**Topic – 2**

EARTHQUAKE

Name :- Akash Mandal

Roll No. :- 16MA20007

Course :- CS40003

Problem Statement –

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.

b) For the given data, calculate the “Five point summary” and hence draw the box plot.

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.

Reference –

EARTHQUAKE.xlsx (Included in the project folder for reference)

Observation of data:

The database EARTHQUAKE.csv reveals the following basic structure-

* 8086 observations of 5 variables
* Time – POSIXct (date-time), format: “yyyy-mm-dd hh:mm:ss”
* Latitude – num (dbl)
* Longitude – num (dbl)
* Depth(km) – num (dbl)
* Magnitude – num (dbl)

Code Snippet

> head(quake)

# A tibble: 6 × 5

Time Latitude Longitude `Depth/Km` Magnitude

<dttm> <dbl> <dbl> <dbl> <dbl>

1 2016-08-24 03:36:32 42.6983 13.2335 8.1 6.0

2 2016-08-24 03:37:26 42.7123 13.2533 9.0 4.5

3 2016-08-24 03:40:46 42.7647 13.1723 9.7 3.8

4 2016-08-24 03:41:38 42.7803 13.1683 9.7 3.9

5 2016-08-24 03:42:07 42.7798 13.1575 9.7 3.6

6 2016-08-24 03:43:58 42.7298 13.2137 10.7 3.4

> str(quake)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 8086 obs. of 5 variables:

$ Time : POSIXct, format: "2016-08-24 03:36:32" "2016-08-24 03:37:26" "2016-08-24 03:40:46" ...

$ Latitude : num 42.7 42.7 42.8 42.8 42.8 ...

$ Longitude: num 13.2 13.3 13.2 13.2 13.2 ...

$ Depth/Km : num 8.1 9 9.7 9.7 9.7 10.7 10.8 10.6 6.6 12.4 ...

$ Magnitude: num 6 4.5 3.8 3.9 3.6 3.4 2.7 3.2 3.7 2.8 ...

Checking the data-

We run a preliminary check on the data-

1. Time, Magnitude – A scatter-plot of magnitude of quakes against time shows a roughly continuous spread over time, and no irregularities are apparent in the Time variable.  
   As for the magnitude, we notice majority of values below 5, and 4-5 values above 5. These are worth checking while deciding if they are outliers. However all values are below 7 (max : 6.5), and are hence, possible.
2. Time, Depth – A scatter plot of Depth(km) against Time reveals most quakes have depth less than 30km. A few lie at a depth greater than 30, and are worth investigating for possible outliers.
3. Latitude, Longitude – A scatter plot of Latitude vs Longitude shows majority of data concentrated between Latitudes 42 and 43.5, and Longitudes 12.7 to 13.6. Some scattered points also exist beyond these ranges, which are worth investigation.

Code Snippet

> #check for discrepancies

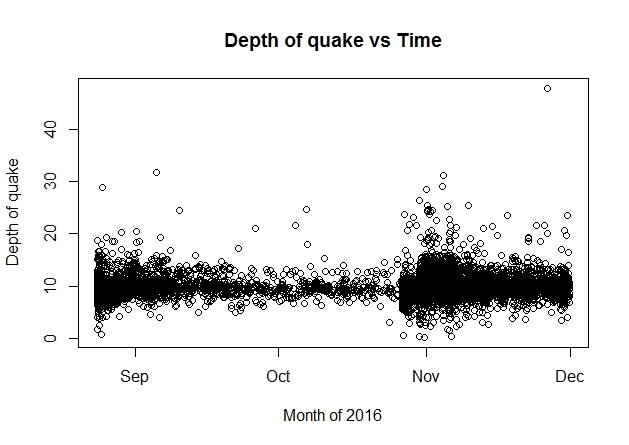
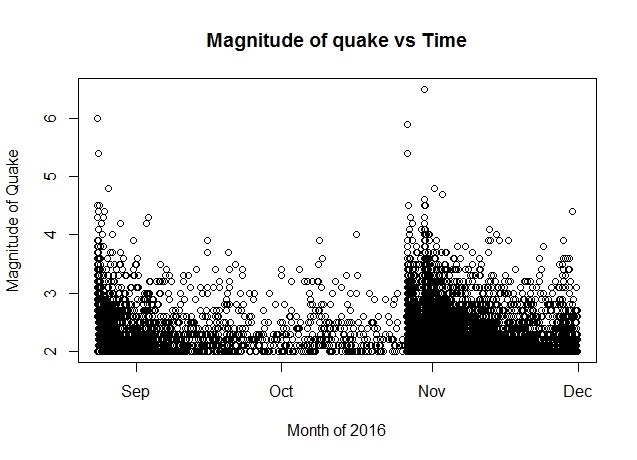
> plot(quake$Time, quake$Magnitude, xlab = "Month of 2016", ylab = "Magnitude of Quake", main = "Magnitude of quake vs Time")

