Homework #1, Miles Tweed

DUE: Monday, August 31st, at 9pm via Canvas Submission.

Please submit the solution on Canvas into the corresponding assignment (e.g. "Homework #1") in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide only code and output. NO NEED TO COPY THE PROBLEM FORMULATION (!)

Please see the partial solutions (both .pdf and .Rmd documents) to get an idea of what's expected, plus an idea of how R markdown code should look like.

Problem 1 (make sure to include code with outputs)

For the *fl_student_survey.csv* file, proceed to:

1. Pick a categorical variable (besides those already utilized in lecture code) and describe its distribution via appropriate numerical and graphical summaries introduced in class. Accompany it with a description in your own words.

I looked at the ordinal categorical variable, 'religiosity', which had the values of no religiosity (0), some religiosity (1), moderately religious (2), very religious (3). The modal category for this variable was 1 which indicates that a majority of students have some level of religiosity. The graphical representation indicates that the majority of the sample subjects either had low (1) or no (0) religiosity. Below are the number of individuals who reported each level of religiosity. The total number of subject in the sample is 60.

```
student_survey <-read.csv('http://sites.williams.edu/bklingen/files/2015/07/fl_student_survey.csv')
table(student_survey$religiosity)
```

```
##
##  0  1  2  3
## 15 29  7  9
```

The following shows the percentage of sample subjects that fall into each category.

```
table(student_survey$religiosity)/60*100
```

```
##
##         0        1        2        3
## 25.00000 48.33333 11.66667 15.00000
```

One quarter (25% or 15 individuals) of the subjects stated that they were not religious at all, 29 subjects (48.33%) claimed to have a low level of religious affiliation, 7 subject (11.667%) claimed to be moderately religious, and 9 subject (15%) claimed to be very religious. Below is a graphical representation of the sample which shows that a majority of the sample fall into the lower two categories (1 and 0).

```
barplot(table(student_survey$religiosity)/60*100,
        main = 'Religiosity of Subjects',
        col = c("mistyrose", "lightgreen", 'lavender', 'cornsilk'),
        xlab = 'Level of Religiosity',
        ylab = '% of Sample Subjects')
```

Religiosity of Subjects
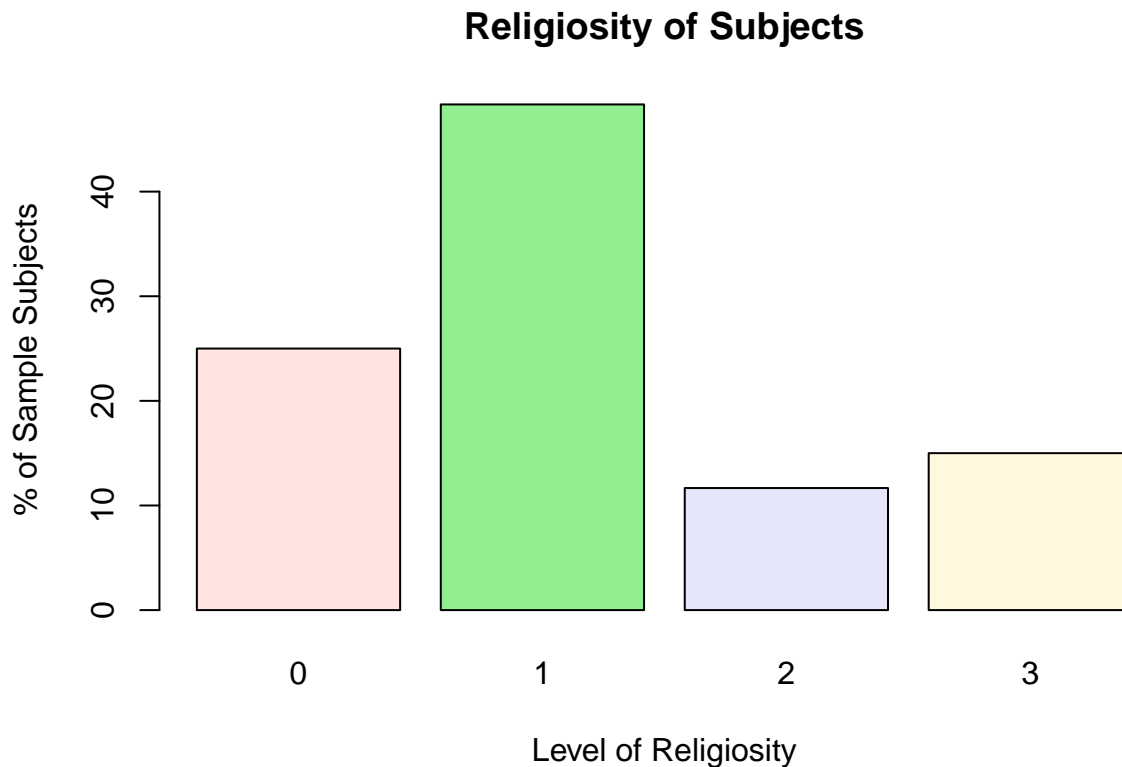
2. Pick a quantitative variable (besides those already utilized in lecture code). Is it discrete or continuous? Describe its distribution via an appropriate graphical summary. What shape does it take? Accompany it with a description in your own words.

