Google Data Analytics – Capstone Project: Creating a Case Study

Track 01: Working with existing questions and datasets

Case Study 1: How does a bike-share navigate speedy success?

Report

I. Introduction:

The goal of this analysis is to understand how annual members and casual riders use Cyclistic bikes differently. The task will be done by analyzing the public Cyclistic trip data in 2021. This information will be used to develop a strategy to convert casual riders into annual members.

II. Data Sources:

The data used in this analysis comes from the official public data of Cyclistic, which is available at https://divvy-tripdata.s3.amazonaws.com/index.html. This data source provides detailed information on the usage of Cyclistic bikes, including trip duration, trip frequency, trip start and end times, and also trip start and end locations.

The data is organized in CSV files, with each file containing information on Cyclistic trips for a specific time period. The files are named according to the time period they cover, for example, 202109-divvy-tripdata.csv contains data for September 2021 or 202112-divvy-tripdata.csv contains data for December 2021. As I have mentioned before in the introduction part, for this analysis, I will mainly use the data of 2021 so that the name of attributes in the data files will be the same. This make it easier to analyze the data since we could take less steps in the data pre-processing section.

The data comes from a reputable source (Cyclistic) and is regularly updated, so there are no major concerns about its credibility. However, as with any data source, there may be some bias or limitations that need to be taken into account when analyzing the data. For example, the data only covers trips made using Cyclistic bikes and does not include information on trips made using other modes of transportation. To ensure that the data ROCCC (Reliable, Original, Comprehensive, Current, Cited), it is important to verify the integrity of the data and to use multiple sources of information when possible.

The data used in this analysis is publicly available and does not contain any personally identifiable information, so there are no major concerns about licensing, privacy, or security. To ensure accessibility, the data can be downloaded from the website and analyzed using commonly available tools such as Excel or R. To verify the integrity of the data, you can check for missing or inconsistent values and compare the data to other sources of information to ensure that it is accurate and reliable.

The Cyclistic trip data provides detailed information on the usage patterns of annual members and casual riders, such as trip start and end times, and trip start and end locations. By analyzing this data, I can identify patterns and trends in the usage of Cyclistic bikes by these two groups of users and draw conclusions about how they use the bikes differently. In addition, there are some attributes that we can added to the data in order to calculate the statistic to the data. For example, in the next section of using R to work with the data, I will added two columns day_of_week and ride_length to the data in order to find the trends people when they using Cyclistic bike.

As with any data source, there may be some limitations or issues that need to be taken into account when analyzing the Cyclistic trip data. For example, the data only covers trips made using Cyclistic bikes and does not include information on trips made using other modes of transportation. Additionally, there may be some missing or inconsistent values in the data that need to be addressed before conducting the analysis.

III. Processing and Analysis:

III.1. Processing:

For this project, I am using R Programming to clean, analyze, and create visualizations for the data. The reason why I chose R is that it is a powerful and versatile tool for data analysis, with a wide range of built-in functions and libraries for cleaning, manipulating, and visualizing data. Additionally, R is widely used in the data analysis community, which means that there are many resources and examples available to help me with analysis. Another reason is that since the data used in this project is very large that spreadsheets cannot handle it easily so using R is the better option in this situation.

To ensure the integrity of my data, I have taken several steps to verify its accuracy and reliability. First, I checked the data for missing or inconsistent values and addressed any issues that I found. I also compared the data to other sources of information to ensure that it is accurate and reliable. To ensure that my data is clean and ready for analysis, I have taken several steps to prepare it. First, I used R's built-in functions to check for missing or inconsistent values and addressed any issues that I found. I also used R's data manipulation functions to transform the data into a format that is suitable for analysis. To verify that my data is clean and ready for analysis, I can use several techniques. For example, I can use summary statistics and visualizations to check for any unusual or unexpected patterns in the data. I can also use R's built-in functions to check for missing or inconsistent values and ensure that the data is in the correct format for analysis. For more detail information about this section, please read the below section: "Data cleaning and manipulation".

