earthquake

HImanshu Pareek

August 27, 2017

## Box plot of EARTHQUAKE data

loading data into R. Filename is "EARTHQUAKE.xlsx"

filename <- "EARTHQUAKE.xlsx"  
  
library(readxl)

## Warning: package 'readxl' was built under R version 3.3.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.3.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

earthquakes <- read\_excel(filename, na = c("Not Available", "not available", "NOT AVAILABLE", "na", "NA", "-", "\_", "null", "NULL"))  
head (earthquakes)

## # A tibble: 6 x 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

number of discrepancy is number of rows with na values. This can be calculated by is.na() function

dataOfNAValues <- earthquakes[rowSums(is.na(earthquakes)) > 0,]  
head (dataOfNAValues)

## # A tibble: 0 x 5  
## # ... with 5 variables: Time <dttm>, Latitude <dbl>, Longitude <dbl>,  
## # Depth/Km <dbl>, Magnitude <dbl>

Here, dataOfNAValues has 0 rows. It means, there is not any discrepancy in our cata. If we somehow forgot to keep names of possible na strings while reading excel file, we have to use omit function to remove all na or not suitable values.

earthquakes <- na.omit(earthquakes)  
head (earthquakes)

## # A tibble: 6 x 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

We will rename columns to ease of writing

colNames <- c("time", "lat", "long", "depth", "mag")  
colnames(earthquakes) <- colNames  
head (earthquakes)

## # A tibble: 6 x 5  
## time lat long depth mag  
## <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

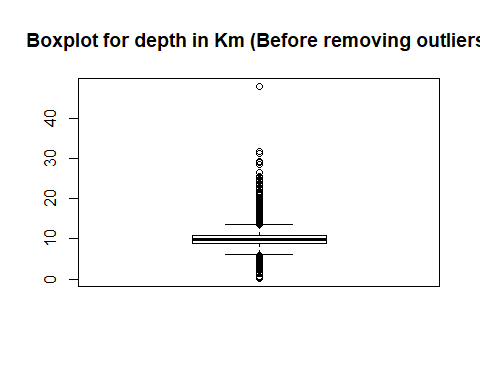
Subsetting depth and magnitude columns.

new\_data <- subset(earthquakes, select = c(depth, mag))  
head (new\_data)

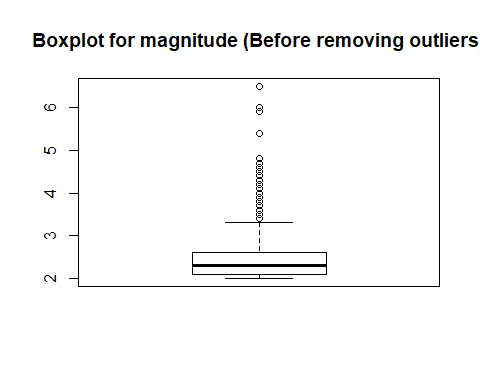
## # A tibble: 6 x 2  
## depth mag  
## <dbl> <dbl>  
## 1 8.1 6.0  
## 2 9.0 4.5  
## 3 9.7 3.8  
## 4 9.7 3.9  
## 5 9.7 3.6  
## 6 10.7 3.4

Boxplot before removing outliers

boxplot(new\_data$depth, main = "Boxplot for depth in Km (Before removing outliers)")



boxplot(new\_data$mag, main = "Boxplot for magnitude (Before removing outliers)")



Calculating 5 point summary

quantileDepth <- quantile(new\_data$depth, probs = c(0, 0.25, 0.50, 0.75, 1.0))  
quantileMag <- quantile(new\_data$mag, probs = c(0, 0.25, 0.50, 0.75, 1.0))  
quantileDepth

## 0% 25% 50% 75% 100%   
## 0.2 8.9 9.9 10.8 47.9

quantileMag

## 0% 25% 50% 75% 100%   
## 2.0 2.1 2.3 2.6 6.5

Calculating minimum and maximum of depth and magnitude to be used.

q1Depth <- quantileDepth[2]  
q3Depth <- quantileDepth[4]  
q1Mag <- quantileMag[2]  
q3Mag <- quantileMag[4]  
  
IQRDepth <- q3Depth - q1Depth  
IQRMag <- q3Mag - q1Mag  
  
minDepth <- q1Depth - (1.5 \* IQRDepth)  
maxDepth <- q3Depth + (1.5 \* IQRDepth)  
minMag <- q1Mag - (1.5 \* IQRMag)  
maxMag <- q3Mag + (1.5 \* IQRMag)  
  
sprintf("Range of %s should be %f - %f\n", c("Depth", "Magnitude"), c(minDepth, minMag), c(maxDepth, maxMag))

## [1] "Range of Depth should be 6.050000 - 13.650000\n"   
## [2] "Range of Magnitude should be 1.350000 - 3.350000\n"

Now, we will remove data below minimum limit and above maximum limit.

new\_data <- filter(new\_data, depth >= minDepth & depth <= maxDepth & mag >= minMag & mag <= maxMag)

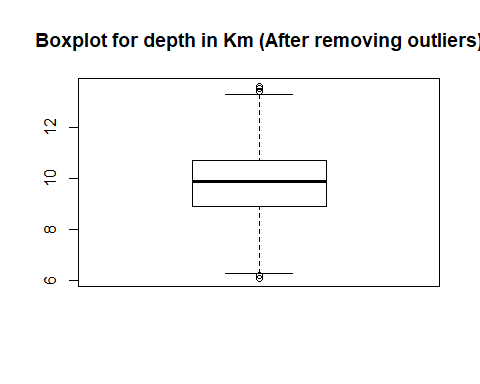
## Warning: package 'bindrcpp' was built under R version 3.3.3

Boxplot after removing outliers

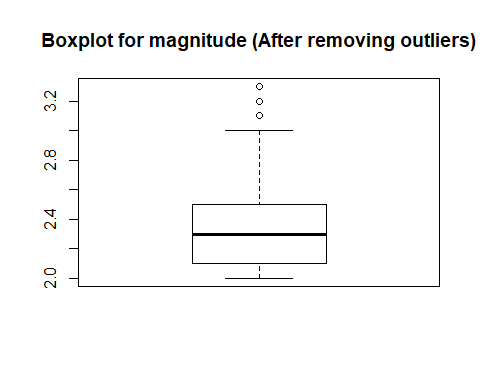
sprintf("After removing outliers")

## [1] "After removing outliers"

boxplot(new\_data$depth, main = "Boxplot for depth in Km (After removing outliers)")



boxplot(new\_data$mag, main = "Boxplot for magnitude (After removing outliers)")



Calculating 5 point summary

quantileDepth <- quantile(new\_data$depth, probs = c(0, 0.25, 0.50, 0.75, 1.0))  
quantileMag <- quantile(new\_data$mag, probs = c(0, 0.25, 0.50, 0.75, 1.0))  
quantileDepth

## 0% 25% 50% 75% 100%   
## 6.1 8.9 9.9 10.7 13.6

quantileMag

## 0% 25% 50% 75% 100%   
## 2.0 2.1 2.3 2.5 3.3

We can see that after removing outliers, boxplot is much more symmetric and we don't have any data in outliers so that we can accept all the data in new\_data.