Assignment - 1
ARNAB BIR - 14MA20009

Topic 1 : CAR data with 50 observations

a) Carefully observe the data. Apparently what it seems?

This data represents the distance to be covered to stop and speeds of cars. The distance increases as the speed increases.

b) It is proposed to analyse the data with the calculations of AM (Arithmetic Mean), GM (Geometric Mean) and HM (Harmonic Mean). Calculate such mean measures. Which mean calculation has a real significance to the data? Justify your answers.

The AM , GM and HM of the data are as follows.

AM of Speed = 15.4 GM of Speed = 14.32501 HM of Speed = 12.96153

AM of Dist = 42.98 GM of Dist = 34.32615 HM of Dist = 22.18214

We can easily visualize that due to positive values of speed and distance AM > GM > HM. Now, Median Speed = 15 which is close to AM of speed. The calculation of AM seems more significant. Again if either of the values is 0 (zero) the value of arithmetic and geometric mean becomes insignificant. Hence AM is a more significant parameter.

c) Do you suggest any other measurement(s) which might be useful implications?

Standard deviation (or variance), median, quantile values, correlation coefficient between speed and distance are some of the important measurements apart from the mentioned means.

Script:

```
CARS <- read.csv("../Data/CARS.csv")
head(CARS)
am_speed <- mean(CARS$speed)
am_dist <- mean(CARS$dist)
gm_speed <- prod(CARS$speed)^(1/length(CARS$speed))
gm_dist <- prod(CARS$dist)^(1/length(CARS$dist))
hm_speed <- 1/mean(1/CARS$speed)
hm_dist <- 1/mean(1/CARS$dist)
me_speed <- median (cars$speed)
me_dist <- median (cars$dist)
```

Topic 2 : EARTHQUAKE data with 8086 observations

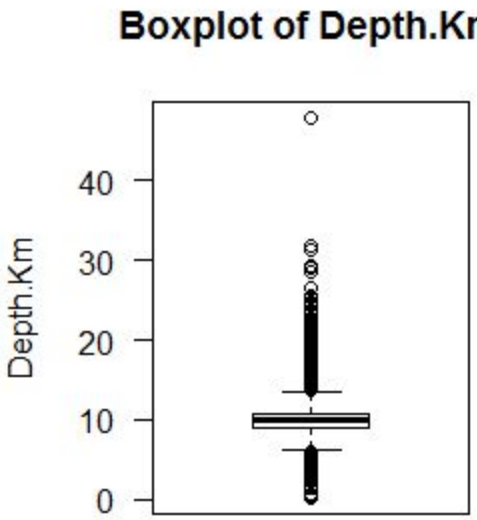
a) The table includes the severity of earthquakes at different places in India during the year 2016. You are advised to browse the data carefully. Point out the discrepancy(ies), if any.

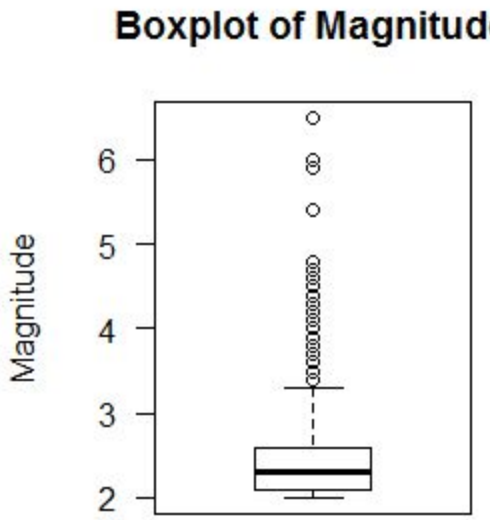
Although it is mentioned that the data contains the severity of earthquakes at different places during India, the earthquakes which happened in between 24 th of August and 30th of November are mentioned. Again the distributions are platykurtic and do not have high concentration towards centre

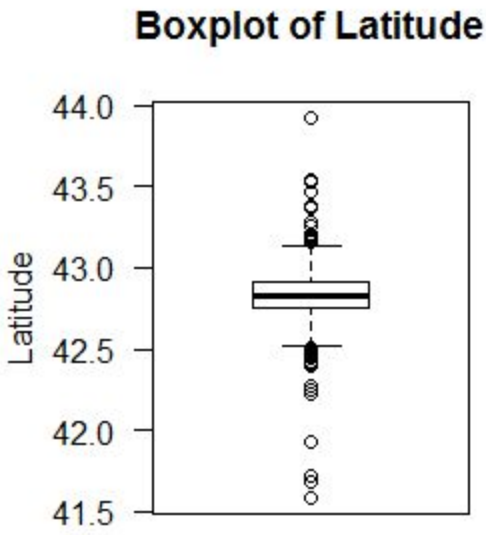
.

