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.

Various answers accepted. A few examples: healthcare expenditures, investment in education, fertility rates (average number of children)

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.

Various answers accepted. One (inter-item) reliability issue with life expectancy is that vital record systems vary from country to country and have also been changing over time. For example, even the most economically developed countries have people who are alive today who do not have birth certificates. In other words, the data sources that life expectancy has been estimated from may be different between different countries.

One (content) validity issue with GDP is its use as the summary of an economy. It may be *useful* to have a single number, but there is no way a single number can adequately represent all aspects of something as large and complex as a country's economy. Another (criterion-related) validity issue with GDP

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.

$\bar{\mathbf{x}} = (-1+0+1)/3 = 0$ and $\overline{c\mathbf{x}} = (-2+0+2)/3 = 0$ (notice this is 2 times $\bar{\mathbf{x}}$).

$SD(\mathbf{x}) = \sqrt{\frac{1}{2}\left[(-1-0)^2 + (0-0)^2 + (1-0)^2\right]} = \sqrt{\frac{1}{2}(2)} = 1$

$SD(c\mathbf{x}) = \sqrt{\frac{1}{2}\left[(-2-0)^2 + (0-0)^2 + (2-0)^2\right]} = \sqrt{\frac{1}{2}(8)} = 2$ (notice this is $|2|$ times $SD(\mathbf{x})$).

(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.

$\overline{c\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}(cx_i) = \frac{c}{n}\sum_{i=1}^{n} x_i = c\bar{\mathbf{x}}$

(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.

$$\begin{aligned}
SD(c\mathbf{x}) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(cx_i - c\bar{x})^2} \\
&= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}c^2(x_i - \bar{x})^2} \\
&= \sqrt{\frac{c^2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= |c|\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}c^2(x_i - \bar{x})^2} \\
&= |c| SD(\mathbf{x})
\end{aligned} \tag{1}$$

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?

After using library(tidyverse) and library(nycflights13) to load the packages, typing airlines to see F9 is the code for Frontier, we do:

```
frontier <- filter(flights, carrier == "F9")
# Origin airport(s)
unique(frontier$origin)
```

```
## [1] "LGA"
```

```
# Destination airport(s)
unique(frontier$dest)
```

```
## [1] "DEN"
```

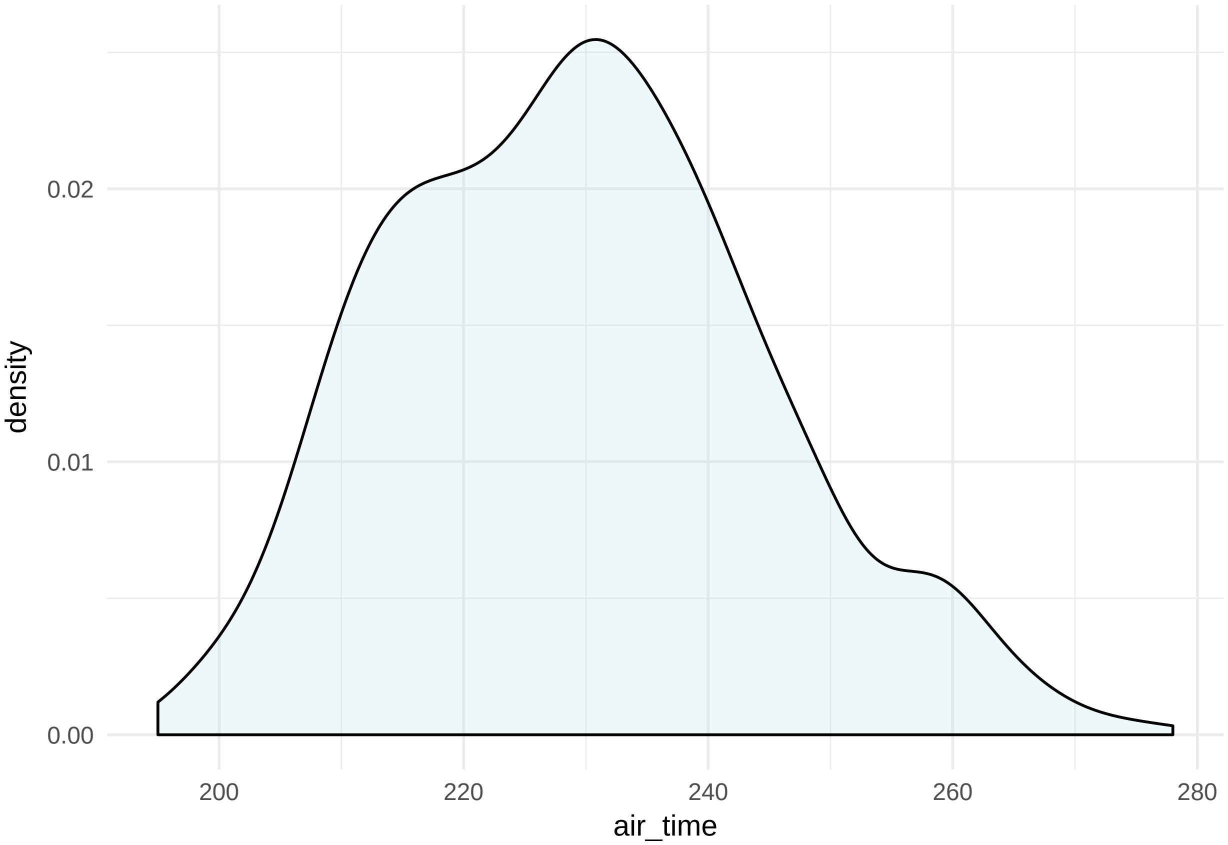
```
# Number of flights (number of rows in the dataset)
nrow(frontier)
```

```
## [1] 685
```

(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?

```
ggplot(frontier, aes(air_time)) + geom_density(fill = "lightblue", alpha = .2) + theme_minimal()
```

```
## Warning: Removed 4 rows containing non-finite values (stat_density).
```



```
mean(frontier$air_time, na.rm = TRUE)
```

```
## [1] 229.5991
```

The mode seems to be slightly larger than 230, and the mean is about 229.6, so these are very close.

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

```
xbar <- mean(frontier$air_time, na.rm = TRUE)
s <- sd(frontier$air_time, na.rm = TRUE)
xbar - 2*s
```

```
## [1] 199.2735
```

```
xbar + 2*s
```

```
## [1] 259.9248
```

Using the 68-95-99 rule, the mean plus or minus two standard deviations (usually) contains the middle 95% of values, i.e. from about 199.3 to about 259.9