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/discrete Average score (grade) - Quantitative/discrete Height (in) - Quantitative/continuous Newspaper readership (times/wk) - Quantitative/continuous

Question 3

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

Answer:

```
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
```

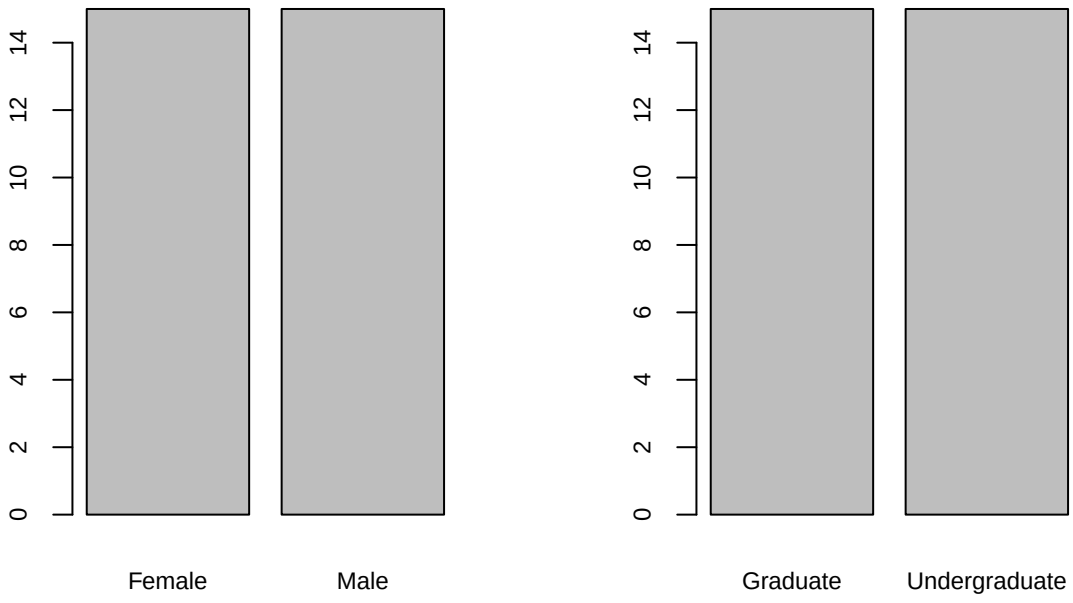
```
##  China    : 1   3rd Qu.:30.0   3rd Qu.:2032   3rd Qu.:88.00
##  Holland  : 1   Max.   :39.0   Max.   :2309   Max.   :96.00
##  (Other)  : 5
##   Height..in.    Newspaper.readership..times.wk.
##  Min.   :59.00   Min.   :3.000
##  1st Qu.:63.00   1st Qu.:4.000
##  Median :66.50   Median :5.000
##  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.

```
barsplots <-
par(mfrow=c(1,2))
par(cex.axis=0.75)
barplot(table(scores$Gender))
barplot(table(scores$Student.Status))
```



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.

```
grads <- subset(scores, Student.Status == "Graduate")
undergrads <- subset(scores, Student.Status == "Undergraduate")

gradsavg <- mean(grads$SAT)
underavg <- mean(undergrads$SAT)
gradsavg
```

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

```
underavg
```

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