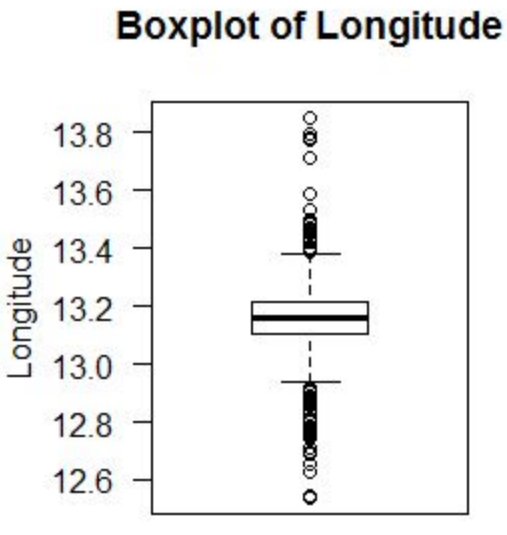
b) For the given data, calculate the "Five point summary" and hence draw the boxplot.

Time	Latitude	Longitude	Depth.Km	Magnitude
Min. :2016-08-24 03:36:32	Min. :41.58	Min. :12.54	Min. : 0.200	Min. :2.000
1st Qu.:2016-09-16 09:53:05	1st Qu.:42.75	1st Qu.:13.10	1st Qu.: 8.900	1st Qu.:2.100
Median :2016-10-31 12:17:52	Median :42.82	Median :13.16	Median : 9.900	Median :2.300
Mean :2016-10-18 12:19:09	Mean :42.83	Mean :13.16	Mean : 9.968	Mean :2.398
3rd Qu.:2016-11-07 05:02:03	3rd Qu.:42.91	3rd Qu.:13.22	3rd Qu.:10.800	3rd Qu.:2.600
Max. :2016-11-30 20:54:35	Max. :43.93	Max. :13.85	Max. :47.900	Max. :6.500









c) Use the IQR calculation and then decide any data as outlier(s). Remove the outlier(s), if found. Taking the cleaned data, obtain the box plot? Compare the two box plots.

The quartiles of different variables are as follows.

Time	Latitude	Longitude	Depth.Km	Magnitude
1st Qu.:2016-09-16 09:53:05	1st Qu.:42.75	1st Qu.:13.10	1st Qu.: 8.900	1st Qu.:2.100
3rd Qu.:2016-11-07 05:02:03	3rd Qu.:42.91	3rd Qu.:13.22	3rd Qu.:10.800	3rd Qu.:2.600

