Cyclistic Bike-Share Analysis

Lilian Nwafor

2022-10-28

```
#LOAD PACKAGES
library(tidyr)
library(dplyr)
library(readr)
library(lubridate)
library(tidyverse)
library(ggplot2)
```

#ANALYSIS FOR THE MONTH OF JANUARY, 2021

#THE LINK TO THE DATASET CAN BE FOUND HERE ()

```
# IMPORT DATA
Jan <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Jan-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Jan <- subset(Jan, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,End_Station_ID,X))
```

```
#TOTAL NUMBER OF ROWS
nrow(Jan)
```

```
## [1] 131573
```

```
# TOTAL NUMBER OF COLUMNS
ncol(Jan)
```

```
## [1] 9
```

```
#duplicated(Jan)
sum(duplicated(Jan)%>%
          head(n = 20))
```

```
#add new column to the dataset
Jan %>%
          add_column(Time_Diff1 = NA,Lat1 = NA, Ride_Lng1 = NA) %>%
          head(n = 20)
```

```
#RENAME COLUMNS
Jan <- Jan %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Jan$Ride_Lng1 = (Jan$Start_Lng - Jan$End_Lng)
Jan$Lat1 = (Jan$Start_Lat - Jan$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Jan$Time_Diff1 = difftime(Jan$Ended_at, Jan$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Jan$Ride_Lng1, na.rm = TRUE)
```

```
## [1] 0.0001779732
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Jan[Jan$Status == 'member','Ride_Lng1'],na.rm = TRUE)
```

```
## [1] 0.0001196993
```

```r
mean(Jan[Jan$Status == 'casual','Ride_Lng1'],na.rm = TRUE)
```

```
## [1] 0.0003746773
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Jan[Jan$Status == 'casual','Time_Diff1'],na.rm = TRUE)
```

```
## Time difference of 747393.6 secs
```

```r
mean(Jan[Jan$Status == 'member','Time_Diff1'],na.rm = TRUE)
```

```
## Time difference of 909641.2 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Jan[Jan$Time_Diff1 >= 400,'Ride_Lng1'],na.rm = TRUE)
```

```
## [1] 0.0002839189
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Jan[Jan$Ride_Lng1 >= 0.1,'Time_Diff1'],na.rm = TRUE)
```

```
## Time difference of 2563.2 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Jan[Jan$Rideable_Type == 'docked_bike','Ride_Lng1'],na.rm = TRUE)
```

```
## [1] 7.919911e-05
```

```r
mean(Jan[Jan$Rideable_Type == 'electric_bike','Ride_Lng1'],na.rm = TRUE)
```

```
## [1] 0.0002377885
```

```r
mean(Jan[Jan$Rideable_Type == 'classic_bike','Ride_Lng1'],na.rm = TRUE)
```

```
## [1] 0.000155473
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Jan[Jan$Rideable_Type == 'docked_bike','Time_Diff1'],na.rm = TRUE)
```

```
## Time difference of 4836923 secs
```

```r
mean(Jan[Jan$Rideable_Type == 'electric_bike','Time_Diff1'],na.rm = TRUE)
```

```
## Time difference of 1183431 secs
```

```r
mean(Jan[Jan$Rideable_Type == 'classic_bike','Time_Diff1'],na.rm = TRUE)
```

```
## Time difference of -68604.85 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALIES
max(Jan$Ride_Lng1, na.rm = TRUE)
```

```
## [1] 0.1391249
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALIES
min(Jan$Ride_Lng1, na.rm = TRUE)
```

```
## [1] -0.13
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Jan$Ride_Lng1)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Jan$Status)
```

```
##
## casual member
##  30080 101493
```

```r
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Jan$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##         70616         13004         47953
```

```r
#EXTRACT COLUMNS OF INTEREST
Jan_df <- Jan %>%
  select(Ride_ID, Rideable_Type, Time_Diff1, Lat1, Ride_Lng1, Status)
```

```r
#preview of the dataset
head(Jan_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff1        Lat1   Ride_Lng1 Status
## 1 70B6A9A437D4C30D  classic_bike   660 secs -0.01098991 -0.01033958 member
## 2 158A465D4E74C54A electric_bike   420 secs  0.02000000  0.00000000 member
## 3 5262016E0F1F2F9A electric_bike   420 secs -0.02000000  0.01000000 member
```

```r
#preview the data summary
glimpse(Jan_df)
```

```
## Rows: 131,573
## Columns: 6
## $ Ride_ID       <chr> "70B6A9A437D4C30D", "158A465D4E74C54A", "5262016E0F1F2F9~
## $ Rideable_Type <chr> "classic_bike", "electric_bike", "electric_bike", "elect~
## $ Time_Diff1    <drtn> 660 secs, 420 secs, 420 secs, 360 secs, 120 secs, 480 s~
## $ Lat1          <dbl> -0.01098991, 0.02000000, -0.02000000, 0.01000000, 0.0000~
## $ Ride_Lng1     <dbl> -0.01033958, 0.00000000, 0.01000000, 0.00000000, 0.00000~
## $ Status        <chr> "member", "member", "member", "member", "member", "membe~
```

```r
summary(Jan_df)
```

```
##    Ride_ID          Rideable_Type       Time_Diff1            Lat1
##  Length:131573      Length:131573      Length:131573     Min.   :-0.20246
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.00876
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.00000
##                                                          Mean   :-0.00009
##                                                          3rd Qu.: 0.00859
##                                                          Max.   : 0.26220
##                                                          NA's   :111
##    Ride_Lng1           Status
##  Min.   :-0.13000   Length:131573
##  1st Qu.:-0.00859   Class :character
##  Median : 0.00000   Mode  :character
##  Mean   : 0.00018
##  3rd Qu.: 0.00897
##  Max.   : 0.13912
##  NA's   :111
```

```r
max(Jan_df$Ride_Lng1,na.rm = TRUE)
```

```
## [1] 0.1391249
```

```r
min(Jan_df$Ride_Lng1,na.rm = TRUE)
```

```
## [1] -0.13
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type

RideLength_Jan_df <- filter(Jan_df,Ride_Lng1 >= 0.1)%>%
          head(n = 20)
```

```
#pie plot
#for Less than 0.1
LessRideLength_Jan_df <- filter(Jan_df,Ride_Lng1 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Jan_df <- filter(Jan_df,Time_Diff1 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Jan_df <- filter(Jan_df,Time_Diff1 < 400)%>%
          head(n = 20)
```

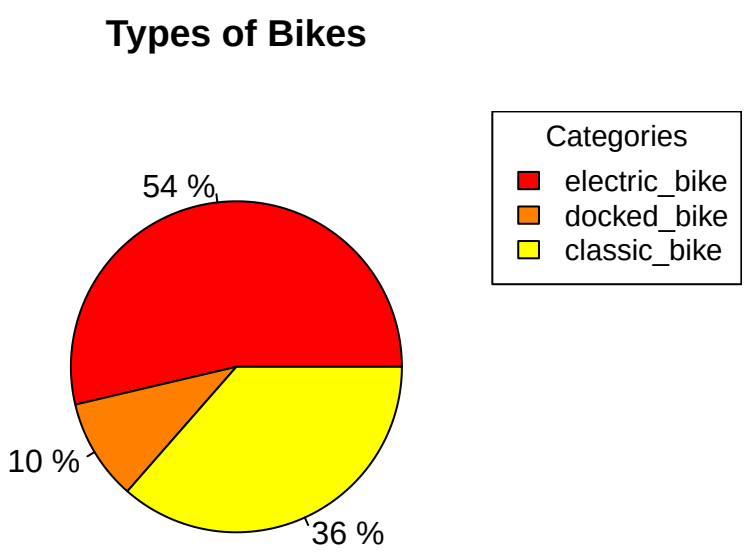
VISUALIZATION

PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Jan_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "54 %" "10 %" "36 %"
```

```
#pie plot
pie(table(Jan_df$Rideable_Type), labels =
      paste(round(prop.table(table(Jan_df$Rideable_Type))*100),"%", sep = " "),
   col = heat.colors(3), main = "Types of Bikes", radius= 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories",cex = 0.9)
```

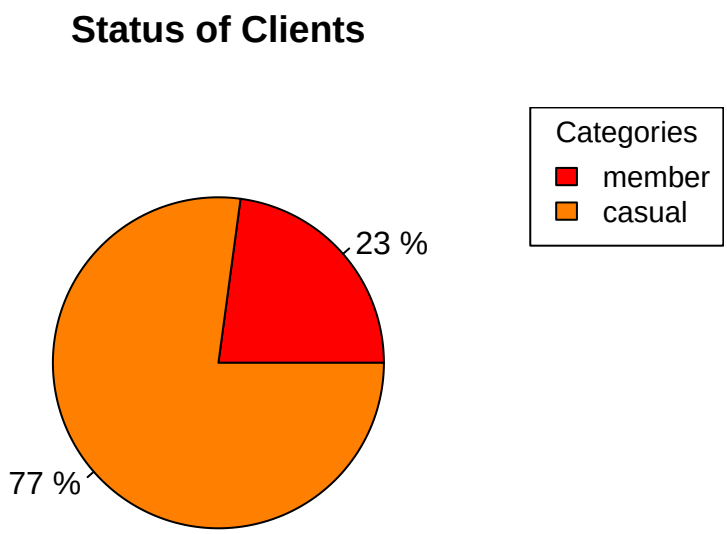


PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Jan_df$Status))*100),"%", sep = " ")
```

```
## [1] "23 %" "77 %"
```

```
#pie plot
pie(table(Jan_df$Status), labels =
      paste(round(prop.table(table(Jan_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients", radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



EXPORT DATA FOR MERGING AND VISUALIZATION

```
#library(rio)
```

```
#write.csv(Jan_df,"C:/Users/LILIAN/Desktop/Exported Datasets from #R/Jan_df.csv")
```

#ANALYSIS FOR THE MONTH OF FEBUARY, 2021

```
# IMPORT DATASETS
Feb <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Feb-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Feb <- subset(Feb, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID,X))
```

```
#TOTAL NUMBER OF ROWS
nrow(Feb)
```

```
## [1] 96834
```

```r
#TOTAL NUMBER OF COLUMNS
ncol(Feb)
```

```
## [1] 9
```

```r
#duplicate check
duplicated(Feb)
sum(duplicated(Feb))%>%
          head(n = 20)
```

```r
#add new column to the dataset
Feb %>%
          add_column (Time_Diff2 = NA,Lat2 = NA, Ride_Lng2 = NA,eval=FALSE) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
Feb <- Feb %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Feb$Ride_Lng2 = (Feb$Start_Lng - Feb$End_Lng)
Feb$Lat2 = (Feb$Start_Lat - Feb$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Feb$Time_Diff2 = difftime(Feb$Ended_at, Feb$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Feb$Ride_Lng2, na.rm = TRUE)
```

```
## [1] 0.0002088121
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Feb[Feb$Status == 'member','Ride_Lng2'],na.rm = TRUE)
```

```
## [1] 0.0001653809
```

```r
mean(Feb[Feb$Status == 'casual','Ride_Lng2'],na.rm = TRUE)
```

```
## [1] 0.0003975525
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Feb[Jan$Status == 'casual','Time_Diff2'],na.rm = TRUE)
```

```
## Time difference of 86984.58 secs
```

```
mean(Feb[Jan$Status == 'member','Time_Diff2'],na.rm = TRUE)
```

```
## Time difference of 49067.42 secs
```

```
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Feb[Feb$Time_Diff2 >= 400,'Ride_Lng2'],na.rm = TRUE)
```

```
## [1] 0.0003045826
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Feb[Feb$Ride_Lng2 >= 0.1,'Time_Diff2'],na.rm = TRUE)
```

```
## Time difference of 2615.172 secs
```

```
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Feb[Feb$Rideable_Type == 'docked_bike','Ride_Lng2'],na.rm = TRUE)
```

```
## [1] 0.0003706937
```

```
mean(Feb[Feb$Rideable_Type == 'electric_bike','Ride_Lng2'],na.rm = TRUE)
```

```
## [1] 0.0002537196
```

```
mean(Feb[Feb$Rideable_Type == 'classic_bike','Ride_Lng2'],na.rm = TRUE)
```

```
## [1] 0.0001791982
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Feb[Feb$Rideable_Type == 'docked_bike','Time_Diff2'],na.rm = TRUE)
```

```
## Time difference of 1274547 secs
```

```
mean(Feb[Feb$Rideable_Type == 'electric_bike','Time_Diff2'],na.rm = TRUE)
```

```
## Time difference of 45554.99 secs
```

```
mean(Feb[Feb$Rideable_Type == 'classic_bike','Time_Diff2'],na.rm = TRUE)
```

```
## Time difference of 22839.93 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALIES
max(Feb$Ride_Lng2, na.rm = TRUE)
```

```
## [1] 0.1736013
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALIES
min(Feb$Ride_Lng2, na.rm = TRUE)
```

```
## [1] -0.138942
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Feb$Ride_Lng2)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Feb$Status)
```

```
##
## casual member
##  18117  78717
```

```r
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Feb$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##         61700          2106         33028
```

```r
#EXTRACT COLUMNS OF INTEREST
Feb_df <- Feb %>%
  select(Ride_ID, Rideable_Type, Time_Diff2, Lat2, Ride_Lng2, Status)
```

```r
#preview of the dataset
head(Feb_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff2       Lat2   Ride_Lng2 Status
## 1 E19E6F1B8D4C42ED electric_bike   600 secs 0.01034067  0.02325700 member
## 2 DC88F20C2C55F27F electric_bike   240 secs 0.00033283 -0.00670700 member
## 3 EC45C94683FE3F27 electric_bike   120 secs 0.00031267  0.00335733 member
```