III.2. Data cleaning and manipulation:

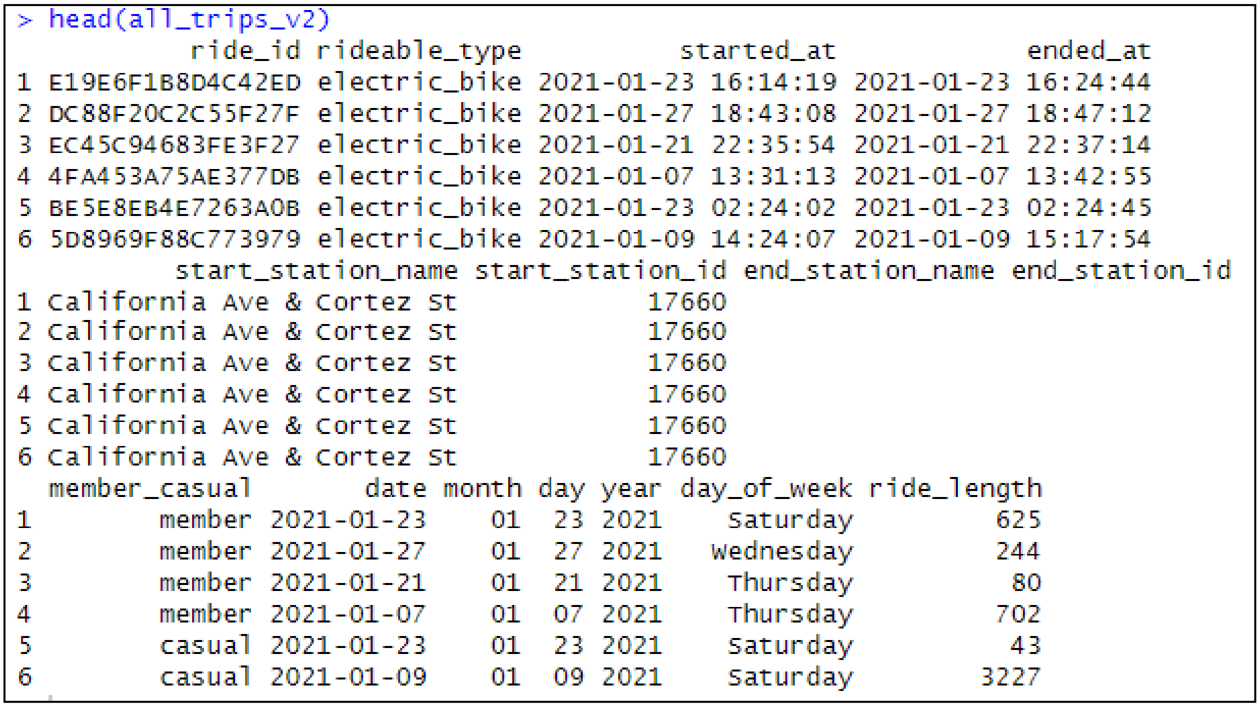
Before analyzing the data, it was necessary to clean and manipulate it to ensure its accuracy and reliability. This involved several steps, including checking for missing or inconsistent values, transforming the data into a suitable format for analysis, and addressing any issues that were identified.

First and foremost, before starting the pre-processing section, I have loaded and checked the data first. There are total 12 data files that have been stored in .csv format. These data contain information of Cyclistic's customers from January to December of the year 2021. The data contains a total of 13 attributes which are ride_id, rideable_type, started_at, ended_at, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng, and member_casual (label of the data). After look through some of the contents in the data, I have combined all of 12 data into only one dataframe and called it all_trips. To be more convinient, I have also removed some variables that I had not used during the analysis process which are start_lat, start_lng, end_lat, end_lng.

After that, I started the cleaned up amd added data process. At the begin of this section, I have, first, split the information about date of started_at attribute into date attribute (a new added attribute), then I continue to split the date into month, day, year and stored these information into new separated variables. Also, I used the date attribute to define the day of the week including Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday. This information will be useful for creating visulizations in the later section. After finished working with the date data, I continued to calculate the ride length (the amout of time that customers spend on riding a bike). Using this variable together with day_of_week variable can help me create useful charts which can be used to help making decisions about the project question: "How do annual members and casual riders use Cyclistic bikes differently?". One notice about the ride_length variable is that the data is mentioned in second and we need to convert it to numeric type before using. After computing the ride_length, I checked the data once more time and founded that there are some bad data in the dataframe. The 'start_station_name' column contains invalid entries of HQ QR {HQ QR means that the bike was taken out by the Divvy's team for maintenance'} and the ride_length contains some negative values so I need to remove it before moving to next part. The code of this part is very simple:

```
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" | all_trips$ride_length<0),]
```

