Exploratory Data Analysis

Manoj Kumar Nagabandi

16 07 2022

```
knitr::opts_chunk$set(warning = FALSE)
```

```
library(ggplot2)
```

1. Download data

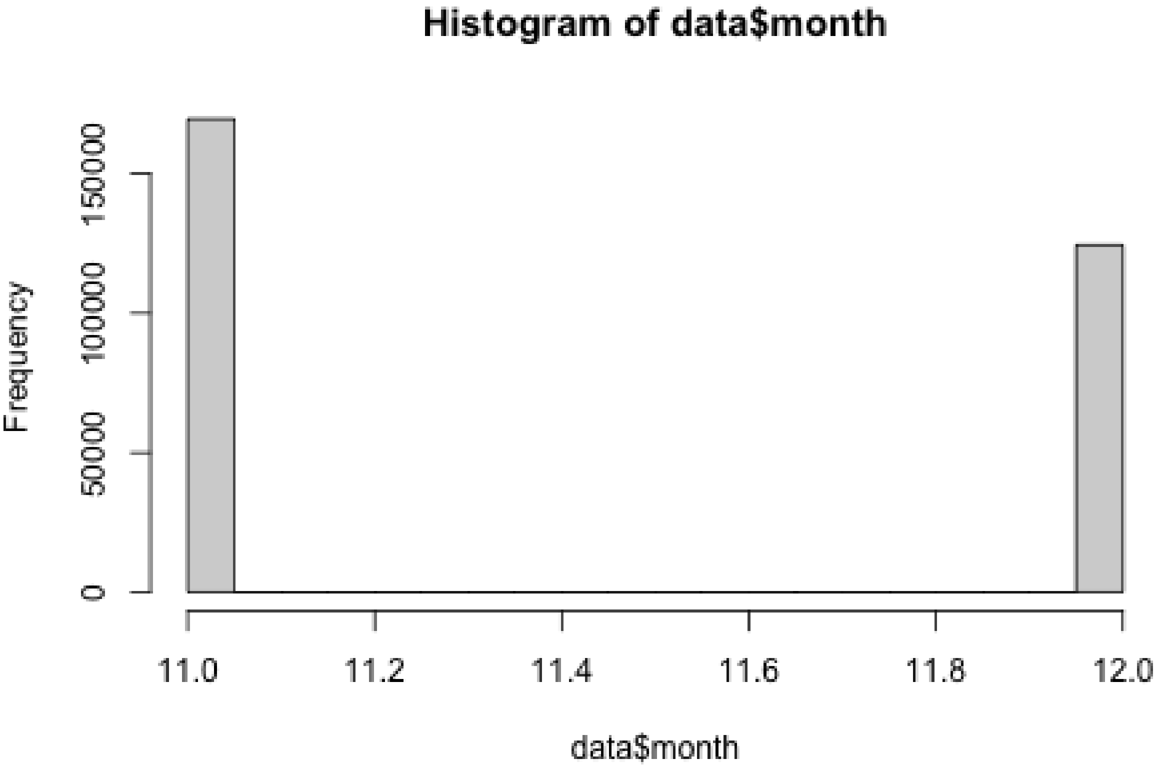
```
file_path = 'rideshare_kaggle_modified.csv'
data = read.csv(file_path)
```

2. Explore data

At this step, after cleaning data, we defined several questions to answer, with help of statistical tests.