```r
glimpse(Feb_df)
```

```
## Rows: 96,834
## Columns: 6
## $ Ride_ID       <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683FE3F2~
## $ Rideable_Type <chr> "electric_bike", "electric_bike", "electric_bike", "elec~
## $ Time_Diff2    <drtn> 600 secs, 240 secs, 120 secs, 660 secs, 0 secs, 3180 se~
## $ Lat2          <dbl> 0.01034067, 0.00033283, 0.00031267, -0.01960133, 0.00032~
## $ Ride_Lng2     <dbl> 0.02325700, -0.00670700, 0.00335733, -0.00666217, 0.0033~
## $ Status        <chr> "member", "member", "member", "member", "casual", "casua~
```

```
#preview the data summary
summary(Feb_df)
```

```
##    Ride_ID           Rideable_Type       Time_Diff2            Lat2
##  Length:96834       Length:96834       Length:96834      Min.   :-0.18974
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.00858
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.00000
##                                                          Mean   :-0.00011
##                                                          3rd Qu.: 0.00836
##                                                          Max.   : 0.20444
##                                                          NA's   :103
##    Ride_Lng2           Status
##  Min.   :-0.13894   Length:96834
##  1st Qu.:-0.00845   Class :character
##  Median : 0.00000   Mode  :character
##  Mean   : 0.00021
##  3rd Qu.: 0.00881
##  Max.   : 0.17360
##  NA's   :103
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Feb_df <- filter(Feb_df,Ride_Lng2 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Feb_df <- filter(Feb_df,Ride_Lng2 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Feb_df <- filter(Feb_df,Time_Diff2 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Feb_df <- filter(Feb_df,Time_Diff2 < 400)%>%
          head(n = 20)
```

VISUALIZATION

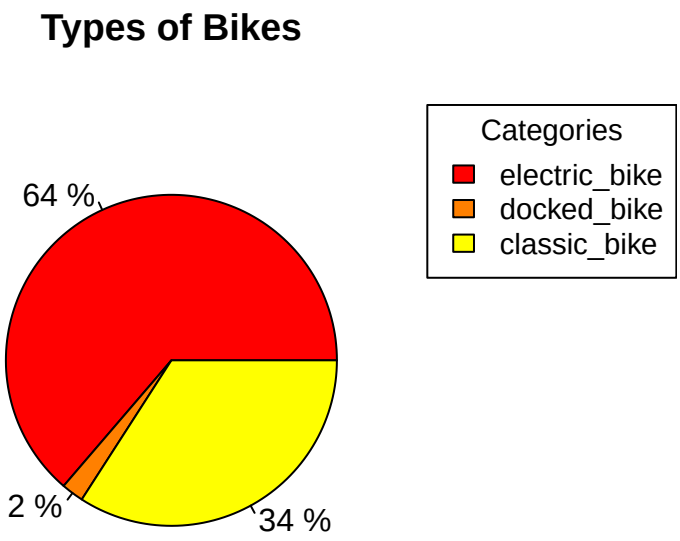
PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Feb_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "64 %" "2 %"  "34 %"
```

```
#pie plot
pie(table(Feb_df$Rideable_Type), labels =
      paste(round(prop.table(table(Feb_df$Rideable_Type))*100),"%", sep = " "),
```

```
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

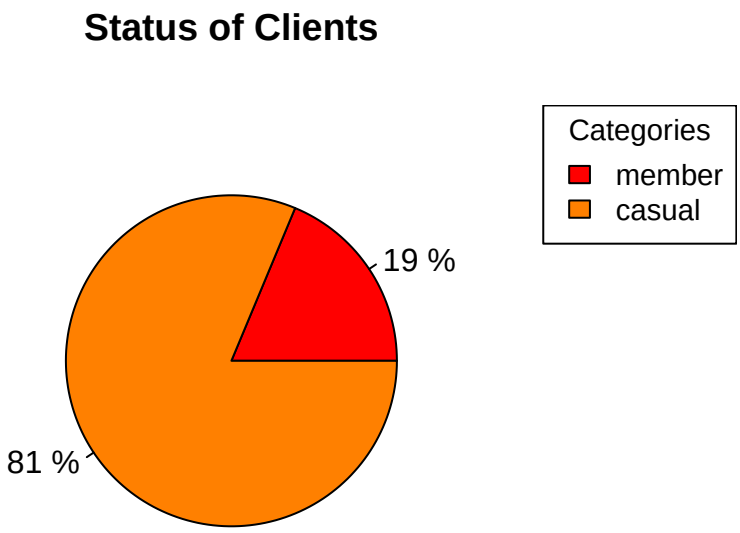


PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Feb_df$Status))*100),"%", sep = " ")
```

```
## [1] "19 %" "81 %"
```

```
#pie plot
pie(table(Feb_df$Status), labels =
      paste(round(prop.table(table(Feb_df$Status))*100),"%", sep = " "),
   col = heat.colors(3), main = "Status of Clients", radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



ANALYSIS FOR THE MONTH OF MARCH, 2021

```
# IMPORT DATASETS
Mar <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Mar-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Mar <- subset(Mar, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID,X))
```

```
#TOTAL NUMBER OF ROWS
nrow(Mar)
```

```
## [1] 49622
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Mar)
```

```
## [1] 9
```

```
#duplicate check
duplicated(Mar)
sum(duplicated(Mar))%>%
          head(n = 20)
```

```
#add new column to the dataset
Mar %>%
          add_column (Time_Diff3 = NA,Lat2 = NA, Ride_Lng3 = NA,eval=FALSE) %>%
          head(n = 20)
```

```
#RENAME COLUMNS
Mar <- Mar %>%
  rename(Status = Member_Casual)
```

```
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Mar$Ride_Lng3 = (Mar$Start_Lng - Mar$End_Lng)
Mar$Lat3 = (Mar$Start_Lat - Mar$End_Lat)
```

```
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Mar$Time_Diff3 = difftime(Mar$Ended_at, Mar$Started_at, units = "secs")
```

```
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Mar$Ride_Lng3, na.rm = TRUE)
```

```
## [1] 0.0002843936
```

```
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Mar[Mar$Status == 'member','Ride_Lng3'],na.rm = TRUE)
```

```
## [1] 0.000189369
```

```
mean(Mar[Mar$Status == 'casual','Ride_Lng3'],na.rm = TRUE)
```

```
## [1] 0.0006554641
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Mar[Mar$Status == 'casual','Time_Diff3'],na.rm = TRUE)
```

```
## Time difference of -6376.139 secs
```

```
mean(Mar[Mar$Status == 'member','Time_Diff3'],na.rm = TRUE)
```

```
## Time difference of 67200.02 secs
```

```
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Mar[Mar$Time_Diff3 >= 400,'Ride_Lng3'],na.rm = TRUE)
```

```
## [1] 0.0003554766
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Mar[Mar$Ride_Lng3 >= 0.1,'Time_Diff3'],na.rm = TRUE)
```

```
## Time difference of 3608.571 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Mar[Mar$Rideable_Type == 'docked_bike','Ride_Lng3'],na.rm = TRUE)
```

```
## [1] 0.00100732
```

```r
mean(Mar[Mar$Rideable_Type == 'electric_bike','Ride_Lng3'],na.rm = TRUE)
```

```
## [1] 0.0004532967
```

```r
mean(Mar[Mar$Rideable_Type == 'classic_bike','Ride_Lng3'],na.rm = TRUE)
```

```
## [1] 0.0001932436
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Mar[Mar$Rideable_Type == 'docked_bike','Time_Diff3'],na.rm = TRUE)
```

```
## Time difference of -980698.3 secs
```

```r
mean(Mar[Mar$Rideable_Type == 'electric_bike','Time_Diff3'],na.rm = TRUE)
```

```
## Time difference of -29808.95 secs
```

```r
mean(Mar[Mar$Rideable_Type == 'classic_bike','Time_Diff3'],na.rm = TRUE)
```

```
## Time difference of 120909.6 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALIES
max(Mar$Ride_Lng3, na.rm = TRUE)
```

```
## [1] 0.139261
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALIES
min(Mar$Ride_Lng3, na.rm = TRUE)
```

```
## [1] -0.144819
```

```
#CALCULATE THE MODE OF RIDE LENGHT
mode(Mar$Ride_Lng3)
```

```
## [1] "numeric"
```

```
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Mar$Status)
```

```
##
## casual member
##  10131  39491
```

```
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Mar$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##         35012          1271         13339
```

```
#EXTRACT COLUMNS OF INTEREST
Mar_df <- Mar %>%
  select(Ride_ID, Rideable_Type, Time_Diff3, Lat3, Ride_Lng3, Status)
```

```
#preview of the dataset
head(Mar_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff3        Lat3 Ride_Lng3 Status
## 1 89E7AA6C29227EFF  classic_bike   420 secs  0.00811800 -0.004652 member
## 2 0FEFDE2603568365  classic_bike  1200 secs -0.00683600  0.003505 casual
## 3 E6159D746B2DBB91 electric_bike   540 secs  0.00092867 -0.003603 member
```

```
glimpse(Mar_df)
```

```
## Rows: 49,622
## Columns: 6
## $ Ride_ID       <chr> "89E7AA6C29227EFF", "0FEFDE2603568365", "E6159D746B2DBB9~
## $ Rideable_Type <chr> "classic_bike", "classic_bike", "electric_bike", "classi~
## $ Time_Diff3    <drtn> 420 secs, 1200 secs, 540 secs, 300 secs, 900 secs, 360 ~
## $ Lat3          <dbl> 0.00811800, -0.00683600, 0.00092867, -0.00748500, -0.003~
## $ Ride_Lng3     <dbl> -0.00465200, 0.00350500, -0.00360300, 0.00186600, 0.0192~
## $ Status        <chr> "member", "casual", "member", "member", "member", "casua~
```

```
summary(Mar_df)
```

```
##    Ride_ID          Rideable_Type       Time_Diff3            Lat3
##  Length:49622       Length:49622       Length:49622      Min.   :-0.20112
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.00886
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.00000
##                                                          Mean   :-0.00017
```

```
##                                                        3rd Qu.: 0.00868
##                                                        Max.   : 0.18837
##                                                        NA's   :214
##    Ride_Lng3           Status
##  Min.   :-0.14482   Length:49622
##  1st Qu.:-0.00848   Class :character
##  Median : 0.00000   Mode  :character
##  Mean   : 0.00028
##  3rd Qu.: 0.00905
##  Max.   : 0.13926
##  NA's   :214
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Mar_df <- filter(Mar_df,Ride_Lng3 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Mar_df <- filter(Mar_df,Ride_Lng3 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Mar_df <- filter(Mar_df,Time_Diff3 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Mar_df <- filter(Mar_df,Time_Diff3 < 400)%>%
          head(n = 20)
```

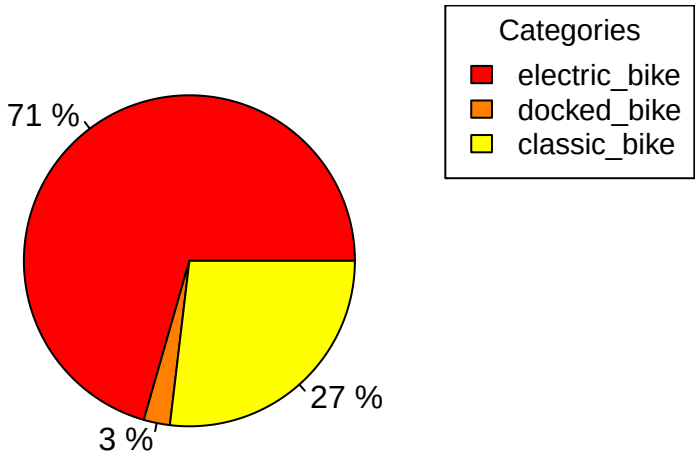
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Mar_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "71 %" "3 %"  "27 %"
```

```
#pie plot
pie(table(Mar_df$Rideable_Type), labels =
      paste(round(prop.table(table(Mar_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

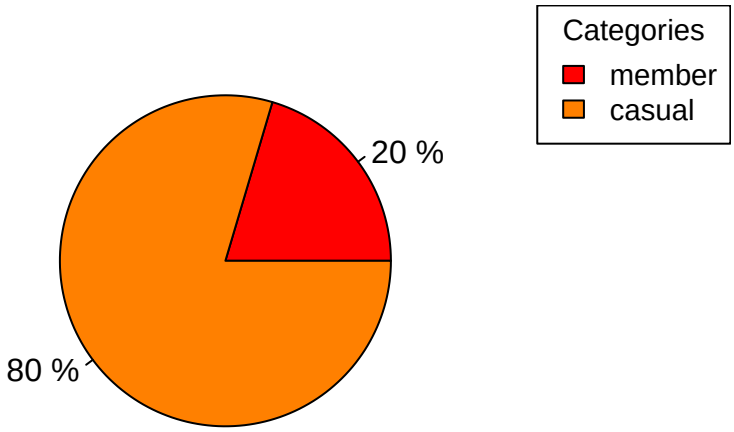


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Mar_df$Status))*100),"%", sep = " ")
```

```
## [1] "20 %" "80 %"
```

```
#pie plot
pie(table(Mar_df$Status), labels =
      paste(round(prop.table(table(Mar_df$Status))*100),"%", sep = " "),
   col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#EXPORTATION OF DATASETS FOR MERGING AND VISUALIZATION

