Taxi_Data.R

hp

Sat Feb 24 19:11:51 2018

```
###
##Part a)

#Loading the green_tripdata_2015-09 dataset into taxi_df

library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.2.5
```

```
## data.table 1.10.4
```

```
##   The fastest way to learn (by data.table authors): https://www.datacamp.com/courses/data-analysis-the-data-table-way
```

```
##   Documentation: ?data.table, example(data.table) and browseVignettes("data.table")
```

```
##   Release notes, videos and slides: http://r-datatable.com
```

```
taxi_df <- fread("https://s3.amazonaws.com/nyc-tlc/trip+data/green_tripdata_2015-09.csv")
```

```
##
Read 9.4% of 1494926 rows
Read 16.7% of 1494926 rows
Read 24.1% of 1494926 rows
Read 29.4% of 1494926 rows
Read 36.1% of 1494926 rows
Read 42.1% of 1494926 rows
Read 49.5% of 1494926 rows
Read 56.9% of 1494926 rows
Read 61.5% of 1494926 rows
Read 68.9% of 1494926 rows
Read 76.3% of 1494926 rows
Read 80.3% of 1494926 rows
Read 87.6% of 1494926 rows
Read 95.0% of 1494926 rows
Read 1494926 rows and 21 (of 21) columns from 0.223 GB file in 00:00:17
```

```
##Part b)

# Reporting number of rows and columns in the dataframe taxi_df

number_of_row=nrow(taxi_df)

number_of_column=ncol(taxi_df)

### DATA CLEANING

min(taxi_df$Trip_distance)
```

```
## [1] 0
```

```
# We can observe that the minimum value of trip distance is 0 which can not be possible, henc
e we will remove all those
#values from our dataframe for which we have trip distance as 0.

min(taxi_df$Fare_amount)
```

```
## [1] -475
```

```
# The minimum value of fare amount is -475, it shows that there are certain values for which
 the fare amount is negative. We will remove
#all those values from our dataframe.

min(taxi_df$Tip_amount)
```

```
## [1] -50
```

```
# The minimum value of fare amount is -50, it shows that there are certain values for which t
he tip amount is negative. We will remove
#all those values from our dataframe.

min(taxi_df$Total_amount)
```

```
## [1] -475
```

```
# The minimum value of total amount is also -475, it shows that there are certain values for
 which the total amount is negative. We will remove
#all those values from our dataframe.

#Now we will filter our dataframe based on these values which are mentioned above

#nrow(taxi_df)
#min(taxi_df$Tip_amount)

taxi_df<-taxi_df[with(taxi_df, Trip_distance >0), ]
taxi_df<-taxi_df[with(taxi_df, Fare_amount >=2.5), ]
taxi_df<-taxi_df[with(taxi_df, Tip_amount >=0), ]

nrow(taxi_df)
```

```
## [1] 1468507
```

```
###
##Part a)

#Plotting a Histogram using ggplot

install.packages("ggplot2")
```

```
## Installing package into 'C:/Users/hp/Documents/R/win-library/3.2'
## (as 'lib' is unspecified)
```

```
## also installing the dependencies 'utf8', 'pillar', 'rlang', 'tibble'
```

```
## 
##   There is a binary version available but the source version is
##   later:
##        binary source needs_compilation
## tibble  1.3.0  1.4.2              TRUE
## 
##   Binaries will be installed
```

```
## Packages which are only available in source form, and may need
##   compilation of C/C++/Fortran: 'utf8' 'rlang'
```

```
##   These will not be installed
## package 'tibble' successfully unpacked and MD5 sums checked
## package 'ggplot2' successfully unpacked and MD5 sums checked
## 
## The downloaded binary packages are in
##  C:\Users\hp\AppData\Local\Temp\Rtmp4anKjS\downloaded_packages
```

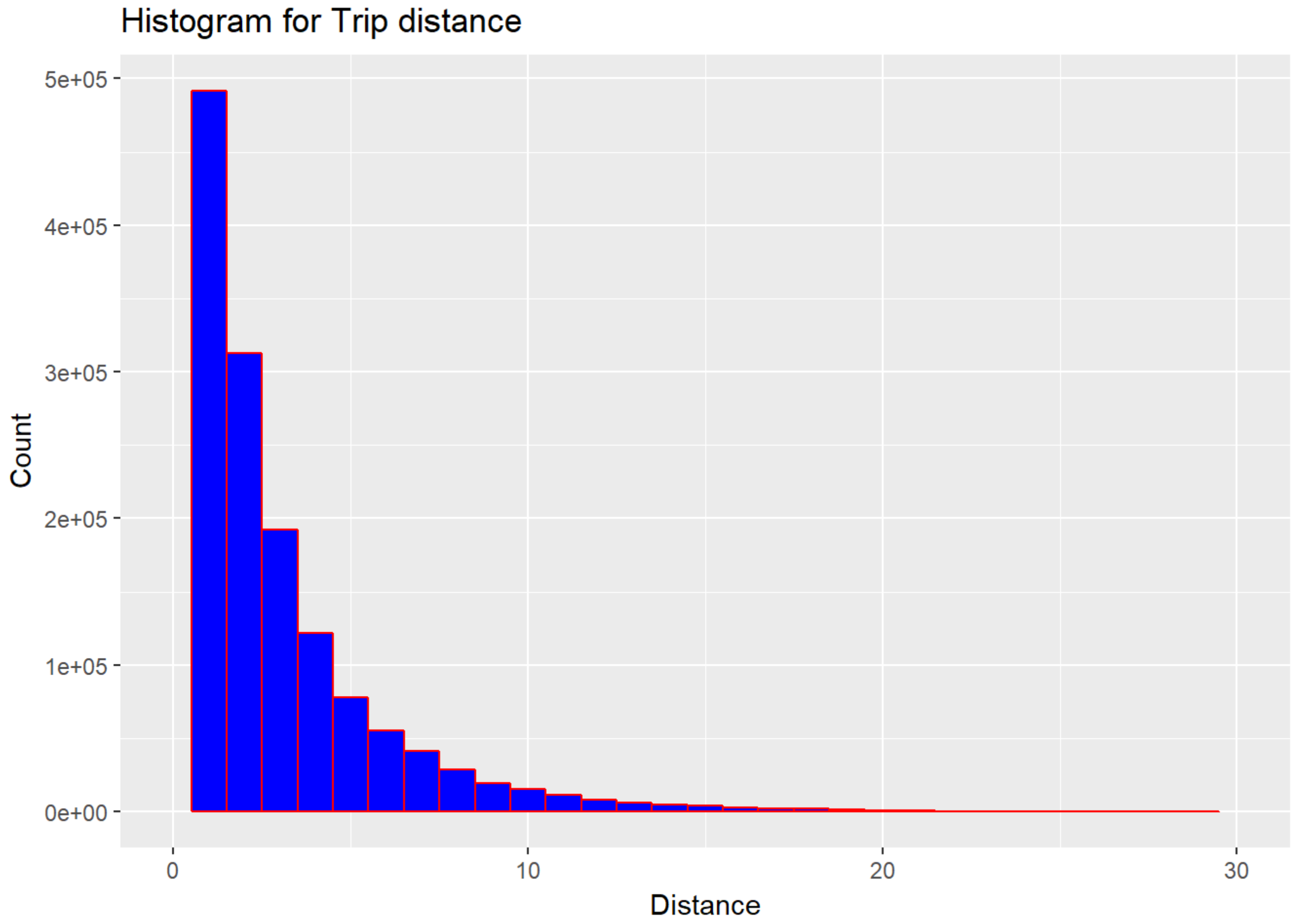
```
## installing the source package 'pillar'
## Warning in install.packages :
##   running command '"C:/PROGRA~1/R/R-32~1.1/bin/x64/R" CMD INSTALL -l "C:\Users\hp\Document
s\R\win-library\3.2" C:\Users\hp\AppData\Local\Temp\Rtmp4anKjS/downloaded_packages/pillar_1.
1.0.tar.gz' had status 1
## Warning in install.packages :
##   installation of package 'pillar' had non-zero exit status
##
## The downloaded source packages are in
##  'C:\Users\hp\AppData\Local\Temp\Rtmp4anKjS\downloaded_packages'
```

