Data Science Challenge: Trips!

Question 1

Programmatically download and load into your favorite analytical tool the trip data for September 2015.
Report how many rows and columns of data you have loaded.

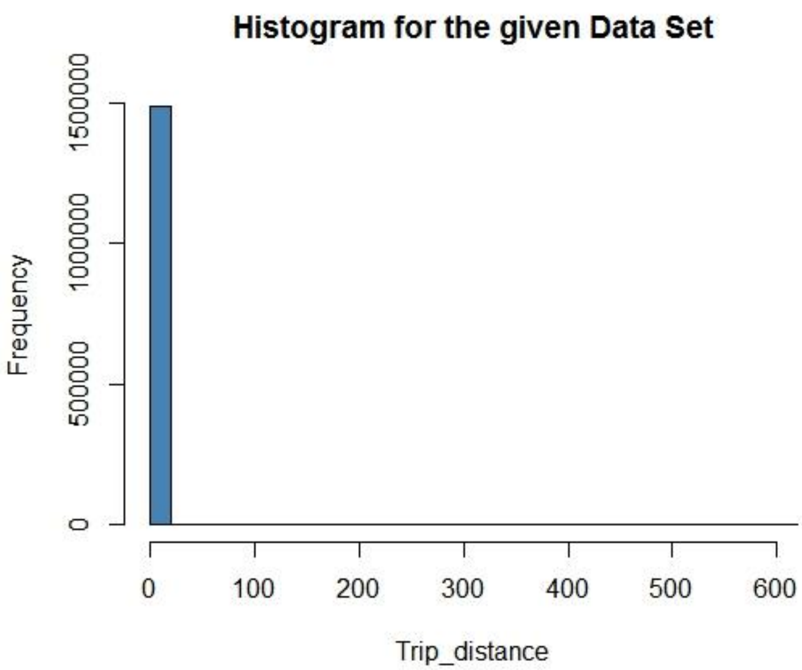
```
> download.file("https://s3.amazonaws.com/nyc-tlc/trip+data/green_tripdata_2015
-09.csv", "green_taxi.csv")

trying URL 'https://s3.amazonaws.com/nyc-tlc/trip+data/green_tripdata_2015-09.c
sv'
Content type 'application/octet-stream' length 239035648 bytes (228.0 MB)
downloaded 228.0 MB

> green_taxi<-read.csv("green_taxi.csv")
> dim(green_taxi)

[1] 1494926      21
```

The given data set (green taxi, September 2015) contains 1494926 rows and 21 columns

Question 2

Plot a histogram of the number of the trip distance ("Trip Distance").
Report any structure you find and any hypotheses you have about that structure.

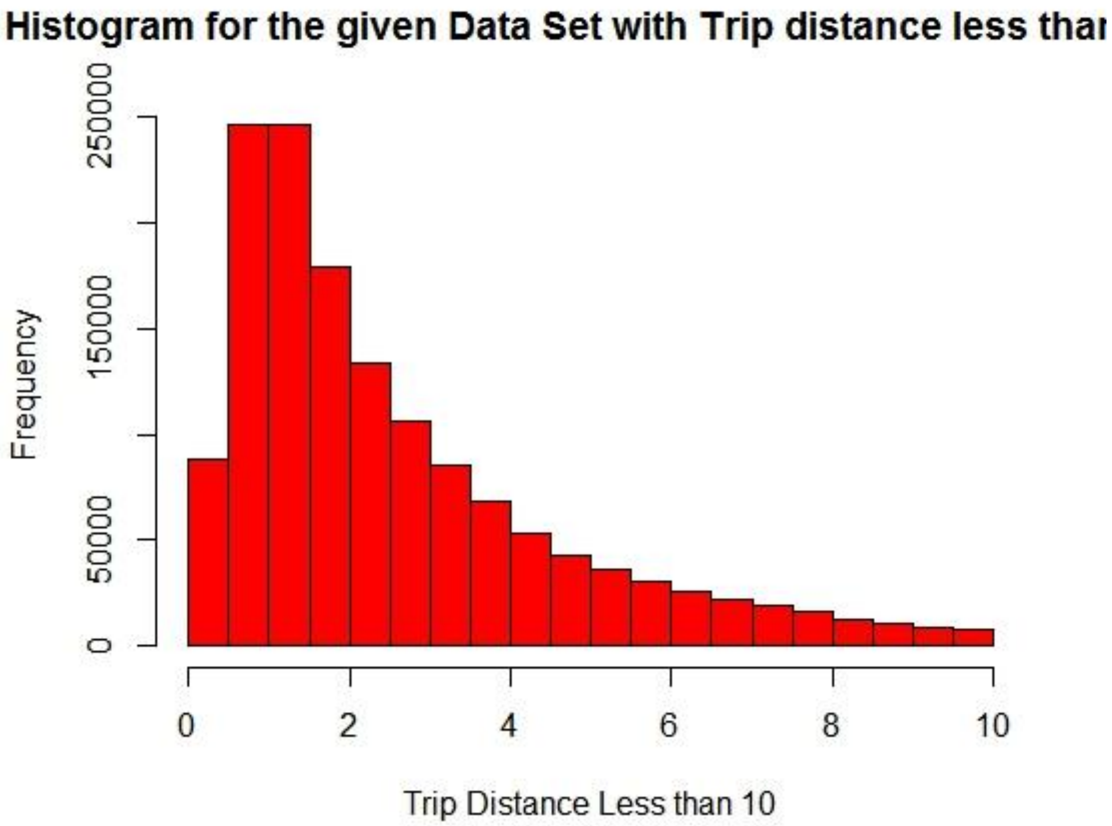
```
> hist(Trip_distance,main = "Histogram for the given Data Set",col=c("steelblu
e"))

> range(Trip_distance)
[1]   0.0 603.1
> summary(Trip_distance)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.100   1.980   2.968   3.740 603.100
```