2.1. In Which months did most of the rides occured?

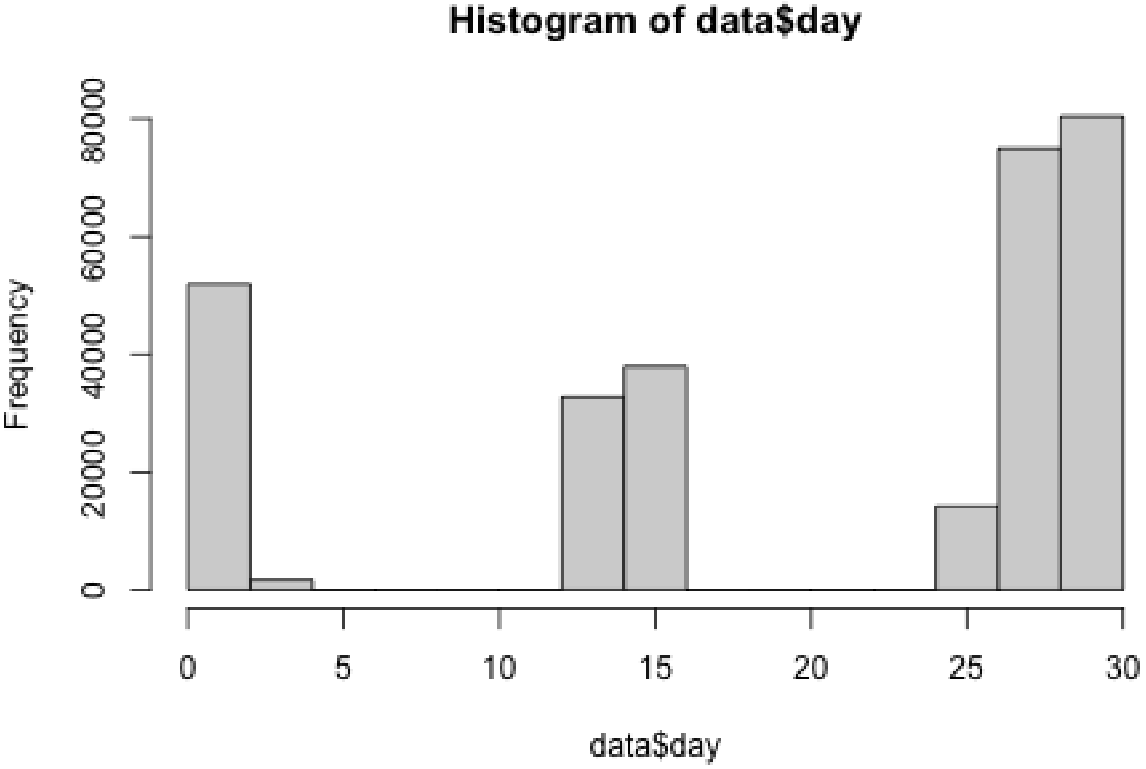
```
hist(data$month)
```



It appears that we only have november and december in our month data. It means the data is only recorded or taken in november and december with november data dominating.

