Weekly Assignment 2

Sean Leggett - BDA201 February 6, 2020

Question 1

Examine the provided csv file and perform basic data inspection.

Answer: No gaps exist in the data. The subjects have been anonymized. Fourteen fields exist. Column titles include some spaces. Last names are duplicated for Jane/John Doe.

Question 2

What is the datatype of each feature?

Answer:
* ID - Qualitative/ordinal
* Last Name - Qualitative/nominal
* First Name - Qualitative/nominal
* City - Qualitative/nominal
* State - Qualitative/nominal
* Gender - Qualitative/nominal
* Student Status - Qualitative/nominal
* Major - Qualitative/nominal
* Country - Qualitative/nominal
* Age - Quantitative/interval/continuous
* SAT - Quantitative/ratio/discrete
* Average score (grade) - Quantitative/ratio/continuous
* Height (in) - Quantitative/ratio/continuous
* Newspaper readership (times/wk) - Quantitative/ration/discrete

Question 3

Use summary() function to display a summary of the features.

Answer:

```r
scores <- read.csv("Assignment_2_data.csv")
summary(scores)
```

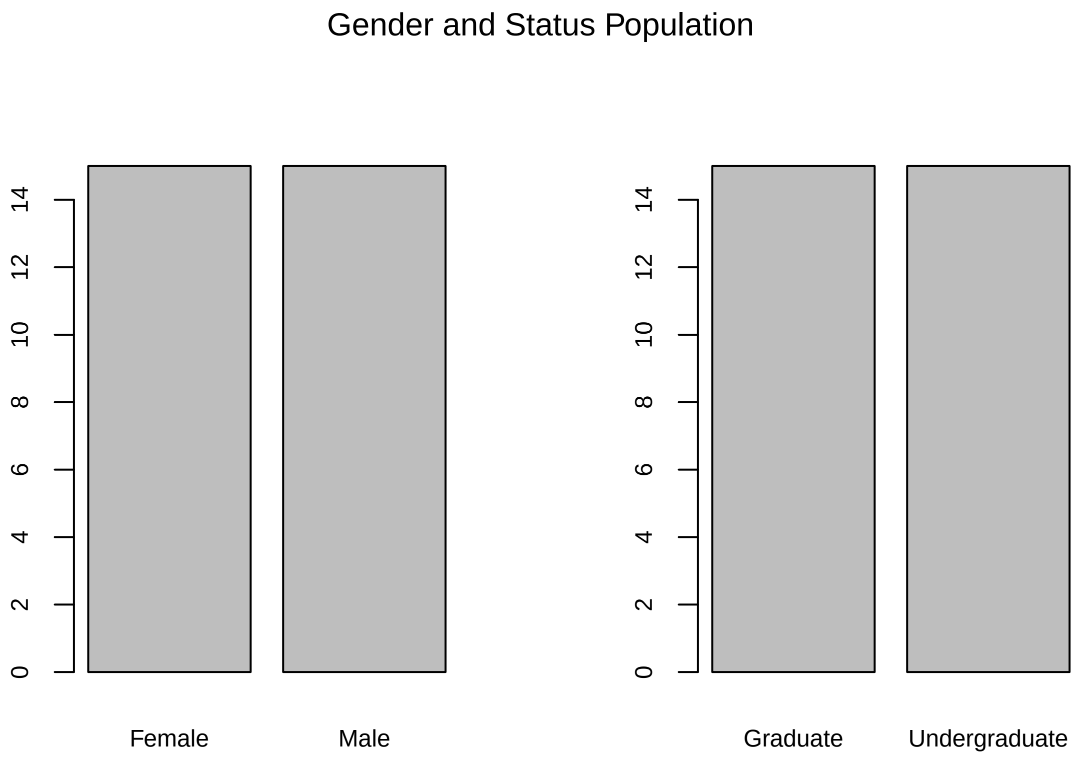
```
##      ï..ID         Last.Name    First.Name           City
##  Min.   : 1.00   DOE01  : 2   JANE01 : 1   New York    : 2
##  1st Qu.: 8.25   DOE02  : 2   JANE02 : 1   Acme        : 1
##  Median :15.50   DOE03  : 2   JANE03 : 1   Amsterdam   : 1
##  Mean   :15.50   DOE04  : 2   JANE04 : 1   Beijing     : 1
##  3rd Qu.:22.75   DOE05  : 2   JANE05 : 1   Buenos Aires: 1
##  Max.   :30.00   DOE06  : 2   JANE06 : 1   Caracas     : 1
##                  (Other):18   (Other):24   (Other)     :23
##         State       Gender         Student.Status      Major
##  New York  : 5   Female:15   Graduate     :15     Econ    :10
##  Argentina : 1   Male  :15   Undergraduate:15     Math    :10
##  Arizona   : 1                                    Politics:10
##  Bulgaria  : 1
##  California: 1
##  Canada    : 1
##  (Other)   :20
##       Country        Age            SAT       Average.score..grade.
##  US       :20   Min.   :18.0   Min.   :1338   Min.   :63.00
##  Argentina: 1   1st Qu.:19.0   1st Qu.:1658   1st Qu.:72.00
##  Bulgaria : 1   Median :23.0   Median :1817   Median :79.50
##  Canada   : 1   Mean   :25.2   Mean   :1849   Mean   :80.37
##  China    : 1   3rd Qu.:30.0   3rd Qu.:2032   3rd Qu.:88.00
##  Holland  : 1   Max.   :39.0   Max.   :2309   Max.   :96.00
##  (Other)  : 5
##   Height..in.    Newspaper.readership..times.wk.
##  Min.   :59.00   Min.   :3.000
##  1st Qu.:63.00   1st Qu.:4.000
##  Median :66.50   Median :5.000
##  Mean   :66.43   Mean   :4.867
##  3rd Qu.:70.75   3rd Qu.:6.000
##  Max.   :75.00   Max.   :7.000
##
```