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

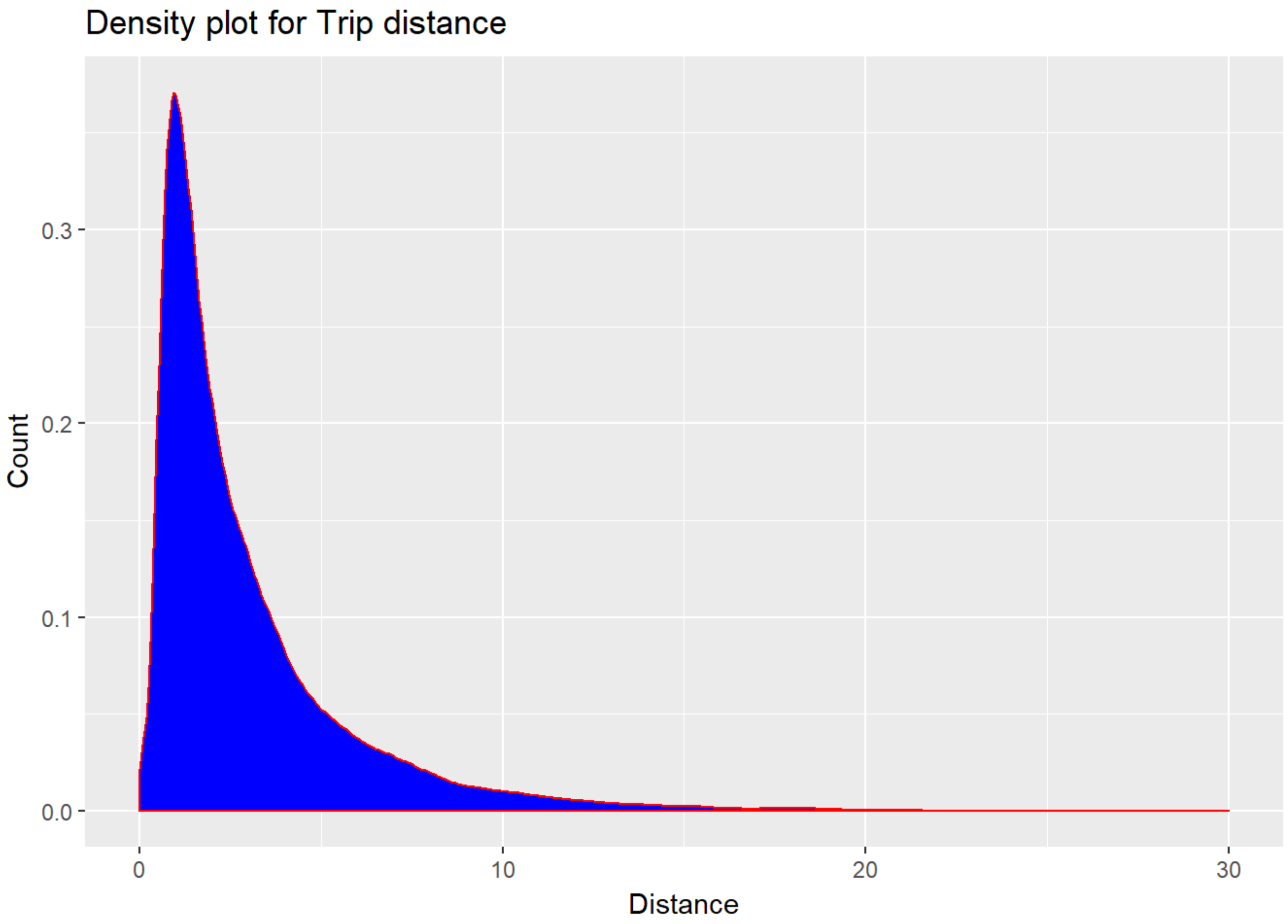
```
#dev.off()---use this function in case ggplot does not work after loading

ggplot(data=taxi_df, aes(taxi_df$Trip_distance))+
  geom_histogram( binwidth =1,color = "red", fill = "blue") + xlim(0, 30)+ labs(title="Histog
ram for Trip distance", x="Distance", y="Count")
```

```
## Warning: Removed 452 rows containing non-finite values (stat_bin).
```

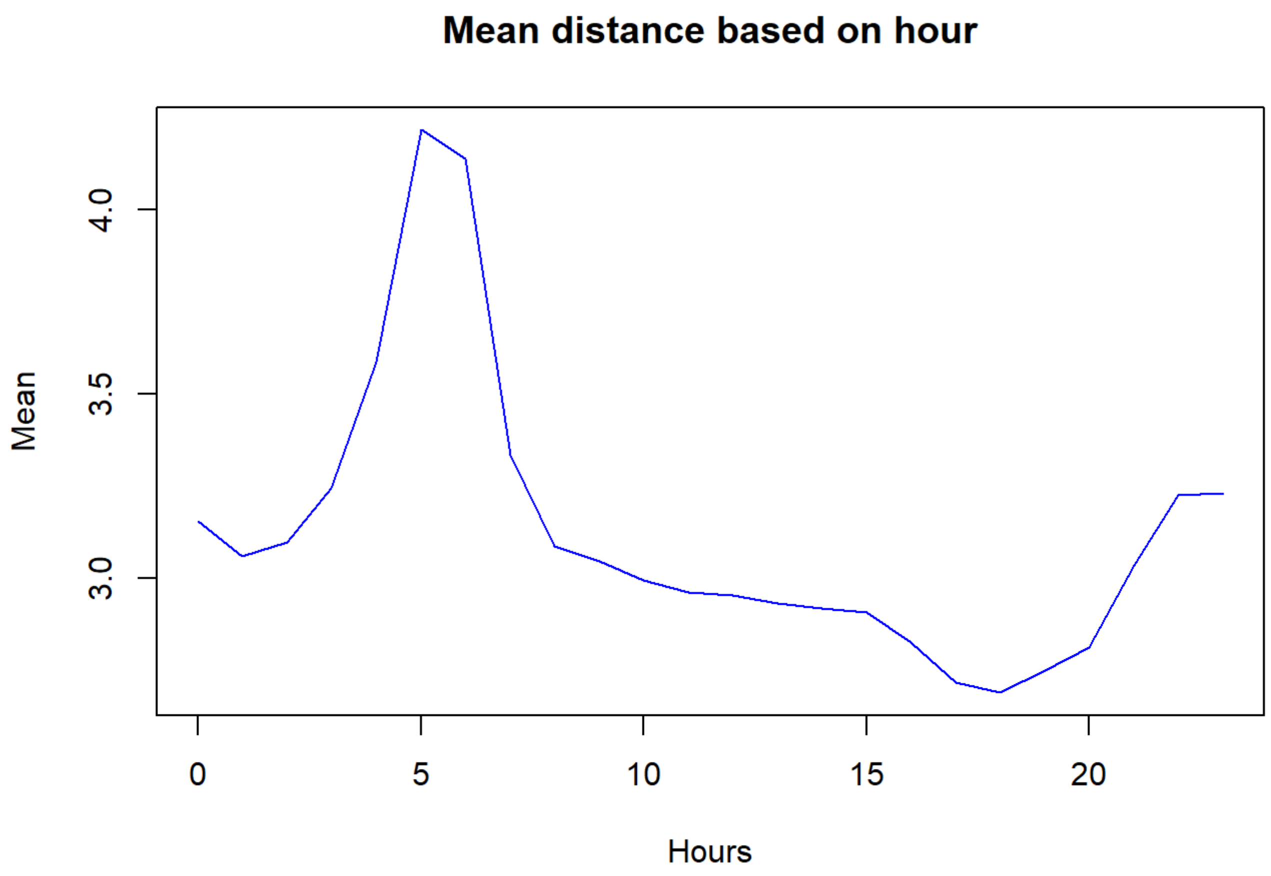


```
ggplot(data=taxi_df, aes(taxi_df$Trip_distance))+
  geom_density( color = "red", fill = "blue" ) + xlim(0, 30)+ labs(title="Density plot for Tr
ip distance", x="Distance", y="Count")
```