The initial histogram plot of Trip_distance shows the maximum frequency in the range 0 to 10. But the above statistics show max value to be 603.1 From the above plot it is clearly evident that values above 10 i.e trip_distance greater than 10 are outliers.

```
> green_taxi_ten <- green_taxi[Trip_distance < 10,]
```

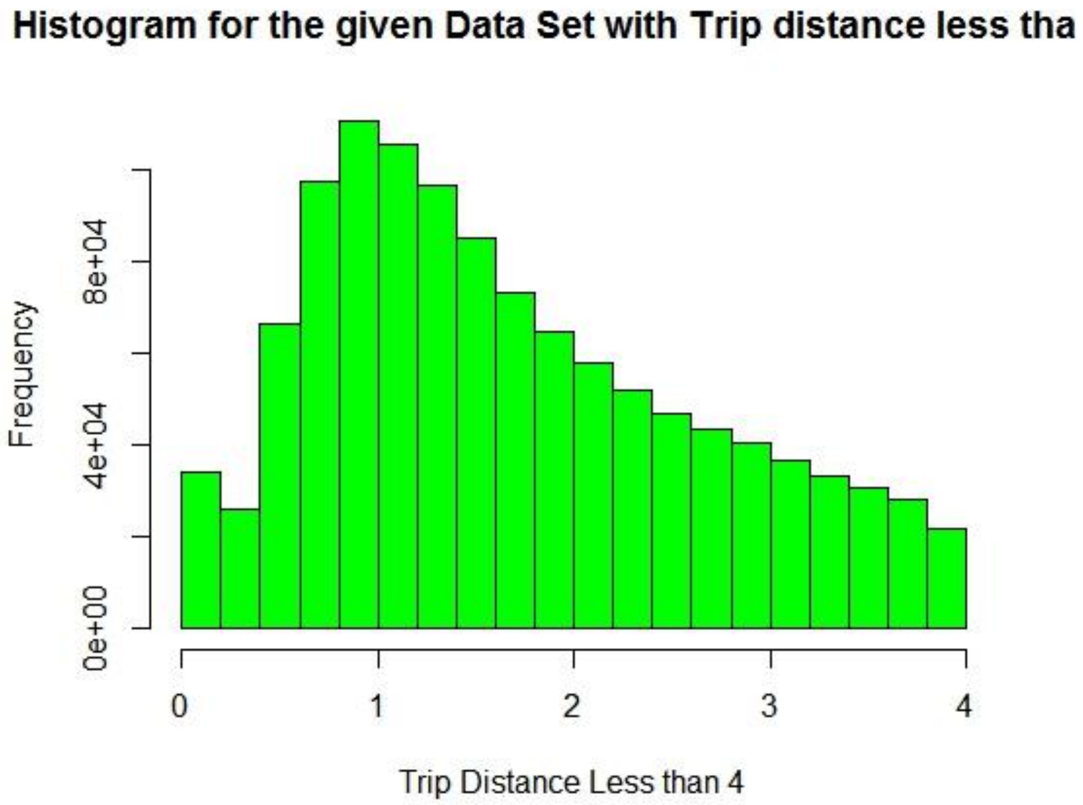
```
> hist(green_taxi_ten$Trip_distance,main="Histogram for the given Data Set with Trip d
istance less than 10",xlab="Trip Distance Less than 10",col=c("red"))
```



Since the previous histogram plot clearly gave an evidence that the maximum frequency was between o to 10. The above plot is a histogram for trip distance less than 10 units

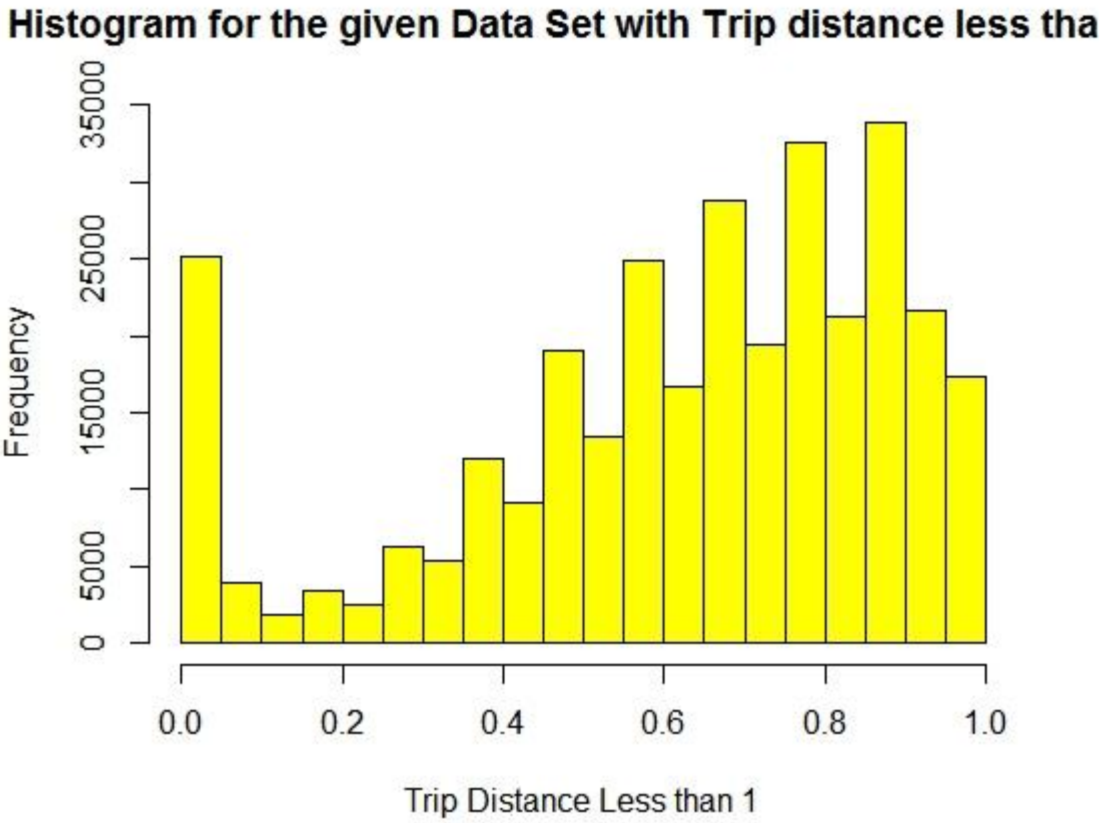
The above plot further tells that within the range 0 to 10, there is higher frequency for trip distance between 0.8 to 1.6

```
> green_taxi_four <- green_taxi[Trip_distance < 4,]
> hist(green_taxi_four$Trip_distance,main="Histogram for the given Data Set with Trip
distance less than 4",xlab="Trip Distance Less than 4",col=c("green"))
```