2.2. In which dates did most of the rides did not occur?

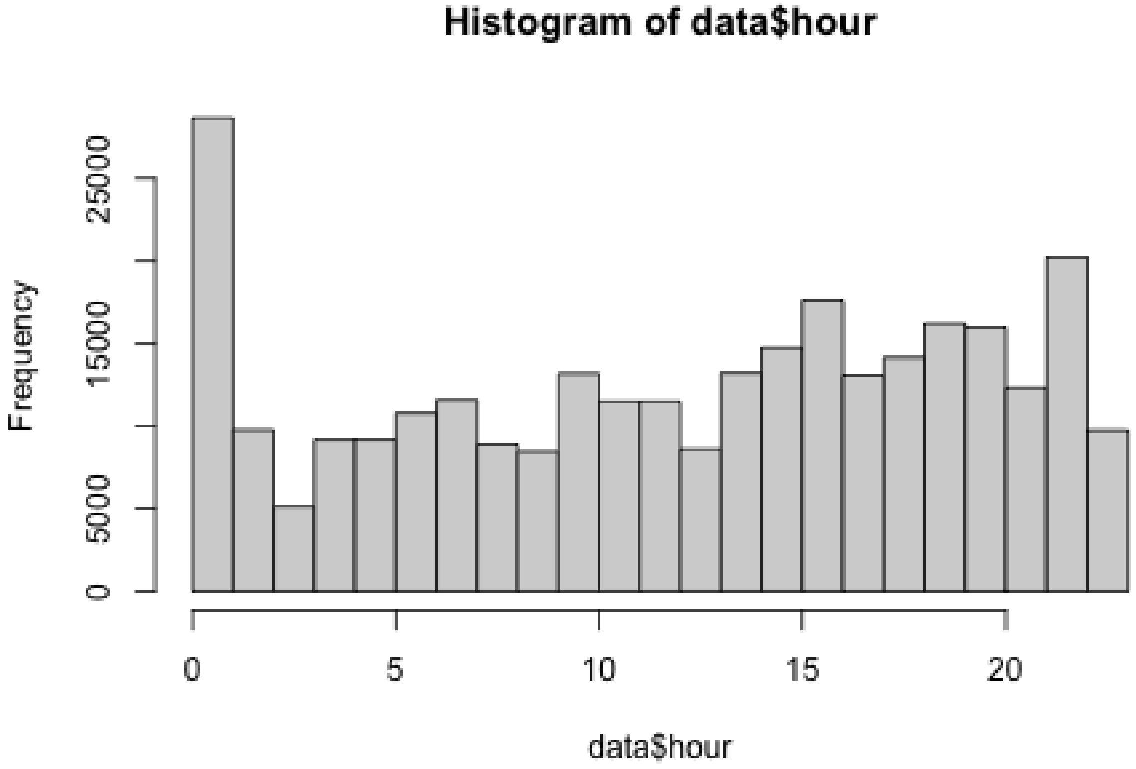
```
hist(data$day)
```



We have many gaps in data in 2 months 4th day to 12th day and from 17th to 25th data are not present in each month.

2.2. How many hours is the data recorded?

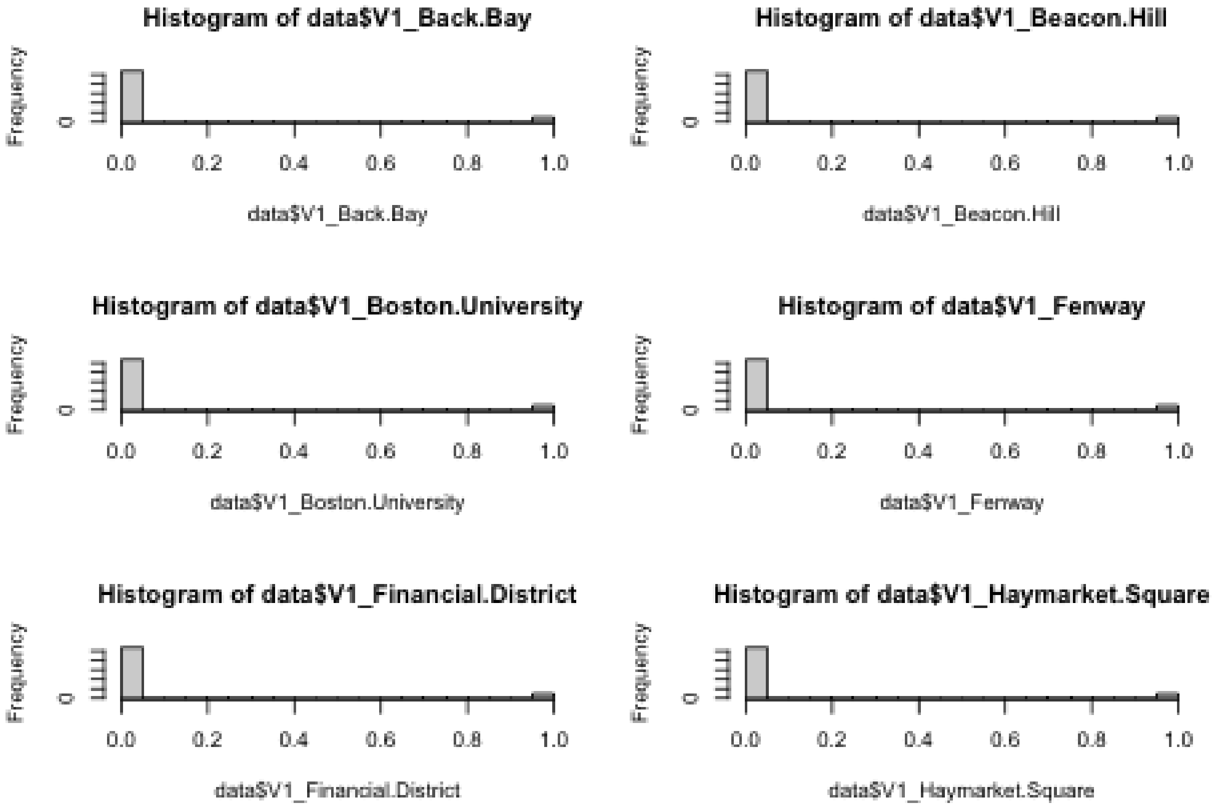
```
hist(data$hour)
```



We have recorderd data of 24hrs.

