Cyclistic: A Case Study

Yuvraj

Introduction

This case study is a portfolio project made by Yuvraj, to showcase the skills of various BI tools and his analytical prowess. This case study was the capstone project of Google Data Analytics Professional Certificate.

About The Case Study

In 2016, Cyclistic successfully launched a bike-share program in Chicago, which has since expanded to include 5,824 geotracked bicycles across 692 stations. The innovative system allows users to unlock bikes from one station and return them to any other station within the network. Cyclistic's initial marketing strategy focused on creating widespread awareness and appealing to diverse consumer segments. The key to this approach lay in the flexible pricing plans, including :

- single-ride passes - Includes the casual users who subscribe for only one ride.
- full-day passes - These people are also categorized under casual users who subscribe for a single day.
- annual memberships - Annual members subscribe to the yearly pass of the company.

This flexibility in pricing has been instrumental in accommodating various customer preferences and promoting the program's growth.

Characters

- Cyclistic: Bike-share program: 5,800+ bikes, 600 stations. Inclusive options: reclining, tricycles, cargo bikes (8% riders). Usage: 30% commute, majority for leisure. Lily Moreno: Marketing Head
- Cyclistic Marketing Analytics Team:
- Cyclistic Executive Team
- Decision-makers for marketing program approval.

Problem Statement

- Finance analysts favor annual members for higher profits.
- Lily Moreno prioritizes annual memberships for growth.
- Focus on converting existing casual riders to members.
- Casual riders already choose Cyclistic for mobility needs.
- Goal: Develop strategies for casual to annual member conversion.
- Marketing team to analyze differences, motivations, and digital impact.
- Historical bike trip data is to be analysed for trend identification.

Questions to be answered Three questions 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?

- Only 1st question is in the purview of a data analyst.

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 your analysis.
5. Supporting visualizations and key findings.
6. Your top three recommendations based on your analysis.

Business Goal To analyse the historical data of 1 year and find out how the Members and casuals use the bikes differently.

Hypothesis

- Members are the group of people who use the bikes for work related travel.
- Casuals include the group who uses the bikes majorly for leisure purposes.
- People maybe taking the bikes out of station for trips.

Tasks

1. Find out total number of member and casual users.
2. Find out the average ride time of both members and casuals.
3. Find out the average distance traveled by members and casuals respectively.
4. Find out the number of member and casual users who are there daily.
5. Find out the average distance traveled by both types of users daily.
6. Find out the top 50 boarding stations of members and casuals respectively.
7. Find out the top 50 ending stations of both types respectively.
8. Find out the number of people who return bikes after one day.
9. Find the average distance traveled by people who return the bike after one day.
10. Find the average distance traveled by members and casuals every month.
11. Find the number of member and casuals traveling on a monthly basis.
12. Find the number of member and casuals using it at different hours of the day.

Procedure of work on R

1. Importing and making the data ready for performing tasks :

```
library(tidyverse)
tripdata<- list.files(path = "C:/Users/yuvip/Desktop/data analytics/cyclist case study/2_yr_data/2021",
lapply(read.csv) %>%
bind_rows
```

*** Importing dataset:**

*** Inspecting and cleaning the dataset:**

1) having a glimpse at the Dataset.

```r
library(dplyr)
glimpse(tripdata)
```

```
## Rows: 5,595,063
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "2021-01-23 16:14:19", "2021-01-27 18:43:08", "2021~
## $ ended_at           <chr> "2021-01-23 16:24:44", "2021-01-27 18:47:12", "2021~
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor~
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660~
## $ end_station_name   <chr> "", "", "", "", "", "", "", "", "", "Wood St & Augu~
## $ end_station_id     <chr> "", "", "", "", "", "", "", "", "", "657", "13258",~
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4~
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696~
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4~
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

2) checking for unique entries with columns having unique entries

```r
unique(tripdata$rideable_type)
```

```
## [1] "electric_bike" "classic_bike"  "docked_bike"
```

```r
unique(tripdata$member_casual)
```

```
## [1] "member" "casual"
```

3) filtering out the blank and NA cells:

```r
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(dplyr)
sapply(tripdata, function(x) any(x == ""))
```

```
##           ride_id      rideable_type         started_at           ended_at
##             FALSE              FALSE              FALSE              FALSE
## start_station_name  start_station_id   end_station_name     end_station_id
##              TRUE               TRUE               TRUE               TRUE
##         start_lat          start_lng            end_lat            end_lng
##             FALSE              FALSE                 NA                 NA
##     member_casual
##             FALSE
```

Checking if any columns has " " or NA as inputs

```
tripdata<- replace(tripdata, tripdata == "", NA)
```

Replacing the " " cells with NA.

```
tripdata<- tripdata[complete.cases(tripdata), ]
```

*Removing all the NA cells from dataset.

*** Mutating the dataset:**

1) Converting datetime columns into POSIXct format for ease of calculations:

```
tripdata$started_at<- as.POSIXct(tripdata$started_at)
tripdata$ended_at<- as.POSIXct(tripdata$ended_at)
```

2) Calculating the duration

```
tripdata$duration_sec = as.numeric(difftime(tripdata$ended_at,tripdata$started_at,units = "secs"))
# This would make another column which subtracts the ended_at column from started_at column to calculat
```

```
tripdata$duration_hr <- tripdata$duration_sec/(60*24)
# This would convert seconds unit to hours and create a new column.
```

```
tripdata$duration_hr <- round(tripdata$duration_hr,2)
# This will round off the decimals in duration_hr column to 2 digits.
```

- Lets check if the results are clean.

```
min(tripdata$duration_sec)
```

```
## [1] -3354
```

Here we see that the minimum time is negative. This means there are cells which have negative input, which implies that there are some inputs in which end_time < start_time. We must delete all such rows, also those where both the times are same.

```r
tripdata<- tripdata[tripdata$duration_sec> 0, ]
# This would filter out any row in which duration is less than or equal to 0.
```

```r
min(tripdata$duration_sec)
```

```
## [1] 1
```

```r
# Checking again.
```

Now we have a clean dataset to work with.

3) Separating the datetime columns into date and time columns:

```r
tripdata<- tripdata %>% mutate(start_date = as.Date(started_at), start_time = format(started_at, format=
# This would create new columns named start_datee and start_time which has separated values.
```

```r
tripdata<- tripdata %>% mutate(end_date = as.Date(ended_at), end_time = format(ended_at, format= "%H:%M
# Same for ended_at column.
```

Lets check the current status of our dataset

```r
glimpse(tripdata)
```

```
## Rows: 4,588,104
## Columns: 19
## $ ride_id            <chr> "B9F73448DFBE0D45", "457C7F4B5D3DA135", "57C750326F~
## $ rideable_type      <chr> "classic_bike", "electric_bike", "electric_bike", "~
## $ started_at         <dttm> 2021-01-24 19:15:38, 2021-01-23 12:57:38, 2021-01-~
## $ ended_at           <dttm> 2021-01-24 19:22:51, 2021-01-23 13:02:10, 2021-01-~
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor~
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660~
## $ end_station_name   <chr> "Wood St & Augusta Blvd", "California Ave & North A~
## $ end_station_id     <chr> "657", "13258", "657", "657", "657", "KA1504000135"~
## $ start_lat          <dbl> 41.90036, 41.90041, 41.90037, 41.90038, 41.90036, 4~
## $ start_lng          <dbl> -87.69670, -87.69673, -87.69669, -87.69672, -87.696~
## $ end_lat            <dbl> 41.89918, 41.91044, 41.89918, 41.89915, 41.89918, 4~
## $ end_lng            <dbl> -87.67220, -87.69689, -87.67218, -87.67218, -87.672~
## $ member_casual      <chr> "member", "member", "casual", "casual", "casual", "~
## $ duration_sec       <dbl> 433, 272, 587, 537, 609, 1233, 360, 268, 1103, 1025~
## $ duration_hr        <dbl> 0.30, 0.19, 0.41, 0.37, 0.42, 0.86, 0.25, 0.19, 0.7~
## $ start_date         <date> 2021-01-24, 2021-01-23, 2021-01-09, 2021-01-09, 20~
## $ start_time         <chr> "19:15:38", "12:57:38", "15:28:04", "15:28:57", "15~
## $ end_date           <date> 2021-01-24, 2021-01-23, 2021-01-09, 2021-01-09, 20~
## $ end_time           <chr> "19:22:51", "13:02:10", "15:37:51", "15:37:54", "16~
```

We can see that we now have a clean dataset with all relevant columns, we would need to perform the aforementioned tasks.

2. Performing the tasks

