Math 660

Sep 2023

Name: Krisha Shah

1. How many bike rides are recorded in the data set?

```
bike_ride_record <- sum(membership_data$Trips.today)
cat("Number of bike rides recorded in the dataset are:", bike_ride_record, "\n")
```

```
## Number of bike rides recorded in the dataset are: 3581986
```

Comment: Here I am using the number of rows of the entire data to find the bike records. Similarly, we can also use number of Bike.Id unique values to find the same record number.

2. What proportion of bike rides are made by male riders? What proportion by female riders?

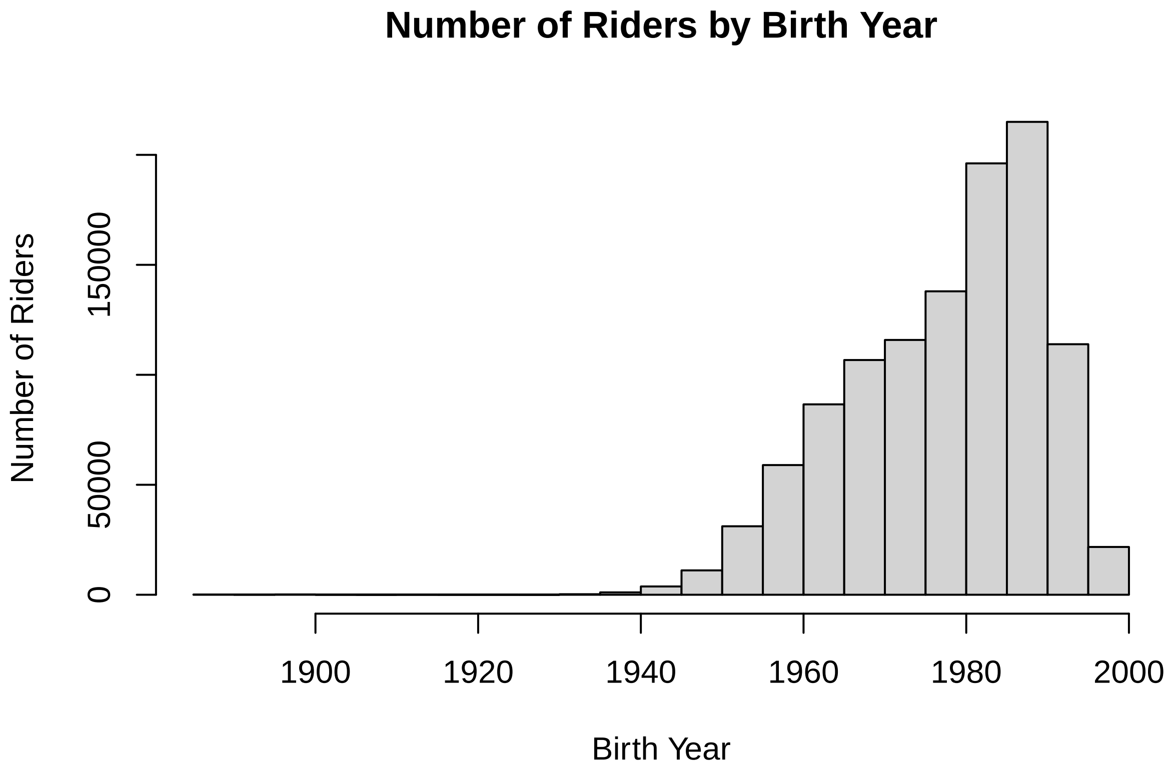
```
male_riders <- sum(data$Gender == "1")
total_riders <- sum(data$Gender)
proportion_male_riders <- male_riders / total_riders

female_riders <- sum(data$Gender == "2")
proportion_female_riders <- female_riders / total_riders

cat("Proportion of bike rides made by male riders is ", proportion_male_riders, "and that of female rid
```

```
## Proportion of bike rides made by male riders is  0.6064485 and that of female riders is  0.1967757
```

Comment: First step is to find the number of records with gender=1 for male riders and gener=2 for female riders. Later to find the proportion, we just divide each record by the total number of bike ride records found in the previous question.

