CIND123 Summer 2019 - Assignment #2

Tasdeed Aziz

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Use R Studio for this assignment. Edit the file `A2-S19-Q` and insert your R code where wherever you see the string "INSERT YOUR ANSWER HERE"

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

When your are done with your answers and before submitting, save the file with the following naming convention :your **Lastname_firstname**

Submit **both** the rmd and the pdf output(or word or html) files, failing to submit **both** will be subject to mark deduction.

Sample Question and Solution

Use `seq()` to create the vector $(1, 2, 3, \ldots, 20)$.

```
seq(1,20)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

Question 1

The following dataset represents the Population by Highest Educational Attainment (Neighbourhood/Ward), provided by the City of Edmonton under the folloing license https://data.edmonton.ca/stories/s/City-of-Edmonton-Open-Data-Terms-of-Use/msh8-if28/

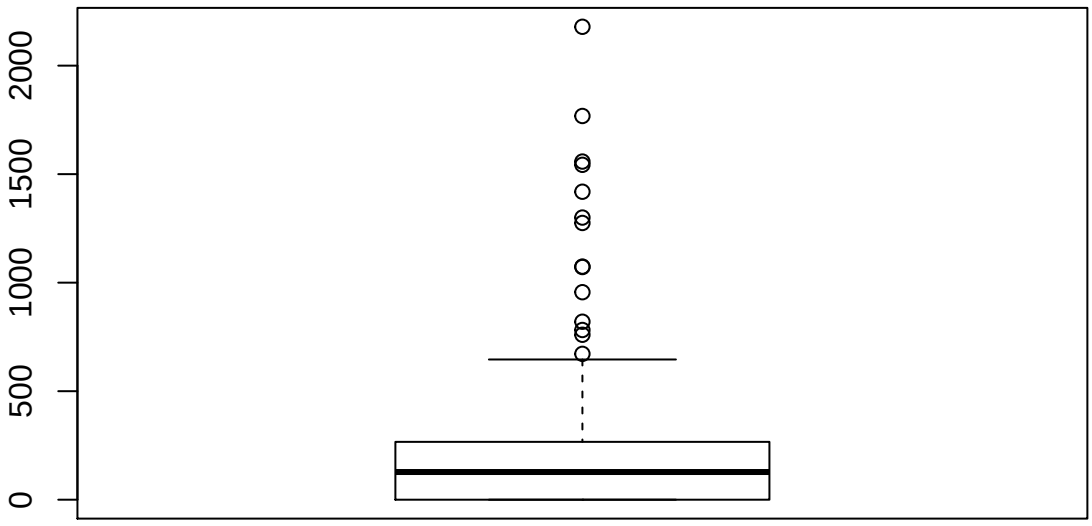
Download and store the datset using the folloing command

```
censusData <- read.csv(file = 'https://data.edmonton.ca/resource/f7ms-98xd.csv', header = T)
str(censusData)
```