```r
library(tidyverse)
sum_count<-tripdata %>% group_by(member_casual) %>% summarise(count = n())
tibble(sum_count)
```

1) Find out total number of member and casual users:

```
## # A tibble: 2 x 2
##   member_casual   count
##   <chr>           <int>
## 1 casual        2048302
## 2 member        2539802
```

```r
Avg_ride_time_hr<- select(tripdata, member_casual, duration_hr)
Avg_ride_time_hr<- Avg_ride_time_hr %>% group_by(member_casual) %>% summarise(avg_ride_time_hr = mean(du
Avg_ride_time_hr<- as.data.frame(Avg_ride_time_hr)
Avg_ride_time_hr$avg_ride_time_hr<- as.numeric(Avg_ride_time_hr$avg_ride_time_hr)
Avg_ride_time_hr$avg_ride_time_hr<-round(Avg_ride_time_hr$avg_ride_time_hr, 2)
```

```r
tibble(Avg_ride_time_hr)
```

2) Find out the average ride time of both members and casuals:

```
## # A tibble: 2 x 2
##   member_casual avg_ride_time_hr
##   <chr>                    <dbl>
## 1 casual                    1.35
## 2 member                    0.55
```

```r
library(geosphere)
calculate_distance <- function(start_lng,start_lat,end_lng,end_lat) {distm(c(start_lng,start_lat), c(end

tripdata$distance<- mapply(calculate_distance,tripdata$start_lng,tripdata$start_lat,tripdata$end_lng,tri
```

```r
average_dist<- tripdata %>% group_by(member_casual) %>% summarise(avg_dist = mean(distance))

average_dist$avg_dist_km<- average_dist$avg_dist/1000
average_dist$avg_dist_km<- round(average_dist$avg_dist_km, 2)
average_dist<- select(average_dist, -avg_dist)
tibble(average_dist)
```

3) Find out the average distance traveled by members and casuals respectively:

```
## # A tibble: 2 x 2
##   member_casual avg_dist_km
##   <chr>               <dbl>
## 1 casual               2.18
## 2 member               2.09
```

```
library(lubridate)
tripdata$start_day<- weekdays(tripdata$start_date)
users_daywise<- select(tripdata, member_casual, start_day)
users_daywise<- table(users_daywise)
users_daywise<-as.data.frame(users_daywise)
users_daywise<- users_daywise %>% mutate(Avg_user_per_day = Freq/52)
# we divided the freq by 52 as there are approx 52 weeks per year.
users_daywise$Avg_user_per_day<-round(users_daywise$Avg_user_per_day,0)
tibble(users_daywise)
```

4) Find out the number of members and casuals users are there on a daily basis:

```
## # A tibble: 14 x 4
##    member_casual start_day   Freq Avg_user_per_day
##    <fct>         <fct>      <int>            <dbl>
##  1 casual        Friday    311694             5994
##  2 member        Friday    375842             7228
##  3 casual        Monday    224416             4316
##  4 member        Monday    345444             6643
##  5 casual        Saturday  477035             9174
##  6 member        Saturday  358883             6902
##  7 casual        Sunday    372195             7158
##  8 member        Sunday    296879             5709
##  9 casual        Thursday  228860             4401
## 10 member        Thursday  375841             7228
## 11 casual        Tuesday   215033             4135
## 12 member        Tuesday   388664             7474
## 13 casual        Wednesday 219069             4213
## 14 member        Wednesday 398249             7659
```

```
average_dist_daywise<- select(tripdata, member_casual,start_day,distance)
average_dist_daywise<- average_dist_daywise %>%
  group_by(member_casual, start_day) %>% summarise(avg_distance = mean(distance), .groups = "drop")

average_dist_daywise$avg_distance_km<- average_dist_daywise$avg_distance/1000
average_dist_daywise$avg_distance_km<- round(average_dist_daywise$avg_distance_km, 2)
average_dist_daywise<- select(average_dist_daywise, -avg_distance)
tibble(average_dist_daywise)
```

5) Find out the average distance traveled by both type of users on a daily basis respectively:

```
## # A tibble: 14 x 3
##    member_casual start_day avg_distance_km
##    <chr>         <chr>               <dbl>
##  1 casual        Friday               2.17
##  2 casual        Monday               2.07
##  3 casual        Saturday             2.28
##  4 casual        Sunday               2.25
##  5 casual        Thursday             2.13
##  6 casual        Tuesday              2.1
##  7 casual        Wednesday            2.12
##  8 member        Friday               2.05
##  9 member        Monday               2.04
## 10 member        Saturday             2.19
## 11 member        Sunday               2.19
## 12 member        Thursday             2.05
## 13 member        Tuesday              2.06
## 14 member        Wednesday            2.07
```

```r
boarding_station<- select(tripdata, member_casual, start_station_name)
boarding_station<- table(boarding_station)
boarding_station<- as.data.frame(boarding_station)
boarding_station<- boarding_station %>% arrange(desc(Freq))
boarding_station<- head(boarding_station, 50)
tibble(boarding_station)
```

6) Find out the top 50 boarding stations of members and casuals respectively:

```
## # A tibble: 50 x 3
##    member_casual start_station_name         Freq
##    <fct>         <fct>                     <int>
##  1 casual        Streeter Dr & Grand Ave   64446
##  2 casual        Millennium Park           32185
##  3 casual        Michigan Ave & Oak St     28661
##  4 member        Clark St & Elm St         23900
##  5 member        Wells St & Concord Ln     22760
##  6 member        Kingsbury St & Kinzie St  22668
##  7 casual        Shedd Aquarium            22544
##  8 casual        Theater on the Lake       20620
##  9 member        Wells St & Elm St         20245
## 10 casual        Lake Shore Dr & Monroe St 19085
## # i 40 more rows
```

```r
ending_station<- select(tripdata, member_casual, end_station_name)
ending_station<- table(ending_station)
ending_station<- as.data.frame(ending_station)
ending_station<- ending_station %>% arrange(desc(Freq))
ending_station<- head(ending_station, 50)
tibble(ending_station)
```

7) Find out the top 50 ending stations of both types respectively.

```
## # A tibble: 50 x 3
##    member_casual end_station_name          Freq
##    <fct>         <fct>                    <int>
##  1 casual        Streeter Dr & Grand Ave  67524
##  2 casual        Millennium Park          33744
##  3 casual        Michigan Ave & Oak St    30364
##  4 member        Clark St & Elm St        23971
##  5 member        Wells St & Concord Ln    23407
##  6 member        Kingsbury St & Kinzie St 22853
##  7 casual        Theater on the Lake      22307
##  8 casual        Shedd Aquarium           21158
##  9 member        Wells St & Elm St        20799
## 10 member        Dearborn St & Erie St    19317
## # i 40 more rows
```

```r
library(dplyr)
tripdata$end_day<- weekdays(tripdata$end_date)
tripdata$not_same_day <- tripdata$start_day != tripdata$end_day
outstation_travellers<- select(tripdata,not_same_day,member_casual)
outstation_table<- table(outstation_travellers)
outstation_table<- as.data.frame((outstation_table))
tibble(outstation_table)
```

8) Find out the number of people who return bikes after one day:

```
## # A tibble: 4 x 3
##   not_same_day member_casual    Freq
##   <fct>        <fct>           <int>
## 1 FALSE        casual        2041039
## 2 TRUE         casual           7263
## 3 FALSE        member        2536066
## 4 TRUE         member           3736
```

```r
outstation_travellers2<- select(tripdata,not_same_day,member_casual,distance)
outstation_travellers2<- outstation_travellers2 %>% mutate(distance_km = distance/1000)
outstation_travellers2$distance_km <- round(outstation_travellers2$distance_km, 2)
outstation_travellers2<- select(outstation_travellers2, -distance)
outstation_table2<-table(outstation_travellers2)
outstation_table2<-as.data.frame(outstation_table2)

outstation_table2_grouped<- outstation_travellers2 %>% mutate(distance_km = as.numeric(distance_km)) %>%

outstation_table_merged<- merge(outstation_table, outstation_table2_grouped, by = c("member_casual","not

outstation_table_merged<- as.data.frame(outstation_table_merged)
outstation_table_merged$avg_dist_km <- outstation_table_merged$total_dist_km /outstation_table_merged$F
```

