Introduction

Case Study: How Does a Bike-share Navigate speedy Sucess?

This case study represents course 8 “Capstone project” of the Google Data Analytics Professional Certificate on Coursera

About The Company

In 2016, cyclistic Launched a successful Bike-share Offereing. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across chicago. Cyclistic has flexible pricing plans: single day passes, full day passes and annual memberships. Customers who purchase annual memberships are cyclistic member. Therefore, the team wants to understand how casual riders and annual members use Cyclistic bikes differently.

There are three questions that will guide the future marketing program:
1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

The Marketing Director has chosen me to answer the First Question: How Do annual member and casual riders use Cyclistic Bikes Differently?

Deliverables:

- A clear statement of the Business task
- A description of all the data sources Used
- Documentation of any cleaning or manipulation of data
- A summary of analysis
- Supporting Visualizations and key findings
- Your Top three Recommendations based on your analysis

Business Task

Analyse the Cyclistic data set for the year 2021 to understand how annual members and casual riders use Cyclistic bikes differently.

Stakeholders

Cyclistic - The bike-share company with 5824 bikes and 692 docking stations all over Chicago.

Lily Moreno - The director of marketing who has requested the analysis for her new marketing strategy.

Cyclistic Marketing Analytics Team - A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide marketing strategy.

Cyclistic Executive Team - Detail-oriented executive team that will decide whether to approve the recommended marketing program.

Data sources Descritption:

The Cyclistic's historical trip data has been provided by Motivate International Inc. under this License. Due to the limitations of the free version tools I am accessible to use and the data being too large data, I tried initially the 12-month data(August 2020 - July 2021), the system(RStudo Cloud free) crushed several times and had reached the available data usage. I eventually chose the 2-month data(June, July 2021) to perform this case study. Each data set is in Csv format and details every ride logged by Cyclistic customer. All user's personal data has been scrubbed for privacy.

Cleaning, Filtering and Manipulation of Data:

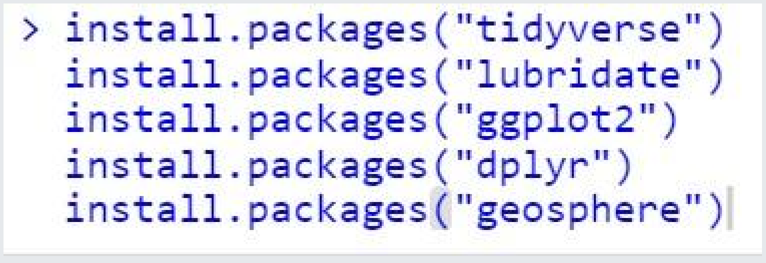
Data cleansing helps avoid various structural errors in data sets which probably could lead to bad Insights. Improving Accuracy and Relevancy in data could aslo Helps save time and resources allocated to poorly designed and targeted marketing campagins.

I will be using Microsoft excel For the Following Steps:

- Removing Duplicates from the Dataset
- With Excel, by filtering out the blank cells in the longitude and latitude info, I deleted the rows with such missing data.
- Removing Extra unnecessary spaces and proper text formatting in the from_station_name and to_station name Columns Using TRIM and PROPER Functions.
- Ensuring Both the start_time and the End_time Columns had the similar Time format('yyyy-mm-dd hh:mm:ss')

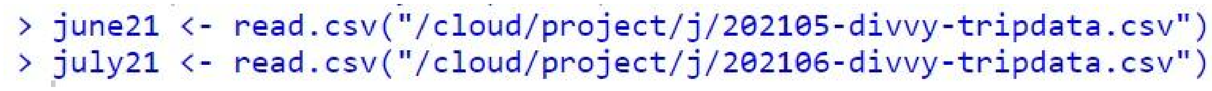
  After completing the above steps, I uploaded the cleaned and rearranged data on RStudio Cloud.

Install and load necessary packages



```
> install.packages("tidyverse")
  install.packages("lubridate")
  install.packages("ggplot2")
  install.packages("dplyr")
  install.packages("geosphere")
```

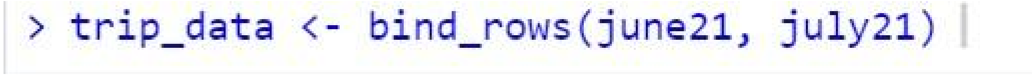
Import data to R studio



```
> june21 <- read.csv("/cloud/project/j/202105-divvy-tripdata.csv")
> july21 <- read.csv("/cloud/project/j/202106-divvy-tripdata.csv")
```

Ensuring that the data has same number of columns and same column names before going forward to merge.