3. Produce a plot showing the number of riders (y axis) by birth year (x axis).

```
histogram <- hist(data$Birth.Year,
                  main = "Number of Riders by Birth Year",
                  xlab = "Birth Year",
                  ylab = "Number of Riders")
```



Provide summary statistics of the age of the riders.

```
data$age <- 2023 - data$Birth.Year

age_summary_statistics <- summary(data$age)
print(age_summary_statistics)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
##   23.00   36.00   43.00   45.16   53.00  138.00   96366
```

We created a histogram plot showcasing the number of riders according to their birth year. Later to calculate their age, we subtract the birth year from current year (2023). And a summary of their age shows that we have riders aging from 23 to 138. The mean age of riders is approximately 45 and the median is 43.

Comment on any unusual values: Rider born in the years 1985 to 1990 are of the age 38-33. Thus, it makes sense that age has the maximum number of drivers. The number increases gradually throughout all these years, but a sudden drop in the year 1990-1995 is unusual.

4. What is the proportion of customers who are non-annual subscribers? What proportion of rides are made by non-annual subscribers?

```
unique_Cust <- table(data$User.Type)
userType <- data.frame(unique_Cust)
non_annual_customers <- bike_ride_record - userType[c(3),c("Freq")]
proportion_non_annual_customers <- non_annual_customers / bike_ride_record
cat("Proportion of non-annual subscribers is:", proportion_non_annual_customers, "\n")
```

```
## Proportion of non-annual subscribers is: 0.6933355
```

```
total_rides <- sum(membership_data$Trips.today)
proportion_rides_by_non_annual_customers <- non_annual_customers / total_riders
cat("Proportion of rides made by non-annual subscribers is:", proportion_rides_by_non_annual_customers)
```

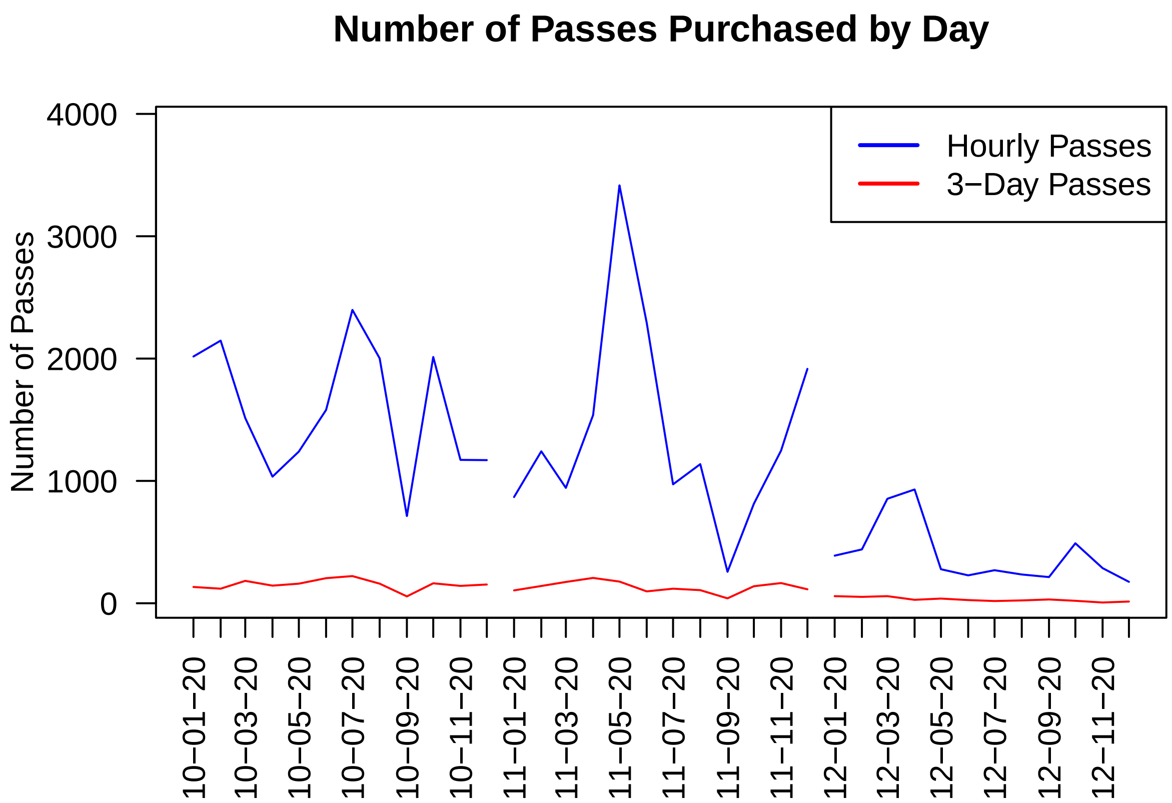
```
## Proportion of rides made by non-annual subscribers is: 1.819067
```