```r
outstation_table_merged$avg_dist_km<- round(outstation_table_merged$avg_dist_km, 2)
tibble(outstation_table_merged)
```

9) Find the average distance traveled by people who return the bike after one day:

```
## # A tibble: 4 x 5
##   member_casual not_same_day    Freq total_dist_km avg_dist_km
##   <fct>         <fct>          <int>         <dbl>       <dbl>
## 1 casual        FALSE        2041039      4453857.        2.18
## 2 casual        TRUE            7263        17962.        2.47
## 3 member        FALSE        2536066      5293338.        2.09
## 4 member        TRUE            3736        13410.        3.59
```

```r
library(lubridate)
tripdata$month <- month(tripdata$start_date)
month_df<- select(tripdata, member_casual, month, distance)
month_df$month<- month.name[month_df$month]
month_df$distance_km<- month_df$distance/1000
month_df<- select(month_df, -distance)
month_df$distance_km<- round(month_df$distance_km, 2)
month_df<- month_df %>% group_by(member_casual, month) %>% summarise(avg_dist_km = mean(distance_km), .g
month_df$avg_dist_km<- round(month_df$avg_dist_km, 2)
tibble(month_df)
```

10) Average distance travelled by both type of customers on a monthly basis:

```
## # A tibble: 24 x 3
##    member_casual month    avg_dist_km
##    <chr>         <chr>          <dbl>
##  1 casual        April           2.05
##  2 casual        August          2.25
##  3 casual        December        1.93
##  4 casual        February        2.02
##  5 casual        January         1.92
##  6 casual        July            2.22
##  7 casual        June            2.19
##  8 casual        March           2.05
##  9 casual        May             2.14
## 10 casual        November        2.01
## # i 14 more rows
```

```r
month_df2<- select(tripdata, member_casual, month)
month_df2$month<- month.name[month_df2$month]
month_df2<- table(month_df2)
month_df2<- as.data.frame(month_df2)
tibble(month_df2)
```

11) Find the number of member and casuals traveling on a monthly basis:

```
## # A tibble: 24 x 3
##    member_casual month      Freq
##    <fct>         <fct>     <int>
##  1 casual        April    120735
##  2 member        April    177972
##  3 casual        August   339902
##  4 member        August   332345
##  5 casual        December  45098
##  6 member        December 131257
##  7 casual        February   8617
##  8 member        February  34409
##  9 casual        January   14628
## 10 member        January   68748
## # i 14 more rows
```

```r
time_df<- select(tripdata, started_at, member_casual)
time_df$started_at <- format(as.POSIXlt(time_df$started_at), format = "%H")
time_df<- table(time_df)
time_df<- as.data.frame(time_df)
time_df<- time_df %>% mutate(Avg_people_per_hour = Freq/365)
# We divided time by 365 as a particular hour will repeat for 365 times a year.
time_df$Avg_people_per_hour = round(time_df$Avg_people_per_hour, 0)
tibble(time_df)
```

12) Number of people using bikes at different hours of day:

```
## # A tibble: 48 x 4
##    started_at member_casual  Freq Avg_people_per_hour
##    <fct>      <fct>         <int>               <dbl>
##  1 00         casual        42321                 116
##  2 01         casual        30667                  84
##  3 02         casual        19579                  54
##  4 03         casual        10220                  28
##  5 04         casual         6688                  18
##  6 05         casual         8823                  24
##  7 06         casual        19406                  53
##  8 07         casual        36024                  99
##  9 08         casual        49678                 136
## 10 09         casual        60885                 167
## # i 38 more rows
```

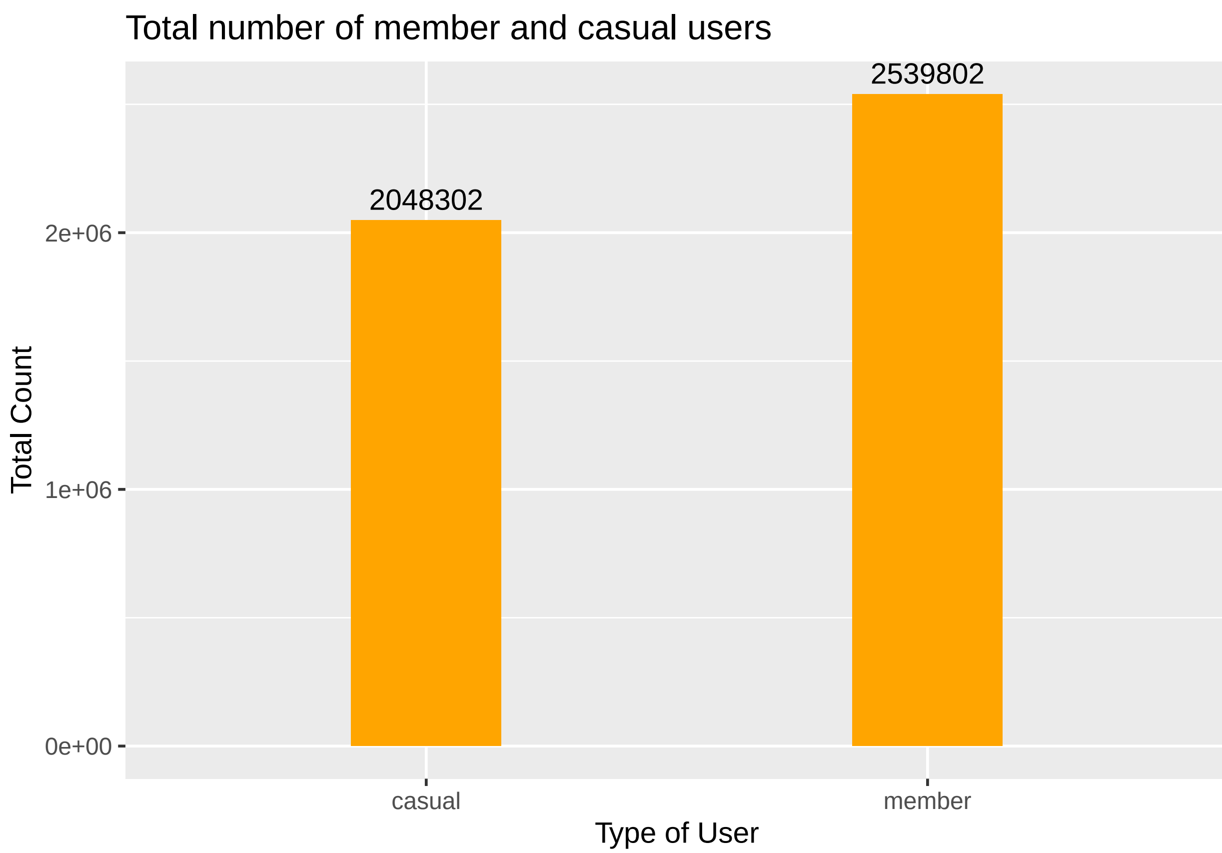
We have successfully completed all the tasks, and now have all the relevant figures needed to make an accurate analysis.

Visualisation and Analysis

- Now it is the time to make graphs from the tables made and analyse them to get useful insights, and determine the best course of action to accomplish the business goal.

1. Total number of member and casual users:

```
library(ggplot2)
ggplot(sum_count, aes(x= member_casual, y = count))+
geom_bar(stat = "identity", fill = "orange", width = 0.3)+
  geom_text(aes(label = count), vjust = -0.5, color = "black")+
labs(title = "Total number of member and casual users",
     x = "Type of User",
     y = "Total Count")
```



Insight: We can clearly see that the annual members are more than the casuals. Although, the number of casual users are also fairly high. If these turn into members, the profits of the firm can skyrocket.

- If we calculate the percentage relation between both we notice that:

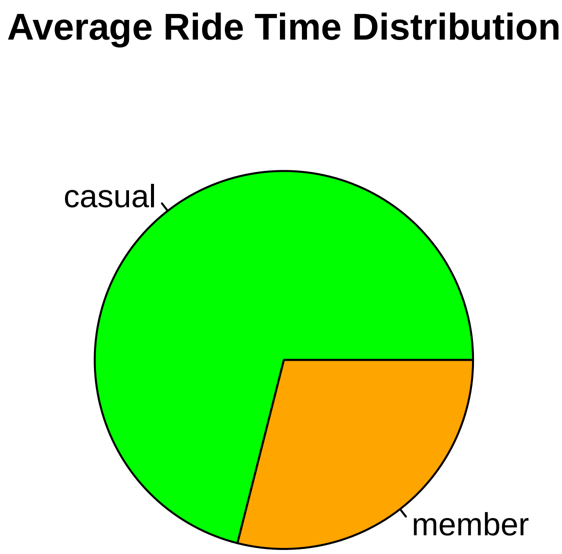
```
print((2048302/2589302)*100)
```

```
## [1] 79.10634
```

- Casuals are approx 20% less than members.

2. Average ride time of both members and casuals:

```
pie(Avg_ride_time_hr$avg_ride_time_hr, labels = Avg_ride_time_hr$member_casual, col = c("green", "orang
```

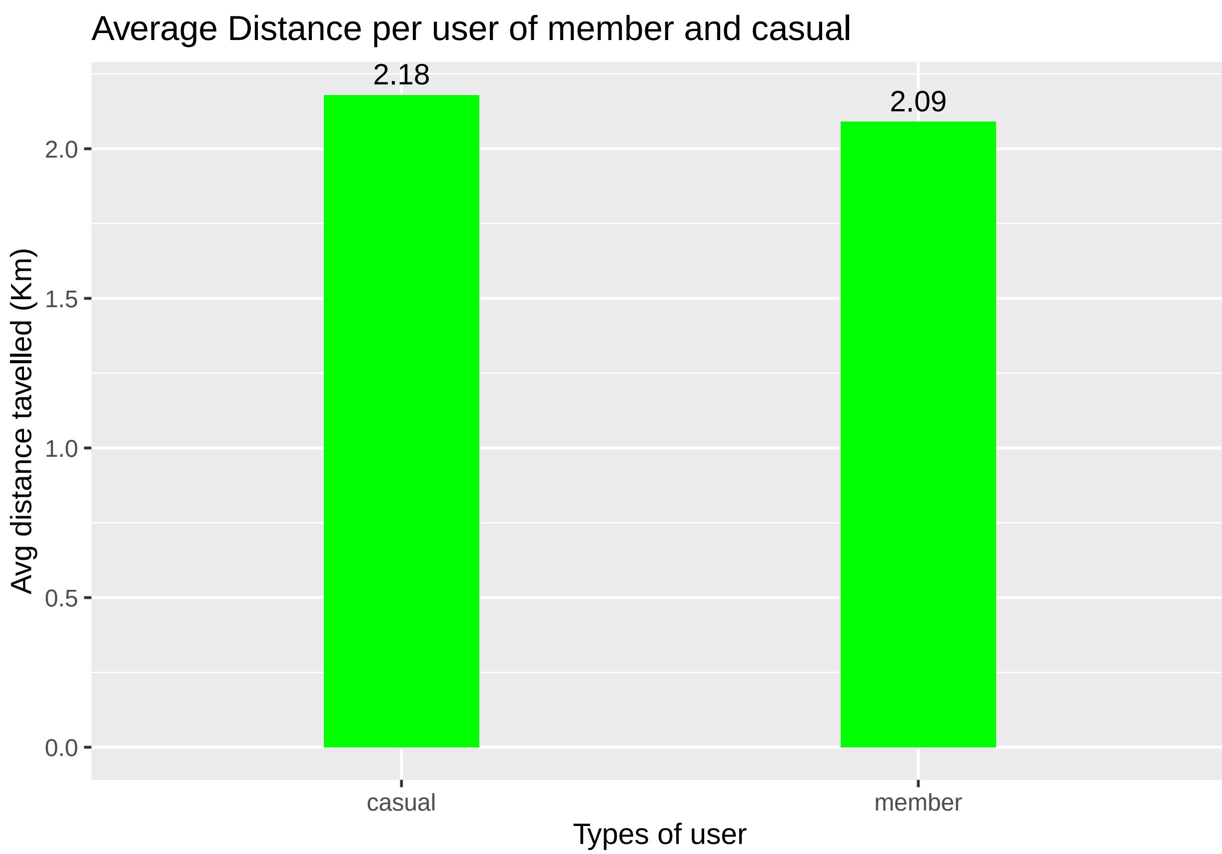


Insight: Although we saw previously that the members are more in number than the casuals, but here we can see that the average ride time per member is almost one fourth of that of a casual.

Analysis This staggering difference was something called for and is in line with the hypothesis. Casuals having such high average ride time despite being less in number an even travelling the same distance may be because, the casuals do not rent bikes on a regular basis but when they do, they cover great distance. Meanwhile, the members, although, don't cover much distance per ride, but due to regular to and fro travel to office and home, increases their average ride time.

3. Average distance traveled by members and casuals respectively:

```
ggplot(average_dist, aes(x= member_casual, y = avg_dist_km))+
  geom_bar(stat = "identity", fill = "green", width = 0.3)+
  geom_text(aes(label = avg_dist_km), vjust = -0.5, color = "black")+
  labs(title = "Average Distance per user of member and casual",
       x = "Types of user",
       y = "Avg distance tavelled (Km)")
```



Insight We can see that the average distance traveled by a casual is also more than that of a member. Although, the difference is very less. We can say that they are almost equal.

- Before analysis of this, we should check the total distance covered by both respectively:

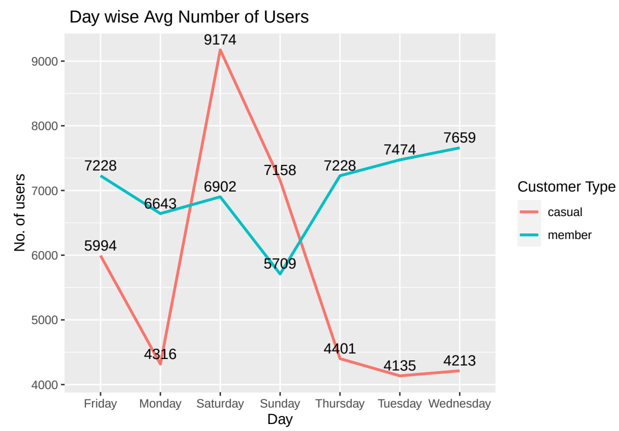
```
d<- select(tripdata, distance, member_casual)
d<- d %>% group_by(member_casual) %>%
  summarise(total_distance = sum(distance), .groups = "drop")
tibble(d)
```

```
## # A tibble: 2 x 2
##   member_casual total_distance
##   <chr>                  <dbl>
## 1 casual           4471885284.
## 2 member           5306736650.
```

- Casuals' total distance is approx. 85% of that of members'.

Analysis Although casuals are lesser than members, and also they ride less total distance, but if relatively considering their population, they ride much more, hence, making their average distance covered much higher than members.