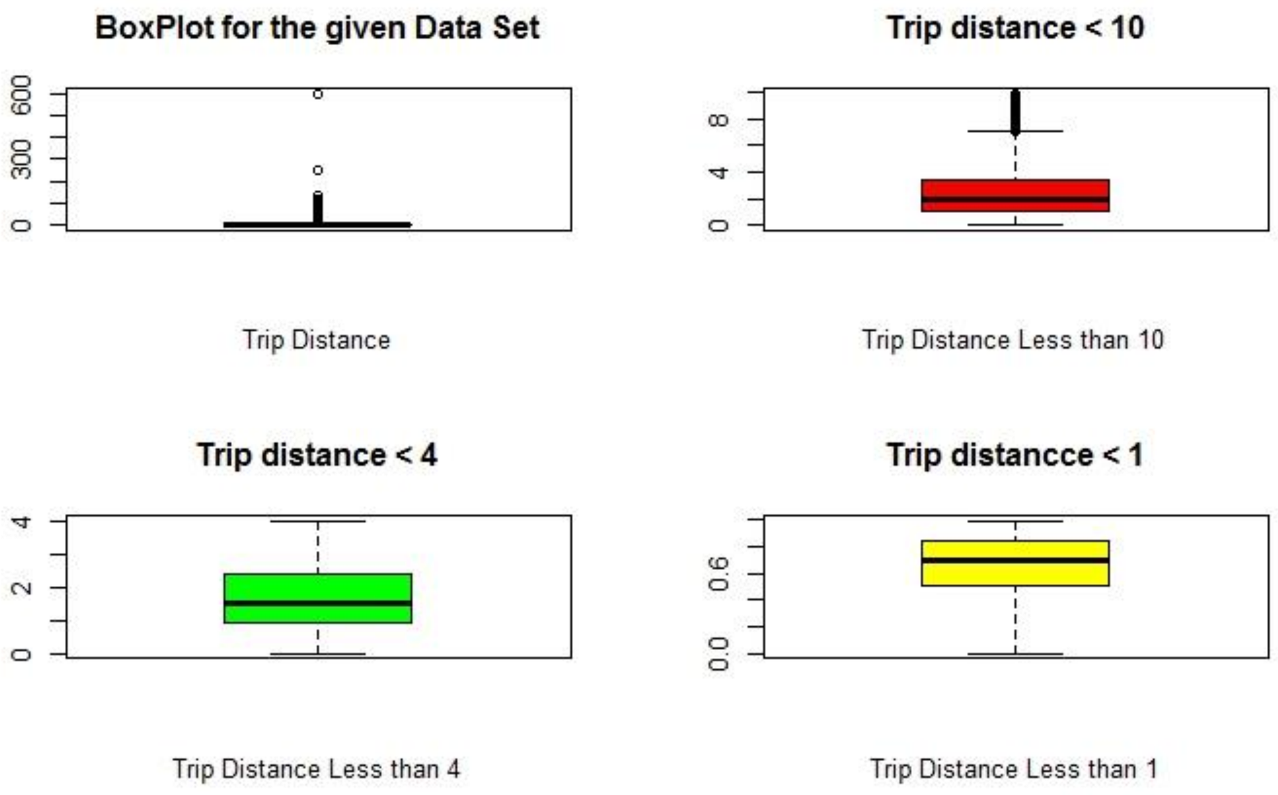
The above plot is a histogram for trip distance less than 4 units. The highest frequency here is between 0.8 a nd 1-unit distance.

```
> green_taxi_one <- green_taxi[Trip_distance < 1,]
> hist(green_taxi_one$Trip_distance,main="Histogram for the given Data Set with Trip d
istance less than 1",xlab="Trip Distance Less than 1",col=c("yellow"))
```



For getting a better estimate, on plotting a histogram for trip distance less than 1 unit. It is observed that the maximum number trips range between distances of 0.85 to 0.90

```
> par(mfrow=c(2,2))

> boxplot(Trip_distance,main = "Histogram for the given Data Set",col=c("steelblue"))
> boxplot(green_taxi_ten$Trip_distance,main="Histogram for the given Data Set with Tri
p distance less than 10",xlab="Trip Distance Less than 10",col=c("red"))
> boxplot(green_taxi_four$Trip_distance,main="Histogram for the given Data Set with Tr
ip distance less than 4",xlab="Trip Distance Less than 4",col=c("green"))
> boxplot(green_taxi_one$Trip_distance,main="Histogram for the given Data Set with Tri
p distance less than 1",xlab="Trip Distance Less than 1",col=c("yellow"))
```



The above plot is a box plot of the 4 different conditions for which the four different histograms were plotted above.

