Data Exploration and Data Visualisation

Siti Khalidah

6/7/2021

## Data Exploration

data <- read.csv("Unemployment\_Rate\_Clean.csv")  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## 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

#structure of dataset  
str(data)

## 'data.frame': 17 obs. of 5 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ State: chr "Johor" "Kedah" "Kelantan" "Malaysia" ...  
## $ X2017: num 3.4 2.8 3.6 3.4 1 2.8 2.9 3.7 3.5 2.1 ...  
## $ X2018: num 3 2.9 4 3.3 1.1 3 2.6 3.3 3.4 2.2 ...  
## $ X2019: num 2.7 2.8 4 3.3 1.1 3.1 2.9 3.4 3.3 2 ...

#change State as character  
data$State <- as.character(data$State)  
data$State

## [1] "Johor" "Kedah" "Kelantan"   
## [4] "Malaysia" "Melaka" "Negeri Sembilan"   
## [7] "Pahang" "Perak" "Perlis"   
## [10] "Pulau Pinang" "Sabah" "Sarawak"   
## [13] "Selangor" "Terengganu" "W.P Labuan"   
## [16] "W.P Putrajaya" "W.P. Kuala Lumpur"

#descriptive summary of dataset  
summary(data)

## X State X2017 X2018   
## Min. : 1 Length:17 Min. :1.000 Min. :1.100   
## 1st Qu.: 5 Class :character 1st Qu.:2.800 1st Qu.:2.600   
## Median : 9 Mode :character Median :3.100 Median :3.000   
## Mean : 9 Mean :3.424 Mean :3.306   
## 3rd Qu.:13 3rd Qu.:3.600 3rd Qu.:3.400   
## Max. :17 Max. :8.500 Max. :7.100   
## X2019   
## Min. :1.100   
## 1st Qu.:2.700   
## Median :3.100   
## Mean :3.212   
## 3rd Qu.:3.400   
## Max. :6.900

#names of variables in dataset  
names(data)

## [1] "X" "State" "X2017" "X2018" "X2019"

#type of dataset  
typeof(data)

## [1] "list"

typeof(data$X2017)

## [1] "double"

typeof(data$X2018)

## [1] "double"

typeof(data$X2019)

## [1] "double"

#number of rows in dataset  
nrow(data)

## [1] 17

#number of columns in dataset  
ncol(data)

## [1] 5

head(data)

## X State X2017 X2018 X2019  
## 1 1 Johor 3.4 3.0 2.7  
## 2 2 Kedah 2.8 2.9 2.8  
## 3 3 Kelantan 3.6 4.0 4.0  
## 4 4 Malaysia 3.4 3.3 3.3  
## 5 5 Melaka 1.0 1.1 1.1  
## 6 6 Negeri Sembilan 2.8 3.0 3.1

tail(data)

## X State X2017 X2018 X2019  
## 12 12 Sarawak 3.0 3.2 3.1  
## 13 13 Selangor 2.8 2.8 2.9  
## 14 14 Terengganu 4.5 4.8 3.5  
## 15 15 W.P Labuan 8.5 7.1 6.9  
## 16 16 W.P Putrajaya 1.5 1.3 1.2  
## 17 17 W.P. Kuala Lumpur 3.1 2.4 2.6

#detect any missing values in dataset  
is.na(data)

## X State X2017 X2018 X2019  
## [1,] FALSE FALSE FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE FALSE FALSE  
## [3,] FALSE FALSE FALSE FALSE FALSE  
## [4,] FALSE FALSE FALSE FALSE FALSE  
## [5,] FALSE FALSE FALSE FALSE FALSE  
## [6,] FALSE FALSE FALSE FALSE FALSE  
## [7,] FALSE FALSE FALSE FALSE FALSE  
## [8,] FALSE FALSE FALSE FALSE FALSE  
## [9,] FALSE FALSE FALSE FALSE FALSE  
## [10,] FALSE FALSE FALSE FALSE FALSE  
## [11,] FALSE FALSE FALSE FALSE FALSE  
## [12,] FALSE FALSE FALSE FALSE FALSE  
## [13,] FALSE FALSE FALSE FALSE FALSE  
## [14,] FALSE FALSE FALSE FALSE FALSE  
## [15,] FALSE FALSE FALSE FALSE FALSE  
## [16,] FALSE FALSE FALSE FALSE FALSE  
## [17,] FALSE FALSE FALSE FALSE FALSE