```r
ggplot(users_daywise, aes(x= start_day, y = Avg_user_per_day, color = member_casual, group = member_casu
geom_line(linewidth = 1, aes(group = member_casual))+
  geom_text(aes(label = Avg_user_per_day), vjust = -0.5, color = "black")+
labs(title = " Day wise Avg Number of Users",
x = "Day",
y = "No. of users",
color = "Customer Type")
```



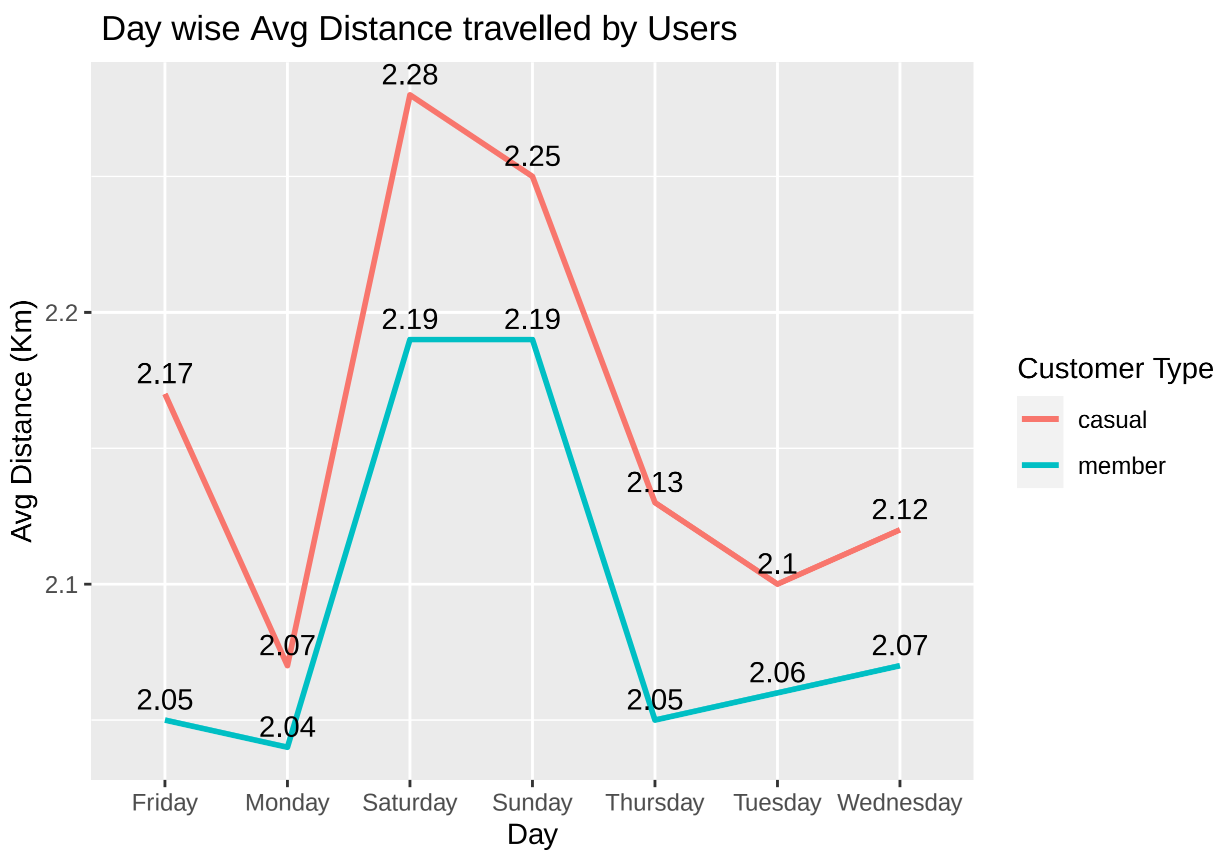
4. Number of members and casuals users on a daily basis:

Insight It can be seen that in weekdays the number of members is significantly more than the casuals. Also, there is a drop in the number of member users during the weekend. But, when we look at the casuals graph the pattern is exactly opposite. Here we can see that the number of riders significantly increase during the weekend and also passes that of the members.

Analysis It is not that there is only an increase in the number of casual user, this figure is skyrocketing and also passes the members, who are greater than the prior.. This is in line with our hypothesis. Weekends are the days of holidays, and working population (who are the majority of the members), do not go office, hence, the graph drops. On the other hand, the casual group utilizes the weekend for leisure activities and hence the graph rises.

5. Average distance traveled by both type of users on a daily basis:

```r
ggplot(average_dist_daywise, aes(x= start_day, y = avg_distance_km, color = member_casual, group = membe
geom_line(linewidth = 1, aes(group = member_casual))+
  geom_text(aes(label = avg_distance_km), vjust = -0.5, color = "black")+
labs(title = " Day wise Avg Distance travelled by Users",
x = "Day",
y = "Avg Distance (Km)",
color = "Customer Type")
```

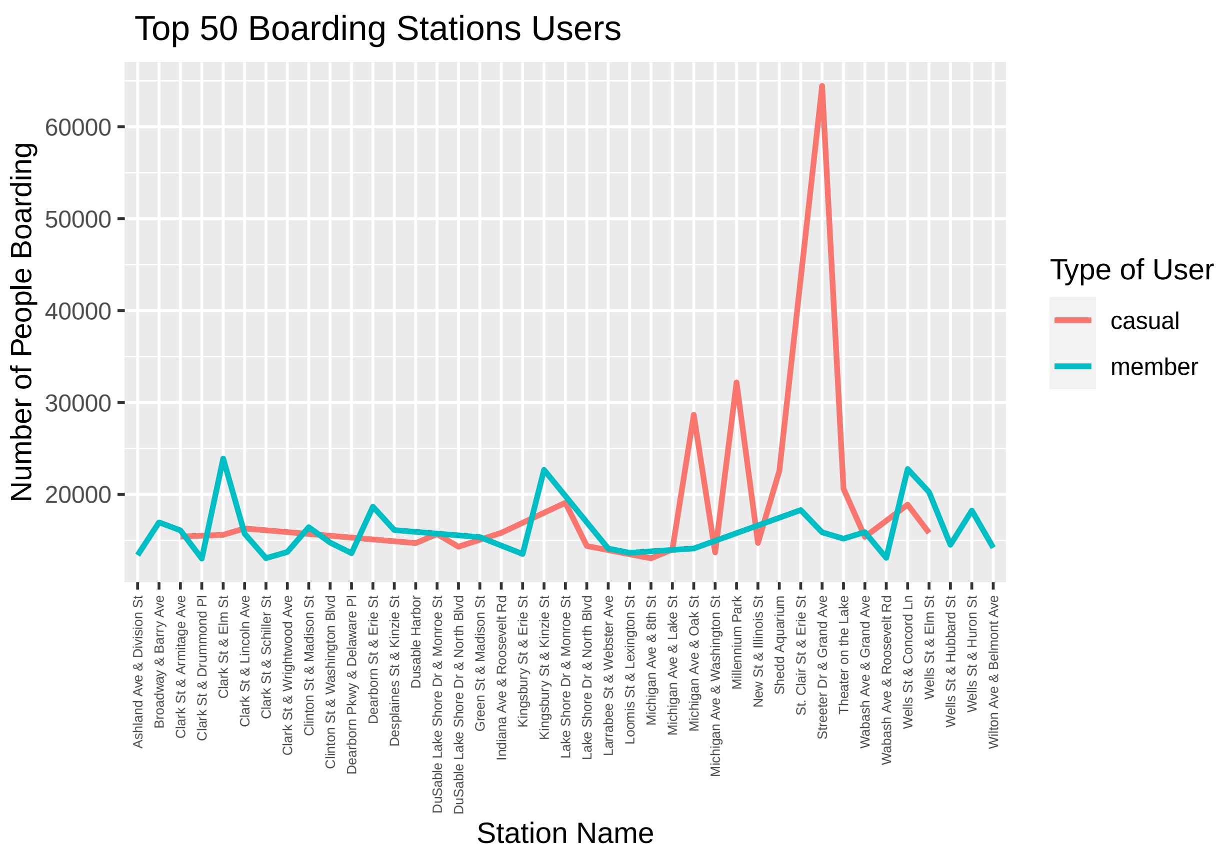


Insight The average distance traveled by casuals is more than that of members. Also this even increases during weekends. Although, average distance traveled by members is less overall, but their graph also increases during the weekend.

Analysis It was expected that the avg dist of members would be lesser than that of casuals, but a new pattern is observed that the avg dist traveled by members also increase during the weekend, although the number of members travelling during the weekend was decreasing. This implies that, there are people in members group also who like leisure activities, are doing so in such amount, that even though their numbers are less, still they increase the overall average distance traveled.

6. Top 50 boarding stations of members and casuals:

```
ggplot(boarding_station, aes(x= start_station_name, y = Freq, color = member_casual, group = member_casu
geom_line(linewidth = 1, aes(group = member_casual))+
  labs(title = " Top 50 Boarding Stations Users",
x = "Station Name",
y = "Number of People Boarding",
color = "Type of User")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 5))
```



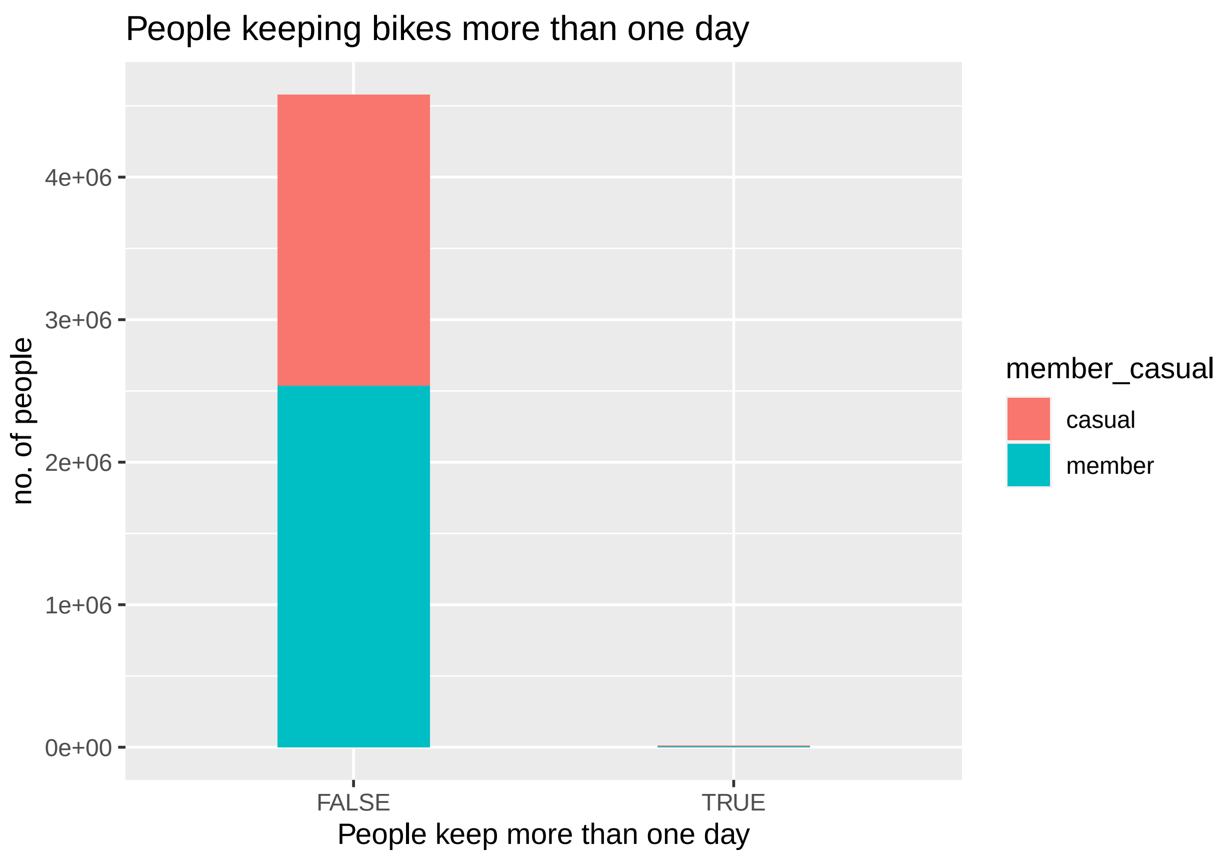
Insights These are the 50 stations from where maximum number member and casuals board the bikes. These are the top 3 stations from where maximum number of casuals board- Streeter Dr & Grand Ave, Millennium Park, Michigan Ave & Oak.

Analysis Our business goal is to convert these casuals into members. Also provided that a good marketing strategy has to be developed, it is important that we consider these stations in a priority list in locations of advertisements.

The same can be done for exit stations. Both the lists gives us 100 stations of Chicago from where majority of casuals board and exit.

7. Number of people who return bikes after one day:

```
ggplot(outstation_table_merged, aes(x= not_same_day, y = Freq, fill = member_casual))+
  geom_bar(position = "stack", stat = "identity", width = 0.4)+
  labs(title = "People keeping bikes more than one day",
       x= "People keep more than one day",
       y = "no. of people",
       color = "type of customer")
```



Insights It can be seen that compared to people who return the bikes in one day, the number of people who return the bikes after a day is negligible. Still, there are people who keep it for more than one day.

- These many people as solved earlier keep bikes for more than a day :

```
tibble(outstation_table_merged)
```

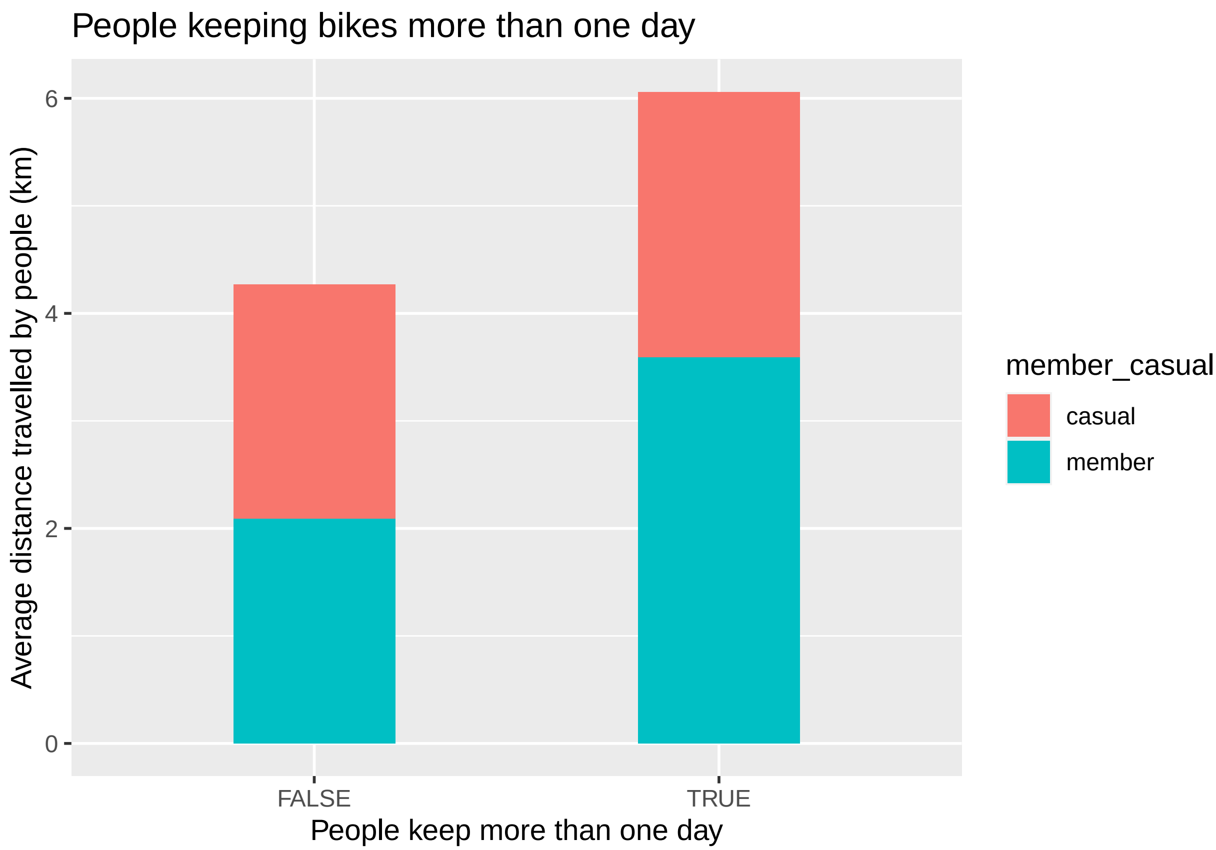
```
## # A tibble: 4 x 5
##   member_casual not_same_day    Freq total_dist_km avg_dist_km
##   <fct>         <fct>          <int>         <dbl>       <dbl>
## 1 casual        FALSE        2041039      4453857.        2.18
## 2 casual        TRUE            7263        17962.        2.47
## 3 member        FALSE        2536066      5293338.        2.09
## 4 member        TRUE            3736        13410.        3.59
```

*We can see that the number of casuals keeping the bikes for multiple days is approximately double to that of the members.

8. Average Distance traveled by people returning it after one day.