Question 3

Report mean and median trip distance grouped by hour of day.
We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fair, and any other interesting characteristics of these trips.

```
> green_taxi_hour <- green_taxi
> green_taxi_hour$Lpep_dropoff_datetime <- as.numeric(substr(as.character(gree
n_taxi_hour$Lpep_dropoff_datetime),12,13))
> mean_hour_dist <- c()
> med_hour_dist <- c()
> returnMeanTripDistance <-function(){
+
+   for (a in c(0:23)) {
+     mean_hour_dist <<- c(mean_hour_dist , mean(green_taxi_hour[green_taxi_ho
ur$Lpep_dropoff_datetime == a,]$Trip_distance))
+     med_hour_dist <<- c(med_hour_dist , median(green_taxi_hour[green_taxi_ho
ur$Lpep_dropoff_datetime == a,]$Trip_distance))
+
+   }
+
+ }
> returnMeanTripDistance()
> mean_hour_dist<-data.frame(c(0:23),mean_hour_dist,med_hour_dist)
> colnames(mean_hour_dist) <- c("Hour","MeanDistance","MedianDistance")
> View(mean_hour_dist)
```

mean_hour_dist

	Hour	MeanDistance	MedianDistance
1	0	3.239356	2.28
2	1	3.130531	2.20
3	2	3.143715	2.21
4	3	3.225473	2.26
5	4	3.463798	2.37
6	5	4.187841	2.92
7	6	3.932660	2.76
8	7	3.217171	2.05
9	8	2.923779	1.87
10	9	3.016995	1.98
11	10	3.010999	1.98
12	11	2.879559	1.86
13	12	2.907513	1.90
14	13	2.889431	1.85
15	14	2.767255	1.80
16	15	2.771893	1.79
17	16	2.766404	1.80
18	17	2.690678	1.76
19	18	2.673491	1.80
20	19	2.746890	1.86
21	20	2.800049	1.92
22	21	2.931069	2.00
23	22	3.147097	2.17
24	23	3.231312	2.25

This table gives the mean and median trip distance respectively grouped by hour of day starti ng from 00 hrs to 23 hrs

Amongst the different airports in NYC, the one that I have chosen is JFK. The latitude and longitude values for which are taken from google maps.

```
> jfk <- green_taxi[(Pickup_latitude >= 40.63 & Pickup_latitude <= 40.66 & Pickup_longitude <= -73.76 & Pickup_longitude >= -73.78)|(Dropoff_latitude >= 40.63 & Dropoff_latitude <= 40.66 & Dropoff_longitude <= -73.76 & Dropoff_longitude >= -73.78),]
```

Since the coordinates of JFK airport is 40.6413,-73.7781 (source: Google Maps), I have considered the coordinates which fall within the boundary of JFK airport so that all the data points pertaining to JFK airport pick up and drop off are considered correctly

```
> dim(jfk)

[1] 2844   21
```

2844 rows match the criteria (i.e either pick up or drop off at JFK airport)

```
> avg_fair <- mean(jfk$Fare_amount)
> View(jfk)
```