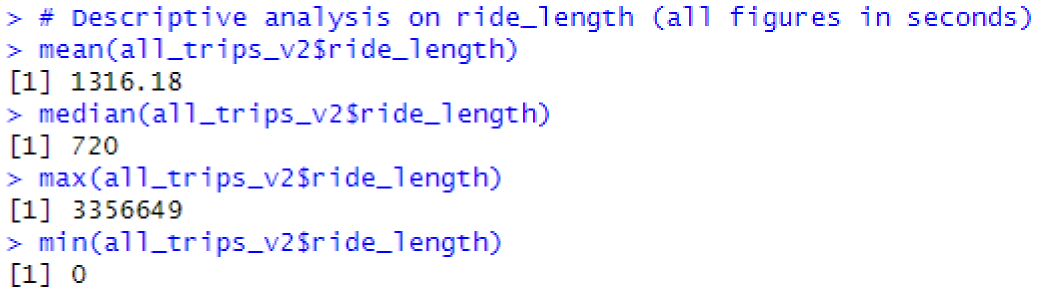
The all_trips_v2 is the final data that I worked on so as to calculate the statistic and create some useful visualizations with it. Now, we should look at the contents of the data before moving to next part:



```
> head(all_trips_v2)
           ride_id rideable_type          started_at            ended_at
1 E19E6F1B8D4C42ED electric_bike 2021-01-23 16:14:19 2021-01-23 16:24:44
2 DC88F20C2C55F27F electric_bike 2021-01-27 18:43:08 2021-01-27 18:47:12
3 EC45C94683FE3F27 electric_bike 2021-01-21 22:35:54 2021-01-21 22:37:14
4 4FA453A75AE377DB electric_bike 2021-01-07 13:31:13 2021-01-07 13:42:55
5 BE5E8EB4E7263A0B electric_bike 2021-01-23 02:24:02 2021-01-23 02:24:45
6 5D8969F88C773979 electric_bike 2021-01-09 14:24:07 2021-01-09 15:17:54
          start_station_name start_station_id end_station_name end_station_id
1 California Ave & Cortez St            17660
2 California Ave & Cortez St            17660
3 California Ave & Cortez St            17660
4 California Ave & Cortez St            17660
5 California Ave & Cortez St            17660
6 California Ave & Cortez St            17660
  member_casual       date month day year day_of_week ride_length
1        member 2021-01-23    01  23 2021    Saturday         625
2        member 2021-01-27    01  27 2021   Wednesday         244
3        member 2021-01-21    01  21 2021    Thursday          80
4        member 2021-01-07    01  07 2021    Thursday         702
5        casual 2021-01-23    01  23 2021    Saturday          43
6        casual 2021-01-09    01  09 2021    Saturday        3227
```