Merge monthly data frames into a large data frame

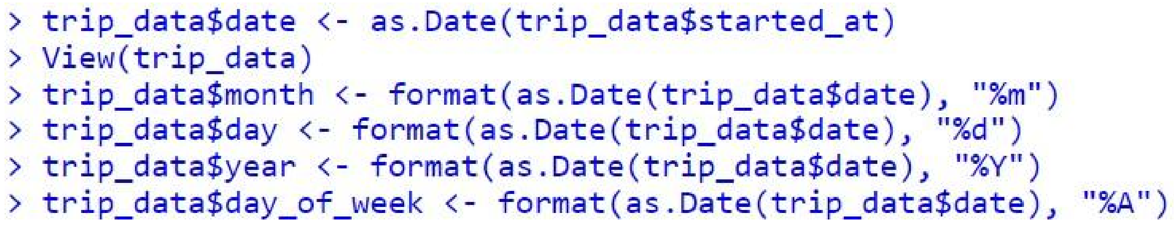


```
> trip_data <- bind_rows(june21, july21)
```

Preparing and Documentation of Data

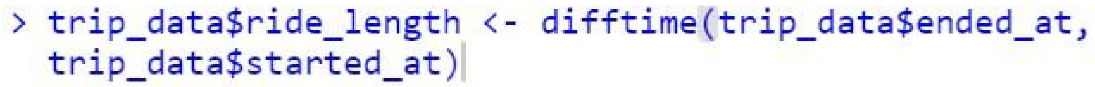
- Check the data for errors.
- Choose your tools.
- Transform the data so you can work with it effectively.
- Document the cleaning process.

Adding columns for date, month, year, day_of_ week into the data frame.



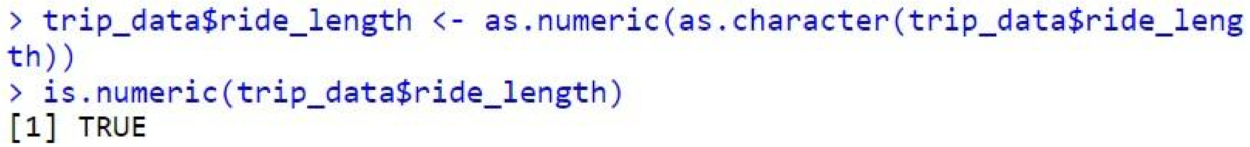
```
> trip_data$date <- as.Date(trip_data$started_at)
> View(trip_data)
> trip_data$month <- format(as.Date(trip_data$date), "%m")
> trip_data$day <- format(as.Date(trip_data$date), "%d")
> trip_data$year <- format(as.Date(trip_data$date), "%Y")
> trip_data$day_of_week <- format(as.Date(trip_data$date),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

Add a ride_length calculation to trip_data



```
> trip_data$ride_length <- difftime(trip_data$ended_at,
  trip_data$started_at)
```

Convert ride_length from Factor to Numeric in order to run calculations