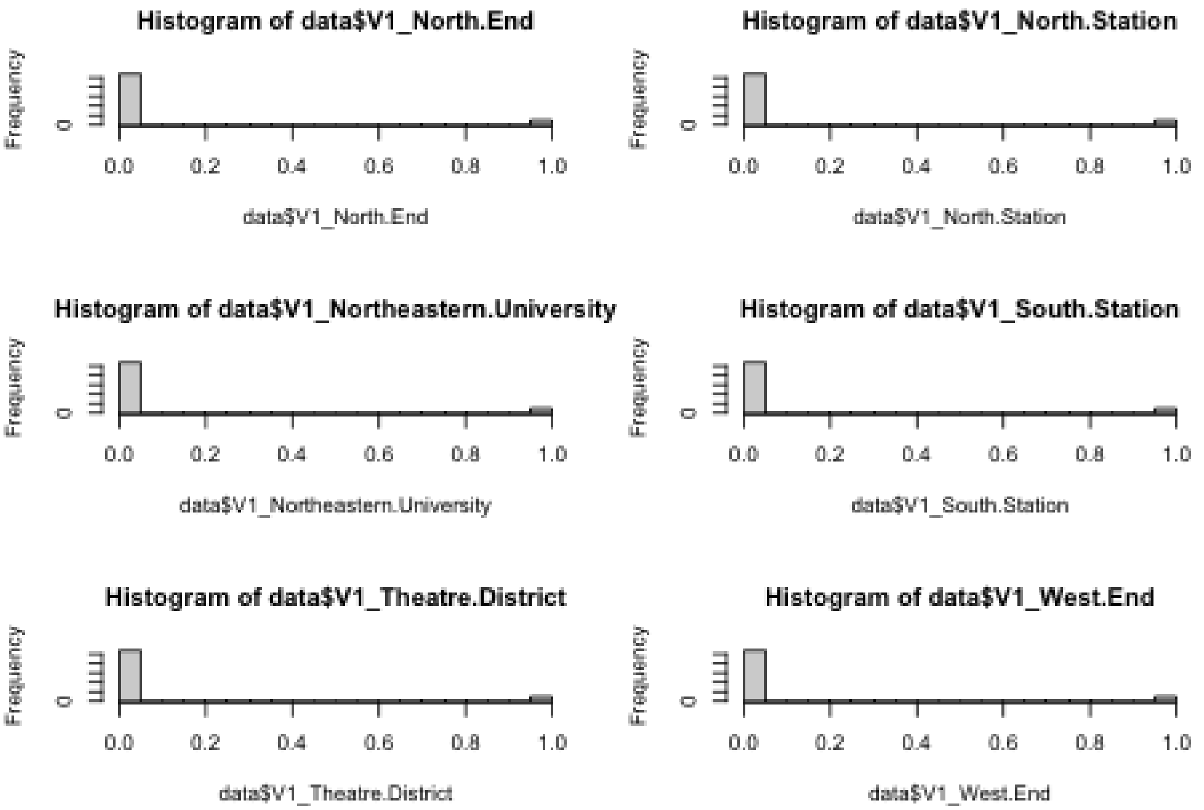
2.3. How many rides are taken from the different source points?

```
par(mfrow = c(3, 2))

hist(data$V1_Back.Bay)
hist(data$V1_Beacon.Hill)
hist(data$V1_Boston.University)
hist(data$V1_Fenway)
hist(data$V1_Financial.District)
hist(data$V1_Haymarket.Square)
```



```r
par(mfrow = c(1, 1))
```

```r
par(mfrow = c(3, 2))

hist(data$V1_North.End)
hist(data$V1_North.Station)
hist(data$V1_Northeastern.University)
hist(data$V1_South.Station)
hist(data$V1_Theatre.District)
hist(data$V1_West.End)
```

