maintain the integrity of column data types and to facilitate efficient data handling for future analyses.

4 Exploratory Data Analysis

I show the main results of my analysis in Figures 2, 3, 4.

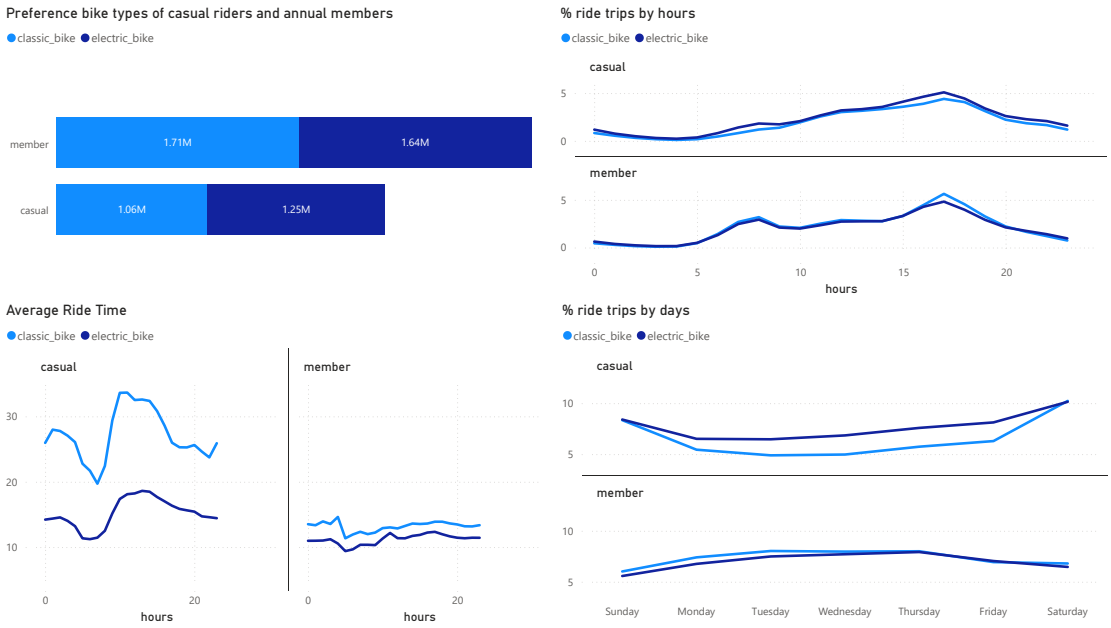

Figure 2: The figure show the preference bike types, average ride time and percentages of ride trips by hours/days for each casual riders and annual members.

4.1 Preference bike types analysis

As shown in Fig.2 and Tab.2, casual riders tend to prefer electric bikes over classic bikes (the margin is 8.16%). In contrast, there is no significant difference in bike preference among annual members (the margin is 2.18%). Overall, casual riders favor electric bikes more, while annual members slightly prefer classic bikes.

Table 2: Preference bike types of casual riders and annual members.

	Classic bikes	Electric bikes
Casual riders	45.92	**54.08**
Annual member	**51.09**	48.91

4.2 Time series analysis

Average ride time analysis The analysis of average ride times in Fig.2 reveals distinct behavioral patterns between casual riders and members. Casual riders tend to use bicycles for longer, leisurely trips, as evidenced by their extended ride durations throughout various times of the day. In contrast, members demonstrate shorter and more consistent ride times, indicating a pattern that aligns with regular commuting behavior. This suggests that members primarily use bicycles for routine transportation needs rather than leisure. Additionally, the data shows that electric bikes are

generally favored for quicker trips by both casual riders and members, underscoring their efficiency and appeal for short-distance travel.

Percentages of ride trips by hours The analysis of bike rental patterns in Fig.2 reveals a distinct dichotomy between casual riders and members with respect to the timing of their rentals. Casual riders predominantly engage in bike rentals during the afternoon, particularly after work hours, suggesting a preference for leisure activities. In contrast, members demonstrate a bimodal distribution, with significant rental activities occurring both in the morning and afternoon. This pattern indicates that members utilize bike rentals not only for leisure purposes but also for commuting, reflecting a more integrated use of bike-sharing services in their daily routines.

Percentages of ride trips by days The analysis of bike rental patterns across days in Fig.2 reveals distinct preferences between casual riders and members. Casual riders exhibit a marked tendency to use electric bikes on weekdays while switching to classic bikes during weekends, with the highest number of trips occurring over the weekend. In contrast, members demonstrate a consistent utilization of both electric and classic bikes predominantly for commuting purposes, with peak usage concentrated on workdays.

4.3 Spatial analysis

As shown in Fig.3 and 4, The spatial analysis of bike rental patterns indicates that both casual riders and members predominantly utilize electric bikes for urban transportation and classic bikes for coastal and recreational destinations. Notably, members exhibit a clear preference for using bikes, particularly electric ones, for commuting purposes, with a significant concentration of start and end stations in business districts and residential areas. Conversely, casual riders tend to favor recreational use, with trip clusters often located near parks and tourist attractions. These spatial distributions underscore the diverse utilization patterns between user groups, emphasizing the importance of location-specific strategies to enhance service efficiency and user satisfaction.

