ALP Data Science Final Project Report

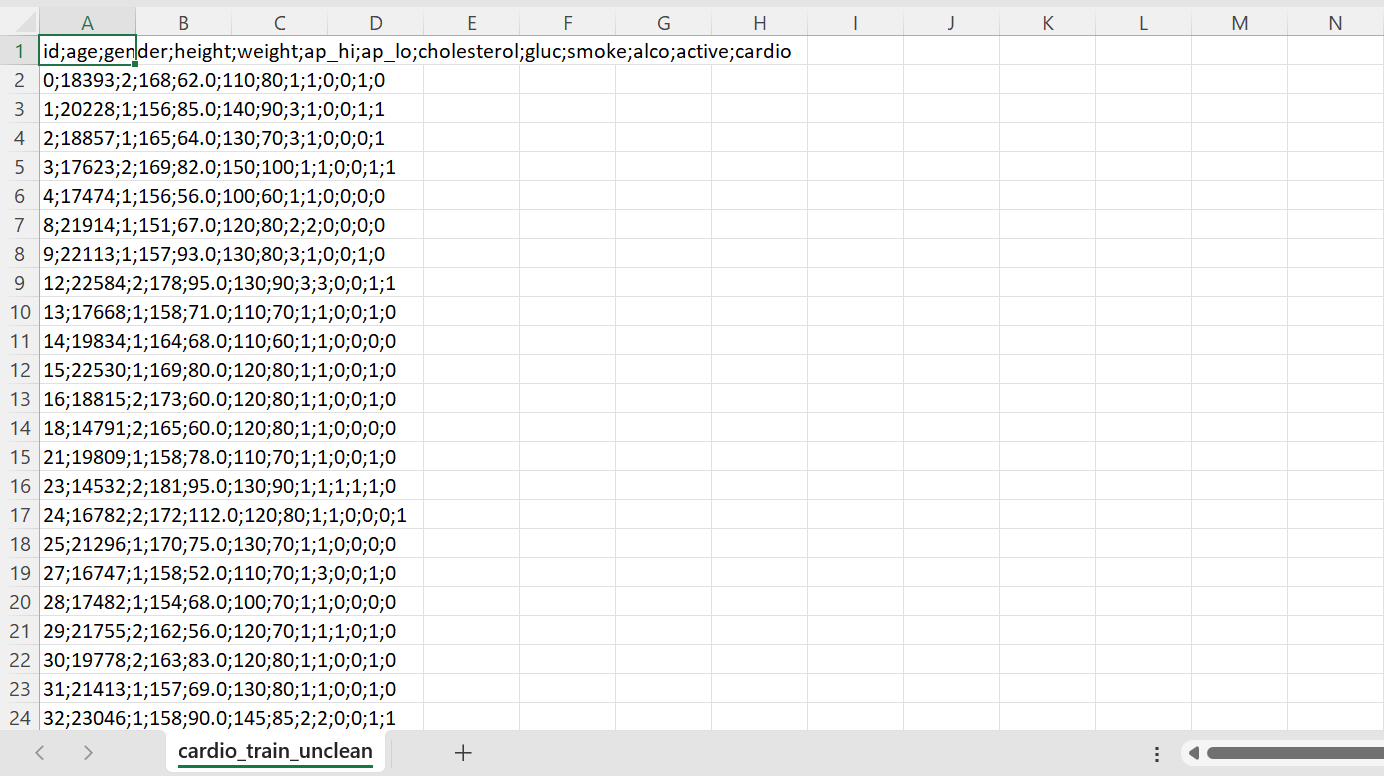
By Lewen Feng and Srikar Munagavalasa

**Abstract:**

Data science is the study of data to extract knowledge and insights. Data scientists are in charge of organizing the mountains of data generated each day into clear and concise information, helping organizations make decisions to improve effectiveness and efficiency. The impact that data scientists have on the world today along with our combined passion for the medical field and computer science, inspired Srikar and I to find, analyze, and produce information from a practice cardiovascular dataset ourselves. In this project, we analyzed how gender, smoking, alcohol consumption, physical activity, and cardiovascular diseases affected one’s blood pressure, weight, cholesterol, and glucose levels. Our goal for this project was to analyze the statistics of four categories: blood pressure, weight, cholesterol, and glucose to learn what aspects of life affect the heart the most.

