---

title: "Case Study: Bike-Sharing-Company Cyclistic"

author: "Reven Cristian"

date: "2023-11-13"

output:

html\_document: default

pdf\_document: default

editor\_options:

chunk\_output\_type: console

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

I load libraries tidyverse

```{r libraries, echo=FALSE, warning=FALSE}

library(knitr)

library(officer)

library(tidyverse)

library(readxl)

library(ggplot2)

```

I load the tables

```{r loadDB, echo=FALSE}

file1<- "~/daten\_bikes/db202305.xlsx"

file2<- "~/daten\_bikes/db202306.xlsx"

file3<- "~/daten\_bikes/db202307.xlsx"

file4<- "~/daten\_bikes/db202308.xlsx"

file5<- "~/daten\_bikes/db202309.xlsx"

file6<- "~/daten\_bikes/db202310.xlsx"

tabl1<- read\_excel(file1)

tabl2<- read\_excel(file2)

tabl3<- read\_excel(file3)

tabl4<- read\_excel(file4)

tabl5<- read\_excel(file5)

tabl6<- read\_excel(file6)

#I solved an error in a data type

tabl2$end\_station\_id <- as.character(tabl2$end\_station\_id)

lt\_comb<- bind\_rows(tabl1,tabl2,tabl3,tabl4,tabl5,tabl6)

```

I extract only time from ride\_length

```{r timeFromDate, warning=FALSE}

lt\_comb$ride\_length <- format(as.POSIXct(lt\_comb$ride\_length),"%H:%M:%S")

```

I add a new column with time values as numeric

```{r time-to-numeric, echo=FALSE}

lt\_comb$duration\_min <- as.numeric(difftime(lt\_comb$ended\_at,lt\_comb$started\_at, units="min"))

```

I look duplicated ride\_id

```{r summary, echo=FALSE}

#summary(lt\_comb)

doppel\_values<- duplicated(lt\_comb$ride\_id)

lt\_comb[doppel\_values, ]

```

I look how many values there are bigger as 12 hours

```{r maxValues1, echo=FALSE}

max\_val<- lt\_comb%>%

filter(lt\_comb$duration\_min > 720)

summary(max\_val)

```

I look how many values are lower than 0 minute

```{r minValues, echo=FALSE}

min\_val<- lt\_comb%>%

filter(lt\_comb$duration\_min < 0)

summary(min\_val)

```

I founded that the data is bad loaded exchanging the start and end data of each trip, I solve this changing the difftime formula for this table

```{r convertin-to+, echo=FALSE}

min\_val$duration\_min <- as.numeric(difftime(min\_val$started\_at,min\_val$ended\_at , units="min"))

```

I create an array with ride\_id with the solved values

```{r newTable1.1.1, echo=FALSE}

min\_val\_id\_ride<- min\_val$ride\_id

```

I create a new table with solved values to trip duration

```{r newTable1.1, echo=FALSE}

lt\_N\_comb<- lt\_comb%>%

filter(!(lt\_comb$ride\_id %in% min\_val\_id\_ride))

lt\_N\_comb<- rbind(lt\_N\_comb,min\_val)

```

I look how many values there are bigger as 4 hours

```{r maxValues, echo=FALSE}

max\_values<- lt\_N\_comb%>%

filter(lt\_N\_comb$duration\_min> 240)

print(max\_values)

```

I filter the data set by extreme values, less than 1 minute and more than 240 minutes

```{r selLowValues,echo=FALSE}

lt\_extrem\_values <- filter(lt\_N\_comb,lt\_N\_comb$duration\_min < 1 | lt\_N\_comb$duration\_min > 240)

print(lt\_extrem\_values)

```

I create an array with ride\_id with extreme values less than 1 minute and more than 240 minutes (12 hours)

```{r newTable, echo=FALSE}

lv\_ext\_id\_ride<- lt\_extrem\_values$ride\_id