Values	
avg_fair	40.4339486638537

<u>Question 4</u>

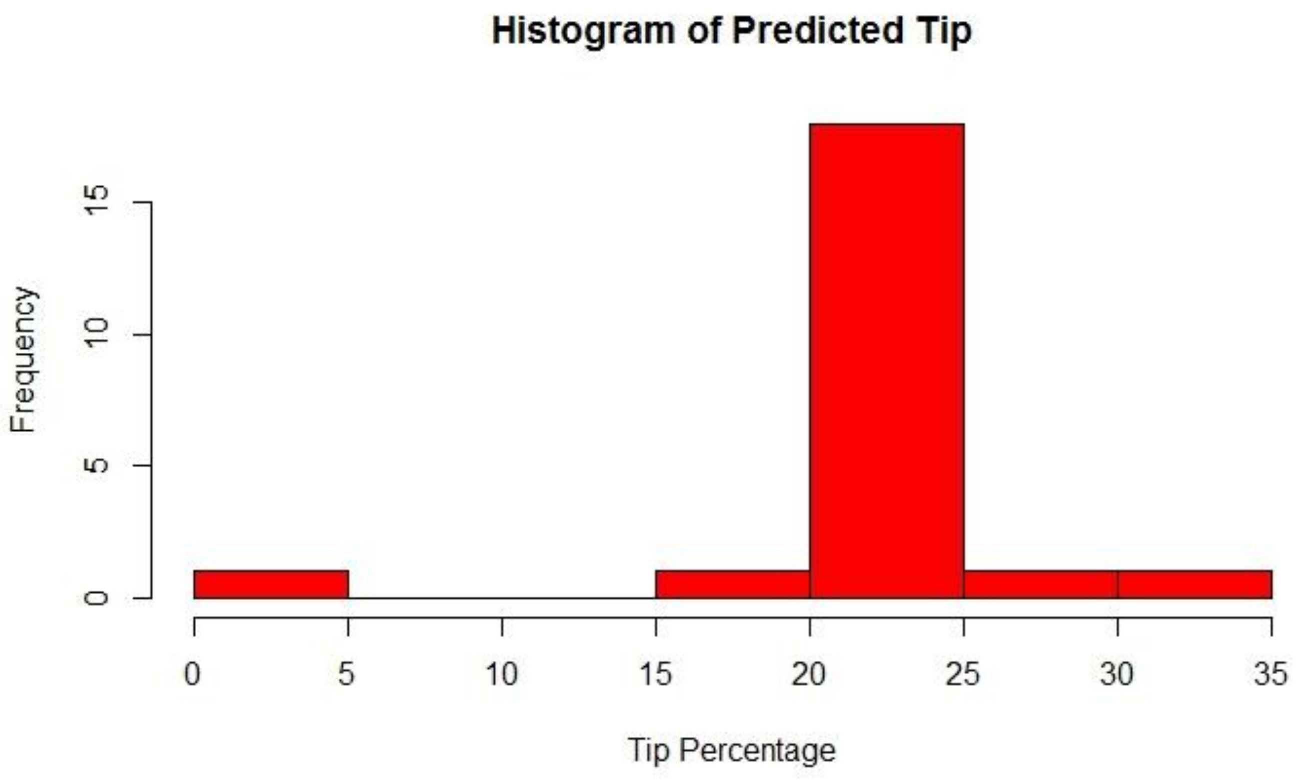
Build a derived variable for tip as a percentage of the total fare.
Build a predictive model for tip as a percentage of the total fare. Use as much of the data as you like (or all of it). We will validate a sample.

```
> tip_percentage <- Tip_amount/Fare_amount*100
> green_taxi$tip_percentage <- tip_percentage
> green_taxi_fare_amount<-green_taxi[tip_percentage>0&tip_percentage<100&Fare_amount>0,]
> summary(green_taxi_fare_amount$Fare_amount)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.05    7.50   11.50   14.27   18.00  427.00

> hist(tip_percentage)
> predicted_tip <- c()
> predictTip <-function(){
+
+   for (a in c(0:20)) {
+     tip_vector<-green_taxi_fare_amount[green_taxi_fare_amount$Fare_amount>=a&green_taxi_fare_amount$Fare_amount<a+1,]$tip_percentage
+     tip<- names(sort(table(tip_vector),decreasing=TRUE))[1]
+     predicted_tip <<- c(predicted_tip , tip)
+   }
```

```
+   tip_vector<-green_taxi_fare_amount[green_taxi_fare_amount$Fare_amount>21,]
$tip_percentage
+   tip<- names(sort(table(tip_vector),decreasing=TRUE))[1]
+   predicted_tip <<- c(predicted_tip , tip)
+ }
> predictTip()
> range <- c()
> fareRange <-function(){
+
+   for (a in c(0:20)) {
+     r<- paste(paste(a,' - '),a+1)
+     range <<- c(range , r)
+   }
+   range <<- c(range , '>21')
+ }
> fareRange()
> tip_prediction <- data.frame(range,predicted_tip)
> colnames(tip_prediction) <- c("MeanDistance","Tip")
> View(tip_prediction)
```

	MeanDistance	Tip
1	0 - 1	20
2	1 - 2	1
3	2 - 3	30.4
4	3 - 4	24.5714285714286
5	4 - 5	25.7777777777778
6	5 - 6	24.7272727272727
7	6 - 7	24.3333333333333
8	7 - 8	23.7142857142857
9	8 - 9	23.25
10	9 - 10	21.7777777777778
11	10 - 11	21.6
12	11 - 12	22.3636363636364
13	12 - 13	22.1666666666667
14	13 - 14	22
15	14 - 15	21.8571428571429
16	15 - 16	21.7333333333333
17	16 - 17	21.625
18	17 - 18	21.5294117647059
19	18 - 19	21.4444444444444
20	19 - 20	20.8421052631579
21	20 - 21	20.8
22	>21	20.6666666666667

```
> hist(as.numeric(as.character(tip_prediction$Tip)),main = "Histogram of
Predicted Tip",col=c("red"),xlab="Tip Percentage")
```