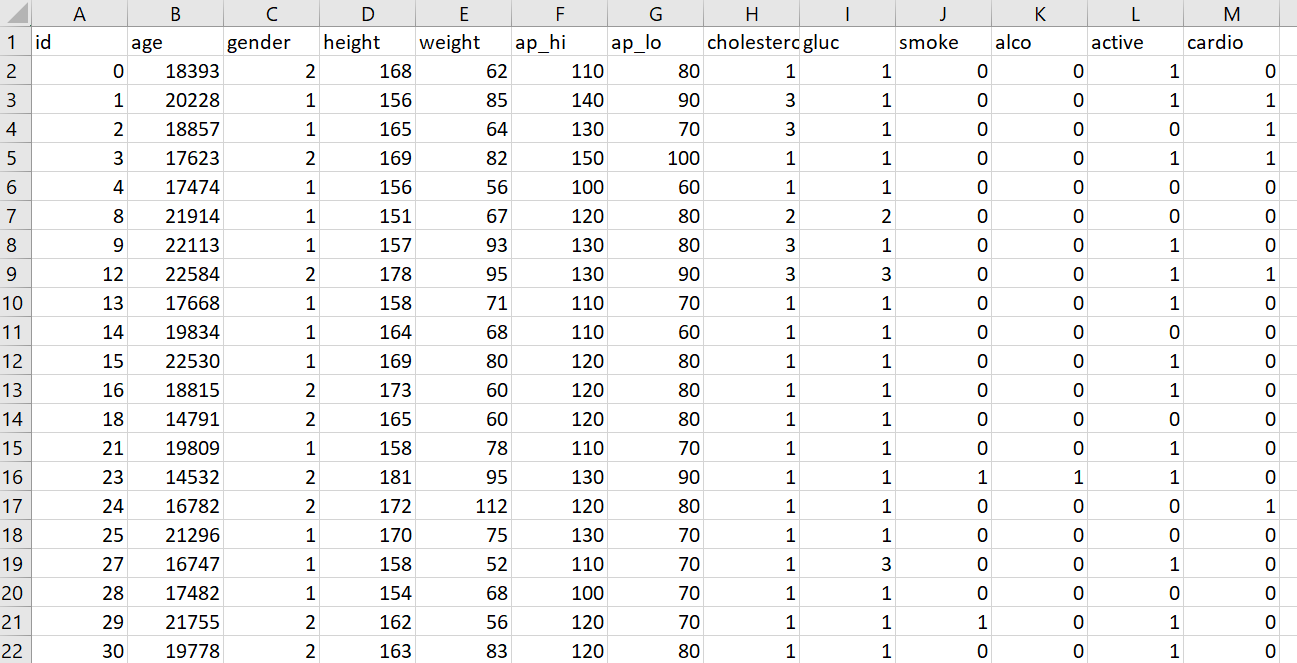
**Methodology:**

After deciding to research heart diseases and conditions, we searched on Kaggle, an open-source online platform that provides tons of free data sets. After discovering a practice dataset with over 70,000 records of patient data on 12 features, we ran into an issue. Each feature’s data was separated by semicolons and we needed to clean it.