```
## 'data.frame':    388 obs. of  15 variables:
##  $ ward                                                : Factor w/ 12 levels "WARD 1","WARD 10",..
##  $ neighbourhood_number                                : int  3140 3330 2690 4720 3381 6510 4060 6
##  $ neighbourhood_name                                  : Factor w/ 388 levels "ABBOTTSFIELD",..: 8
##  $ no_certificate_diploma_or_degree                    : int  63 55 14 105 0 0 261 126 0 153 ...
##  $ high_school_diploma_or_equivalent                   : int  280 445 126 692 0 0 658 565 0 497 ..
##  $ trades_certificate                                  : int  36 75 24 156 0 0 119 143 0 120 ...
##  $ registered_apprenticeship_certificate               : int  6 14 3 23 0 0 9 39 0 13 ...
##  $ college_certificate_or_diploma                      : int  256 257 64 624 0 0 343 484 0 355 ...
##  $ university_certificate_below_bachelor_s_level       : int  60 67 14 89 0 0 56 123 0 143 ...
##  $ bachelor_s_degree                                   : int  415 552 38 581 0 0 210 166 0 365 ...
##  $ university_certificate_or_diploma_above_bachelor_level: int  61 137 4 75 0 0 22 20 0 33 ...
##  $ medical_degree                                      : int  38 58 1 9 0 0 1 5 0 13 ...
##  $ master_s_degree                                     : int  137 194 4 125 0 0 46 33 0 46 ...
##  $ earned_doctorate                                    : int  42 64 2 25 0 0 0 10 0 5 ...
##  $ no_response                                         : int  25 141 42 2218 0 0 843 305 0 102 ...
```

a) Remove the all the outliers from the `bachelor_s_degree` variable, then store it as `bachelor_s_degree_without_outlie`

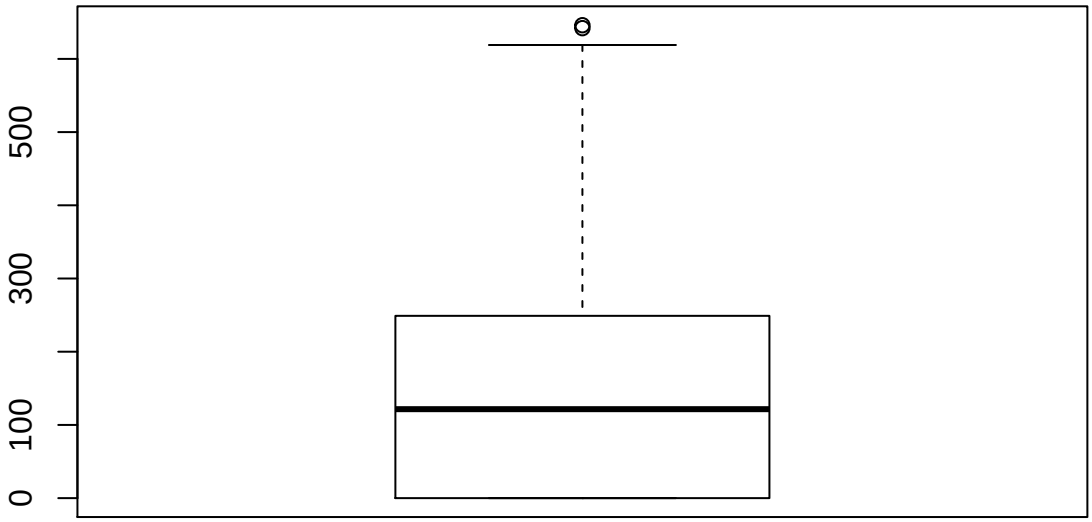
```
#Assigning the dataset with variable 'a'
a <- censusData

#Detecting outliers
outliers <- boxplot(a$bachelor_s_degree)$out
```



```
outliers
```

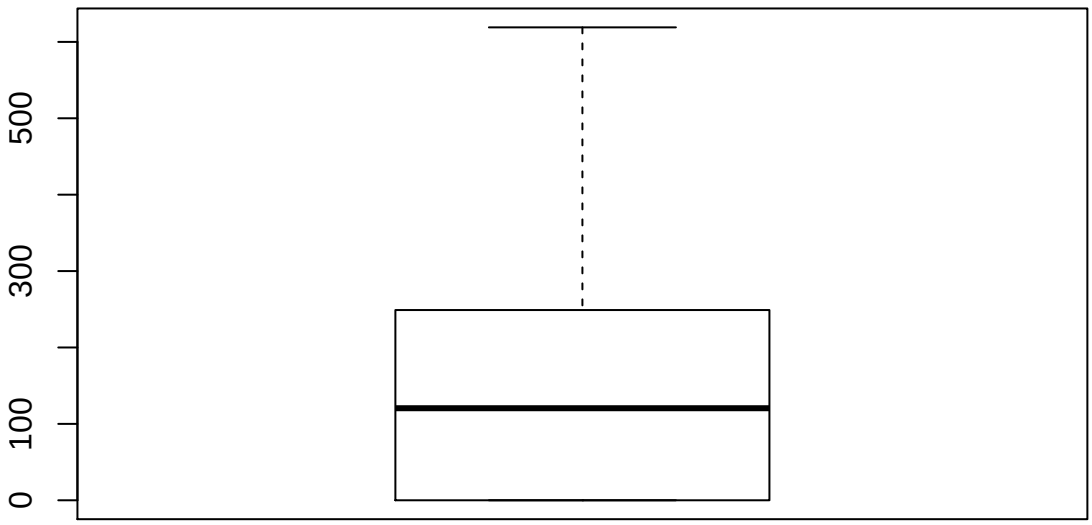
```
##  [1] 1275 1072  956 2179 1768 1543  760  672  782 1558  820 1074 1300 1419
```

