q1. Is there a correlation between alcohol consumption and heart disease?

We conducted a two sample t-test of AlcoholDrinker/Non-Drinker & heart disease. The t-statistic of 18.1505142963367 indicates a substantial difference between the means of the two sample groups. A higher t-statistic suggests a larger difference between the sample means. However, the P-value is 1.38 and is greater than alpha (0.05) so we fail to reject the null hypothesis and there is no significant difference between the two groups. A limitation for the column of data for alcohol was that it was categorical and only about 7% of the dataset drank alcohol and that's not representative of the general population trends for alcohol drinking according to the CDC since about one third of the population doesn't drink alcohol, about half lightly drinks and a fifth are moderate to heavy drinkers. Having a column of data that is numeric and measures levels of alcohol consumption would help to give a more accurate representation of the correlation between alcohol and heart disease. The data could also be broken down by the type of alcohol consumed in order to test for differences by beer, liquor or wine may have on heart disease or other health outcomes.

q2. How does heart disease differ by sex?

We conducted multiple charts on the large data set where we had split the data up first by sex and then further more by heart disease categories. When you looked at the initial graph it did show a difference in males and females heart disease rates and that men did have a higher chance of heart disease. We then further broke the data down by specific categories such as stroke, smoking, alcohol, difficulty walking, health, etc. These graphs further showed that males had a higher chance of cardiovascular disease than females.

q3. How does body fat (BMI) / level of inactivity influence heart disease?

One dataset did not include BMI therefore code was written to calculate BMI using a person's height and weight. Then using 'bins' in python code, BMI classifications were created for the datasets. For all datasets, a higher percentage of the population that had heart disease had BMI classifications in the obese categories. Interestingly, the data from the CDC / BRFSS surveys showed that people in underweight BMI classifications also had a higher percentage of heart disease than people classified as normal weight (BMI). An analysis was performed looking at the relationship between the BMI classifications and whether a person was active in the population of those who had heart disease. The BRFSS 2022 dataset, which had a relatively small percentage of

those with heart disease, clearly showed that those underweight and overweight had lower levels of activity on average. The Kaggle dataset showed a similar trend, but not as pronounced. Since the data collection for all datasets asking respondents their level of activity was binary: either yes active or no inactive, Venn diagrams were created for the BRFSS 2022 and Kaggle datasets to determine whether any trends were present. In summary, there was no trend because with the BRFSS data, one third of those with heart disease was inactive, whereas in the Kaggle population, only one fifth were both inactive and had heart disease. However with the BRFSS data, there were significantly more people that were inactive and did not have heart disease.

q4.Relationship Between Daily Food Consumption and Likelihood of Developing Heart Disease

We have taken the dataset from Kaggle. The dataset contains the data with general health conditions for different age groups and the different factors that affect heart disease. Taking into consideration how the fruits intake, veggies intake and the potato intake effect the heart health created the visualizations using the histograms. The data showed the population with fruit intake and veggie intake with heart disease has slight decrease in heart disease. While the population with the potato intake has slightly more risk of heart disease. Significantly the data showed the food has less impact developing the heart disease or reducing the heart disease in comparison with other factors.

https://www.kaggle.com/code/poshanbelbase/cvd-prediction-mutiple-model-comparison/notebook

q5. Is smoking the leading cause of heart disease?

We conducted a two sample t-test of Smoker/Non-smoker & heart disease. The t-statistic of -61.297861168863285 indicates a substantial difference between the sample data and the population data. The magnitude of the t-statistic suggests a significant discrepancy in means. The P-value was 0.0 and suggests strong evidence against the null hypothesis. With a t-statistic of -61.297861168863285 and a p-value of 0.0, we reject the null hypothesis. The results are highly statistically significant, indicating that the observed difference between the sample and population means is not due to random chance. The findings support the alternative hypothesis, suggesting that there is a real and significant difference between the smoker and non-smoker data. This data was limiting however since it was categorical and only gave Yes or No for responses, this could be circumvented by providing data for consumption levels through amount smoked a given week or day in order better test how closely correlated smoking may be to the presence of heart disease.40% of the dataset answered Yes to smoking & ‘Smoking’ is also a wide ranging category than can include cigarettes, E-cigs and pot. The kind of smoking ought to be specified in a separate column along with the

consumption level given the rate of cigarette smoking has gone down over the years and pot has been legalized in many states so ‘smoking’ ought to be specified. So while we did find that there was a statistically significant correlation between smoking and heart disease, specifying what is smoked and the consumption levels would give a better representation to see correlations between use of a substance and the presence of heart disease.