#observe the number of missing values  
sum(is.na(data))

## [1] 0

sum(is.na(data$X2017))

## [1] 0

sum(is.na(data$X2018))

## [1] 0

sum(is.na(data$X2019))

## [1] 0

#calculate the mean of dataset  
mean(data$X2017, na.rm = TRUE)

## [1] 3.423529

mean(data$X2018, na.rm = TRUE)

## [1] 3.305882

mean(data$X2019, na.rm = TRUE)

## [1] 3.211765

#add mean dataset to data frame  
data2.1 <- select(data, c(X2017, X2018, X2019))  
data2.2 <- mutate(data2.1,Mean = rowMeans(data2.1))  
data2.3 <- select(data, c(X,State))  
data2 <- bind\_cols  
  
#calculate the median of dataset  
median(data$X2017, na.rm = TRUE)

## [1] 3.1

median(data$X2018, na.rm = TRUE)

## [1] 3

median(data$X2019, na.rm = TRUE)

## [1] 3.1

#calculate the variance of dataset  
var(data$X2017, na.rm = TRUE)

## [1] 2.801912

var(data$X2018, na.rm = TRUE)

## [1] 2.211838

var(data$X2019, na.rm = TRUE)

## [1] 2.003603

#calculate the standard deviation of dataset  
sd(data$X2017, na.rm = TRUE)

## [1] 1.673891

sd(data$X2018, na.rm = TRUE)

## [1] 1.487225

sd(data$X2019, na.rm = TRUE)

## [1] 1.415487

#maximum value in each year (2017-2019)  
max(data$X2017, na.rm = TRUE)

## [1] 8.5

max(data$X2018, na.rm = TRUE)

## [1] 7.1

max(data$X2019, na.rm = TRUE)

## [1] 6.9

#minimum value in each year (2017-2019)  
min(data$X2017, na.rm = TRUE)

## [1] 1

min(data$X2018, na.rm = TRUE)

## [1] 1.1

min(data$X2019, na.rm = TRUE)

## [1] 1.1

#sort the value in each year (2017-2019)  
sort(data$X2017)

## [1] 1.0 1.5 2.1 2.8 2.8 2.8 2.9 3.0 3.1 3.4 3.4 3.5 3.6 3.7 4.5 5.6 8.5

sort(data$X2018)

## [1] 1.1 1.3 2.2 2.4 2.6 2.8 2.9 3.0 3.0 3.2 3.3 3.3 3.4 4.0 4.8 5.8 7.1

sort(data$X2019)

## [1] 1.1 1.2 2.0 2.6 2.7 2.8 2.9 2.9 3.1 3.1 3.3 3.3 3.4 3.5 4.0 5.8 6.9

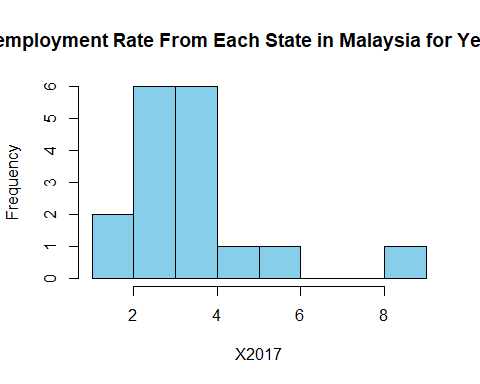
## Data Visualisation

## Histogram

data3 <- select(data, c(X, State))  
data3

## X State  
## 1 1 Johor  
## 2 2 Kedah  
## 3 3 Kelantan  
## 4 4 Malaysia  
## 5 5 Melaka  
## 6 6 Negeri Sembilan  
## 7 7 Pahang  
## 8 8 Perak  
## 9 9 Perlis  
## 10 10 Pulau Pinang  
## 11 11 Sabah  
## 12 12 Sarawak  
## 13 13 Selangor  
## 14 14 Terengganu  
## 15 15 W.P Labuan  
## 16 16 W.P Putrajaya  
## 17 17 W.P. Kuala Lumpur