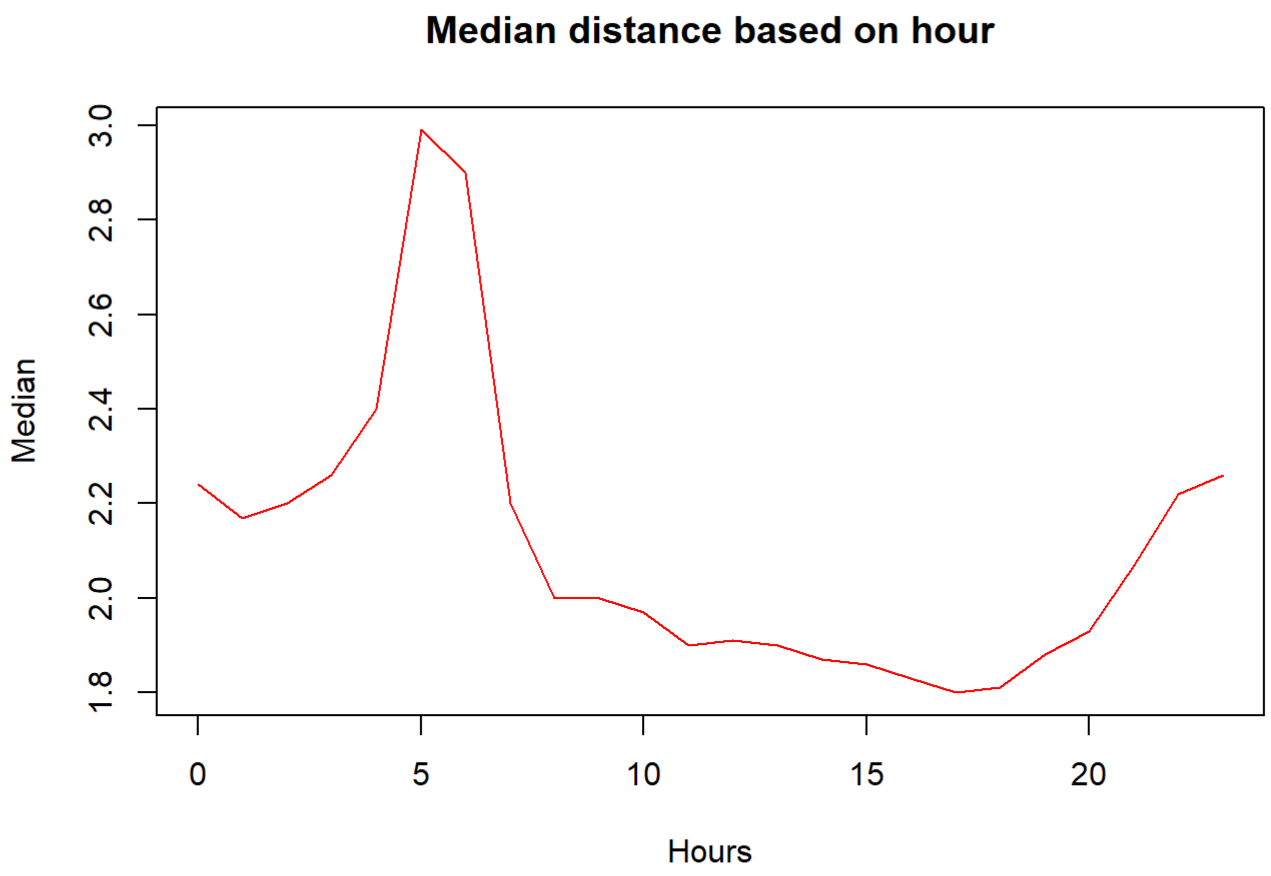
```
## Warning: Removed 452 rows containing non-finite values (stat_density).
```



```
##Part b)

#The Histogram obtained for "Trip distance" is right skewed. It can be observed that
#the mean is larger than the median.

#Hypothesis:The trips are not normally distributed hence they cannot be random. It seems that
 maximum
#number of trips happen at certain time i.e. during peak hours.



###
##Part a)

# Fetching hours from pickup time and adding these hours into a new column "hours" of taxi_d
f.
#NOTE: The 0th hour is corresponding to 12 AM

pickup_time <- as.POSIXlt(taxi_df$lpep_pickup_datetime)
taxi_df$hours<-pickup_time$hour

#Making 2 new columns which have values of trip day and trip week according to month

taxi_df$day_of_trip=pickup_time$mday

#I am assuming that a week is from Sunday to Saturday

taxi_df$week_of_month[taxi_df$day_of_trip >=1 & taxi_df$day_of_trip <= 5]<-1
taxi_df$week_of_month[taxi_df$day_of_trip >=6 & taxi_df$day_of_trip <= 12]<-2
taxi_df$week_of_month[taxi_df$day_of_trip >=13 & taxi_df$day_of_trip <= 19]<-3
taxi_df$week_of_month[taxi_df$day_of_trip >=20 & taxi_df$day_of_trip <= 26]<-4
taxi_df$week_of_month[taxi_df$day_of_trip >=27 & taxi_df$day_of_trip <= 30]<-5

#calculating and plotting mean and median by hours

mean_trip_distance= aggregate(taxi_df$Trip_distance, list(hours=taxi_df$hours), mean)
plot(mean_trip_distance,col = 'blue', pch = 19, cex = 0.5,main = "Mean distance based on hou
r",
     xlab = "Hours",
     ylab = "Mean",type='l')
```



```
median_trip_distance= aggregate(taxi_df$Trip_distance, list(hours=taxi_df$hours), median)
plot(median_trip_distance,col = 'red', pch = 19, cex = 0.5,main = "Median distance based on h
our",
     xlab = "Hours",
     ylab = "Median",type='l')
```



```
##Part b)

#creating a new column which has the value 'Yes' if the ratecodeid is 2(for JFK) or 3(for New
ark) and 'No' for rest of the values.

taxi_df$airport_pick_or_drop[taxi_df$RateCodeID==2 | taxi_df$RateCodeID==3]<-'Yes'
taxi_df$airport_pick_or_drop[taxi_df$RateCodeID==1 | taxi_df$RateCodeID==4 | taxi_df$RateCode
ID==5 | taxi_df$RateCodeID==6]<-'No'

library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.2.5
```

