Introduction to Data Analysis and Mining
Homework #1

Due on Sunday, Jan 21, 2018 10:00 p.m.

Instructor: Hasan Kurban

Kwong Yuet Michael Fadillah Wong

January 21, 2018

Problem 1

Here is data from the *Digest of Education Statistics, 2005,Table 63.* – viewed in a plain text file called `teach.txt`. You are allowed to use R packages and built-in functions for this question.

Year, Ratio
1955, 26.9
1960, 25.8
1965, 24.7
1970, 22.3
1980, 18.7
1985, 17.9
1990, 17.2
1995, 117.3
2000, 16.0
2005, 15.5

Q1.1 Provide the R code that reads the data from teach.txt into an R data.frame?

R Script

```
teach <- read.table("/Users/Fadil/Desktop/school/B365/HW1/hw1/teach.txt", sep = ",", header
```

Q1.2 Suppose you're interested in looking at *only* the Ratios. Give R code that produces this data.

R script

```
RatioData <- teach[,c("Ratio")]
```

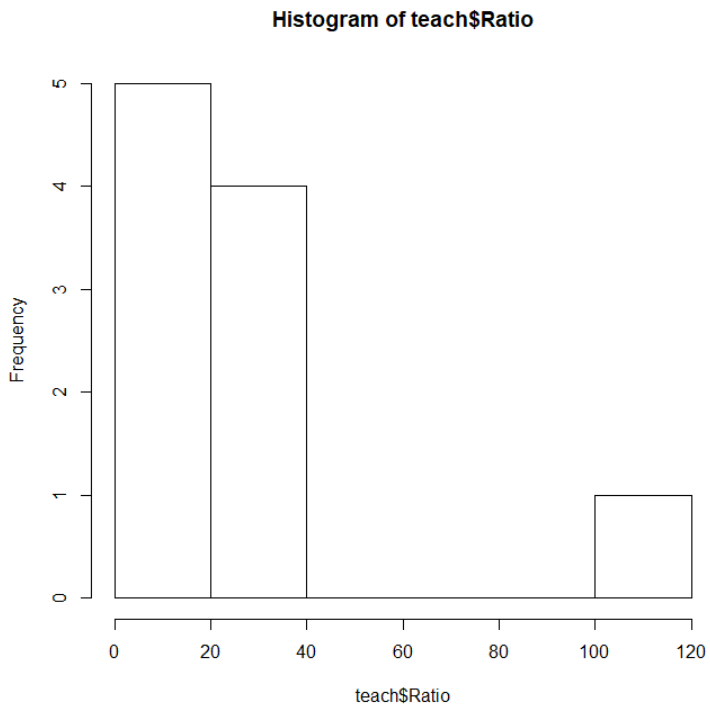
Q1.3 Give a select operation on the data.frame that gives the rows whose ratios are greater than 18, but less than 22. What does this yield?

R script

```
teach[teach$Ratio == (RatioData[18 < RatioData & RatioData < 22]), ]
```

Problem 1 continued on next page...

Q1.4 Here is the histogram plot of `teach$Ratio`



Give R code that produces this plot.

R script

```
hist(teach$Ratio)
```

Q1.5 Discuss the data including the histogram and this R code:

```
plot(Year, Ratio, type="l")
```

it produce a plot that list year as x axis, ratio as y axis. By this plot, it is easily seen 1955-2005's correspondingly. From the data, year 1995 has the highest ratio, while year 2005 has the lowest.

Problem 2

Load mydata.txt into R and answer the following questions. You are allowed to use R packages and built-in functions for this question.

Q2.1 How many entries are there in the data set? Answer here ...

R script

```
> length(myData$V1) + length(myData$V2) + length(myData$V3) + length(myData$V4) + length(myD
[1] 9910
```

Q2.2 Calculate mean and median of variable V2. Answer here ...

R script

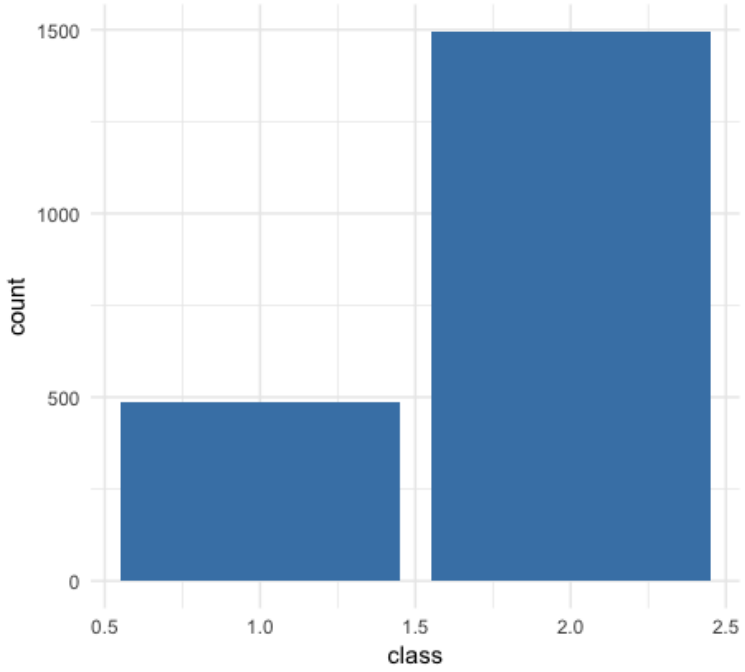
```
> median(myData$V2)
[1] -0.43816771
> mean(myData$V2)
[1] -0.9705022
```

Q2.3 Find variance and standard deviation of variable V1. Answer here ...

R script

```
> sd(myData$V2)
[1] 1.983679
> var(myData$V2)
[1] 3.934984
```

Q2.4 Variable 5, $V5$, is the class the variable and the bar plot below shows the distribution of data points among different classes. Give the R code that produces the below figure. (Color is not required to be the same)



R script

```
> dataplot <- barplot(table(myData$V5), xpd = FALSE, ylim = c(0, 1600), xaxt = 'n', xlab = "
> axis(1, seq(0.5, 2.5, by = 0.5))
```

Problem 3

Create an R function that calculates Euclidean distance between same dimensional two vectors (data points). Call this function dist.euclidean.R. Assume three pieces of data $x_1 = (1, 2); x_2 = (3, 4); x_3 = (6, 4)$ (x_1, x_2, x_3 are two dimensional data points). Using your R function, determine which two are the least dissimilar. Answer here ....

R script

```
dist.euclidean.R <- function(x,y){
  return(sqrt((x[1] - y[1])^2 + (x[2] - y[2])^2))
}
> dist.euclidean.R(c(1,2), c(3,4))
[1] 2.828427
> dist.euclidean.R(c(1,2), c(6,4))
[1] 5.385165
> dist.euclidean.R(c(3,4), c(6,4))
[1] 3
```

Problem 4

In this question, you are asked to implement two R functions to calculate mean and variance. Call this functions sample.mean.R and sample.variance.R. You're given a sample of data: 15,2,44,21,40,20,19,18. Calculate the sample mean and sample variance using your functions. Answer here...

R script

```
sample.variance.R <- function(x){
  return(sum((x-mean(x))^2/(length(x)-1)))
}
sample.mean.R <- function(x){
  return(sum(x)/length(x))
}
> sample.mean.R(c(15,2,44,21,40,20,19,18))
[1] 22.375
> sample.variance.R(c(15,2,44,21,40,20,19,18))
[1] 183.6964
```