The topic that our team decided to pick for our project is “Predictive Indicators of Heart Disease”. We found a dataset on the data science platform Kaggle which gave us the idea for our project as well as the datasets that we analyzed. This dataset was provided originally by the CDC and was gathered by doing telephone interviews with nearly 320,000 individuals. After reading in the CSV and viewing the variables that our dataset contained we came up with 4 research questions:

**What age group is heart disease most prevalent?**

Our team viewed our age variable and found that there were 13 different age categories starting at 18 and going to 80 or older. We hypothesized that individuals in our oldest age category (80 or older) would have the highest population of heart disease as it makes logical sense that a persons chance of having heart disease would increase as they age. To test our hypothesis, we made a series of graphs showing our age categories and the amount of individuals with and without heart disease. Our hypothesis was ultimately correct that the 80 or older age category had the highest number of individuals with heart disease. To take our hypothesis a step further, we calculated the correlation between age and presence of heart disease and found a correlation coefficient of .23%. This would indicate that while our dataset is showing a positive correlation between age and heart disease, it is a relatively weak one.

A graph of heart disease

Description automatically generated

A graph with numbers and a number of people

Description automatically generated with medium confidence

**Does the amount of sleep that someone gets correlate with the presence of heart disease?**

The sleep variable in this dataset was simply measured by asking respondents how many hours they typically slept in a given night. Our team hypothesized that there would be a negative correlation between heart disease and sleep. Specifically, we thought that the fewer hours someone slept, the more likely they would have heart disease. We began analyzing this question by forming two data frames: one that captured mean values of individuals with heart disease and one that captured mean values of individuals that did not have heart disease. The group with heart disease had an average sleep time of 7.14 hours per night and our group without heart disease had an average sleep time of 7.09 hours per night. This would suggest that our hypothesis is not correct as the group that has heart disease actually slept more on average than our group without heart disease. We wanted to investigate this relationship further, so we calculated the correlation between sleep time and presence of heart disease and got a correlation coefficient of .0083%. This would imply that there is next to no correlation between sleep time and the presence of heart disease.

A graph with numbers and lines

Description automatically generated with medium confidence

**Are people with higher BMI more likely to have heart disease?**

BMI, or Body Mass Index is a measure of body fat based on a person’s weight in relation to their height. We hypothesized that there would be a positive relationship between BMI and the presence of heart disease. Specifically, we thought that the higher an individuals BMI, the greater the chance they had heart disease. We began analyzing this question by forming two data frames: one that captured mean values of individuals with heart disease and one that captured mean values of individuals that did not have heart disease. The group with heart disease had an average BMI of 29.40 and the group that did not have heart disease had an average BMI of 28.22. This would suggest that our hypothesis is correct as the group with heart disease has a higher BMI on average. We wanted to investigate this relationship further, so we calculated the correlation between BMI and the presence of heart disease and got a correlation coefficient of .0518%. This indicates that while there is a small positive correlation shown between BMI and presence of heart disease, it is not a statistically significant correlation.

A diagram of a number of bmi distribution

Description automatically generated

A chart of a group of people

Description automatically generated with medium confidence

**Which lifestyle choice correlates higher with the presence of heart disease? Smoking or drinking alcohol?**

Our team chose this question because we wanted to see the impact of lifestyle choices on the development of heart disease. Our team discussed that while alcohol is dangerous in large amounts, there are studies that point to the idea that small amounts (a glass of wine) can actually be good for our heart. With this in mind, we hypothesized that smoking would correlate higher with the presence of heart disease than drinking alcohol would. We began analyzing this question by creating a data frame that showed individuals with heart disease and whether they reported smoking or drinking alcohol. Of the 27,373 people with heart disease in our data set, we found that nearly 59% of them reported smoking and only 1,141 of them reported drinking alcohol. While our team felt this was a significant finding, we wanted to determine how significant it was by calculating the correlation coefficient between the presence of heart disease and smoking / drinking alcohol. We found that smoking had a correlation coefficient of .11% which would indicate a positive, although weak, association between smoking and the presence of heart disease. Interestingly enough, we found that drinking alcohol had a correlation coefficient of -.03%, suggesting that there is a very weak negative correlation between drinking alcohol and the presence of heart disease. These findings would indicate that although the relationship is not very strong, statistically speaking, our hypothesis is correct in that smoking does correlate higher with the presence of heart disease than drinking alcohol.

A graph of different colored squares

Description automatically generated

A graph with blue and orange lines

Description automatically generated

**Limitations of our Data**

The biggest limitation of our dataset that we found is the sheer number of respondents without heart disease compared to the respondents with heart disease. We had roughly 300,000 respondents that did not have heart disease and 27,373 respondents with heart disease. We believe that a more balanced dataset might’ve yielded higher correlation values. Another limitation of our dataset is that researchers from the CDC mostly called people with landlines, they collected minimal data by calling cell phones. We believe that this potentially skewed our data towards older people as well as homeowners who are more likely to have a landline. In this dataset, we also had a significant amount of outliers in our BMI as well as sleep time which skewed our mean values higher. A reason we might’ve seen this is that phone interviews are self-reported surveys where we do not necessarily know the respondent is telling the truth. For instance, there were several people who reported sleeping 24 hours a day as well as people who reported sleeping 0 hours per day. Lastly, we found that of the respondents that had heart disease, we only had 4.1% of individuals that reported drinking alcohol while we had nearly 59% that reported smoking. Our belief is that these values most likely do not reflect the true population as the National Survey on Drug Use and Health reported that 69.5% of Americans drank alcohol in the last year.