run\_analysis\_script\_output

Ritesh Ingale

December 25, 2015

Load required libraries

suppressMessages(library(dplyr))  
library(chron)

Acquire the data

tripdata\_01 <- read.csv("train\_tripdata\_01.csv", header = FALSE, sep = ",", stringsAsFactors = FALSE, quote = "")  
tripdata\_02 <- read.csv("train\_tripdata\_02.csv", header = FALSE, sep = ",", stringsAsFactors = FALSE, quote = "")  
train\_tripdata <- dplyr::bind\_rows(tripdata\_01,tripdata\_02)  
rm(tripdata\_01, tripdata\_02)

Give the variable names

names(train\_tripdata) <- c("VendorID", "pickup\_time", "dropoff\_time", "Passenger\_count", "Trip\_distance", "Pickup\_longitude", "Pickup\_latitude", "RateCodeID", "Store\_and\_fwd\_flag", "Dropoff\_longitude", "Dropoff\_ latitude", "Payment\_type", "Fare\_amount", "Extra", "MTA\_tax", "Improve\_surcharge", "Tip\_amount", "Tolls\_amount", "Total\_amount")  
head(train\_tripdata, 10)

## Source: local data frame [10 x 19]  
##   
## VendorID pickup\_time dropoff\_time Passenger\_count  
## (int) (chr) (chr) (int)  
## 1 1 2015-01-07 20:40:05 2015-01-07 20:50:35 2  
## 2 1 2015-01-07 20:40:05 2015-01-07 20:45:40 1  
## 3 1 2015-01-07 20:40:05 2015-01-07 20:43:41 1  
## 4 1 2015-01-07 20:40:05 2015-01-07 20:53:31 1  
## 5 1 2015-01-07 20:40:05 2015-01-07 21:02:08 1  
## 6 1 2015-01-07 20:40:06 2015-01-07 20:53:38 1  
## 7 2 2015-01-07 21:01:42 2015-01-07 21:17:28 2  
## 8 2 2015-01-07 21:01:42 2015-01-07 21:12:14 2  
## 9 2 2015-01-07 21:01:43 2015-01-07 21:05:19 2  
## 10 2 2015-01-07 21:01:43 2015-01-07 21:15:45 1  
## Variables not shown: Trip\_distance (dbl), Pickup\_longitude (dbl),  
## Pickup\_latitude (dbl), RateCodeID (int), Store\_and\_fwd\_flag (chr),  
## Dropoff\_longitude (dbl), Dropoff\_ latitude (dbl), Payment\_type (int),  
## Fare\_amount (dbl), Extra (dbl), MTA\_tax (dbl), Improve\_surcharge (dbl),  
## Tip\_amount (dbl), Tolls\_amount (dbl), Total\_amount (dbl)

Drop the observations which we dont required in our analysis

train\_tripdata <- train\_tripdata %>% filter(RateCodeID != 0, Payment\_type != 0, Passenger\_count >= 0, Trip\_distance >= 0)

Change the values of some variable from negative to positive

train\_tripdata$Fare\_amount <- abs(train\_tripdata$Fare\_amount)  
train\_tripdata$Extra <- abs(train\_tripdata$Extra)  
train\_tripdata$MTA\_tax <- abs(train\_tripdata$MTA\_tax)  
train\_tripdata$Improve\_surcharge <- abs(train\_tripdata$Improve\_surcharge)  
train\_tripdata$Tip\_amount <- abs(train\_tripdata$Tip\_amount)  
train\_tripdata$Tolls\_amount <- abs(train\_tripdata$Tolls\_amount)  
train\_tripdata$Total\_amount <- abs(train\_tripdata$Total\_amount)

convert time values from char to datetime object

train\_tripdata$pickup\_time <- as.POSIXct(train\_tripdata$pickup\_time, format = "%Y-%m-%d %H:%M:%S", tz = "EST")  
train\_tripdata$dropoff\_time <- as.POSIXct(train\_tripdata$dropoff\_time, format = "%Y-%m-%d %H:%M:%S", tz = "EST")

Calculate the total time in minute of each ride and store it as separate variable in data frame

train\_tripdata <- train\_tripdata %>% mutate(Total\_Time\_min = ((dropoff\_time - pickup\_time)/60))  
train\_tripdata$Total\_Time\_min <- as.numeric(round(train\_tripdata$Total\_Time\_min, digits = 2))

Calculate the average speed of car in each ride Convert trip distance from miles to kilometers

train\_tripdata$Trip\_distance <- round(as.numeric(train\_tripdata$Trip\_distance / 0.6214), digits = 2)  
train\_tripdata <- train\_tripdata %>% mutate(avg\_speed = Trip\_distance / Total\_Time\_min)  
train\_tripdata <- train\_tripdata %>% select(1:3, Total\_Time\_min, avg\_speed, 4:19)

