UNDERSTANDING HEART DISEASE AND HOW TO AVOID IT

Smit Patel

Abstract –
Datasets are often hard to interpret based on the amount of variables and rows they usually include, but with the use of data visualization tools such as Matplotlib, Seaborn and Plotly, one can process, manipulate, and visualize the data in order to better understand and better interpret the data.

In this report, I will present the process that was gone through in order to understand the variables that are highly correlated to heart disease, and provide visualizations that can be understood by any audience to help better understand heart disease and how to keep individual health in good standing.

Table of Contents

1 **Introduction** **1**
1.1. Problem Definition 1
1.2. Research Questions 2
1.3. Suggested Solution 2
2 **Dataset** **2**
2.1. Data Description 2
2.2. Data Cleaning 3
3 **Exploratory Data Analysis** **3**
3.1. Gender and Heart Disease 3
3.2. Age and Heart Disease 4
3.3. Kidney Disease and Heart Disease .. 4
3.4. Stroke and Heart Disease 5
3.5. Diabetes and Heart Disease 5
3.6. BMI and Heart Disease 6
3.7. General Health and Heart Disease .. 6
3.8. Physical Health and Heart Disease . 7
3.9. Mental Health and Heart Disease ... 7
3.10. Smoking and Heart Disease 7
3.11. Alcohol and Heart Disease 8
3.12. Sleep and Heart Disease 8
4 **Conclusion and Future Work** **9**
5 **References** **9**

1 Introduction

1.1 Problem Definition

According to a study done by PHAC (Public Health Agency of Canada), about 1 in 12 (~2.6 million) Canadians aged 20 and older live with diagnosed heart disease. This shocking number aligns with the visualization in Figure 1, where approximately 8% of people per 100 people from the dataset got heart disease. Further, they have reported that every hour ~14 Canadian adults aged 20 and older with diagnosed heart disease die. Although these numbers are very high, heart disease is avoidable and there are many factors that can aid in this prevention. This project aims to identify what these factors are as well as the impact they have on heart disease, in order to help individuals better understand what things they can do in their day to day life to avoid getting heart disease as well as things they can do to reduce its effects if they do already have it.

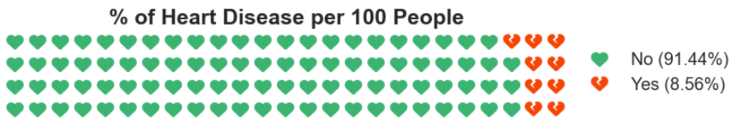

Figure 1: Waffle Chart – Heart Disease % per 100 people

1.2 Research Questions

To give a better idea of what this project is trying to achieve, a few research questions will be formulated in advance, that will be answered through data visualizations.

1. Does being a certain gender or age lead to a higher chance of getting heart disease?

2. Are those with other health issues (diabetic, asthma, kidney disease) more likely to get heart disease than those with none? If so, which of these issues is most linked to heart disease?
3. Does general health and BMI have any effect on getting heart disease?
4. Is there any correlation between physical and mental health and getting heart disease?
5. Does alcohol consumption and smoking have any effect on a person getting heart disease?
6. Is there any correlation between sleeping and getting a heart disease? Are those that sleep less still able to avoid heart disease through other factors (such as physical activity, etc.)?

1.3 Suggested Solution

Through the use of data visualizations and the obtained dataset, many of these questions can be answered and actual solutions can be presented to those that would like to improve their health and avoid health issues such as heart disease. Some techniques that will be applied include the creation of visualizations using libraries such as Matplotlib and Seaborn to show trends in the data, creation of interactive plots through Plotly that would allow a user to see how changes in factors either increase or decrease the chance of getting heart disease, and creation of plots that involve distribution and probabilities to give insight into how likely an individual is to get heart disease based on their current conditions.

2 Dataset

The dataset chosen for this analysis is an open source dataset available through Kaggle[1], which contains Personal Key Indicators of Heart Disease data. The dataset contains the 2020 annual survey data of 400,000 adults related to their health status, physical activity, and if they have health disease or not. The available data consists of 18 columns, where the columns contain features describing BMI, Alcohol Drinking Status, Smoking Status, Stroke Status, Physical Health Status, Sex, Age Category, Race, Diabetic Status, Physical Activity Status Health, Sleep Time, Asthma Status, Kidney Disease Status, and Heart Disease Status. Of these variables, 9 were of type boolean, 5 were of type string, and 4 were of type decimal.

2.1 Data Description

For the datasets that were used for any further analysis, a description has been given in Table 1, which highlights the variable name for the dataset, the type of each variable, and a short description of what each variable consists of.

