```
library(XLS)
library(xlsx)
library(readxl)

# PART 1a
# generate central tendency values (mean median mode) for air_e and water_e. also boxplots

epi <- read_excel('EPI2010_data.xls', sheet="EPI2010_onlyEPIcountries")
epi_data <- data.frame(epi)

getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# air_e central tendency values:
AIR_E <- as.numeric(epi_data$AIR_E)
mean(AIR_E)
median(AIR_E)
getmode(AIR_E)

# water_e central tendency values:
WATER_E <- as.numeric(epi_data$WATER_E)
mean(WATER_E)
median(WATER_E)
getmode(WATER_E)

boxplot(AIR_E,WATER_E, names=c("AIR_E","WATER_E"))

# central tendency for NOX_pt and SO2_pt
NOX_PT <- as.numeric(epi_data$NOX_pt)
mean(NOX_PT)
median(NOX_PT)
getmode(NOX_PT)

SO2_PT <- as.numeric(epi_data$SO2_pt)
mean(SO2_PT)
median(SO2_PT)
getmode(SO2_PT)

# box for OZONE_pt and WQI_pt
OZONE_PT <- as.numeric(epi_data$OZONE_pt)
WQI_PT <- as.numeric(epi_data$WQI_pt)
boxplot(OZONE_PT, WQI_PT, names=c("OZONE_PT","WQI_PT"))
```

```
# central tendency for climate and agriculture
CLIM <- as.numeric(epi_data$CLIMATE)
mean(CLIM)
median(CLIM)
getmode(CLIM)

AGRI <- as.numeric(epi_data$AGRICULTURE)
mean(AGRI)
median(AGRI)
getmode(AGRI)

# box for FISHERIES and NMVOC_pt
FISH <- as.numeric(epi_data$FISHERIES)
tf <- is.na(FISH)
FISH <- FISH[!tf]   #filter out NA values
NM <- as.numeric(epi_data$NMVOC_pt)
boxplot(FISH, NM, names=c("FISHERIES","NMVOC_pt"))


ENV <- as.numeric(epi_data$ENVHEALTH)
ECO <- as.numeric(epi_data$ECOSYSTEM)
boxplot(ENV, ECO, names=c("ENVHEALTH","ECOSYSTEM"))
qqplot(ENV, ECO)

# 1b

EPI_data <- read.csv("EPI_data.csv")
SA_data <-EPI_data[EPI_data$EPI_regions=='South Asia',]
SA_data <-SA_data[SA_data$Country!='Afghanistan',]

View(SA_data)

SA_EPI <- as.numeric(SA_data$EPI)
tf <- is.na(SA_EPI)
SA_EPI <- SA_EPI[!tf]   #filter out NA values

SA_ENVHEALTH <- as.numeric(SA_data$ENVHEALTH)
tf <- is.na(SA_ENVHEALTH)
SA_ENVHEALTH <- SA_ENVHEALTH[!tf]   #filter out NA values

SA_ECOSYSTEM <- as.numeric(SA_data$ECOSYSTEM)
tf <- is.na(SA_ECOSYSTEM)
SA_ECOSYSTEM <- SA_ECOSYSTEM[!tf]   #filter out NA values
```

```
SA_DALY <- as.numeric(SA_data$DALY)
tf <- is.na(SA_DALY)
SA_DALY <- SA_DALY[!tf]   #filter out NA values

SA_AIR_H <- as.numeric(SA_data$AIR_H)
tf <- is.na(SA_AIR_H)
SA_AIR_H <- SA_AIR_H[!tf]   #filter out NA values

SA_WATER_H <- as.numeric(SA_data$WATER_H)
tf <- is.na(SA_WATER_H)
SA_WATER_H <- SA_WATER_H[!tf]   #filter out NA values

SA_AIR_E <- as.numeric(SA_data$AIR_E)
tf <- is.na(SA_AIR_E)
SA_AIR_E <- SA_AIR_E[!tf]   #filter out NA values

SA_WATER_E <- as.numeric(SA_data$WATER_E)
tf <- is.na(SA_WATER_E)
SA_WATER_E <- SA_WATER_E[!tf]   #filter out NA values

SA_BIODIVERSITY <- as.numeric(SA_data$BIODIVERSITY)
tf <- is.na(SA_BIODIVERSITY)
