***Uber Data Analysis***

Data Import and sanity checks

>install.packages(“tidyverse”)

>library(tidyverse)

Read data into R

uber = read.csv("uber.csv")

Check the dimension of data set

dim(uber)

29101 13

#Uber dataset is of 29101 uber rides (for 6 six months) for 13 different variables

View top and bottom rows to make sure no formatting issues are there or header and footer is included in data set

head(uber)

pickup\_dt borough pickups spd vsb temp dewp slp pcp01 pcp06 pcp24 sd

1 2015-01-01 01:00:00 Bronx 152 5 10 30 7 1023.5 0 0 0 0

2 2015-01-01 01:00:00 Brooklyn 1519 5 10 30 7 1023.5 0 0 0 0

3 2015-01-01 01:00:00 EWR 0 5 10 30 7 1023.5 0 0 0 0

4 2015-01-01 01:00:00 Manhattan 5258 5 10 30 7 1023.5 0 0 0 0

5 2015-01-01 01:00:00 Queens 405 5 10 30 7 1023.5 0 0 0 0

6 2015-01-01 01:00:00 Staten Island 6 5 10 30 7 1023.5 0 0 0 0

hday

1 Y

2 Y

3 Y

4 Y

5 Y

6 Y

tail(uber)

pickup\_dt borough pickups spd vsb temp dewp slp pcp01 pcp06 pcp24 sd

29096 2015-06-30 23:00:00 Brooklyn 990 7 10 75 65 1011.8 0 0 0 0

29097 2015-06-30 23:00:00 EWR 0 7 10 75 65 1011.8 0 0 0 0

29098 2015-06-30 23:00:00 Manhattan 3828 7 10 75 65 1011.8 0 0 0 0

29099 2015-06-30 23:00:00 Queens 580 7 10 75 65 1011.8 0 0 0 0

29100 2015-06-30 23:00:00 Staten Island 0 7 10 75 65 1011.8 0 0 0 0

29101 2015-06-30 23:00:00 <NA> 3 7 10 75 65 1011.8 0 0 0 0

hday

29096 N

29097 N

29098 N

29099 N

29100 N

29101 N

0 0 0 0 N

This looks fine, let us now check for data types and structure

str(uber)

'data.frame': 29101 obs. of 13 variables:

$ pickup\_dt: Factor w/ 4343 levels "2015-01-01 01:00:00",..: 1 1 1 1 1 1 1 2 2 2 ...

$ borough : Factor w/ 6 levels "Bronx","Brooklyn",..: 1 2 3 4 5 6 NA 1 2 3 ...

$ pickups : int 152 1519 0 5258 405 6 4 120 1229 0 ...

$ spd : num 5 5 5 5 5 5 5 3 3 3 ...

$ vsb : num 10 10 10 10 10 10 10 10 10 10 ...

$ temp : num 30 30 30 30 30 30 30 30 30 30 ...

$ dewp : num 7 7 7 7 7 7 7 6 6 6 ...

$ slp : num 1024 1024 1024 1024 1024 ...

$ pcp01 : num 0 0 0 0 0 0 0 0 0 0 ...

$ pcp06 : num 0 0 0 0 0 0 0 0 0 0 ...

$ pcp24 : num 0 0 0 0 0 0 0 0 0 0 ...

$ sd : num 0 0 0 0 0 0 0 0 0 0 ...

$ hday : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...

* Pickup date is date & time stamp and taken as factor
* Borough and hday are factors, rest all are numeric variables

Check summary statistics

summary(uber)