```
> trip_data$ride_length <- as.numeric(as.character(trip_data$ride_leng
th))
> is.numeric(trip_data$ride_length)
[1] TRUE
```

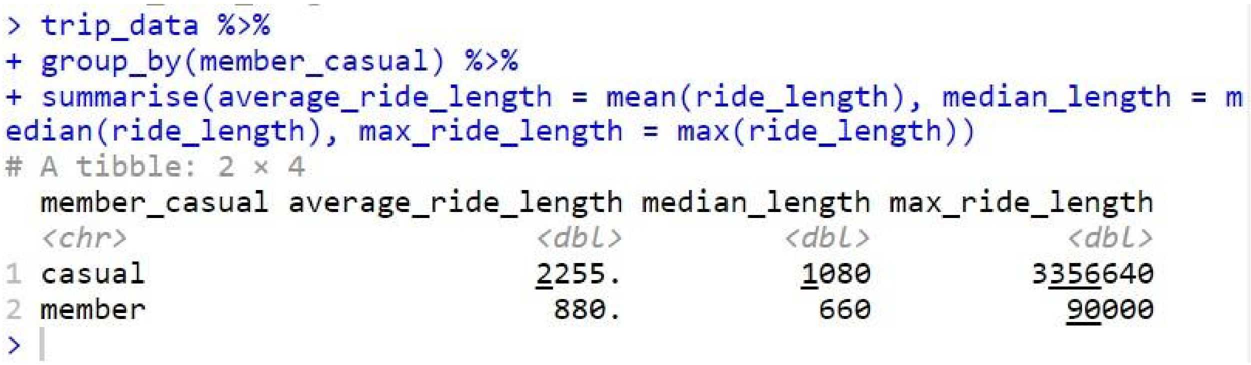
Analyze

Mostly my data had been reorganized in the previous phase with Excel sheet. I only reorgnized the data here as the purpose of generating the reasonable graphs. All the required information are now in one place and ready for exploration.

Conduct descriptive analysis on ride_length

Mean = Straight average (total ride_length/ total rides)

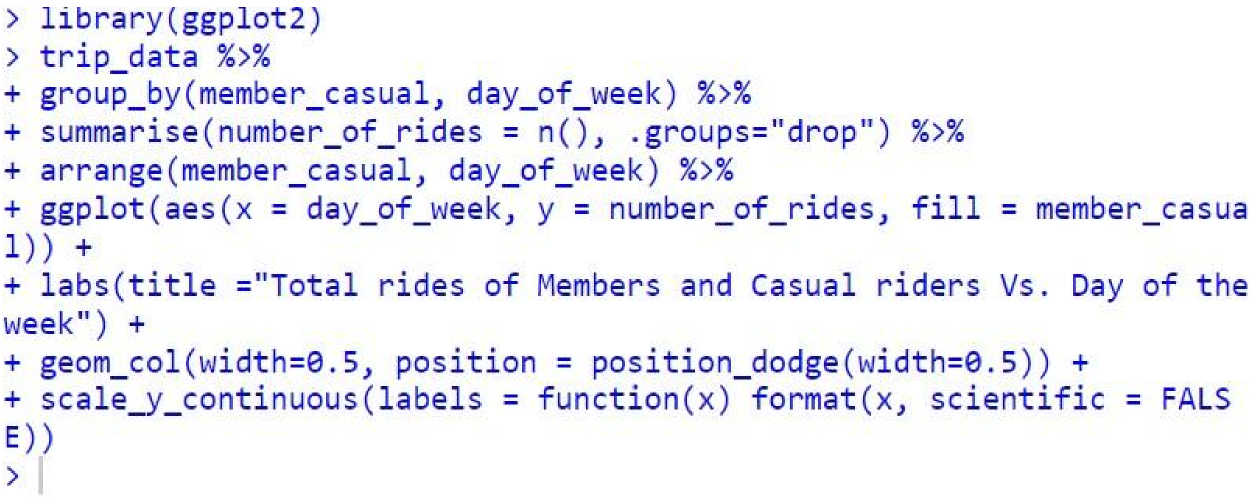
Median = Median Value of ride_length
Max = The Longest Ride



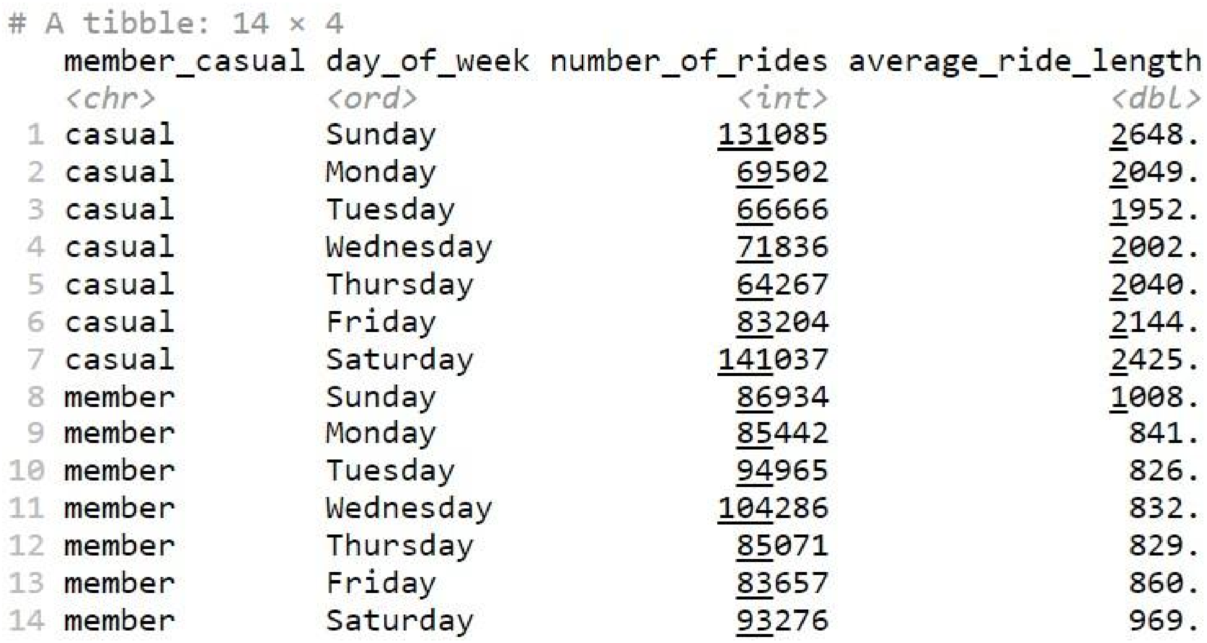
```
> trip_data %>%
+ group_by(member_casual) %>%