*Caption: The data was formatted by semicolons in a single cell instead of splitting the data into individual cells across many columns.*

We cleaned the data set using Excel’s built-in tool, Power Query, which allowed us to split cells based on delimiters. Delimiters are characters that are used to denote the bounds between regions in a text (e.g. commas and periods in English or parenthesis and lines in Python). Then we considered using SQL, or structured query language to extract data. SQL is a programming language that is designed specifically to access, store, and process information from databases easily. However, we quickly encountered numerous issues with the Replit databases, Replit PostgreSQL, SQLite library, and relational database management systems (RDBMS) like Amazon AWS or Microsoft Azure. After consulting the instructors for help, we found another built-in Excel tool, Pivot Table. Pivot Table had the same capabilities as SQL, being able to extract data, but it could also visualize it into a table and analyze it using Excel’s functions. Lastly, we pasted the analyzed data collected from Excel into a Google Spreadsheet since it offered cleaner graphs and a fresher look.

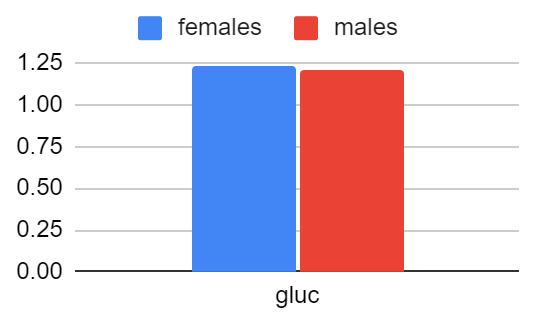
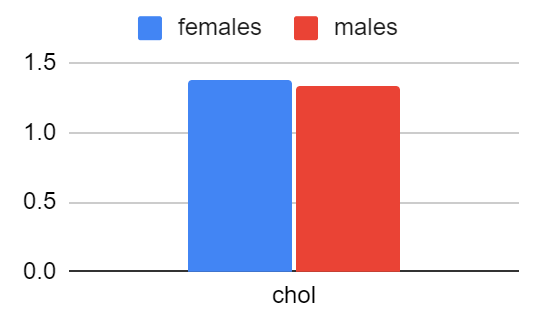
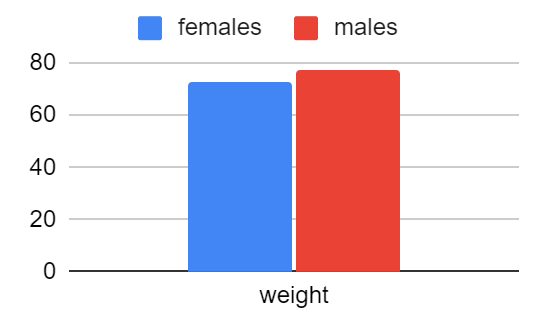
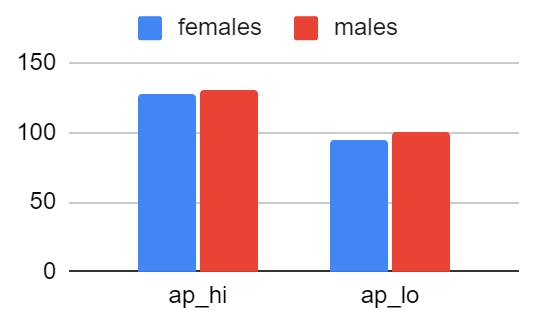


*Caption: The cleaned data is formatted correctly through the rows and columns, making it easier to read and analyze.*