```
#write.csv(Mar_df,"C:/Users/LILIAN/Desktop/Exported Datasets from #R/Mar_df.csv")
```

#ANALYSIS FOR THE MONTH OF APRIL, 2021

```
# IMPORT DATASETS
Apr <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/April-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Apr <- subset(Apr, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID))
```

```
#TOTAL NUMBER OF ROWS
nrow(Apr)
```

```
## [1] 214272
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Apr)
```

```
## [1] 9
```

```r
#duplicate check
duplicated(Apr)
sum(duplicated(Apr))%>%
          head(n = 20)
```

```r
#add new column to the dataset
Apr %>% add_column (Time_Diff4 = NA,Lat4 = NA, Ride_Lng4 = NA,eval=FALSE) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
Apr <- Apr %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Apr$Ride_Lng4 = (Apr$Start_Lng - Apr$End_Lng)
Apr$Lat4 = (Apr$Start_Lat - Apr$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Apr$Time_Diff4 = difftime(Apr$Ended_at, Apr$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Apr$Ride_Lng4, na.rm = TRUE)
```

```
## [1] 0.0001724059
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Apr[Apr$Status == 'member','Ride_Lng4'],na.rm = TRUE)
```

```
## [1] 0.0001394471
```

```r
mean(Apr[Apr$Status == 'casual','Ride_Lng4'],na.rm = TRUE)
```

```
## [1] 0.0002292261
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Apr[Apr$Status == 'casual','Time_Diff4'],na.rm = TRUE)
```

```
## Time difference of 185823.5 secs
```

```r
mean(Apr[Apr$Status == 'member','Time_Diff4'],na.rm = TRUE)
```

```
## Time difference of -14708.17 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Apr[Apr$Time_Diff4 >= 400,'Ride_Lng4'],na.rm = TRUE)
```

```
## [1] 0.0002030166
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Apr[Apr$Ride_Lng3 >= 0.1,'Time_Diff4'],na.rm = TRUE)
```

```
## Time difference of NaN secs
```

```
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Apr[Apr$Rideable_Type == 'docked_bike','Ride_Lng4'],na.rm = TRUE)
```

```
## [1] -9.206096e-05
```

```
mean(Apr[Apr$Rideable_Type == 'electric_bike','Ride_Lng4'],na.rm = TRUE)
```

```
## [1] 0.000256429
```

```
mean(Apr[Apr$Rideable_Type == 'classic_bike','Ride_Lng4'],na.rm = TRUE)
```

```
## [1] 0.0001651722
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Apr[Apr$Rideable_Type == 'docked_bike','Time_Diff4'],na.rm = TRUE)
```

```
## Time difference of 333857.6 secs
```

```
mean(Apr[Apr$Rideable_Type == 'electric_bike','Time_Diff4'],na.rm = TRUE)
```

```
## Time difference of 17368.2 secs
```

```
mean(Apr[Apr$Rideable_Type == 'classic_bike','Time_Diff4'],na.rm = TRUE)
```

```
## Time difference of 47888.5 secs
```

```
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALIES
max(Apr$Ride_Lng4, na.rm = TRUE)
```

```
## [1] 0.3532472
```

```
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALIES
min(Apr$Ride_Lng4, na.rm = TRUE)
```

```
## [1] -0.1391249
```

```
#CALCULATE THE MODE OF RIDE LENGHT
mode(Apr$Ride_Lng4)
```

```
## [1] "numeric"
```

```
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Apr$Status)
```

```
##
## casual member
##  78711 135561
```

```
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Apr$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        142115         14451         57706
```

```
#EXTRACT COLUMNS OF INTEREST
Apr_df <- Apr %>%
  select(Ride_ID, Rideable_Type, Time_Diff4, Lat4, Ride_Lng4, Status)
```

```
#preview of the dataset
head(Apr_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff4  Lat4 Ride_Lng4 Status
## 1 34E6463B89188D1C electric_bike   480 secs  0.00     -0.03 member
## 2 F0B55B0D2DD86D3A electric_bike  1020 secs -0.01      0.06 member
## 3 77D7653614151D4E electric_bike   840 secs -0.01      0.03 member
```

```
#preview the data summary
summary(Apr)
```

```
##    Ride_ID          Rideable_Type       Started_at          Ended_at
##  Length:214272      Length:214272      Length:214272      Length:214272
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    Start_Lat       Start_Lng         End_Lat         End_Lng
##  Min.   :41.65   Min.   :-87.78   Min.   :41.64   Min.   :-88.07
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.90   Median :-87.64   Median :41.90   Median :-87.64
##  Mean   :41.90   Mean   :-87.65   Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.07   Max.   :-87.53   Max.   :42.08   Max.   :-87.53
##                                   NA's   :165     NA's   :165
```

```
##     Status            Ride_Lng4             Lat4           Time_Diff4
##  Length:214272      Min.   :-0.13912   Min.   :-0.20246   Length:214272
##  Class :character   1st Qu.:-0.00881   1st Qu.:-0.00950   Class :difftime
##  Mode  :character   Median : 0.00000   Median : 0.00000   Mode  :numeric
##                     Mean   : 0.00017   Mean   :-0.00025
##                     3rd Qu.: 0.00920   3rd Qu.: 0.00917
##                     Max.   : 0.35325   Max.   : 0.20246
##                     NA's   :165        NA's   :165
```

```
glimpse(Apr)
```

```
## Rows: 214,272
## Columns: 12
## $ Ride_ID       <chr> "34E6463B89188D1C", "F0B55B0D2DD86D3A", "77D7653614151D4~
## $ Rideable_Type <chr> "electric_bike", "electric_bike", "electric_bike", "elec~
## $ Started_at    <chr> "04-03-21 11:26", "08-03-21 12:00", "23-03-21 16:30", "3~
## $ Ended_at      <chr> "04-03-21 11:34", "08-03-21 12:17", "23-03-21 16:44", "3~
## $ Start_Lat     <dbl> 41.92000, 41.92000, 41.94000, 41.80000, 41.82000, 41.830~
## $ Start_Lng     <dbl> -87.75000, -87.69000, -87.68000, -87.59000, -87.62000, -~
## $ End_Lat       <dbl> 41.92000, 41.93000, 41.95000, 41.80000, 41.83000, 41.740~
## $ End_Lng       <dbl> -87.72000, -87.75000, -87.71000, -87.59000, -87.61000, -~
## $ Status        <chr> "member", "member", "member", "member", "member", "membe~
## $ Ride_Lng4     <dbl> -0.03, 0.06, 0.03, 0.00, -0.01, 0.03, 0.00, 0.03, -0.01,~
## $ Lat4          <dbl> 0.00, -0.01, -0.01, 0.00, -0.01, 0.09, 0.06, -0.02, 0.00~
## $ Time_Diff4    <drtn> 480 secs, 1020 secs, 840 secs, 1020 secs, 2400 secs, 28~
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Apr_df <- filter(Apr_df,Ride_Lng4 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Apr_df <- filter(Apr_df,Ride_Lng4 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Apr_df <- filter(Apr_df,Time_Diff4 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Apr_df <- filter(Apr_df,Time_Diff4 < 400)%>%
          head(n = 20)
```

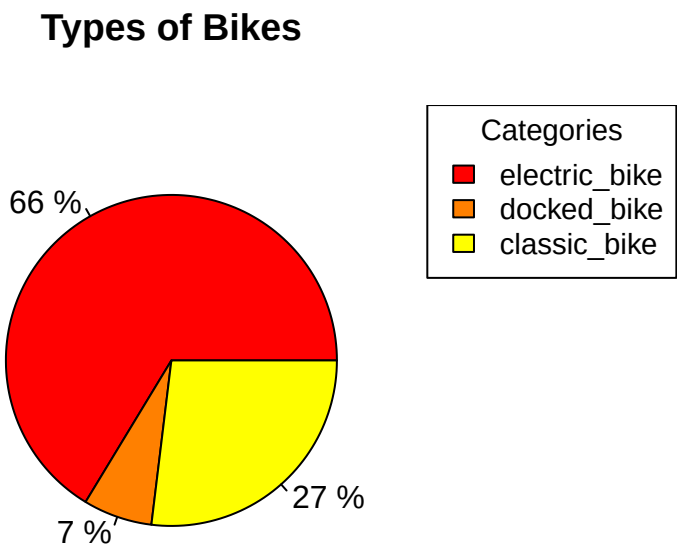
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Apr_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "66 %" "7 %"  "27 %"
```

```
#pie plot
pie(table(Apr_df$Rideable_Type), labels =
      paste(round(prop.table(table(Apr_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes", radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

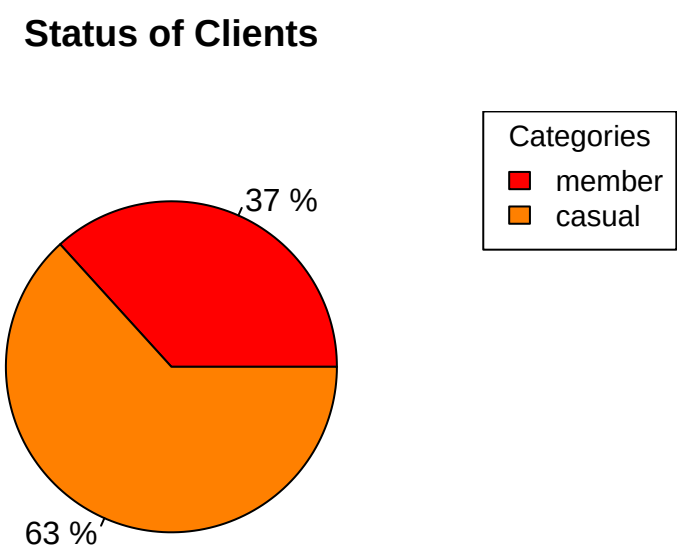


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Apr_df$Status))*100),"%", sep = " ")
```

```
## [1] "37 %" "63 %"
```

```
#pie plot
pie(table(Apr_df$Status), labels =
      paste(round(prop.table(table(Apr_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#ANALYSIS FOR THE MONTH OF MAY, 2021

```
# IMPORT DATASETS
May <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/May-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
May <- subset(May, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID))
```

```
#TOTAL NUMBER OF ROWS
nrow(May)
```

```
## [1] 337230
```

```
#TOTAL NUMBER OF COLUMNS
ncol(May)
```

```
## [1] 9
```

```
#duplicate check
duplicated(May)
sum(duplicated(May))%>%
          head(n = 20)
```

```r
#add new column to the dataset
May %>% add_column (Time_Diff5 = NA,Lat5 = NA, Ride_Lng5 = NA,eval=FALSE) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
May <- May %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
May$Ride_Lng5 = (May$Start_Lng - May$End_Lng)
May$Lat5 = (May$Start_Lat - May$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
May$Time_Diff5 = difftime(May$Ended_at, May$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(May$Ride_Lng5, na.rm = TRUE)
```

```
## [1] 0.0002859008
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(May[May$Status == 'member','Ride_Lng5'],na.rm = TRUE)
```

```
## [1] 0.0002674697
```

```r
mean(May[May$Status == 'casual','Ride_Lng5'],na.rm = TRUE)
```

```
## [1] 0.0003129972
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(May[May$Status == 'casual','Time_Diff5'],na.rm = TRUE)
```

```
## Time difference of 55305.65 secs
```

```r
mean(May[May$Status == 'member','Time_Diff5'],na.rm = TRUE)
```

```
## Time difference of -47741.51 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(May[May$Time_Diff5 >= 400,'Ride_Lng5'],na.rm = TRUE)
```

```
## [1] 0.0003523027
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(May[May$Ride_Lng5 >= 0.1,'Time_Diff5'],na.rm = TRUE)
```

```
## Time difference of 392713.3 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(May[May$Rideable_Type == 'docked_bike','Ride_Lng5'],na.rm = TRUE)
```

```
## [1] 0.0003866173
```

```r
mean(May[May$Rideable_Type == 'electric_bike','Ride_Lng5'],na.rm = TRUE)
```

```
## [1] 0.0001934047
```

```r
mean(May[May$Rideable_Type == 'classic_bike','Ride_Lng5'],na.rm = TRUE)
```

```
## [1] 0.0003165326
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(May[May$Rideable_Type == 'docked_bike','Time_Diff5'],na.rm = TRUE)
```

```
## Time difference of 187171.7 secs
```

```r
mean(May[May$Rideable_Type == 'electric_bike','Time_Diff5'],na.rm = TRUE)
```

```
## Time difference of -38841.94 secs
```

```r
mean(May[May$Rideable_Type == 'classic_bike','Time_Diff5'],na.rm = TRUE)
```

```
## Time difference of -13264.33 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALIES
max(May$Ride_Lng5,na.rm = TRUE)
```

```
## [1] 0.2057282
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALIES
min(May$Ride_Lng5,na.rm = TRUE)
```

```
## [1] -0.133806
```

```
#CALCULATE THE MODE OF RIDE LENGHT
mode(May$Ride_Lng5)
```

```
## [1] "numeric"
```

```
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(May$Status)
```

```
##
## casual member
## 136601 200629
```

```
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(May$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        214619         24714         97897
```

```
#EXTRACT COLUMNS OF INTEREST
May_df <- May %>%
   select(Ride_ID, Rideable_Type, Time_Diff5, Lat5, Ride_Lng5, Status)
```