+ summarise(average_ride_length = mean(ride_length), median_length = m
edian(ride_length), max_ride_length = max(ride_length))
# A tibble: 2 × 4
  member_casual average_ride_length median_length max_ride_length
  <chr>                       <dbl>         <dbl>           <dbl>
1 casual                      2255.          1080         3356640
2 member                       880.           660           90000
> |
```

The system was crashing simultaneously so I could not make a graphical representation for the average_ride_length by member and casual, Tried again several times until midnight then went to bed.

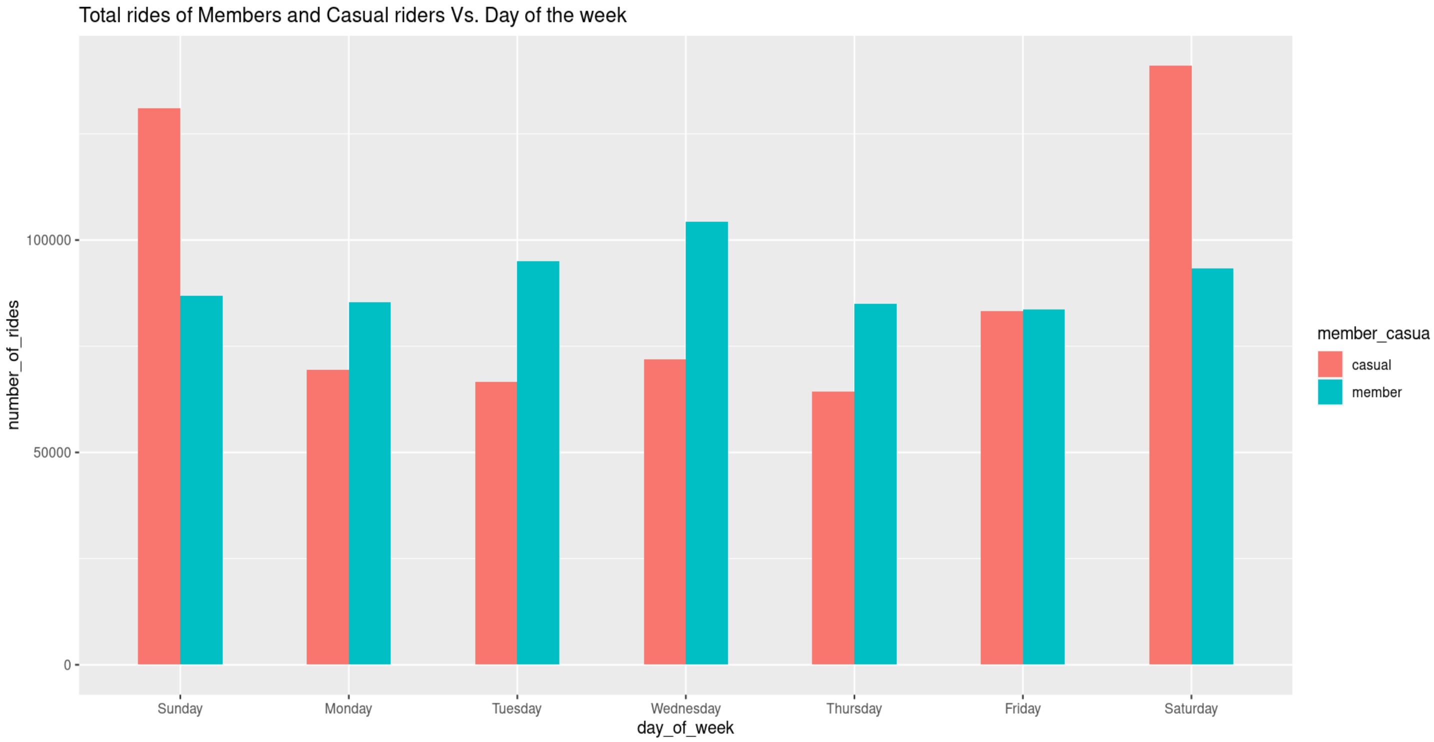
Total rides taken on each day by members vs casual riders