```
ggplot(outstation_table_merged, aes(x= not_same_day, y = avg_dist_km, fill = member_casual))+
  geom_bar(position = "stack", stat = "identity", width = 0.4)+
```

```r
  labs(title = "People keeping bikes more than one day",
       x= "People keep more than one day",
       y = "Average distance travelled by people (km)",
       color = "type of user")
```



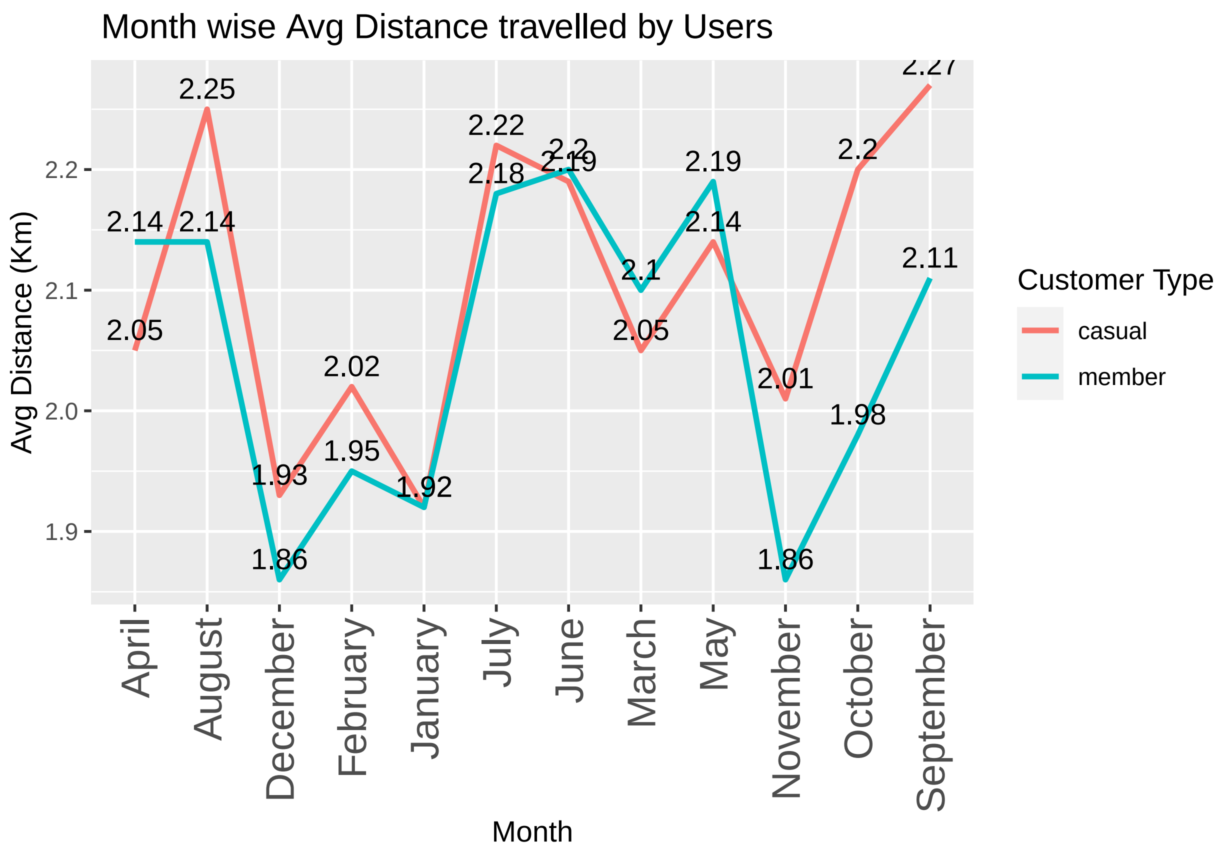
Insights Although, we previously saw that the the number of people keeping the bikes for more than one day is negligible, yet the avg distance traveled by them is significantly high.

Analysis The purpose of analyzing this factor is that, if the people are keeping it for more than one day, hey may be taking it for trips. We can see that the average distance traveled by people not returning their bike on the same day is fairly higher than those who do. Also, there are a fairly good number of casuals who do so. Hence, we can say that casuals are also using bikes for going on trips, and exploring the city and tourist spots.

9. Average distance traveled by members and casuals on a monthly basis

```r
ggplot(month_df, aes(x= month, y = avg_dist_km, color = member_casual, group = member_casual))+
geom_line(linewidth = 1, aes(group = member_casual))+
  geom_text(aes(label = avg_dist_km), vjust = -0.5, color = "black")+
labs(title = " Month wise Avg Distance travelled by Users",
x = "Month",
y = "Avg Distance (Km)",
```

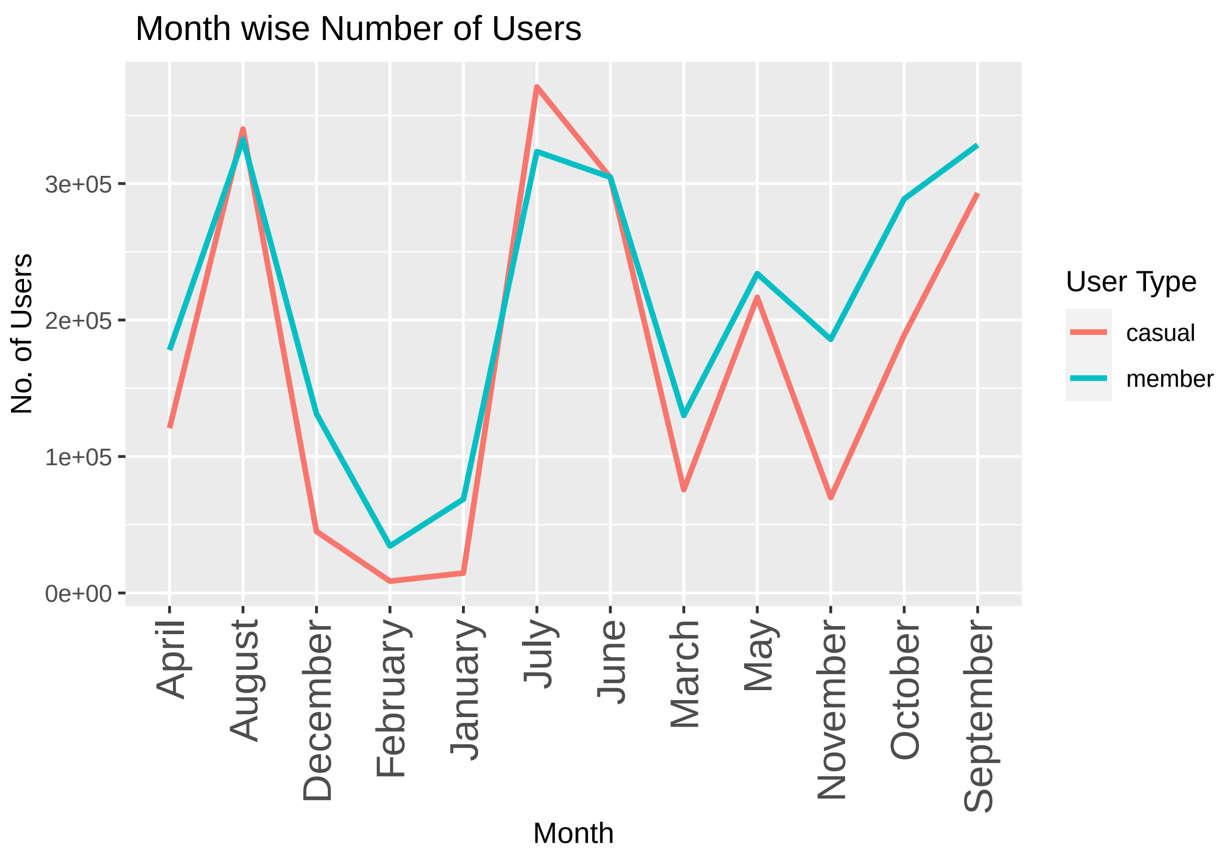
```
color = "Customer Type")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 15))
```



Insight The monthly distance traveled by members and casuals appear to be in the same pattern. We can see that casuals are always above than members, except in the months of March, April and May.

10. Number of member and casuals traveling on a monthly basis

