CSCI-B 365 Introduction to Data Analysis and Mining
Homework 1
Computer Science
Spring 2018
Indiana University,
Bloomington, IN

Siyi Xian
siyixian@indiana.edu

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

All the work herein is solely mine.

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("teach.txt", header = TRUE, sep = ",")
data.frame(teach)
> table
   Year Ratio
1  1955  26.9
2  1960  25.8
3  1965  24.7
4  1970  22.3
5  1980  18.7
6  1985  17.9
7  1990  17.2
8  1995 117.3
9  2000  16.0
10 2005  15.5
```

Problem 1 continued on next page. . .

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

R script

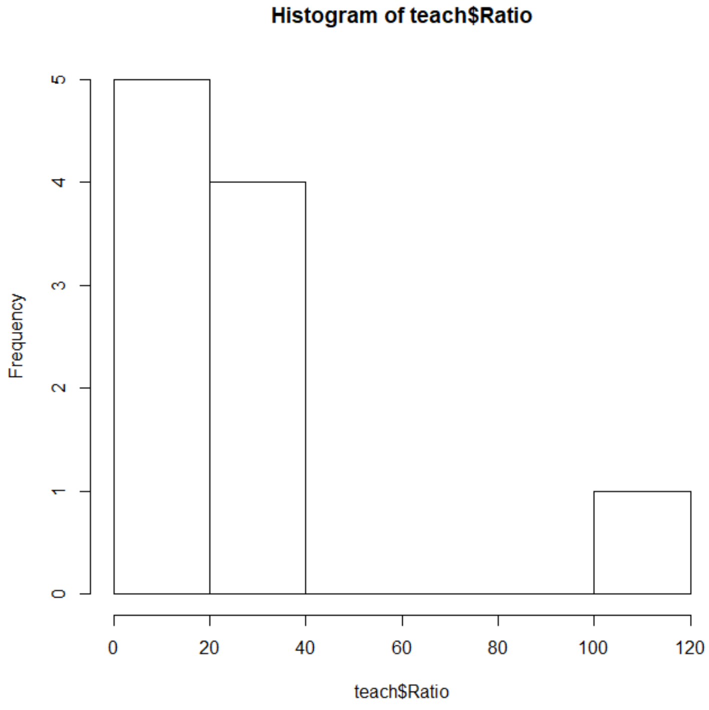
```
> teach[2]
   Ratio
1   26.9
2   25.8
3   24.7
4   22.3
5   18.7
6   17.9
7   17.2
8  117.3
9   16.0
10  15.5
```

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

```
> subset(teach, Ratio > 18 & Ratio < 22)
  Year Ratio
5 1980  18.7
```

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")
```

```
Histogram shows the frequency of the Ration value show up.
The image which show by the code shows the tend of change by years.
```

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

```
mydata <- read.table("mydata.txt", header = TRUE, sep = ",")
data.frame(mydata)
length(mydata$V1)
> length(mydata$V1)
[1] 1982
```

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

R script

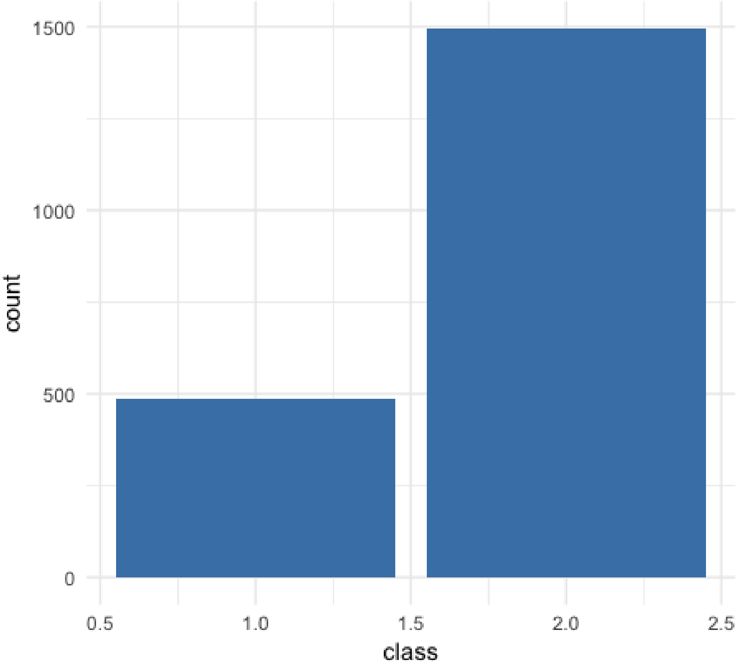
```
> mean(mydata$V2)
[1] -0.9705022
> median(mydata$V2)
[1] -0.4381677
```

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

R script

```
> var(mydata$V1)
[1] 2.676754
> sd(mydata$V1)
[1] 1.636079
```

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

```
barplot(table(mydata$V5), xlab = "class", ylab = "count", col = "blue")
```

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

```
source("dist.euclidean.R")

x1 <- c(1, 2)
x2 <- c(3, 4)
x3 <- c(6, 4)

d12 <- dis(x1, x2)
d23 <- dis(x2, x3)
d13 <- dis(x1, x3)

if (d12 < d23 & d12 < d13)
  "x1 and x2 is the least dissmilar"
if (d13 < d23 & d13 < d12)
  "x1 and x3 is the least dissmilar"
if (d23 < d12 & d23 < d13)
  "x2 and x3 is the least dissmilar"

[1] "x1 and x2 is the least dissmilar"
```

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

```
source("sample.mean.R")
source("sample.variance.R")

number = c(15, 2, 44, 21, 40, 20, 19, 18)

mean.value (number)
variance.value (number)

> mean.value (number)
[1] 22.375
> variance.value (number)
[1] 183.6964
```