```
#Cyclistic Bike Share Data Analysis Capstone Project


#load packages
library(tidyverse)
library(geosphere)
library(ggrepel)
library(ggmap)
library(sf)
library(mapview)
library(plyr)


#import data
data1 <- read.csv(file = '202004-divvy-tripdata.csv', header = TRUE, sep = ",")
data2 <- read.csv(file = '202005-divvy-tripdata.csv', header = TRUE, sep = ",")
data3 <- read.csv(file = '202006-divvy-tripdata.csv', header = TRUE, sep = ",")
data4 <- read.csv(file = '202007-divvy-tripdata.csv', header = TRUE, sep = ",")
data5 <- read.csv(file = '202008-divvy-tripdata.csv', header = TRUE, sep = ",")
data6 <- read.csv(file = '202009-divvy-tripdata.csv', header = TRUE, sep = ",")
data7 <- read.csv(file = '202010-divvy-tripdata.csv', header = TRUE, sep = ",")
data8 <- read.csv(file = '202011-divvy-tripdata.csv', header = TRUE, sep = ",")
data9 <- read.csv(file = '202012-divvy-tripdata.csv', header = TRUE, sep = ",")
data10 <- read.csv(file = '202101-divvy-tripdata.csv', header = TRUE, sep = ",")
data11 <- read.csv(file = '202102-divvy-tripdata.csv', header = TRUE, sep = ",")
data12 <- read.csv(file = '202103-divvy-tripdata.csv', header = TRUE, sep = ",")
data13 <- read.csv(file = '202104-divvy-tripdata.csv', header = TRUE, sep = ",")


#merge files into one full data set
bike_data <- merge(data1, data2, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data3, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data4, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data5, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data6, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data7, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data8, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data9, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data10, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data11, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data12, all.x = TRUE, all.y = TRUE)
bike_data <- merge(bike_data, data13, all.x = TRUE, all.y = TRUE)
```

```
#view data to understand how it is organized
View(bike_data)
head(bike_data)
colnames(bike_data)
glimpse(bike_data)
table(bike_data$member_casual)


#check for and drop null values
is.na(bike_data)
sum(is.na(bike_data))
bike_data <- bike_data[!sapply(bike_data, is.null)]


#create new columns as needed
#create length of ride column
bike_data$ride_length_mins <- difftime(bike_data$ended_at, bike_data$started_at, units = "mins")
bike_data$ride_length_mins <- as.numeric(bike_data$ride_length_mins)


#create distance of ride column
bike_data <- mutate(bike_data, ride_distance_miles = distHaversine(cbind(start_lng, start_lat),
cbind(lag(end_lng), lag(end_lat)), r = 3958.8))
bike_data$ride_distance_miles <- as.numeric(bike_data$ride_distance_miles)


#create weekday column
bike_data$weekday <- strftime(bike_data$started_at, '%A')


#create month column
bike_data$month <- format(as.Date(bike_data$started_at), format = "%B-%Y")


#change column names for clarity
#change "member_casual" column to "rider_type"
names(bike_data)[names(bike_data) == "member_casual"] <- "rider_type"

#change "rideable_type" column to "bike_type"
names(bike_data)[names(bike_data) == "rideable_type"] <- "bike_type"


#check data
head(bike_data)
```

```
str(bike_data)


#sort and filter data as needed
bike_data %>%
  arrange(ride_length_mins)


bike_data %>%
  arrange(ride_distance_miles)

bike_data <- bike_data %>%
  filter(ride_length_mins > 0)


bike_data <- bike_data %>%
  filter(ride_distance_miles > 0)


#summarize data
bike_data %>%
  group_by(rider_type) %>%
  summarize(mean_ride_length = mean(ride_length_mins), mean_ride_distance =
mean(ride_distance_miles))


#determine most popular start and end stations by rider type
casual_rider <- bike_data %>%
  filter(rider_type == "casual")

data <- count(casual_rider, 'start_station_name')

head(sort(data$start_station_name, decreasing = TRUE), n = 10)


data.end <- count(casual_rider, 'end_station_name')

head(sort(data.end$end_station_name, decreasing = TRUE), n = 10)



member_rider <- bike_data %>%
  filter(rider_type == "member")
```

```
data.mem <- count(member_rider, 'start_station_name')

head(sort(data.mem$start_station_name, decreasing = TRUE), n = 10)


data.mem.end <- count(member_rider, 'end_station_name')

head(sort(data.mem.end$end_station_name, decreasing = TRUE), n = 10)



#visualize data
#create bar chart to compare rider types: members versus casual riders
ggplot(data = bike_data) +
  geom_bar(mapping = aes(x = rider_type, fill = rider_type)) +
  labs(title = "Counts of Cyclistic Bike Riders by Type", caption = "Source: Motivate International Inc.")


#create boxplots to compare rider types
ggplot(data = bike_data) +
  geom_boxplot(mapping = aes(x = rider_type, y = ride_length_mins, color = rider_type)) +
  labs(title = "Boxplot Comparison of Cyclistic Bike Riders' Length of Ride in Minutes by Rider Type",
caption = "Source: Motivate International Inc.")


ggplot(data = bike_data) +
  geom_boxplot(mapping = aes(x = rider_type, y = ride_distance_miles, color = rider_type)) +
  labs(title = "Boxplot Comparison of Cyclistic Bike Riders' Distance of Ride in Miles by Rider Type",
caption = "Source: Motivate International Inc.")


#create plots of rider use by day of the week and month
bike_data$weekday_order <- factor(bike_data$weekday, levels = c("Monday", "Tuesday",
"Wednesday", "Thursday", "Friday", "Saturday", "Sunday"))

bike_data$month_order <- factor(bike_data$month, levels = c("April-2020", "May-2020", "June-2020",
"July-2020", "August-2020", "September-2020", "October-2020", "November-2020", "December-2020",
"January-2021", "February-2021", "March-2021", "April-2021"))



ggplot(data = bike_data) +
  geom_bar(mapping = aes(x = weekday_order, fill = rider_type), position = "dodge") +
  labs(title = "Counts of Cyclistic Bike Riders by Day of Week by Rider Type", caption = "Source: Motivate International Inc.")


ggplot(data = bike_data) +
  geom_bar(mapping = aes(x = month_order, fill = rider_type), position = "dodge") +
  labs(title = "Counts of Cyclistic Bike Riders by Month by Rider Type", caption = "Source: Motivate International Inc.") +
  theme(axis.text.x = element_text(size = 7, angle = 45))



#create plots of rider use by bike type
ggplot(data = bike_data) +
  geom_bar(mapping = aes(x = bike_type, fill = rider_type), position = "dodge") +
  labs(title = "Counts of Cyclistic Bike Riders by Bike Type by Rider Type", caption = "Source: Motivate International Inc.")


#create scatterplot to compare ride time and ride distance between rider types
ggplot(data = bike_data) +
  geom_point(mapping = aes(x = ride_distance_miles, y = ride_length_mins, color = rider_type)) +
  facet_wrap(~rider_type) +
  labs(title = "Relationship of Cyclistic Bike Riders' Time and Distance Traveled by Rider Type", caption = "Source: Motivate International Inc.")
```