```
#preview of the dataset
head(May_df,3)
```

```
##            Ride_ID Rideable_Type     Time_Diff5      Lat5 Ride_Lng5 Status
## 1 6C992BD37A98A63F  classic_bike      1860 secs -0.050702  0.035218 member
## 2 1E0145613A209000   docked_bike      3840 secs  0.000000  0.000000 casual
## 3 E498E15508A80BAD   docked_bike 126226680 secs  0.000000  0.000000 casual
```

```
glimpse(May_df)
```

```
## Rows: 337,230
## Columns: 6
## $ Ride_ID       <chr> "6C992BD37A98A63F", "1E0145613A209000", "E498E15508A80BA~
## $ Rideable_Type <chr> "classic_bike", "docked_bike", "docked_bike", "classic_b~
## $ Time_Diff5    <drtn> 1860 secs, 3840 secs, 126226680 secs, 1500 secs, 5460 s~
## $ Lat5          <dbl> -0.05070200, 0.00000000, 0.00000000, -0.04503100, 0.0000~
## $ Ride_Lng5     <dbl> 0.03521800, 0.00000000, 0.00000000, -0.00999500, 0.00000~
## $ Status        <chr> "member", "casual", "casual", "member", "casual", "casua~
```

```
summary(May_df)
```

```
##    Ride_ID          Rideable_Type       Time_Diff5            Lat5
##  Length:337230      Length:337230      Length:337230     Min.   :-0.26175
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.00994
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.00000
##                                                          Mean   :-0.00038
```

```
##                                                        3rd Qu.: 0.00922
##                                                        Max.   : 0.26539
##                                                        NA's   :267
##    Ride_Lng5           Status
##  Min.   :-0.13381   Length:337230
##  1st Qu.:-0.00877   Class :character
##  Median : 0.00000   Mode  :character
##  Mean   : 0.00029
##  3rd Qu.: 0.00944
##  Max.   : 0.20573
##  NA's   :267
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_May_df <- filter(May_df,Ride_Lng5 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_May_df <- filter(May_df,Ride_Lng5 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_May_df <- filter(May_df,Time_Diff5 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_May_df <- filter(May_df,Time_Diff5 < 400)%>%
          head(n = 20)
```

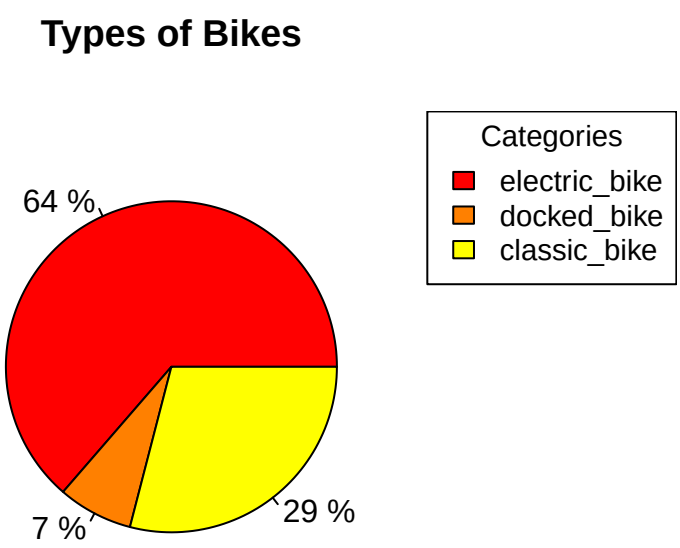
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(May_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "64 %" "7 %"  "29 %"
```

```
#pie plot
pie(table(May_df$Rideable_Type), labels =
      paste(round(prop.table(table(May_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

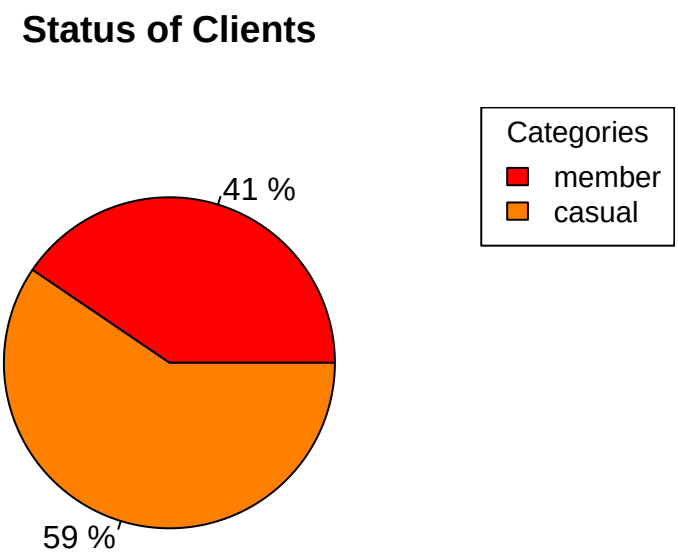


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(May_df$Status))*100),"%", sep = " ")
```

```
## [1] "41 %" "59 %"
```

```
#pie plot
pie(table(May_df$Status), labels =
      paste(round(prop.table(table(May_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#ANALYSIS FOR THE MONTH OF JUNE, 2021

```
# IMPORT DATASETS
Jun <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/June-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Jun <- subset(Jun, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,End_Station_ID,X))
```

```
#TOTAL NUMBER OF ROWS
nrow(Jun)
```

```
## [1] 337230
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Jun)
```

```
## [1] 9
```

```
#duplicate check
duplicated(Jun)
sum(duplicated(Jun))%>%
          head(n = 20)
```

```r
#add new column to the dataset
Jun %>% add_column (Time_Diff6 = NA,Lat6 = NA, Ride_Lng6 = NA) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
Jun <- Jun %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Jun$Ride_Lng6 = (Jun$Start_Lng - Jun$End_Lng)
Jun$Lat6 = (Jun$Start_Lat - Jun$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Jun$Time_Diff6 = difftime(Jun$Ended_at, Jun$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Jun$Ride_Lng6, na.rm = TRUE)
```

```
## [1] 0.0002859008
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Jun[Jun$Status == 'member','Ride_Lng6'],na.rm = TRUE)
```

```
## [1] 0.0002674697
```

```r
mean(Jun[Jun$Status == 'casual','Ride_Lng6'],na.rm = TRUE)
```

```
## [1] 0.0003129972
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #STATUS COLUMN IS EQUALS
mean(Jun[Jun$Status == 'casual','Time_Diff6'],na.rm = TRUE)
```

```
## Time difference of 55305.65 secs
```

```r
mean(Jun[Jun$Status == 'member','Time_Diff6'],na.rm = TRUE)
```

```
## Time difference of -47741.51 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Jun[Jun$Time_Diff6 >= 400,'Ride_Lng6'],na.rm = TRUE)
```

```
## [1] 0.0003523027
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Jun[Jun$Ride_Lng6 >= 0.1,'Time_Diff6'],na.rm = TRUE)
```

```
## Time difference of 392713.3 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Jun[Jun$Rideable_Type == 'docked_bike','Ride_Lng6'],na.rm = TRUE)
```

```
## [1] 0.0003866173
```

```r
mean(Jun[Jun$Rideable_Type == 'electric_bike','Ride_Lng6'],na.rm = TRUE)
```

```
## [1] 0.0001934047
```

```r
mean(Jun[Jun$Rideable_Type == 'classic_bike','Ride_Lng6'],na.rm = TRUE)
```

```
## [1] 0.0003165326
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #RIDEABLE TYPE COLUMN IS
mean(Jun[Jun$Rideable_Type == 'docked_bike','Time_Diff6'],na.rm = TRUE)
```

```
## Time difference of 187171.7 secs
```

```r
mean(Jun[Jun$Rideable_Type == 'electric_bike','Time_Diff6'],na.rm = TRUE)
```

```
## Time difference of -38841.94 secs
```

```r
mean(Jun[Jun$Rideable_Type == 'classic_bike','Time_Diff6'],na.rm = TRUE)
```

```
## Time difference of -13264.33 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALIES
max(Jun$Ride_Lng6, na.rm = TRUE)
```

```
## [1] 0.2057282
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALIES
min(Jun$Ride_Lng6, na.rm = TRUE)
```

```
## [1] -0.133806
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Jun$Ride_Lng6)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Jun$Status)
```

```
##
## casual member
## 136601 200629
```

```r
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Jun$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        214619         24714         97897
```

```r
#EXTRACT COLUMNS OF INTEREST
Jun_df <- Jun %>%
  select(Ride_ID, Rideable_Type, Time_Diff6, Lat6, Ride_Lng6, Status)
```

```r
#preview of the dataset
head(Jun_df,3)
```

```
##            Ride_ID Rideable_Type     Time_Diff6      Lat6 Ride_Lng6 Status
## 1 6C992BD37A98A63F  classic_bike      1860 secs -0.050702  0.035218 member
## 2 1E0145613A209000   docked_bike      3840 secs  0.000000  0.000000 casual
## 3 E498E15508A80BAD   docked_bike 126226680 secs  0.000000  0.000000 casual
```

```r
glimpse(Jun_df)
```

```
## Rows: 337,230
## Columns: 6
## $ Ride_ID       <chr> "6C992BD37A98A63F", "1E0145613A209000", "E498E15508A80BA~
## $ Rideable_Type <chr> "classic_bike", "docked_bike", "docked_bike", "classic_b~
## $ Time_Diff6    <drtn> 1860 secs, 3840 secs, 126226680 secs, 1500 secs, 5460 s~
## $ Lat6          <dbl> -0.05070200, 0.00000000, 0.00000000, -0.04503100, 0.0000~
## $ Ride_Lng6     <dbl> 0.03521800, 0.00000000, 0.00000000, -0.00999500, 0.00000~
## $ Status        <chr> "member", "casual", "casual", "member", "casual", "casua~
```

```r
#preview the data summary
summary(Jun_df)
```

```
##    Ride_ID          Rideable_Type       Time_Diff6            Lat6
##  Length:337230      Length:337230      Length:337230     Min.   :-0.26175
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.00994
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.00000
##                                                          Mean   :-0.00038
##                                                          3rd Qu.: 0.00922
##                                                          Max.   : 0.26539
##                                                          NA's   :267
##    Ride_Lng6           Status
##  Min.   :-0.13381   Length:337230
##  1st Qu.:-0.00877   Class :character
##  Median : 0.00000   Mode  :character
##  Mean   : 0.00029
##  3rd Qu.: 0.00944
##  Max.   : 0.20573
##  NA's   :267
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Jun_df <- filter(Jun_df,Ride_Lng6 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Jun_df <- filter(Jun_df,Ride_Lng6 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Jun_df <- filter(Jun_df,Time_Diff6 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Jun_df <- filter(Jun_df,Time_Diff6 < 400)%>%
          head(n = 20)
```

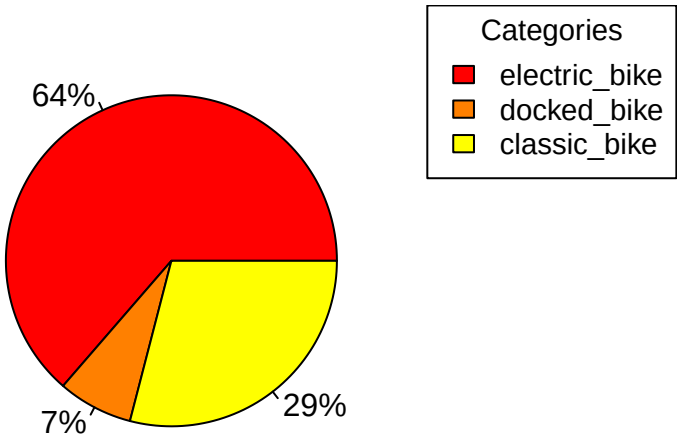
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Jun_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "64 %" "7 %"  "29 %"
```

```
#pie plot
pie(table(Jun_df$Rideable_Type), labels =
      paste(round(prop.table(table(Jun_df$Rideable_Type))*100),"%", sep = ""),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

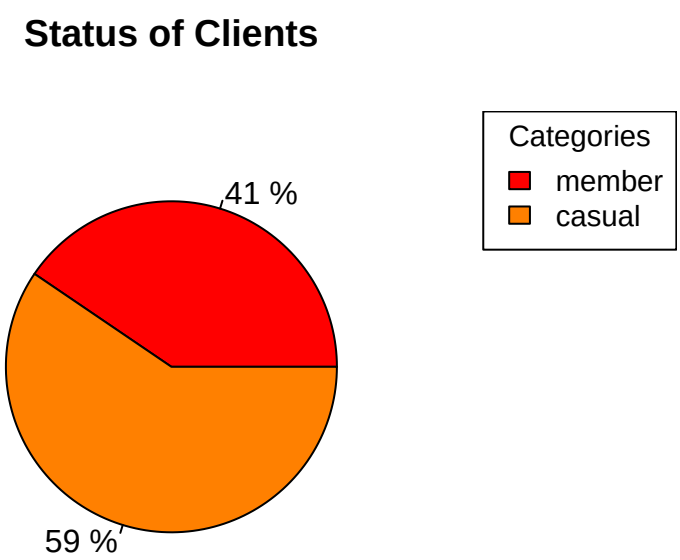


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Jun_df$Status))*100),"%", sep = " ")
```

```
## [1] "41 %" "59 %"
```

```
#pie plot
pie(table(Jun_df$Status), labels =
     paste(round(prop.table(table(Jun_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#ANALYSIS FOR THE MONTH OF JUly, 2021

```
# IMPORT DATASETS
Jul <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/July-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Jul <- subset(Jul, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,End_Station_ID,X))
```