```
> library(ggplot2)
> trip_data %>%
+ group_by(member_casual, day_of_week) %>%
+ summarise(number_of_rides = n(), .groups="drop") %>%
+ arrange(member_casual, day_of_week) %>%
+ ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casua
l)) +
+ labs(title ="Total rides of Members and Casual riders Vs. Day of the
week") +
+ geom_col(width=0.5, position = position_dodge(width=0.5)) +
+ scale_y_continuous(labels = function(x) format(x, scientific = FALS
E))
> |
```



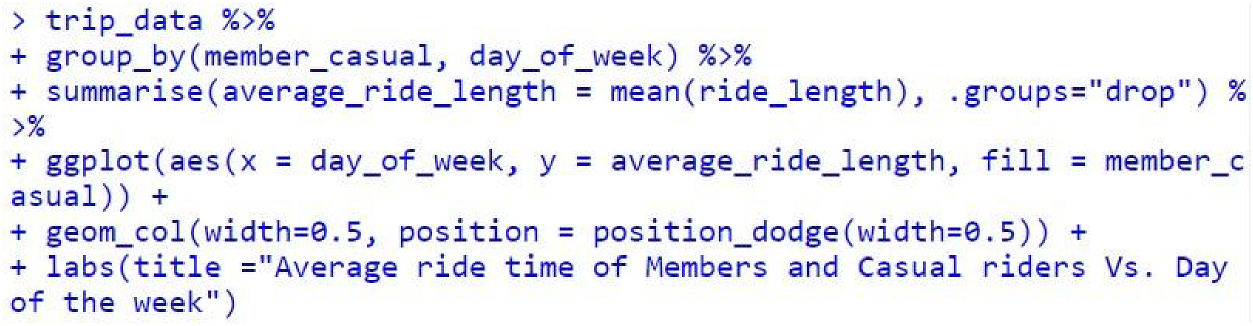
```
# A tibble: 14 × 4
   member_casual day_of_week number_of_rides average_ride_length
   <chr>         <ord>                 <int>               <dbl>
 1 casual        Sunday               131085               2648.
 2 casual        Monday                69502               2049.
 3 casual        Tuesday               66666               1952.
 4 casual        Wednesday             71836               2002.
 5 casual        Thursday              64267               2040.
 6 casual        Friday                83204               2144.
 7 casual        Saturday             141037               2425.
 8 member        Sunday                86934               1008.
 9 member        Monday                85442                841.
10 member        Tuesday               94965                826.
11 member        Wednesday            104286                832.
12 member        Thursday              85071                829.
13 member        Friday                83657                860.
14 member        Saturday              93276                969.
```



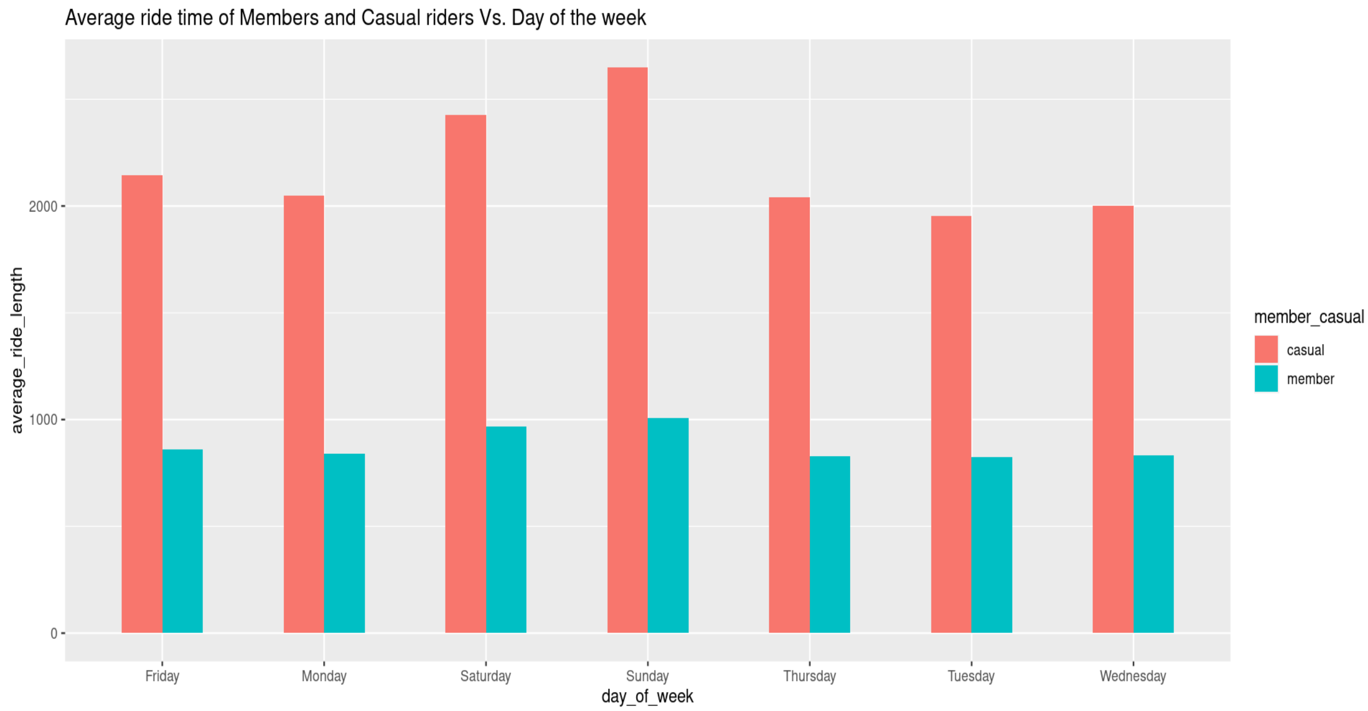