Now the IQR can be represented as,

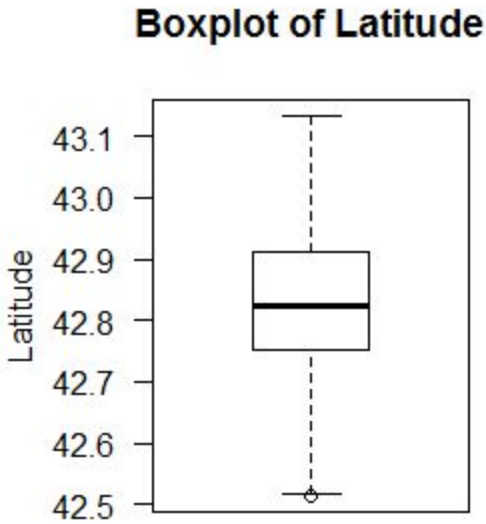
$$IQR = Q\_3 - Q\_1$$

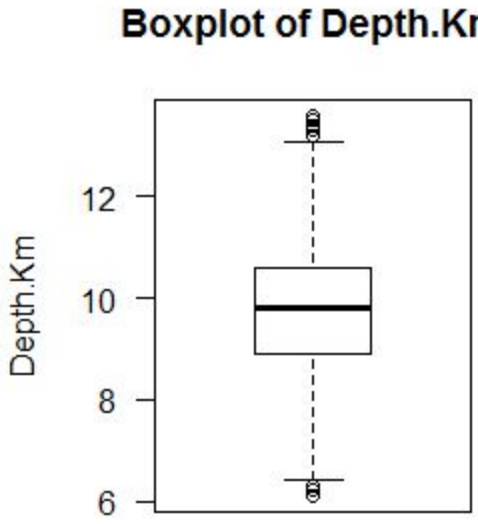
In order to remove the outliers, we need to calculate the the maximum and the minimum value. We calculate maximum and minimum as,

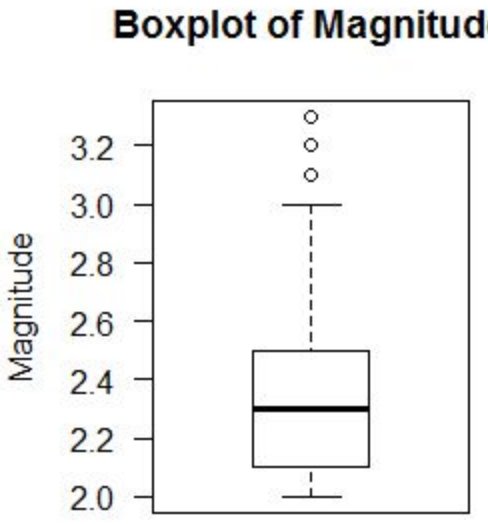
$$Max = Q\_3 + 1.5 * IQR$$
$$Min = Q\_1 - 1.5 * IQR$$

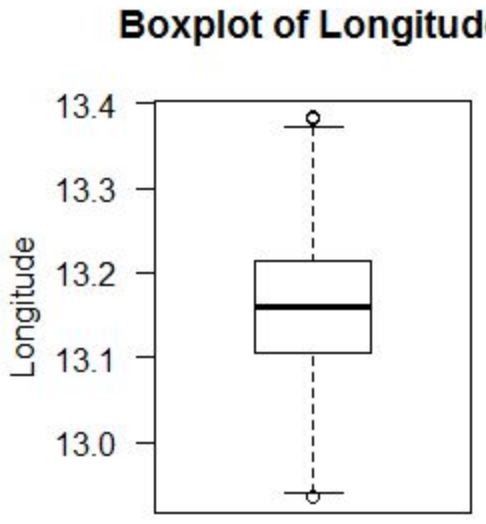
The values higher than the Max and lower than the Min are the outliers.

After removing the outliers, the boxplots become as follows.


