Cyclistic Bike Sharing Analysis

Introduction

This notebook is a step by step walk through of my capstone case study project of the Google Data Analytics Professional Certificate.

This notebooks serves as a report with the following deliverables:

1. A clear statement of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of analysis
5. Supporting visualizations and key findings
6. The top three recommendations based on the analysis

Scenario

The director of marketing at Cyclistic, a bike-share company in Chicago, believes the company's future success depends on maximizing the number of annual memberships. Therefore,your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Phase 1 - Ask

This phase covers identifying the key stakeholders, the requirements of the analysis and then forming the business task in light of both of these findings.

Key Stakeholders

Lily Moreno: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.

Cyclistic executive team: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

Business Task

The business task is defined as finding key differences between annual members and casual riders and to see how we can leverage these differences to help casual riders opt for the annual memberships which the executives believe are the key to the company's future success.

Phase 2 - Prepare

In this phase we collect and organize the data and determine its credibility.

Data Source

The data we are using is Cyclistic's 12 month historical trip data from July 2022 - Jun 2023 available here. The datasets have been deemed okay and relevant to our business task prior to this analysis.

Licensing

The data has been made available by Motivate International Inc. under this license.) This is public data that you can use to explore how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit you from using riders' personally identifiable information. This means that you won't be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Phase 3 - Process

In this phase we cleaned and manipulated data to make it more suited to the analysis. The tool we decided to use for our analysis was our because of its ease of use and power.

Data Cleaning and Manipulation

Here is a step by step breakdown for explanation and reproducibility of the cleanings and manipulations we applied to the data.

1. We loaded all the necessary libraries and loaded each months data separately.

```
library(tidyverse)
library(lubridate)
library(ggplot2)

jul<-read_csv('202207-divvy-tripdata.csv')
aug<-read_csv('202208-divvy-tripdata.csv')
sep<-read_csv('202209-divvy-tripdata.csv')
oct<-read_csv('202210-divvy-tripdata.csv')
nov<-read_csv('202211-divvy-tripdata.csv')
dec<-read_csv('202212-divvy-tripdata.csv')
jan<-read_csv('202301-divvy-tripdata.csv')
feb<-read_csv('202302-divvy-tripdata.csv')
mar<-read_csv('202303-divvy-tripdata.csv')
apr<-read_csv('202304-divvy-tripdata.csv')
may<-read_csv('202305-divvy-tripdata.csv')
jun<-read_csv('202306-divvy-tripdata.csv')
```

2. The started_at and ended_at columns for July and August datasets were in chr format and for the sake of consistency and ease of analysis later down the line we had to convert them to the proper datetime format which matched the other months' datasets.

```
str(jul$started_at)
str(jul$ended_at)

str(aug$started_at)
str(aug$ended_at)

str(sep$started_at)
str(sep$ended_at)

jul[["started_at"]] <- strptime(jul[["started_at"]],format = "%m/%d/%Y %H:%M")
jul <- mutate(jul,started_at = as_datetime(started_at),format="%Y-%m-%d %H:%M:%S")
jul[["ended_at"]] <- strptime(jul[["ended_at"]],format = "%m/%d/%Y %H:%M")
jul <- mutate(jul,ended_at = as_datetime(ended_at),format="%Y-%m-%d %H:%M:%S")

aug[["started_at"]] <- strptime(aug[["started_at"]],format = "%m/%d/%Y %H:%M")
aug <- mutate(aug,started_at = as_datetime(started_at),format="%Y-%m-%d %H:%M:%S")
aug[["ended_at"]] <- strptime(aug[["ended_at"]],format = "%m/%d/%Y %H:%M")
aug <- mutate(aug,ended_at = as_datetime(ended_at),format="%Y-%m-%d %H:%M:%S")
```