pickup\_dt borough pickups spd

2015-01-01 01:00:00: 7 Bronx :4343 Min. : 0.0 Min. : 0.000

2015-01-01 02:00:00: 7 Brooklyn :4343 1st Qu.: 1.0 1st Qu.: 3.000

2015-01-01 03:00:00: 7 EWR :4343 Median : 54.0 Median : 6.000

2015-01-01 04:00:00: 7 Manhattan :4343 Mean : 490.2 Mean : 5.985

2015-01-01 05:00:00: 7 Queens :4343 3rd Qu.: 449.0 3rd Qu.: 8.000

2015-01-01 10:00:00: 7 Staten Island:4343 Max. :7883.0 Max. :21.000

(Other) :29059 NA's :3043

vsb temp dewp slp pcp01

Min. : 0.000 Min. : 2.00 Min. :-16.00 Min. : 991.4 Min. :0.00000

1st Qu.: 9.100 1st Qu.:32.00 1st Qu.: 14.00 1st Qu.:1012.5 1st Qu.:0.00000

Median :10.000 Median :46.00 Median : 30.00 Median :1018.2 Median :0.00000

Mean : 8.818 Mean :47.67 Mean : 30.82 Mean :1017.8 Mean :0.00383

3rd Qu.:10.000 3rd Qu.:64.50 3rd Qu.: 50.00 3rd Qu.:1022.9 3rd Qu.:0.00000

Max. :10.000 Max. :89.00 Max. : 73.00 Max. :1043.4 Max. :0.28000

pcp06 pcp24 sd hday

Min. :0.00000 Min. :0.00000 Min. : 0.000 N:27980

1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.: 0.000 Y: 1121

Median :0.00000 Median :0.00000 Median : 0.000

Mean :0.02613 Mean :0.09046 Mean : 2.529

3rd Qu.:0.00000 3rd Qu.:0.05000 3rd Qu.: 2.958

Max. :1.24000 Max. :2.10000 Max. :19.000

# Almost all borough has identical distribution, few NA's are observed

# pickup shows possibility of outliers

# visibility of 0 shows extreme conditions, but cannot be ruled out

# temperatures are in Fahrenheit so given range of 2 to 89 translates roughly -16 to 31 Celsius

#NYC's borough - for six areas (Bronx, Brooklyn, EWR, Manhattan , Queens & Staten Island)

#pickups: Number of pickups - from 0 to 7883

#Wind speed in miles/hour - from 0 to 21

# Snow depth in inches - from 0 to 19

# hday: showing 1121 rides on holidays as compared to 27980 rides on working days

#Dew point in Fahrenheit - from -16 to 73

# Sea level pressure - from 991.4 to 1043.4

# Snow depth in inches - from 0 to 19

# liquid precipitation from 0 to 2.1

# Different scales and different variations in weather and local conditions effecting uber rides.

Check for any missing Values : To find NAs in the dataset

anyNA(uber)

[1] TRUE

sum(is.na(uber))

[1] 3043

* This corresponds to missing value of borough as seen in summary output

sapply(uber, function(x) sum(is.na(x)))

pickup\_dt borough pickups spd vsb temp dewp slp pcp01

0 3043 0 0 0 0 0 0 0

pcp06 pcp24 sd hday

0 0 0 0

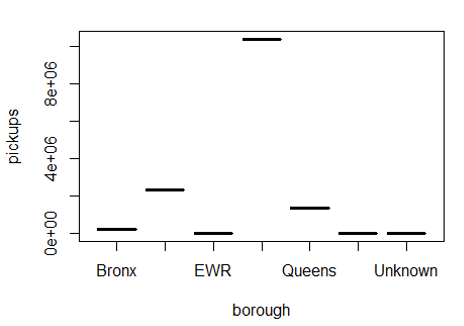
Sapply() iterates over all columns and checks for NA values in given command

This confirm only one column (borough) has NA

Also, borough contains high number of NA values, imputing with any technique might introduce bias. We would instead create a new category called “Unknown” for missing values here.

> uber$borough = as.factor(replace(as.character(uber$borough), is.na(uber$borough),"Unknown"))

> plot(aggregate(pickups~borough,data=uber, sum), type="b")



> table(uber$borough)

Bronx Brooklyn EWR Manhattan Queens

4343 4343 4343 4343 4343

Staten Island Unknown

4343 3043

# To inspect the proportions of different areas

* notice all areas have equally represented excluding Unknowns
* Plot shows Manhattan highest number of rides and almost equally distributed rides to Bronx, EWR, Queens and unknowns borough

**Generate features from date variable**:

Given date variable is in factor form which might not provide meaningful insights.