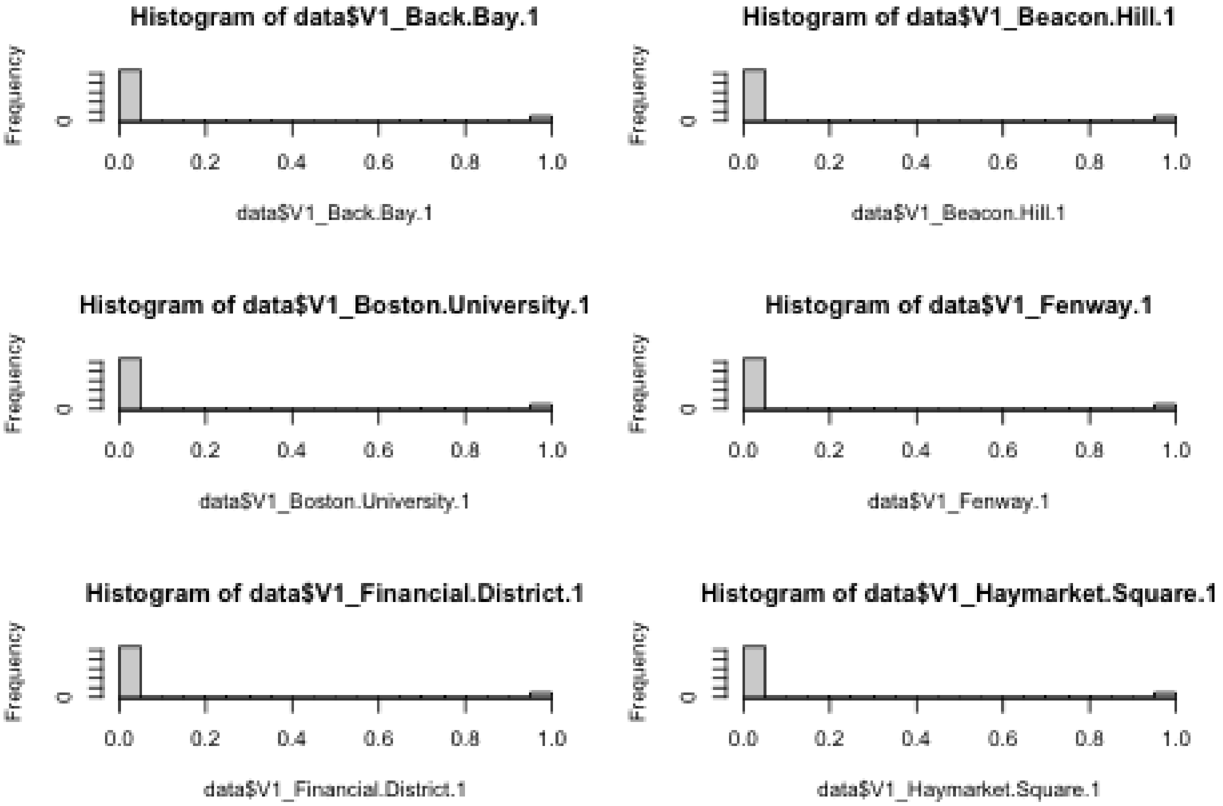


```
par(mfrow = c(1, 1))
```

It seems that all sources are almost equal in number. There are about 50k data in each source feature (Back Bay, Beacon Hill, Boston University, etc)