```
#TOTAL NUMBER OF ROWS
nrow(Jul)
```

```
## [1] 531633
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Jul)
```

```
## [1] 9
```

```
#add new column to the dataset
Jul %>% add_column (Time_Diff7 = NA,Lat7 = NA, Ride_Lng7 = NA) %>%
          head(n = 20)
```

```
#RENAME COLUMNS
Jul <- Jul %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Jul$Ride_Lng7 = (Jul$Start_Lng - Jul$End_Lng)
Jul$Lat7 = (Jul$Start_Lat - Jul$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Jul$Time_Diff7 = difftime(Jul$Ended_at, Jul$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Jul$Ride_Lng7, na.rm = TRUE)
```

```
## [1] 0.0002723528
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Jul[Jul$Status == 'member','Ride_Lng7'],na.rm = TRUE)
```

```
## [1] 0.0002457503
```

```r
mean(Jul[Jul$Status == 'casual','Ride_Lng7'],na.rm = TRUE)
```

```
## [1] 0.0003008244
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #STATUS COLUMN IS EQUALS
mean(Jul[Jul$Status == 'casual','Time_Diff7'],na.rm = TRUE)
```

```
## Time difference of -31134.71 secs
```

```r
mean(Jul[Jul$Status == 'member','Time_Diff7'],na.rm = TRUE)
```

```
## Time difference of 18743.86 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Jul[Jul$Time_Diff7 >= 400,'Ride_Lng7'],na.rm = TRUE)
```

```
## [1] 0.0003243817
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Jul[Jul$Ride_Lng7 >= 0.1,'Time_Diff7'],na.rm = TRUE)
```

```
## Time difference of 6414854 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Jul[Jul$Rideable_Type == 'docked_bike','Ride_Lng7'],na.rm = TRUE)
```

```
## [1] 0.0002067668
```

```r
mean(Jul[Jul$Rideable_Type == 'electric_bike','Ride_Lng7'],na.rm = TRUE)
```

```
## [1] 0.0003361215
```

```r
mean(Jul[Jul$Rideable_Type == 'classic_bike','Ride_Lng7'],na.rm = TRUE)
```

```
## [1] 0.0002445433
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #RIDEABLE TYPE COLUMN IS
mean(Jul[Jul$Rideable_Type == 'docked_bike','Time_Diff7'],na.rm = TRUE)
```

```
## Time difference of -251299.1 secs
```

```r
mean(Jul[Jul$Rideable_Type == 'electric_bike','Time_Diff7'],na.rm = TRUE)
```

```
## Time difference of 26884.62 secs
```

```r
mean(Jul[Jul$Rideable_Type == 'classic_bike','Time_Diff7'],na.rm = TRUE)
```

```
## Time difference of 10441.68 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALIES
max(Jul$Ride_Lng7, na.rm = TRUE)
```

```
## [1] 0.2147897
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALIES
min(Jul$Ride_Lng7, na.rm = TRUE)
```

```
## [1] -0.164488
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Jul$Ride_Lng7)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Jul$Status)
```

```
##
## casual member
## 256916 274717
```

```r
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Jul$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        309093         43353        179187
```

```r
#EXTRACT COLUMNS OF INTEREST
Jul_df <- Jul %>%
  select(Ride_ID, Rideable_Type, Time_Diff7, Lat7, Ride_Lng7, Status)
```

```r
#preview of the dataset
head(Jul_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff7 Lat7 Ride_Lng7 Status
## 1 C809ED75D6160B2A electric_bike   720 secs 0.01     -0.02 casual
## 2 DD59FDCE0ACACAF3 electric_bike  2700 secs 0.09     -0.04 casual
## 3 0AB83CB88C43EFC2 electric_bike    60 secs 0.00      0.00 casual
```

```r
glimpse(Jul_df)
```

```
## Rows: 531,633
## Columns: 6
## $ Ride_ID       <chr> "C809ED75D6160B2A", "DD59FDCE0ACACAF3", "0AB83CB88C43EFC~
## $ Rideable_Type <chr> "electric_bike", "electric_bike", "electric_bike", "elec~
## $ Time_Diff7    <drtn> 720 secs, 2700 secs, 60 secs, 960 secs, 420 secs, 1440 ~
## $ Lat7          <dbl> 0.0100000, 0.0900000, 0.0000000, -0.0200000, 0.0000000, ~
## $ Ride_Lng7     <dbl> -0.02000000, -0.04000000, 0.00000000, -0.01000000, 0.010~
## $ Status        <chr> "casual", "casual", "casual", "casual", "casual", "casua~
```

```r
summary(Jul_df)
```

```
##    Ride_ID          Rideable_Type       Time_Diff7            Lat7
##  Length:531633      Length:531633      Length:531633     Min.   :-0.2536
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.0100
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.0000
##                                                          Mean   :-0.0004
##                                                          3rd Qu.: 0.0096
##                                                          Max.   : 0.3036
##                                                          NA's   :452
##    Ride_Lng7          Status
##  Min.   :-0.1645   Length:531633
##  1st Qu.:-0.0092   Class :character
##  Median : 0.0000   Mode  :character
##  Mean   : 0.0003
##  3rd Qu.: 0.0097
##  Max.   : 0.2148
##  NA's   :452
```

```r
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Jul_df <- filter(Jul_df,Ride_Lng7 >= 0.1)%>%
          head(n = 20)
```

```r
#for Less than 0.1
LessRideLength_Jul_df <- filter(Jul_df,Ride_Lng7 < 0.1)%>%
          head(n = 20)
```

```r
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Jul_df <- filter(Jul_df,Time_Diff7 >= 400)%>%
          head(n = 20)
```

```r
LessTimeDiff_Jul_df <- filter(Jul_df,Time_Diff7 < 400)%>%
          head(n = 20)
```

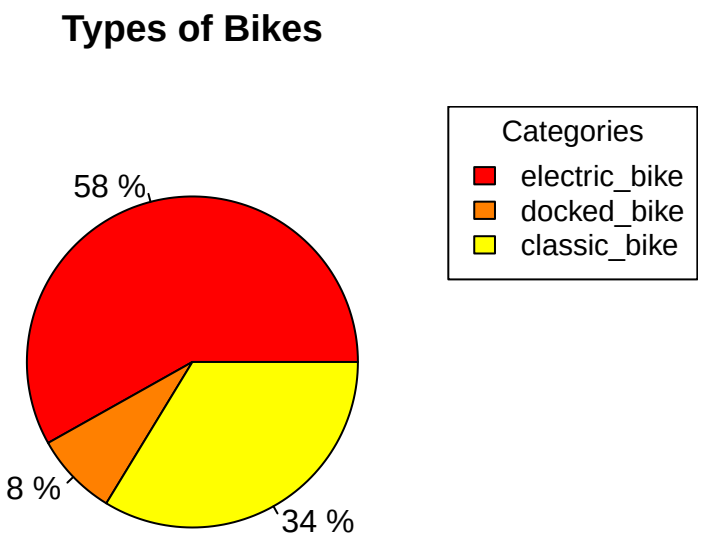
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```r
#To get the % of each variable
paste(round(prop.table(table(Jul_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "58 %" "8 %"  "34 %"
```

```r
#pie plot
pie(table(Jul_df$Rideable_Type), labels =
      paste(round(prop.table(table(Jul_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike","classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

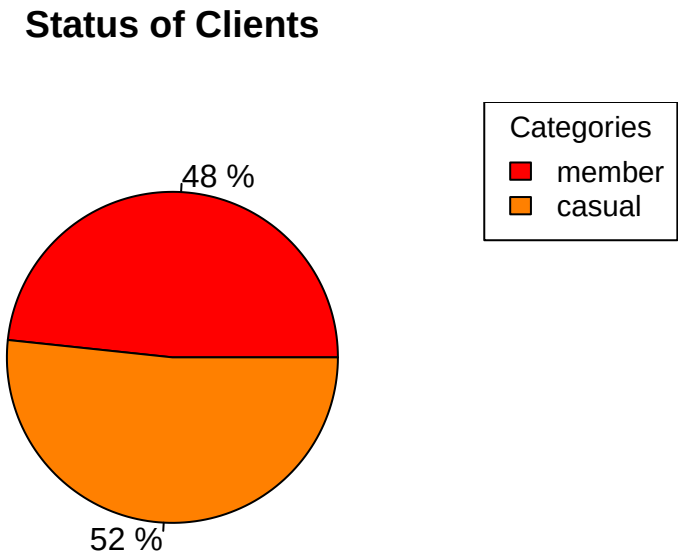


#PLOTTING PIE CHART FOR Status COLUMNS

```r
#To get the % of each variable
paste(round(prop.table(table(Jul_df$Status))*100),"%", sep = " ")
```

```
## [1] "48 %" "52 %"
```

```r
#pie plot
pie(table(Jul_df$Status), labels =
      paste(round(prop.table(table(Jul_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```




```r
# IMPORT DATASETS
Aug <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Aug-divvy-tripdata-21.csv")
```

```r
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Aug <- subset(Aug, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,End_Station_ID,X))
```

```r
#TOTAL NUMBER OF ROWS
nrow(Aug)
```

```
## [1] 729595
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Aug)
```

```
## [1] 9
```

```
#duplicate check
duplicated(Aug)
sum(duplicated(Aug))%>%
          head(n = 20)
```

```
#add new column to the dataset
Aug %>% add_column (Time_Diff8 = NA,Lat8 = NA, Ride_Lng8 = NA) %>%
          head(n = 20)
```

```
#RENAME COLUMNS
Aug <- Aug %>%
  rename(Status = Member_Casual)
```

```
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Aug$Ride_Lng8 = (Aug$Start_Lng - Aug$End_Lng)
```

```
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Aug$Time_Diff8 = difftime(Aug$Ended_at, Aug$Started_at, units = "secs")
```

```
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Aug$Ride_Lng8, na.rm = TRUE)
```

```
## [1] 0.00026723
```

```
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Aug[Aug$Status == 'member','Ride_Lng8'],na.rm = TRUE)
```

```
## [1] 0.000214137
```

```
mean(Aug[Aug$Status == 'casual','Ride_Lng8'],na.rm = TRUE)
```

```
## [1] 0.000318683
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #STATUS COLUMN IS EQUALS
mean(Aug[Aug$Status == 'casual','Time_Diff8'],na.rm = TRUE)
```

```
## Time difference of 293620 secs
```

```
mean(Aug[Aug$Status == 'member','Time_Diff8'],na.rm = TRUE)
```

```
## Time difference of 26755.37 secs
```

```
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Aug[Aug$Time_Diff8 >= 400,'Ride_Lng8'],na.rm = TRUE)
```

```
## [1] 0.0003320006
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Aug[Aug$Ride_Lng8 >= 0.1,'Time_Diff8'],na.rm = TRUE)
```

```
## Time difference of -1801633 secs
```

```
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Aug[Aug$Rideable_Type == 'docked_bike','Ride_Lng8'],na.rm = TRUE)
```

```
## [1] 0.0003471111
```

```
mean(Aug[Aug$Rideable_Type == 'electric_bike','Ride_Lng8'],na.rm = TRUE)
```

```
## [1] 0.0002399849
```

```
mean(Aug[Aug$Rideable_Type == 'classic_bike','Ride_Lng8'],na.rm = TRUE)
```

```
## [1] 0.0002729532
```

```
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #RIDEABLE TYPE COLUMN IS
mean(Aug[Aug$Rideable_Type == 'docked_bike','Time_Diff8'],na.rm = TRUE)
```

```
## Time difference of 1038049 secs
```

```
mean(Aug[Aug$Rideable_Type == 'electric_bike','Time_Diff8'],na.rm = TRUE)
```

```
## Time difference of 80512.99 secs
```

```
mean(Aug[Aug$Rideable_Type == 'classic_bike','Time_Diff8'],na.rm = TRUE)
```

```
## Time difference of 103915.1 secs
```

```
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALUES
max(Aug$Ride_Lng8, na.rm = TRUE)
```

```
## [1] 0.2342447
```

```
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALUES
min(Aug$Ride_Lng8, na.rm = TRUE)
```

```
## [1] -0.1643892
```

```
#CALCULATE THE MODE OF RIDE LENGHT
mode(Aug$Ride_Lng8)
```

```
## [1] "numeric"
```

```
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Aug$Status)
```

```
##
## casual member
## 370681 358914
```

```
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Aug$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        435020         51716        242859
```

```
#EXTRACT COLUMNS OF INTEREST
Aug_df <- Aug %>%
  select(Ride_ID, Rideable_Type, Time_Diff8, Ride_Lng8, Status)
```

```
#preview of the dataset
head(Aug_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff8 Ride_Lng8 Status
## 1 99FEC93BA843FB20 electric_bike   180 secs      0.01 member
## 2 06048DCFC8520CAF electric_bike   360 secs      0.01 member
## 3 9598066F68045DF2 electric_bike   360 secs     -0.01 member
```

```
glimpse(Aug_df)
```

```
## Rows: 729,595
## Columns: 5
## $ Ride_ID       <chr> "99FEC93BA843FB20", "06048DCFC8520CAF", "9598066F68045DF~
## $ Rideable_Type <chr> "electric_bike", "electric_bike", "electric_bike", "elec~
## $ Time_Diff8    <drtn> 180 secs, 360 secs, 360 secs, 1500 secs, 240 secs, 360 ~
## $ Ride_Lng8     <dbl> 0.010000, 0.010000, -0.010000, 0.020000, 0.000000, 0.000~
## $ Status        <chr> "member", "member", "member", "member", "member", "membe~
```

```
summary(Aug_df)
```

```
##    Ride_ID           Rideable_Type       Time_Diff8         Ride_Lng8
##  Length:729595      Length:729595      Length:729595     Min.   :-0.1644
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.0096
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.0000
##                                                          Mean   : 0.0003
##                                                          3rd Qu.: 0.0100
##                                                          Max.   : 0.2342
##                                                          NA's   :717
##     Status
##  Length:729595
##  Class :character
##  Mode  :character
##
##
##
##
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Aug_df <- filter(Aug_df,Ride_Lng8 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Aug_df <- filter(Aug_df,Ride_Lng8 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Aug_df <- filter(Aug_df,Time_Diff8 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Aug_df <- filter(Aug_df,Time_Diff8 < 400)%>%
          head(n = 20)
```