```
#removing outliers using which function
a<-a[-which(a$bachelor_s_degree %in% outliers),]
boxplot(a$bachelor_s_degree)$out
```




```
## [1] 642 646
```

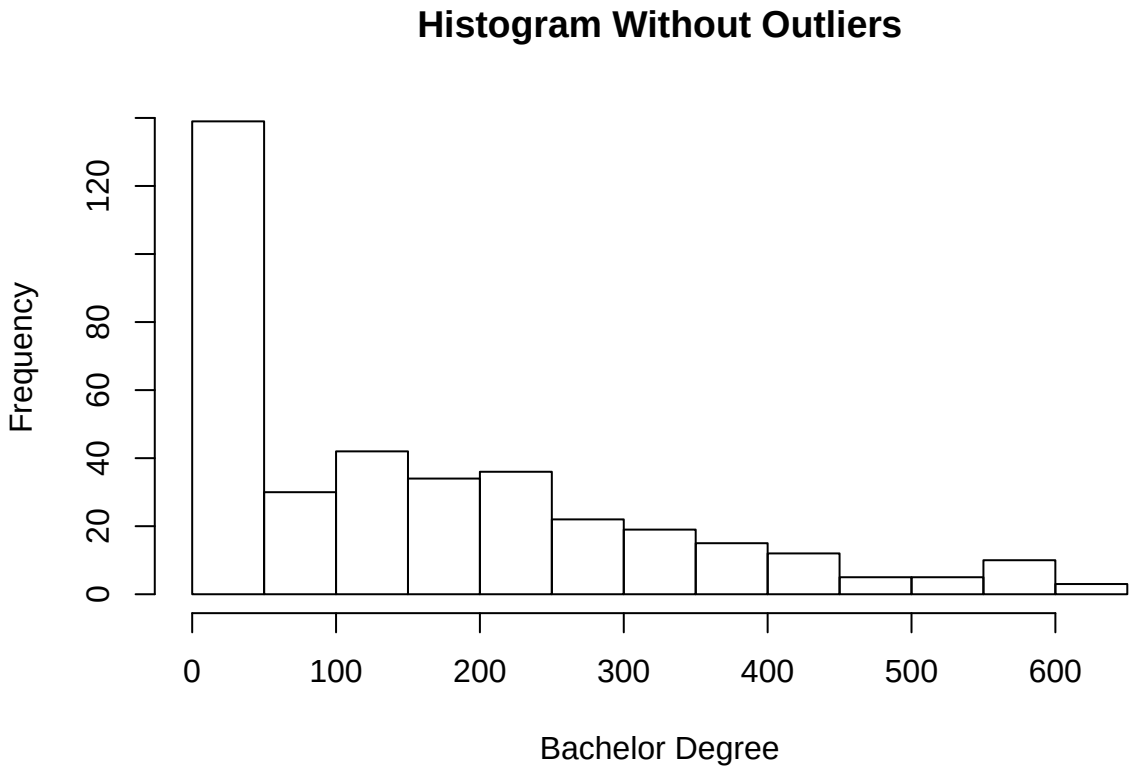
```
#Removing the  all outliers using selector as two outliers were not detected by boxplot but exsists in

bachelor_s_degree_without_outliers <- a$bachelor_s_degree[a$bachelor_s_degree < 642]
boxplot(bachelor_s_degree_without_outliers)
```



b) Plot two histograms that show the distribution of the `bachelor_s_degree` and `bachelor_s_degree_without_outliers` variables.

