Exploratory Data Analysis:

Cyclistic Bike-share

Table of content

1.Introduction

1.1 Background of the case study

Welcome to the Cyclistic bike-share analysis case study! As a junior data analyst on Cyclistic's marketing team in Chicago, your mission is to unravel the biking behaviors of annual members and casual riders. In the pursuit of enhancing annual memberships, you'll follow the data analysis process—Ask, Prepare, Process, Analyze, Share, and Act. Throughout this journey, the Case Study Roadmap tables, equipped with guiding questions and key tasks, will be your compass.

In this scenario, the director of marketing, Lily Moreno, and her team are driven to understand the distinctions in how casual riders and annual members utilize Cyclistic bikes. The goal is to design a targeted marketing

strategy to convert casual riders into loyal annual members. To gain the approval of Cyclistic executives, your recommendations must be supported by compelling data insights and professional visualizations.

Let's meet the key players: Cyclistic, a bike-share program with a diverse fleet of 5,800 bicycles, including assistive options for riders with disabilities; Lily Moreno, the director of marketing; the Cyclistic marketing analytics team; and the detail-oriented Cyclistic executive team.

Since its launch in 2016, Cyclistic has grown to a fleet of 5,824 geotracked bicycles across 692 stations in Chicago. The company's marketing strategy has traditionally focused on broad consumer segments and flexible pricing plans. Annual members, identified as more profitable, have become a target for future growth.

Your mission, assigned by Lily Moreno, is to answer the first question: How do annual members and casual riders use Cyclistic bikes differently? This analysis is crucial for informing the development of a targeted marketing program. Join us on this data-driven journey as we explore historical bike trip data to uncover trends and insights that will shape Cyclistic's future marketing strategies."

1.2 Business task

Analyze the usage patterns of annual members and casual riders of Cyclistic bikes to identify key distinctions. The goal is to develop targeted marketing strategies aimed at converting casual riders into annual members.

2. Data Sources

The data sets used were; Divvy 2019 Q1 and Divvy 2020 Q1.Both datasets were provided by the google data analytics course.

The data set consists of the following data

Provided Variables The following variables in the Cyclistic bike-share dataset were provided directly:

- 'ride_id':representing the unique identifier for each ride.
- 'started_at': indicating the start time of the ride.
- 'ended_at':the end time of the ride.
- 'rideable_type':describing the type of rideable used (e.g., bike model).
- 'start_station_id': Numeric value representing the ID of the start station.
- 'start_station_name': Character vector indicating the name of the start station.
- 'end_station_id':representing the ID of the end station.
- 'end_station_name': indicating the name of the end station.
- 'member_casual':"member" is used to describe annual customers and "casual" for everyday customers
- 'date':date of the ride.
- **Derived Variables**

  The following variables were derived or created through calculations:

- 'year': the year of the ride.
- 'month': month of the ride.
- 'ride_length': duration of the ride in seconds.
- 'day_of_week': the day of the week.

These derived variables provide additional insights into temporal aspects and ride duration, contributing to a more comprehensive analysis of Cyclistic bike-share patterns.

This dataset comprises a total of 788,189 records and is a valuable source for understanding the usage patterns of Cyclistic bike-share riders. The data spans the first quater of the years 2019 and 2020.It covers various aspects, including both provided and derived variables.

3. Description of Data Sources Data Cleaning and Manipulation

3.1 Data collection

The necessary R packages (`"Tidyverse"` and `"Conflicted"`) were installed and used to load the data sets.

3.2 Data cleaning and wrangling

the data cleaning process involved analyzing the `q1_2020` and `q1_2019` data-sets for discrepancies.

The **`colnames`** function was used to analyze the column names of both datasets as they had to be the same for merging to occur. It was identified that the column names in the data-sets were mismatched.The column names in `q1_2019` were then renamed in order to match that of `q1_2020`.

