

# Google Capstone

Zazzini Marco

2023-10-20

*# Load data from March 2023 to August 2023*

```
trip23_Mar <- read.csv("C:/Users/Utente/Downloads/Downloads/divvy/202302-divvy-tripdata.csv")
trip23_Apr <- read.csv("C:/Users/Utente/Downloads/Downloads/divvy/202303-divvy-tripdata.csv")
trip23_May <- read.csv("C:/Users/Utente/Downloads/Downloads/divvy/202304-divvy-tripdata.csv")
trip23_Jun <- read.csv("C:/Users/Utente/Downloads/Downloads/divvy/202305-divvy-tripdata.csv")
trip23_Jul <- read.csv("C:/Users/Utente/Downloads/Downloads/divvy/202306-divvy-tripdata.csv")
trip23_Aug <- read.csv("C:/Users/Utente/Downloads/Downloads/divvy/202307-divvy-tripdata.csv")
```

*# Combine together*

```
trips23<- rbind( trip23_Mar, trip23_Apr, trip23_May, trip23_Jun, trip23_Jul, trip23_Aug)
```

*# Drop useless columns*

```
trips23 <- trips23 %>%
  select(-c(start_lat, start_lng, end_lat, end_lng,
start_station_id,end_station_id, end_station_name))
```