```

I create a new table without trips with duration extreme values and I show a summary

```{r newTable1, echo=FALSE}

lt\_NN\_comb <- subset(lt\_N\_comb, !(lt\_N\_comb$ride\_id %in% lt\_extrem\_values$ride\_id))

summarize(lt\_NN\_comb)

```

Filtered table start and end time, duration and users

```{r filter\_per\_duration, message=FALSE, warning=FALSE, echo=FALSE}

lt\_NN\_comb\_filt<- lt\_NN\_comb%>%

select(ride\_id, rideable\_type, started\_at, ended\_at, member\_casual, ride\_length, duration\_min)

summary(lt\_NN\_comb\_filt)

```

##Graphics

I show An Histogram with the ride duration

```{r graphic9, echo=FALSE, warning=FALSE, message=FALSE}

average <- mean(lt\_NN\_comb$duration\_min, na.rm = TRUE)

ggplot(lt\_NN\_comb, aes(x=duration\_min))+

geom\_histogram(fill= "lightblue", color="blue")+

geom\_vline(xintercept = average, color = "red", linetype = "dashed", size = 1) +

#scale\_x\_continuous(breaks = seq(0,240, by=10) )+

scale\_x\_continuous(breaks = seq(0,240, by=10), limits = c(1,240))+

labs(title = "Distribution of trip duration ",x ="Duration in Minutes",xintercept="mean", y="count")+

annotate("text", x = average + 5, y = 1500000, label = paste("Mean =", round(average, 2)),

color = "red", size = 4, hjust = 0)

```

I graph the proportion beetwen casual users and members

```{r graphics, echo=FALSE, warning=FALSE}

ggplot(lt\_NN\_comb)+

geom\_bar(mapping = aes(x= lt\_NN\_comb$member\_casual, fill=member\_casual))+

labs(x="Type of Member", title = "Type of Member", fill= "Type of member")

```

I graph the type of bicycle vs the type of user

```{r graphics1, echo=FALSE, warning=FALSE}

# user vs kind of bike

ggplot(lt\_NN\_comb)+

geom\_bar(mapping = aes(x= lt\_NN\_comb$rideable\_type, fill= lt\_NN\_comb$member\_casual))+

facet\_wrap(~member\_casual)+

labs(x="Type of bike", fill= "Type of member",title = "type of Bike vs type of Member")

```

I graph the length of the trip vs the type of bike

```{r graphic2, echo=FALSE, warning=FALSE}

ggplot(lt\_NN\_comb, aes(x = rideable\_type, y = duration\_min, color= rideable\_type)) +

geom\_boxplot()+

facet\_grid(~member\_casual)+

scale\_y\_continuous( breaks = seq(0,260, by= 20))+

labs(title = "Type of Bike vs Duration of Trip vs User Type", y="Duration in minutes", x="Type of Bike",color="Type of Bike")

```

I graph the length of the trip vs the type of bike limiting the ride duration until 60 minutes

```{r graphic3, echo=FALSE, warning=FALSE}

ggplot(lt\_NN\_comb, aes(x = rideable\_type, y = duration\_min, color= rideable\_type)) +

geom\_boxplot()+

facet\_grid(~member\_casual)+

scale\_y\_continuous(limits = c(0,60), breaks = seq(0,60, by= 10))+

labs(title = "Type of Bike vs Duration of Trip vs User Type", y="Duration in minutes", x="Type of Bike",color="Type of Bike")

```

I graph the length of the trip vs the type of bike

```{r graphic4, echo=FALSE, warning=FALSE}

#ggplot(lt\_NN\_comb)+

# geom\_jitter(mapping = aes(x= duration\_min, y= member\_casual, color= member\_casual))+

# labs(x="Ride length", y="Type of Member", color= "Type of Member", title = "Trip duration per #User Type")+

# scale\_x\_continuous(limits = c(0,240), breaks = seq(0,240, by=20))