Let is try to breaking pickup\_dt them into features like month, day , hour etc

# convert date into date form first

##study strptime function..advance functions for treating time stamp variable

> ?strptime

|  |  |
| --- | --- |
| strptime {base} | R Documentation |

Date-time Conversion Functions to and from Character

**Description**

Functions to convert between character representations and objects of classes "POSIXlt" and"POSIXct" representing calendar dates and times.

> uber$start\_date = strptime(uber$pickup\_dt,'%Y-%m-%d %H:%M')

> library(lubridate) #Lubridate is an R package that makes it easier to work with dates and times

> uber$start\_month = month(uber$start\_date)

> uber$start\_day = day(uber$start\_date)

> uber$start\_hour = hour(uber$start\_date)

> uber$wday = weekdays(uber$start\_date)

> uber = uber[,-14]

* We have added new features for month of ride, day of month and hour of ride. Also

wday represent which day of week it is.

**Check for number of holidays each month**

> # try to get no. of holidays in each month

> #unique function used to keep only unique/distinct rows from a data frame

> unique(uber[which(uber$hday=="Y"),c("start\_day","start\_month")])

start\_day start\_month

1 1 1

2848 19 1

6649 12 2

7293 16 2

20608 10 5

23055 25 5

24526 3 6

* We can see that we have two holidays in Jan (1st & 19th), 2 in Feb (12th & 16th), 2 in May(10th & 25th) and 1 in June(3rd). No holidays in March and April

>table(uber$hday,uber$start\_month)

1 2 3 4 5 6

N 4588 4169 4957 4798 4730 4738

Y 309 323 0 0 328 161

* No trips in 3rd and 4th month…Looks like no holidays in these month

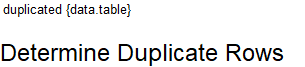
This shows number of trips in holidays vs non-holidays in month

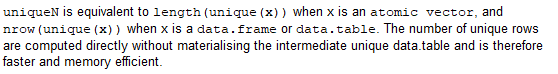
We will come again to check the effect on trips on holidays vs non-holidays

Before that let us do some univariate analysis

> library(data.table) #widely used for fast aggregation of large datasets

> ?uniqueN



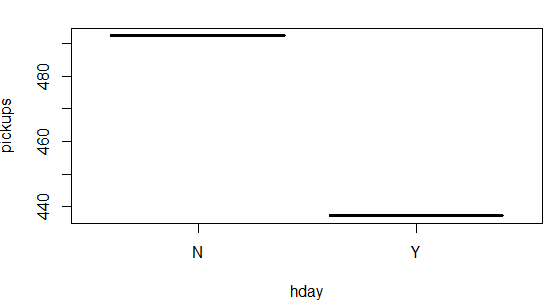


> uniqueN(uber, by=c('start\_month', 'start\_day'))

[1] 181

* In total, our days is for 181 days in Jan 2015 to June 2015

> plot(aggregate(pickups~hday,data=uber, mean), type="b")

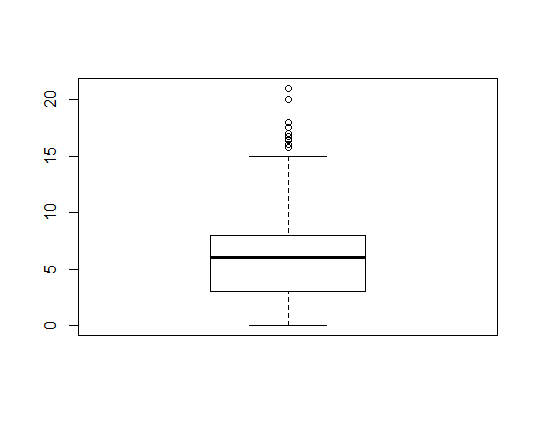


* Rides on working days is higher than on holidays

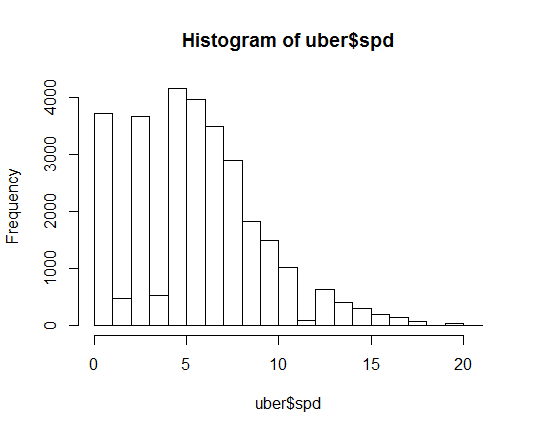
**Uni-Variate Analysis**

Speed:

>boxplot(uber$spd) # outlier present



> hist(uber$spd)# skewed histogram



* Boxplot shows there are outliers in data set.
* Histogram also shows the right skew in distribution
* On an average speed is 5 miles/hour

**Check the distribution for pickups**

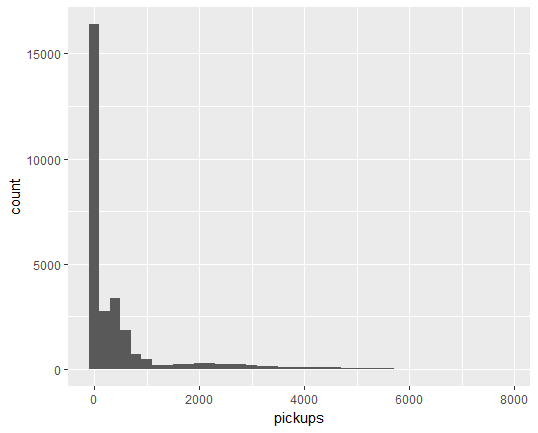
> library(ggplot2) #a system for declaratively creating graphics

> #for pick up counts

> ggplot(uber, aes(pickups)) +

+ geom\_histogram(binwidth = 200)

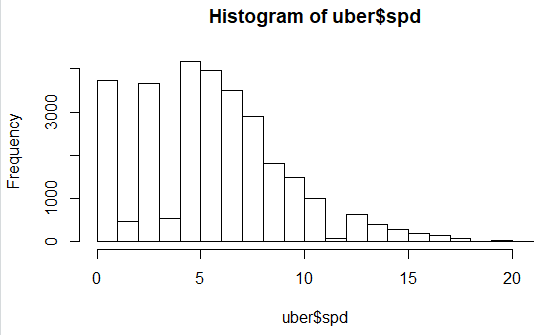
**# Histogram is heavily skewed**



* Many have 0 rides or close to it.
* But skew is clearly visible
* **check for outliers in other variable as well**

**For wind speed:**

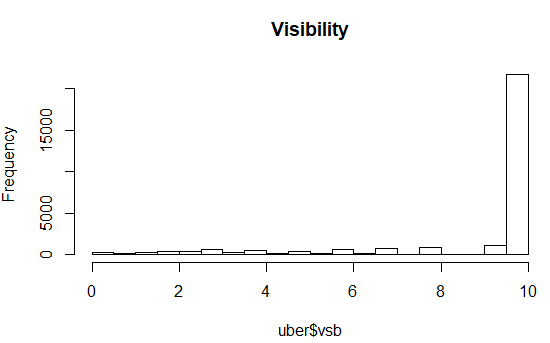
> hist(uber$spd)



* Low speed for duration, except few outliers, avg is around 5

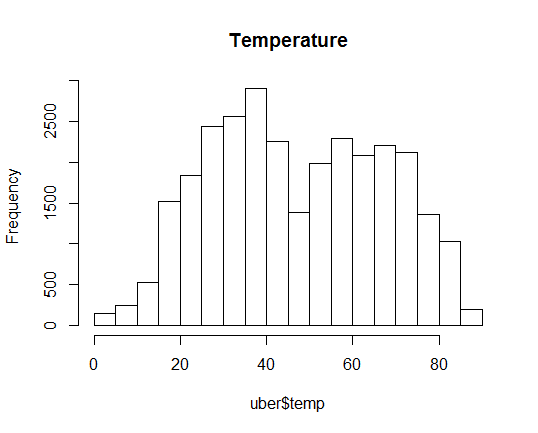
**Visibility**

> hist(uber$vsb, main= "Visibility")