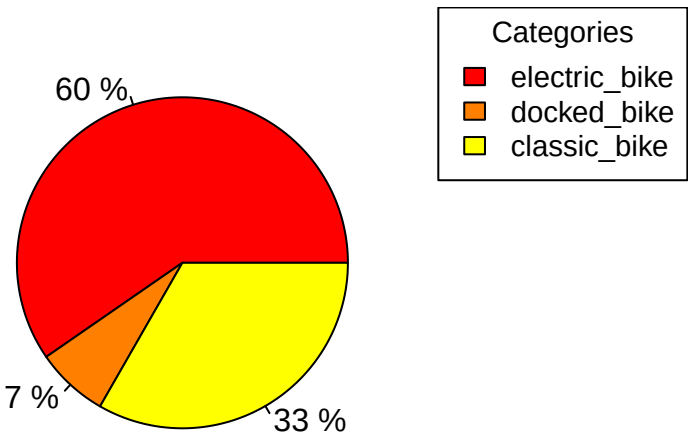
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Aug_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "60 %" "7 %"  "33 %"
```

```
#pie plot
pie(table(Aug_df$Rideable_Type), labels =
      paste(round(prop.table(table(Aug_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

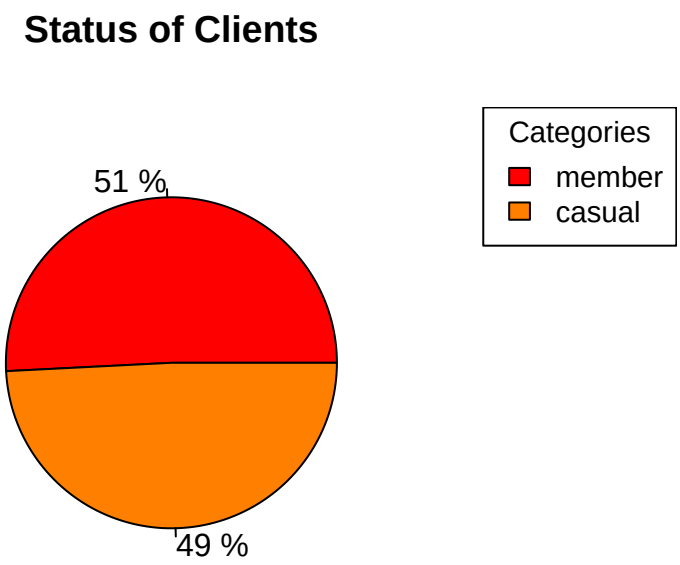


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Aug_df$Status))*100),"%", sep = " ")
```

```
## [1] "51 %" "49 %"
```

```
#pie plot
pie(table(Aug_df$Status), labels =
      paste(round(prop.table(table(Aug_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#ANALYSIS FOR THE MONTH OF SEPTEMBER, 2021

```
# IMPORT DATASETS
Sep <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Sept-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Sep <- subset(Sep, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID))
```

```
#TOTAL NUMBER OF ROWS
nrow(Sep)
```

```
## [1] 822410
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Sep)
```

```
## [1] 9
```

```
#duplicate check
duplicated(Sep)
sum(duplicated(Sep))%>%
          head(n = 20)
```

```r
#add new column to the dataset
Sep %>% add_column (Time_Diff9 = NA,Lat9 = NA, Ride_Lng9 = NA,eval=FALSE) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
Sep <- Sep %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Sep$Ride_Lng9 = (Sep$Start_Lng - Sep$End_Lng)
Sep$Lat9 = (Sep$Start_Lat - Sep$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Sep$Time_Diff9 = difftime(Sep$Ended_at, Sep$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Sep$Ride_Lng9, na.rm = TRUE)
```

```
## [1] 0.000123108
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Sep[Sep$Status == 'member','Ride_Lng9'],na.rm = TRUE)
```

```
## [1] 0.0001013588
```

```r
mean(Sep[Sep$Status == 'casual','Ride_Lng9'],na.rm = TRUE)
```

```
## [1] 0.0001418389
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #STATUS COLUMN IS EQUALS
mean(Sep[Sep$Status == 'casual','Time_Diff9'],na.rm = TRUE)
```

```
## Time difference of -481092.4 secs
```

```r
mean(Sep[Sep$Status == 'member','Time_Diff9'],na.rm = TRUE)
```

```
## Time difference of -103904.7 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Sep[Sep$Time_Diff9 >= 400,'Ride_Lng9'],na.rm = TRUE)
```

```
## [1] 0.0001424492
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Sep[Sep$Ride_Lng9 >= 0.1,'Time_Diff9'],na.rm = TRUE)
```

```
## Time difference of -9957893 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Sep[Sep$Rideable_Type == 'docked_bike','Ride_Lng9'],na.rm = TRUE)
```

```
## [1] 0.0001744906
```

```r
mean(Sep[Sep$Rideable_Type == 'electric_bike','Ride_Lng9'],na.rm = TRUE)
```

```
## [1] 5.597178e-05
```

```r
mean(Sep[Sep$Rideable_Type == 'classic_bike','Ride_Lng9'],na.rm = TRUE)
```

```
## [1] 0.0001514443
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #RIDEABLE TYPE COLUMN IS
mean(Sep[Sep$Rideable_Type == 'docked_bike','Time_Diff9'],na.rm = TRUE)
```

```
## Time difference of -373787.8 secs
```

```r
mean(Sep[Sep$Rideable_Type == 'electric_bike','Time_Diff9'],na.rm = TRUE)
```

```
## Time difference of -122227.6 secs
```

```r
mean(Sep[Sep$Rideable_Type == 'classic_bike','Time_Diff9'],na.rm = TRUE)
```

```
## Time difference of -392798.1 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALUES
max(Sep$Ride_Lng9, na.rm = TRUE)
```

```
## [1] 0.23094
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALUES
min(Sep$Ride_Lng9, na.rm = TRUE)
```

```
## [1] -0.1786105
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Sep$Ride_Lng9)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Sep$Status)
```

```
##
## casual member
## 442056 380354
```

```
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Sep$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        506909         57698        257803
```

```
#EXTRACT COLUMNS OF INTEREST
Sep_df <- Sep %>%
  select(Ride_ID, Rideable_Type, Time_Diff9, Lat9, Ride_Lng9, Status)
```

```
#preview of the dataset
head(Sep_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff9      Lat9 Ride_Lng9 Status
## 1 0A1B623926EF4E16   docked_bike  2100 secs -0.015384  0.023796 casual
## 2 B2D5583A5A5E76EE  classic_bike  1140 secs  0.010464 -0.025231 casual
## 3 6F264597DDBF427A  classic_bike  1080 secs -0.029789  0.000372 member
```

```
glimpse(Sep_df)
```

```
## Rows: 822,410
## Columns: 6
## $ Ride_ID       <chr> "0A1B623926EF4E16", "B2D5583A5A5E76EE", "6F264597DDBF427~
## $ Rideable_Type <chr> "docked_bike", "classic_bike", "classic_bike", "classic_~
## $ Time_Diff9    <drtn> 2100 secs, 1140 secs, 1080 secs, 900 secs, 1140 secs, 3~
## $ Lat9          <dbl> -0.01538400, 0.01046400, -0.02978900, 0.00580700, 0.0137~
## $ Ride_Lng9     <dbl> 0.02379600, -0.02523100, 0.00037200, -0.04325500, -0.038~
## $ Status        <chr> "casual", "casual", "member", "member", "casual", "casua~
```

```
summary(Sep_df)
```

```
##    Ride_ID          Rideable_Type       Time_Diff9            Lat9
##  Length:822410      Length:822410      Length:822410     Min.   :-0.2530
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.0104
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.0000
##                                                          Mean   :-0.0002
##                                                          3rd Qu.: 0.0100
##                                                          Max.   : 0.2897
##                                                          NA's   :731
##    Ride_Lng9          Status
##  Min.   :-0.1786   Length:822410
##  1st Qu.:-0.0099   Class :character
##  Median : 0.0000   Mode  :character
##  Mean   : 0.0001
##  3rd Qu.: 0.0100
##  Max.   : 0.2309
##  NA's   :731
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Sep_df <- filter(Sep_df,Ride_Lng9 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Sep_df <- filter(Sep_df,Ride_Lng9 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Sep_df <- filter(Sep_df,Time_Diff9 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Sep_df <- filter(Sep_df,Time_Diff9 < 400)%>%
          head(n = 20)
```

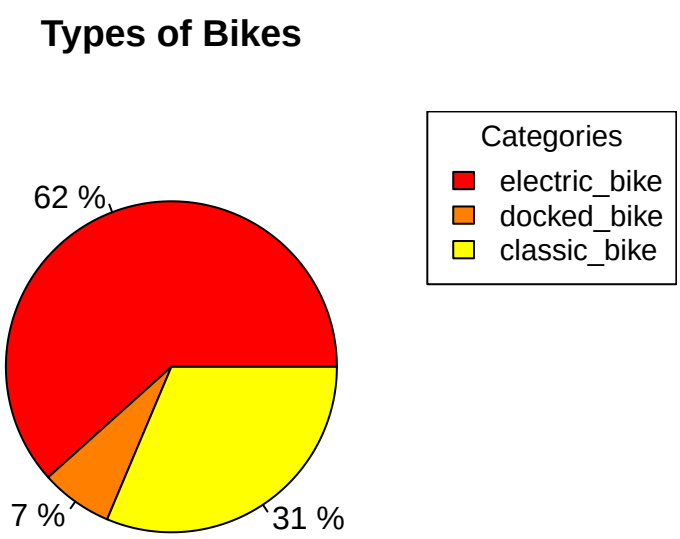
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Sep_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "62 %" "7 %"  "31 %"
```

```
#pie plot
pie(table(Sep_df$Rideable_Type), labels =
      paste(round(prop.table(table(Sep_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike","classic_bike"),
      fill = heat.colors(3), title = "Categories", cex = 0.9)
```

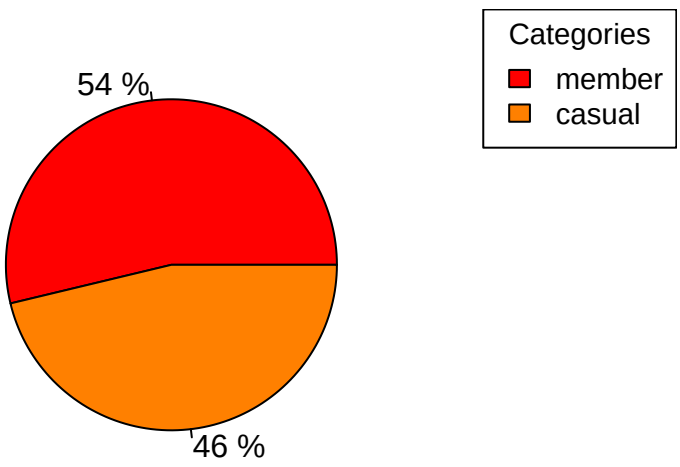


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Sep_df$Status))*100),"%", sep = " ")
```

```
## [1] "54 %" "46 %"
```

```
#pie plot
pie(table(Sep_df$Status), labels =
      paste(round(prop.table(table(Sep_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#ANALYSIS FOR THE MONTH OF OCTOBER, 2021

```
# IMPORT DATASETS
Oct <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Oct-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Oct <- subset(Oct, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID,X))
```

```
#TOTAL NUMBER OF ROWS
nrow(Oct)
```

```
## [1] 804352
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Oct)
```

```
## [1] 9
```

```
#duplicate check
duplicated(Oct)
sum(duplicated(Oct))%>%
          head(n = 20)
```

```r
#add new column to the dataset
Oct %>% add_column (Time_Diff10 = NA,Lat10 = NA, Ride_Lng10 = NA) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
Oct <- Oct %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Oct$Ride_Lng10 = (Oct$Start_Lng - Oct$End_Lng)
Oct$Lat10 = (Oct$Start_Lat - Oct$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Oct$Time_Diff10 = difftime(Oct$Ended_at, Oct$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Oct$Ride_Lng10, na.rm = TRUE)
```

```
## [1] 0.0001709961
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Oct[Oct$Status == 'member','Ride_Lng10'],na.rm = TRUE)
```

```
## [1] 0.000118296
```

```r
mean(Oct[Oct$Status == 'casual','Ride_Lng10'],na.rm = TRUE)
```

```
## [1] 0.0002210653
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #STATUS COLUMN IS EQUALS
mean(Oct[Oct$Status == 'casual','Time_Diff10'],na.rm = TRUE)
```

```
## Time difference of 230913.3 secs
```

```r
mean(Oct[Oct$Status == 'member','Time_Diff10'],na.rm = TRUE)
```

```
## Time difference of 52594.81 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Oct[Oct$Time_Diff10 >= 400,'Ride_Lng10'],na.rm = TRUE)
```