```

I show the average of trips duration per type of user

```{r, sumary\_duration, warning=FALSE, echo=FALSE, message=FALSE}

promedio\_duraciones <- lt\_NN\_comb %>%

group\_by(member\_casual) %>%

summarise(duracion\_promedio = mean(duration\_min))

ggplot(promedio\_duraciones, aes(x = member\_casual, y = duracion\_promedio , fill = member\_casual)) +

geom\_bar(stat = "identity") +

labs(x = "Type of Member", y = "Average Ride Length (minutes)", fill = "Type of Member", title="Average of Trip Duration per User Type") +

theme\_minimal()

```

I graph the length of rides during the Weekday vs the type of user

```{r graphic5, echo=FALSE, warning=FALSE}

week\_day <- factor(lt\_NN\_comb$weekday, levels = 1:7, labels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

ggplot(lt\_NN\_comb)+

geom\_bar(aes(x= week\_day, fill= member\_casual), position="dodge")+

labs(x="weekdays", fill= "Type of Member", title = "User Types per Weekday")

```

I graph the length of the trip vs weekdays------------------------------------

```{r graphic6, echo=FALSE, warning=FALSE}

week\_day <- factor(lt\_NN\_comb$weekday, levels = 1:7, labels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

ggplot(lt\_NN\_comb)+

geom\_jitter(mapping = aes(x= week\_day, y= lt\_NN\_comb$duration\_min, color= member\_casual))+

labs(x="Weekdays", y="Duration", color= "Member type", title = "Trip duration per Weekdays")

```

I graph the Top 10 Start Stations of the trip vs weekdays

```{r graphic7, echo=FALSE, warning=FALSE}

frec<- table(lt\_NN\_comb$start\_station\_name, lt\_NN\_comb$weekday)

top10\_stations<- names(sort(rowSums(frec),decreasing = TRUE)[1:10])

data\_top10\_stations <- lt\_NN\_comb[lt\_NN\_comb$start\_station\_name %in% top10\_stations, ]

week\_day <- factor(data\_top10\_stations$weekday, levels = 1:7, labels = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"))

custom\_colors <- c("#FF0000", "#004800", "#0000FF", "#FFFF00", "#800080", "#00FFFF", "#FFA500", "#FFC0CB", "#00FF00", "#A52A2A")

ggplot(data\_top10\_stations,aes(x= week\_day, fill= start\_station\_name))+

geom\_bar(position="dodge")+

facet\_wrap(~member\_casual)+

labs(x="Weekdays", y="Station", fill="Station Name", title = "Top 10 use of Stations by days of the Week ")+

scale\_fill\_manual(

values= setNames(custom\_colors, top10\_stations),

breaks = top10\_stations

)

```

```{r maps, warning=FALSE, message=FALSE, echo=FALSE}

library(leaflet)

#I show in a map the distribution of the most important station were trips are #starting

estaciones <- data.frame(

nombre = c("Streeter Dr & Grand Ave", "DuSable Lake Shore Dr & Monroe St.",

"DuSable Lake Shore Dr & North Blvd.", "Michigan Ave & Oak St.",

"Theater on the Lake","Clark St & Elm St", "Wells St & Concord Ln",

"Millennium Park", "Kingsbury St & Kinzie St", "Wells St & Elm St"),

lat = c(41.892278, 41.881042, 41.911722, 41.900960, 41.926666,41.921973, 41.912418, 41.882556, 41.889177, 41.903222),

lon = c(-87.612043, -87.616230, -87.618549, -87.623777, -87.631122,-87.631749, -87.633503, -87.624774, -87.638506, -87.634590)

)

mapa <- leaflet(estaciones) %>%

addTiles() %>%

addMarkers(

lat = ~lat,

lng = ~lon,

label = ~nombre,

labelOptions = labelOptions(

noHide = TRUE,

textOnly = TRUE

)

)

```