Andrew Updegrove
661410127

Lab 2

ITWS 6600

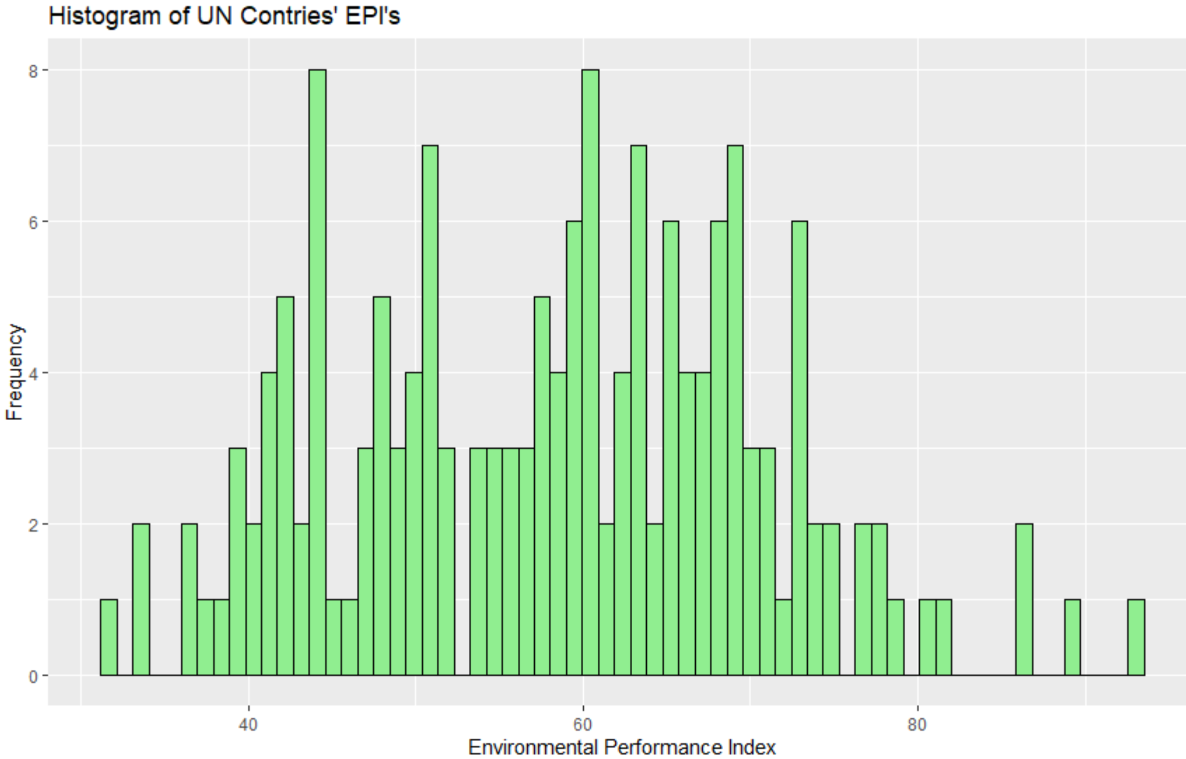
Part 1

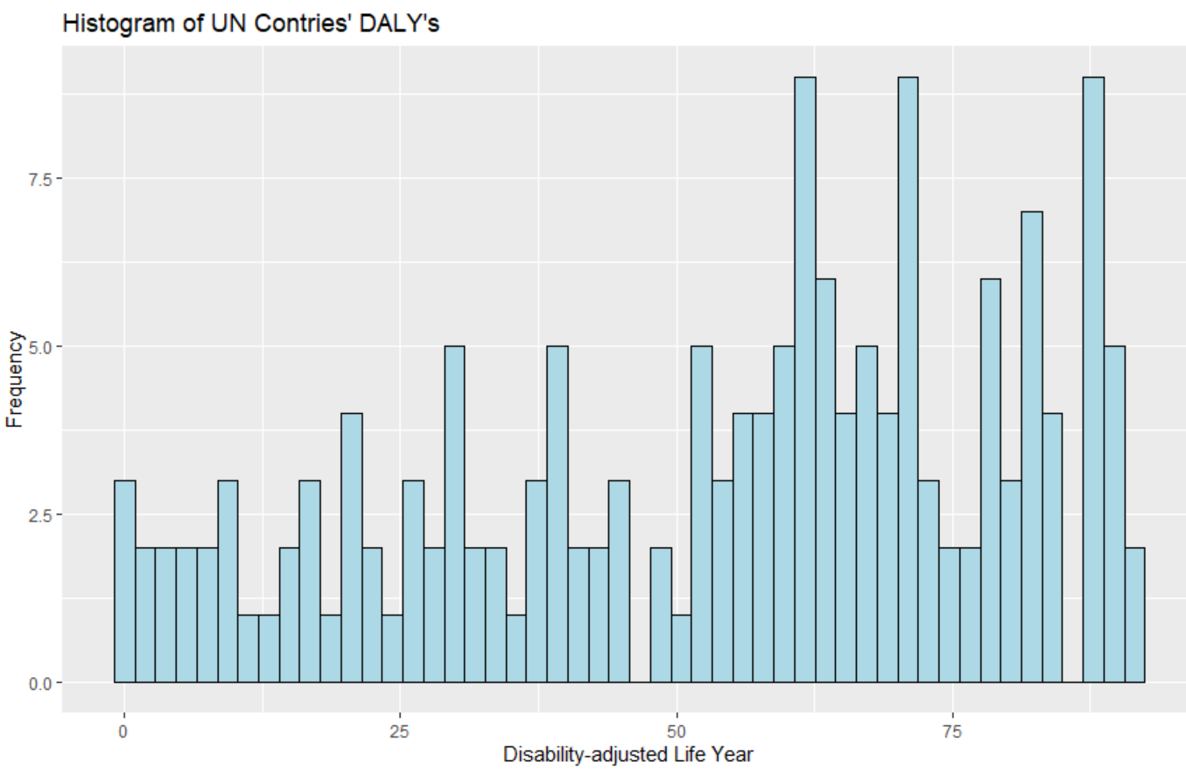
a. The code for this part

```
library(ggplot2)
#choose 2010EPI_data.csv
data1 <- read.csv(file.choose(), skip=1, header=T)
data_clean<-data1[rowSums(is.na(data1)) < "100",]

mode <- function(d){
  uni <- unique(d)
  uni[which.max(tabulate(match(d, uni)))]
}

epi_mean <- mean(data_clean$EPI, na.RM = T)
epi_median <- median(data_clean$EPI, na.RM = T)
epi_mode <- mode(data_clean$EPI)

daly_mean <- mean(data_clean$DALY, na.RM = T)
daly_median <- median(data_clean$DALY, nam.RM = T)
daly_mode <- mode(data_clean$DALY)

epi_hist <- ggplot(data_clean, aes(x=EPI)) + ...
geom_histogram(bins= 65, color="black", fill="lightgreen") + ...
labs(x = "Environmental Performance Index", y = "Frequency", ...
title="Histogram of UN Contries' EPI's")

daly_hist <- ggplot(data_clean, aes(x=DALY)) + ...
geom_histogram(bins= 50, color="black", fill="lightblue") + ...
labs(x = "Disability-adjusted Life Year", y = "Frequency", ...
title="Histogram of UN Contries' DALY's")

boxplot(data_clean$ENVHEALTH, data_clean$ECOSYSTEM, ...
names=c("Environmental Health", "Ecosystem Vitality"), ...
="UN countries Environment Indicators")

qqplot(data_clean$ENVHEALTH, data_clean$ECOSYSTEM, ...
xlab="Environmental Health", ylab="Ecosystem Vitality", ...
main="Q-Q plot for UN Countries")
```

	mean	median	mode
EPI	58.37	59.2	51.3
DALY	53.62	60.35	86.86

Table 1: Central Tendencies


Histogram of UN Contries' EPI's
Frequency
Environmental Performance Index
0
2
4
6
8
40
60
80