Boxplot of Latitude
Latitude
43.1
43.0
42.9
42.8
42.7
42.6
42.5


Boxplot of Depth.Km
Depth.Km
12
10
8
6


Boxplot of Magnitude
Magnitude
3.2
3.0
2.8
2.6
2.4
2.2
2.0


Boxplot of Longitude
Longitude
13.4
13.3
13.2
13.1
13.0

Script:

```
require(xlsx)
EARTHQUAKE <- read.xlsx("../Data/EARTHQUAKE.xlsx", sheetIndex =
1)
head(EARTHQUAKE)
EARTHQUAKE[which(is.na(EARTHQUAKE$Longitude)),]

min(EARTHQUAKE$Time)
max(EARTHQUAKE$Time)
min(EARTHQUAKE$Latitude)
max(EARTHQUAKE$Latitude)
min(EARTHQUAKE$Longitude)
max(EARTHQUAKE$Longitude)
min(EARTHQUAKE$Depth.Km)
max(EARTHQUAKE$Depth.Km)
min(EARTHQUAKE$Magnitude)
max(EARTHQUAKE$Magnitude)

summary(EARTHQUAKE)
boxplot(EARTHQUAKE$Depth.Km, las = 2, ylab = "Depth.Km", main =
"Boxplot of Depth.Km")
boxplot(EARTHQUAKE$Magnitude, las = 2, ylab = "Magnitude", main
= "Boxplot of Magnitude")
boxplot(EARTHQUAKE$Latitude, las = 2, ylab = "Latitude", main =
"Boxplot of Latitude")
boxplot(EARTHQUAKE$Longitude, las = 2, ylab = "Longitude", main
= "Boxplot of Longitude")

max_Depth.Km <- quantile(EARTHQUAKE$Depth.Km, 0.75, na.rm=TRUE)
+ (IQR(EARTHQUAKE$Depth.Km, na.rm=TRUE) * 1.5 )
min_Depth.Km <- quantile(EARTHQUAKE$Depth.Km,0.25, na.rm=TRUE) -
(IQR(EARTHQUAKE$Depth.Km, na.rm=TRUE) * 1.5 )

max_Magnitude <- quantile(EARTHQUAKE$Magnitude, 0.75,
na.rm=TRUE) + (IQR(EARTHQUAKE$Magnitude, na.rm=TRUE) * 1.5 )
min_Magnitude <- quantile(EARTHQUAKE$Magnitude,0.25, na.rm=TRUE)
- (IQR(EARTHQUAKE$Magnitude, na.rm=TRUE) * 1.5 )
```

```
max_Latitude <- quantile(EARTHQUAKE$Latitude, 0.75, na.rm=TRUE)
+ (IQR(EARTHQUAKE$Latitude, na.rm=TRUE) * 1.5 )
min_Latitude <- quantile(EARTHQUAKE$Latitude,0.25, na.rm=TRUE) -
(IQR(EARTHQUAKE$Latitude, na.rm=TRUE) * 1.5 )

max_Longitude <- quantile(EARTHQUAKE$Longitude, 0.75,
na.rm=TRUE) + (IQR(EARTHQUAKE$Longitude, na.rm=TRUE) * 1.5 )
min_Longitude <- quantile(EARTHQUAKE$Longitude,0.25, na.rm=TRUE)
- (IQR(EARTHQUAKE$Longitude, na.rm=TRUE) * 1.5 )

Depth.Km_idx <- which(EARTHQUAKE$Depth.Km < max_Depth.Km &
EARTHQUAKE$Depth.Km > min_Depth.Km)
Magnitude_idx <- which(EARTHQUAKE$Magnitude < max_Magnitude &
EARTHQUAKE$Magnitude > min_Magnitude)
Latitude_idx <- which(EARTHQUAKE$Latitude < max_Latitude &
EARTHQUAKE$Latitude > min_Latitude)
Longitude_idx <- which(EARTHQUAKE$Longitude < max_Longitude &
EARTHQUAKE$Longitude > min_Longitude)

boxplot(EARTHQUAKE$Depth.Km[Depth.Km_idx], las = 2, ylab =
"Depth.Km", main = "Boxplot of Depth.Km")
boxplot(EARTHQUAKE$Magnitude[Magnitude_idx], las = 2, ylab =
"Magnitude", main = "Boxplot of Magnitude")
boxplot(EARTHQUAKE$Latitude[Latitude_idx], las = 2, ylab =
"Latitude", main = "Boxplot of Latitude")
boxplot(EARTHQUAKE$Longitude[Longitude_idx], las = 2, ylab =
"Longitude", main = "Boxplot of Longitude")
```

Topic 3 : AUTOMOBILE data with 205 observations

a) Categorize all the attributes listed in the table according to the NOIR topology?

The attributes are the following.