*# Statistics*

```
colnames(trips23)
```

```
## [1] "ride_id"           "rideable_type"      "started_at"
## [4] "ended_at"          "start_station_name" "member_casual"
```

```
nrow(trips23)
```

```
## [1] 2967808
```

```
dim(trips23)
```

```
## [1] 2967808      6
```

```
head(trips23, 6)
```

```
##           ride_id rideable_type      started_at      ended_at
## 1 CBCD0D7777F0E45F  classic_bike 2023-02-14 11:59:42 2023-02-14 12:13:38
```

```
## 2 F3EC5FCE5FF39DE9 electric_bike 2023-02-15 13:53:48 2023-02-15 13:59:08
## 3 E54C1F27FA9354FF classic_bike 2023-02-19 11:10:57 2023-02-19 11:35:01
## 4 3D561E04F739CC45 electric_bike 2023-02-26 16:12:05 2023-02-26 16:39:55
## 5 0CB4B4D53B2DBE05 electric_bike 2023-02-20 11:55:23 2023-02-20 12:05:48
## 6 C67EB62172C472EB classic_bike 2023-02-24 18:50:16 2023-02-24 18:56:40
##           start_station_name member_casual
## 1 Southport Ave & Clybourn Ave      casual
## 2   Clarendon Ave & Gordon Ter      casual
## 3 Southport Ave & Clybourn Ave      member
## 4 Southport Ave & Clybourn Ave      member
## 5  Prairie Ave & Garfield Blvd      member
## 6      Wells St & Concord Ln        member
```

```
str(trips23)
```

```
## 'data.frame':   2967808 obs. of  6 variables:
## $ ride_id      : chr  "CBCD0D777F0E45F" "F3EC5FCE5FF39DE9"
## $ rideable_type: chr  "classic_bike" "electric_bike" "classic_bike"
## $ started_at   : chr  "2023-02-14 11:59:42" "2023-02-15 13:53:48" "2023-
## $ ended_at     : chr  "2023-02-14 12:13:38" "2023-02-15 13:59:08" "2023-
## $ start_station_name: chr  "Southport Ave & Clybourn Ave" "Clarendon Ave &
## $ member_casual : chr  "casual" "casual" "member" "member" ...
```

```
summary(trips23)
```

```
##   ride_id      rideable_type      started_at      ended_at
## Length:2967808 Length:2967808 Length:2967808 Length:2967808
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
## start_station_name member_casual
## Length:2967808 Length:2967808
## Class :character Class :character
## Mode :character Mode :character
```

```
#The default format is yyyy-mm-dd
```

```
trips23$date <- as.Date(trips23$started_at)
trips23$month <- format(as.Date(trips23$date), "%m")
trips23$day <- format(as.Date(trips23$date), "%d")
trips23$year <- format(as.Date(trips23$date), "%Y")
trips23$day_of_week <- format(as.Date(trips23$date), "%A")
```

```
# Convert start and end time in hours and minutes
```

```
trips23$time <- format(trips23$started_at, format= "%H:%M:%S")
trips23$time <- as_hms(ymd_hms(trips23$time))
trips23$time2 <- format(trips23$ended_at, format= "%H:%M:%S")
trips23$time2 <- as_hms(ymd_hms(trips23$time2))
```

```
# Time length of a bike walk
```

```
trips23$ride_length <- as.double(difftime(trips23$time2, trips23$time))/60
```

```
#change datatype to numeric for further analysis
```

```
trips23$ride_length <- as.numeric(as.character(trips23$ride_length))
```

```
# View and check changed dataset
```

```
str(trips23)
```

```
## 'data.frame':    2967808 obs. of  14 variables:
## $ ride_id      : chr  "CBCD0D7777F0E45F" "F3EC5FCE5FF39DE9"
## $ rideable_type: chr  "classic_bike" "electric_bike" "classic_bike"
## $ started_at   : chr  "2023-02-14 11:59:42" "2023-02-15 13:53:48" "2023-
## $ ended_at     : chr  "2023-02-14 12:13:38" "2023-02-15 13:59:08" "2023-
## $ start_station_name: chr  "Southport Ave & Clybourn Ave" "Clarendon Ave &
## $ member_casual : chr  "casual" "casual" "member" "member" ...
## $ date         : Date, format: "2023-02-14" "2023-02-15" ...
## $ month        : chr  "02" "02" "02" "02" ...
## $ day          : chr  "14" "15" "19" "26" ...
## $ year         : chr  "2023" "2023" "2023" "2023" ...
## $ day_of_week  : chr  "martedì" "mercoledì" "domenica" "domenica" ...
## $ time         : 'hms' num  11:59:42 13:53:48 11:10:57 16:12:05 ...
## ..- attr(*, "units")= chr  "secs"
## $ time2        : 'hms' num  12:13:38 13:59:08 11:35:01 16:39:55 ...
## ..- attr(*, "units")= chr  "secs"
## $ ride_length  : num  13.93 5.33 24.07 27.83 10.42 ...
```

```
# Remove all blank
```

```
trips23 <- trips23[!(trips23$start_station_name == "HQ QR" |
trips23$ride_length<0),]
head(trips23)
```

```
##           ride_id rideable_type      started_at      ended_at
## 1 CBCD0D7777F0E45F  classic_bike 2023-02-14 11:59:42 2023-02-14 12:13:38
## 2 F3EC5FCE5FF39DE9  electric_bike 2023-02-15 13:53:48 2023-02-15 13:59:08
## 3 E54C1F27FA9354FF  classic_bike 2023-02-19 11:10:57 2023-02-19 11:35:01
## 4 3D561E04F739CC45  electric_bike 2023-02-26 16:12:05 2023-02-26 16:39:55
## 5 0CB4B4D53B2DBE05  electric_bike 2023-02-20 11:55:23 2023-02-20 12:05:48
## 6 C67EB62172C472EB  classic_bike 2023-02-24 18:50:16 2023-02-24 18:56:40
##           start_station_name member_casual      date month day year
## 1 Southport Ave & Clybourn Ave      casual 2023-02-14    02  14 2023
## 2   Clarendon Ave & Gordon Ter      casual 2023-02-15    02  15 2023
## 3 Southport Ave & Clybourn Ave      member 2023-02-19    02  19 2023
## 4 Southport Ave & Clybourn Ave      member 2023-02-26    02  26 2023
## 5   Prairie Ave & Garfield Blvd      member 2023-02-20    02  20 2023
```

```
## 6      Wells St & Concord Ln      member 2023-02-24    02  24 2023
##   day_of_week      time      time2 ride_length
## 1   martedì 11:59:42 12:13:38   13.933333
## 2   mercoledì 13:53:48 13:59:08    5.333333
## 3   domenica 11:10:57 11:35:01   24.066667
## 4   domenica 16:12:05 16:39:55   27.833333
## 5   lunedì 11:55:23 12:05:48   10.416667
## 6   venerdì 18:50:16 18:56:40    6.400000
```