```r
hist(censusData$bachelor_s_degree, xlab = "Bachelors Degree", main = 'Histogram With Outliers')
```


```r
hist(bachelor_s_degree_without_outliers, xlab = 'Bachelor Degree', main = 'Histogram Without Outliers')
```



c) Use the aggregate function to determine the sum of `medical_degree` holders grouped by `ward`.

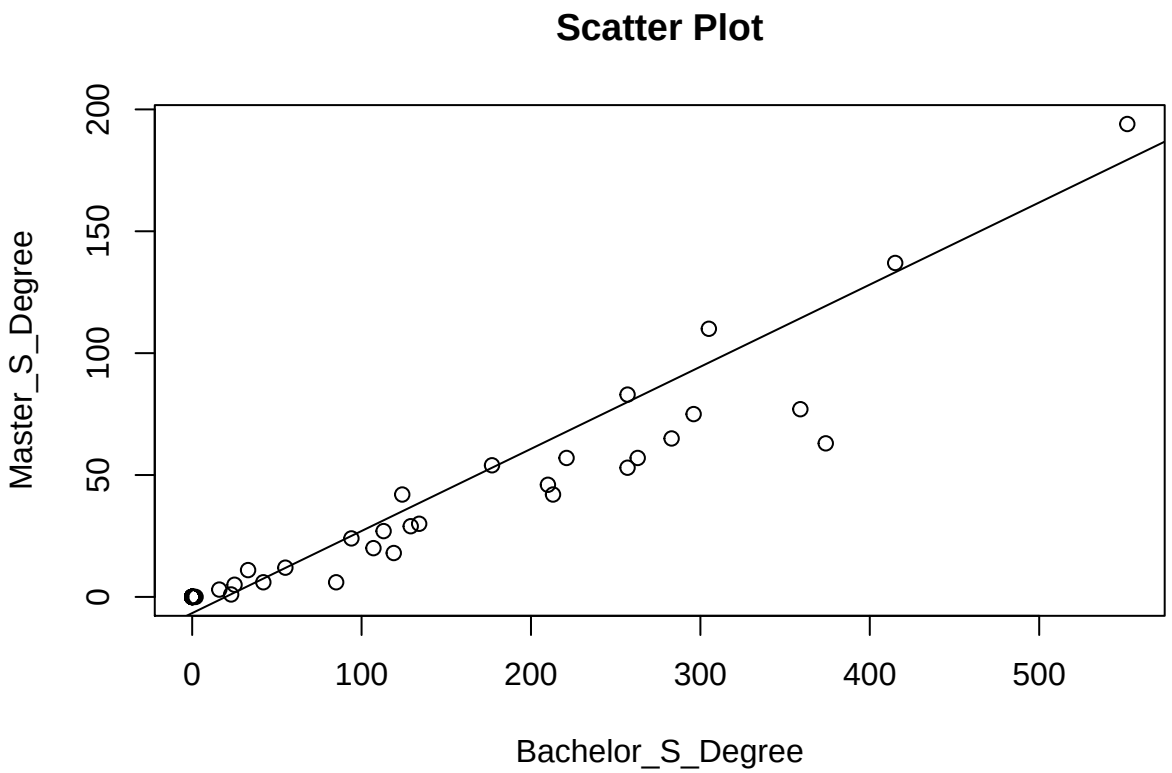
```
aggregate(list(Medical_Degree = censusData$medical_degree),list(Ward = censusData$ward), FUN = sum)
```

```
##       Ward Medical_Degree
## 1   WARD 1            254
## 2  WARD 10            312
## 3  WARD 11            161
## 4  WARD 12            234
## 5   WARD 2            223
## 6   WARD 3            100
## 7   WARD 4            120
## 8   WARD 5            477
## 9   WARD 6            363
## 10  WARD 7             79
## 11  WARD 8            535
## 12  WARD 9            962
```

d) Draw a scatterplot for the `bachelor_s_degree` and `master_s_degree` holders in `WARD 1`. Describe this relationship in terms of strength and direction.

Description: The variables has positive correlation as most of the point falls on the straight line as shown by the plot.