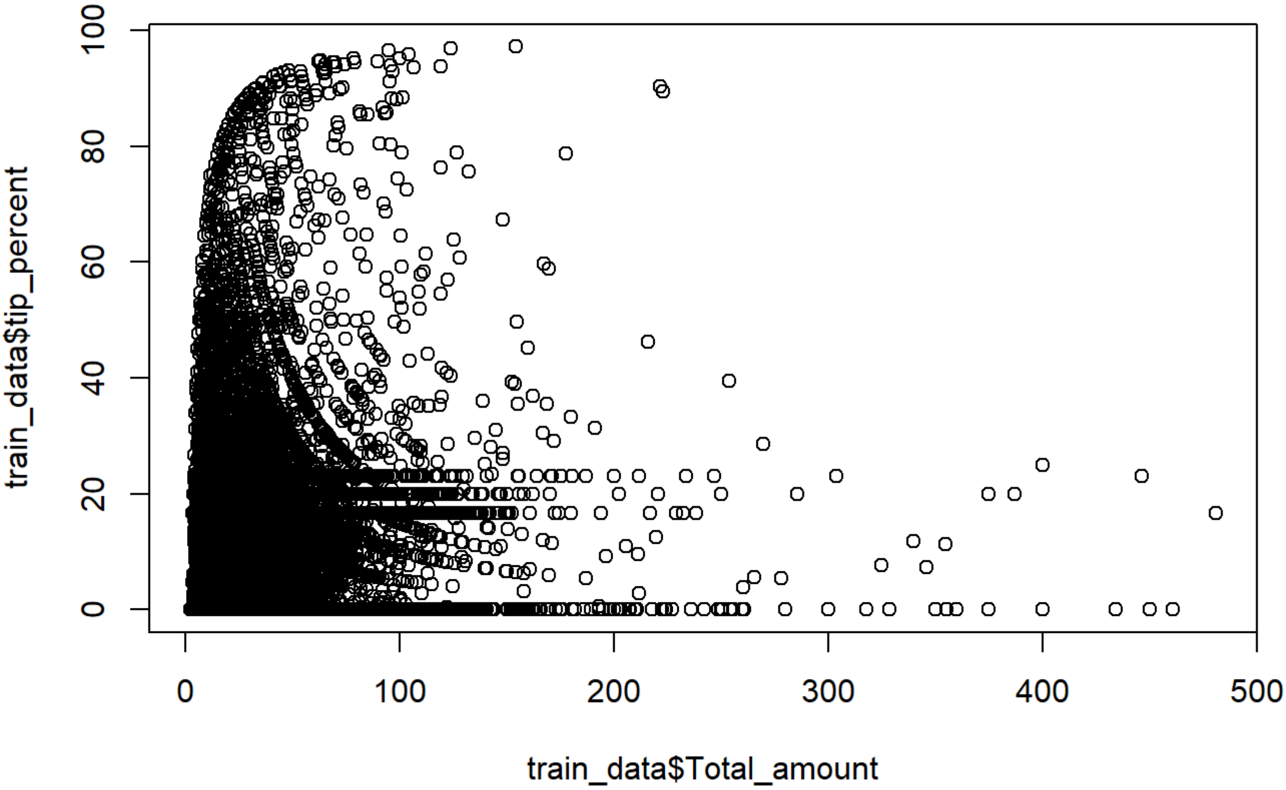
```
count(taxi_df$airport_pick_or_drop)
```

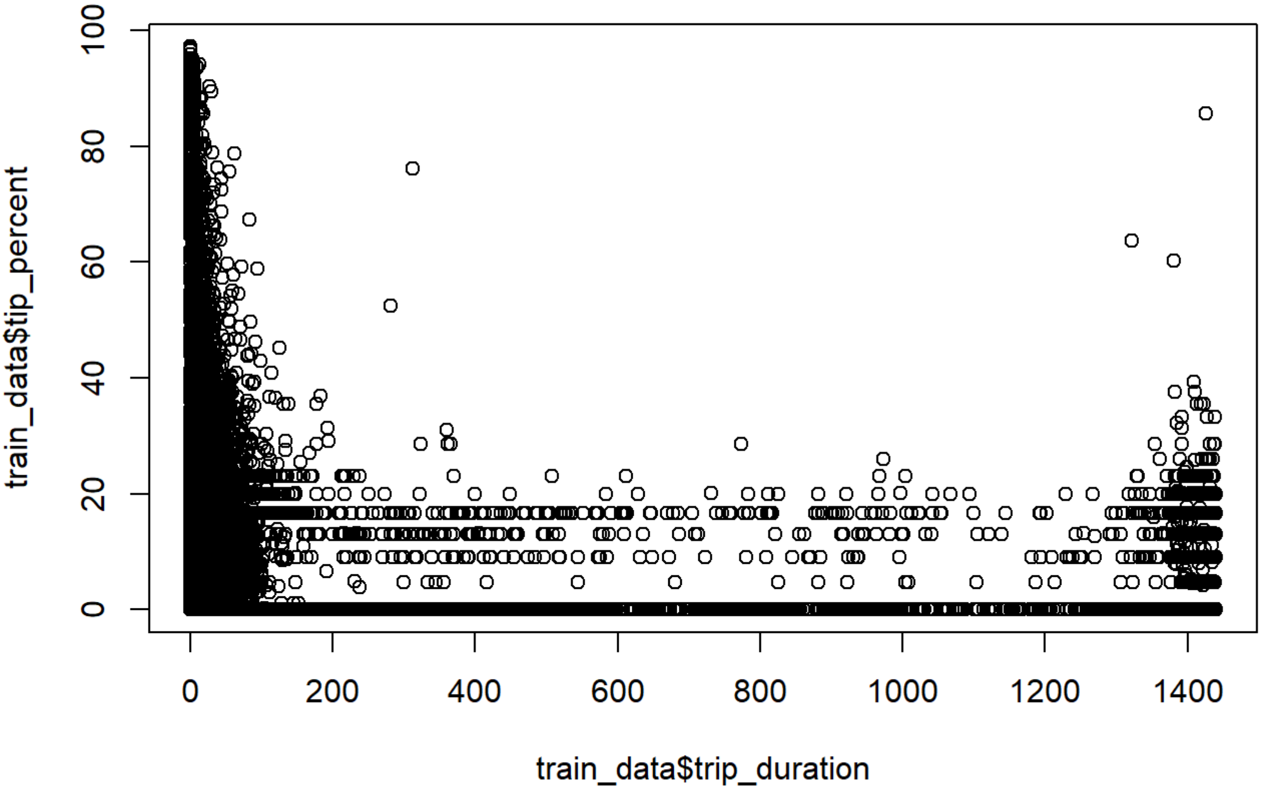
```
##     x    freq
## 1  No 1464227
## 2 Yes    4280
```