**Important Notes:**

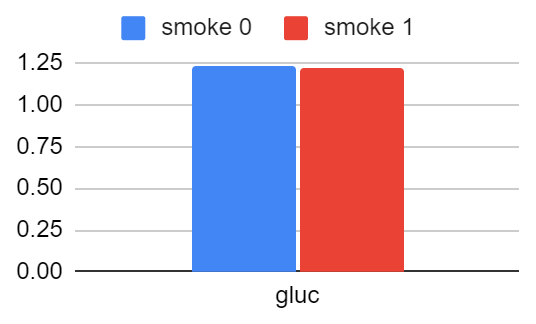
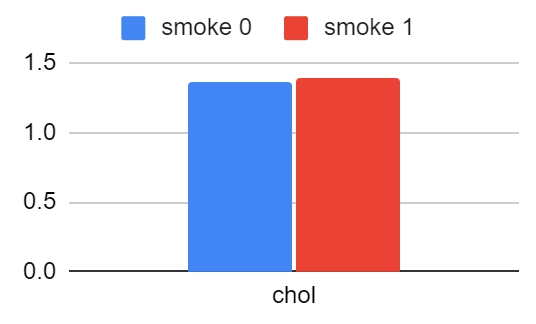
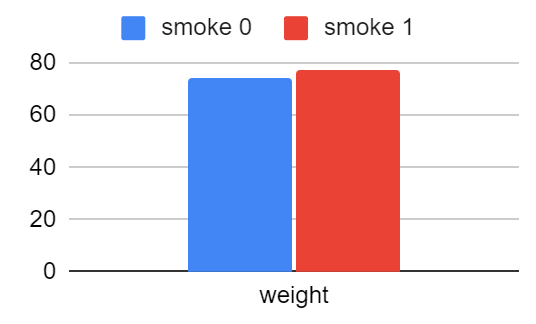
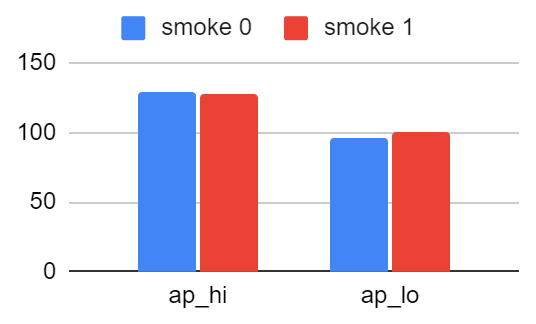
Cholesterol and glucose levels are measured on a scale of 1 to 3, with 1 being normal levels and 3 being well above normal. Smoking, alcohol consumption, physical activity, and cardiovascular diseases are marked with 0 and 1. Blood pressure is separated into ap\_hi (systolic blood pressure) and ap\_lo (diastolic blood pressure). Systolic pressure is the pressure caused by a contracting heart, pushing blood through the body. Diastolic pressure is the pressure caused by a relaxed heart, filling the heart with blood. Systolic pressure is higher than diastolic pressure because blood is being pushed into the body instead of sucked out of it. These blood pressure levels are measured with a high and low of mmHg (millimeters of mercury or roughly 133.322 pascals). Weight is measured in kg (kilograms).

