**--------------------------------------------------------------------**

**STEP 1: CREATING WORK ENVIRONMENT**

**--------------------------------------------------------------------**

**Installing packages**

install.packages("readr")

library(readr)

install.packages("dplyr")

library(dplyr)

install.packages("tidyverse")

library(tidyverse)

install.packages("lubridate")

library(lubridate)

**Setting work directory**

getwd()

setwd("/Users/Irina/Documents/DATA ANALYST/COURSE 8 - Capstone Project/My\_Project/CSV\_files\_for\_R/CSV\_files")

knitr::opts\_knit$set(root.dir = "/Users/Irina/Documents/DATA ANALYST/COURSE 8 - Capstone Project/My\_Project/CSV\_files\_for\_R/CSV\_files")

**------------------------------------------------------------------**

**STEP 2: COLLECTING DATA**

**--------------------------------------------------------------------**

**Reading csv files**

october<-read.csv("2021\_10\_cyclistic\_tripdata.csv", TRUE, ",")

november<-read.csv("2021\_11\_cyclistic\_tripdata.csv", TRUE, ",")

december<-read.csv("2021\_12\_cyclistic\_tripdata.csv", TRUE, ",")

january<-read.csv("2022\_01\_cyclistic\_tripdata.csv", TRUE, ",")

february<-read.csv("2022\_02\_cyclistic\_tripdata.csv", TRUE, ",")

march<-read.csv("2022\_03\_cyclistic\_tripdata.csv", TRUE, ",")

april<-read.csv("2022\_04\_cyclistic\_tripdata.csv", TRUE, ",")

may<-read.csv("2022\_05\_cyclistic\_tripdata.csv", TRUE, ",")

june<-read.csv("2022\_06\_cyclistic\_tripdata.csv", TRUE, ",")

july<-read.csv("2022\_07\_cyclistic\_tripdata.csv", TRUE, ",")

august<-read.csv("2022\_08\_cyclistic\_tripdata.csv", TRUE, ",")

september<-read.csv("2022\_09\_cyclistic\_tripdata.csv", TRUE, ",")

**Combinig 12 datasets into 1**

combined\_data<-rbind(october,november,december,january,february,march,april,may,june,july,august,september)

------------------------------------------------------------------------------------

**STEP 3: CLEANING UP AND PREPARING DATA FOR ANALYSIS**

------------------------------------------------------------------------------------

**Inspecting the new table that has been created**

head(combined\_data)

str(combined\_data)

summary(combined\_data)

**We need to convert character to date/time format in "started\_at" and "ended\_at" columns. Also calculating a new column for trip length in minutes, changing its format from chr to numeric, and deleting "started\_at" and "ended\_at" columns**

combined\_data<-combined\_data %>%

  mutate(start=mdy\_hm(started\_at)) %>%

  mutate(end=mdy\_hm(ended\_at)) %>%

  mutate(trip\_length\_min=(end-start)/60) %>%

  mutate(trip\_length\_min= as.numeric(trip\_length\_min)) %>%

  select(-started\_at, -ended\_at)

**Adding columns that list the date, month, day, and year of each ride. This will allow us to aggregate ride data for each month, day, or year.**

combined\_data$date<-as.Date(combined\_data$start)

combined\_data$month<-format(as.Date(combined\_data$date), "%m")

combined\_data$day<-format(as.Date(combined\_data$date), "%d")

combined\_data$year<-format(as.Date(combined\_data$date), "%Y")

combined\_data$day\_of\_week<-format(as.Date(combined\_data$date), "%A")

------------------------------------------------------------------------------------

**STEP 4: CONDUCTING DESCRIPTIVE ANALYSIS ON DURATION OF RIDES**

------------------------------------------------------------------------------------

**Descriptive analysis on ride length for members and casual riders combined (all figures in minutes)**

Notes: We know from the number of observations that the total number of rides is 5763638. Now we know the total trip length in minutes - 114260086, average trip length - 19.8, and maximum trip length - 40705

sum(combined\_data$trip\_length\_min)

mean(combined\_data$trip\_length\_min)

max(combined\_data$trip\_length\_min)

**Comparing members and casual riders: grouping by "mem\_casual” and summarizing number of trips, trip hours, avg trip length, and max trip length**

combined\_data %>%

  group\_by(member\_casual) %>%

 summarise(trips\_hours=sum(trip\_length\_min)/60,avr\_trip\_min=mean(trip\_length\_min), max\_trip\_length=max(trip\_length\_min))