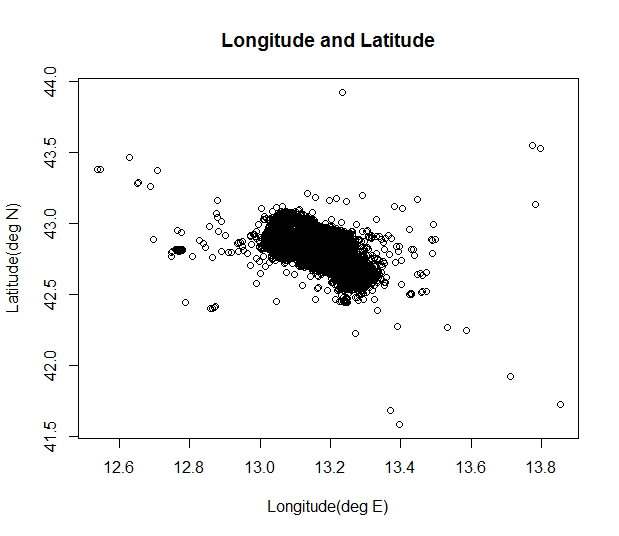
> plot(quake$Time, quake$`Depth/Km`, xlab = "Month of 2016", ylab = "Depth of quake", main = "Depth of quake vs Time")

> max(quake$Magnitude)

[1] 6.5

> plot(quake$Longitude, quake$Latitude, xlab = "Longitude(deg E)", ylab = "Latitude(deg N)", main = "Longitude and Latitude")





Five point summary –

We conduct the five number summary for the variables Depth, and Magnitude, as they are the measured variables of interest in our analysis.

**Min Q1 Q2 Q3 Max**

**Depth(km)** – 0.2 8.9 9.9 10.8 47.9

**Magnitude** – 2.0 2.1 2.3 2.6 6.5

Code Snippet

> #five number summaries

> fnum\_depth <- fivenum(quake$`Depth/Km`)

> fnum\_magnitude <- fivenum(quake$Magnitude)

> print(fnum\_depth)

[1] 0.2 8.9 9.9 10.8 47.9

> print(fnum\_magnitude)

[1] 2.0 2.1 2.3 2.6 6.5

Box plots –

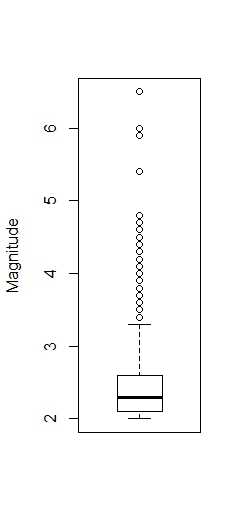
The box plots are constructed for Depth and Magnitude respectively as follows. The entire data is used for these plots.

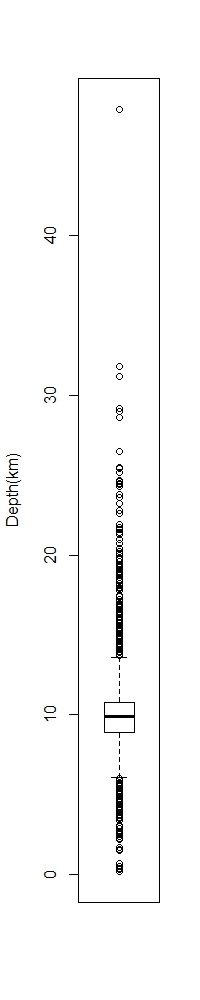
Code Snippet

> #boxplots, entire data

> boxplot(quake$`Depth/Km`, ylab = "Depth(km)")

> boxplot(quake$Magnitude, ylab = "Magnitude")





IQR, Finding outliers and cleaning data

1. We find the IQR as follows-
   1. Depth(km) – 1.9
   2. Magnitude – 0.5
2. Using this IQR we define outliers as follows
   1. For observations beyond 1.5 times IQR on either side of Q1 or Q3, called ‘minor’ outliers.
   2. For observations beyond 3 times IQR on either side of Q1 or Q3, called ‘major’ outliers.
3. For depth
   1. Cut-off for minor outliers – 6.05km below, and 13.65km above.
   2. Cut-off for major outliers – 3.2km below, and 16.5km above.
4. For magnitude
   1. Cut-off for minor outliers – 1.35 below, and 3.35 above.
   2. Cut-off for major outliers – 0.6 below, and 4.1 above.
5. We take a look at how many observations are excluded
   1. For depth
      1. Minor outliers – 502 rows
      2. Major outliers – 116 rows

We note that 502 is about 6.21% of the total number of observations, and so many outliers are unexpected. Thus for depth, we will consider only major outliers, which form 1.43% of the observations.

