Cox Deliverable 1

CIS 368

Data Dictionary:

```
- gender: Male or Female response from participant
- age: Age in years of the individual
- avg_glucose: Average glucose levels of the individual
- bmi: Body mass index of the individual
- smoking_status: is the individual a former smoker, never smoked, or currently smokes
- stroke_poss: is the individual likely to have a stroke
```

Problem Statement:

How likely are males and females likely to have a stroke based on their smoking status, average glucose levels, and their body mass index.

Data Clean-up:

For the data set there are more categorical variables than numerical ones. So I only wanted to use a few, such as smoking status and gender, as to not have too many variables that could skew the results. I will take smoking staus and gender and assign them numbers to get a more accurate spread on information. Since there is over 4,000 entries I will remove those that have any null values and still have enough entries remaining to do the exploratory data analysis. I will also use a standardize the data to help make the data more accurate. Usually I would say age is a categorical variable but since the data has the exact age and not a bracket range where their age is I see it as a useful numerical variable.

Predictor Variables:

```
- gender
- age
- avg_glucose
- bmi
- smoking_status
```

Target Variable:

```
- stroke_poss
```

In [ ]: