Homework 1: Causality and Expressions

Please complete this notebook by filling in the cells provided.

Recommended Reading:

- What is Data Science (http://www.inferentialthinking.com/chapters/01/what-is-data-science.html)
- Causality and Experiments (http://www.inferentialthinking.com/chapters/02/causality-and-experiments.html)
- Programming in Python (http://www.inferentialthinking.com/chapters/03/programming-in-python.html)

For all problems that you must write explanations and sentences for, please provide your answer in the designated space. Moreover, throughout this homework and all future ones, please be sure to not re-assign variables throughout the notebook! For example, if you use `max_temperature` in your answer to one question, do not reassign it later on.

Deadline:

This assignment is due Thursday, January 23 at 11:59 P.M. Late work will not be accepted as per the course policies (see the Syllabus and Course policies on Canvas (https://canvas.yale.edu).

Directly sharing answers is not okay, but discussing problems with the course staff or with other students is encouraged. Refer to the policies page to learn more about how to learn cooperatively.

You should start early so that you have time to get help if you're stuck. The drop-in office hours schedule appears on can be found on Canvas (https://canvas.yale.edu). You can also post questions or start discussions on Piazza (https://piazza.com/yale/spring2020/sds123/home).

Submission:

Submit your assignment both as a .pdf and .ipynb (Jupyter notebook) in Canvas.

To produce the .pdf, please do the following in order to preserve the cell structure of the notebook:

1. Go to "File" at the top-left of your Jupyter Notebook
2. Under "Download as", select "HTML (.html)"
3. After the .html has downloaded, open it and then select "File" and "Print" (note you will not actually be printing)
4. From the print window, select the option to save as a .pdf

To produce the .ipynb, please do the following:

1. Go to "File" at the top-left of your Jupyter Notebook
2. Under "Download as", select "Notebook (.ipynb)"

1. Scary Arithmetic

An ad for ADT Security Systems says,

> "When you go on vacation, burglars go to work [...] According to FBI statistics, over 25% of home burglaries occur between Memorial Day and Labor Day."

Does the data in the ad support the claim that burglars are more likely to go to work during the time between Memorial Day and Labor Day? Please explain your answer.

Write your answer here, replacing this text.

2. Characters in Little Women

In lecture, we counted the number of times that the literary characters were named in each chapter of the classic book, [Little Women (https://www.inferentialthinking.com/chapters/01/3/1/literary-characters)](https://www.inferentialthinking.com/chapters/01/3/1/literary-characters). In computer science, the word "character" also refers to a letter, digit, space, or punctuation mark; any single element of a text. The following code generates a scatter plot in which each dot corresponds to a chapter of Little Women. The horizontal position of a dot measures the number of periods in the chapter. The vertical position measures the total number of characters.

In [ ]:

```python
# This cell contains code that hasn't yet been covered in the course,
# but you should be able to interpret the scatter plot it generates.

from datascience import *
from urllib.request import urlopen
import numpy as np
%matplotlib inline

little_women_url = 'https://www.inferentialthinking.com/data/little_women.txt'
chapters = urlopen(little_women_url).read().decode().split('CHAPTER ')[1:]
text = Table().with_column('Chapters', chapters)
Table().with_columns(
    'Periods',    np.char.count(chapters, '.'),
    'Characters', text.apply(len, 0)
    ).scatter(0)
```

Question 1. Around how many periods are there in the chapter with the most characters? Assign either 1, 2, 3, 4, or 5 to the name `characters_q1` below.

1. 250
2. 390
3. 440
4. 32,000
5. 40,000

In [ ]:

```
characters_q1 = ...
```

Question 2. Which of the following chapters has the most characters per period? Assign either 1, 2, or 3 to the name `characters_q2` below.