* 1. For magnitude
     1. Minor outliers – 333 rows
     2. Major outliers – 30 rows

We note that 333 is about 4.11% of the total number of observations and major outliers form 0.37% of the observations. We can investigate by excluding both types of outliers.

1. We consider 4 refined data frames
   1. Only major outliers of Depth variable removed
   2. Only minor outliers of Magnitude variable removed
   3. Only major outliers of Magnitude variable removed
   4. Minor outliers of Magnitude, and major outliers of Depth removed.

Code Snippet

> #quartiles

> q1\_depth <- fnum\_depth[2]

> q2\_depth <- fnum\_depth[3]

> q3\_depth <- fnum\_depth[4]

> q1\_magnitude <- fnum\_magnitude[2]

> q2\_magnitude <- fnum\_magnitude[3]

> q3\_magnitude <- fnum\_magnitude[4]

> #IQR

> iqr\_depth <- IQR(quake$`Depth/Km`)

> iqr\_magnitude <- IQR(quake$Magnitude)

> print(iqr\_depth)

[1] 1.9

> print(iqr\_magnitude)

[1] 0.5

Code Snippet

> #cutoff values for outliers

> min\_cutoff\_depth\_minor <- q1\_depth - 1.5 \* iqr\_depth

> max\_cutoff\_depth\_minor <- q3\_depth + 1.5 \* iqr\_depth

> min\_cutoff\_depth\_major <- q1\_depth - 3 \* iqr\_depth

> max\_cutoff\_depth\_major <- q3\_depth + 3 \* iqr\_depth

>

> print(min\_cutoff\_depth\_minor)

[1] 6.05

> print(max\_cutoff\_depth\_minor)

[1] 13.65

> print(min\_cutoff\_depth\_major)

[1] 3.2

> print(max\_cutoff\_depth\_major)

[1] 16.5

> min\_cutoff\_magnitude\_minor <- q1\_magnitude - 1.5 \* iqr\_magnitude

> max\_cutoff\_magnitude\_minor <- q3\_magnitude + 1.5 \* iqr\_magnitude

> min\_cutoff\_magnitude\_major <- q1\_magnitude - 3 \* iqr\_magnitude

> max\_cutoff\_magnitude\_major <- q3\_magnitude + 3 \* iqr\_magnitude

>

> print(min\_cutoff\_magnitude\_minor)

[1] 1.35

> print(max\_cutoff\_magnitude\_minor)

[1] 3.35

> print(min\_cutoff\_magnitude\_major)

[1] 0.6

> print(max\_cutoff\_magnitude\_major)

[1] 4.1

|  |
| --- |
|  |
|  |
|  |

Code Snippet

> #number of rows excluded

> nrow(quake[quake$`Depth/Km` < min\_cutoff\_depth\_minor | quake$`Depth/Km` > max\_cutoff\_depth\_minor, ])

[1] 502

> nrow(quake[quake$`Depth/Km` < min\_cutoff\_depth\_major | quake$`Depth/Km` > max\_cutoff\_depth\_major, ])

[1] 116

> nrow(quake[quake$Magnitude < min\_cutoff\_magnitude\_minor | quake$Magnitude > max\_cutoff\_magnitude\_minor, ])

[1] 333

> nrow(quake[quake$Magnitude < min\_cutoff\_magnitude\_major | quake$Magnitude > max\_cutoff\_magnitude\_major, ])