3. We combined all the data into quarters, and also a single consolidated dataset and removed the longitudinal and latitudinal data from the data. The cleaning steps were applied to all of the above mentioned datasets, but are only being shown for the all trips data for the sake of readability.

```
q3_2022 <- bind_rows(jul,aug,sep)
q4_2022 <- bind_rows(oct,nov,dec)
q1_2023 <- bind_rows(jan,feb,mar)
q2_2023 <- bind_rows(apr,may,jun)

all_rows<-bind_rows(q3_2022,q4_2022,q1_2023,q2_2023)

all_trips <- all_rows %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
  str(all_trips)
```

4. We added date, month, day, year and day_of_week columns of the trips by extracting them from the started_at columns, as we'll be doing some analysis based on these columns later.

```
all_trips$date <- as.Date(all_trips$started_at)
all_trips$month <- format(as.Date(all_trips$date), "%m")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips$year <- format(as.Date(all_trips$date), "%Y")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
```

5. Added a ride_length column to the data by calculating the difference in the started_at and the ended_at time.

```
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
str(all_trips)
```

6. We transformed the ride_length column to numeric for ease of use.

```
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
```

7. And for the sake of data validity we went ahead and removed the ride lengths that were negative or zero. We cleaned up 20889 rows with this step.

```
all_trips_v2 <- all_trips[!(all_trips$ride_length<=0),]
```

Phase 4 - Analyze

In this phase we had the core part of our analysis after the data was made ready in the previous step.

Summary of Analysis

Here is a step by step breakdown of our analysis steps.

1. We first performed a descriptive analysis of our data set and found the min, max, median and mean for the ride length.

```
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     334     590    1104    1037 2483235
```

2. Furthermore we compared the descriptive statistics for different rider types.

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

all_trips_v2$member_casual <chr>	all_trips_v2$ride_length <dbl>
casual	1670.8120
member	745.3942

2 rows

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
```

all_trips_v2$member_casual <chr>	all_trips_v2$ride_length <dbl>
casual	720
member	517

2 rows

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
```

all_trips_v2$member_casual <chr>	all_trips_v2$ride_length <dbl>
casual	2483235
member	93580

2 rows

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)
```

all_trips_v2$member_casual <chr>	all_trips_v2$ride_length <dbl>
casual	1
member	1

2 rows

3. We also compared the average ride duration for each day of the week across different rider types.

```
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesda
y", "Thursday", "Friday", "Saturday"))

aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
```

all_trips_v2$member_casual <chr>	all_trips_v2$day_of_week <ord>	all_trips_v2$ride_length <dbl>
casual	Sunday	1975.2954
member	Sunday	823.0356
casual	Monday	1633.4171
member	Monday	707.6603
casual	Tuesday	1460.8541
member	Tuesday	712.7872
casual	Wednesday	1399.3720
member	Wednesday	711.7820
casual	Thursday	1412.1802
member	Thursday	715.2788

1-10 of 14 rows   Previous **1** 2 Next

4. We repeated the same steps for Q3 of 2022 and Q1 of 2023 to enable our comparative analysis later.

Phase 5 - Share

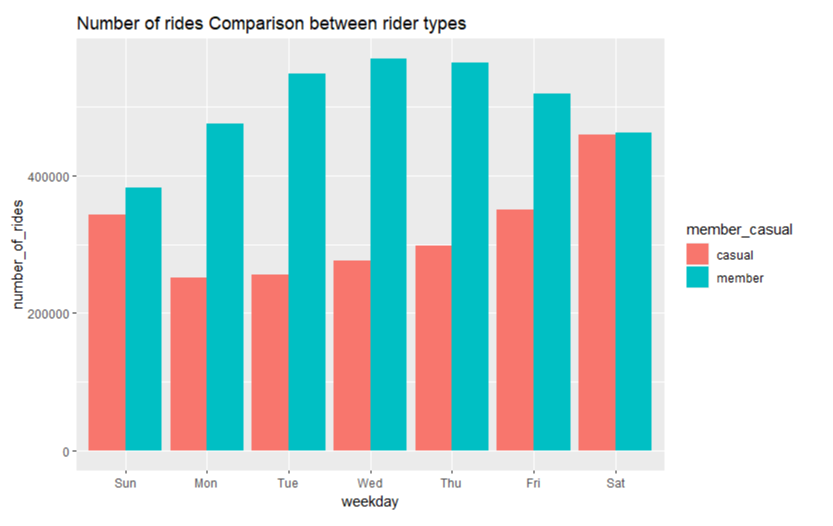
In this phase we summarize and analysis and present key findings with the help of supporting visualizations.

Supporting Visualizations and Key Findings

Here we have some key findings and visualizations by analyzing the average duration and number of rides in 3 key scenarios.

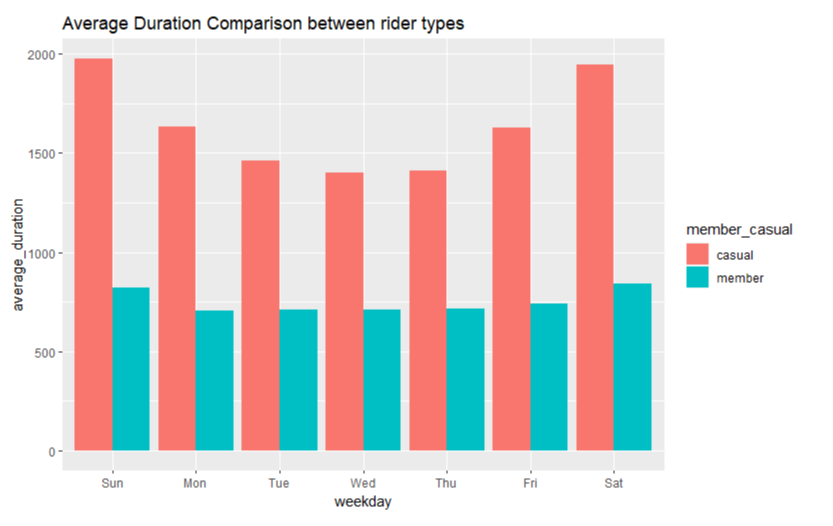
1. We present the comparison between the number of rides between members and casual riders.

- We can see that the members had more number of rides throughout the weeks and the casual riders had their spikes over the weekends.



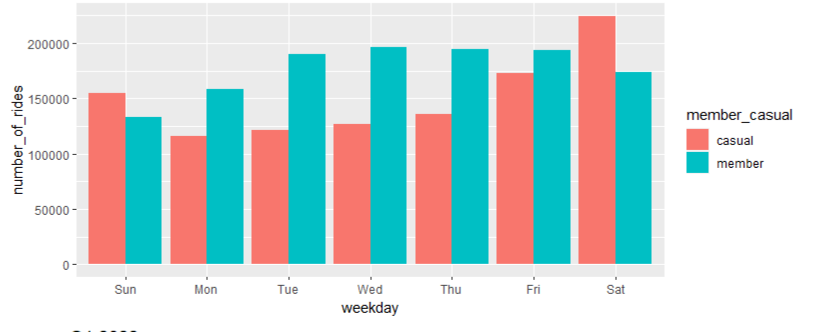
2. We present a comparison between the average duration of a ride between members and casual riders.

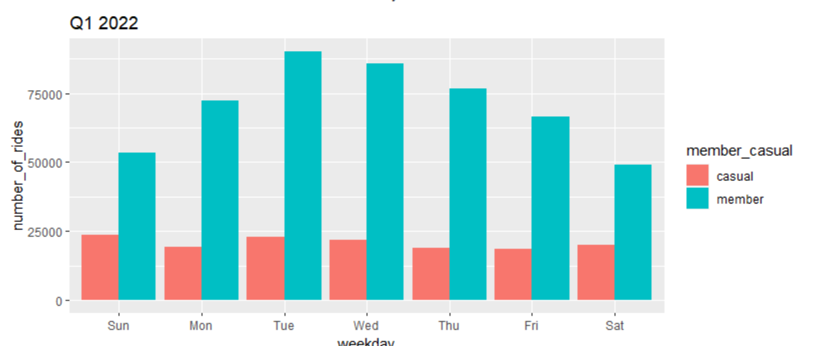
- In the average duration comparison we see that the casual riders had much higher average durations compared to the members. The average duration for members remains consistent indicating that they had set routes which may have been a key factor in making them opt for the membership.