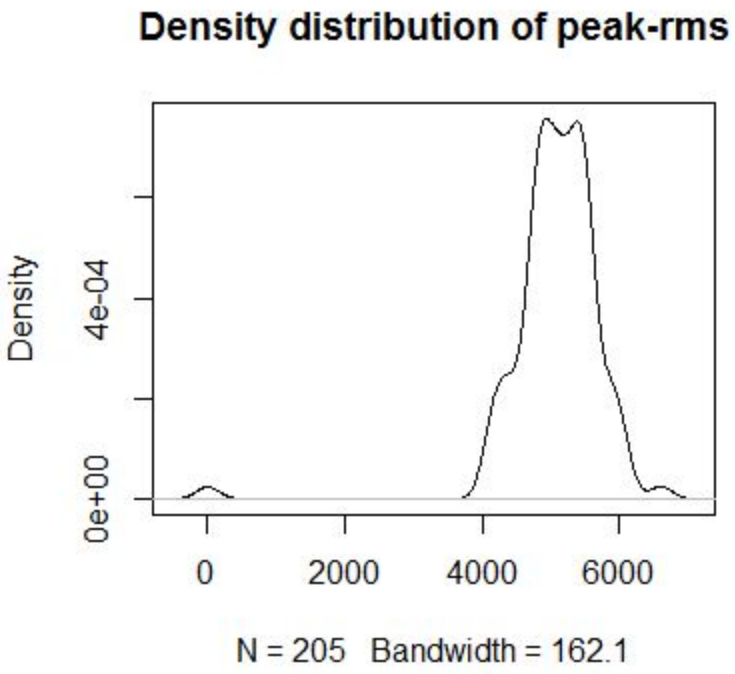
```
[1] "symboling"   "normalized.losses"   "make"   "fuel.type"  "aspiration"   "num.of.doors"
 [7] "body.style"  "drive.wheels"  "engine.location"  "wheel.base"  "length"    "width"
[13] "height"    "curb.weight"   "engine.type"   "num.of.cylinders"  "engine.size"
"fuel.system"
[19] "bore"    "stroke"   "compression.ratio"    "horsepower"   "peak.rpm"    "city.mpg"
[25] "highway.mpg"       "price"
```

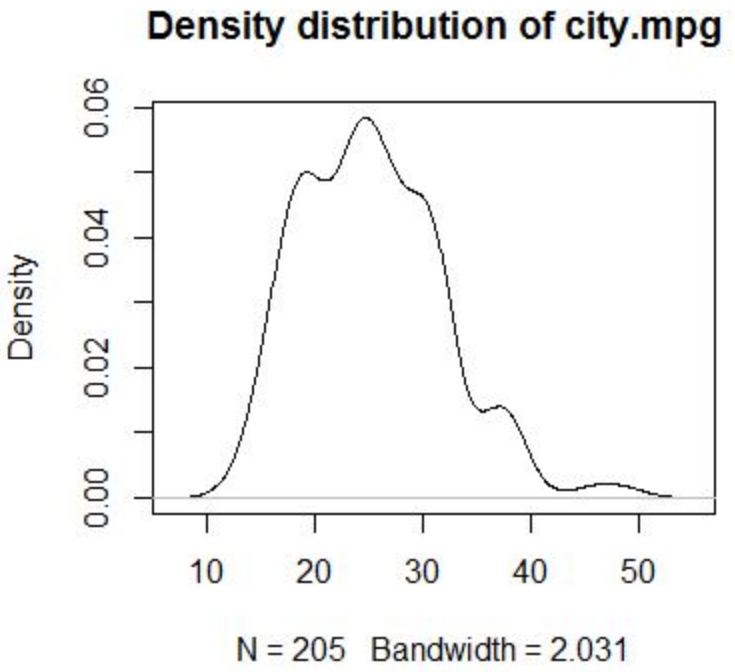
Nominal Variables : make, fuel.type, aspiration, body.style, drive.wheels, engine.location, engine.type, fuel.system.
Ordinal Variables : symboling, num.of.doors, num.of.cylinders
Interval Variables : normalized.loses, wheel.base, length, width, height, curb.weight, engine.size, horsepower, peak.rpm, city.mpg, highway,.mpg, price
Ratio Variables : compression.ratio

b) Apply the applicable central tendency measures to any four attributes taking one attribute from each category.

Nominal Variables :: Central Tendency Measures : aspiration
Mode = Gas
Ordinal Variables :: Central Tendency Measures : num.of.doors
Mean = 4.380488
Median = 4
Mode = 4
Interval Variables :: Central Tendency Measures : wheel.base
Mean = 98.75659
Median = 97
Mode = 94.5
Ratio Variables :: Central Tendency Measures : compression.ratio
Mean = 10.14254
Median = 9
Mode = 9

c) Consider the attribute "peak-rpm" and "city-mpg"? Find which probability distribution(s) they are likely to follow?





Both of the variables follow gaussian distribution. The distribution of peak-rms is negatively skewed and the distribution of city.mpg is positively skewed.