[1] 30

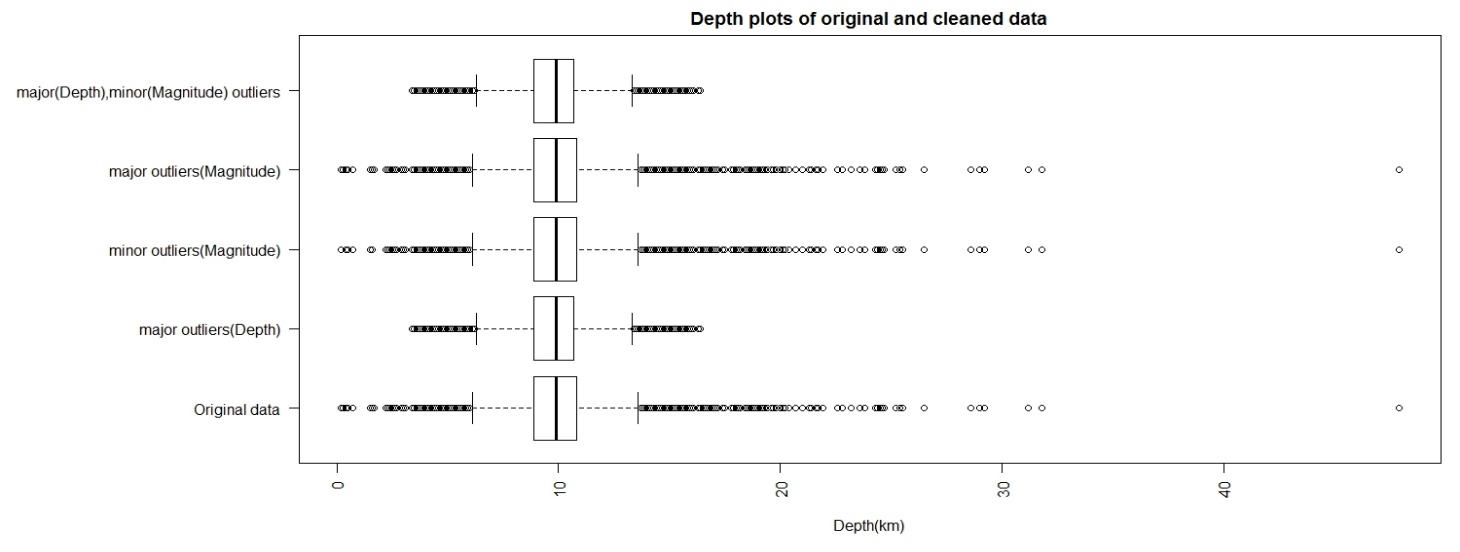
> quake\_depth\_major <- quake[quake$`Depth/Km` > min\_cutoff\_depth\_major & quake$`Depth/Km` < max\_cutoff\_depth\_major, ]

> quake\_mag\_minor <- quake[quake$Magnitude > min\_cutoff\_magnitude\_minor & quake$Magnitude < max\_cutoff\_magnitude\_minor, ]

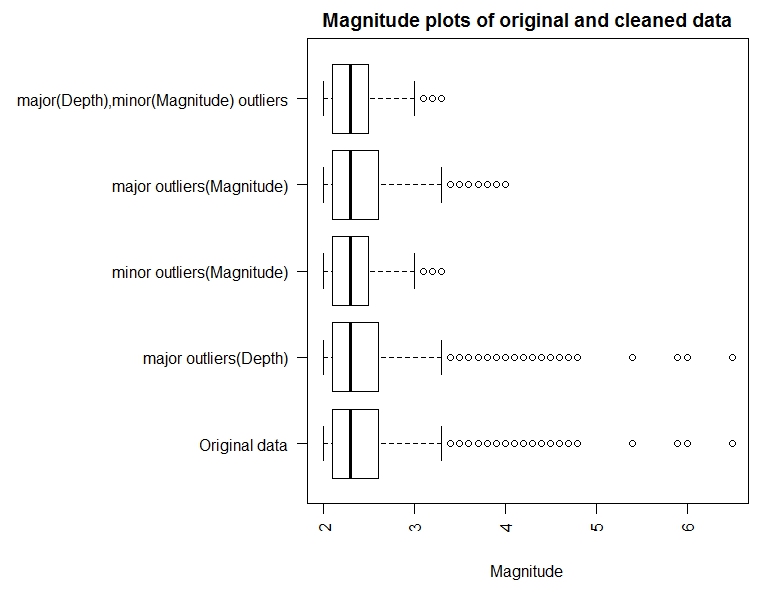
> quake\_mag\_major <- quake[quake$Magnitude > min\_cutoff\_magnitude\_major & quake$Magnitude < max\_cutoff\_magnitude\_major, ]

> quake\_mag\_minor\_depth\_major <- quake[quake$`Depth/Km` > min\_cutoff\_depth\_major & quake$`Depth/Km` < max\_cutoff\_depth\_major & quake$Magnitude > min\_cutoff\_magnitude\_minor & quake$Magnitude < max\_cutoff\_magnitude\_minor, ]

Boxplots with cleaned data –

Depth –

Magnitude –



Comparison of box plots –

1. Depth –

The large number of observations, and spread of data, makes the boxplots look similar, except for the outliers, which are removed in the cleaned data set.

1. Magnitude –

After removal of outliers, median does not appear to shift much, however the position of the third quartile has changed, along with the length of the right whisker, as compared to the original data set.

Conclusion –

1. The data was browsed and certain possible areas of discrepancies were pointed out
2. Five – number summary was found for the data, and boxplots obtained.
3. It was observed that we cannot afford to remove all minor outliers from ‘Depth’ variable, so only major outliers are filtered out.
4. Further, we obtain data sets by removing first the minor, and then the major outliers of the Magnitude variables.
5. We draw boxplots for cleaned datasets, and compare with original data.