*# Calculate values to determine membership type propagation.*

```
aggregate(trips23$ride_length ~ trips23$member_casual, FUN = mean)
```

```
##   trips23$member_casual trips23$ride_length
## 1             casual      20.13142
## 2             member      11.96180
```

```
aggregate(trips23$ride_length ~ trips23$member_casual, FUN = median)
```

```
##   trips23$member_casual trips23$ride_length
## 1             casual      12.26667
## 2             member       8.60000
```

```
aggregate(trips23$ride_length ~ trips23$member_casual, FUN = max)
```

```
##   trips23$member_casual trips23$ride_length
## 1             casual     1372.95
## 2             member     1139.75
```

```
aggregate(trips23$ride_length ~ trips23$member_casual, FUN = min)
```

```
##   trips23$member_casual trips23$ride_length
## 1             casual           0
## 2             member           0
```

*# Check day of week*

```
trips23$day_of_week <- ordered( trips23$day_of_week, levels=c("domenica",
"lunedì", "martedì", "mercoledì", "giovedì", "venerdì", "sabato"))
```

```
trips23 %>%
  mutate(day_of_week = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, day_of_week ) %>%
  summarise(number_of_rides = n())
```

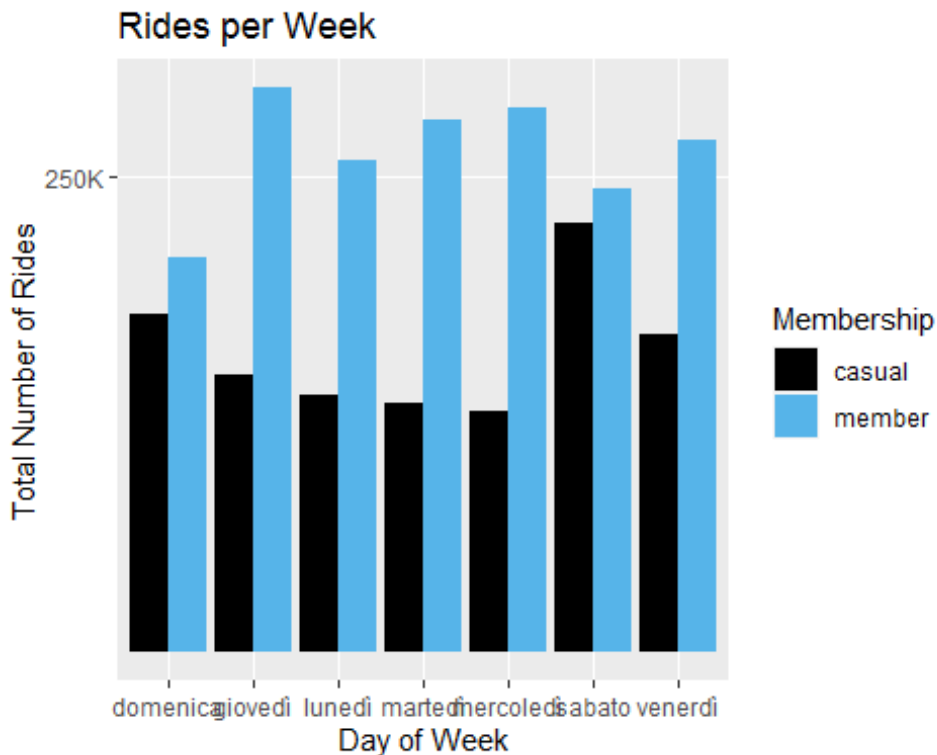
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 14 x 3
## # Groups:   member_casual [2]
##   member_casual day_of_week number_of_rides
##   <chr>         <ord>         <int>
## 1 casual       dom           178237
## 2 casual       lun           135583
```