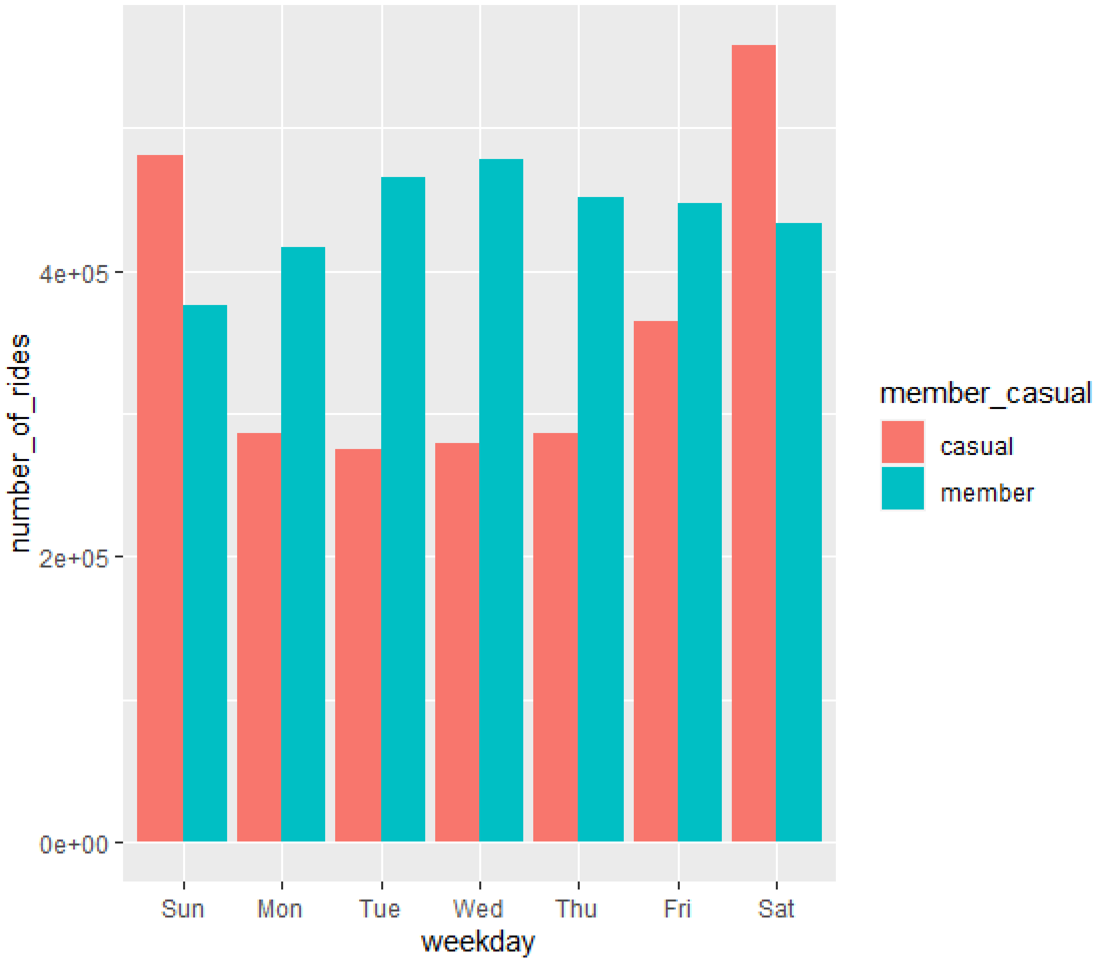
III.3. Analysis:

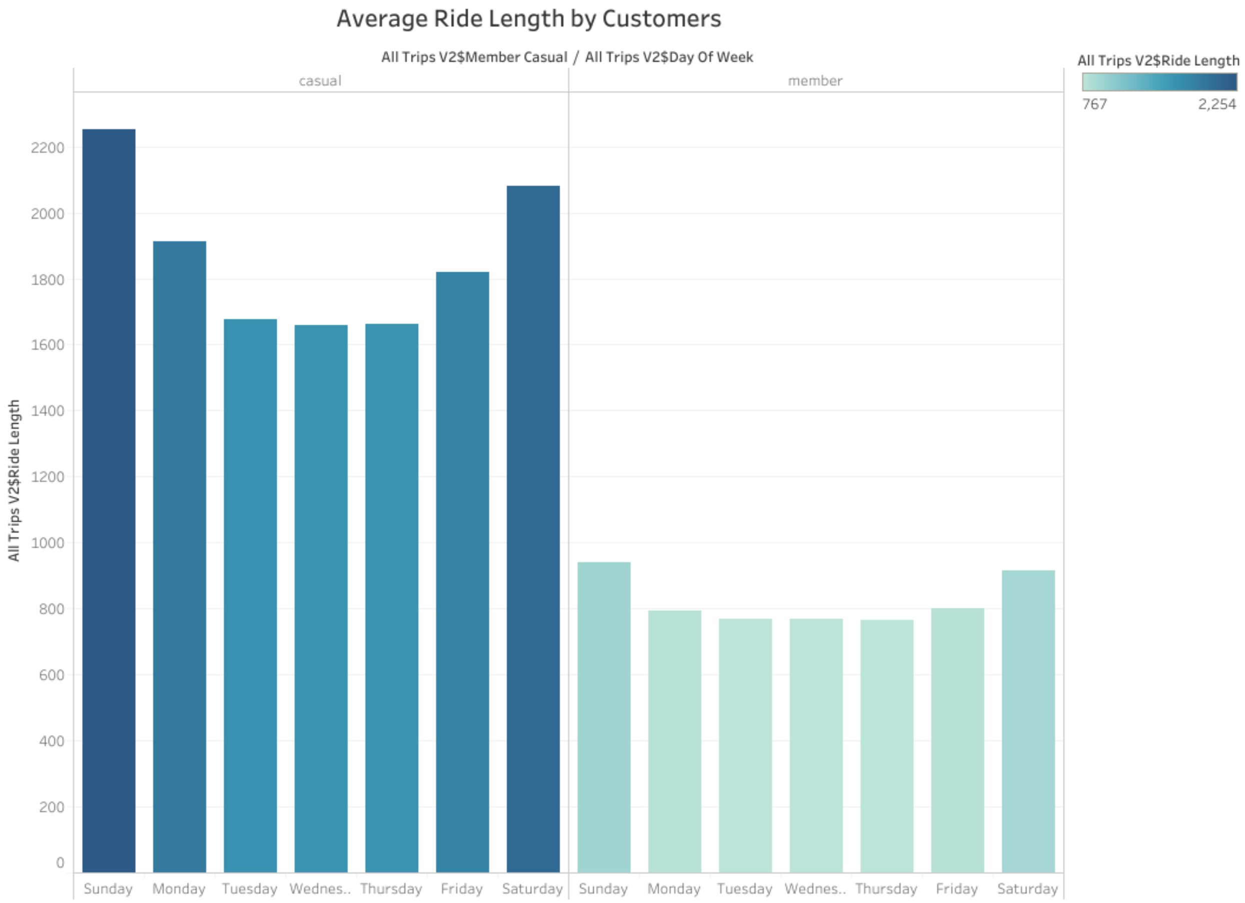
Now, let’s calculate some descriptive analysis on ride_length.



```
> # Descriptive analysis on ride_length (all figures in seconds)
> mean(all_trips_v2$ride_length)
[1] 1316.18
> median(all_trips_v2$ride_length)
[1] 720
> max(all_trips_v2$ride_length)
[1] 3356649
> min(all_trips_v2$ride_length)
[1] 0
```