SA_BIODIVERSITY <- SA_BIODIVERSITY[!tf]   #filter out NA values

SA_FORESTRY <- as.numeric(SA_data$FORESTRY)
tf <- is.na(SA_FORESTRY)
SA_FORESTRY <- SA_FORESTRY[!tf]   #filter out NA values

SA_CLIMATE <- as.numeric(SA_data$CLIMATE)
tf <- is.na(SA_CLIMATE)
SA_CLIMATE <- SA_CLIMATE[!tf]   #filter out NA values

# DALY, AIR_H, WATER_H, AIR_E, WATER_E, BIODIVERSITY, FORESTRY, AGRICULTURE,
CLIMATE

cor(SA_EPI, SA_ENVHEALTH)
cor(SA_EPI, SA_ECOSYSTEM)
cor(SA_EPI, SA_DALY)
cor(SA_EPI, SA_AIR_H)
cor(SA_EPI, SA_WATER_H)
cor(SA_EPI, SA_AIR_E)
cor(SA_EPI, SA_WATER_E)
cor(SA_EPI, SA_BIODIVERSITY)
```

```
cor(SA_EPI, SA_FORESTRY)
cor(SA_EPI, SA_CLIMATE)

# The most important variable for EPI is the ECOSYSTEM variable for South Asia, because it has the highest correlation

boxplot(ENVHEALTH,DALY,AIR_H,WATER_H, names=c("ENVHEALTH", "DALY", "AIR_H", "WATER_H"))

lmENVH <- lm(ENVHEALTH~DALY+AIR_H+WATER_H)

lmENVH

summary(lmENVH)

cENVH <- coef(lmENVH)

#Predict
DALYNEW <- c(seq(5,95,5))
AIR_HNEW <- c(seq(5,95,5))
WATER_HNEW <- c(seq(5,95,5))

NEW <- data.frame(DALYNEW,AIR_HNEW,WATER_HNEW)
View(NEW)

pENV <- predict(lmENVH,NEW,interval="prediction")

#'newdata' had 19 rows but variables found have 163 rows. not sure what i'd be preducting here
# this is from slide 13

cENV <- predict(lmENVH,NEW,interval="confidence")

cENV

# using the response variable as: AIR_E
Model1 <- lm(AIR_E ~DALY+AIR_H+WATER_H)

summary(Model1)

# using the response variable as: CLIMATE
Model2 <- lm(CLIM ~DALY+AIR_H+WATER_H)
summary(Model2)

shapiro.test(ENVHEALTH)
```

```
shapiro.test(ECOSYSTEM)

# run shapiro wilik for:
#ENVHEALTH, DALY, AIR_H,WATER_H
# check with dim(), if over 5000 elements, get a 5000 length slice
# note how normal they are
# Repeat the same exercise using the EPI.csv (not 2010EPI.csv) for the same 4 variables.

View(ENVHEALTH)
shapiro.test(ENVHEALTH) # W = 0.92019, p-value = 8.179e-08, very likely normal

View(DALY)
shapiro.test(DALY) # W = 0.93784, p-value = 1.522e-06, very likely normal

View(AIR_H)
shapiro.test(AIR_H) # W = 0.92875, p-value = 3.204e-07, very likely normal

View(WATER_H)
shapiro.test(WATER_H) # W = 0.87183, p-value = 1.348e-10, likely normal

epi <- read.csv('EPI_Data.csv')
epi_data <- data.frame(epi)
View(epi_data)

ENVHEALTH <- as.numeric(epi_data$ENVHEALTH)
DALY <- as.numeric(epi_data$DALY)
AIR_H <- as.numeric(epi_data$AIR_H)
WATER_H <- as.numeric(epi_data$WATER_H)

View(ENVHEALTH)
shapiro.test(ENVHEALTH) # W = 0.91613, p-value = 1.083e-08, very likely normal

View(DALY)
shapiro.test(DALY) # W = 0.93654, p-value = 1.891e-07, very likely normal

View(AIR_H)
shapiro.test(AIR_H) # W = 0.92138, p-value = 8.994e-09, very likely normal

View(WATER_H)
shapiro.test(WATER_H) # W = 0.8597, p-value = 1.679e-12, likely normal
```