Variable	Type	Description
HeartDisease	boolean	If respondent has ever had coronary heart disease
BMI	float	Body Mass Index
Smoking	boolean	If respondent has smoked at least 100 cigarettes in lifetime
AlcoholDrinking	boolean	Heavy drinkers (14 drinks/week for men and 7/week for women)
Stroke	boolean	If respondent has ever had a stroke
PhysicalHealth	int	Days for which physical health was NOT good
MentalHealth	int	Days for which mental health was NOT good
DiffWalking	boolean	If respondent has serious difficulty walking or climbing stairs
Sex	string	Gender – male or female
AgeCategory	string	Fourteen-level age categories
Race	string	Race of respondent
Diabetic	string	If respondent has ever had diabetes
PhysicalActivity	boolean	If respondent has partaken in physical activity in past 30 days
GenHealth	string	Individual general health measure of respondent
SleepTime	int	Hours of sleep in 24-hour period
Asthma	boolean	If respondent ever had asthma
KidneyDisease	boolean	If respondent ever had kidney disease
SkinCancer	boolean	If respondent ever had skin cancer

Table 1: “heart_2020_cleaned” Dataset

2.2 Data Cleaning

The original dataset consisted of over 300 variables, and was dwindled down to 18 variables and cleaned prior to being uploaded on Kaggle. With this, the dataset was already relatively cleaned and there were no N/A values that needed to be imputed. Thus, only a few small changes needed to be made to the dataset in order to produce data visualizations. The first change was to create a second dataset where each of the columns with boolean values of “Yes” or “No” were to be replaced with “1” or “0”. The reason for this was to create catplots, as well as bar plots that relied on frequency, rather than count. This was segmented into a second dataset so that visualizations could also be made with the total count of the dataset with boolean values. The second and final change made was to create bins for the “BMI” column, in order to group the BMI’s into a certain category for each individual. Doing this would help generalize the overall BMI’s for the consensus and allow for more conclusions to be drawn from the visualizations. The bins that were created were “underweight”, “normal weight”, “overweight”, “obese”, and “extremely obese”, and the values for each bin were taken from CDC’s website[2], outlining what ranges BMI’s are categorized by. Upon making these two changes, the dataset was now cleaned, prepared and ready to be worked with.

3 Exploratory Data Analysis

Through the majority of this section, we will take a look at the different data visualizations that were produced with the dataset, as well as create some insights, which will be used to conclude on what factors are prevalent in those that have had heart disease, as well as what can be done to avoid getting it.

3.1 Gender and Heart Disease

The first factor to be explored is whether gender plays any role in heart disease and if this disease is more common in males or females. To do so, a bar plot was created with gender on the x-axis and heart disease percentage on the y-axis, which can be seen in Figure 2. In this figure, it can clearly be seen that for the data, males are indeed more likely to get heart disease than females, where about 7% of females in the dataset had gotten the disease, while 11% of the males had gotten it.

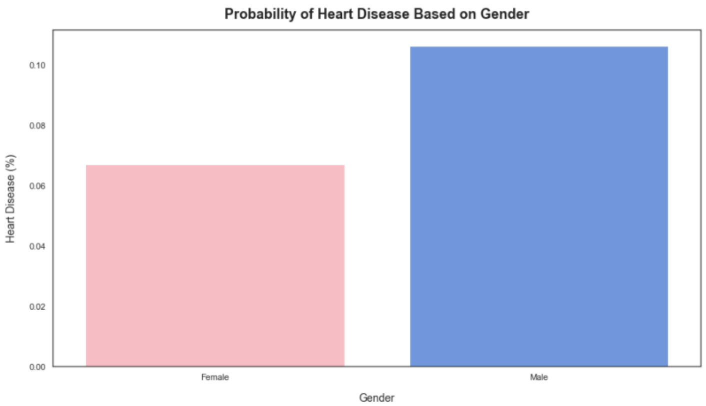

Figure 2: Bar Plot of Heart Disease percentage based on gender

Next, I wanted to target in on those that didn’t have heart disease, to better understand how this disparity looked. To do this, a pie chart was made, and split by both gender as well as those that had and didn’t have heart disease. The results of this plot can be seen in Figure 3, and it can clearly be seen once again that out of the total participants who experienced heart disease, ~ 18% was seen more for males. Further, we can also see that of those that didn’t get heart disease, the genders are pretty split, thus telling us that the dataset is balanced in terms of participants. With both Figures 2 and 3, we can conclude that males are indeed more likely to get heart disease than females.