Create train dataset and test dataset

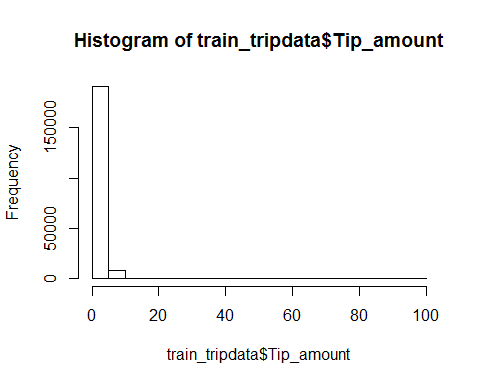
Lets do EDA - Exploratory Data Analysis!

summary(train\_tripdata$Tip\_amount)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.2523 0.0000 95.3300

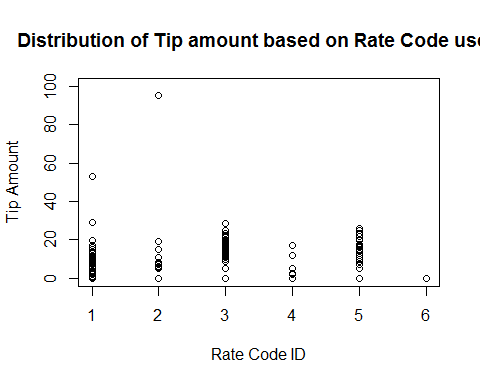
Lets explore by ploting the tip amount values using graphs Histograph will give the frequency of Tip amount, in out case is 0.00 value is more frequent

hist(train\_tripdata$Tip\_amount)



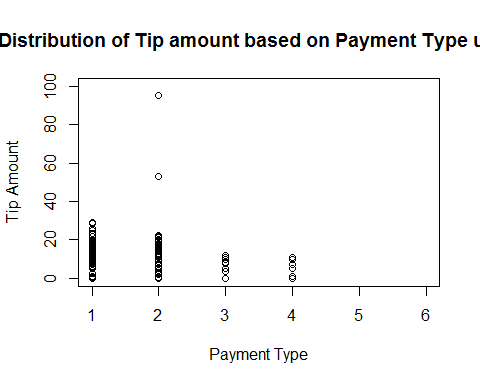
This plot will explore the pattern of customers who likely to give tip amount when they traveled to specific destination

plot(train\_tripdata$RateCodeID, train\_tripdata$Tip\_amount, main = "Distribution of Tip amount based on Rate Code used", xlab = "Rate Code ID", ylab = "Tip Amount", xlim = c(1,6), ylim = c(0.00, 100.00))



This plot will explore the pattern of customers who likely to give tip amount when they used payment method

plot(train\_tripdata$Payment\_type, train\_tripdata$Tip\_amount, main = "Distribution of Tip amount based on Payment Type used", xlab = "Payment Type", ylab = "Tip Amount", xlim = c(1,6), ylim = c(0.00, 100.00))

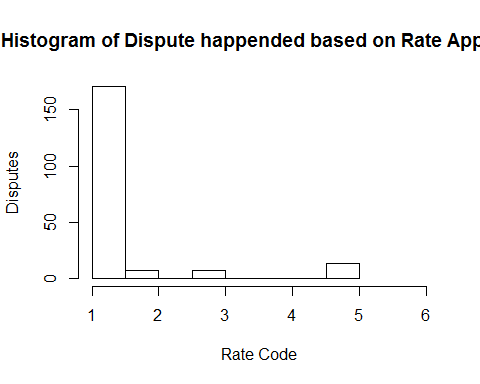


How many time Dispute happended in past two months

dispute\_happend <- train\_tripdata %>% filter(Payment\_type == 4)  
pickup\_time\_list <- as.POSIXlt(dispute\_happend$pickup\_time, format = "%Y-%m-%d %H:%M:%S", tz = "EST")  
dropoff\_time\_list <- as.POSIXlt(dispute\_happend$dropoff\_time, format = "%Y-%m-%d %H:%M:%S", tz = "EST")

For which type of Rate the dispute likely to be happended

hist(dispute\_happend$RateCodeID, main = "Histogram of Dispute happended based on Rate Applied", xlab = "Rate Code", ylab = "Disputes", xlim = c(1,6))



For how much count of passenger the dispute likely to be happended

hist(dispute\_happend$Passenger\_count, main = "Histogram of Dispute happended based on Passenger Count", xlab = "Passengers", ylab = "Disputes")

