Problem set 1

I agree to abide by the Stern Code of Conduct (name & NYU ID) ____________________

Question 1

In this question you will imagine an example of a study and use that example to answer each part below.

Think of an example study related to your interests where the goal is to discover the relationship between a predictor (or explanatory variable) and an outcome variable. What is the *outcome variable*, and what is the *predictor variable*?

Are there any potential *reliability* or *validity* issues with either of these variables? Explain why or why not.

Think of a *confounding variable* that might be associated with both the outcome and the predictor, and explain how we might make the wrong conclusion if we studied the predictor and outcome variable without controlling for this confounding variable.

Question 2

- We write $\mathbf{x}$ (bold, lower case) to refer to the list $x_1, x_2, \ldots, x_n$ of n observations of the variable x.
- For a single number denoted c, we write $c\mathbf{x}$ for the list $cx_1, cx_2, \ldots, cx_n$.
- In words, each element of $\mathbf{x}$ gets multiplied by c, forming a new list. We might say this is *scaled by a factor of c.*
- Remember that "$\sum_{i=1}^{n}$ something involving i" just means to add up the expression "something involving i" for all values of i starting at 1 and ending at n.
- For example, $\sum_{i=1}^{3} 2(i-1)^2 = 2(0)^2 + 2(1)^2 + 2(2)^2 = 10$.

(a) For the list $\mathbf{x} = -1, 0, 1$, and $c = 2$, compute the means and SDs of both $\mathbf{x}$ and $c\mathbf{x}$. Answers should be numbers.

(b) Remembering that $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean of $\mathbf{x}$, what is the mean of $c\mathbf{x}$? Answer should be an expression. Show your work.

(c) Remembering that $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the standard deviation (SD) of $\mathbf{x}$, what is the SD of $c\mathbf{x}$? Answer should be an expression. Show your work.

Question 3

In R, install and load the nycflights13 package, and also load the tidyverse package. After loading the packages, type airlines to see the abbreviation codes for each airline. Find the code for Frontier Airlines Inc. Next, use filter() on the flights data to create a new variable containing the subset of flights operated by Frontier. You can name this variable anything you like, but it's generally a good idea for variable names to be self-descriptive.

For hints, see **Section 3.3** in the ModernDive book linked on the course page: http://moderndive.com/3-viz.html

(a) What are the origins and destinations for these flights, and how many flights are in the dataset?

(b) Use ggplot() on the Frontier data, along with geom_density(), to create a density plot of the air_time variable. Find the highest point of the density–the *mode*–what value of the x-axis corresponds to this highest point? Is this value close to the mean() of air_time?

(c) Use the mean and SD to find a range that contains the middle 95% of air_time values.