Project Summary

Heart attacks are a medical emergency that occurs when a part of the heart is not receiving enough blood to function. The more time the blockage is not treated, the greater the damage caused to the heart muscle. There are many factors that can affect the risk of a heart attack in a subject. In this project, we will be using a dataset to determine which factors increase a subject's risk of a heart attack occurring.

The dataset can be found here: https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/data. This dataset is a synthetic creation to simulate a realistic experience, with the aim of fostering learning and experimentation to encourage a deeper understanding of data analysis and interpretation. Though this dataset is not real data, this project will be written under the pretense that it is to simulate a real-world scenario.

Key Questions

I hope to solve some key questions related to how different factors affect the risk of a heart attack. Even though there is data on many different factors, this project will not be using all of them. First, I will conduct comparisons on two different but related factors and analyze how they affect the risk. The comparisons I conducted are:

- Sex vs Age
- Smoking status vs Alcohol consumption status
- Previous heart problems vs Family history
- Diabetes vs Obesity
- Diet vs Cholesterol levels

I selected these comparisons based on similarities between factors or their general grouping with one another. Then, I will calculate averages across different geographic location using the following categories:

- Country
- Continent
- Hemisphere

Data Preparation

The data set was already very usable and required minimal changes. First, I checked for duplicates and did not find any. Then, I found that some categories made more sense to be grouped to conduct analysis and used a nested-if statement to achieve this. Those categories were age, which I grouped into "Old" (55 and older), "Middle Age" (31 to 54), and "Young" (30 and younger), and cholesterol, which I grouped into "Dangerous" (240 and greater), "At-Risk" (200 to 239), and "Heart-Healthy" (less than 200)[1].

Approach

I created several pivot tables according to the factors I wanted to analyze. Each pivot table compares two different factors and how they influence a person's risk of a heart attack. The values of the table are the sum of patients with a risk of a heart attack of each grouping of factors. Sum was used here because a patient with a risk of a heart attack is indicated by a value of 1. Once the pivot tables were created, I then created corresponding figures to provide visualizations for the findings. I decided to choose bar graphs as I found they best demonstrated the patterns in the data. These can be found in the "Sum Pivot Tables" sheet.

Along with these comparisons, I also conducted analysis by geographic categories. I created pivot tables that calculated the average risk of a heart attack in each geographic location where data was taken. From the pivot tables, I created more bar graphs that provide some interesting insights to the data. These can be found in the "Average Pivot Tables" sheet.

Results and Takeaways

The analysis demonstrated some very interesting, and even surprising, results. The Sex vs Age comparison shows us that males are more than twice as likely to be at risk of a heart attack than females and that as you age, risk of a heart attack increases. The Smoking status vs Alcohol consumption status comparison shows us that while alcohol consumption only slightly increases the risk, smoking increases the risk of a heart attack by more than 8 times than the risk for a non-smoker. In a more surprising result (at least to me), the Previous heart problems vs Family history comparison shows to us that previous heart issues does not have an effect on risk of a heart attack, but family history does. The data tells us that those who do not have a family history

[1] Cholesterol groups are based on common guidelines for cholesterol levels

of heart issues are actually more likely to be at risk of a heart attack. This could be attributed to those that have a family history of health issues are more aware of the risk and lead healthier lives. In another surprising result, the Diabetes vs Obesity tells us that while having diabetes increases the risk of a heart attack by more than double, obesity has a minimal effect. Lastly, our Diet vs Cholesterol levels comparison tells us that diet has a minimal effect on risk. How cholesterol affects risk is very interesting in that the "Heart-Healthy" group has more risk of heart attack subjects than the "At-Risk" group, even though the "Dangerous" group has the highest number of risk of heart attack subjects. This can likely be attributed to less subjects falling into the "At-Risk" category than the other two. Nevertheless, it is very likely that the higher your cholesterol levels are, the more likely you will be at risk of a heart attack.

Our geographic data tables also demonstrate some insightful trends in the data. Heart attack risk averages by country have very little variation of the countries where data was collected. Heart attack risk averages by continent are a little more insightful with North America, South America, and Africa having the highest averages and Europe having the smallest. Lastly, the northern hemisphere has a slightly higher average of heart attack risk than the southern hemisphere. The data also tells us that the worldwide average of a subject having a risk of a heart attack is about 35.82% of the population. Overall, heart attack risk is not greatly influenced by geographic location.

This dataset and analysis has provided many insights into factors affecting risk of heart attacks. Healthcare professions and researchers can use these results to make better informed decisions and strategies to provide the best care possible. Knowing what increases risks can influence people to lead healthier and live longer lives.