Case Study: Cyclistic Bike Share

Capstone Project of Google Data Analytics Certification

Summary Report

Created By: Kamran Shafi Khan

Contents

1 Introduction **2**

2 Business Task **2**

3 Data Sources **2**

4 Data Cleaning and Transformation **2**

4.1 Descriptive Analysis 2
4.2 Checking Data for Errors 3
4.3 Removing Irrelevant Data 4

5 Summary of Analysis **5**

6 Key Findings **5**

6.1 Total Ride Count 6
6.2 Ride Count by Each Month 6
6.3 Median Ride Duration by Season 7
6.4 Ride Count by Weekday 7
6.5 Ride Count by Time of Day 7

7 Recommendations **8**

7.1 Target Peak Season (Jun-Aug) 8
7.2 Weekend Packages for Casual Riders 8
7.3 Targeted Marketing by Time of Day 8

8 Appendix **9**

8.1 Dashboard 9

List of Figures

1 Creation of ride_length and day_of_week columns 3
2 Histogram showing right skewed data for ride_length 4
3 Total Ride Count and Percentage of each Category 6
4 Ride Count Distribution by Month 6
5 Median Ride Duration by Each Season 7
6 Ride Count by Weekday 7
7 Ride Count by Time of Day 8

1 Introduction

Cyclistic is a bike-share fictional company in Chicago. There are two categories of riders according to its pricing plan - casual riders and annual members. Those who purchase single or full ride passes are referred to as casual riders and those who purchase annual memberships are called Cyclistic members.
The Cylclistic financial team believes that the company can make significant growth by increasing the number of annual memberships. For this purpose, the marketing analytics team, lead by Lily Moreno, have to devise a marketing strategy with the aim to convert the casual riders into annual members. This case study aims to analyze how the casual riders and annual members use Cyclistic Bikes differently.

2 Business Task

To analyze how casual riders and annual members use Cyclistic Bikes differently which will help guide the future marketing campaign.

3 Data Sources

1. The dataset provided in the case study is the historical data of the previous 12 months of Cyclistic rides. In our present case, we are using the data from Febuary 2023 - January 2024.
2. The data is organized in 12 CSV files where file each represents its particular month.
3. The dataset posseses the qualities of 'ROCCC' as supported by the following facts:
   (a) Data is provided by a trusted organization (Google) - Reliable and Original
   (b) Data from past 12 months which is sufficient for analysis - Comprehensive
   (c) Data updated with the latest additions e.g. using January 2024 as our most recent month - Current.
   (d) Data provided by Lyft Bikes and Scooters that operates the City of Chicago's ("City") Divvy bicycle sharing service - Cited
4. Data is made available under the license of Motivate International Inc. It has prohibited to use rider's personally identifiable information (PII) or for any unlawful purpose.

4 Data Cleaning and Transformation

Following were the data tools that were used for cleaning and manipulation:

- MS Excel - Data Exploration and Cleaning
- MS SQL - Data Cleaning, Transformation and Analysis

4.1 Descriptive Analysis

1. From the months of Febuary 2023 to January 2024 there are a total of 5,674,449 rides recorded.

2. The dataset is composed of the following 13 fields:
    - **ride_id** - (string) unique ID for each ride record
    - **rideable_type** - (string) type of bike used. Three types - classic, electric and docked.
    - **started_at** - (datetime) starting date and time of ride
    - **ended_at** - (datetime) ending date and time of ride
    - **start_station_name** - (string) name of station from where the ride started
    - **start_station_id** - (string) ID of the start station
    - **end_station_name** - (string) name of station from where the ride ended
    - **end_station_id** - (string) ID of the end station
    - **start_lat and start_lng** - (float) Latitude and Longitude value of position from where the ride started
    - **end_lat and end_lng** - (float) Latitude and Longitude value from of position where the ride ended
    - **member_casual** - (string) type of ride member i.e. member and casual

4.2 Checking Data for Errors

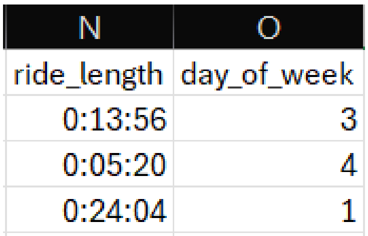
The whole data is present in 12 csv files. First, each csv file was converted to xlsx (excel) format. After that each file was inspected for errors and following observations were made:

1. Using Excel's ”Remove Duplicate” function, it was concluded that there are no duplicate rows present in the whole dataset.