Comment: To calculate the proportion of non-annual customers, we need to find the sum of total customers including annual, hourly and three day pass. Later divide the sum of hourly and three day pass customers from the total customers.

Similarly for calculating the proportion of rides made by non-annual subscribers, we have to find the total rides made on each day and divide it from the former.

5. Produce a time plot of the number of 24-hr pass purchased by day. On the same plot, include a time plot of the number of 3-day pass purchased.

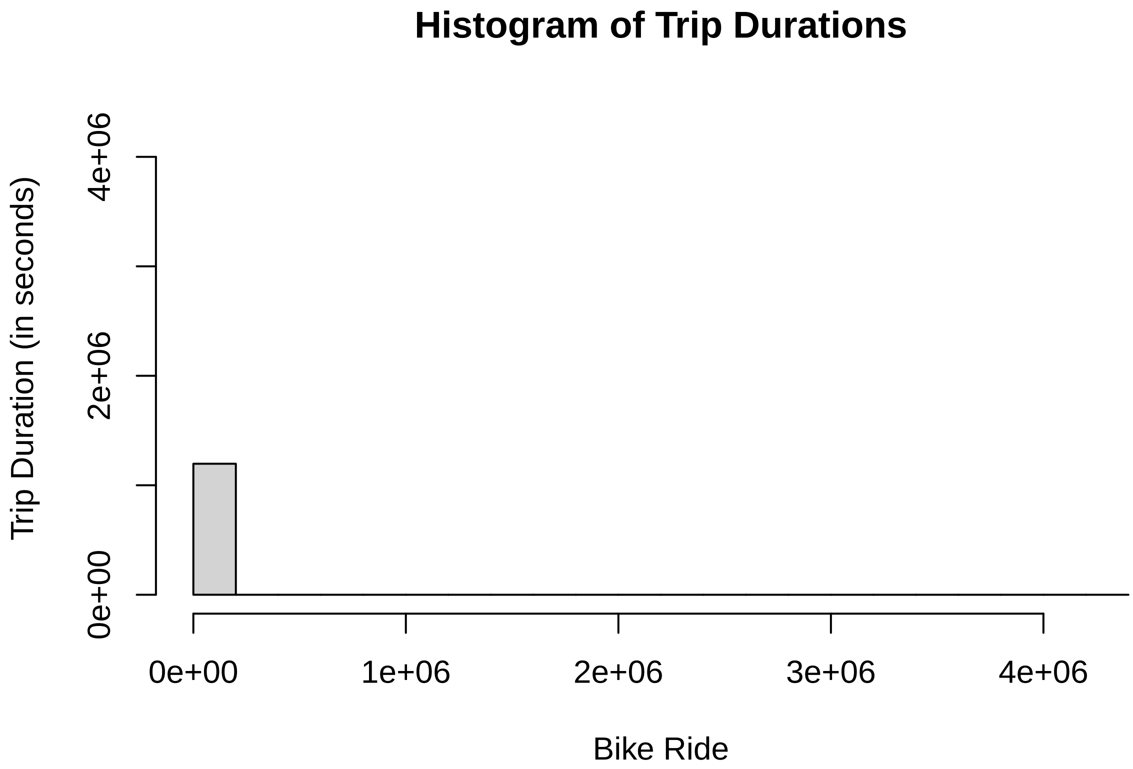
```
membership_data$Date <- as.Date(membership_data$Date)
par(las =2)

plot(membership_data$Date,
     membership_data$X24.Hour.Passes.Purchased.Today,
     type = "l",
     col = "blue",
     xlab = " ",
     ylab = "Number of Passes",
     main = "Number of Passes Purchased by Day",
     xaxt = "n")

lines(membership_data$Date,
      membership_data$X3.Day.Passes.Purchased.Today,
      col = "red")

axis(1, at = membership_data$Date, labels = membership_data$Date)

legend("topright", legend = c("Hourly Passes", "3-Day Passes"), col = c("blue", "red"), lwd = 2)
```