The str() function was used to inspect the structure of both datasets

The **`str()`** function confirmed that the **`q1_2019`** columns **`ride_id`** and **`rideable_type`** were identified as numerical data. They were converted into character data using the **`mutate`** function to ensure compatibility with the **`q1_2020`** data.

The `bind_rows` unction was used to bind the `q1_2019` and `q1_2020` data-sets into one big data frame.

```
all_trips <- bind_rows(q1_2019, q1_2020)
```

The lat, long, birthyear, and gender fields in this data were removed as this data was dropped beginning in 2020.

3.3 Data Manipulation

A summary of the newly created data frame was made.

```
colnames(q1_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

In the “member_casual” column, there were two names for members (“member” and “Subscriber”) and two names for casual riders (“Customer” and “casual”). They were consolidated from four to two labels;“member” and “casual”.

The data could only be aggregated at the ride-level, which was too granular. Some additional columns of data were added such as day, month, year – to provide additional opportunities to aggregate the data.

`ride_length` column was calculated by taking the diffrence between the `ended_at` and `started_at` columns from the `all_trips` data-set

The structure of the `all_trips data` was identified using the `str()` function

`ride_length` was converted to numeric so calculations could be run on the data

A new data frame `all_rides_v2` was created since the previous data frame includes a few hundred entries when bikes were taken out of docks and checked for quality by Divvy or `ride_length` was negative

The summary statistics of the dataset was made

The summary statistics for the **ride_length** variable are as follows:

- Minimum (**Min.**): 1
- 1st Quartile (**1st Qu.**): 331
- Median: 539
- Mean: 1189
- 3rd Quartile (**3rd Qu.**): 912
- Maximum (**Max.**): 10632022

the data was aggregated based on mean,median,max and min

The average ride length per day of the week for each member and casual member was calculated;

The average ride length for casual riders is approximately 5,372.78 seconds.In contrast, the average ride length for members is significantly shorter at around 795.25 seconds.

This suggests a notable difference in riding patterns between casual and member riders, with casual riders tending to have longer rides on average.

In order to ensure logical and consistent representation of days of the week in visualizations and analyses, the levels of the 'day_of_week' variable have been explicitly ordered. This ordering follows the standard sequence: Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday. This step is crucial for maintaining the proper chronological order when exploring patterns or trends associated with specific days of the week in the dataset.

To understand how ride lengths vary based on both the member type and the day of the week, an analysis was conducted using the 'aggregate' function. The mean ride length was calculated for each combination of member type ('member' or 'casual') and day of the week. The results provide insights into potential patterns or differences in ride lengths based on these factors.

Some observations:

- On **Sunday**, **casual** rides tend to have a longer average duration (5061.3044) compared to **member** rides (972.9383).
- **Thursday** has the highest average duration for both **casual** (8451.6669) and **member** (707.2093) rides.