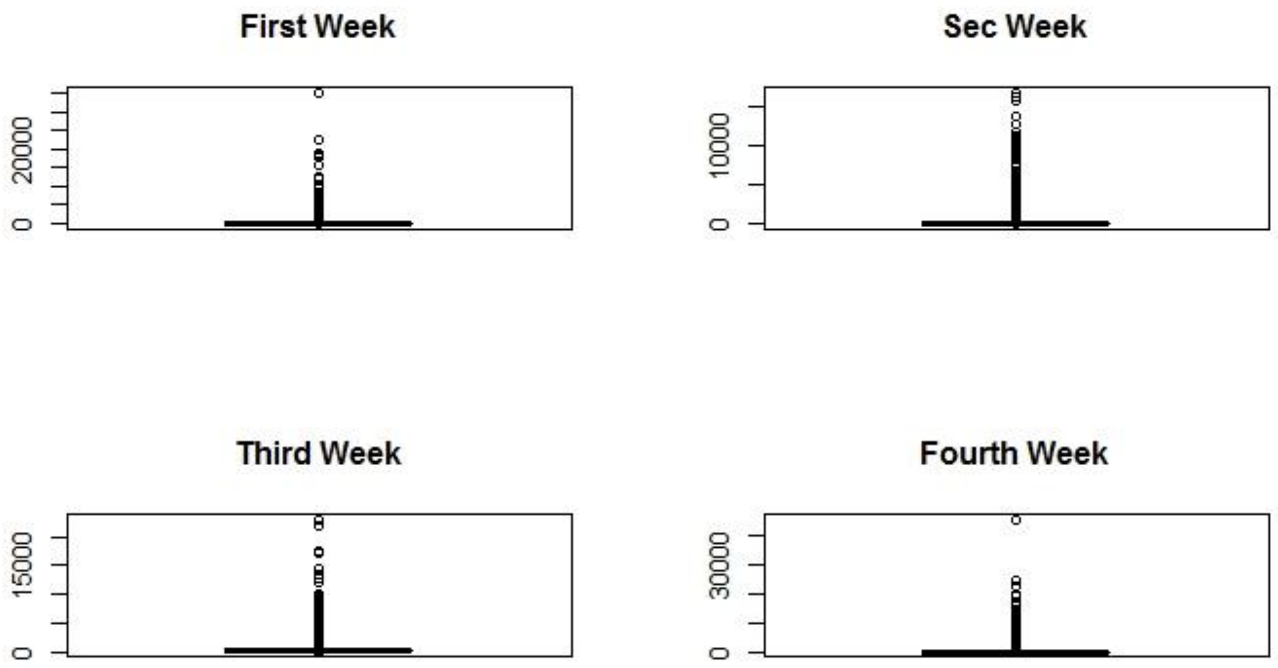
Majority of the customer pay a tip in the range of 20 to 25% of the total far according to this histogram

<u>Question 5</u>

Option A: Distributions
Build a derived variable representing the average speed over the course of a trip.
Can you perform a test to determine if the average trip speeds are materially the same in all weeks of September? If you decide they are not the same, can you form a hypothesis regarding why they differ?
Can you build up a hypothesis of average trip speed as a function of time of day?

```
> x<-as.POSIXct(as.character(green_taxi$Lpep_dropoff_datetime),format="%Y-%m-%
d %H:%M:%S")
> y<-as.POSIXct(as.character(green_taxi$lpep_pickup_datetime),format="%Y-%m-%d
%H:%M:%S")
> time<-as.numeric(difftime(x, y,units="hours"))
> green_taxi$Trip_Time <- time
> avgspeed<-green_taxi$Trip_distance/time
> day<-as.numeric(substr(as.character(green_taxi$Lpep_dropoff_datetime),9,10))
> hours<-as.numeric(substr(as.character(green_taxi$Lpep_dropoff_datetime),12,1
3))
> green_taxi$Hours <- hours
> week<-ceiling(day/7)
> avgspeed[!is.finite(avgspeed)] <- 0
> green_taxi$Average_Speed <- avgspeed
> green_taxi$Week <- as.factor(week)
> green_taxi[green_taxi$Hours==0,]$Hours<-24
> summary(green_taxi[green_taxi$Week==1,]$Average_Speed)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.00     9.75    12.23    16.23    15.74 35370.00
> summary(green_taxi[green_taxi$Week==2,]$Average_Speed)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.00     9.00    11.46    15.12    14.66 16740.00
> summary(green_taxi[green_taxi$Week==3,]$Average_Speed)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    0.000     9.107    11.520    15.200    14.700 22880.000
> summary(green_taxi[green_taxi$Week==4,]$Average_Speed)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.00     9.58    12.03    16.34    15.39 45720.00

> green_taxi[green_taxi$Average_Speed==35370,]
       VendorID lpep_pickup_datetime Lpep_dropoff_datetime Store_and_fwd_flag
RateCodeID
309325        2  2015-09-07 05:43:59   2015-09-07 05:44:01                  N
5
       Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude Pas
senger_count
309325        -73.93752        40.83756         -73.79018         40.64682
1
       Trip_distance Fare_amount Extra MTA_tax Tip_amount Tolls_amount Ehail_f
ee
309325         19.65          63     0       0       12.6            0
NA
       improvement_surcharge Total_amount Payment_type Trip_type tip_percentag
e    Trip_Time Hours
309325                     0         75.6            1         2             2
0 0.0005555556     5
       Average_Speed Week
309325         35370    1
```

```
> par(mfrow=c(2,2))
> boxplot(green_taxi[green_taxi$Week==1,]$Average_Speed,col="steelblue",main="
First Week")
> boxplot(green_taxi[green_taxi$Week==2,]$Average_Speed,col="red",main="Sec We
ek")
> boxplot(green_taxi[green_taxi$Week==3,]$Average_Speed,col="Green",main="Thir
d Week")
> boxplot(green_taxi[green_taxi$Week==4,]$Average_Speed,col="Yellow",main="Fou
rth Week")
```