Histogram of UN Contries' DALY's
Frequency
Disability-adjusted Life Year
0.0
2.5
5.0
7.5
0
25
50
75

b. The most important factor for EPI in the European region is CLIMATE with a coefficient of .4407. The code to generate it looks like this. The code to predict the values for ENVHEALTH, AIR_E, and the CLIMATE variables were generated with this code below that.

```
#choose EPI data
library(dplyr)
EPI_data <- read.csv(file.choose(), header=T)
just_EU <- filter(EPI_data,EPI_data$EPI_regions == 'Europe')
eu_lm <- lm(EPI~DALY+AIR_H+WATER_H+AIR_E+WATER_E+...
BIODIVERSITY+FORESTRY+FISHERIES+AGRICULTURE+...
CLIMATE, data=just_EU)

eeu_coefs <- eu_lm$coefficients
eu_coefs[which.max(eu_coefs)]


boxplot(EPI_data$ENVHEALTH,EPI_data$DALY, ...
EPI_data$AIR_H, EPI_data$WATER_H)
lmENVH <- lm(ENVHEALTH~DALY+AIR_H+WATER_H, data= EPI_data)
summary(lmENVH)
cENVH <- coef(lmENVH)
DALYNEW <- c(seq(5,95,5))
AIR_HNEW <- c(seq(5,95,5))
WATER_HNEW <- c(seq(5,95,5))

#Predicting for the ENVHEALTH variable
NEW <- data.frame(DALY=DALYNEW, AIR_H=AIR_HNEW, WATER_H=WATER_HNEW)
pENV <- predict(lmENVH, newdata=NEW, interval="prediction")
cENV <- predict(lmENVH, newdata=NEW, interval="confidence")

#Predicting the Air_E variable
lmAIR_E <- lm(AIR_E~DALY+AIR_H+WATER_H, data= EPI_data)
pAIR <- predict(lmAIR_E, newdata=NEW, interval="prediction")
cAIR <- predict(lmAIR_E, newdata=NEW, interval="confidence")

#Predicting the CLIMATE variable
lmCLIMATE <- lm(CLIMATE~DALY+AIR_H+WATER_H, data= EPI_data)
pCLIMATE <- predict(lmCLIMATE, newdata=NEW, interval="prediction")
cCLIMATE <- predict(lmCLIMATE, newdata=NEW, interval="confidence")
```