```
ggplot(month_df2, aes(x= month, y = Freq, color = member_casual, group = member_casual))+
geom_line(linewidth = 1, aes(group = member_casual))+
labs(title = " Month wise Number of Users",
x = "Month",
y = "No. of Users",
color = "User Type")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 15))
```

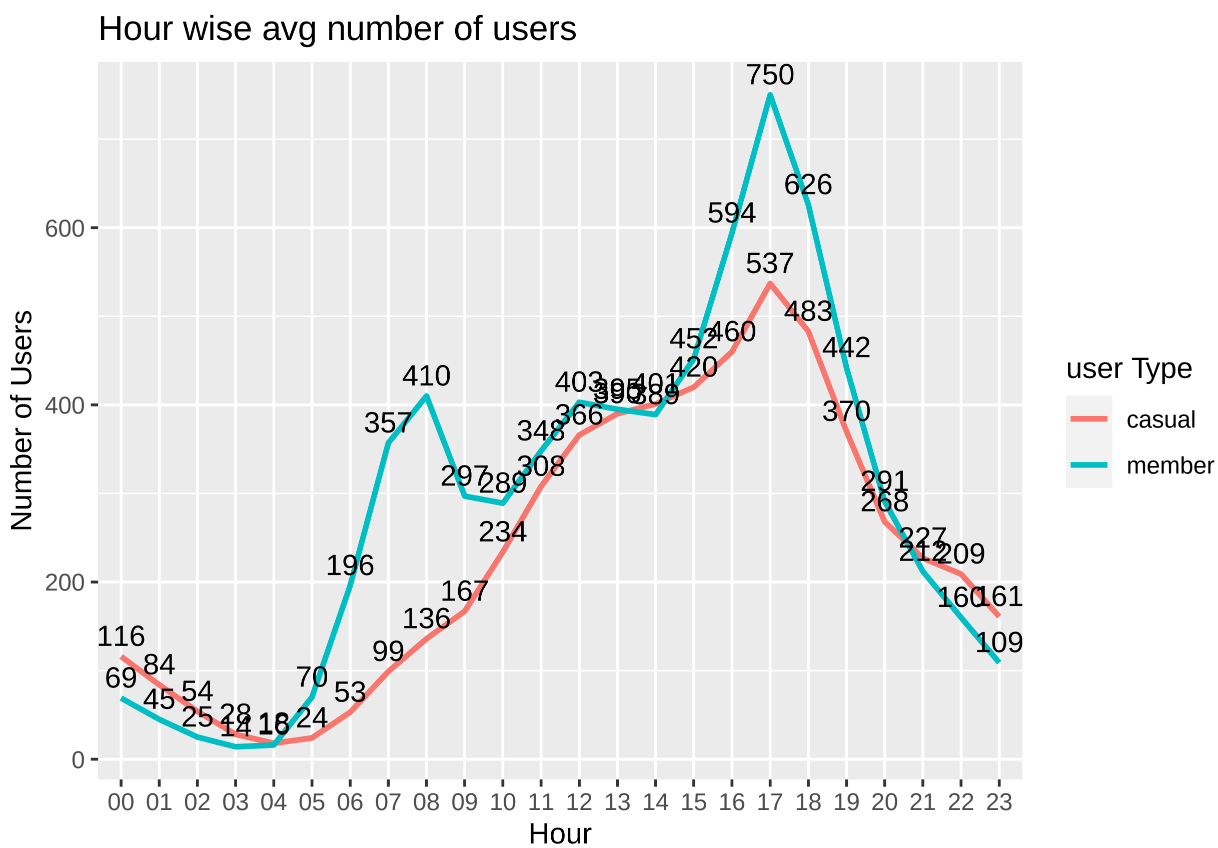


Insight As with previous graph, here also,the user pattern is same in both groups. Here the least users are in February, but most users are in July.

Analysis We can see that in months of vacations like July and September both average distance and number of users increases. It can be seen that in the month of April, May the number of users rises yet the average distance falls. This may be because the of starting of new academic year and opening of offices, people go to offices and use for leisure less. Hence, the casual users' avg distance falls even more than usual.

11. People using bikes at different hours of the day

```
ggplot(time_df, aes(x= started_at, y = Avg_people_per_hour, color = member_casual, group = member_casual
geom_line(linewidth = 1, aes(group = member_casual))+
labs(title = "Hour wise avg number of users",
x = "Hour",
y = "Number of Users",
color = "user Type")+
geom_text(aes(label = Avg_people_per_hour), vjust = -0.5, color = "black")
```



Insight The members as expected increase during the office going, lunch and office end hours. A similar trend can be observed with casuals also.

Analysis Looking at the casual trends and comparing it with members, we can say that there are also a good number of office going people in this group also. Since, there is a spike in members at early morning, we can say that even school/ college going people use bikes for their commute.

This morning spike is not seen in the casuals group, rather there is a steep growth. Although during the returning hours, the casuals trends mimics that of the members. We can safely say that the college going people use bikes for their commutes.

Also at night the number of casuals is greater than that of members. It may be said that, night shift workers may be using the bikes for their commutes.

Final Analysis

The analysis of all the above visualizations can be summarized as follows:-

Our hypothesis that members are majorly the working population is true. Although members also use bikes for leisure activities, the percentage of such members is less.

Whereas it would not be correct to say that casuals are entirely the people using the bikes for leisure activities. According to the analysis, people also use bikes for work commutes, and other than that, those people who use it entirely for leisure activities may also be in the working population, but use other modes of travel for work commutes.

A good diversity in the usage patterns of casuals can be seen: 1. A good number of casuals are college students, who use the bikes for their commute to colleges. 2. People going on trips and sightseeing are also the ones in this group. 3. Night shift workers are also using the bikes for their commute.

People who use the bikes for other leisure activities in their free time and on holidays form the majority of this group.

Recommendation

1.Tapping the majority of people using bikes for leisure activities:

- There can be a discount for members during holidays and weekends when the number of casuals is in the majority. This would persuade them to take memberships.
- A dynamic pricing system must be made based on an hourly basis. At 1 PM both the members and casuals are high and equal in number. That may act as sweet spot to provide special pricing for members. It would persuade the casuals to take membership as it the hour when they also need bikes. If we provide it at 5 PM when the casuals peak, then the number of Members is even higher and we would have to give discounts to even existing members, draining the profit.
- Advertisement signifying how using Cyclistic bikes enables the person to explore the city and go on trips more cheaply and easily.

Tapping the college students:

- There can be special discounts for students upon showing their college IDs on membership.
- Use of social media and youth-relevant influencer marketing can render good results.
- Use of hoardings near college areas and at stations with higher casual footfalls. Following is the list of top 10 stations with highest casual footfall.

```r
library(tidyverse)
data.frame(Top_10_stations = c("Streeter Dr & Grand Ave",
"Millennium Park",
"Michigan Ave & Oak St",
"Theater on the Lake",
"Shedd Aquarium",
"Wells St & Concord Ln",
"Lake Shore Dr & Monroe St",
"Lake Shore Dr & North Blvd",
"DuSable Lake Shore Dr & North Blvd",
"Wabash Ave & Grand Ave"))
```

```
##                          Top_10_stations
## 1            Streeter Dr & Grand Ave
## 2                    Millennium Park
## 3              Michigan Ave & Oak St
## 4                Theater on the Lake
## 5                     Shedd Aquarium
## 6              Wells St & Concord Ln
## 7          Lake Shore Dr & Monroe St
## 8         Lake Shore Dr & North Blvd
## 9  DuSable Lake Shore Dr & North Blvd
## 10            Wabash Ave & Grand Ave
```

Tapping the casual working segment and night time workers:

- Special advertisements can be made that indicate how using Cyclistic bikes eases the commute of people to their offices, and is cheaper than other modes of travel.
- Special nighttime discounts and membership facilities can be drawn upon for the nighttime workers.

Conclusion and Sources

- All the tasks were accomplished, and due analysis was given. It is hoped that a good marketing strategy considering all the stated points would help the company to accomplish the Business Goal of converting casual users to annual members, hence, boosting the profits.
- Source of dataset: This dataset was provided along with the problem statement of the capstone project. https://divvy-tripdata.s3.amazonaws.com/index.html

THANK YOU