*Note: as we know from the previous calculation, the average duration of trips is 20 mins. Now, members’ average trip is 13 min, and casual users ride 2.7x longer on average – 30 mins.*

**Calculating number of trips for members vs casual riders**

combined\_data %>%

  count(member\_casual)

*Note: Members take more trips than casual riders. Casual riders account for 41% of the total number of rides.*

**Calculating the average trip time by each day for members vs casual riders**

aggregate(combined\_data$trip\_length\_min~combined\_data$member\_casual+combined\_data$day\_of\_week, FUN=mean)

**Fixing the order of the days of the week**

combined\_data$day\_of\_week<-ordered(combined\_data$day\_of\_week,levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

**Running the average trip time by days of the week for members vs casual riders again**

aggregate(combined\_data$trip\_length\_min~combined\_data$member\_casual+combined\_data$day\_of\_week, FUN=mean) %>%

  print(14)

**Analyzing ridership data by rider type and weekday**

combined\_data %>%

  mutate(weekday=wday(start,label=TRUE)) %>%

  group\_by(member\_casual,weekday) %>%

 summarize(number\_of\_rides=n(),average\_duration\_min=mean(trip\_length\_min)) %>%

  arrange(member\_casual,weekday)

**----------------------------------------------------**

**STEP 5: VISUALIZING DATA**

**----------------------------------------------------**

**Visualizing the number of rides by rider and weekday**

combined\_data %>%

  mutate(weekday=wday(start,label=TRUE)) %>%

  group\_by(member\_casual,weekday) %>%

 summarize(number\_of\_rides=n(),average\_duration\_min=mean(trip\_length\_min)) %>%

  arrange(member\_casual,weekday) %>%

  ggplot(aes(x=weekday, y=number\_of\_rides, fill=member\_casual))+

  geom\_col(position ="dodge")+

  labs(title = "Number of Rides by Rider Type and Weekday")

*Notes: For members the highest number of rides falls on Tuesday, Wednesday, and Thursday. Casual riders, on the other hand, make more rides on weekends, with the lowest numbers on Monday, Tuesday, and Wednesday.*

**Visualizing average duration of rides by rider and weekday**

combined\_data %>%

  mutate(weekday=wday(start,label=TRUE)) %>%

  group\_by(member\_casual,weekday) %>%

 summarize(number\_of\_rides=n(),average\_duration\_min=mean(trip\_length\_min)) %>%

  arrange(member\_casual,weekday) %>%

  ggplot(aes(x=weekday, y=average\_duration\_min, fill=member\_casual))+

  geom\_col(position ="dodge")+

  labs(title = "Average Duration of Rides by Rider Type and Weekday")

*Note: The average duration of rides for casual riders is the highest on Sundays, Saturdays and Mondays.  The average duration of rides for members is very consistent all through the week with a small uptick on Saturday and Sunday.*

**Analyzing number of rides by rider type and month**

combined\_data %>%

  mutate(month=month(start,label=TRUE)) %>%

  group\_by(member\_casual,month) %>%

 summarize(number\_of\_rides=n(),average\_duration\_min=mean(trip\_length\_min)) %>%

  arrange(member\_casual,month)

**Visualizing the number of rides by rider and month (bar chart)**

combined\_data %>%

  mutate(month=month(start,label=TRUE)) %>%

  group\_by(member\_casual, month) %>%

 summarize(number\_of\_rides=n(),average\_duration\_min=mean(trip\_length\_min)) %>%

  arrange(member\_casual,month) %>%

  ggplot(aes(x=month, y=number\_of\_rides, fill=member\_casual))+

  geom\_col(position ="dodge")+

  labs(title = "Number of Rides by Rider Type and Month")

**Visualizing the number of rides by rider type and month (line chart)**

combined\_data %>%

  mutate(month=month(start,label=TRUE)) %>%

  group\_by(member\_casual, month) %>%

 summarize(number\_of\_rides=n(),average\_duration\_min=mean(trip\_length\_min)) %>%

  arrange(member\_casual,month) %>%

  ggplot(aes(x=month, y=number\_of\_rides, group=member\_casual, color=member\_casual))+

  geom\_line(size=1)+

  labs(title = "Number of Rides by Rider Type and Month")

*Note: Both members and casual riders follow a similar pattern: lowest numbers in January and February and a steady climb beginning in March. However, for members, the numbers reach their peak in August; and for casual riders – in July.*