2. Using pivot tables, it was observed no incorrect, outdated or missing values are present for the following fields:
    (a) ride_id
    (b) rideable_type
    (c) started_at and ended_at
    (d) member_casual

3. The remaining columns of start and end station's name, id, latitude and longitude contained missing values and were further inspected in SQL.

4. Created two new columns:
    (a) **ride_length** - ride duration in hh:mm:ss format. The difference between the started_at and ended_at columns.
    (b) **day_of_week** - week day of when the ride was started encoded in digits where 1 = Saturday.



N	O
ride_length	day_of_week
0:13:56	3
0:05:20	4
0:24:04	1

Figure 1: Creation of ride_length and day_of_week columns

5. Plotting the histogram of ride_length column, observed that the data values are highly skewed to the right as shown in the figure:

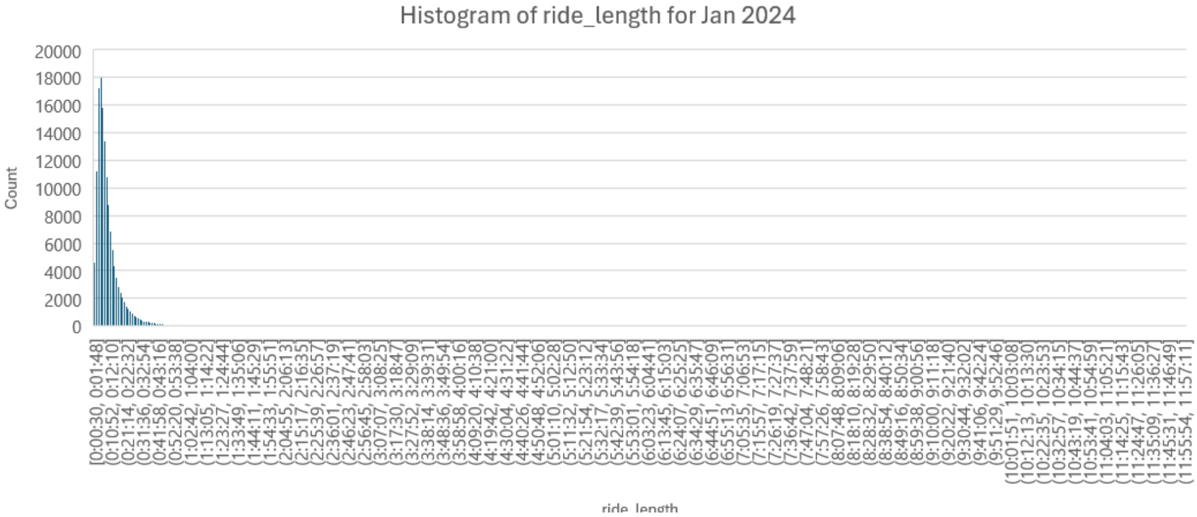


Figure 2: Histogram showing right skewed data for ride_length

This behaviour was observed in all of the remaining months indicating very high outlier values. Hence, we concluded that the **median** of ride_length will be a better for measuring its center than **mean**.

6. Upon sorting records by ride_length, we observed that there are a few records in each spreadsheet (about 2%) of the total, whose ride duration are either very small or very large to be considered practically possible i.e. there are some ride duration with zero seconds while there are also some span over to 5-6 days.

   There are also a few records (less than 0.001%) in each spreadsheet whose ride_length is composed of negative values i.e. ended_at time is lesser than started_at time. All of these rides were from the electric bike category.

4.3 Removing Irrelevant Data

After initial exploration, we had to to merge all the spreadsheets. Since the records of all the spreadsheets added up to more than 5 million rows so using Excel for merging the spreadsheets was not a practical option.
In order to overcome this challenge, we used MS SQL server. As we now had all of the records in a single file so it was easier to perform data processing task to the whole dataset.

1. Setting a threshold of minimum and maximum ride lengths to obtain better accuracy. The minimum threshold selected is 30 seconds while the maximum to 12 hours. These values are selected due to following reasons:

   (a) It is practically impossible for a bike ride to have duration of less than 30 seconds. Upon examination, these rides had the same starting and ending station indicating that the rides were practically not conducted. More than half of these rides were from the category ”electric bikes” indicating that there might be a starting issue with these type of bikes. Other than that, these values might be possible due to some system or technical issues, or human error.