- **Wednesday** and **Tuesday** also show variations in average ride durations between **casual** and **member** categories.

resulting table provides insights into the number of rides and the average duration of rides for each combination of member type and day of the week. This structured summary facilitates a better understanding of how ridership patterns vary across different days and user types.

Some observations:

- On **Sunday**, **casual** rides tend to have a longer average duration (5061.3044) compared to **member** rides (972.9383).
- **Thursday** has the highest average duration for both **casual** (8451.6669) and **member** (707.2093) rides.
- **Wednesday** and **Tuesday** also show variations in average ride durations between **casual** and **member** categories.

Z-score was used in order to identify outliers in the dataset

4. Summary Analysis

4.1 Descriptive Analysis

In analyzing the ride lengths of Cyclistic bike rides, several key descriptive statistics offer insights into the distribution and characteristics of ride durations:

- The minimum ride length is 1 second, indicating the presence of rides with very short durations.
- The first quartile (25th percentile) is 331 seconds, signifying that 25% of rides have a duration of 331 seconds or less.
- The median ride length is 539 seconds, representing the middle value in the dataset. Half of the rides have a duration of 539 seconds or less, and half have a longer duration.
- The mean (average) ride length is 1189 seconds, providing a measure of the central tendency for the dataset.
- The third quartile (75th percentile) is 912 seconds, with 75% of rides having a duration of 912 seconds or less.
- The maximum ride length is 10,632,022 seconds, denoting the longest recorded ride in the dataset.

These statistics collectively illustrate the range and distribution of ride lengths, highlighting both typical patterns and instances of extreme durations. The presence of rides with very short or exceptionally long durations could be further explored in the context of the overall dataset and may contribute to the identification of outliers or specific user behaviors

4.2 Comparative Analysis

1. Number of Rides:

   For Casual Riders:The highest number of rides occurs on Saturday (13,473 rides), followed by Sunday (18,652 rides).The lowest number of rides is on Monday (5,591 rides).

   For Members:Sunday has the highest number of rides (60,197 rides), while Saturday has the lowest (59,413 rides).

2. Average Duration (seconds):

   For Casual Riders:The average ride duration is relatively consistent across days, ranging from 4,480 to 8,452 seconds.Thursday has the highest average duration (8,452 seconds), while Wednesday has the lowest (4,480 seconds).

   For Members:The average ride duration is consistently lower than that of casual riders, ranging from 707 to 974 seconds.Wednesday has the highest average duration (974 seconds), while Tuesday has the lowest (769 seconds).

These statistics provide insights into the patterns of bike rides for both casual riders and members, highlighting differences in usage behavior across days of the week. Casual riders tend to have more variable ride durations, while members generally have shorter and more consistent ride durations throughout the week.

4.3 Outliers