Comment on your plot: We can observe that the number of 24 hour passes are higher than the number of 3-day passes each day. The legend on the top of the plot shows the color coding line plot for each of them.

6. The data set contains information about the duration of each bike trip. Comment on these values. Make a histogram of the duration and provide summary statistics. Comment on your findings.

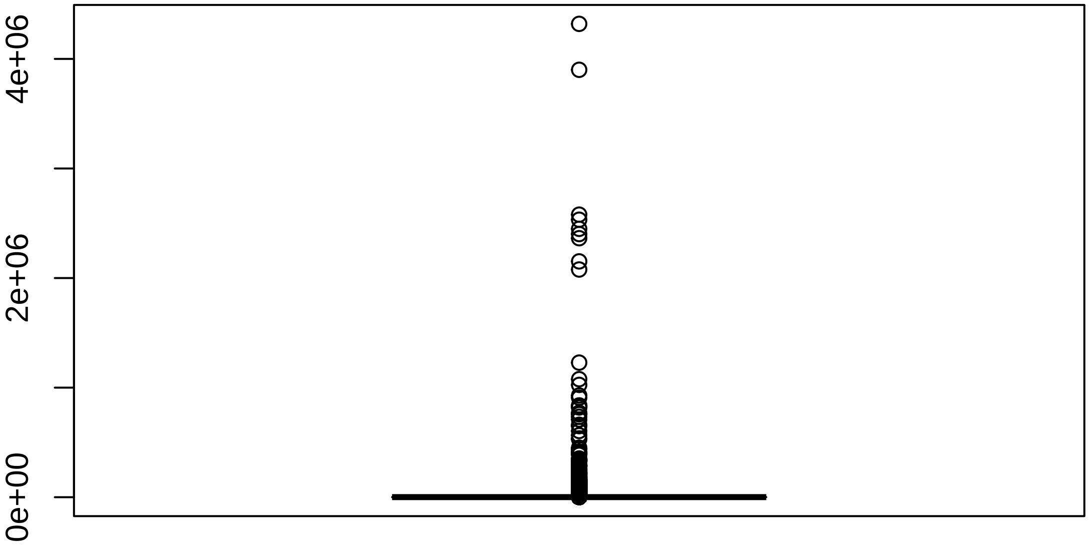
```
histogram <- hist(data$Trip.Duration, ylim=c(0,max(data$Trip.Duration)),
                  main = "Histogram of Trip Durations",
                  xlab = "Bike Ride",
                  ylab = "Trip Duration (in seconds)")
```