```
plot(censusData$bachelor_s_degree[which(censusData$ward == 'WARD 1')],censusData$master_s_degree[which(
abline(lm(master_s_degree ~ bachelor_s_degree,data = censusData))
```



Question 2

In an experiment of rolling 10 dice simultaneously. Use the binomial distribution to calculate the followings:
a) The probability of getting six 6's

```
dbinom(x=6,size =10,p=1/6)
```

```
## [1] 0.002170635
```

b) The probability of getting six, seven, or eight 4's

```
a <- dbinom(x=6:8, size = 10, p=1/6)
a
```

```
## [1] 2.170635e-03 2.480726e-04 1.860544e-05
```

```
sum(a)
```

```
## [1] 0.002437313
```

c) The probability of getting six odd numbers

```
dbinom(x=6,size = 10,p=1/2)
```

```
## [1] 0.2050781
```

Question 3

In a shipment of 20 engines, history shows that the probability of any one engine proving unsatisfactory is 0.1

a) Use the Binomial approximation to calculate the probability that at least three engines are defective?

```
1 - pbinom(q=2, size = 20,p=0.1)
```

```
## [1] 0.3230732
```

b) Use the Poisson approximation to calculate the probability that at least three engines are defective?

```
1-ppois(q=2,lambda = (20*0.1),lower.tail = TRUE)
```

```
## [1] 0.3233236
```

c) Compare the results of parts a and b, then illustrate on how well the Poisson probability distribution approximates the Binomial probability distribution.

```
# The binomial approximation of atleast three engines defective is 32.31%
# The poison distribution of atleast three engine defective is 32.33%
# Both gives similar result with a very small difference. Poisson distribution does give a better appro
# due to lambda which is calculated by np ( size * probability).
```

Question 4

In a shipment of 300 processors, there are 12 defective processors. A quality control consultant randomly collects 6 processors for inspection to determine whether they are defective. Use the Hypergeometric approximation to calculate the following:

a) The probability that there are exactly 2 defectives in the sample

```
dhyper(x=2,m=12,n=(300-12),k=6)
```

```
## [1] 0.01924295
```

b) The probability that there are at most 5 defectives in the sample, P(X<=5).

```
phyper(q=5,m=12,n=(300-12),k=6)
```

```
## [1] 1
```

Question 5

a) Suppose widgit weights produced at Acme Widgit Works have weights that are normally distributed with mean 17.46 grams and variance 375.67 grams. What is the probability that a randomly chosen widgit weighs more than 19 grams?

```
pnorm(q = 19, mean = 17.46, sd = sqrt(375.67),lower.tail = FALSE, log.p = FALSE)
```

```
## [1] 0.4683356
```

b) Suppose IQ scores are normally distributed with mean 100 and standard deviation 15. What is the 95th percentile of the distribution of IQ scores?

```
qnorm(p=0.95,mean=100, sd=15,lower.tail = TRUE, log.p = FALSE)
```

```
## [1] 124.6728
```

c) Suppose widgits produced at Acme Widgit Works have probability 0.005 of being defective. Suppose widgits are shipped in cartons containing 25 widgits. What is the probability that a randomly chosen carton contains exactly one defective widgit?

```
dbinom(x=1,size = 25,p=0.005,log=FALSE)
```

```
## [1] 0.1108317
```

d) Suppose widgits produced at Acme Widgit Works have probability 0.005 of being defective. Suppose widgits are shipped in cartons containing 25 widgits. What is the probability that a randomly chosen carton contains no more than one defective widgit?

```
pbinom(q=1,size = 25,p=0.005,log=FALSE)
```

```
## [1] 0.9930519
```

END of Assignment #2.