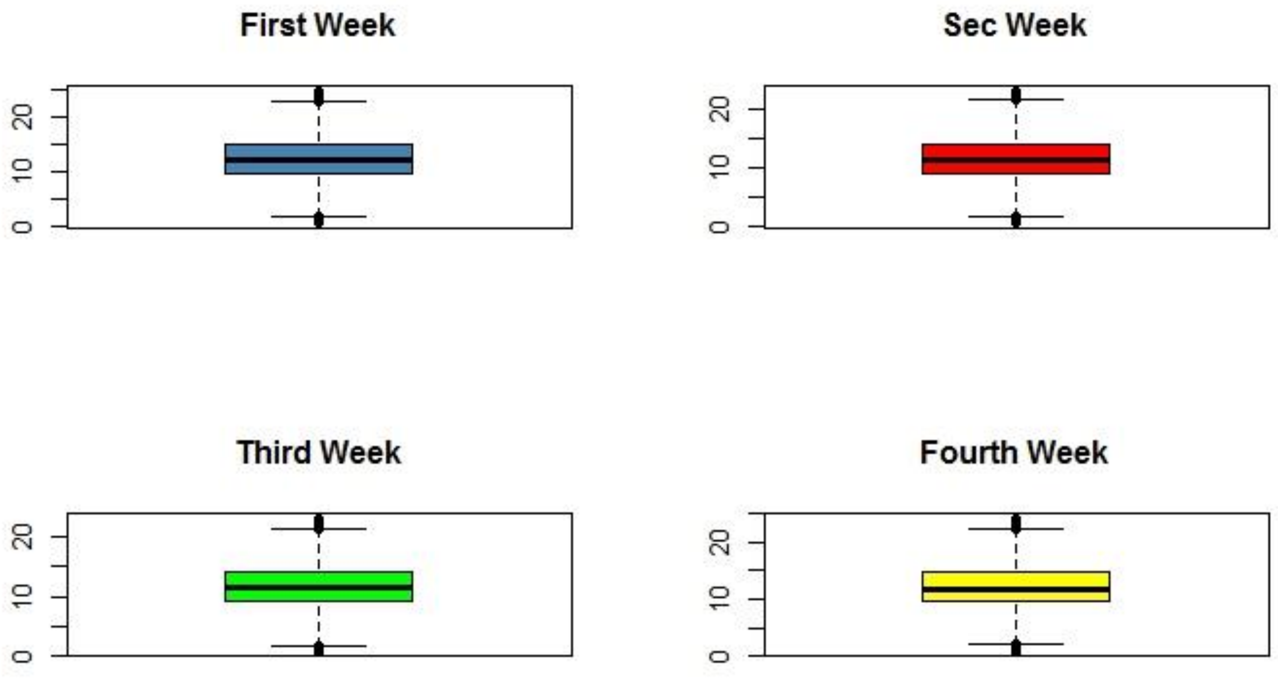


Here q,w,e and r are dataframes containing the average speed of every trip in the first,second , third and fourth week respectively

```
> summary(q)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.761   9.730  12.020  12.630  14.960  24.730
> summary(w)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5143  9.0000 11.2700 11.7300 14.0000 23.1500
> summary(e)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.720   9.125  11.340  11.820  14.070  23.090
> summary(r)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8571  9.5950 11.8500 12.4000 14.6900 24.1100
```

For a trip distance of 19.65 miles, its practically impossible to cover in 3 secs. So I have removed outliers

So below is a plot for the above statistics. I.e Box plot of q,w,e and r respectively




```
> kruskal.test(list(green_taxi[green_taxi$Week==1 ,]$Average_Speed,
+                   green_taxi[green_taxi$Week==2 ,]$Average_Speed,
+                   green_taxi[green_taxi$Week==3 ,]$Average_Speed,
+                   green_taxi[green_taxi$Week==4 ,]$Average_Speed))
```

The Kruskal–Wallis rank sum test or one-way analysis of variance gives the following result s

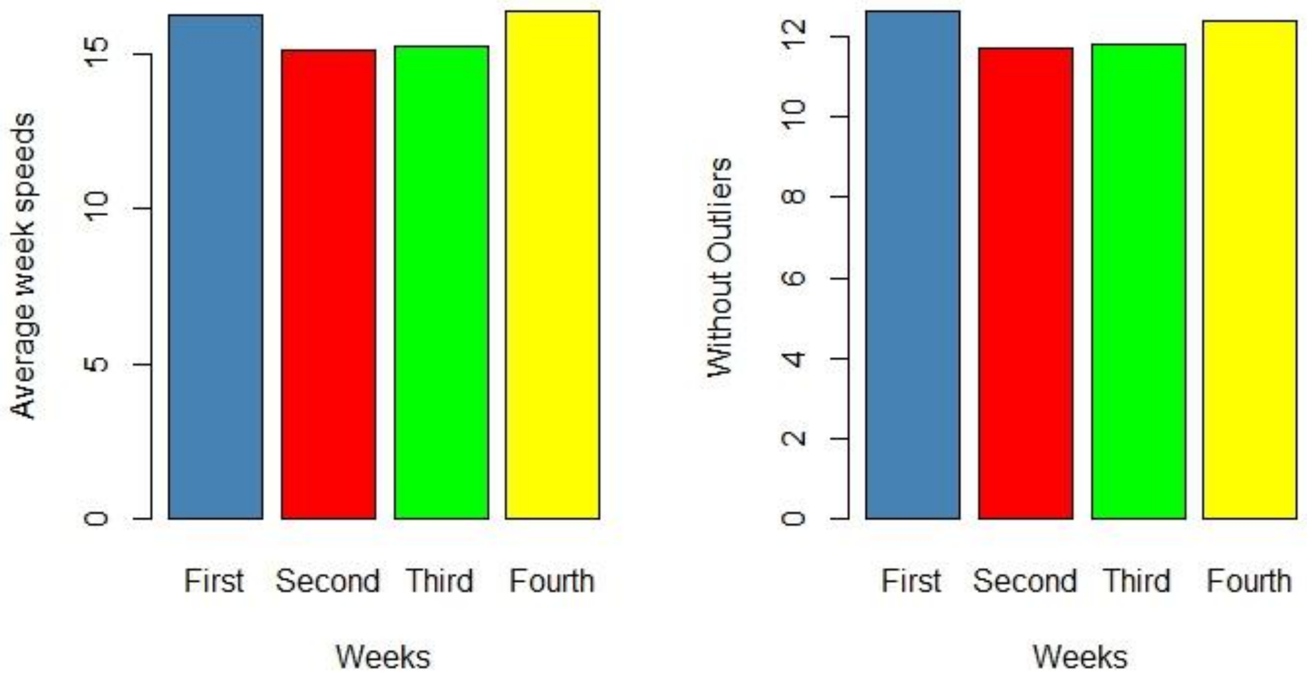
```
	Kruskal-Wallis rank sum test

data:  list(green_taxi[green_taxi$Week == 1, ]$Average_Speed, green_taxi[green_taxi$We
ek ==     2, ]$Average_Speed, green_taxi[green_taxi$Week == 3, ]$Average_Speed,     gr
een_taxi[green_taxi$Week == 4, ]$Average_Speed)
Kruskal-Wallis chi-squared = 8652.3, df = 3, p-value < 2.2e-16
```