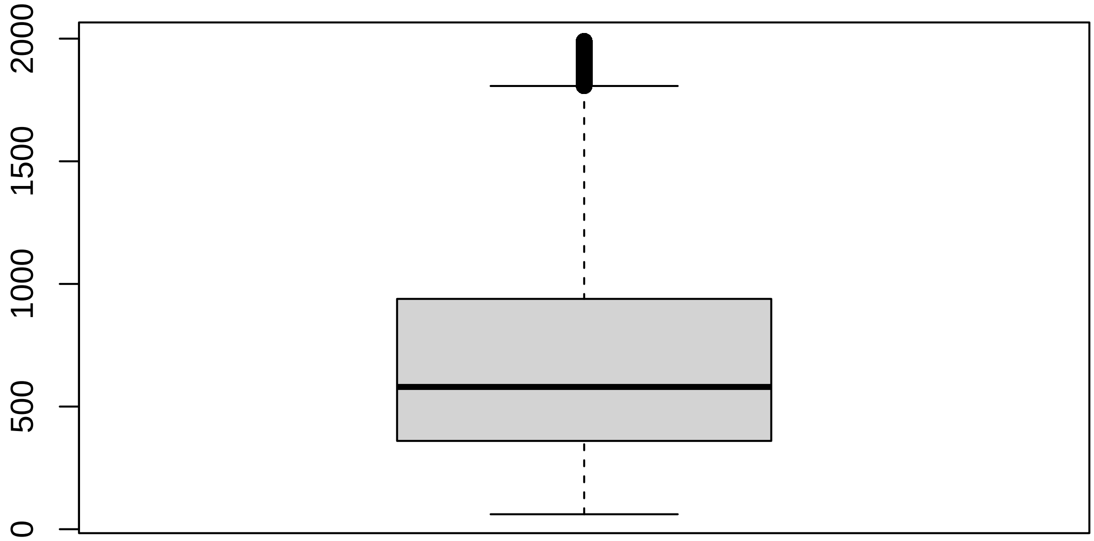
```
# Summary Statistics
summary_statistics <- summary(data$Trip.Duration)
print(summary_statistics)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      61     369     605     879    1017 4319753
```

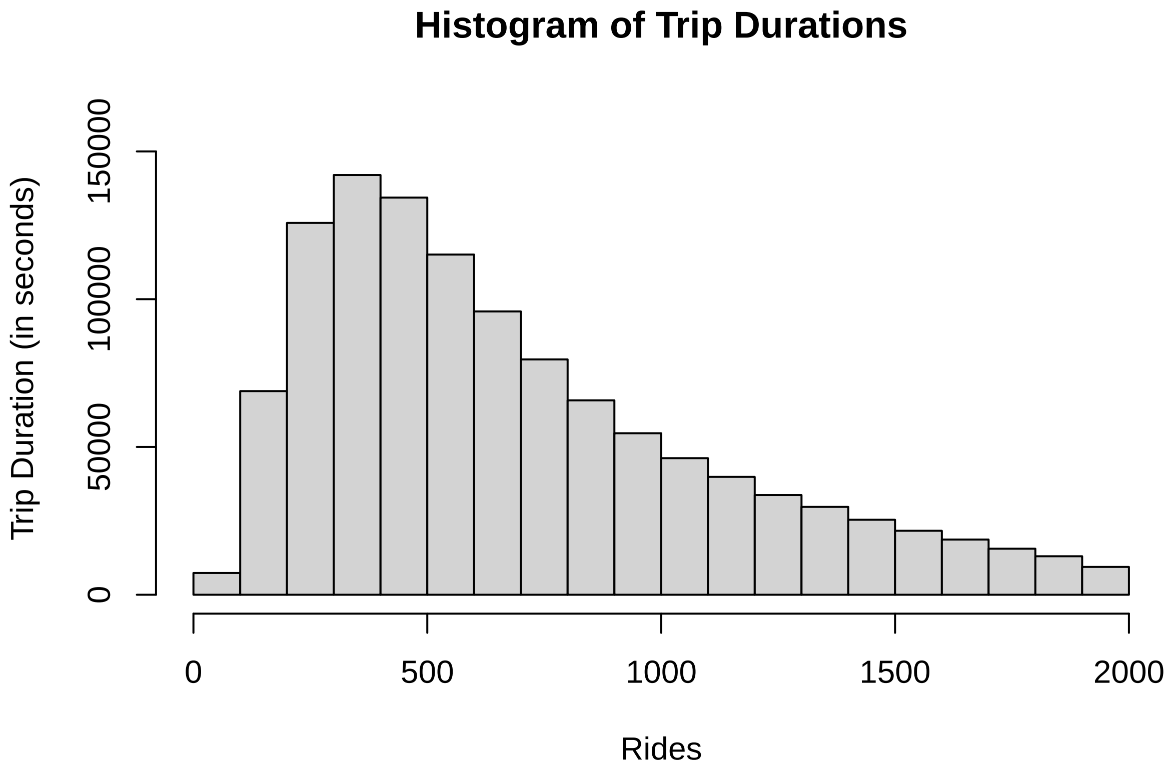
```
x=data$Trip.Duration
boxplot(x)
```



```r
# Removing outliers
x_out_rm <- x[!x %in% boxplot.stats(x)$out]
boxplot(x_out_rm)
```



```
histogram <- hist(x_out_rm, ylim=c(0,160000),
                  main = "Histogram of Trip Durations",
                  xlab = "Rides",
                  ylab = "Trip Duration (in seconds)")
```