```
## 3 casual      mar      130740
## 4 casual      mer      126585
## 5 casual      gio      146313
## 6 casual      ven      167494
## 7 casual      sab      225740
## 8 member      dom      207639
## 9 member      lun      258840
## 10 member     mar      280410
## 11 member     mer      286010
## 12 member     gio      296949
## 13 member     ven      269632
## 14 member     sab      244112
```

```
trips23$day_of_week <- format(as.Date(trips23$date), "%A")
trips23 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("#000000", "#56B4E9")) +
  labs(x='Day of Week', y='Total Number of Rides', title='Rides per Week', fill =
'Membership') +
  scale_y_continuous(breaks = c(250000, 450000, 550000), labels = c("250K",
"450K", "550K"))

## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

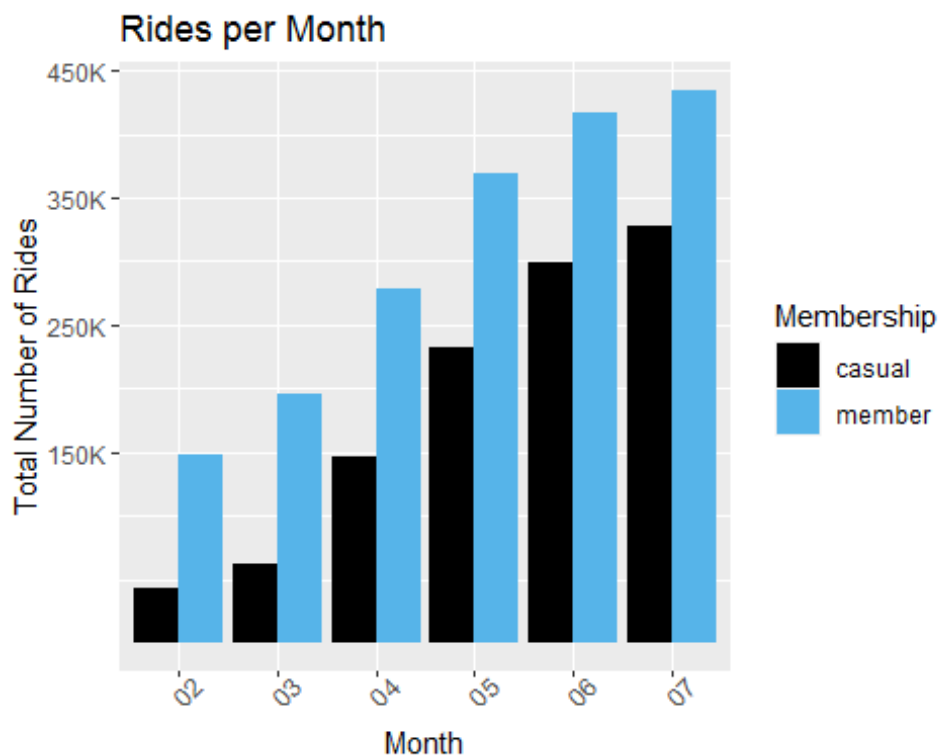


*# We can see that casual use frequently on Friday(venerdì), Saturday(sabato) and Sunday(domenica), and users who own a membership use on average more throughout the week*