Question 4 and Question 5

How many males/females? How many graduate/undergraduate? Plot both using bar plots.

Answer:
Summary shows us that there are 15 each of males and females. Also, 15 each of graduates and undergraduates.

```
par(mfrow=c(1,2)) ## draw layout for frames
par(cex.axis=0.75) ## smaller text on x axis to fit
barplot(table(scores$Gender))
barplot(table(scores$Student.Status))
mtext("Gender and Status Population", outer=TRUE,  cex=1, line = -.9)
```



Question 6

Is the average SAT score same for graduates and undergraduates?

Answer:
Undergraduates have a slightly higher average(mean) SAT score.1,841.2 for graduates vs 1,856.6 for undergraduates.

```r
## subset dataframe based on status
grads <- subset(scores, Student.Status == "Graduate")
undergrads <- subset(scores, Student.Status == "Undergraduate")

## calculate means
gradsavg <- mean(grads$SAT)
underavg <- mean(undergrads$SAT)

## display
gradsavg
```

```
## [1] 1841.2
```

```r
underavg
```

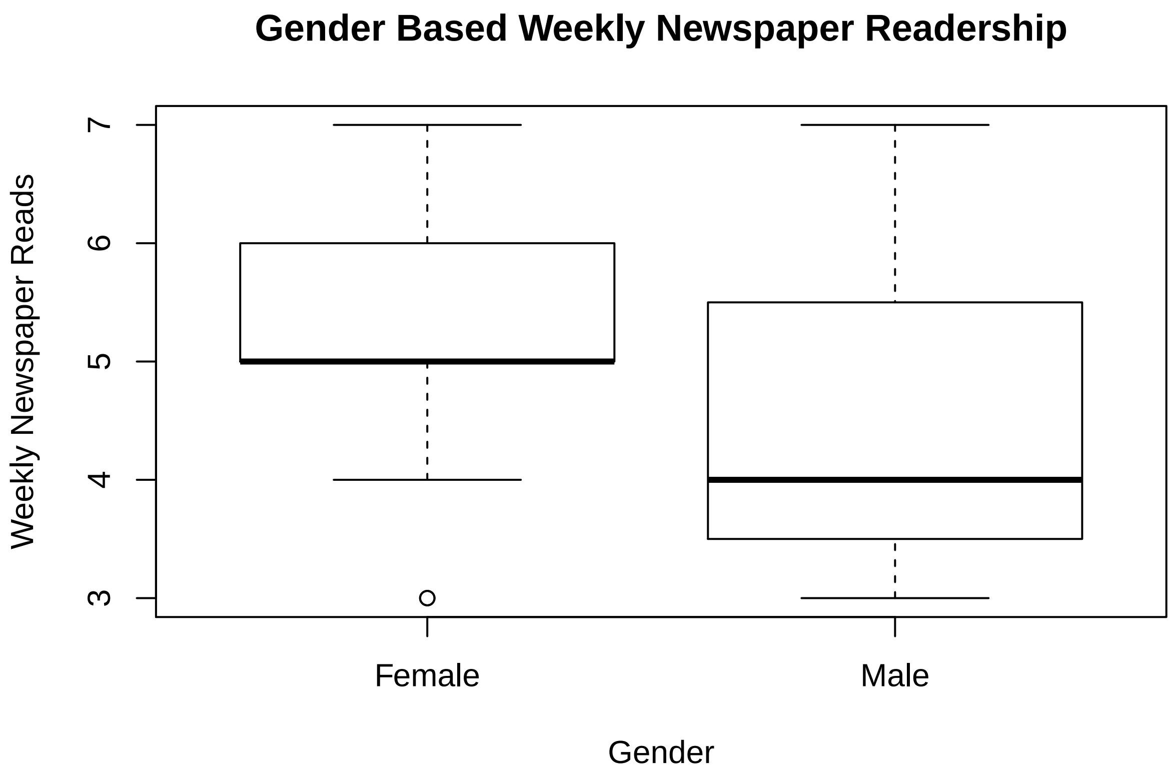
```
## [1] 1856.6
```

Question 7

Between men or women? Who reads the newspaper more frequently and which group has more variation? (show using box plots)

Answer:
The women from this sample read more newspapers weekly than men; the men from the sample have a greater variability in number of times per week they read newspapers. The plot below illustrates this finding.

```
readershipplot <- boxplot(Newspaper.readership..times.wk. ~ Gender, data = scores,
                          main = "Gender Based Weekly Newspaper Readership",
                          xlab = "Gender",
                          ylab = "Weekly Newspaper Reads")
```



Question 8

For age, Height, and Newspaper readership, calculate the following using R: a. Measures of location such as mean, median, and mode (where applicable). b. Measures of variation such as variance, standard deviation and IQR

Answer:
The code below provides the measures requested:

Part A: Mean and Median
Calculated and grouped into dataframe for ease of presentation.

```
## calculate and assign means and medians
agemean <- mean(scores$Age)
heightmean <- mean(scores$Height..in.)
readershipmean <- mean(scores$Newspaper.readership..times.wk.)
agemedian <- median(scores$Age)
heightmedian <- median(scores$Height..in.)
readershipmedian <- median(scores$Newspaper.readership..times.wk.)

