Chicago BikeRide Analysis

Patrick

2022-08-19

Chicago Bike Ride

This is an analysis of Chicago bicycle users. The intention of this analysis is to find out the profiles of the users trying to find some trends about it. Questions: 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?

Libraries

The following libraries were used for this analysis:

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.1     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tidyr)
library(janitor)
```

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

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

```
library(purrr)
library(geosphere)
```

Downloading and the Cleaning Data

The Data are for users over one year, from May 2021 to May 2022.

```
#May 21
bike_may_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202105-divvy-tripdata.csv")
bike_may_2021[bike_may_2021 == ""] <- NA
bike_may_2021 <-na.omit(bike_may_2021)


#June 21
bike_june_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202106-divvy-tripdata.csv")
bike_june_2021[bike_june_2021 == ""] <- NA
bike_june_2021 <-na.omit(bike_june_2021)


#July 21
bike_july_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202107-divvy-tripdata.csv")
bike_july_2021[bike_july_2021 == ""] <- NA
bike_july_2021 <-na.omit(bike_july_2021)


#August 21
bike_august_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202108-divvy-tripdata.csv")
bike_august_2021[bike_august_2021 == ""] <- NA
bike_august_2021 <-na.omit(bike_august_2021)


#September 21
bike_september_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202109-divvy-tripdata.csv")
bike_september_2021[bike_september_2021 == ""] <- NA
bike_september_2021 <-na.omit(bike_september_2021)

#October 21
bike_october_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202110-divvy-tripdata.csv")
bike_october_2021[bike_october_2021 == ""] <- NA
bike_october_2021 <-na.omit(bike_october_2021)


#November 21
bike_november_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202111-divvy-tripdata.csv")
bike_november_2021[bike_november_2021 == ""] <- NA
```

```
bike_november_2021 <-na.omit(bike_november_2021)


#December 21
bike_december_2021 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202112-divvy-tripdata.csv")
bike_december_2021[bike_december_2021 == ""] <- NA
bike_december_2021 <-na.omit(bike_december_2021)


#January 22
bike_january_2022 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202201-divvy-tripdata.csv")
bike_january_2022[bike_january_2022 == ""] <- NA
bike_january_2022 <-na.omit(bike_january_2022)


#February 22
bike_february_2022 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202202-divvy-tripdata.csv")
bike_february_2022[bike_february_2022 == ""] <- NA
bike_february_2022 <-na.omit(bike_february_2022)


#March 22
bike_march_2022 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202203-divvy-tripdata.csv")
bike_march_2022[bike_march_2022 == ""] <- NA
bike_march_2022 <-na.omit(bike_march_2022)

#April 22
bike_april_2022 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202204-divvy-tripdata.csv")
bike_april_2022[bike_april_2022 == ""] <- NA
bike_april_2022 <-na.omit(bike_april_2022)


#May 22
bike_may_2022 <-read.csv("D:/GoogleAnalytcsCourse/CaseBike/202205-divvy-tripdata.csv")
bike_may_2022[bike_may_2022 == ""] <- NA
bike_may_2022 <-na.omit(bike_may_2022)
```

Counting Memnbers and then Joining in one Table