   (b) It is also not possible for a rider to use a bike for more than 12 hours continuously due to limited human capacity. This might also indicate some technical or system errors. It can be also attributed to the fact that the rider might have forgotten to end his/her/zir ride and the timer might have been set on for a long time.

(c) There are about 93,687 rows whose ride duration is less than 30 seconds and 3,922 rows with duration greater than 12 hours. This make up only about 1.7% of the total records. Hence, this data is is very small compared to the whole. So we can safely remove this data on account of irrelevancy.

5 Summary of Analysis

The data was analyzed mostly using SQL server. There were some interesting trends and relationships observed. Following is a brief account of the analysis:

1. All of the data is merged vertically into a single table using the "UNION" statement of SQL. This choice is made to so that aggregation and calculation can be applied to the whole dataset easily.

2. Total ride count for both member groups was calculated. Casual group made 2,008,212 trips while annual members made 3,568,328 bike rides in the past 12 months.

3. The distribution of ride count along the months was also observed. For this purpose, the data records were grouped according to the ride month and member group. The ride count varied along the months for both groups.

4. For finding the median ride time, the months were divided into two groups:

   (a) Summer: April - September

   (b) Winter: October - March

   This division was made due to the fact that a similar trend that was being observed in each group i.e. in summer the median ride times were relatively higher than in the winter months. We took the average to provide a central picture for both summer and winter.

5. Data was aggregated by each week day and member category in terms of ride count.

6. In order to understand the effect of time of day on the number of rides, a new column by the name time_of_day was created, which divided the rides into three categories based on their starting time:

   (a) 6:00 am to 11:00 am - Morning

   (b) 12:00 pm to 4:00 pm - Afternoon

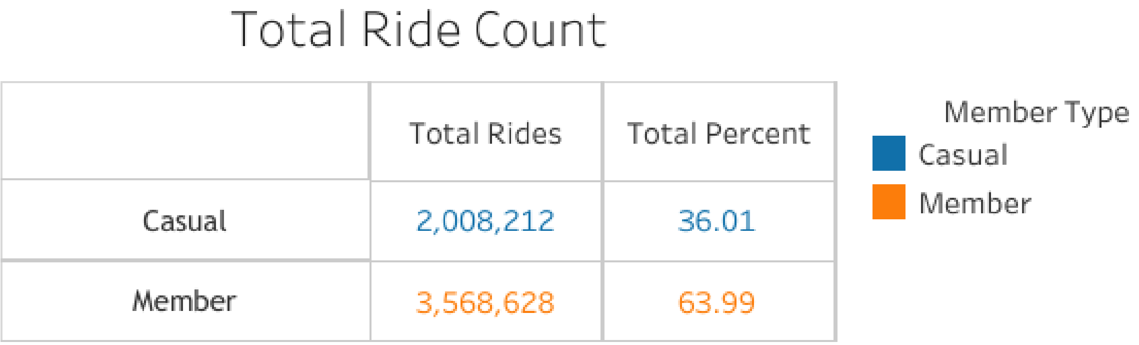
   (c) 5:00 pm to 5:00 am - Evening

6 Key Findings

After performing our analysis, we have come up with the following insights that are supported by the adjoining visuals:

6.1 Total Ride Count

About 63.99% of the total rides were conducted by annual members while 36.01% of them were by casual riders in the past 12 months.



Total Ride Count

	Total Rides	Total Percent
Casual	2,008,212	36.01
Member	3,568,628	63.99

Member Type

🟦 Casual

🟧 Member

Figure 3: Total Ride Count and Percentage of each Category

6.2 Ride Count by Each Month

Most of the rides were conducted in the months of June, July and August for both categories. Annual members ride count was higher in all months. However, the difference in the ride counts was relatively less in the months of June, July and August as compared to remaining months.