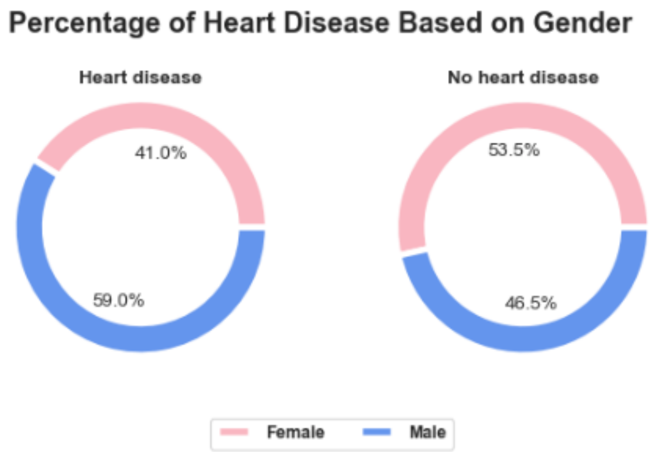

Figure 3: Pie Chart of having/not having heart disease for both genders

3.2 Age and Heart Disease

The next factor to be explored is whether age plays any role in heart disease and if you are likelier to get heart disease at certain ages. To do so, two different bar plots were created and each age range was plotted along the x-axis with

with heart disease percentage on the y-axis. The results of this initial plot can be seen in Figure 4, and it can be seen that the probability of getting heart disease based on age range follows an exponential distribution. With this we can infer that the older you get, the more likely you are to get heart disease. Further, we can also see that once you are above the age of 60, the chances of getting heart disease is around 10%.

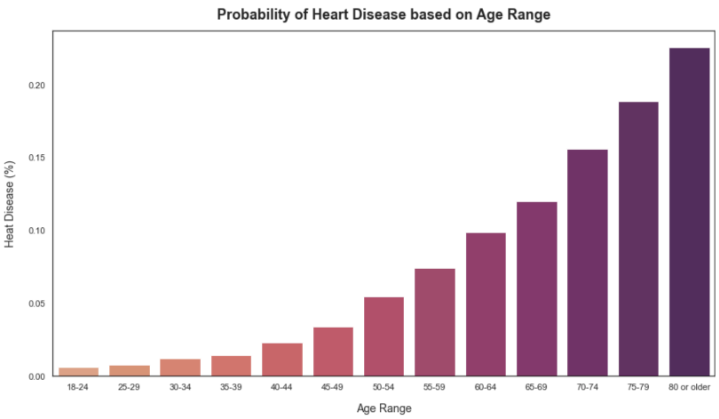

Figure 4: Bar Plot of Heart Disease percentage based on age range

Next, I wanted to look at how adding in the factor of gender changes things, and so another bar plot was created with the same axes, but the hue for each age range was gender. The results of this plot can be seen in figure 5, and we can see once again that for each age range, there are more males that get heart disease than females. We can also see that under the age of 50 (for the ranges of 25-29, 35-39, 40-44), the amount of males and females that get heart disease are almost equal, but as the ages increase, the difference becomes more skewed and is nowhere near equal.

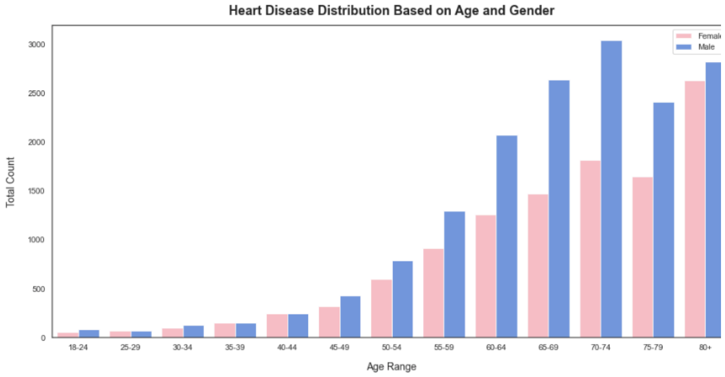

Figure 5: Bar Plot of total count of heart disease based on age range and gender

3.3 Kidney Disease and Heart Disease

There are many diseases that become common as one ages based on their genetics and overall health. The next factor to be explored is whether there is any correlation between having a kidney disease and getting heart disease, and if having a disease such as kidney disease raises the chances of getting heart disease. The most interesting way to visualize this is through a waffle chart, and using the pywaffle package allows for more styling including personalized icon types. The results of this plot can be seen in Figure 6 where there the plots are split by those that have kidney disease and those that do not. For each waffle plot, the red represents the percentage of individuals that do have heart disease, and we can see that out of those that did have kidney disease, almost 30% ended up having heart disease as well. We can also see that for those that didn't have kidney disease, there was almost 20% less people that ended up getting heart disease.