```
## [1] 0.0002061291
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Oct[Oct$Ride_Lng10 >= 0.1,'Time_Diff10'],na.rm = TRUE)
```

```
## Time difference of 915263.2 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Oct[Oct$Rideable_Type == 'docked_bike','Ride_Lng10'],na.rm = TRUE)
```

```
## [1] 0.0001293134
```

```r
mean(Oct[Oct$Rideable_Type == 'electric_bike','Ride_Lng10'],na.rm = TRUE)
```

```
## [1] 0.0001453412
```

```r
mean(Oct[Oct$Rideable_Type == 'classic_bike','Ride_Lng10'],na.rm = TRUE)
```

```
## [1] 0.000187823
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #RIDEABLE TYPE COLUMN IS
mean(Oct[Oct$Rideable_Type == 'docked_bike','Time_Diff10'],na.rm = TRUE)
```

```
## Time difference of 491041.2 secs
```

```r
mean(Oct[Oct$Rideable_Type == 'electric_bike','Time_Diff10'],na.rm = TRUE)
```

```
## Time difference of 96264.64 secs
```

```r
mean(Oct[Oct$Rideable_Type == 'classic_bike','Time_Diff10'],na.rm = TRUE)
```

```
## Time difference of 137356 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALUES
max(Oct$Ride_Lng10, na.rm = TRUE)
```

```
## [1] 0.2127328
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALUES
min(Oct$Ride_Lng10, na.rm = TRUE)
```

```
## [1] -0.2126255
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Oct$Ride_Lng10)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Oct$Status)
```

```
##
## casual member
## 412671 391681
```

```r
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Oct$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        503033         45065        256254
```

```r
#EXTRACT COLUMNS OF INTEREST
Oct_df <- Oct %>%
  select(Ride_ID, Rideable_Type, Time_Diff10, Lat10, Ride_Lng10, Status)
```

```r
#preview of the dataset
head(Oct_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff10 Lat10 Ride_Lng10 Status
## 1 99103BB87CC6C1BB electric_bike    420 secs  0.00       0.00 member
## 2 EAFCCCFB0A3FC5A1 electric_bike    960 secs  0.00      -0.05 member
## 3 9EF4F46C57AD234D electric_bike    960 secs -0.02       0.01 member
```

```r
glimpse(Oct_df)
```

```
## Rows: 804,352
## Columns: 6
## $ Ride_ID       <chr> "99103BB87CC6C1BB", "EAFCCCFB0A3FC5A1", "9EF4F46C57AD234~
## $ Rideable_Type <chr> "electric_bike", "electric_bike", "electric_bike", "elec~
## $ Time_Diff10   <drtn> 420 secs, 960 secs, 960 secs, 960 secs, 540 secs, 360 s~
## $ Lat10         <dbl> 0.00000000, 0.00000000, -0.02000000, 0.02000000, 0.02000~
## $ Ride_Lng10    <dbl> 0.00000000, -0.05000000, 0.01000000, -0.02000000, 0.0200~
## $ Status        <chr> "member", "member", "member", "member", "member", "membe~
```

```r
summary(Oct_df)
```

```
##    Ride_ID          Rideable_Type      Time_Diff10           Lat10
##  Length:804352      Length:804352      Length:804352     Min.   :-0.2713
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.0101
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.0000
##                                                          Mean   :-0.0002
##                                                          3rd Qu.: 0.0100
##                                                          Max.   : 0.2335
##                                                          NA's   :706
##    Ride_Lng10         Status
##  Min.   :-0.2126   Length:804352
##  1st Qu.:-0.0099   Class :character
##  Median : 0.0000   Mode  :character
##  Mean   : 0.0002
##  3rd Qu.: 0.0100
##  Max.   : 0.2127
##  NA's   :706
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Oct_df <- filter(Oct_df,Ride_Lng10 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Oct_df <- filter(Oct_df,Ride_Lng10 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable #type
TimeDiff_Oct_df <- filter(Oct_df,Time_Diff10 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Oct_df <- filter(Oct_df,Time_Diff10 < 400)%>%
          head(n = 20)
```

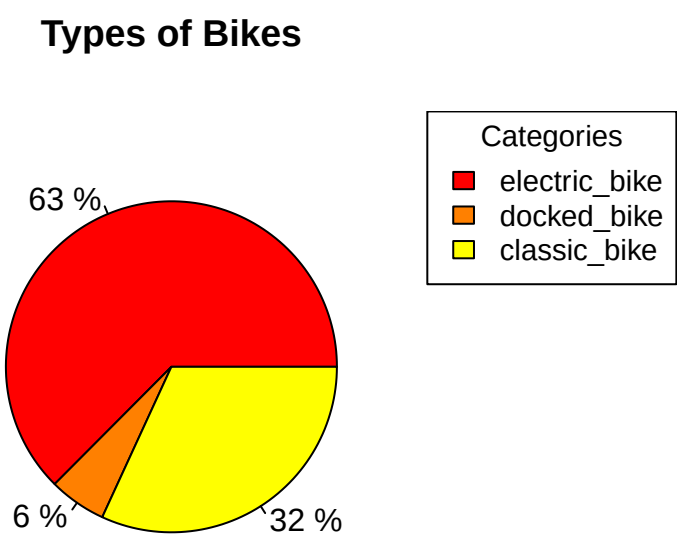
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Oct_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "63 %" "6 %"  "32 %"
```

```
#pie plot
pie(table(Oct_df$Rideable_Type), labels =
      paste(round(prop.table(table(Oct_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

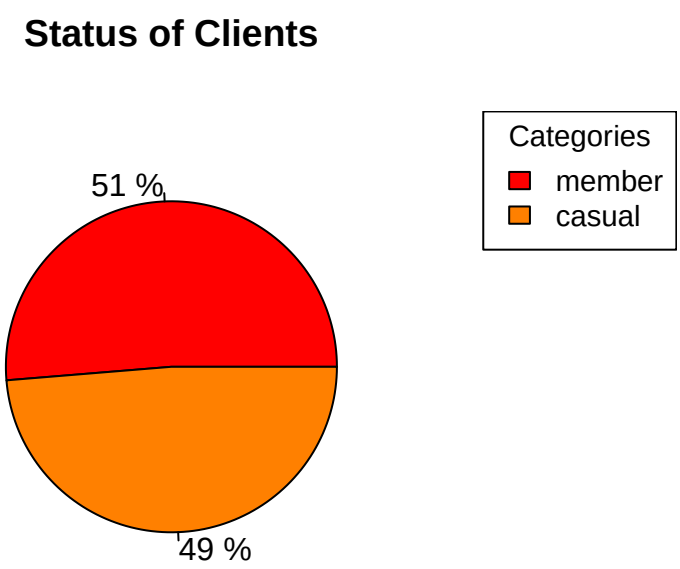


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Oct_df$Status))*100),"%", sep = " ")
```

```
## [1] "51 %" "49 %"
```

```
#pie plot
pie(table(Oct_df$Status), labels =
      paste(round(prop.table(table(Oct_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#ANALYSIS FOR THE MONTH OF NOVEMBER, 2021

```
# IMPORT DATASETS
Nov <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Nov-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Nov <- subset(Nov, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID,X))
```

```
#TOTAL NUMBER OF ROWS
nrow(Nov)
```

```
## [1] 519380
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Nov)
```

```
## [1] 9
```

```
#duplicate check
duplicated(Nov)
sum(duplicated(Nov))%>%
          head(n = 20)
```

```r
#add new column to the dataset
Nov %>% add_column (Time_Diff11 = NA,Lat11 = NA, Ride_Lng11 = NA) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
Nov <- Nov %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Nov$Ride_Lng11 = (Nov$Start_Lng - Nov$End_Lng)
Nov$Lat11 = (Nov$Start_Lat - Nov$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Nov$Time_Diff11 = difftime(Nov$Ended_at, Nov$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Nov$Ride_Lng11, na.rm = TRUE)
```

```
## [1] 0.0001491978
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Nov[Nov$Status == 'member','Ride_Lng11'],na.rm = TRUE)
```

```
## [1] 0.0001574586
```

```r
mean(Nov[Nov$Status == 'casual','Ride_Lng11'],na.rm = TRUE)
```

```
## [1] 0.0001408219
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #STATUS COLUMN IS EQUALS
mean(Nov[Nov$Status == 'casual','Time_Diff11'],na.rm = TRUE)
```

```
## Time difference of 112477.2 secs
```

```r
mean(Nov[Nov$Status == 'member','Time_Diff11'],na.rm = TRUE)
```

```
## Time difference of 42002.22 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Nov[Nov$Time_Diff11 >= 400,'Ride_Lng11'],na.rm = TRUE)
```

```
## [1] 0.0001625854
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Nov[Nov$Ride_Lng11 >= 0.1,'Time_Diff11'],na.rm = TRUE)
```

```
## Time difference of 619757.8 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Nov[Nov$Rideable_Type == 'docked_bike','Ride_Lng11'],na.rm = TRUE)
```

```
## [1] 8.334144e-05
```

```r
mean(Nov[Nov$Rideable_Type == 'electric_bike','Ride_Lng11'],na.rm = TRUE)
```

```
## [1] 1.776344e-05
```

```r
mean(Nov[Nov$Rideable_Type == 'classic_bike','Ride_Lng11'],na.rm = TRUE)
```

```
## [1] 0.0002386684
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #RIDEABLE TYPE COLUMN IS
mean(Nov[Nov$Rideable_Type == 'docked_bike','Time_Diff11'],na.rm = TRUE)
```

```
## Time difference of 708186.9 secs
```

```r
mean(Nov[Nov$Rideable_Type == 'electric_bike','Time_Diff11'],na.rm = TRUE)
```

```
## Time difference of 7119.505 secs
```

```r
mean(Nov[Nov$Rideable_Type == 'classic_bike','Time_Diff11'],na.rm = TRUE)
```

```
## Time difference of 77739.6 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALUES
max(Nov$Ride_Lng11, na.rm = TRUE)
```

```
## [1] 0.21
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALUES
min(Nov$Ride_Lng11, na.rm = TRUE)
```

```
## [1] -0.1880328
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Nov$Ride_Lng11)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Nov$Status)
```

```
##
## casual member
## 257994 261386
```

```r
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Nov$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        302897         21266        195217
```

```r
#EXTRACT COLUMNS OF INTEREST
Nov_df <- Nov %>%
  select(Ride_ID, Rideable_Type, Time_Diff11, Lat11, Ride_Lng11, Status)
```

```r
#preview of the dataset
head(Nov_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff11       Lat11  Ride_Lng11 Status
## 1 9A4A8801674A3995  classic_bike    300 secs  0.00452800 -0.00408100 member
## 2 6310A715D7121C8C electric_bike    900 secs -0.03703917  0.01116717 member
## 3 7EFD1EEF66C09AB3  classic_bike   1200 secs -0.02936600  0.00210300 member
```

```r
glimpse(Nov_df)
```

```
## Rows: 519,380
## Columns: 6
## $ Ride_ID       <chr> "9A4A8801674A3995", "6310A715D7121C8C", "7EFD1EEF66C09AB~
## $ Rideable_Type <chr> "classic_bike", "electric_bike", "classic_bike", "classi~
## $ Time_Diff11   <drtn> 300 secs, 900 secs, 1200 secs, 600 secs, 600 secs, 540 ~
## $ Lat11         <dbl> 0.00452800, -0.03703917, -0.02936600, -0.01145800, -0.01~
## $ Ride_Lng11    <dbl> -0.00408100, 0.01116717, 0.00210300, 0.00899600, 0.00899~
## $ Status        <chr> "member", "member", "member", "member", "member", "membe~
```

```r
summary(Nov_df)
```

```
##    Ride_ID          Rideable_Type      Time_Diff11           Lat11
##  Length:519380      Length:519380      Length:519380     Min.   :-0.2539
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.0100
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.0000
##                                                          Mean   :-0.0003
##                                                          3rd Qu.: 0.0100
##                                                          Max.   : 0.2641
##                                                          NA's   :466
##    Ride_Lng11         Status
##  Min.   :-0.1880   Length:519380
##  1st Qu.:-0.0100   Class :character
##  Median : 0.0000   Mode  :character
##  Mean   : 0.0001
##  3rd Qu.: 0.0100
##  Max.   : 0.2100
##  NA's   :466
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Nov_df <- filter(Nov_df,Ride_Lng11 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Nov_df <- filter(Nov_df,Ride_Lng11 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Nov_df <- filter(Nov_df,Time_Diff11 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Nov_df <- filter(Nov_df,Time_Diff11 < 400)%>%
          head(n = 20)
```

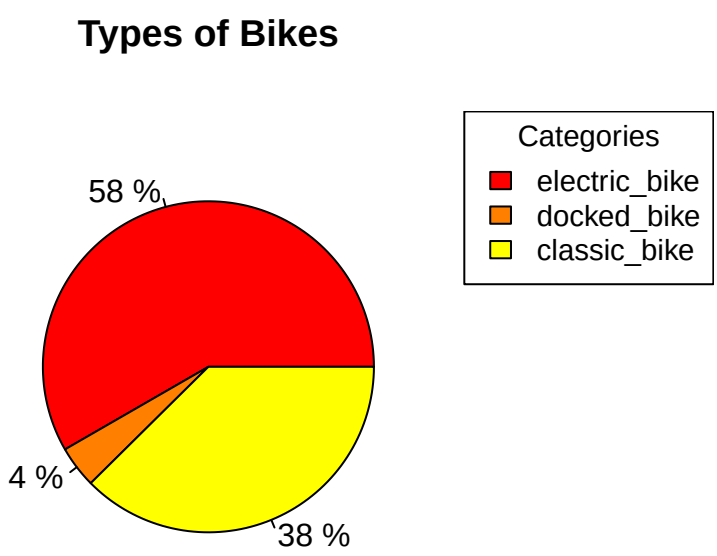
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Nov_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "58 %" "4 %"  "38 %"
```

```
#pie plot
pie(table(Nov_df$Rideable_Type), labels =
     paste(round(prop.table(table(Nov_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes",radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike", "classic_bike"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```

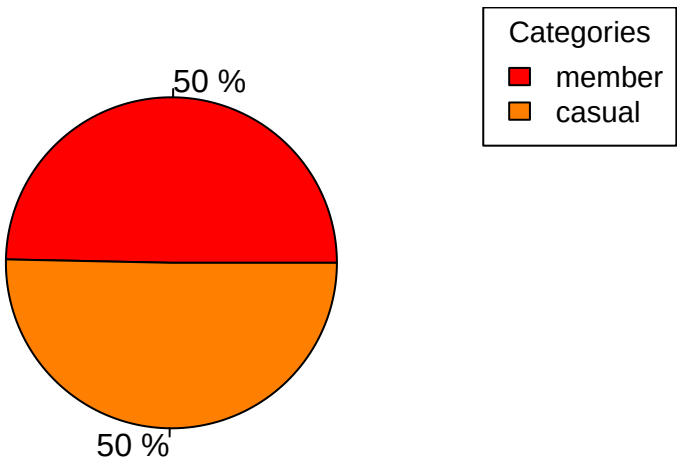


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Nov_df$Status))*100),"%", sep = " ")
```

```
## [1] "50 %" "50 %"
```

```
#pie plot
pie(table(Nov_df$Status), labels =
      paste(round(prop.table(table(Nov_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```



#ANALYSIS FOR THE MONTH OF DECEMBER, 2021

