CSC 315, Fall 2018
Lab #3: Associations

You should create a single R script that covers all problems, and your script should have a heading similar to the following:

```
#######################################################
##  Garrett Dancik
##  CSC 315, Lab #3
#######################################################
```

Your answer to each question should be numbered in a comment in your R script. When your script is complete, create a Notebook and turn in a hardcopy of the Notebook when the assignment is due.

Note: all graphs must be given an appropriate title, x-axis label, and y-axis label

1. Construct the following contingency table in *R*, along with a table showing the appropriate conditional proportions, where *Income* is the explanatory variable and *Happiness* is the response variable. Answer the question: does there appear to be a relationship between Income and happiness? Why or why not?

	Happiness		
Income	Not Too Happy	Pretty Happy	Very Happy
Above average	26	233	164
Average	117	473	293
Below average	172	383	132

2. Import our class survey data (a cleaned version), which is available here: https://gdancik.github.io/CSC-315/data/datasets/csc-315_survey_cleaned.csv

3. Construct a stacked bar graph that shows whether there is an association between whether someone is a Cat or Dog person and their favorite CSC course. Note: ggplot will plot missing values in the response (*y*) variable. In order to prevent this, we must remove rows with missing values from the data frame. This is accomplished with the code below (assuming the survey data is stored in *survey*). You can then generate the plot using *survey2*.

```
library(dplyr)
survey2 <- filter(survey, !is.na(Favorite.CSC.Course))
```

4. In our class, does there appear to be an association between whether someone is a Cat or Dog person and their favorite CSC course? Why or why not? Are there any courses that "Dog People" liked the best that were not mentioned by "Cat People"?

5. In our class, 25% of "Dog People" chose CSC-270 as their favorite CSC course while 0% of "Cat People" chose this course. Later in the semester we will discuss the concept of "statistical significance", which quantifies whether differences like this can be explained by chance. Construct a contingency table for this data and determine how many more "Dog People" chose CSC-270 than "Cat People". Would you expect this difference to hold if we had a larger number of students in the course?

6. Construct a scatterplot of HS GPA vs. College GPA, so that College GPA would be predicted from HS GPA, and add the regression line from the corresponding linear model.

7. Calculate the correlation and describe the association between HS and College GPA.

8. The *mtcars* dataset contains data on 32 cars extracted from the 1974 *Motor Trend US* magazine. This dataset is available in *R* in the data.frame *mtcars*, and can be viewed using the code below:

   ```
   View(mtcars)
   ```

   The two variables we will examine are *wt*, the weight of the car in thousands of pounds, and *mpg*, the gas mileage in miles per gallon from road tests. Additional information about the dataset can be found by typing *?mtcars* in the *R* console. Construct a scatterplot that predicts gas mileage from the vehicle's weight, and add the corresponding regression line. Describe the relationship between weight and miles per gallon based on these results.

9. Find the linear regression line that predicts miles per gallon from weight. Find and interpret the *y*-intercept. Find and interpret the slope.

10. Based on this set of cars (in 1974), what would you predict the miles per gallon to be for a car that weighed 3000 pounds? What would you predict the miles per gallon to be for a car that weighed 7000 pounds? (Remember that if the prediction would be an extrapolation, you should say so and not make this prediction).