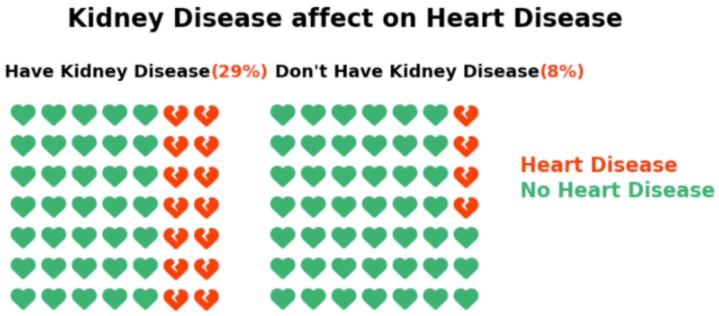

Figure 6: Waffle Chart of heart disease percentage based on kidney disease

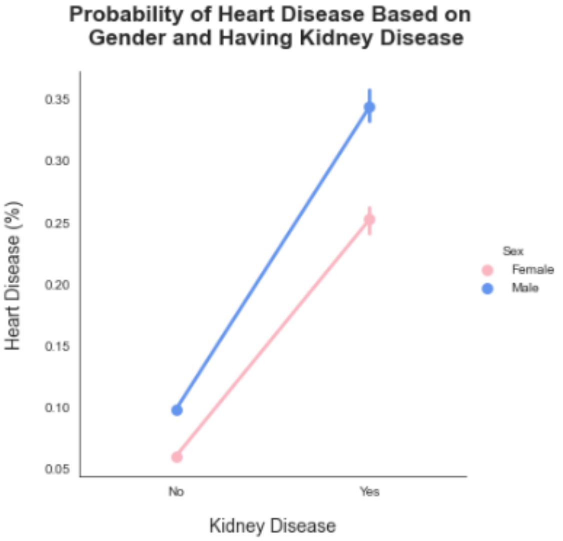

Figure 7: Cat Plot of Heart Disease percentage based on kidney disease and gender

Next, I once again chose to add the gender factor to the above plot in order to get an understanding of how the percentages change based on gender. The results can be seen in the cat plot in Figure 7 and we can see once again that even with the kidney disease factor, males end up making a higher percentage of individuals that get heart disease. Further, we can also see the differential between having kidney disease and not in terms of heart disease percentage is quite high. For those that don't have kidney disease, the percentage of getting heart disease is around 6-10% for both genders, and this percentage rises all the way to 25-35% for both genders. Thus, the conclusion here is that there is a correlation between having kidney disease and getting heart disease and having kidney disease can indeed lead to getting a heart disease.

3.4 Stroke and Heart Disease

Following on the topic of potential health issues leading to heart disease, the next factor to be explored is if having a stroke increases the chances of getting heart disease and if there is any correlation there. Once again, a waffle chart was created using pywaffle and the resulting plot can be seen in Figure 8. Looking at the percentages, we can actually see that for those that have a stroke, 36% ended up getting heart disease, which is actually a higher correlation than having kidney disease. We can almost immediately conclude that having a stroke is significantly linked to getting heart disease or having heart disease may lead to a stroke.

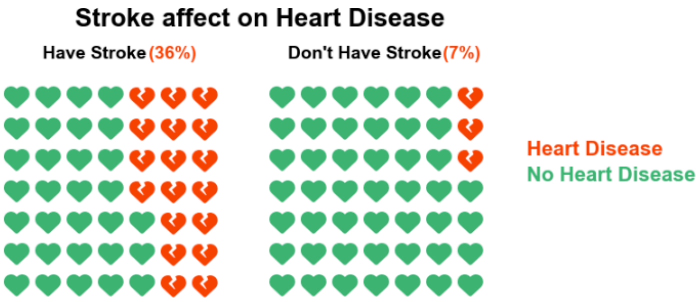

Figure 8: Waffle Chart of heart disease percentage based on stroke

Once again, I wanted to see how having a stroke changed based on gender and if the common trend of males having a higher chance of getting heart disease still holds. A cat plot was formulated, with the results being displayed in Figure 9. The trend of males holding a higher percentage of those getting heart disease still holds, but the most alarming insight is that for males, the correlation between having a stroke and getting heart disease is huge, at almost 45%. In general, females also see a huge jump of getting heart disease from not having a stroke to having one, with the percentage jumping from 5% to 30%. From these results, we can see that taking care of your heart is very important and doing anything to avoid having a stroke is of very high importance.

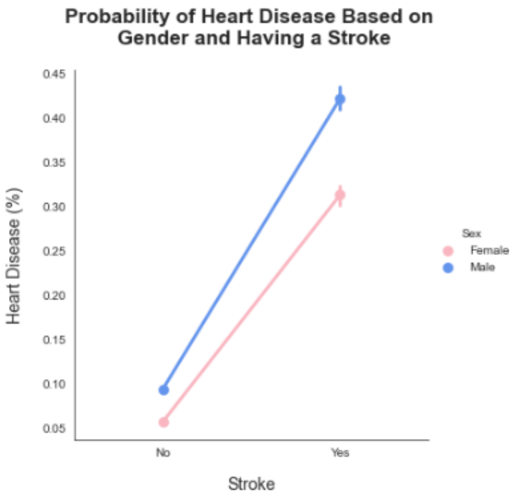

Figure 9: Cat Plot of Heart Disease percentage based on stroke and gender

3.5 Diabetes and Heart Disease

The final health-based factor to look at is diabetes, which is very prevalent and a growing issue year over year. A final waffle chart was created with the diabetes variable being used and the resulting plot can be seen in Figure 10. We can immediately see that 22% of people that did have diabetes, ended up getting heart disease and so there is some correlation. However, the percentage is slightly lower than kidney disease and much lower than having stroke, so diabetes is not as heavily linked to getting heart disease. It seems that almost any health condition can heighten the chance of getting heart disease.

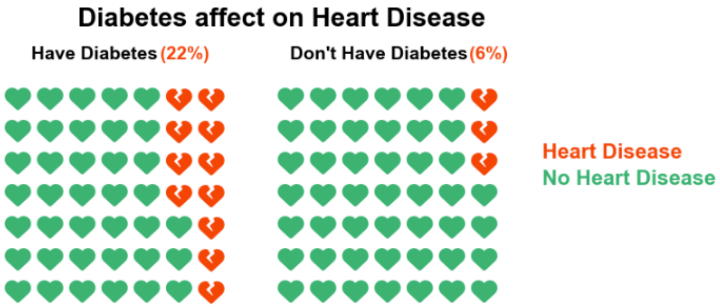

Figure 10: Waffle Chart of heart disease percentage based on diabetes

Looking at the data, it can also be seen that the diabetes type was split into 4 different categories and so this was investigated further to see if the type of diabetes increased the probability of getting heart disease. The results can be seen in the bar plot in Figure 11 with the 4 types of diabetes along the x-axis. It can be seen that those that are on the borderline of having diabetes do see an increase in getting heart disease than those that do not have diabetes at all. The percentage different between having diabetes and not having it, and getting heart disease is ~15%, so those that borderline diabetes can definitely decrease their chances of getting heart disease as well as diabetes through personal prevention strategies.

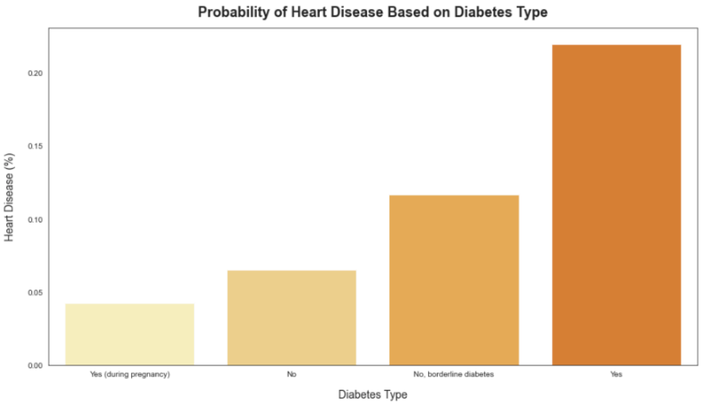

Figure 11: Bar Plot of heart disease percentage based on diabetes type

3.6 BMI and Heart Disease

BMI is a well-known metric used to differentiate weight ranges and the next factor to be compared against heart disease to determine if there is any correlation between the two. For this, an interactive horizontal bar plot was created using Plotly where the bars were expressed as a percentage value out of 100 and plotted. The resulting plot can be seen in Figure 12, with the percentage of each level along the x-axis and having/not having heart disease along the y-axis. From this plot, we can clearly see that those that are overweight make up the majority of those that have heart disease. However, we can see that overweight makes up the largest percentage for both having heart disease and not having it, and so we cannot conclude that being overweight leads to getting heart disease. Looking at obese and extremely obese, we can see that the percentage is increased for those that do have heart disease, and so we can conclude that being either obese or extremely obese will definitely increase the probability of getting heart disease.

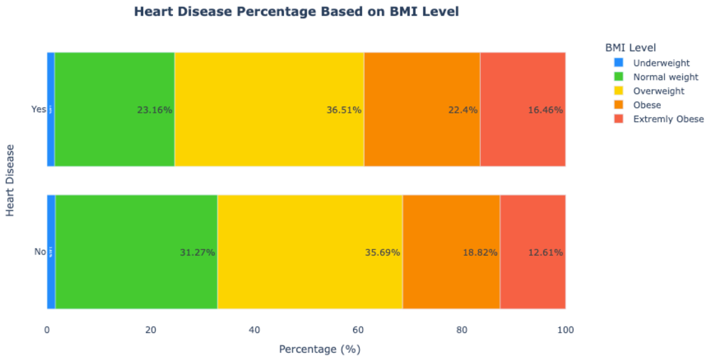

Figure 12: Bar Plot of heart disease percentage based on BMI

3.7 General Health and Heart Disease

General health is also an important metric for an individual to gauge how healthy they are and this factor is the next to be compared to heart disease to determine if there is any correlation between heart disease and having a specific health level. An interactive horizontal bar plot was once again created using Plotly with the plots being represented as a percentage out of 100. The resulting plot can be seen in Figure 13 and it can be observed that those that are of "good" health level make up the majority of those that get heart disease and those that are of "very good" health make up the majority of those that do not get heart disease. With this, we can see that just having a good health level is not enough and to truly avoid heart disease, one would have to have very good overall health. Comparing the percentage of people that have very good or excellent health, we can see that 60% of people that didn't have heart disease had at least very good health, whereas those that did have heart disease had only contained ~20% of people with excellent or very good health, which is a 40% difference. Thus, it can be seen that living an overall healthier lifestyle can significantly decrease the chances of getting heart disease.

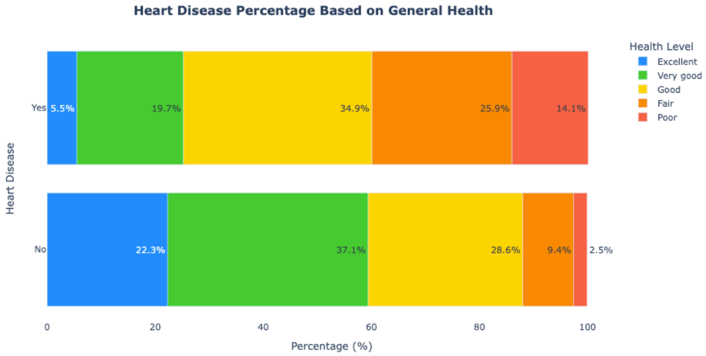

Figure 13: Bar Plot of heart disease percentage based on general health

3.8 Physical Health and Heart Disease

Physical health is an important metric to measure and respondents for the dataset were asked to provide a number for the number of days over the last 30 days that they had bad physical health. This factor is the next to be plotted against heart disease to determine if having bad physical health for a certain amount of days directly correlated to getting heart disease. A KDE plot was created where the distribution of bad physical health days was plotted against the x-axis with the frequency of the number of days plotted on the y-axis and the resulting plot can be seen in Figure 14. It can be seen from the plot that between the areas of 2-4 days of bad physical health, the frequency of those that do get heart disease is higher than those that do not and the same can be observed for the peaks at 10, 15, 20, and 30, with the highest frequency difference at 30 days. Based on this we can conclude that overall, the higher the days of bad physical health, the higher chance of getting heart disease and if you are spending over a month with bad physical health, you are much more likely to get heart disease. Incorporating even some physical activity day-to-day can prove to make a huge difference.

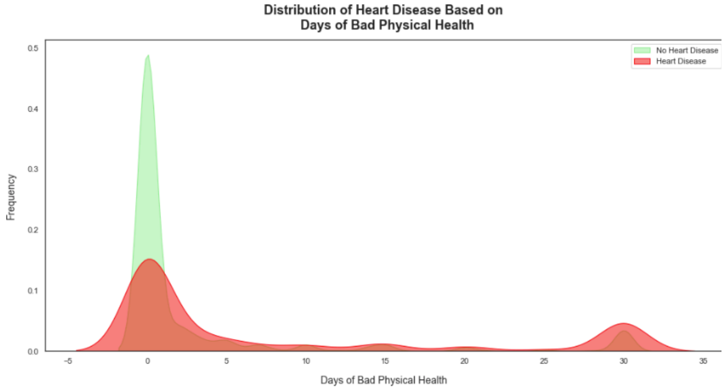

Figure 14: KDE plot of distribution of heart disease based on days of bad physical health

3.9 Mental Health and Heart Disease

Similar to physical healthy, respondents for the dataset were asked to provide a number for the number of days over the last 30 days that they had bad mental health. This factor was plotted similarly against heart disease to determine if having bad mental health for a certain amount of days directly correlated to getting heart disease. The resulting KDE plot can be seen in Figure 15. Comparing this plot to Figure 14, the main difference that can be observed is that at each of the peaks, of 5, 10, 15, 20, 30, the frequency of those getting heart disease is consistently less than that of people getting heart disease. Although this may not be entirely true, based on observing the KDE plots side-by-side, we can conclude that mental health does not have as much of an effect on getting heart disease as physical health does.