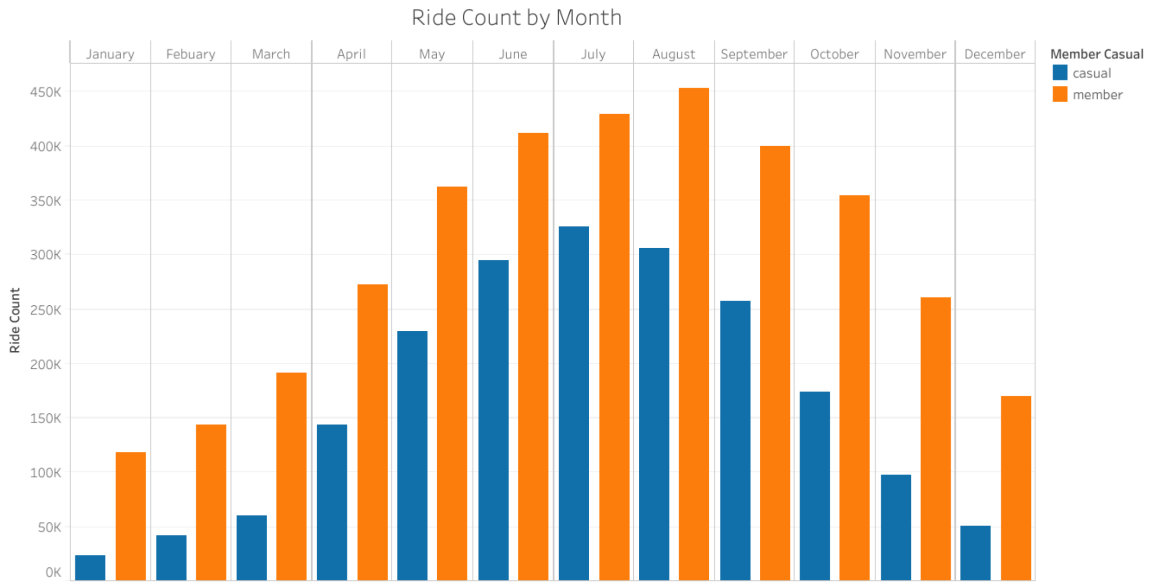


Figure 4: Ride Count Distribution by Month

6.3 Median Ride Duration by Season

Casual members lead the median ride duration in all months. This difference was relatively more in the summer (April - Sept) as compared to winter months (Oct - Mar).
Shorter ride times for annual members and longer ride times for casual members can be attributed to the fact that those who ride with annual subscription most probably use their bikes for regular commute and work purposes whereas the casual members are using bikes for leisure purposes and hence they have larger median duration.

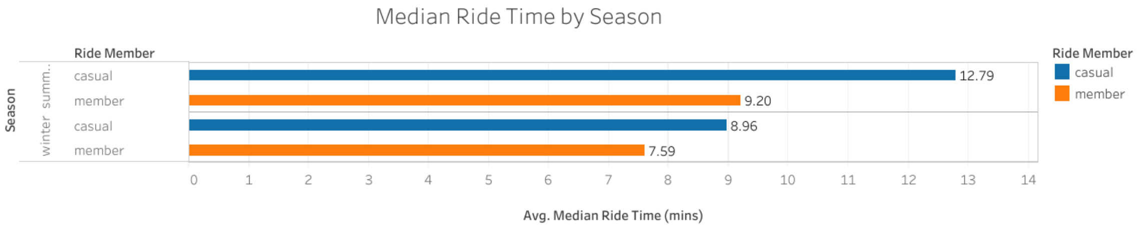


Figure 5: Median Ride Duration by Each Season

6.4 Ride Count by Weekday

Casual members ride more on weekends as compared to weekdays and opposite is true for annual members. This again proves the point that casual members are utilizing their bikes mostly in their free time, recreational and enjoyment activities whereas the annual members are using them for their daily and regular commute activities.
The difference in ride counts is relatively small in weekends as compared to weekdays as shown in the visual below:

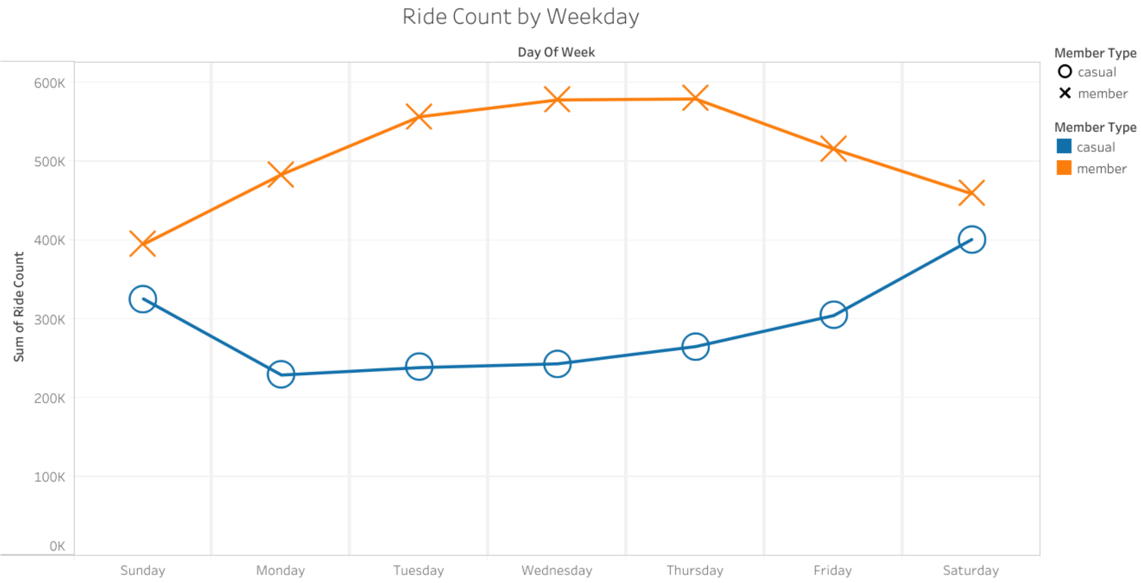


Figure 6: Ride Count by Weekday

6.5 Ride Count by Time of Day

In the morning time (6 - 11 am) the ride count of casual riders is lesser as compared to afternoon and evening.

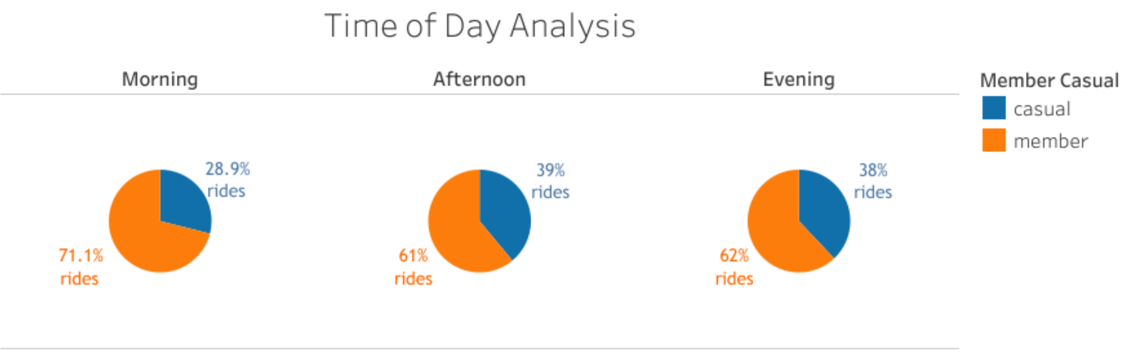


Figure 7: Ride Count by Time of Day

7 Recommendations

Based on the above visuals and the story it is communicating we have concluded the following top three recommendations to develop the future marketing strategy:

7.1 Target Peak Season (Jun-Aug)

Marketing campaigns aimed at converting casual riders to annual memberships can be most effective during peak ridership months: June, July, and August. These months see the highest overall ride counts for both categories, with a narrower gap between casual and annual member usage compared to other months.

7.2 Weekend Packages for Casual Riders

Since annual members take shorter trips primarily on weekdays, while casual riders embark on longer rides on weekends, their usage patterns suggest different motivations. Annual members likely use cycling for regular commutes and work-related trips, while casual riders choose it for leisure and recreational activities.
To capitalize on this, consider offering weekend packages that cater to casual riders' preferences while also showcasing the benefits of annual membership.

7.3 Targeted Marketing by Time of Day

Casual members demonstrate a higher usage rate during evenings and afternoons compared to mornings. Leverage this insight by tailoring marketing campaigns to these specific time frames.

8 Appendix

8.1 Dashboard

For accessing the interactive dashboard you can click on the following link below:
Click me! Cyclistic Case Study Tableau Dashboard