## assign to dataframe for display purposes
dfLocations <- data.frame("Variable" = c("Age", "Height", "Readership"), "Mean" = c(agemean, heightmean

## display
dfLocations
```

```
##     Variable      Mean Median
## 1         Age 25.200000   23.0
## 2      Height 66.433333   66.5
## 3 Readership  4.866667    5.0
```

Part A: Mode

There is no function for mode (statistical mode) in base R. With a small dataset such as this, it is reasonable to simply sort and count.

For Age we have a dual mode of 18 and 19 each with 5 observations. For Height we have a mode of 68 inches with 4 observations. For Readership we have a mode of 5 with 9 observations.

There are packages to expand statistical functions and some include mode calculations. For larger datasets these would be hepful. Or else a programmatical approach leveraging some sort of count and max functions.

```
agemodetest <- sort(scores$Age)
agemodetest
```

```
##  [1] 18 18 18 18 18 19 19 19 19 19 20 20 21 21 21 25 25 26 28 30 30 30 30
## [24] 31 33 33 33 37 38 39
```

```
heightmodetest <- sort(scores$Height..in.)
heightmodetest
```

```
##  [1] 59 59 60 61 62 62 62 63 63 63 64 64 64 65 66 67 67 68 68 68 68 70 71
## [24] 71 71 72 73 73 74 75
```

```
readershipmodetest <- sort(scores$Newspaper.readership..times.wk.)
readershipmodetest
```

```
##  [1] 3 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 7 7 7
```

Part B: Measures of Variation
Calculated and grouped into dataframe for ease of presentation.

```
## calculate and assign variance, SD & IQR
agevariance <- var(scores$Age)
heightvariance <- var(scores$Height..in.)
readershipvariance <- var(scores$Newspaper.readership..times.wk.)
sdage <- sd(scores$Age)
sdheight <- sd(scores$Height..in.)
sdreadership <- sd(scores$Newspaper.readership..times.wk.)
ageIQR <- IQR(scores$Age)
heightIQR <- IQR(scores$Height..in.)
readershipIQR <- IQR(scores$Newspaper.readership..times.wk.)

## combine in dataframe for presentation
dfvariation <- data.frame("Variable" = c("Age", "Height", "Readership"), "Variance" = c(agevariance, hei

dfvariation
```

```
##     Variable  Variance Standard.Deviation   IQR
## 1        Age 47.200000           6.870226 11.00
## 2     Height 21.702299           4.658573  7.75
## 3 Readership  1.636782           1.279368  2.00
```

Reference Material:
1. Course Content
2. Sams Teach Yourself R in 24 Hours, Andy Nicholls, Richard Pugh, Aimee Gott. Sams, 2016.