In the analysis of outliers, it's important to acknowledge that, due to the nature of a case study, direct access to the actual data and contextual information is often limited. The provided analysis reveals distinctive patterns and characteristics associated with the identified outliers, but a comprehensive understanding may require collaboration with data custodians or further investigation by the data providers. Here's a breakdown:

1. Count and Distribution:

   The 379 outliers are identified based on z-scores exceeding a threshold of $|\text{z-score}| > 3$, indicating a considerable deviation from the mean.

Without access to detailed contextual information, these outliers present unique instances that merit further exploration.

2. Start and End Timestamps:

   The wide range of start and end timestamps, from January 1, 2019, to May 19, 2020, signifies a diverse set of outlier instances over the observation period.

   The lack of direct access to underlying reasons emphasizes the need for collaboration with data custodians to unveil potential explanations.

3. Rideable Type and Station Information:

   The absence of a discernible pattern in rideable type, start station, or end station for outliers raises questions that may be clarified through additional insights provided by those managing the data.

4. Ride Length:

   The significant variability in ride lengths, culminating in an average of 737,187 seconds, suggests a departure from typical ride durations.

   Unraveling the root causes behind such prolonged or shortened rides may involve a deeper examination beyond the available dataset.

5. Day of the Week:

   The distribution of outliers across different days underscores the need for domain-specific knowledge to interpret whether certain days exhibit anomalous user behaviors.

6. Z-scores:

   Z-scores exceeding 22 standard deviations from the mean underscore the extreme nature of these outliers. Further exploration is essential to uncover potential anomalies contributing to such substantial deviations.

Overall Implications: The analysis underscores the importance of collaboration with data custodians or domain experts to provide insights into the specific circumstances surrounding these outliers. This collaborative approach will aid in uncovering nuanced explanations and refining the understanding of these unique instances, which is crucial for a comprehensive interpretation of the data.

5.Overview

5.2 Visual Representation of Data

5.3 Interpretation of Visualizations Key Findings

It has been observed that while annual members contribute to a higher count of rides, the average ride length tends to be greater among casual members. This intriguing pattern suggests that although annual members are more frequent riders, casual members, on average, engage in longer rides. This disparity in ride lengths between the two member types could potentially offer valuable insights into their distinct usage behaviors and preferences within the Cyclistic bike-sharing system.

6. Recommendation

- Short-Duration Rides:

The presence of rides with very short durations (minimum of 1 second) suggests the existence of brief trips. Investigating the nature of these short rides and their frequency could reveal opportunities for targeted promotions or adjustments to pricing structures for quick, convenient rides.

- Outliers Investigation:

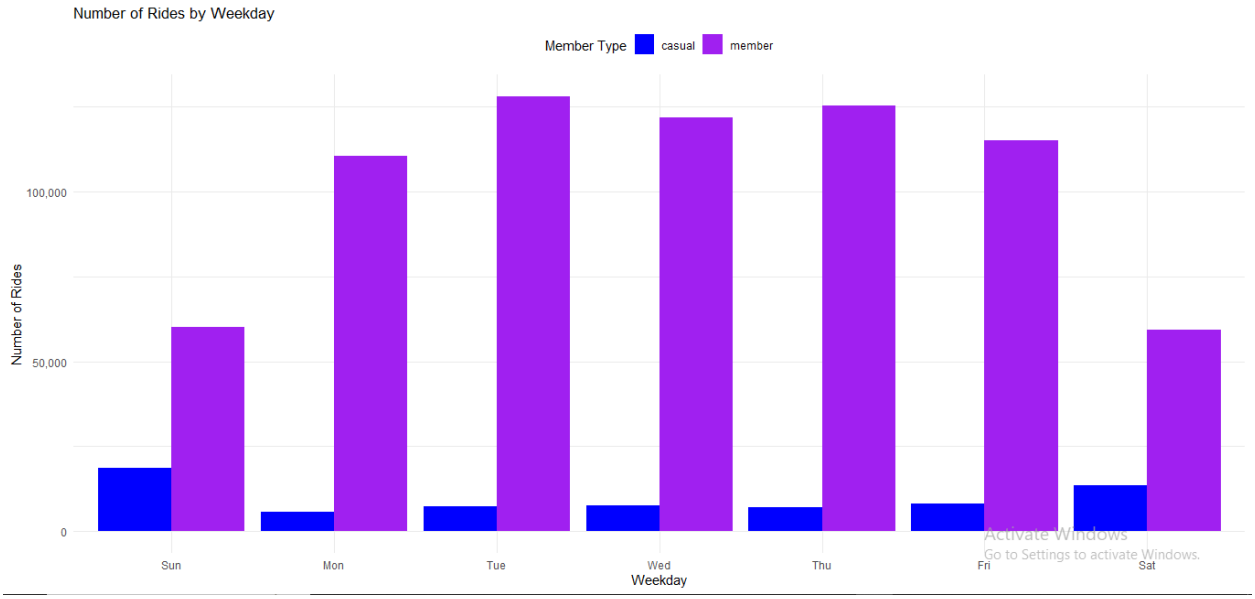

Figure 1: **This chart represents the count of rides for each day of the week. Specifically, Tuesday stands out with the highest number of rides among annual members, while Sunday takes the lead for casual members.**