```
members_may_2021 <- bike_may_2021 %>% count(member_casual)
members_june_2021 <- bike_june_2021 %>% count(member_casual)
members_july_2021 <- bike_july_2021 %>% count(member_casual)
members_august_2021 <- bike_august_2021 %>% count(member_casual)
members_september_2021 <- bike_september_2021 %>% count(member_casual)
members_october_2021 <- bike_october_2021 %>% count(member_casual)
members_november_2021 <- bike_november_2021 %>% count(member_casual)
members_december_2021 <- bike_december_2021 %>% count(member_casual)
members_january_2022 <- bike_january_2022 %>% count(member_casual)
members_february_2022 <- bike_february_2022 %>% count(member_casual)
members_march_2022 <- bike_march_2022 %>% count(member_casual)
members_april_2022 <- bike_april_2022 %>% count(member_casual)
```

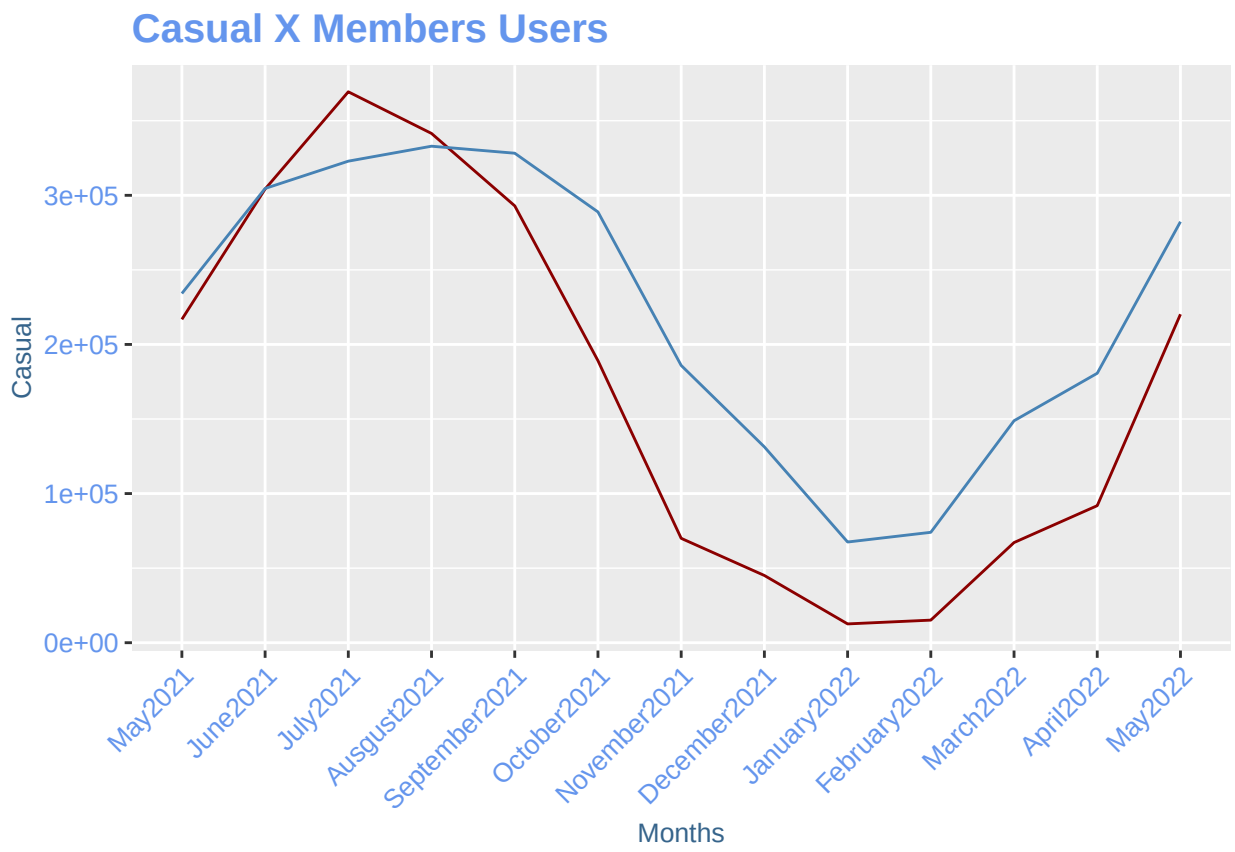
```r
members_may_2022 <- bike_may_2022 %>% count(member_casual)


total_members <-purrr::reduce(list(members_may_2021,members_june_2021,
                                   members_july_2021, members_august_2021,
                                   members_september_2021, members_october_2021,
                                   members_november_2021,
                                   members_december_2021, members_january_2022,
                                   members_february_2022, members_march_2022,
                                   members_april_2022, members_may_2022), dplyr::left_join ,by="member_ca

total_members_table <-as.data.frame(t(total_members))
total_members_table<- total_members_table %>% row_to_names(row_number = 1)