Next, I have create two bar charts for the data:





These two bar charts give an me some idea about how customers spend their time on the activity of riding a bike. The first image provide information about number of rides that two type of customers 'causal' and 'member' use bike to travel during day_of_week. Based on the bar graph I provided, it appears that members took more rides than casual users on every day of the week. The highest number of rides for both groups were taken on Saturday, with members taking slightly more rides than casual users. The lowest number of rides for both groups were taken on Monday, with members taking significantly more rides than casual users.

These trends suggest that annual members and casual riders use Cyclistic bikes differently. Members seem to use the bikes more consistently throughout the week, while casual users tend to use the bikes more on weekends. This could indicate that members are using the bikes for regular commuting or daily activities, while casual users are using the bikes more for leisure or occasional trips.

In terms of the director of marketing’s belief that the company’s future success depends on maximizing the number of annual memberships, these trends could support that idea. Since members are using the bikes more consistently and frequently than casual users, increasing the number of annual memberships could lead to an overall increase in bike usage and revenue for the company.

Turning to the second chart, the image describe the duration of a ride of two classes. Based on the bar graph, it appears that casual users tend to have longer ride durations than members on all days of the week. The highest average duration for casual users is on Saturday, while the highest average duration for member users is on Friday. The lowest average duration for both casual and member users is on Wednesday.

These trends suggest that while members use the bikes more frequently, casual users tend to use them for longer periods of time. This could indicate that casual users are using the bikes for leisurely rides or sightseeing, while members are using them for shorter, more practical trips such as commuting.

In terms of the director of marketing's belief that the company's future success depends on maximizing the number of annual memberships, these trends provide some additional context. While increasing the number of annual memberships could lead to an overall increase in bike usage and revenue, it's also important to consider the different usage patterns of casual users. Strategies could be developed to encourage casual users to take longer, more frequent rides, potentially increasing their likelihood of converting to annual memberships.

IV. Recommendations:

Based on the analysis project, I would see that while annual members use bike more frequently, casual customers use bike for a long run. Therefore, for the believe of the director of marketing: "the company's future success depends on maximizing the number of annual memberships", the project would support this idea. My recommendation is that we would encourage customers to use bike frequently. The company can create some events, activities that encourage all type of customers to use bike for longer trip and make them to become annual membership. In addition, the company can encougage new customers try using bikes to go to work if there office is not so far from their house. Company can tell people the benefits of using bikes when travel in short distance and persuade them to become annual membership.