**Gender:**



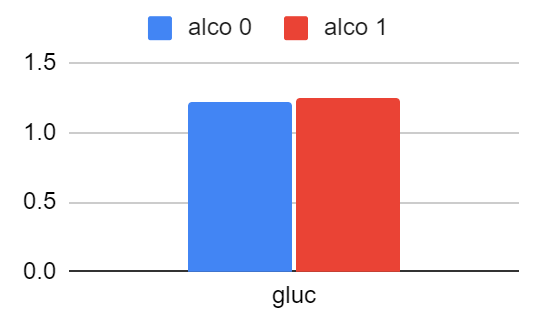
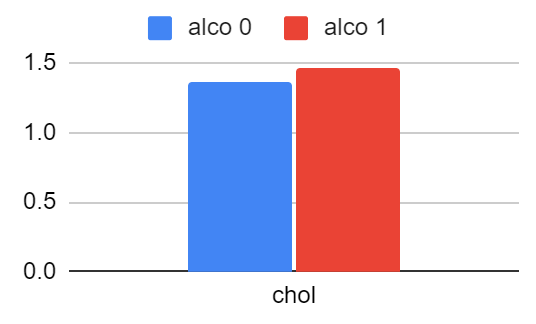
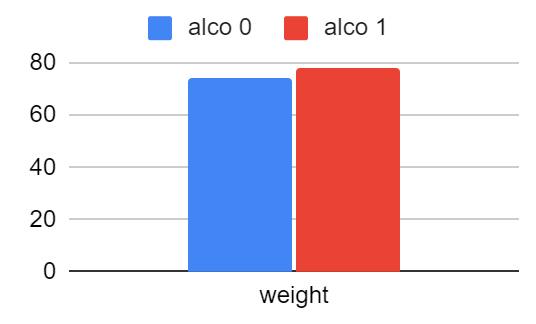
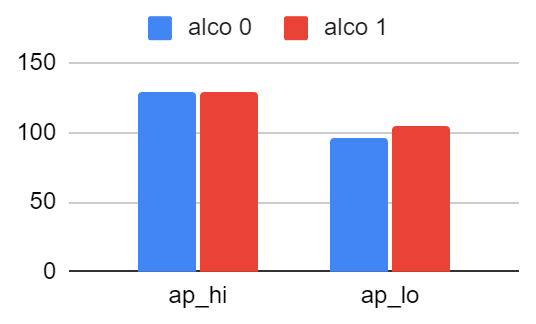
These four graphs show the differences in average statistics of males and females, which are sometimes surprisingly substantial. Most notably, males, on average, are 4.69kg heavier than females. The blood pressure difference is significant as well, with the average systolic blood pressure being 1.94 mmHg (millimetres of mercury) higher and the average diastolic blood pressure being 6.03 mmHg higher for males. Although the first two graphs provide sizable differences, the latter two show the opposite, with females being only 0.05 and 0.02 higher for cholesterol and glucose levels, respectively. Trends were expected and went along with our hypothesis. Males have more testosterone, which increases blood pressure, and have a 33% larger heart than females. Men also are, on average, taller, have more muscle mass, and have a higher bone density, leading to a heavier weight.

**Smoking:**

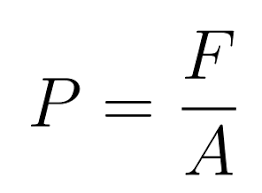


These four graphs display the cardiovascular difference between smokers and non-smokers. The average systolic and diastolic blood pressures are nearly the same ( non-smokers being 0.50 mmHg higher and 3.45 mmHg lower respectively) and smokers are slightly heavier by 3.44kg on average. Similar to the male and female comparisons, the cholesterol and glucose levels changed infinitesimally, with a 0.02 and 0.01 difference. This caught us by surprise as we hypothesized that smoking would have a massive increase in one’s blood pressure and weight. However, nicotine, an addictive chemical common in tobacco products, mainly harms the pathogens, body acids, calcium, and alveoli (tiny air sacs of the lungs that perform rapid gaseous exchange). If we analyzed a dataset testing the lungs, digestive organs, and bones, smoking would show a remarkable effect. Although this analysis yields fewer results, smoking is still a hazardous habit that can have devastating effects.