Total rides of Members and Casual riders Vs. Day of the week

From the above graph, it is clear that casual riders have the highest number of rides on the weekend(saturday and sunday) compared to the other days while members are quite consistent but they have the lowest number of rides on the weekend.

Average Time taken on rides each day by members vs casual riders



```
> trip_data %>%
+ group_by(member_casual, day_of_week) %>%
+ summarise(average_ride_length = mean(ride_length), .groups="drop") %
>%
+ ggplot(aes(x = day_of_week, y = average_ride_length, fill = member_c
asual)) +
+ geom_col(width=0.5, position = position_dodge(width=0.5)) +
+ labs(title ="Average ride time of Members and Casual riders Vs. Day
of the week")
```



From above we can see that casual riders ride for a longer time during the week with the highest ride on the weekends while members drive at a consistent pace during the weeks Thus, casual riders have more money to throw on leisure rides.

Conclusion

- Casual riders travel for a longer time period.
- Members ride less on the weekend compared to casual riders.
- Due to limited R studio access I can say by looking at the data that casual riders ride more in hotter months on june and july as compared to the members.

Deliverables

- Encourage member riders to ride on weekends by giving them various Dicounts and incentives or extending their membership by a period of time.
- Host fun biking competitions with prizes at intervals for members on the weekends. Since there are lot of casual riders on weekends,this will also attract them to get a membership.

Resources and Links

RDocumentation

RStudio and