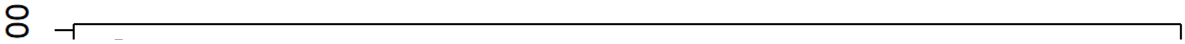
```
# using count from library 'plyr' we can observe that a total of 4280 trips originated or end
ed at new york airports


average_fair= aggregate(taxi_df$Fare_amount, list(taxi_df$airport_pick_or_drop),mean,drop=FAL
SE)

#the average_fair from or to airports is 53.24755

average_tip_amount= aggregate(taxi_df$Tip_amount, list(taxi_df$airport_pick_or_drop),mean,dro
p=FALSE)

#average tip for trips from and to airports is 5.395044


###

##Part a)


#Calculating tip percent and storing all those values in a new column 'taxi_percent' in taxi_
df

taxi_df$tip_percent=round(100*taxi_df$Tip_amount/taxi_df$Total_amount,2)

##Part b)

# Building a model which predict the tip percentge based on the values of Total amount, Trip
 duration and Trip distance

# Calculating the total duration of the trip in minutes

pickup_time = as.POSIXct(taxi_df$lpep_pickup_datetime,format='%Y-%m-%d %H:%M:%S')
drop_time= as.POSIXct(taxi_df$Lpep_dropoff_datetime,format='%Y-%m-%d %H:%M:%S')
taxi_df$trip_duration= round((drop_time-pickup_time)/60,2)
taxi_df$trip_duration= as.numeric(taxi_df$trip_duration)

# Now as few of the values in trip duration were 0 We will remove these values from our datas
et.

taxi_df<-taxi_df[with(taxi_df, trip_duration >0), ]

#Creating a traning and testing data set for model prediction

#Training data set having 1000000 number of rows

train_data=taxi_df[1:1000000]

# Testing data set having rest of the rows from taxi_df

test_data=taxi_df[1000001:1468366]

#creating a variable test count which contains value of tip percent from testing dataset
test_counts= test_data$tip_percent

#Using lm() for prediction
prediction_model = lm(train_data$tip_percent ~ train_data$Total_amount + train_data$trip_dura
tion + train_data$Trip_distance , data = train_data)

summary(prediction_model)
```

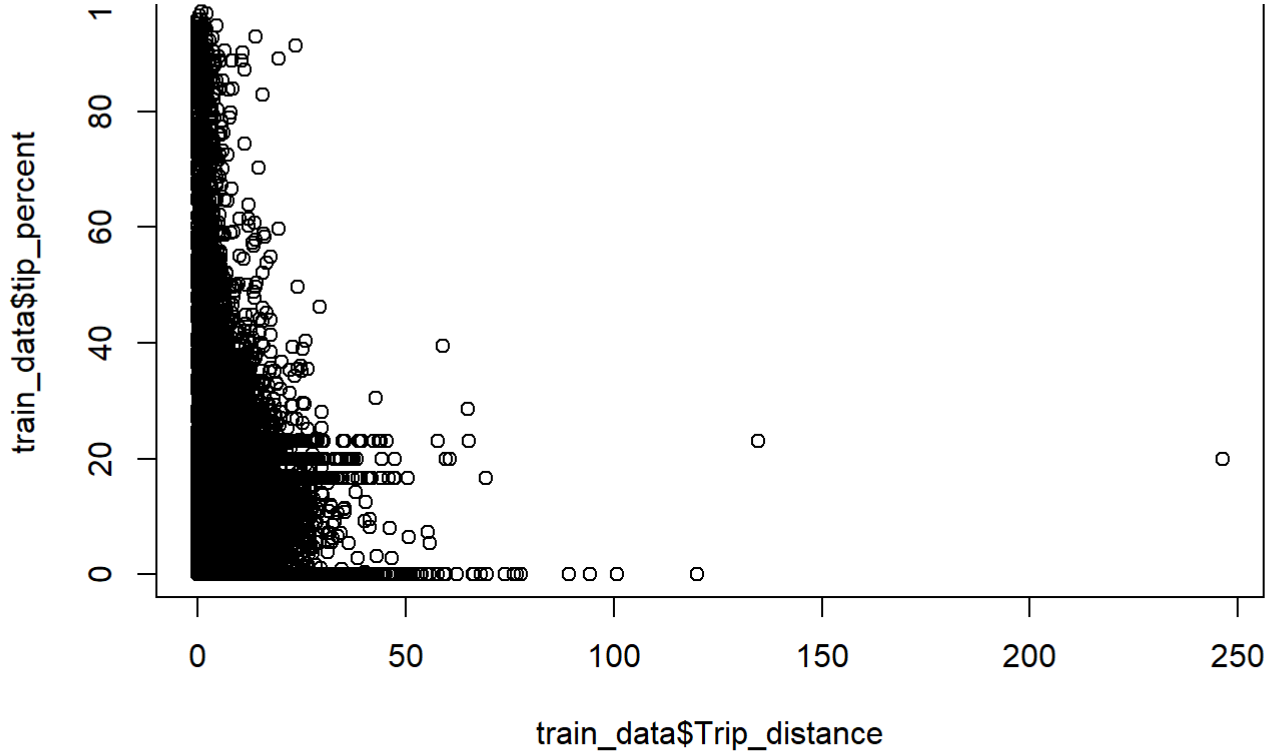
```
##
## Call:
## lm(formula = train_data$tip_percent ~ train_data$Total_amount +
##     train_data$trip_duration + train_data$Trip_distance, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -278.692   -5.484   -4.626    8.070  240.007
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.204e+00  1.449e-02   152.1   <2e-16 ***
## train_data$Total_amount   6.885e-01  1.757e-03   391.8   <2e-16 ***
## train_data$trip_duration -2.712e-03  8.662e-05   -31.3   <2e-16 ***
## train_data$Trip_distance -1.982e+00  6.388e-03  -310.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.016 on 999996 degrees of freedom
## Multiple R-squared:  0.1424, Adjusted R-squared:  0.1424
## F-statistic: 5.536e+04 on 3 and 999996 DF,  p-value: < 2.2e-16
```

```
# Looking at the values obatined from Summary function we can obsereve how the value of tip p
ercent changes when we use total amount,
#trip duration and trip distance as predictor variables.

plot(train_data$tip_percent ~ train_data$Total_amount + train_data$trip_duration + train_data
$Trip_distance )
```


train_data$tip_percent
0
20
40
60
80
100
0
100
200
300
400
500
train_data$Total_amount


train_data$tip_percent
0
20
40
60
80
100
0
200
400
600
800
1000
1200
1400
train_data$trip_duration