**Alcohol Consumption:**



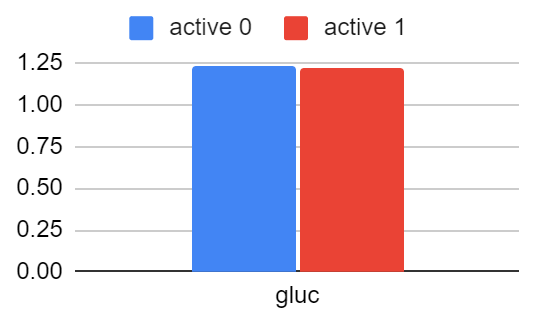
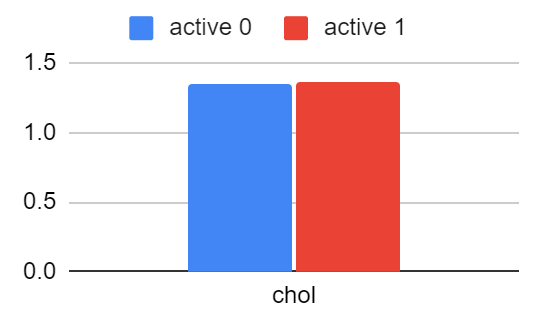
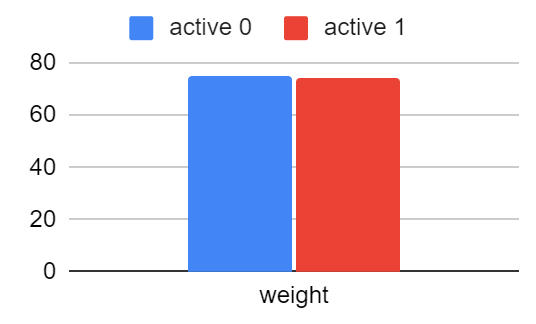
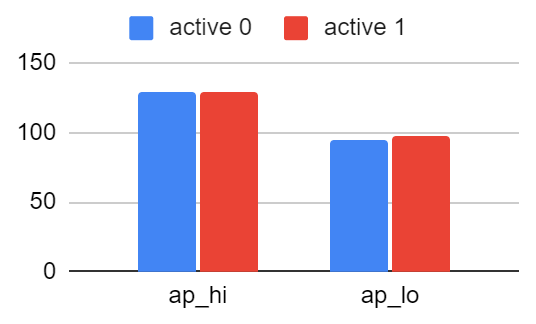
The four graphs above compare the effect of alcohol consumption. People with high alcohol consumption had a 0.96 mmHg higher average systolic blood pressure and an 8.86 mmHg higher average diastolic blood pressure, a massive difference. Alcohol consumption also leads to a 4.28 kg weight gain. The cholesterol and glucose levels have seen a large rise, about 0.11 and 0.03 respectively. We correctly hypothesized most of these results except for the weight increase, which was unexpected. This is because alcohol consumption increases the calcium levels that are connected to the blood vessels. These blood vessels are overwhelmed by the calcium, causing them to constrict and not allow blood and oxygen to travel smoothly. Pressure builds up when blood vessels constrict due to the pressure formula.



*Caption: If the area decreases, the fraction becomes greater, and therefore the pressure increases.*

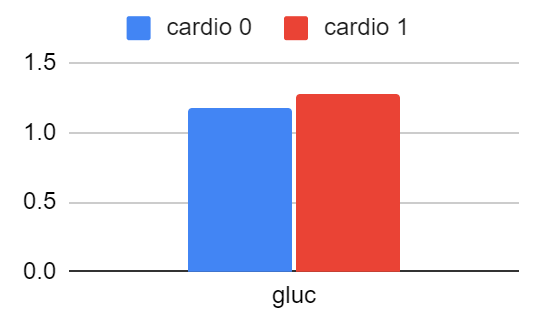
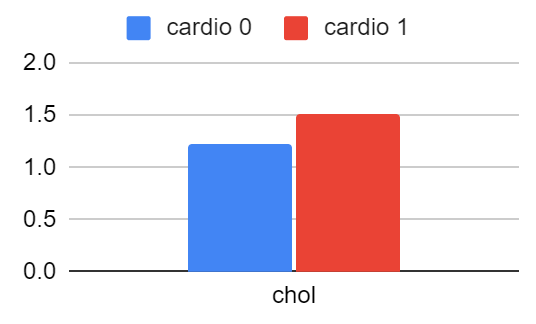
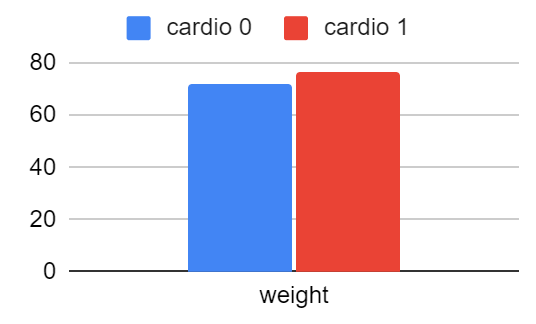
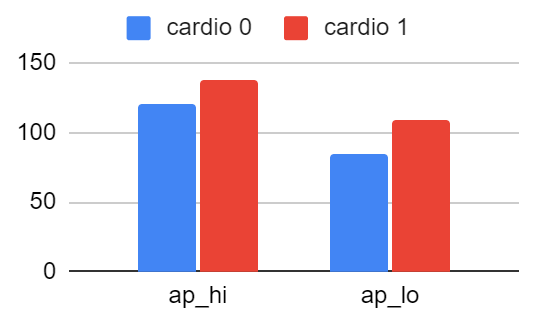
Similarly, regular alcohol consumption can make people overweight and raise cholesterol levels since alcohol is broken down in the liver and reconstructed into cholesterol, triglycerides, and other fats.