1. The chapter with about 60 periods
2. The chapter with about 350 periods
3. The chapter with about 440 periods

In [ ]:

```
characters_q2 = ...
```

To discover more interesting facts from this plot, read Section 1.3.2 (https://www.inferentialthinking.com/chapters/01/3/2/another-kind-of-character) of the textbook.

3. Names and Assignment Statements

Question 1. When you run the following cell, Python produces a cryptic error message.

In [ ]:

```
4 = 2 + 2
```

Choose the best explanation of what's wrong with the code, and then assign 1, 2, 3, or 4 to `names_q1` below to indicate your answer.

1. Python is smart and already knows `4 = 2 + 2`.
2. `4` is already a defined number, and it doesn't make sense to make a number be a name for something else. In Python, "`x = 2 + 2`" means "assign `x` as the name for the value of `2 + 2`."
3. It should be `2 + 2 = 4`.
4. I don't get an error message. This is a trick question.

In [ ]:

```
names_q1 = ...
```

Question 2. When you run the following cell, Python will produce another cryptic error message.

In [ ]:

```
two = 3
six = two plus two
```

Choose the best explanation of what's wrong with the code and assign 1, 2, 3, or 4 to `names_q2` below to indicate your answer.

1. The `plus` operation only applies to numbers, not the word "two".
2. The name "two" cannot be assigned to the number 3.
3. Two plus two is four, not six.
4. The name `two` cannot be followed directly by another name.

In [ ]:

```
names_q2 = ...
```

Question 3. When you run the following cell, Python will, yet again, produce another cryptic error message.

In [ ]:

```
x = print(5)
y = x + 2
```

Choose the best explanation of what's wrong with the code and assign 1, 2, 3, or 4 to `names_q3` below to indicate your answer.

1. You cannot add the letter x with 2.
2. The `print` operation is meant for displaying values to the programmer, not for assigning values!
3. Python doesn't want `y` to be assigned.
4. What error message?

```
In [ ]: names_q3 = ...
```

4. Job Opportunities & Education in Rural India

A study (http://www.nber.org/papers/w16021.pdf) at UCLA investigated factors that might result in greater attention to the health and education of girls in rural India. One such factor is information about job opportunities for women. The idea is that if people know that educated women can get good jobs, they might take more care of the health and education of girls in their families, as an investment in the girls' future potential as earners. Without the knowledge of job opportunities, the author hypothesizes that families invest less in their women, and instead, invest in their men.

The study focused on 160 villages outside the capital of India, all with little access to information about call centers and similar organizations that offer job opportunities to women. In 80 of the villages chosen at random, recruiters visited the village, described the opportunities, recruited women who had some English language proficiency and experience with computers, and provided ongoing support free of charge for three years. In the other 80 villages, no recruiters visited and no other intervention was made.

At the end of the study period, the researchers recorded data about the school attendance and health of the children in the villages.

Question 1. Which statement best describes the *treatment* and *control* groups for this study? Assign either 1, 2, or 3 to the name `jobs_q1` below.

1. The treatment group was the 80 villages visited by recruiters, and the control group was the other 80 villages with no intervention.
2. The treatment group was the 160 villages selected, and the control group was the rest of the villages outside the capital of India.
3. There is no clear notion of *treatment* and *control* group in this study.

```
In [ ]: jobs_q1 = ...
```

Question 2. Was this an observational study or a randomized controlled experiment? Assign either 1, 2, or 3 to the name `jobs_q2` below.

1. This was an observational study.
2. This was a randomized controlled experiment.
3. This was a randomized observational study.

```
In [ ]: jobs_q2 = ...
```

Question 3. The study reported, “Girls aged 5-15 in villages that received the recruiting services were 3 to 5 percentage points more likely to be in school and experienced an increase in Body Mass Index, reflecting greater nutrition and/or medical care. However, there was no net gain in height. For boys, there was no change in any of these measures.” Why do you think the author points out the lack of change in the boys?

Hint: Remember the original hypothesis. The author believes that educating women in job opportunities will cause families to invest in their women more.