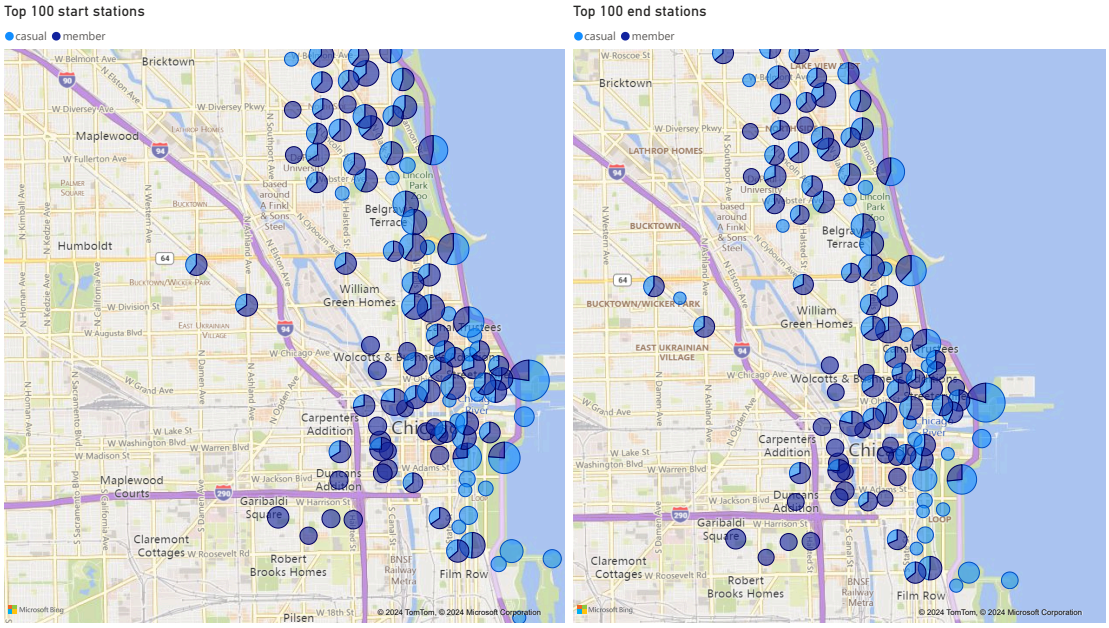

Figure 3: Top 100 start and end stations by classic bikes.

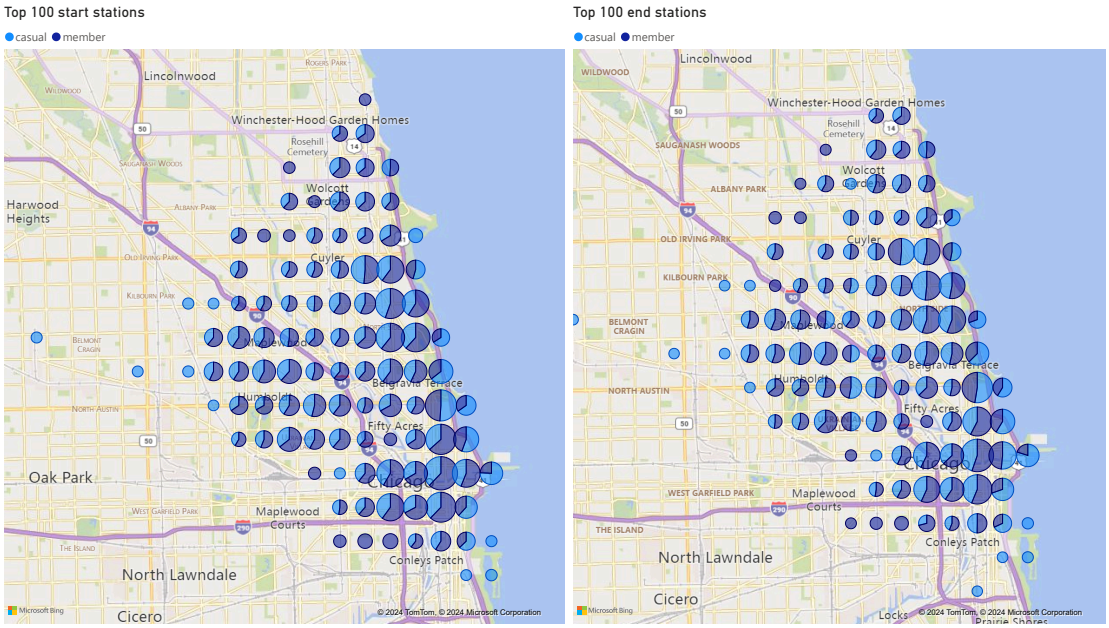

Figure 4: Top 100 start and end stations by electric bikes.

5 Proposed marketing strategies

The proposed marketing strategies aim to enhance the conversion of casual customers to annual members, leveraging the identified usage patterns. First, the implementation of special promotions and discounts is recommended. Weekday discounts for annual memberships can attract casual riders who typically use bikes for leisure post-work, encouraging them to commit to longer-term usage. Additionally, weekend specials can capitalize on peak casual usage times, making the transition to membership more appealing through cost savings. Second, enhancing membership perks is crucial. Offering exclusive access to electric bikes during peak hours and weekends ensures availability for members, thereby addressing their commuting needs effectively. Furthermore, introducing priority reservations for members, particularly for electric bikes, can cater to their preference for efficient and quicker trips. These strategies are designed to align with the behavioral patterns of casual riders, thereby increasing the likelihood of converting them into annual members.