00

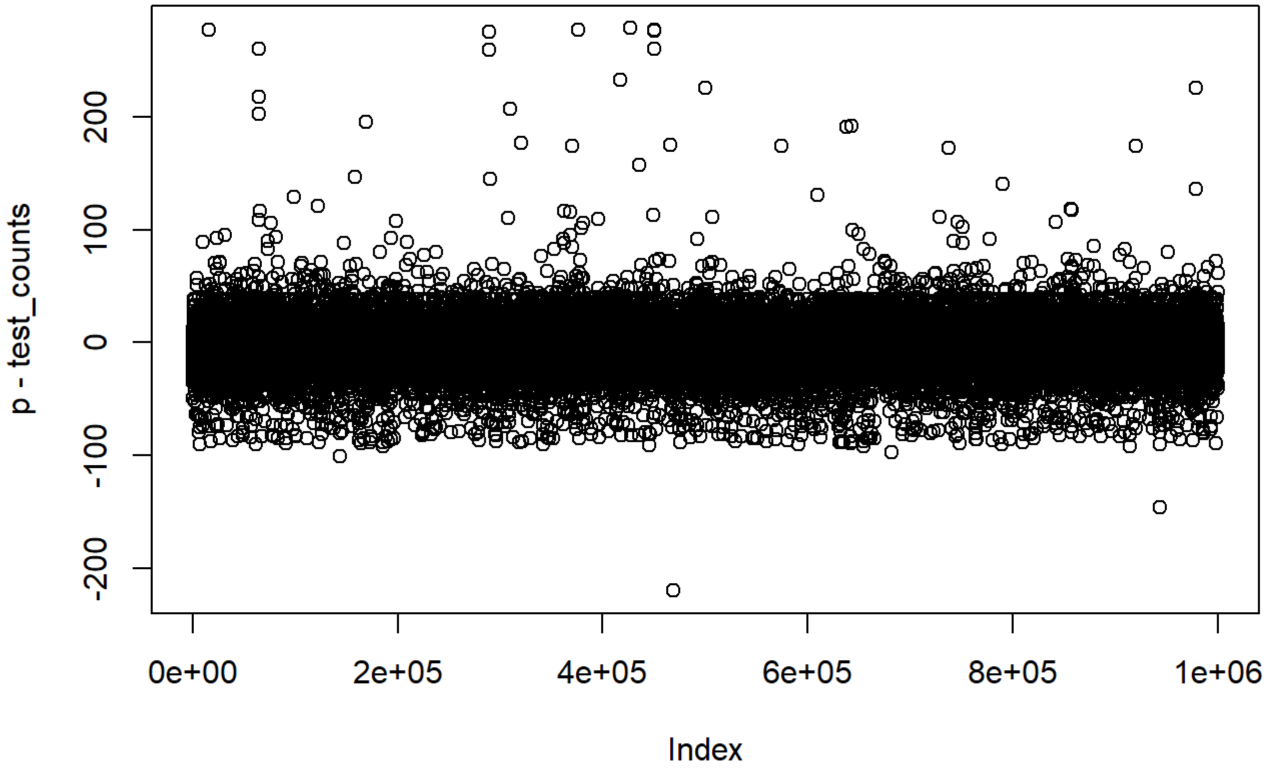


```
#Predicting the model
p = predict(prediction_model, test_data)
```

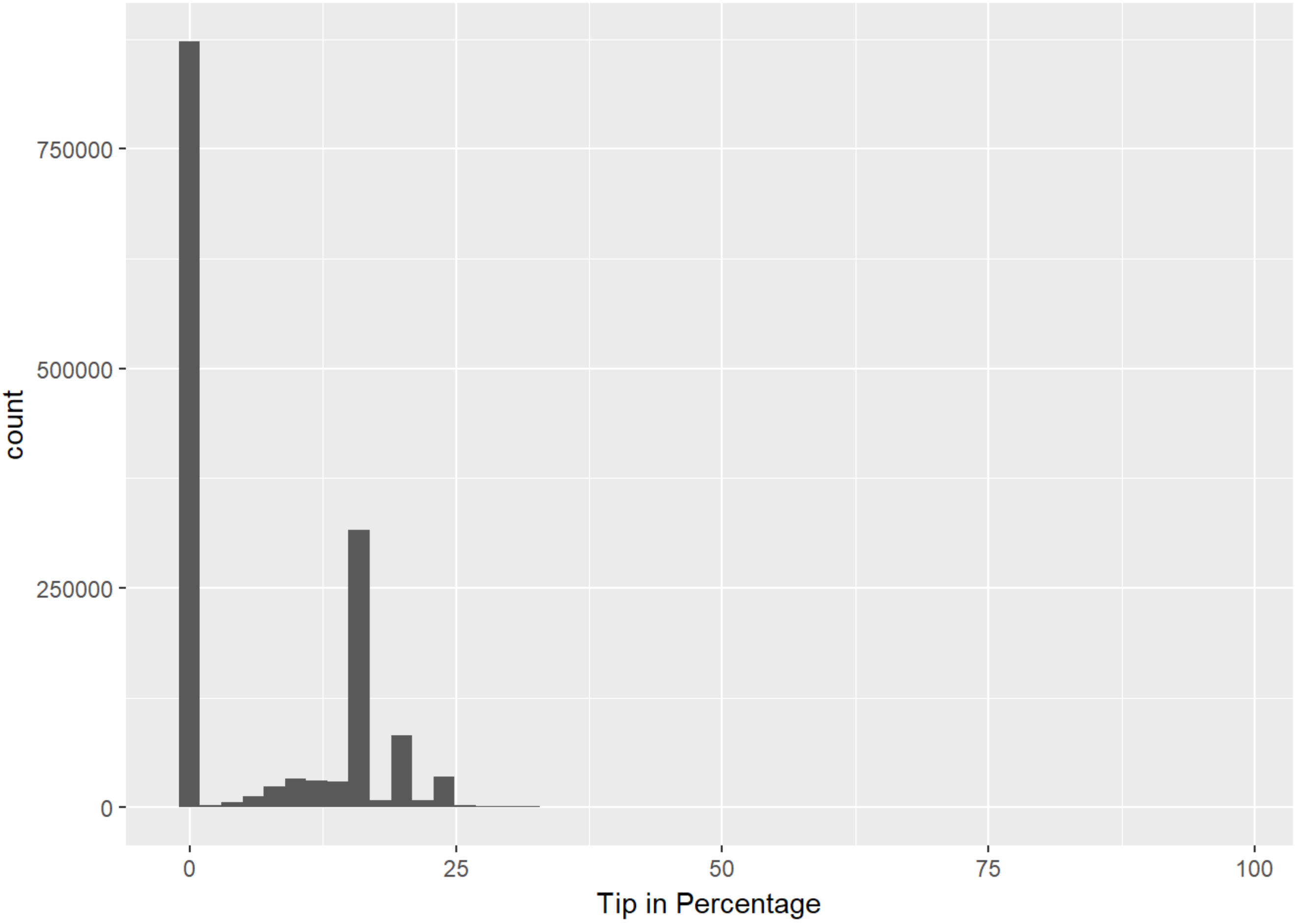
```
## Warning: 'newdata' had 468366 rows but variables found have 1000000 rows
```

```
#Plotting the predicted model
plot(p - test_counts)
```

```
## Warning in p - test_counts: longer object length is not a multiple of
## shorter object length
```



```
#Also plotting a histogram to showcase the tip percentage

ggplot(data=taxi_df, aes(taxi_df$tip_percent)) +
  geom_histogram(bins=50)+labs( x = "Tip in Percentage", y = "count")
```