#Naming the rows
row.names(total_members_table)[1] <- "May2021"
row.names(total_members_table)[2] <- "June2021"
row.names(total_members_table)[3] <- "July2021"
row.names(total_members_table)[4] <- "August2021"
row.names(total_members_table)[5] <- "September2021"
row.names(total_members_table)[6] <- "October2021"
row.names(total_members_table)[7] <- "November2021"
row.names(total_members_table)[8] <- "December2021"
row.names(total_members_table)[9] <- "January2022"
row.names(total_members_table)[10] <- "February2022"
row.names(total_members_table)[11] <- "March2022"
row.names(total_members_table)[12] <- "April2022"
row.names(total_members_table)[13] <- "May2022"

total_members_table$Meses <- c("May2021","June2021","July2021","Ausgust2021",
                               "September2021","October2021","November2021","December2021",
                              "January2022","February2022","March2022","April2022","May2022")

total_members_table$Meses <- factor(total_members_table$Meses,levels= c("May2021","June2021","July2021"
                               "September2021","October2021","November2021","December2021",
                              "January2022","February2022","March2022","April2022","May2022"))



#Transforming the column type in integer to summary them
total_members_table <- transform(total_members_table, casual=as.numeric(casual), member=as.numeric(membe
summary(total_members_table)
```

```
##      casual           member                 Meses
##  Min.   : 12605   Min.   : 67523   May2021      :1
##  1st Qu.: 67156   1st Qu.:148827   June2021     :1
##  Median :189117   Median :234165   July2021     :1
##  Mean   :172005   Mean   :221710   Ausgust2021  :1
##  3rd Qu.:292931   3rd Qu.:304586   September2021:1
##  Max.   :369415   Max.   :332933   October2021  :1
##                                    (Other)      :7
```

Casual X Members Chart

After summarizing and cleaning the data, it is time to visualize the relationship between casual and regular members

```
themes <- theme(
  plot.title = element_text(colour ="cornflowerblue", face = "bold", size = (15)),
  axis.title = element_text(size = (10), colour = "steelblue4"),
  axis.text = element_text(colour = "cornflowerblue", size = (10))
)

CMplot<- ggplot(total_members_table, aes(x=Meses)) +
  geom_line(aes(y = casual,group = 1), color = "darkred") +
  geom_line(aes(y = member,group = 2), color="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

print(CMplot+themes+labs(title = "Casual X Members Users", y= "Casual", x= "Months"))
```



The first thing we can observe is the months when we have more users is during the summer, since the data is from Chicago, we can notice that in the winter the number of users decreases significantly. One of the solutions that the company could propose in this period would be to offer breaks for members during this time or propose the exchange for other services, or think in a sustainable way to propose to the user to donate this money to charity since mainly in winter the demand for this type of activity should be strongly encouraged.

##Checking the most type of bike in a year

```r
type_of_bikes_may21 <- bike_may_2021 %>%count(rideable_type)
type_of_bikes_june21 <- bike_june_2021 %>%count(rideable_type)
type_of_bikes_july21 <- bike_july_2021 %>%count(rideable_type)
type_of_bikes_august21 <- bike_august_2021 %>%count(rideable_type)
type_of_bikes_september21 <- bike_september_2021 %>%count(rideable_type)
type_of_bikes_october21 <- bike_october_2021 %>%count(rideable_type)
type_of_bikes_november21 <- bike_november_2021 %>%count(rideable_type)
type_of_bikes_december21 <- bike_december_2021 %>%count(rideable_type)
type_of_bikes_january22 <- bike_january_2022 %>%count(rideable_type)
type_of_bikes_february22 <- bike_february_2022 %>%count(rideable_type)
type_of_bikes_march22 <- bike_march_2022 %>%count(rideable_type)
type_of_bikes_april22 <- bike_april_2022 %>%count(rideable_type)
type_of_bikes_may22 <- bike_may_2022 %>%count(rideable_type)