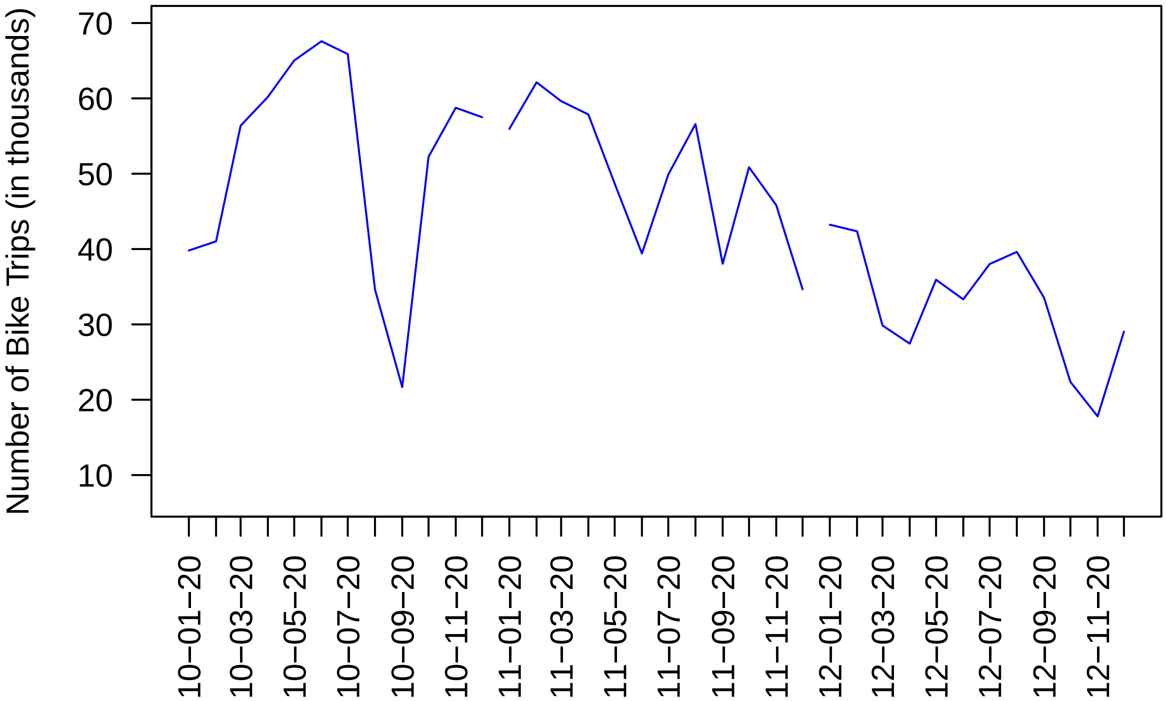


Comment: The bike trip duration typically ranges from 61 seconds to 4319753 seconds. The mean however is in the hundreth unit range. Thus the maximum values looks like an outlier. If we plot the graph considering the range from minimum to maximum values, the graph will not potray the right results. It gives us a deceiving interpretation. Thus I tried to remove the outlier values and print the second histogram for a clear interpretation.

Reference: https://statisticsglobe.com/remove-outliers-from-data-set-in-r

7. Produce a time plot of the number of bike trips by day, from Oct 1, 2016 to Dec 31, 2016. Provide summary statistics.

```
par(las =2)
plot(membership_data$Date,
     membership_data$Trips.today/1000,
     type = "l",
