Raymond Lin
304937942

Homework 1

1. **Instructions**
    a. Study Partners: Derek Hu

Raymond Lin
304937942

2. **Data Collection With Transit Tweets**
    a. One issue with the approach is that the data collected will not be representative of the entire population of people who ride the Los Angeles Transit system. This is because the data collected is only from Twitter users. Most likely, not all transit riders will be Twitter users. If we take a sample from only riders who use Twitter, it may cause a bias in the data. An example is that Twitter users are generally young adults, who may have an opinion about the transit system that is not representative of the whole population. Additionally, if we only collect data from Tweets, it will be biased, because people who Tweet about a particular topic usually have some existing opinions with the topic. For example, people will only Tweet about the transit system if they have an issue with it. Thus, this will skew the data towards people who have an opinion about the LA transit system.

3. **Model Extensibility**
    a. No, it would not be a good idea to use the model on Brazil’s population. The model based on the UK population worked well on the USA population, because both countries have similar characteristics or features. We consider factors like socioeconomic status of the population, race and genetics, culture etc… Both are first-world countries, have a majority white population, and have similar culture in terms of food consumed etc… In contrast, Brazil is not a first-world country. Brazil does not have a majority white population, but instead, it is split somewhat evenly between whites, blacks and hispanics. Finally, Brazil’s culture differs from that of the US and UK, and people living in Brazil may be consuming different food. There may be a correlation between food eaten/quality of food and Breast Cancer rates. There may be more factors that contrast the US and UK women population from that of Brazil. Because of these factors, we cannot use the model to accurately predict breast cancer in the Brazilian population.

4. Experiment Design

a. Here is a list of features that we could use to predict whether or not a student will stop attending lecture:

   i. year in school
   ii. professor for course
   iii. student interest in course content/material
   iv. major requirement that the course satisfies
   v. student's history of attendance
   vi. GPA
   vii. distance traveled to reach lecture hall
   viii. time of lecture

b. First, we would use a binary classification to label student attendance - 1 for "stopped attending class" and 0 for "didn't stop attending class". Then, we would classify a student as 1, if they failed to attend more than 80% of their classes within a given time period. For example, consider a class with two lectures a week. Within three weeks (6 lectures), if the student misses 5 or more of the lectures, he or she will be labeled as a 1. Even if the student decides to return to class and improve attendance, it will not change the label, because the student stopped attending lectures during a given period of time. We can collect this data by taking roll for the particular class.

c. For many of the features, we would have the survey the students. Features such as GPA, major requirement, and history of attendance are confidential and thus must be asked directly to the student. Subjective information, such as student interest in course content and distance traveled, must also be asked directly. The only features we can publicly obtain would be the professor teaching the course and the time of lecture. This information can be found on the university course offerings webpage or catalog.

5. **True or False**
    a. False. If there is no existing dataset on a particular question that we would like to find answers to, we must do some data collection and gathering.
    b. False. While Python is popular among data scientists for its extensive data processing and visualization libraries, there are many other tools that data scientists will find useful. For example, R is another popular language for data scientists.
    c. False. While developing new models is part of a data scientist's job, most of their time is actually spent exploring the data and cleaning it.
    d. True. If there is a bias in the historical data, and we use that data to make a prediction, the prediction will also be biased.
    e. False. We are categorizing each income range as a number, not using the actual number itself. Thus, the data is categorical.

6. Probability

a. $P(X = 0) = \frac{3}{6} \times \frac{2}{5} = \frac{1}{5}$

b. $P(X = 0, Y = 1) = (\frac{1}{6} \times \frac{2}{5}) + (\frac{2}{6} \times \frac{1}{5}) = \frac{2}{15}$

7. **Imputation**
    a. There are many ways to impute missing data, such as using median, regression imputation, hot-deck imputation, and cold-deck imputation.
        i. Median: We can replace missing data with the median value of the particular feature across the entire dataset. This is a good way to impute data because it is fast and easy. However, it is inaccurate if the distribution of the data is very sparse. It is possible that we impute the value with the median, but in reality, the value was supposed to be in one end of the distribution, far from the median.
        ii. Regression: We can replace the missing data with a value that is predicted using the other features of the dataset. We can train a model that predicts the value of the missing feature and then use this model to predict the missing value based on the existing features. The advantages of this method are that it takes into consideration the correlation between the other features and the missing feature. This will result in an accurate predicted value that follows a particular trend. However, this method does not include the error associated with prediction. Predicted values will fall directly on the linear regression line without variance. This results in the model to be overfitted and over identifies relationships.
        iii. Hot-deck: We can impute data by selecting another random record in the dataset that has similar feature values. Then, we can simply replace the missing value with that record’s value. Hot-deck imputation is used because it is simple and accurate for the most part. However, it has some limitations. For example, it is possible that there are no other records in the dataset that closely matches the one with the value that we need to impute. Then, we cannot use this method for imputing data.
        iv. Cold-deck: This method is the same as hot-deck imputation, but instead of choosing from our own dataset, we choose from an external source. It has the same benefits as hot-deck imputation. However, its drawbacks are that there may not be an external source with such information, or such data is difficult to obtain.