*# Total rides per months*

```
trips23 %>%
  group_by(member_casual, month) %>%
  summarise(total_rides = n(), `average_duration_(mins)` = mean(ride_length)) %>%
  arrange(member_casual) %>%
  ggplot(aes(x=month, y=total_rides, fill = member_casual)) + geom_col(position =
"dodge") +
  scale_fill_manual(values = c("#000000", "#56B4E9")) +
  labs(x= "Month", y= "Total Number of Rides", title = "Rides per Month", fill =
"Membership") +
  scale_y_continuous(breaks = c(150000, 250000, 350000, 450000), labels =
c("150K", "250K", "350K", "450K")) + theme(axis.text.x = element_text(angle = 45))

## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



*# During the summer months, casual users predominated.*

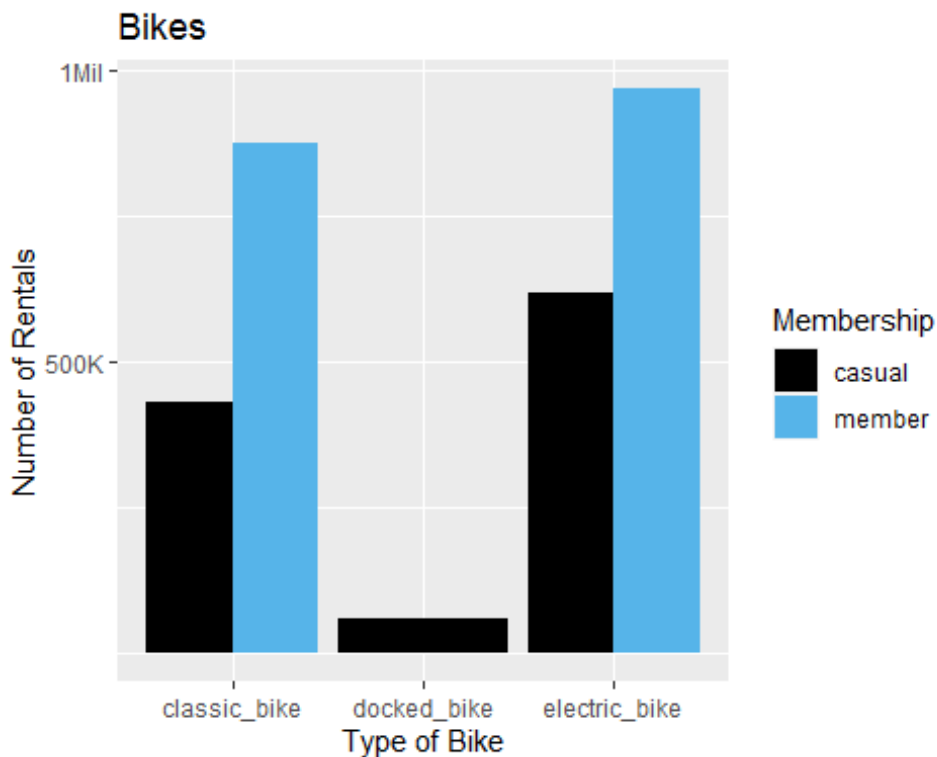
*# Conversely, during the winter months there is very little activity from regular users.*

*# Over the long term, membership users outperform regular/casual users*

*# We know that the company owns two types of bicycles. Let's analyze which type is used more often*

```
trips23 %>%
```

```
ggplot(aes(x = rideable_type, fill = member_casual)) + geom_bar(position =
"dodge") +
  scale_fill_manual(values = c("#000000", "#56B4E9")) +
  labs(x= 'Type of Bike', y='Number of Rentals', title='Bikes', fill =
'Membership') +
  scale_y_continuous(breaks = c(500000, 1000000, 1500000), labels = c("500K",
"1Mil", "1.5Mil"))
```



```
trips23 %>%
  mutate(day_of_week = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n() ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_manual(values = c("#000000",
"#56B4E9")) +
  labs(x='Days of the week', y='Average Mins', title='Average ride time',
fill='Membership')
```

## `summarise()` has grouped output by 'member\_casual'. You can override using the  
## `.groups` argument.