The distance_residence variable is quantitative and continuous. For this analysis distance is assumed to be in miles. The average distance that students from the sample population travel to and from school is:

```
sum(student_survey$distance_residence)/60
```

```
## [1] 3.818333
```

The maximum distance that students from the sample population travel to and from school is:

```
max(student_survey$distance_residence)
```

```
## [1] 20
```

The minimum distance that students from the sample population travel to and from school is:

```
min(student_survey$distance_residence)
```

```
## [1] 0.2
```

Overall, the graphical representation, shows a right skewed distribution with a majority of student residing closer to school (within 5 miles).

```
hist(student_survey$distance_residence,
     col = c("mistyrose", "lightgreen", 'lavender', 'cornsilk'),
     xlab = 'Distance',
     main = "Student's distance to residence")
```

Student's distance to residence
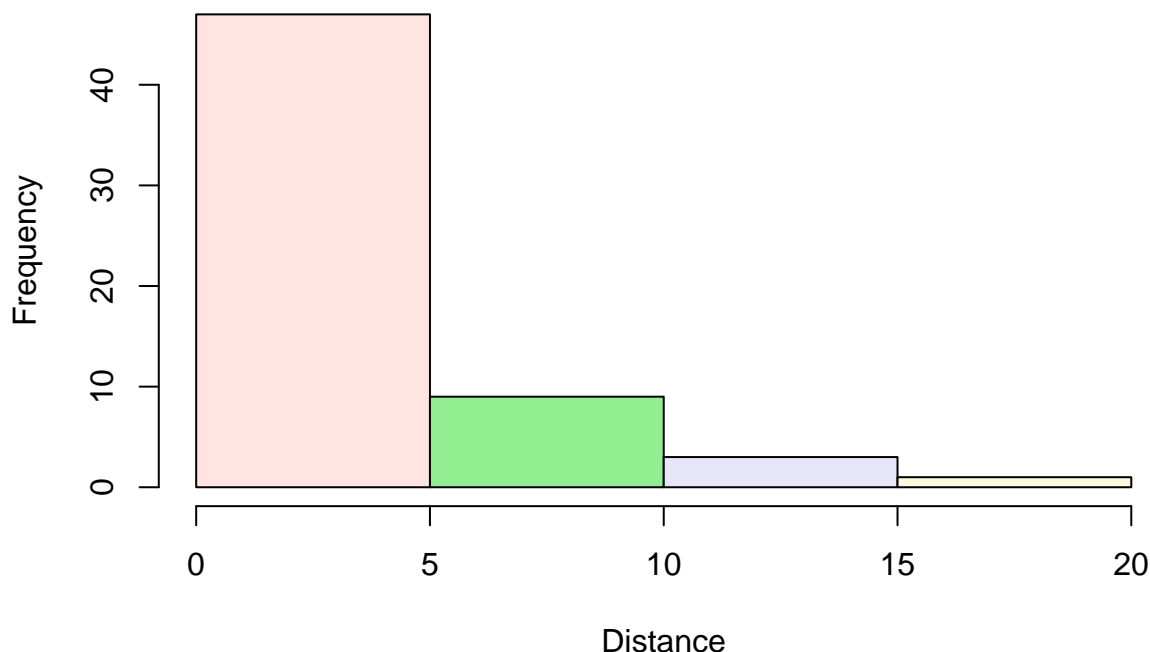
Problem 2

1.7

   a) The sample is the collection of 3077 people who responded to the survey.

   b) The population is the population of the United States in general.

   c) The statistics reported is the percentage of the sample subjects who responded that they believed the anti-religious books should be removed (23%).

1.8

   a) The sample is the 1200 Floridians that were surveyed while the population is the entire population of Florida.

   b) The reported percentages are statistics because they represent a numerical summary of the sample.

1.9

   a) The subject would be one of the individual graduate students that were interviewed.

   b) The sample would be the collection of students from the one graduate program that were interviewed.

   c) The population consists of graduate programs in the liberal arts student's field.

1.10

   a) The samples are the collections of random citizens from each country. The population is the collection of citizens from the countries that make up a region (i.e. Africa, North America, etc.)

   b) As stated ("74% of Africans felt", "whereas 38% of North Americans felt"), the percentages are population parameters. This is because the percentages are being used to characterize the populous of the entire region instead of just the collection of individuals from each country.

2.3

a) quantitative (discrete)

b) categorical (nominal)

c) categorical (nominal)

d) quantitative (continuous)

2.6

a) continuous

b) discrete

c) continuous

d) discrete