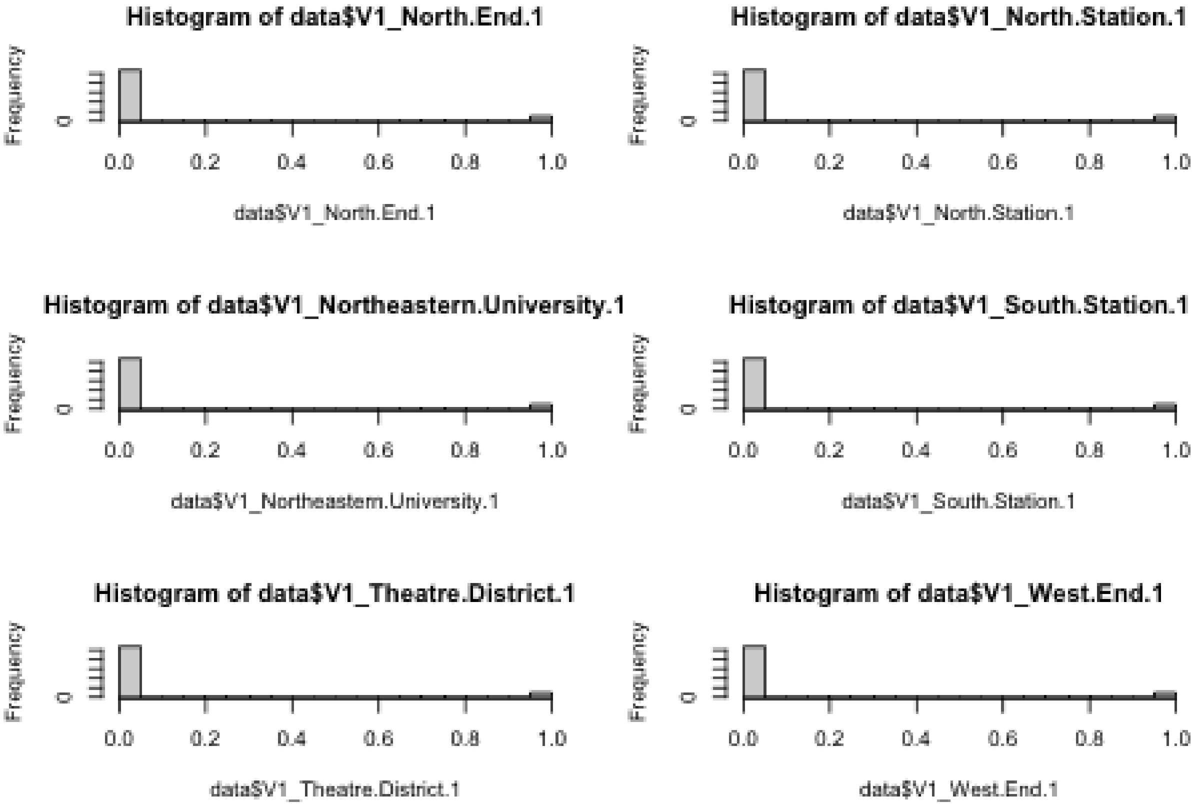
2.4. How many rides are taken from the different destinantion points?

```
par(mfrow = c(3, 2))

hist(data$V1_Back.Bay.1)
hist(data$V1_Beacon.Hill.1)
hist(data$V1_Boston.University.1)
hist(data$V1_Fenway.1)
hist(data$V1_Financial.District.1)
hist(data$V1_Haymarket.Square.1)
```



```r
par(mfrow = c(1, 1))
```

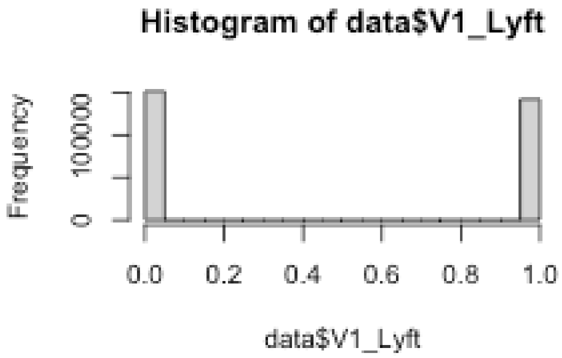
```r
par(mfrow = c(3, 2))

hist(data$V1_North.End.1)
hist(data$V1_North.Station.1)
hist(data$V1_Northeastern.University.1)
hist(data$V1_South.Station.1)
hist(data$V1_Theatre.District.1)
hist(data$V1_West.End.1)
```

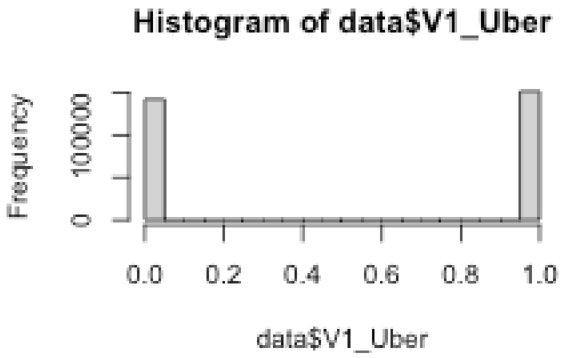


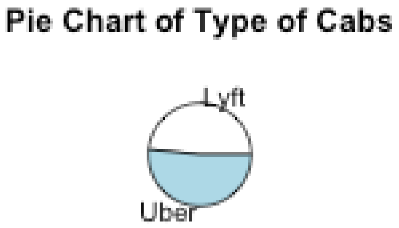
```
par(mfrow = c(1, 1))
```

2.5 How many cab types are used?

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

hist(data$V1_Lyft)
hist(data$V1_Uber)

dcab<- c(sum(data$V1_Lyft==1),sum(data$V1_Uber==1))
ncab <- c("Lyft","Uber")