#joining the type of bikes in one table
total_type_bikes <-purrr::reduce(list(type_of_bikes_may21,type_of_bikes_june21,
                                      type_of_bikes_july21, type_of_bikes_august21,
                                      type_of_bikes_september21, type_of_bikes_october21,
                                      type_of_bikes_november21,
                                      type_of_bikes_december21, type_of_bikes_january22,
                                      type_of_bikes_february22, type_of_bikes_march22,
                                      type_of_bikes_april22, type_of_bikes_may22), dplyr::left_join ,by=


total_type_bikes <-as.data.frame(t(total_type_bikes))


#Making the first row as the columns names

total_type_bikes<- total_type_bikes %>% row_to_names(row_number = 1)


#Naming the rows
row.names(total_type_bikes)[1] <- "May2021"
row.names(total_type_bikes)[2] <- "June2021"
row.names(total_type_bikes)[3] <- "July2021"
row.names(total_type_bikes)[4] <- "August2021"
row.names(total_type_bikes)[5] <- "September2021"
row.names(total_type_bikes)[6] <- "October2021"
row.names(total_type_bikes)[7] <- "November2021"
row.names(total_type_bikes)[8] <- "December2021"
row.names(total_type_bikes)[9] <- "January2022"
row.names(total_type_bikes)[10] <- "February2022"
row.names(total_type_bikes)[11] <- "March2022"
row.names(total_type_bikes)[12] <- "April2022"
row.names(total_type_bikes)[13] <- "May2022"



total_type_bikes$Meses <- c("May2021","June2021","July2021","Ausgust2021",
                            "September2021","October2021","November2021","December2021",
                            "January2022","February2022","March2022","April2022","May2022")
```

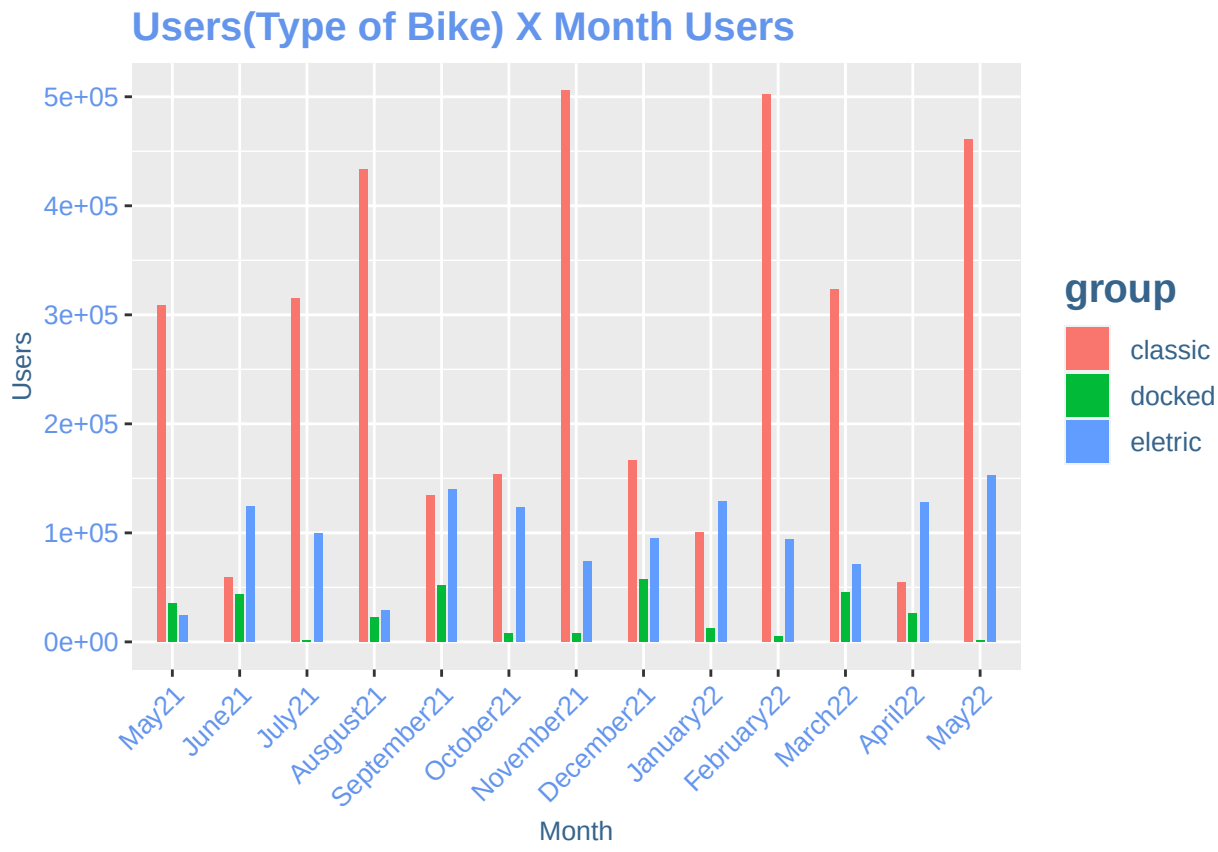
```
total_type_bikes$Meses <- factor(total_members_table$Meses,levels= c("May2021","June2021","July2021","Au
                                                                     "September2021","October2021","
                                                                     "January2022","February2022","Ma