750000
500000
250000
0
count
0
25
50
75
100
Tip in Percentage

```
###

##Part a)

#calculating the average speed in miles per hour and adding that value in our dataframe

taxi_df$average_speed = round(taxi_df$Trip_distance/(taxi_df$trip_duration/60),2)

#Now as the maximum speed should be below or equal to 25mph in NYC "http://nymag.com/daily/in
telligencer/2014/11/things-to-know-about-nycs-new-speed-limit.html".
#We will clean the data based on the average speed.

taxi_df<-taxi_df[with(taxi_df, average_speed <= 25), ]

##Part b)

#HYPOTHESIS TESING

#Finding the averrage speed according to the week of month

average_speed_weekwise= aggregate(taxi_df$average_speed, list(week=taxi_df$week_of_month),mea
n)

#Hypothesis Testing using Paired student t-test. We are using this method as we are concerned
 with average speeds for each week.

#Let's define null and alternate hypothesis

#Null Hypothesis(H0): Average speed for each week is same and true difference in mean is zero
 for both samples.

#Alternate Hypothesis(H1): Average speeds are different for different week and true differenc
e is not zero for the samples.

#making samples of 1st 100 elements from each week

week1_sample=taxi_df$average_speed[taxi_df$week_of_month==1][1:100]
week2_sample=taxi_df$average_speed[taxi_df$week_of_month==2][1:100]
week3_sample=taxi_df$average_speed[taxi_df$week_of_month==3][1:100]
week4_sample=taxi_df$average_speed[taxi_df$week_of_month==4][1:100]
week5_sample=taxi_df$average_speed[taxi_df$week_of_month==5][1:100]

#performing Hypothesis testing for week1 and week2 sample
t.test(week1_sample,week2_sample, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  week1_sample and week2_sample
## t = 2.52, df = 99, p-value = 0.01334
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3217753 2.7052247
## sample estimates:
## mean of the differences
##                  1.5135
```

```r
#performing Hypothesis testing for week1 and week3 sample
t.test(week1_sample,week3_sample, paired=TRUE)
```

```
## 
##  Paired t-test
## 
## data:  week1_sample and week3_sample
## t = 3.327, df = 99, p-value = 0.001233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7579522 2.9980478
## sample estimates:
## mean of the differences 
##                   1.878
```

```r
#performing Hypothesis testing for week1 and week4 sample
t.test(week1_sample,week4_sample, paired=TRUE)
```

```
## 
##  Paired t-test
## 
## data:  week1_sample and week4_sample
## t = 2.3104, df = 99, p-value = 0.02294
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1969283 2.5928717
## sample estimates:
## mean of the differences 
##                  1.3949
```

```r
#performing Hypothesis testing for week1 and week5 sample
t.test(week1_sample,week5_sample, paired=TRUE)
```

```
## 
##  Paired t-test
## 
## data:  week1_sample and week5_sample
## t = 2.7494, df = 99, p-value = 0.007099
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.45443 2.81137
## sample estimates:
## mean of the differences 
##                  1.6329
```

```r
#performing Hypothesis testing for week2 and week3 sample
t.test(week2_sample,week3_sample, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  week2_sample and week3_sample
## t = 0.69303, df = 99, p-value = 0.4899
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6791013  1.4081013
## sample estimates:
## mean of the differences
##                  0.3645
```

```
#performing Hypothesis testing for week2 and week4 sample
t.test(week2_sample,week4_sample, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  week2_sample and week4_sample
## t = -0.18984, df = 99, p-value = 0.8498
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.358198  1.120998
## sample estimates:
## mean of the differences
##                 -0.1186
```

```
#performing Hypothesis testing for week2 and week5 sample
t.test(week2_sample,week5_sample, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  week2_sample and week5_sample
## t = 0.21561, df = 99, p-value = 0.8297
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9794113  1.2182113
## sample estimates:
## mean of the differences
##                  0.1194
```

```
#performing Hypothesis testing for week3 and week4 sample
t.test(week3_sample,week4_sample, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  week3_sample and week4_sample
## t = -0.88476, df = 99, p-value = 0.3784
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5665306  0.6003306
## sample estimates:
## mean of the differences
##                 -0.4831
```

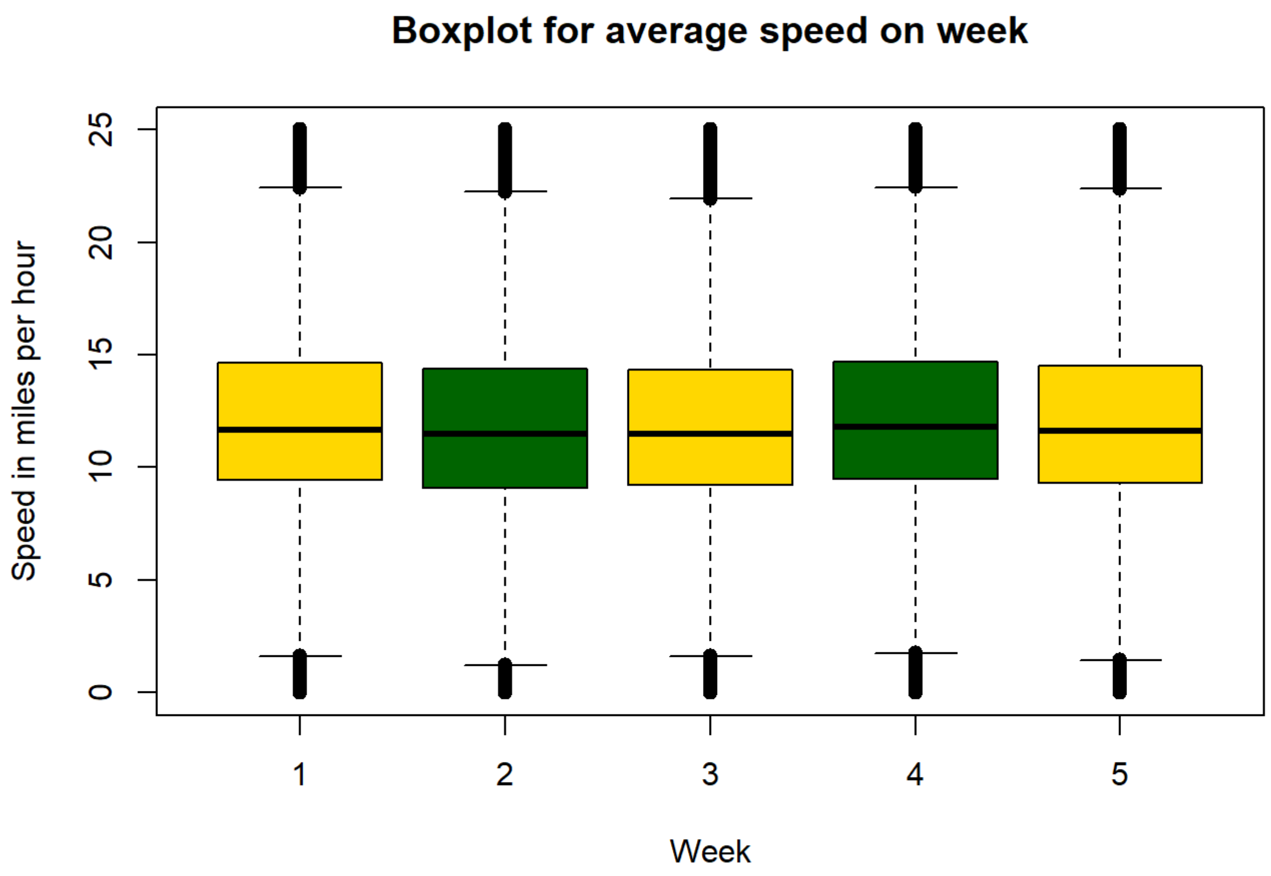
```
#performing Hypothesis testing for week3 and week5 sample
t.test(week3_sample,week5_sample, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  week3_sample and week5_sample
## t = -0.49645, df = 99, p-value = 0.6207
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2247236  0.7345236
## sample estimates:
## mean of the differences
##                 -0.2451
```