NULL HYPOTHESIS:
There is **no significant** difference in the average speeds among the 4 weeks in September

ALTERNATE HYPOTHESIS:
There is **significan**t difference in the average speeds among the 4 weeks in September

Since the p value obtained above is clearly way below 0.1. We reject the null hypothesis at a confidence interval of 99%. i.e It is pretty clear that there is significant difference in the aver age speeds among the 4 weeks in september

Below is a plot of the mean of q,w,e and respectively. Right one is without outliers. (q,w,e an d already defined above)



Here fweek,sweek,tweek and forweek are dataframes pertaining to subsets of the main data set containing data of week1, week2, week3 and week 4 respectively

```
> dim(fweek)
[1] 341656     26
> dim(sweek)
[1] 361069     26
> dim(tweek)
[1] 363312     26
> dim(forWeek)
[1] 338708     26
```

The plot above shows that the average week speed is higher in the first and last week compar ed to the other two. The dimension statistics above also reveals that there are lesser trips in th e first and last week compared to the other two.

So, It can be inferred that since there are lesser number of trips in the first and last week (dull business period), cab drivers would have driven fast inorder to get more customers and there by more money. This is why there is a difference in the average speed among the four weeks.

Now the same Kruskal–Wallis rank sum test or one-way analysis of variance is performed fo r the mean values of average speed for every hour of the day. The results are as follows:

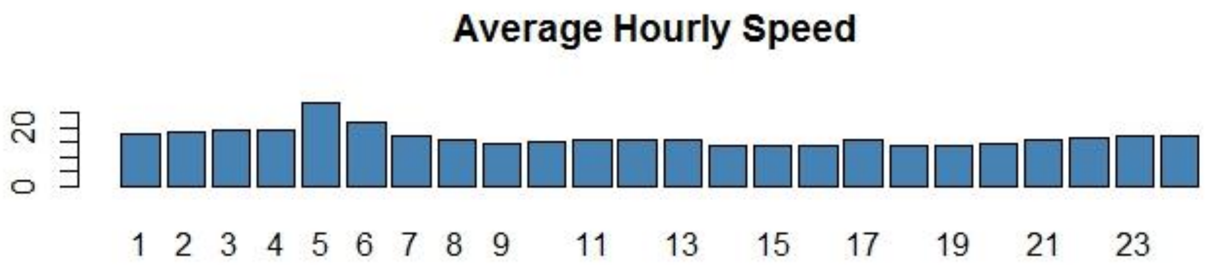
```
> kruskal.test(hourlySpeed)

        Kruskal-Wallis rank sum test

data:  hourlySpeed
Kruskal-Wallis chi-squared = 160900, df = 23, p-value < 2.2e-16
```

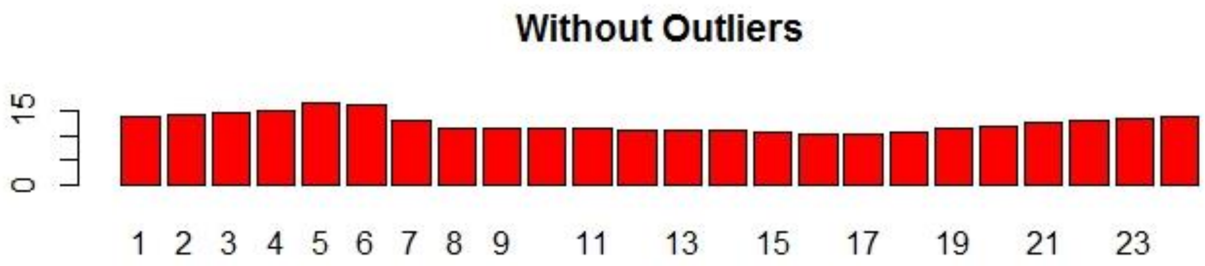
NULL HYPOTHESIS:
There is **no significant** difference in the average speeds among the 24 hours of the day

ALTERNATE HYPOTHESIS:
There is **significan**t difference in the average speeds among 24 hours of the day

Since the p value obtained is clearly below 0.1, we reject the null hypothesis at 90% confiden ce interval.

The below plot is plot of mean values of average speed for every hour of the day.





From the plots above its very evident that it hits a peak at 5:00 hrs and hits the lowest point at 18:00 hrs. This could be reasoned as traffic being very less in the morning hours at 5:00 am and hence high speeds at that time and same way heavy traffic at 18:00 hrs (peak time) could be the reason why speeds are rock bottom at 18:00 hrs.