#Transforming the column type in integer to summary them
total_type_bikes <- transform(total_type_bikes, classic=as.numeric(classic_bike), docked= as.numeric(do
summary(total_type_bikes)
```

```
##  classic_bike       docked_bike        electric_bike                Meses
##  Length:13          Length:13          Length:13          May2021      :1
##  Class :character   Class :character   Class :character   June2021     :1
##  Mode  :character   Mode  :character   Mode  :character   July2021     :1
##                                                           Ausgust2021  :1
##                                                           September2021:1
##                                                           October2021  :1
##                                                           (Other)      :7
##     classic           docked         electric
##  Min.   : 54697   Min.   :  943   Min.   : 24488
##  1st Qu.:134292   1st Qu.: 7565   1st Qu.: 73469
##  Median :308330   Median :22689   Median : 99311
##  Mean   :270614   Mean   :24378   Mean   : 98723
##  3rd Qu.:433787   3rd Qu.:43353   3rd Qu.:127515
##  Max.   :505544   Max.   :57698   Max.   :152824
##
```

```
View(total_type_bikes)

#Reshaping the results of the total type of bike to plot in a ggplot bar

dfbike <- data.frame(group = c("classic", "docked","eletric"),
                     May21=c(308330,43353,99311),
                     June21=c(433787,51716,123275),
                     July21=c(505544,57698,129079),
                     August21=c(501829,45065,127515),
                     september21=c(461077,35337,124736),
                     October21=c(315180,22689,140103),
                     November21=c(153630,7565,94709),
                     December21=c(100272,4878,71221),
                     January22=c(54697,943,24488),
                     February22=c(59223,1344,28611),
                     March22=c(134292,8222,73469),
                     April22=c(166524,11980,94056),
                     May22=c(323601,26120,152824))
rdtbike <- dfbike %>% gather(key = Month, value = Users, May21:May22)
View(rdtbike)
rdtbike$Month <- c("May21","June21","July21","Ausgust21",
                           "September21","October21","November21","December21",
                           "January22","February22","March22","April22","May22")
rdtbike$Month <- factor(rdtbike$Month,levels= c("May21","June21","July21","Ausgust21",
                                                "September21","October21","Novembe
                                                "January22","February22","March22"
```

##Graph (Type of Bike in a year)