**Visualizing average duration of rides by rider type and month**

combined\_data %>%

  mutate(month=month(start,label=TRUE)) %>%

  group\_by(member\_casual, month) %>%

 summarize(number\_of\_rides=n(),average\_duration\_min=mean(trip\_length\_min)) %>%

  arrange(member\_casual,month) %>%

  ggplot(aes(x=month, y=average\_duration\_min,fill=member\_casual))+

  geom\_col(position="dodge")+

  labs(title = "Average Duration of Rides by Rider Type and Month")

*Note: We notice right away that the average duration of casual riders' trips is significantly higher than that of members, with the highest numbers in March and June and the lowest in November. For members, the highest average duration falls on May, June, July, and the lowest - November, December.*

**------------------------------------------------------------------------------------**

**STEP 6: CONDUCTING DESCRIPTIVE ANALYSIS ON DISTANCE**

**-----------------------------------------------------------------------------------**

**Creating vectors**

x<-combined\_data %>%

  select(c(start\_lng, start\_lat))

y<-combined\_data %>%

  select(c(end\_lng, end\_lat))

**Adding a distance column**

combined\_data<-combined\_data %>%

  mutate(distance=distHaversine(x,y,r=3961))

*Note: right away I saw some observations with 0 for the distance variable, even though the trip duration (trip\_length\_min) for the same observations wasn't 0. I decided to see how many observations in the dataset have 0 in the distance column.*

**Counting rows with "0" for distance**

 sum(combined\_data$distance==0,na.rm=TRUE)

*Note: 4091058 rows don't show any distance. Conclusion: I decided to exclude the distance data from the analysis. If I worked on a team, I would find out more about the tracking system for miles that Cyclistic uses for their bikes (if any).*

**------------------------------------------------------------------------------------**

**STEP 7: CONDUCT DESCRIPTIVE ANALYSIS OF BIKE PREFERENCES**

**------------------------------------------------------------------------------------**

combined\_data %>%

  group\_by(member\_casual, rideable\_type) %>%

  summarize(count=n()) %>%

  arrange(member\_casual, rideable\_type)

**Visualizing bike preferences for casual riders vs members**

combined\_data %>%

  group\_by(member\_casual, rideable\_type) %>%

  summarize(count=n()) %>%

  arrange(rideable\_type, member\_casual) %>%

  ggplot(aes(rideable\_type, count, fill=member\_casual))+

  geom\_col()+

  labs(title="Bike Preferences by Rider Type")

*Note: There is a difference in bike preferences between members and casual riders. Unlike members whose number 1 preference is classic bikes followed by electric bikes, more casual riders prefer electric bikes over classic bikes, and they are the only ones that use docked bikes.*

**------------------------------------------------------------------------------------**

**STEP 8: ANALYSIS OF TOP 5 START AND END STATIONS**

**------------------------------------------------------------------------------------**

**Top 5 start stations for members**

*Note: Kingsbury St & Kinzie St, Clark St & Elm St, Wells St & Concord Ln, Clinton St & Washington Blvd, Ellis Ave & 60th St*

combined\_data %>%

  filter(member\_casual=="member") %>%

  group\_by(start\_station\_name) %>%

  summarize(count=n()) %>%

  arrange(-count) %>%

  print(n=6)

**Top 5 end stations for members**

*Note: Kingsbury St & Kinzie St, Clark St & Elm St, Wells St & Concord Ln, University Ave & 57th St, Clinton St & Washington Blvd*

combined\_data %>%

  filter(member\_casual=="member") %>%

  group\_by(end\_station\_name) %>%

  summarize(count=n()) %>%

  arrange(-count) %>%

  print(n=6)

**Top 5 start stations for casual riders**

*Note: Streeter Dr & Grand Ave, DuSable Lake Shore Dr & Monroe St, Millennium Park, Michigan Ave & Oak St, DuSable Lake Shore Dr & North Blvd*

combined\_data %>%

  filter(member\_casual=="casual") %>%

  group\_by(start\_station\_name) %>%

  summarize(count=n()) %>%

  arrange(-count) %>%

  print(n=6)

**Top 5 end stations for casual riders**

*Note: Streeter Dr & Grand Ave, DuSable Lake Shore Dr & Monroe St, Millennium Park, Michigan Ave & Oak St, DuSable Lake Shore Dr & North Blvd*

combined\_data %>%

  filter(member\_casual=="casual") %>%

  group\_by(end\_station\_name) %>%

  summarize(count=n()) %>%

  arrange(-count) %>%

  print(n=6)

*Note: Top 5 start and end stations for casual riders are all popular touristic location with an easy access to places like Adler Planetarium, Navy Pier, Field Museum, Shedd Aquarium, a variety of beaches, etc*