Homework 1

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

Question 1

In the `gapminder` dataset, we saw an association between GDP per capita and life expectancy. Countries with higher GDP per capita also tend to have higher life expectancy. But we also know that **association is not causation**.

Can you think of any *confounding variables* that may be associated with both GDP per capita and life expectancy? Give two examples.

Can you think of any measurement issues regarding GDP per capita or life expectancy? Give two examples, and explain how they relate to reliability or validity.

Question 2

- We write **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 **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 x and cx. Answers should be numbers.

(b) Remembering that $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean of x, what is the mean of cx? 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 x, what is the SD of cx? 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.