     col = "blue",
     xlab = " ",
     ylab = "Number of Bike Trips (in thousands)",
     xaxt = "n")

axis(1, at = membership_data$Date, labels = membership_data$Date)
```



```
# Summary Statistics
summary_statistics <- summary(membership_data$Trips.today)
print(summary_statistics)
```

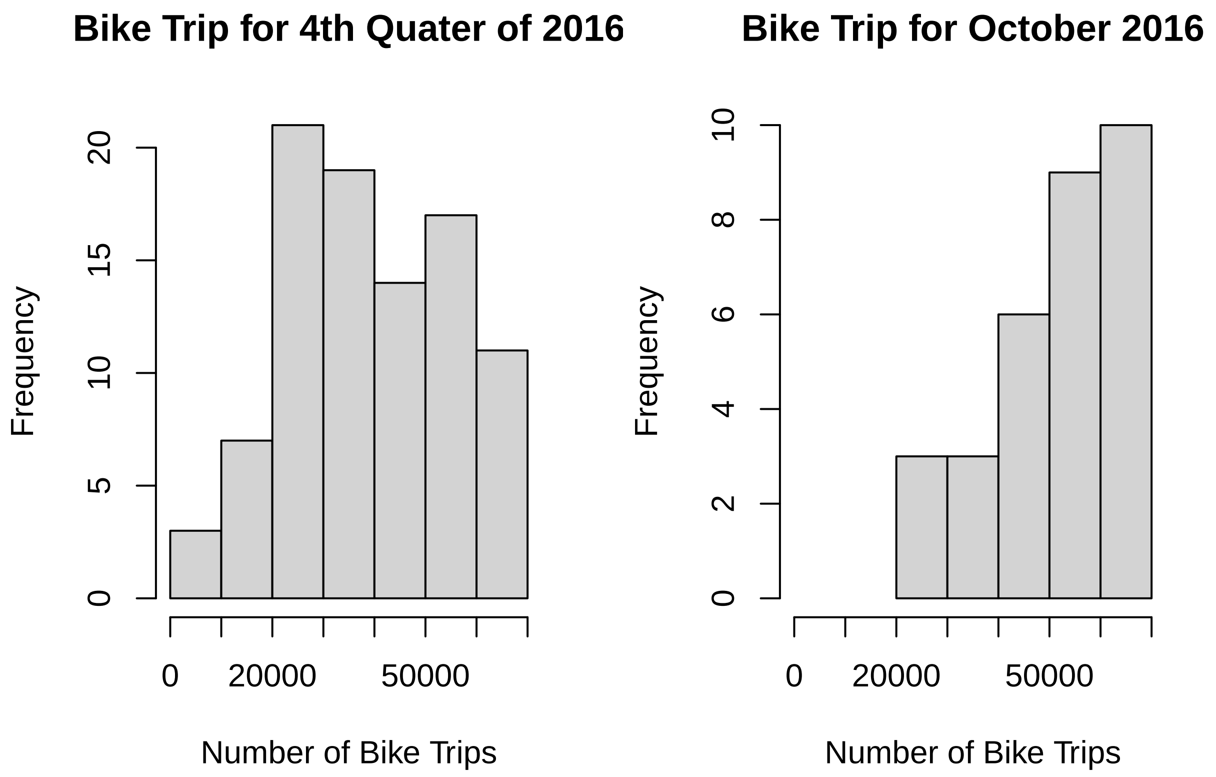
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6993   25226   38537   38935   52516   69758
```

Comment: The summary statistics shows the min, max, median and mean values for the bike rides each day. And the same result can be seen visually in the line plot.

8. Produce a histogram of the daily number of trips for the 4th quarter of 2016, and another histogram for only Oct 2016. Make it so that both histograms show the same range on the x-axis.