Part 2

Exercise 1. The first predicted value for enrollment was 81437. The second predicted value, when including income, was 137452. The code was:

```
#dataset_multipleRegression.csv
theData <- read.csv(file.choose(),header=T)
```

```
summary(theData)
attach(theData)
lmROLL <- lm(ROLL ~ UNEM + HGRAD)
new.theData <- data.frame(UNEM = c(7),HGRAD = c(90000))
pROLL <- predict(lmROLL, newdata=new.theData, interval="prediction")
pROLL #Predicted value is 81437

#Again with per capita income
lmROLL2 <- lm(ROLL ~ UNEM + HGRAD + INC)
new.roll <- data.frame(UNEM = 7, HGRAD = 90000, INC = 25000)
pROLL2 <- predict(lmROLL2, newdata=new.roll)
pROLL2 #Predicted value is 137452
```

Exercise 2. The KNN classification was performed on the abalone dataset

```
# choose abalone.csv
library(class)


abalone <- read.csv(file.choose(), header=T)
head(abalone)
abalone$Rings <- as.numeric(abalone$Rings)
abalone$Rings <- cut(abalone$Rings, breaks=c(-1,8,11,35), ...
labels=c('young', 'adult', 'old'))

abalone$Rings <- as.factor(abalone$Rings)
summary(abalone$Rings)
aba <- abalone
aba$Sex <- NULL

hemin_max_normalize <- function(x) {
  return (x - min(x)) / (max(x) - min(x))
}

aba[1:7] <- as.data.frame(lapply(aba[1:7], min_max_normalize))
split_index <- sample(2, nrow(aba), replace=TRUE, prob=c(.7, .3))
training_samp <- aba[split_index == 1,]
testing_sampl <- aba[split_index == 2,]

round_odd <- function(x) {
  return (2*floor(x/2)+1)
}

kay = round_odd(sqrt(nrow(training_samp)))
KNNpred <- knn(train = training_samp[1:7], test = testing_sampl[1:7], ...
cl = training_samp$Rings, k = kay)

# finding accuracy
accuracy <- length(which(KNNpred == testing_sampl[,8])) / length(KNNpred)
# 69.4% accuracy for the KNN
```

Exercise 3. The output of the clustering was:

	1	2	3
setosa	50	0	0
versicolor	0	2	48
verginica	0	36	14

The code looks like:

```
library(ggplot2)
head(iris)
summary(iris)
data_iris = iris[,-5]

set.seed(12657)
k.max <- 12

k_cluster <- kmeans(data_iris,3,nstart=20,iter.max = 1000)

table(iris[,5], k_cluster$cluster)
```

Exercise 4. The outputs from some of the functions appear in the code snippet below.

```
library(dplyr)
#choose EPI_data.csv
EPI_data <- read.csv(file.choose(), header=T)
attach(EPI_data)
EPI_data %>% sample_n(5,replace= FALSE) %>% select(EPI,DALY)
# Output:
# EPI  DALY
# 1 55.3 64.40
# 2 47.9 18.16
# 3 69.2 80.96
# 4   NA 40.88
# 5 62.9 67.82

EPI_data %>% sample_frac(size = .1, replace = FALSE) %>% select(EPI, DALY)
# Output:
# EPI  DALY
# 1   41.0 20.31
# 2   58.8 59.41
# 3     NA    NA
# 4   78.2 82.81
# 5   44.0 39.85
# 6   39.4  1.35
# 7   67.0 64.40
# 8   69.4 69.04
# 9   76.8 63.34
# 10    NA    NA
```

```
# 11  63.5  65.50
# 12  37.6   0.00
# 13  42.3  51.99
# 14    NA  63.34
# 15  89.1  89.10
# 16  44.6   5.81
# 17    NA     NA
# 18  73.4  82.81
# 19  59.1  70.31
# 20    NA  44.18
# 21  60.6  69.04
# 22    NA     NA
# 23    NA  73.01

new_decs_EPI <- arrange(EPI_data, desc(EPI))
new_decs_DALY <- arrange(EPI_data, desc(EPI))

EPI_data <- mutate(EPI_data, double_EPI = EPI * 2, double_DALY = DALY * 2)

summarise(EPI_data, avg_EPI = mean(EPI, na.rm = TRUE), avg_DALY = ...
mean(EPI, na.rm = TRUE))
#Output:
#     avg_EPI avg_DALY
# 1 58.37055 58.37055
```