Write your answer here, replacing this text.

5. Differences between Universities

Question 1. Suppose you'd like to *quantify* how *dissimilar* two universities are, using three quantitative characteristics. The US Department of Education data on UW (https://collegescorecard.ed.gov/school/?236948-University-of-Washington-Seattle-Campus) and Cal (https://collegescorecard.ed.gov/school/?110635-University-of-California-Berkeley) describes the following three traits (among many others):

Trait	UW	Cal
Average annual cost to attend ($)	13,566	13,707
Graduation rate (percentage)	83	91
Socioeconomic Diversity (percentage)	25	31

You decide to define the dissimilarity between two universities as the maximum of the absolute values of the 3 differences in their respective trait values.

Using this method, compute the dissimilarity between UW and CAL. Name the result `dissimilarity`. Use a single expression (a single line of code) to compute the answer. Let Python perform all the arithmetic (like subtracting 91 from 83) rather than simplifying the expression yourself. The built-in `abs` function takes absolute values.

In [ ]:

```
dissimilarity = ...
dissimilarity
```

6. Nearsightedness Study

Myopia, or nearsightedness, results from a number of genetic and environmental factors. In 1999, Quinn et al studied the relation between myopia and ambient lighting at night (for example, from nightlights or room lights) during childhood.

Question 1. The data were gathered by the following procedure, reported in the study. “Between January and June 1998, parents of children aged 2-16 years [...] that were seen as outpatients in a university pediatric ophthalmology clinic completed a questionnaire on the child’s light exposure both at present and before the age of 2 years.” Was this study observational, or was it a controlled experiment? Explain.

Write your answer here, replacing this text.

Question 2. The study found that of the children who slept with a room light on before the age of 2, 55% were myopic. Of the children who slept with a night light on before the age of 2, 34% were myopic. Of the children who slept in the dark before the age of 2, 10% were myopic. The study concluded that, "The prevalence of myopia [...] during childhood was strongly associated with ambient light exposure during sleep at night in the first two years after birth."

Do the data support this statement? You may interpret “strongly” in any reasonable qualitative way.

Write your answer here, replacing this text.

Question 3. On May 13, 1999, CNN reported the results of this study under the headline, “Night light may lead to nearsightedness.” Does the conclusion of the study claim that night light causes nearsightedness?

Write your answer here, replacing this text.

Question 4. The final paragraph of the CNN report said that “several eye specialists” had pointed out that the study should have accounted for heredity.

Myopia is passed down from parents to children. It's reasonable to suppose that myopic parents are more likely to leave lights on in their children's rooms than other parents. In what way do you think this might have affected the data?

Write your answer here, replacing this text.

7. Studying the Survivors

The Reverend Henry Whitehead was skeptical of John Snow’s conclusion about the Broad Street pump. After the Broad Street cholera epidemic ended, Whitehead set about trying to prove Snow wrong. (The history of the event is detailed here (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1034367/pdf/medhist00183-0026.pdf).)

He realized that Snow had focused his analysis almost entirely on those who had died. Whitehead, therefore, investigated the drinking habits of people in the Broad Street area who had not died in the outbreak.

What is the main reason it was important to study this group?

1) If Whitehead had found that many people had drunk water from the Broad Street pump and not caught cholera, that would have been evidence against Snow's hypothesis.

2) Survivors could provide additional information about what else could have caused the cholera, potentially unearthing another cause.

3) Through considering the survivors, Whitehead could have identified a cure for cholera.

In [ ]:

```
# Assign survivor_answer to 1, 2, or 3
survivor_answer = ...
```

Note: Whitehead ended up finding further proof that the Broad Street pump played the central role in spreading the disease to the people who lived near it. Eventually, he became one of Snow’s greatest defenders.

8. Submission

Once you're finished, follow the instructions at the top of this notebook to save as a .pdf and .ipynb. Then submit the two files through Canvas.