```
# IMPORT DATASETS
Dec <- read.csv("C:/Users/LILIAN/Desktop/Google Capstone project/Dec-divvy-tripdata-21.csv")
```

```
#DROP INSIGNIFICANT COLUMNS OF THE DATASET
Dec <- subset(Dec, select = -c(Start_Station_Name,Start_Station_ID,End_Station_Name,
                               End_Station_ID))
```

```
#TOTAL NUMBER OF ROWS
nrow(Dec)
```

```
## [1] 631226
```

```
#TOTAL NUMBER OF COLUMNS
ncol(Dec)
```

```
## [1] 9
```

```
#duplicate check
duplicated(Dec)
sum(duplicated(Dec))%>%
          head(n = 20)
```

```r
#add new column to the dataset
Dec %>% add_column (Time_Diff12 = NA,Lat12 = NA, Ride_Lng12 = NA,eval=FALSE) %>%
          head(n = 20)
```

```r
#RENAME COLUMNS
Dec <- Dec %>%
  rename(Status = Member_Casual)
```

```r
#TO GET A DIFFERENCE OF Lat and LENGHT columns
Dec$Ride_Lng12 = (Dec$Start_Lng - Dec$End_Lng)
Dec$Lat12 = (Dec$Start_Lat - Dec$End_Lat)
```

```r
#DIFFERENCE OF STARTED TIME AND ENDED TIME
Dec$Time_Diff12 = difftime(Dec$Ended_at, Dec$Started_at, units = "secs")
```

```r
#CALCULATE MEAN MEAN RIDE LENGHT USING COLUMN NAME, IGNORE MISSING VALUES
mean(Dec$Ride_Lng12, na.rm = TRUE)
```

```
## [1] 0.0002816721
```

```r
#THE MEAN OF THE RIDE lENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#STATUS COLUMN IS EQUALS TO EACH OF CASUAL AND MEMBER.
mean(Dec[Dec$Status == 'member','Ride_Lng12'],na.rm = TRUE)
```

```
## [1] 0.0002158595
```

```r
mean(Dec[Dec$Status == 'casual','Ride_Lng12'],na.rm = TRUE)
```

```
## [1] 0.0003774895
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #STATUS COLUMN IS EQUALS
mean(Dec[Dec$Status == 'casual','Time_Diff12'],na.rm = TRUE)
```

```
## Time difference of 176875.6 secs
```

```r
mean(Dec[Dec$Status == 'member','Time_Diff12'],na.rm = TRUE)
```

```
## Time difference of 52830.35 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN WHERE THE TIME DIFFERENCE COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 400
mean(Dec[Dec$Time_Diff12 >= 400,'Ride_Lng12'],na.rm = TRUE)
```

```
## [1] 0.0003744893
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN WHERE THE RIDE LENGHT COLUMN HAS
#VALUE GREATER THAN OR EQUALS TO 0.1
mean(Dec[Dec$Ride_Lng12 >= 0.1,'Time_Diff12'],na.rm = TRUE)
```

```
## Time difference of 658568.3 secs
```

```r
#THE MEAN OF THE RIDE LENGHT COLUMN FOR EVERY ROW IN THE DATASET WHERE THE
#RIDEABLE TYPE COLUMN IS EQUALS TO EACH OF ELECTRIC, CLASSIC AND DOCKED BIKE.
mean(Dec[Dec$Rideable_Type == 'docked_bike','Ride_Lng12'],na.rm = TRUE)
```

```
## [1] 0.000315072
```

```r
mean(Dec[Dec$Rideable_Type == 'electric_bike','Ride_Lng12'],na.rm = TRUE)
```

```
## [1] 0.0001880042
```

```r
mean(Dec[Dec$Rideable_Type == 'classic_bike','Ride_Lng12'],na.rm = TRUE)
```

```
## [1] 0.000365928
```

```r
#THE MEAN OF THE TIME DIFFERENCE COLUMN FOR EVERY ROW IN THE DATASET WHERE THE #RIDEABLE TYPE COLUMN IS
mean(Dec[Dec$Rideable_Type == 'docked_bike','Time_Diff12'],na.rm = TRUE)
```

```
## Time difference of 406004.4 secs
```

```r
mean(Dec[Dec$Rideable_Type == 'electric_bike','Time_Diff12'],na.rm = TRUE)
```

```
## Time difference of 95833.07 secs
```

```r
mean(Dec[Dec$Rideable_Type == 'classic_bike','Time_Diff12'],na.rm = TRUE)
```

```
## Time difference of 88454.21 secs
```

```r
# CALCULATE THE MAX RIDE LENGHT WITH MISSING VALUES
max(Dec$Ride_Lng12, na.rm = TRUE)
```

```
## [1] 0.29
```

```r
# CALCULATE THE MIN RIDE LENGHT WITH MISSING VALUES
min(Dec$Ride_Lng12, na.rm = TRUE)
```

```
## [1] -0.2
```

```r
#CALCULATE THE MODE OF RIDE LENGHT
mode(Dec$Ride_Lng12)
```

```
## [1] "numeric"
```

```r
#SUM OF EACH VARIABLE IN THE STATUS COLUMN
table(Dec$Status)
```

```
##
## casual member
## 257242 373984
```

```
#SUM OF EACH VARIABLE IN THE RIDEABLE COLUMN
table(Dec$Rideable_Type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        316139         22884        292203
```

```
#EXTRACT COLUMNS OF INTEREST
Dec_df <- Dec %>%
  select(Ride_ID, Rideable_Type, Time_Diff12, Lat12, Ride_Lng12, Status)
```

```
#preview of the dataset
head(Dec_df,3)
```

```
##            Ride_ID Rideable_Type Time_Diff12       Lat12  Ride_Lng12 Status
## 1 620BC6107255BF4C electric_bike    180 secs -0.00081367 -0.00849533 member
## 2 4471C70731AB2E45 electric_bike    120 secs  0.00000000  0.01000000 member
## 3 26CA69D43D15EE14 electric_bike    480 secs -0.02000000  0.02000000 member
```

```
glimpse(Dec_df)
```

```
## Rows: 631,226
## Columns: 6
## $ Ride_ID       <chr> "620BC6107255BF4C", "4471C70731AB2E45", "26CA69D43D15EE1~
## $ Rideable_Type <chr> "electric_bike", "electric_bike", "electric_bike", "elec~
## $ Time_Diff12   <drtn> 180 secs, 120 secs, 480 secs, 120 secs, 540 secs, 840 s~
## $ Lat12         <dbl> -0.00081367, 0.00000000, -0.02000000, 0.00000000, 0.0000~
## $ Ride_Lng12    <dbl> -0.00849533, 0.01000000, 0.02000000, 0.00000000, -0.0200~
## $ Status        <chr> "member", "member", "member", "member", "member", "membe~
```

```
summary(Dec_df)
```

```
##    Ride_ID          Rideable_Type      Time_Diff12           Lat12
##  Length:631226      Length:631226      Length:631226     Min.   :-0.2510
##  Class :character   Class :character   Class :difftime   1st Qu.:-0.0098
##  Mode  :character   Mode  :character   Mode  :numeric    Median : 0.0000
##                                                          Mean   :-0.0002
##                                                          3rd Qu.: 0.0092
##                                                          Max.   : 0.2292
##                                                          NA's   :484
##    Ride_Lng12         Status
##  Min.   :-0.2000   Length:631226
##  1st Qu.:-0.0094   Class :character
##  Median : 0.0000   Mode  :character
##  Mean   : 0.0003
##  3rd Qu.: 0.0099
##  Max.   : 0.2900
##  NA's   :484
```

```
#FILTER PARAMETERS
#filter ride lenght greater or equals to 0.1 against status and rideable type
RideLength_Dec_df <- filter(Dec_df,Ride_Lng12 >= 0.1)%>%
          head(n = 20)
```

```
#for Less than 0.1
LessRideLength_Dec_df <- filter(Dec_df,Ride_Lng12 < 0.1)%>%
          head(n = 20)
```

```
#filter for Time difference greater than 400, plot against status & rideable type
TimeDiff_Dec_df <- filter(Dec_df,Time_Diff12 >= 400)%>%
          head(n = 20)
```

```
LessTimeDiff_Dec_df <- filter(Dec_df,Time_Diff12 < 400)%>%
          head(n = 20)
LessTimeDiff_Dec_df
```

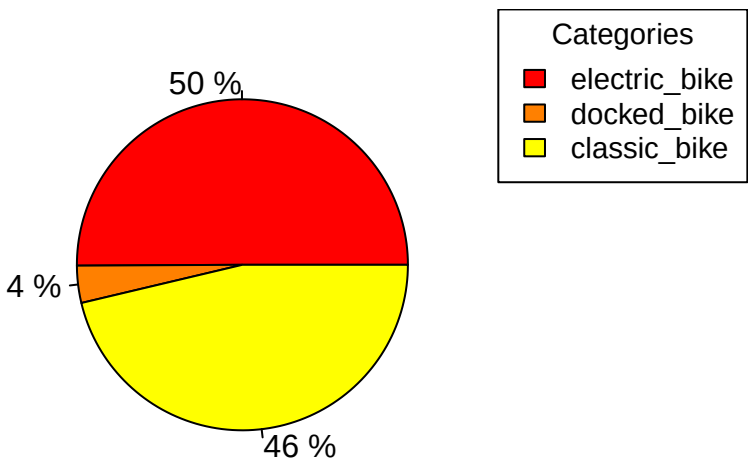
#VISUALIZATION

#PLOTTING PIE CHART FOR RIDEABLE BIKE COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Dec_df$Rideable_Type))*100),"%", sep = " ")
```

```
## [1] "50 %" "4 %"  "46 %"
```

```
#pie plot
pie(table(Dec_df$Rideable_Type), labels =
      paste(round(prop.table(table(Dec_df$Rideable_Type))*100),"%", sep = " "),
    col = heat.colors(3), main = "Types of Bikes", radius = 0.7)
legend("topright", legend = c("electric_bike", "docked_bike","classic_bike"),
      fill = heat.colors(3), title = "Categories", cex = 0.9)
```

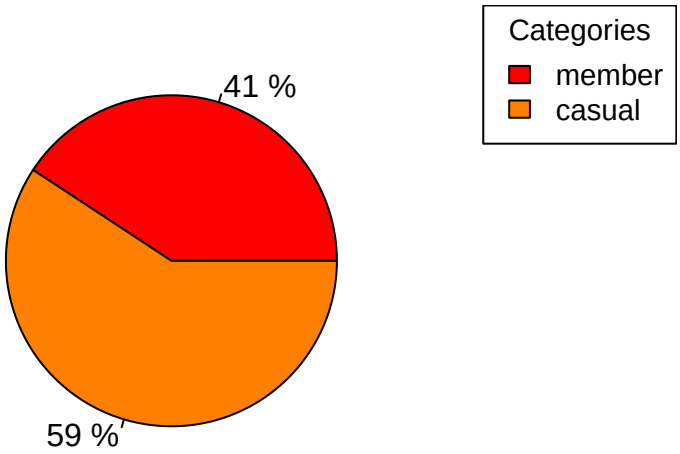


#PLOTTING PIE CHART FOR Status COLUMNS

```
#To get the % of each variable
paste(round(prop.table(table(Dec_df$Status))*100),"%", sep = " ")
```

```
## [1] "41 %" "59 %"
```

```
#pie plot
pie(table(Dec_df$Status), labels =
     paste(round(prop.table(table(Dec_df$Status))*100),"%", sep = " "),
    col = heat.colors(3), main = "Status of Clients",radius = 0.7)
legend("topright", legend = c("member", "casual"),
       fill = heat.colors(3), title = "Categories", cex = 0.9)
```


Status of Clients
Categories
member
casual
41 %
59 %