3. We present a comparison between the number of rides between members and casual riders between Q3 2022 and Q1 2023

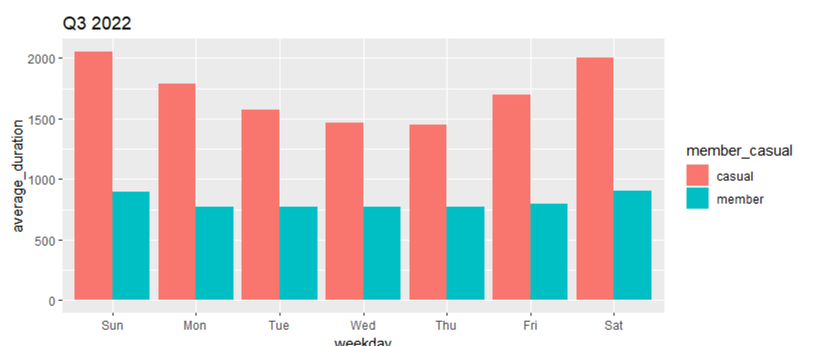
- We see again the same pattern, members had higher and consistent number of throughout the week, whereas casual riders were peaking on the weekends.
- Number of rides dropped for both in the colder months of the year, but the fall was especially noticable in casual rider numbers.

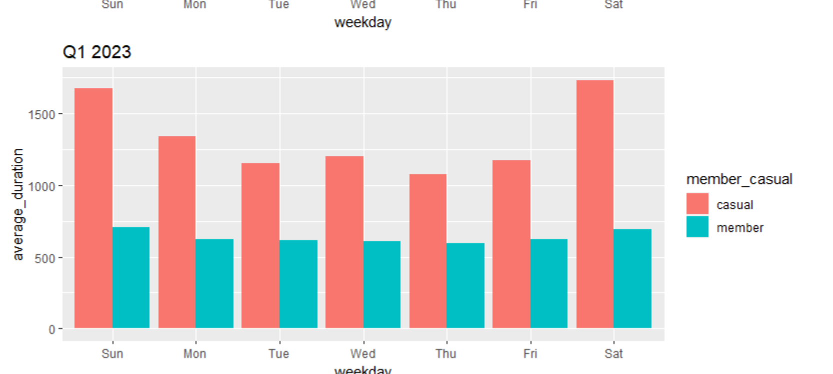




4. We present a comparison between the average duration of a ride between members and casual riders between Q3 2022 and Q1 2023.

- Comparing the average ride durations across the two quarters, again indicated that casual riders had higher ride times
- However in the colder months, their numbers dropped more than the members, who again displayed consistency indicative of a routine





Phase 6 - Act

In this phase we wrap up our findings and provide actionable insights based on our these findings.

Top Recommendations

Here are three recommendations for acting on the findings of this analysis.

1. As casual riders are having higher riding durations, a way of marketing to them would be to show them how much they could save on a membership plan as to what they spend on individual rides. Showing them actual potential savings would be super effective
2. Target the winter months with special promotions and offers which will lead to increased rides from both member types.
3. Offer varying levels on memberships. Instead of just an annual plan, introduce, monthly, weekly, weekend, and daily membership. This would be a great way to convert a significant chunk of the weekend casual riders into seeing the benefits of a membership.