STAT 345 Fall 2018
Homework 8 - Data Exploration Name: _____

Please adhere to the homework rules as given in the Syllabus.

1. Sample Statistics. Timothy and Jimothy run competing lemonade stands. Their sales (in dollars) for this past week are given in the following table. **All calculations here should be done by hand**.

Timothy (x_i)	Jimothy (y_i)
12	10
13	14
9	9
10	11
6	8
11	14
14	13

a) Calculate the sample mean, variance and standard deviation for Timothy sales. Do the same for Jimothy.

b) Calculate the sample correlation between Timothy and Jimothy's sales.

c) Combine Timothy and Jimothys sales into a single variable (14 observations total now). Find the 5 number summary of these data. Interpret the 3^{rd} quartile.

2. Guess the Correlation. Go to `guessthecorrelation.com`. Enter your full name as a username. Play the game until you get at least 50 points. Submit a screenshot showing your high-score and user name for credit. Top 3 highest scores get challenge points!

3. Bonferroni Correction. Check out the cartoon at `xkcd.com/882`. The point of this comic is that, often, one can find a "significant" result if you search hard enough. If done incorrectly, this is sometimes known as *p-hacking* (or data-dredging, data snooping etc.)

Suppose we have a collection of data $x_1, x_2, \cdots x_n$, and we would like to determine if any of them are outliers. We have a procedure that tests if x_i is an outlier. If x_i is **not** an outlier, the procedure will lie to us (this is called a *false positive*) with probability α (usually α is small like 0.05). Suppose that none of our data points are outliers, and define A_i to be the event that x_i is (wrongly) identified as an outlier so that $P(A_i) = \alpha$. Assume that each A_i is independent?

a) What is the probability that at least 1 point is (wrongly) identified as an outlier? What happens to this probability as n gets large?
Hint: $P(\text{at least one false positive}) = P(A_1 \cup A_2 \cup \cdots A_n)$.

Bonferroni's Inequality states that

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots P(A_n).$$

(Prove this using induction for challenge points). Now suppose we use a different procedure to test if x_i is an outlier. This procedure only lies to us (false positive) with probability α/n. Let B_i denote the event that x_i is (wrongly) identified as an outlier using this new procedure so that $P(B_i) = \alpha/n$.

b) Show that the probability of at least 1 (wrongly) identified outlier, using this new procedure, is always less than or equal to α no matter how large n gets. [1]

[1] Comment: This is a useful tool that allows us to control the false positive rate, but nothing comes for free. Using Bonferronis procedure gives us less "power", meaning that we are less likely to identify real outliers.

4. Data Analysis. The CDI dataset contains demographic information on the 440 largest counties in the US. The dataset can be read into R (from the course webpage) with the following command.

```
read.csv('http://math.unm.edu/~knrumsey/cdi_sample.csv')
```

Extract the column named "PerCapitaIncome" and divide it by 1000 so the units are $1000. Your answer to this problem should be in the form of a typed report.

a) Create a Histogram of the data. Comment on the skew of the data. Does the data look normally distributed? (*Note: I recommend using the argument* `breaks=12` *to get a better looking histogram*

b) Create a Boxplot of the data.

c) Report the mean, variance, and standard deviation. Also give the five number summary of the data and the IQR.

d) Give the mode of the data. Do this again after rounding the data to 1 decimal place.

e) Use a Kernel Density Estimate (KDE) to estimate the mode of the data. Which of the three estimates of the Mode seems to fit the data the best?

f) Create a QQ-plot of the data. What does this say about the Normality of the data?

g) Use the $1.5 \times IQR$ rule to identify potential outliers.

Challenge Problems

h) Create a Box-Cox plot for the data. Choose a transformation of the data based on the Box-Cox plot. Make histograms and QQ-plots to assess the Normality of the transformed data.

i) Now that the transformed data is (theoretically) more Normal, use the z-score method on the transformed data to identify potential outliers.

j) Repeat the process of outlier identification using a Bonferroni correction. Do the results differ?