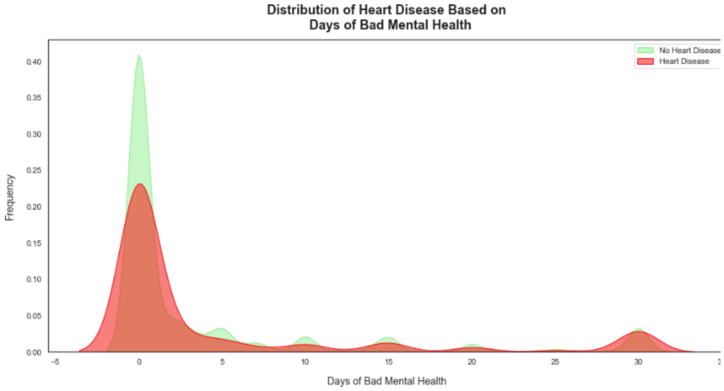

Figure 15: KDE plot of distribution of heart disease based on days of bad mental health

3.10 Smoking and Heart Disease

Smoking has been known for a very long time to be a deterrent to good health and affecting the overall longevity of a person's life. The idea is now to see if it also has an effect on an individual getting heart disease. To do this, the smoking variable was plotted along the x-axis using a bar chart with the heart disease percentage along the y-axis. The resulting plot is seen in Figure 15 and the main insight gained from this visualization is that the percentage of those that get heart disease is doubled for those that do smoke versus those that do not. This is not too surprising and is expected as smoking is known to be bad for both your lungs and heart, and it is promising to see the data showing this as well.

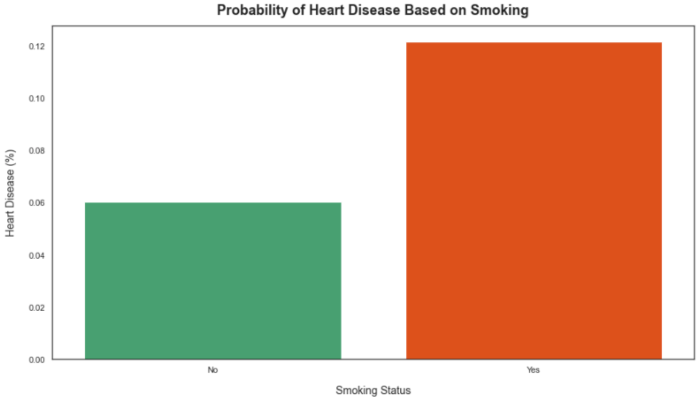

Figure 15: Bar plot of percentage of heart disease based on smoking

The next step was to see how the percentages differed for those that did have heart disease and those that didn't. With this in mind, a pie chart was created with a pie chart for those with heart disease and one created for those with no heart disease, with associated percentages for smoking and not smoking. The resulting plot is seen in Figure 16 and we can see for those that do have heart disease, almost 60% were smokers and for those that did not have heart disease, 60% were not smokers. This further confirms our assumption that smoking is definitely correlated to heart disease and it significantly raises the chances of getting heart disease.

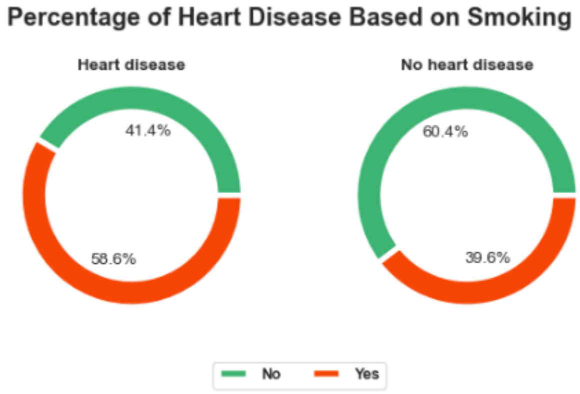

Figure 16: Pie chart of percentage of heart disease based on smoking

3.11 Alcohol and Heart Disease

Another common vice is alcohol consumption, and although it hasn't been tied to heart disease, the goal was to compare this variable against heart disease to determine if alcohol consumption does indeed have an effect on getting heart disease. Similarly to the smoking plots created, a bar plot was created based on alcohol consumption and can be seen in Figure 17. Further, a pie chart was created and can be seen in Figure 18. In Figure 17, we can see that the drinking alcohol versus not, does not have much of an effect on heart disease percentage, with only ~ 3% increase in getting heart disease for those that drink. From this insight, it seems as though alcohol consumption does not have as much of an impact on heart disease. Looking at Figure 18, we can see that out of those that have heart disease, only 4% consume alcohol, and out of those that don't have heart disease, only 7% consume alcohol. These numbers seem very low and it seems as though a lot of individuals are not drinking alcohol and this is leading to them not getting heart disease. Based on this, we can conclude that either drinking alcohol does not heavily impact getting heart disease or that there is not enough census data on those that consume alcohol to come to a conclusion on this.