Script:

```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

AUTOMOBILES <- read.csv("../Data/AUTOMOBILES.csv")
head(AUTOMOBILES)
names(AUTOMOBILES)

length(AUTOMOBILES$fuel.type[AUTOMOBILES$fuel.type == "gas"])
length(AUTOMOBILES$fuel.type[AUTOMOBILES$fuel.type == "diesel"])
unique(AUTOMOBILES$num.of.cylinders)
AUTOMOBILES$num_of_cylinders[AUTOMOBILES$num.of.cylinders ==
"four"] = as.numeric(4)
AUTOMOBILES$num_of_cylinders[AUTOMOBILES$num.of.cylinders ==
"six"] = as.numeric(6)
AUTOMOBILES$num_of_cylinders[AUTOMOBILES$num.of.cylinders ==
"five"] = as.numeric(5)
AUTOMOBILES$num_of_cylinders[AUTOMOBILES$num.of.cylinders ==
"twelve"] = as.numeric(12)
AUTOMOBILES$num_of_cylinders[AUTOMOBILES$num.of.cylinders ==
"two"] = as.numeric(2)
AUTOMOBILES$num_of_cylinders[AUTOMOBILES$num.of.cylinders ==
"eight"] = as.numeric(8)
AUTOMOBILES$num_of_cylinders[AUTOMOBILES$num.of.cylinders ==
"three"] = as.numeric(3)

mean(AUTOMOBILES$num_of_cylinders)
median(AUTOMOBILES$num_of_cylinders)
Mode(AUTOMOBILES$num_of_cylinders)

mean(AUTOMOBILES$compression.ratio)
median(AUTOMOBILES$compression.ratio)
Mode(AUTOMOBILES$compression.ratio)
```

```
AUTOMOBILES$peak.rpm <- gsub("?",0,AUTOMOBILES$peak.rpm, fixed =
TRUE)
AUTOMOBILES$peak.rpm <-
as.numeric(as.character(AUTOMOBILES$peak.rpm))
d <- density(as.numeric(as.character(AUTOMOBILES$peak.rpm))) #
returns the density data
plot(d, main = "Density distribution of peak-rms")

AUTOMOBILES$city.mpg <- gsub("?",0,AUTOMOBILES$city.mpg, fixed =
TRUE)
AUTOMOBILES$city.mpg <-
as.numeric(as.character(AUTOMOBILES$city.mpg))
d <- density(as.numeric(as.character(AUTOMOBILES$city.mpg))) #
returns the density data
plot(d, main = "Density distribution of city.mpg")
```

Topic 4 : IRIS data with 50 observations

a) Consider the 150 observations as very close to population data. Find the population Mean.

Sepal Length Mean = 5.843333 cm
Sepal Width Mean = 3.054 cm
Petal length Mean = 3.758667 cm
Petal Width Mean = 1.198667 cm

b) Assume a sample of size 50 chosen at random, find the population variance.

Sepal Length Variance of a sample = 0.6401673
Sepal Width Variance of a sample = 0.1310571
Petal Length Variance of a sample = 2.924098
Petal Width Variance of a sample = 0.5788776

c) Compare the sample variance with that of population variance?

Sepal Length Variance of population = 0.6856935

Sepal Width Variance of population = 0.188004
Petal Length Variance of population = 3.113179
Petal Width Variance of population = 0.5824143

Population variance > sample variance. Due to more number of observations, the population variance always has higher value than sample variance.

Script:

```
library(xlsx)
IRIS = read.xlsx("../Data/IRIS.xlsx",sheetIndex = 1)

#(a)Calculate the population mean
s.len.mean = mean(IRIS$SepalLengthCm)
s.wid.mean = mean(IRIS$SepalWidthCm)
p.len.mean = mean(IRIS$PetalLengthCm)
p.wid.mean = mean(IRIS$PetalWidthCm)


IRIS.sample = data[sample(nrow(data), 50),]
sample.s.len.var = var(IRIS.sample$SepalLengthCm)
sample.s.wid.var = var(IRIS.sample$SepalWidthCm)
sample.p.len.var = var(IRIS.sample$PetalLengthCm)
sample.p.wid.var = var(IRIS.sample$PetalWidthCm)

population.s.len.var = var(IRIS$SepalLengthCm)
population.s.wid.var = var(IRIS$SepalWidthCm)
population.p.len.var = var(IRIS$PetalLengthCm)
population.p.wid.var = var(IRIS$PetalWidthCm)
```