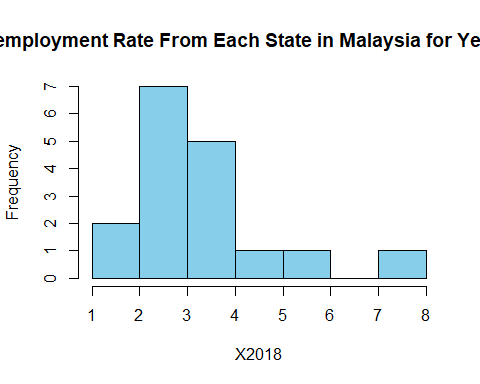
#Histogram  
attach(data)  
hist(X2017,main="Unemployment Rate From Each State in Malaysia for Year 2017",col="skyblue")



attach(data)

## The following objects are masked from data (pos = 3):  
##   
## State, X, X2017, X2018, X2019

hist(X2018,main="Unemployment Rate From Each State in Malaysia for Year 2018",col="skyblue")

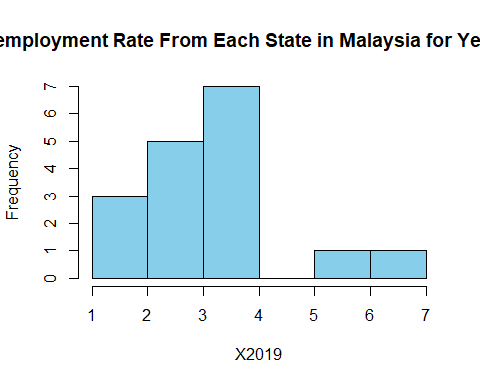


attach(data)

## The following objects are masked from data (pos = 3):  
##   
## State, X, X2017, X2018, X2019

## The following objects are masked from data (pos = 4):  
##   
## State, X, X2017, X2018, X2019

hist(X2019,main="Unemployment Rate From Each State in Malaysia for Year 2019",col="skyblue")

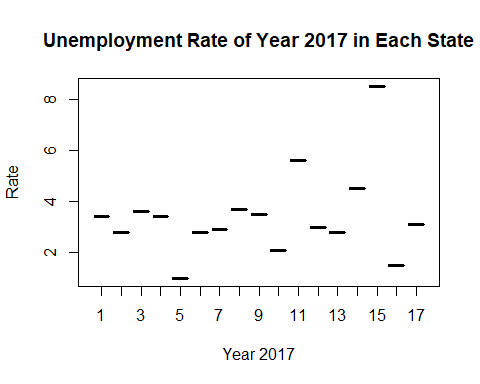


## Boxplot

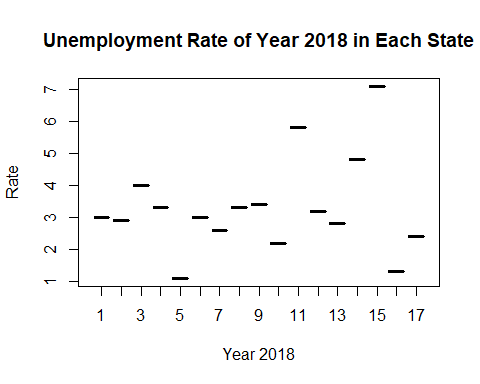
#Boxplot  
data3 <- select(data, c(X, State))  
data3

## X State  
## 1 1 Johor  
## 2 2 Kedah  
## 3 3 Kelantan  
## 4 4 Malaysia  
## 5 5 Melaka  
## 6 6 Negeri Sembilan  
## 7 7 Pahang  
## 8 8 Perak  
## 9 9 Perlis  
## 10 10 Pulau Pinang  
## 11 11 Sabah  
## 12 12 Sarawak  
## 13 13 Selangor  
## 14 14 Terengganu  
## 15 15 W.P Labuan  
## 16 16 W.P Putrajaya  
## 17 17 W.P. Kuala Lumpur

#Boxplot  
boxplot(X2017 ~ X, data = data3, xlab = "Year 2017",  
 ylab = "Rate", main = "Unemployment Rate of Year 2017 in Each State")



boxplot(X2018 ~ X, data = data3, xlab = "Year 2018",  
 ylab = "Rate", main = "Unemployment Rate of Year 2018 in Each State")



boxplot(X2019 ~ X, data = data3, xlab = "Year 2019",  
 ylab = "Rate", main = "Unemployment Rate of Year 2019 in Each State")