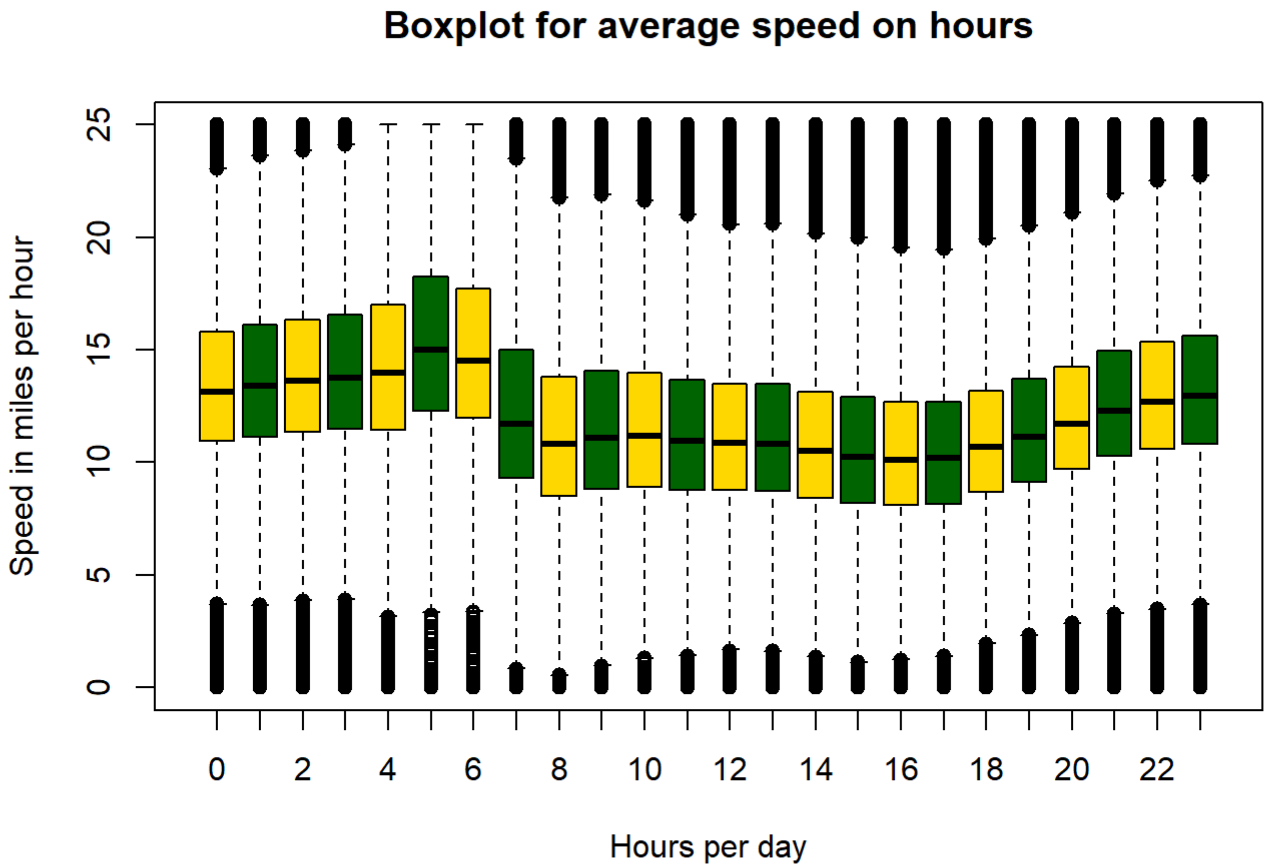
```
#performing Hypothesis testing for week4 and week5 sample
t.test(week4_sample,week5_sample, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  week4_sample and week5_sample
## t = 0.47409, df = 99, p-value = 0.6365
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7581039  1.2341039
## sample estimates:
## mean of the differences
##                   0.238
```

```
#After performing all the above mentioned tests it is evident that p-value for all the test i
s greater than 0.05 and hence
# we can REJECT NULL HYPOTHESIS in favor of Alternative Hypothesis.

#Conclusion: The average speeds of NYC green taxi is different for each week.

# Plotting a boxplot for the average speed based on each week.

boxplot(taxi_df$average_speed~taxi_df$week_of_month,data=taxi_df, col=(c("gold","darkgreen"))
,main="Boxplot for average speed on week",
        xlab="Week", ylab="Speed in miles per hour")
```



```
#part c)

#Plotting a boxplot for the average speed based on each hour of day

boxplot(taxi_df$average_speed~taxi_df$hours,data=taxi_df, col=(c("gold","darkgreen")),main="Boxplot for average speed on hours",
        xlab="Hours per day", ylab="Speed in miles per hour")
```



```
#HYPOTHESIS TESTING

#Hypothesis Testing using One-way Analysis of Variance (ANOVA). We are using this method as we are concerned with average speeds for each hour.

#Let's define null and alternate hypothesis

#Null Hypothesis(H0): Average speed for each hour is same and true difference in mean is zero.

#Alternate Hypothesis(H1): Average speeds are different for different hour and true difference is not zero for the samples.

model_on_hours <- lm(taxi_df$average_speed~taxi_df$hours,data=taxi_df)

summary(model_on_hours)
```

```
##
## Call:
## lm(formula = taxi_df$average_speed ~ taxi_df$hours, data = taxi_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9193  -2.8375  -0.5413   2.3085  13.3550
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.9193293  0.0081840  1578.6   <2e-16 ***
## taxi_df$hours -0.0554035  0.0005375  -103.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.283 on 1398838 degrees of freedom
## Multiple R-squared:  0.007538,   Adjusted R-squared:  0.007537
## F-statistic: 1.062e+04 on 1 and 1398838 DF,  p-value: < 2.2e-16
```

```
anova(model_on_hours)
```

```
## Analysis of Variance Table
##
## Response: taxi_df$average_speed
##                    Df   Sum Sq Mean Sq F value    Pr(>F)
## taxi_df$hours       1   194901  194901   10624 < 2.2e-16 ***
## Residuals     1398838 25661200      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Using the ANOVA test it is clear that the speeds are different for each hour. Hence we will
 reject Null Hypothesis in favor of
#Alertnate Hypothesis.

#Using the boxplot we can also infer that the speed is maximum at 5 AM and gradually decrease
s from there. The speed starts increasing slowly
#after 6 PM.


######## THE END
```