```
themes <- theme(
  plot.title = element_text(colour ="cornflowerblue", face = "bold", size = (15)),
  axis.title = element_text(size = (10), colour = "steelblue4"),
  axis.text = element_text(colour = "cornflowerblue", size = (10)),
  legend.title = element_text(colour ="steelblue4", face = "bold", size = (15)),
  legend.text = element_text(colour = "steelblue4", size = (10)))


UMplot <- ggplot(data = rdtbike, aes(x = Month, y=Users, fill = group )) +
  geom_bar(stat = "identity",  width = 0.4,
           position=position_dodge(width = 0.5))+
  theme(axis.text.x = element_text(angle = 45, hjust=1))

print(UMplot+themes+labs(title = "Users(Type of Bike) X Month Users", y= "Users", x= "Month"))
```



By analyzing this graph we can see that the classic bicycle is the most used. Electric and the docked type has an increasing number during the months of December, August and September which could indicate a vacation user since it is a family type of bicycle.

##Verifying the km by user

```
bike_may_2021 <- bike_may_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_june_2021 <- bike_june_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_july_2021 <- bike_july_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_august_2021 <- bike_august_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_september_2021 <- bike_september_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_october_2021 <- bike_october_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_november_2021 <- bike_november_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_december_2021 <- bike_december_2021 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_january_2022 <- bike_january_2022 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_february_2022 <- bike_february_2022 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_march_2022 <- bike_march_2022 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
```

```
  )

bike_april_2022 <- bike_april_2022 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

bike_may_2022 <- bike_may_2022 %>%
  mutate(
    dist= geosphere::distHaversine(cbind(start_lng, start_lat)/1000, cbind(end_lng,end_lat)/1000)
  )

countbmay_2021 <- sum(bike_may_2021$dist == 0.0000000)
countbjun_2021 <- sum(bike_june_2021$dist == 0.0000000)
countbjul_2021 <- sum(bike_july_2021$dist == 0.0000000)
countbaug_2021 <- sum(bike_august_2021$dist == 0.0000000)
countbsep_2021 <- sum(bike_september_2021$dist == 0.0000000)
countboct_2021 <- sum(bike_october_2021$dist == 0.0000000)
countbnov_2021 <- sum(bike_november_2021$dist == 0.0000000)
countbdec_2021 <- sum(bike_december_2021$dist == 0.0000000)
countbjan_2022 <- sum(bike_january_2022$dist == 0.0000000)
countbfev_2022 <- sum(bike_february_2022$dist == 0.0000000)
countbmarch_2022 <- sum(bike_march_2022$dist == 0.0000000)
countbapril_2022 <- sum(bike_april_2022$dist == 0.0000000)
countbmay_2022 <- sum(bike_may_2022$dist == 0.0000000)

df_distance <- data.frame(c("may21",
                            "jun21","jul21","aug21","sep21","oct21",
                            "nov21","dec21","jan/22","fev/22",
                            "mar/22","apr22","may22"),c(countbmay_2021,countbjun_2021,countbjul_2021,
                                                        countbaug_2021,countbsep_2021,countboct_2021,
                                                        countbnov_2021,countbdec_2021,countbjan_2022,
                                                        countbfev_2022,countbmarch_2022,countbapril_202
                                                        countbmay_2022));
names(df_distance)<- c("Month","Count")

View(df_distance)
summary(df_distance)
```

```
##     Month               Count
##  Length:13          Min.   : 2616
##  Class :character   1st Qu.: 8500
##  Mode  :character   Median :20656
##                     Mean   :22086
##                     3rd Qu.:35154
##                     Max.   :44349
```

Analyzing this data, we could notice the high number of member users with the number 0km, this occurs because they tend to return to the same station (remaining the same longitude and latitude), but when looking for the time, it is found to be different from zero.

##Conclusion

Analyzing the data, we see that the number of casuals and members for a year starting in May 2021 and May 2022 tends to be approximate. To increase these numbers, a good approach would be to work with

points, the user could earn points per mile or km and these could exchange for products and all this boosted through digital advertising, influencers riding and showing the practicality of getting to know the city with the bike.