**Computational Thinking Project**

Hunter Lawless, Zack Lasek, Max Stevens

**Problem**

We want to better understand the effects of diabetes and the groups of people who are more likely to get diabetes. This includes seeing the breakdown of age group in the data and preexisting conditions that lead to being diagnosed more frequently, as well as what makes someone more likely to be diagnosed with diabetes, such as being obese or not.

**Data**

The original data has 11 columns and 768 rows in it. The data is about diabetes, factors that go into the likelihood of having diabetes, and some age description. The columns are “Glucose”, “BloodSugar”, “BloodPressure”, “SkinThickness”, “Insulin”, “BMI”, “Obesity”, “DiabetesPedigreeFunction”, “Age”, “AgeGroup”, and “Outcome”. The first 6 columns in that order go into determining the patient's likelihood of having diabetes. The next column is the likelihood of the patient having diabetes. The next two columns are about the patient's age and which age group they fall into. Finally, the last column is the outcome, which is if the patient does or does not have diabetes. The data in the columns are either a number (like Glucose and BloodPressure) or a yes or no question that is represented as a 1 or 0 (like AgeGroup and Outcome), respectively.

The original data was relatively simple and didn’t need many changes and preprocessing done on it since there are only 11 columns and 768 rows. However, we felt the “SkinThickness” column was unnecessary, so we removed it in the code using the drop method in Pandas before the function. Additionally, we decided to drop the “Insulin” column in the same code using the same method because it had a lot of missing values that were represented as just 0s in the data. We felt that a column with that many missing values was a poor column to keep and use in the data. Therefore, we dropped those two columns and that was the only preprocessing of the data that was needed and done within the code in Python. We now have 9 columns in the data.

**Questions**

1. What percentage of patients in the data have diabetes?
2. What is the distribution of the age group for the diabetes data?
3. Is there a correlation between BMI and Diabetes Pedigree Function?
4. Which age group is more likely to be diagnosed with diabetes?
5. Does a preexisting condition, like obesity, make someone more likely to be diagnosed with diabetes?

**Functions**

The input for the functions of the data, which are the 4 graphs that we coded and displayed, is the data on diabetes. The 1st graph/function is a breakdown of our data and showing it visually into the ratio of patients with diabetes to patients without. It uses the data in the “Obesity” column of the data. The 2nd graph/function is a second breakdown of the data. It is the ratio of respondents who are categorized as adults versus young adults in the data. This function uses the “AgeGroup” column of the data to code the graph. The 3rd graph/function is a scatter plot showing if there is a correlation between BMI and Diabetes Pedigree Function using the “BMI” and “DiabetesPedigreeFunction” columns of our dataset to produce the plot output. The 4th graph/function is a bar graph comparing the percentage of adults and young adults who have diabetes. This function is using the “AgeGroup” and “Outcome” columns of the data to produce the graph output. The 5th graph/function is comparing the percentage of obese and not obese people who have diabetes. This function is using the “Obesity” and “Outcome” columns to output the bar graph.