* Almost clear weather

**Temperature**

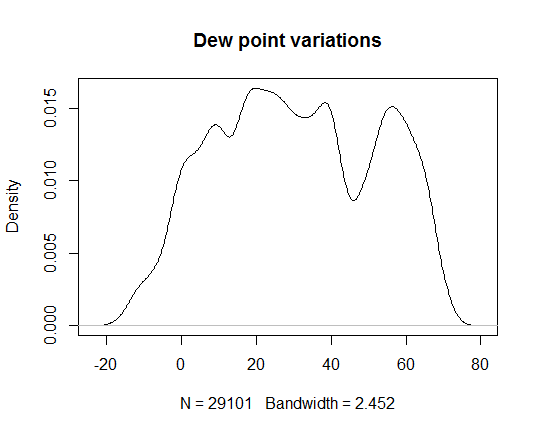


* Two peaks can be seen, one at around 35F and other one at around 60F (bi-modal)
* It peaks at 35 (~1.5 C) suggest cold weather conditions, summers are not so

intense

**Dew Point**

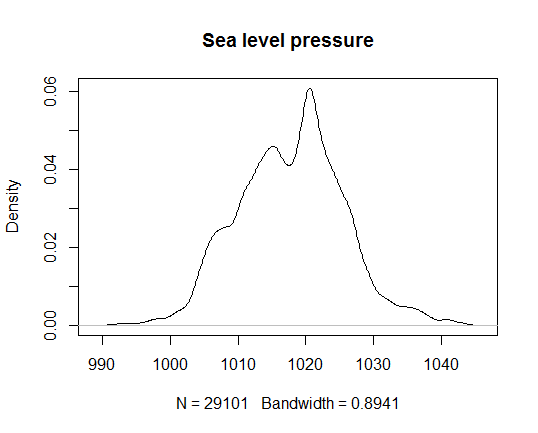
> plot(density(uber$dewp), main="Dew point variations")



* Distribution is quite like that of temperature(bi-modal)

**Sea level pressure**

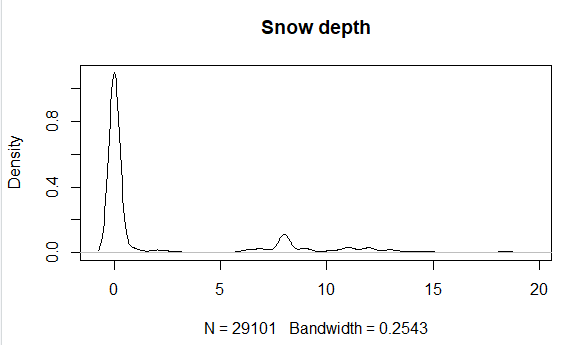
> plot(density(uber$slp), main="Sea level pressure")



* This resembles normal distribution
* We would expect pressure, temperature and dew points to show some correlation, hence we can expect similar distribution for them.

**Snow depth**

> plot(density(uber$sd), main="Snow depth")



* No snow for majority of times

**Bi-variate Analysis**

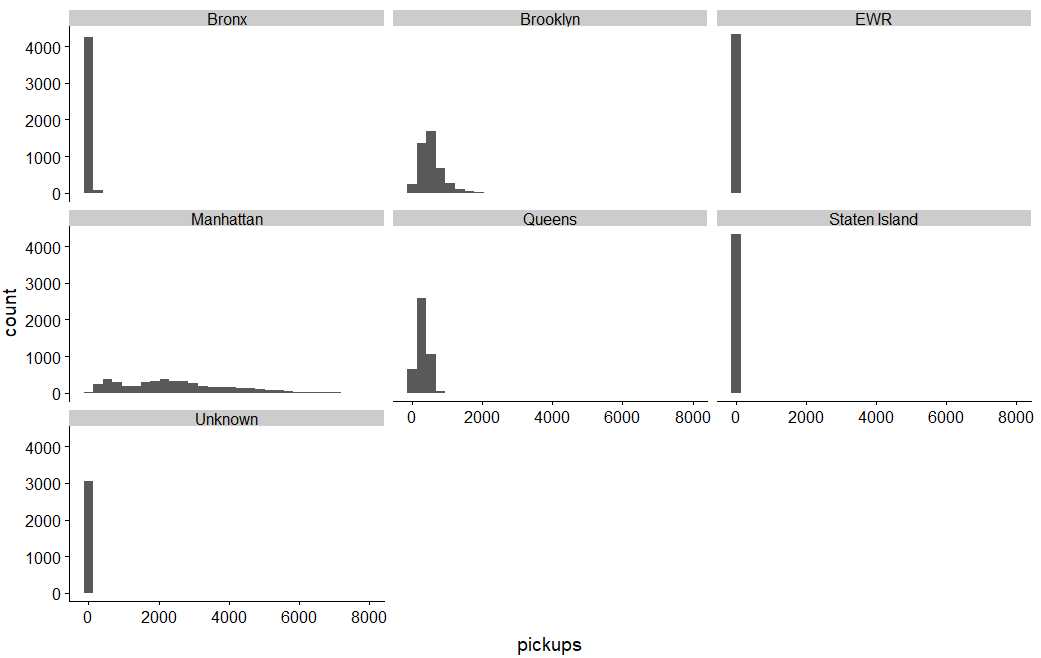
> #Borough wise pickup

> ggplot(uber, aes(pickups)) +

+ geom\_histogram() +

+ facet\_wrap(~ borough, ncol = 3)

Pickups broken by boroughs



> uber %>% group\_by(borough) %>%

+ summarise(`Total Pickups` = sum(pickups)) %>%

+ arrange(desc(`Total Pickups`))

# A tibble: 7 x 2

borough `Total Pickups`

*<fct>* *<int>*

1 Manhattan 10367841

2 Brooklyn 2321035

3 Queens 1343528

4 Bronx 220047

5 Staten Island 6957

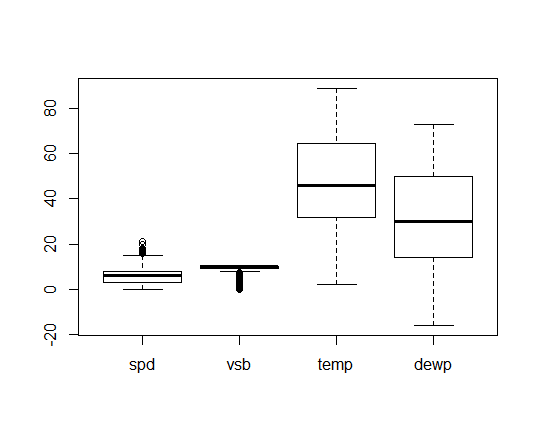
6 Unknown 6260

7 EWR 105

* Majority of 0 rides are in unknown, Staten Island, EWR and Bronx
* Manhattan seems to have highest demand and then Brooklyn

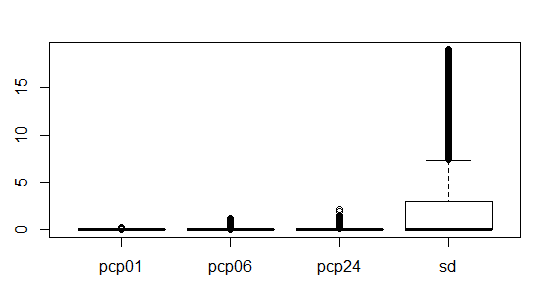
**Multivariate Analysis:**

> boxplot(uber[,c(4:7)])



* Temperature and dew points doesn’t show any outliers

> boxplot(uber[,c(9:12)])



* Pcp01 has less of outliers, sd shows plenty
* Check variable distributions

**Inference:**

We did univariate, bivariate and multivariate analysis to examine each variable and with other variable contributing toward uber rides.

Our analysis is for six boroughs in NYC for 181 total days and also looked at number of holiday every month.

In terms of borough, Manhattan contributes to largest share in bookings done

Holiday was another variable which show number of bookings on non-holidays compared to holidays. Point to note is that holidays and non-holidays does not include week day off. It just compares 6 holidays against the regular days. We can stretch this by considering all Sundays as holiday and replotting the difference

We used different libraries to plot and examine these variables like tidyverse, ggplot2, data.table and lubridata(for date variable).