pie(dcab, labels = ncab, main="Pie Chart of Type of Cabs")
par(mfrow = c(1, 1))
```







Uber data and Lyft data are almost of same size.

2.5 Which cab gives beter fare per mile?

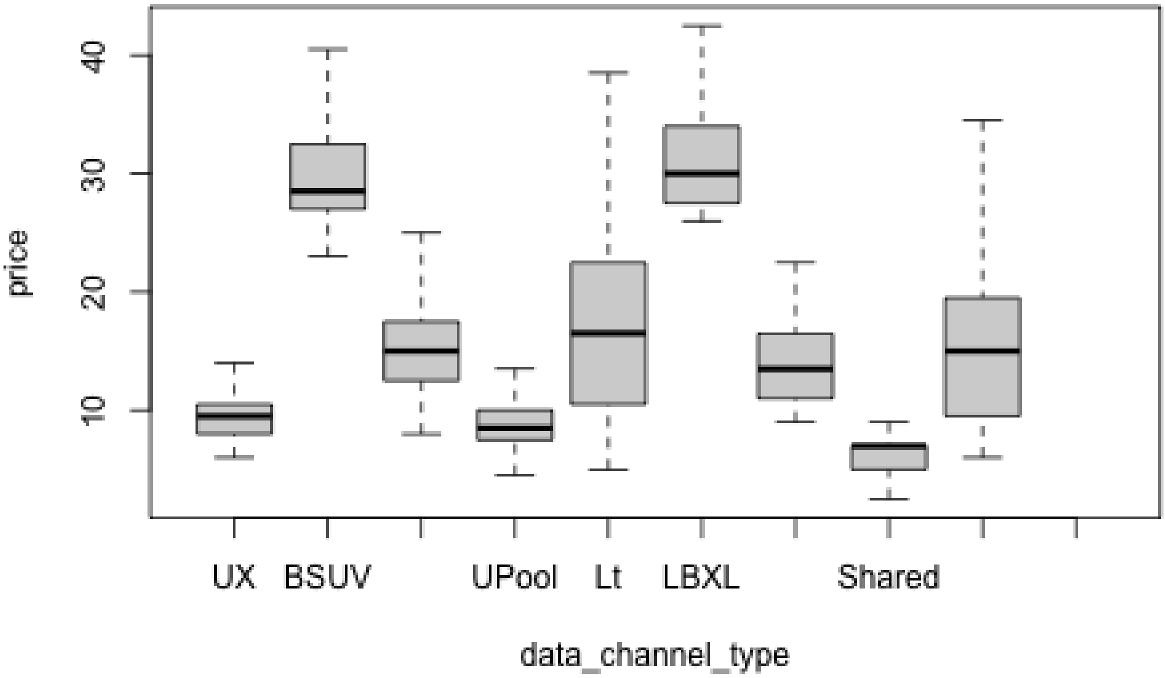
```
data["fare_per_mile"]=round(data$price/data$distance,2)
```

```
vec = seq(1:dim(data)[1])# vector(length = dim(data)[1])

vec[which(data$V1_UberX == T)] = 'UX'
vec[which(data$V1_Black.SUV == T)] = 'BSUV'
vec[which(data$V1_UberXL == T)] = 'UXL'
vec[which(data$V1_UberPool == T)] = 'UPool'
vec[which(data$V1_Lyft == T)] = 'Lt'
vec[which(data$V1_Lux.Black.XL == T)] = 'LBXL'
vec[which(data$V1_Lyft.XL == T)] = 'LXL'
vec[which(data$V1_Shared == T)] = 'Shared'

vec[which(
      data$V1_UberX == F &
      data$V1_Black.SUV == F &
      data$V1_UberXL == F &
      data$V1_UberPool == F &
      data$V1_Lyft == F &
      data$V1_Lux.Black.XL == F &
      data$V1_Lyft.XL == F &
      data$V1_Shared == F
)] = 'NA'


vec_factor = factor(
  vec,
  levels = c('UX', 'BSUV', 'UXL', 'UPool', 'Lt', 'LBXL','LXL','Shared', 'NA', labels = levels)
)
```

```
data$data_channel_type = vec_factor
```

On the boxplots below, we can visualize readers' preferences

```
boxplot(price ~ data_channel_type, data = data, outline = F)
```



Lyft has a better rate for carpool category. Lyft XL has a slightly lower fare per mile than UberXL. Uber Black SUV shows lower rate than Lyft Black XL. Lyft ordinary ride when compared to UberX has higher fare per mile.