```
par(mfrow = c(1, 2))
membership_data$Date <- as.Date(membership_data$Date)

x_limits <- c(0, 70000)

histogram1 <- hist(membership_data$Trips.today,
                   main = "Bike Trip for 4th Quater of 2016",
                   xlab = "Number of Bike Trips",
                   xlim = x_limits)
```

```r
histogram2 <- hist(Oct_data$Trips.today,
                   main = "Bike Trip for October 2016",
                   xlab = "Number of Bike Trips",
                   xlim = x_limits)
```



Comment: I referred for ways to group data as per date. But found it difficult to do without using dplyr package. Hence went with another method.

Reference: https://stackoverflow.com/questions/49669862/how-to-group-by-in-base-r

9. In addition, answer the following questions (you might not be able to answer some of the questions using these two datasets; if so, just point that out). Use text to answer the questions. For (c) to (f) however, in addition to text, also include the appropriate R code in your Rmd file, so that your pdf file also shows these answers as a result of running the R code.

(a) Do any of the variables have unusual values? Which ones, and why are the values unusual?

Answer: Trip Duration (in seconds) in the dataset 201611-citibike-tripdata has unusual/high values that fall into outliers. These values totally disrupt the mean data. The subscribers column has certain blank values which leads to unusual data. Gender variable in 201611-citibike-tripdata has unusual values marked as 0. They neither fall into male nor female category.

(b) On which day were there the most number of unique riders?

Answer: Did not quite understand the term unique riders. We need more data on the term unique specifically pointing on a variable. Then it is possible to find the unique values in that variable.

(c) How many unique Citi Bike stations were there in the dataset ?

Answer: There were 606 unique Citi Bike Stations in the dataset.

```r
unique_Citi_Bike_Station <- unique(data$Start.Station.Name)
cat(length(unique_Citi_Bike_Station))
```

```
## 606
```

(d) How many unique Citi Bikes were used at least once in the dataset?

Answer: There are 10082 unique Citi Bikes used at least once in the dataset.

```r
unique_Citi_Bike <- unique(data$Bike.ID)
cat(length(unique_Citi_Bike))
```

```
## 10082
```

(e) On which day(s) were the most 3 day passes purchased?

Answer: Maximum 3 day passes were sold on 10/17/2016

```r
max_3_day_pass <- max(membership_data$X3.Day.Passes.Purchased.Today)
selected_date <- membership_data$Date[membership_data$X3.Day.Passes.Purchased.Today == max_3_day_pass]
```

(f) How many bike "trips" were there from the launch of the Citi Bike program till Sep 30, 2016?

Answer: There were a total of 34619850 bike trips from the launch of Citi Bike program till Spt 30. To find this value, we take the cumulative trips on Oct 1 minus the total trips on Oct 1.

```
Sept30 <- membership_data[1,3] - membership_data[1,2]
cat(Sept30)
```

```
## 34619850
```