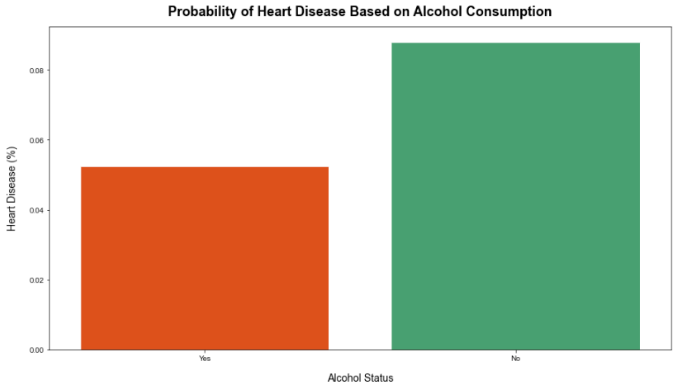

Figure 17: Bar plot of percentage of heart disease based on alcohol consumption

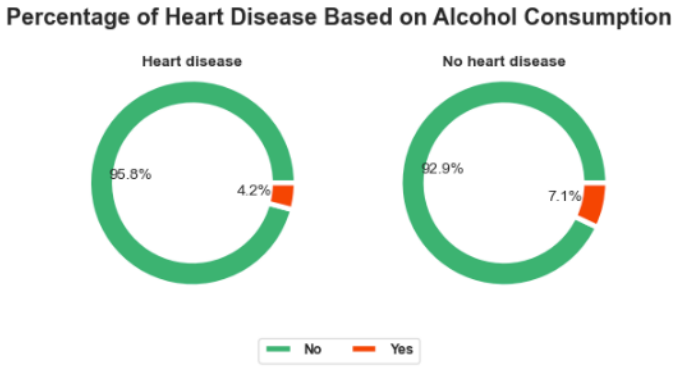

Figure 18: Pie chart of percentage of heart disease based on alcohol consumption

3.12 Sleep and Heart Disease

The final factor to be looked into is sleep, as sleep is something very important and each individual partakes in. We all know the amount of sleep you get in a day is very important for body recovery, but sleeps ties to heart disease is unknown so seeing if there is a correlation is of great importance. To do this, a bar plot was created with sleep time in hours plotted on the x-axis and heart disease percentage was plotted on the y-axis with each bar plot representing if an individual had heart disease or not. The resulting plot can be seen in Figure 19 and looking at this plot, it can be seen that for those that sleep in the ranges outside of 6-8 hours, there is a higher percentage probability of getting heart disease. With this, we can conclude that sleep only has an effect on heart disease when sleeping outside the 6-8 hour range. Within the 6-8 hour range, the proportion of those that get heart disease versus not is nearly identical, with the exception of 7 hours, where there are more individuals that don’t get heat disease than do.

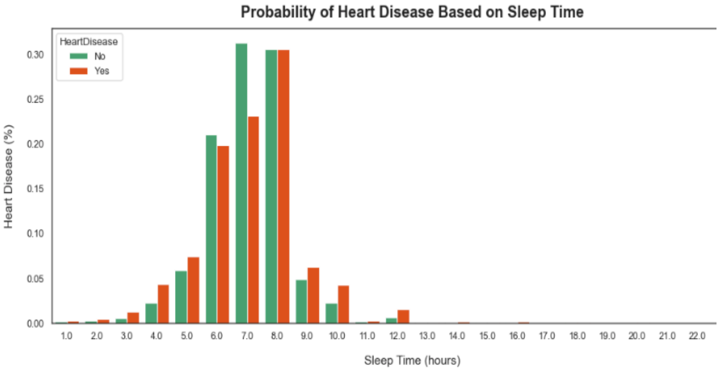

Figure 19: Bar chart of percentage of heart disease based on sleep time in hours

4 Conclusion and Future Work

Through this EDA project, I was able to gain many valuable insights on what factors do and do not have effect on heart disease. Some final conclusions are:

- Males predominantly have a higher chance of getting heart disease than females
- An individual’s chances of getting heart disease exponentially increase as you get older
- Having a kidney disease, diabetes or a stroke is very correlated to getting heart disease
- Having bad physical health for 30 days or more can lead to heart disease
- Alcohol consumption does not have an effect on getting heart disease

Further, I am now also able to provide some prevention tips, which are backed by the dataset and insights gained. These include:

- A BMI of 18.5-24 (considered normal weight) as well as having very good health are two very important factors that can lead to prevention of heart disease
- Smoking will double your chances getting heart disease. Avoid it at all costs
- Getting sleep outside of 6-8 hours can also negatively impact your chances of getting heart disease

Keeping all these things in mind, there is still room for future work, as this project focused mainly on EDA and identifying commonalities to heart disease through visualizations. Some possible future work would include applying a range of machine learning methods such as classifier models (logistic regression, SVM, random forest, etc.) in order to make predictions such as if a subset of variables will lead to getting heart disease. Using this, a person would be able to enter personal attributes into a model (such as BMI, sleep in hours, physical health, etc.) and get a prediction on whether they would get heart disease or not.

5 References

[1] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[2] https://www.cdc.gov/obesity/basics/adult-defining.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obesity%20range