**Physical Activity:**



The 4 graphs above examined people who participate in physical activity and those who do not. The average systolic blood pressure of both populations differed by 0.01 mmHg, whereas the average diastolic blood pressure of active people was 2.27 mmHg higher. Surprisingly the weight is relatively the same, only 0.61 kg less. Akin to most of the other tests, the cholesterol and glucose levels increased by 0.02 and decreased by 0.01, respectively. Overall, being active had a very minor change in the grand scheme compared to the other tests. The major surprise was how little influence physical activity had on people’s weight. A potential explanation for this may be that as one exercise, they lose fat but gain muscle, which has a 15% higher density. This means that, despite losing fat, the muscle that active people gain counterbalances that effect on weight. While weight may not change much due to this reason, losing body fat and gaining muscle has various long-term benefits for other aspects of the body.

**Cardiovascular Diseases:**



The final 4 graphs we created were on people who had cardiovascular diseases to healthy people. Cardiovascular diseases had the most impactful outcome in every field. The average systolic blood pressure had an enormous margin of 16.78 mmHg and the average diastolic blood pressure was an even greater increase of 24.77 mmHg. In addition, the weight drastically changed, increasing by 5.23 kg. The cholesterol levels changed by an all-time high in the tests we did going up by 0.30 and the glucose levels increased by 0.10. We expected cardiovascular diseases to have some effect on the features we studied, but we did not predict this vast of an impact. Cardiovascular diseases were, by far, the biggest factor in one’s blood pressure, weight, cholesterol, and glucose levels. This is due to the symptom of blood clots, because they narrow the artery, giving blood and oxygen flow several challenges. Narrow arteries cause the pressure inside them to build up due to the pressure formula (refer to the alcohol consumption section). It also explains the increase in the other three values because blood clots build up fatty acids in blood vessels which hold fat, cholesterol, and glucose. Although the consequences of cardiovascular diseases are disastrous, one can get tested and treated to lessen the effects.

**Conclusion:**

The purpose of this project was to analyze the four different categories, blood pressure, weight, cholesterol level, and glucose level. The first test was the gender test, which showed the differences in weight (due to the average male’s more muscular mass, denser bones, and taller body). The blood pressure also changed due to males having a bigger heart to pump blood and have more testosterone. The next test, smoking, yielded a slight increase in the average diastolic blood pressure and a larger increase in weight. This was because smoking primarily affects the lungs and less of the heart. The third test, alcohol consumption, had mostly normal graphs, showing exceptional harm (specifically the weight). Alcohol has several negative repercussions on the body by introducing an excess amount of something that would be beneficial under normal conditions and in turn making it dangerous. The least effective out of the tests was the physical activity test since the tested populations differed very little. The strangest result was the lack of effect on weight, but burning fat while gaining muscle could be an explanation for this. Finally, the test with the greatest effect on all fields was the cardiovascular diseases test, which had the greatest influence on every field. Overall, this experience of analyzing a practice data set helped us dive into the field of data science. Srikar and I enjoyed the process and look forward to continuing our journey as data scientists by working on future projects.

Statistics Spreadsheet:

<https://docs.google.com/spreadsheets/d/1D7RyBPR4DJaZ0neinXetLeDFcrGD9xX_Rl2-qc9zc8U/edit?usp=sharing>