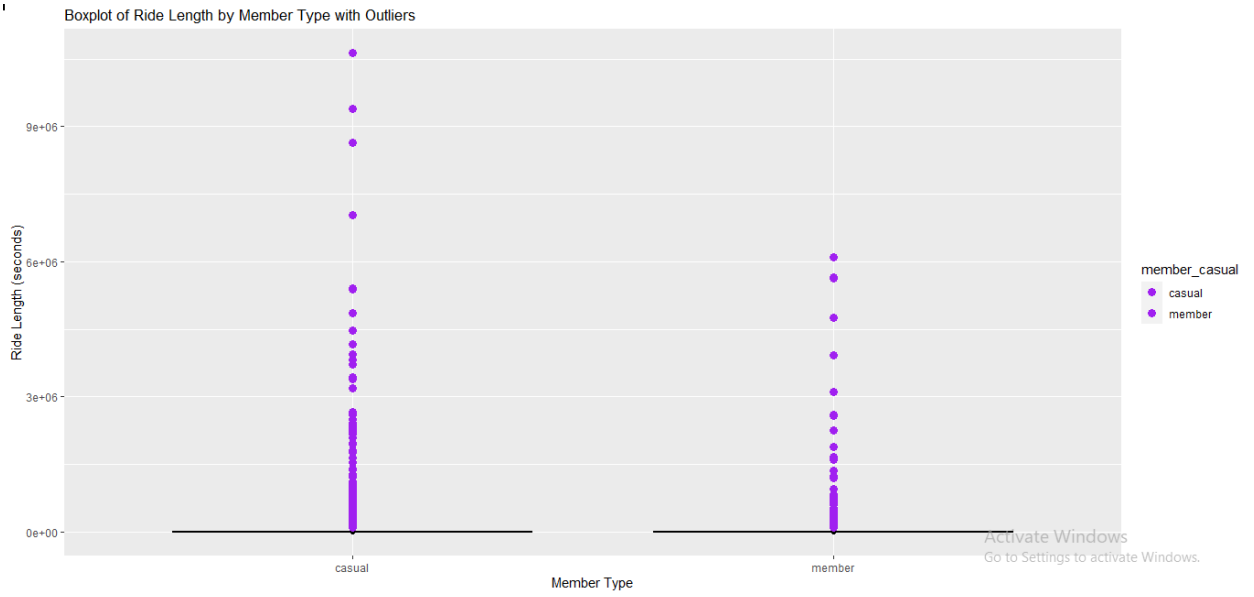

Figure 2: **This boxplot illustrates the distribution of ride lengths categorized by member type, showcasing a notable presence of outliers, particularly among casual members.**

Understanding the outliers identified in the dataset is crucial. More data will be needed in order to identify the potential causes

- Usage Patterns for Marketing:

promotions encouraging casual riders on Mondays could capitalize on the lower ride count, while targeted incentives on Saturdays might boost engagement with annual members.

- Ride Duration Considerations:

Acknowledging the differences in ride durations between casual riders and annual members is essential.Tailoring marketing messages based on these distinctions could involve emphasizing the flexibility and leisurely aspects of casual rides and the efficiency and convenience of shorter rides for annual members. Promotions could highlight the ease of incorporating bike rides into daily routines.

- Weekday-Specific Strategies:

Considering the variability in average ride durations across weekdays for both casual riders and members, weekday-specific marketing strategies can be devised. For example, promotions on Thursdays could focus on encouraging longer rides for casual riders, while campaigns on Wednesdays might highlight the quick and efficient nature of rides for members.

These recommendations provide a starting point for Cyclistic to refine its marketing approach based on observed patterns and trends. Further exploration and collaboration will contribute to a more comprehensive understanding of user behaviors and potential areas for service optimization.

7. Conclusion

In conclusion, the analysis of Cyclistic's bike-sharing dataset has revealed valuable insights into user behaviors, usage patterns, and potential areas for targeted marketing strategies. The descriptive statistics provided a comprehensive overview of ride lengths, highlighting variations in durations, while the comparative analysis delineated distinctions between casual riders and annual members across different days of the week. The identification and exploration of outliers underscored the need for further investigation and collaboration to understand the root causes of extreme deviations.

Despite the inherent limitations of a case study in accessing detailed contextual information, the findings offer actionable recommendations for Cyclistic. The presence of short-duration rides suggests opportunities for tailored promotions, and the investigation of outliers could uncover areas for service improvements. The recognition of usage patterns among different user groups allows for the development of targeted marketing campaigns, maximizing engagement on specific days and emphasizing the unique benefits for casual riders and annual members.

The collaborative exploration with data custodians is crucial to unraveling the complexities behind outliers and gaining a more nuanced understanding of user behaviors. As Cyclistic moves forward, these insights can inform strategic decisions to enhance user experiences, refine marketing approaches, and optimize service offerings.

In summary, the analysis provides a foundation for Cyclistic to leverage data-driven insights in refining its operations and marketing strategies. The collaborative and exploratory nature of further investigations will be essential in unlocking the full potential of the dataset and ensuring the effectiveness of future initiatives.