**Project Report**

# GitHub URL

<https://github.com/aoifej8/UCDPA_aoifejohnston/upload/main>

# Abstract

The project is based on a key dataset regarding the key indicators of heart disease. This project includes a csv file imported from Kaggle.com (<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>).

Introduction

I chose to base my project on the key indicators of heart disease because it is a topic that is personal to me, and I am very interested in exploring the data. Heart disease is the leading cause of death in many countries, including the United States of America (*Heart Disease,* 2020). The term ‘heart disease’ can refer to numerous heart conditions such as a heart attack or coronary artery disease. A heart attack occurs when a section of the heart is not receiving enough blood, and therefore causes damage to the heart. Coronary disease is caused by a buildup of plaque on the walls of the arteries that pump blood to the heart and other parts of the body. Without the tireless effort of heart pumping blood, both of these conditions are highly likely to cause death.

The heart is one of the most important organs in the body and it is the main organ in the cardiovascular system. This system is made up of a network of blood vessels which pumps blood all around the body. The heart also controls the rhythm and speed of the heart rate in the body, along with maintaining blood pressure. Furthermore, undesirable carbon dioxide and waste products are carried away by the blood filled with nutrients and oxygen.

The heart is evidently important; therefore, it needs to be understood and taken care of. This dataset explores the indicators of heart disease, which sparks my interest as I feel this is vital information for everyone to know. Staying informed and being aware of the indicators keeps a person’s risk of heart disease low. Some risk factors are uncontrollable, such as race or family background. However, many key risk factors are controllable and being aware of them can significantly reduce one’s risk of heart disease. Additionally, the detection and prevention of heart disease is vital to healthcare.

# Dataset

The dataset I chose was on the key indicators of heart disease was extracted from Kaggle.com. Kaggle is an online platform with countless amounts of datasets, targeted towards data scientists and machine learning fanatics. It grants its users access to GPU integrated notebooks, it allows them to interact and work together with other Kaggle users to tackle data science-based challenges. Moreover, Kaggle is a platform where anyone can find and publish datasets (Uslu, 2022).

This dataset came from the Centers for Disease Control and Prevention (CDC). The CDC gathers data on the health status of American residents, and it describes itself as “the largest continuously conducted health survey system in the world”. There are 18 columns in this dataset, 5 are numerical while the other 13 are categorical.

1. HeartDisease: Respondents that have ever reported having experienced any heart disease. (Yes/No)
2. BMI: Body mass index of respondents.
3. Smoking: Respondents that have smoked at least 100 cigarettes in their life. (Yes/No)
4. AlcoholDrinking: Respondents who are adult men who consume more than 14 drinks per week, and respondents who are adult women who consume more than 7 drinks per week. (Yes/No)
5. Stroke: Have you suffered from a stroke? (Yes/No)
6. PhysicalHealth: How many days, in the last 30 days, were respondents’ physical health not good? (0-30 Days)
7. MentalHealth: How many days, in the last 30 days, were respondents’ mental health not good? (0-30 Days)
8. DiffWalking: Do you have difficulty walking or climbing stairs? (Yes/No)
9. Sex: Are you male or female?
10. AgeCategory: 13 level age category:
    * 18-24
    * 25-29
    * 30-34
    * 35-39
    * 40-44
    * 45-49
    * 50-54
    * 55-59
    * 60-64
    * 65-69
    * 70-74
    * 75-79
    * 80 or older
11. Race: Ethnicity
12. Diabetic: (Ever told) (you had) diabetes? (Yes/No)
13. PhysicalActivity: Respondents who did physical activity or exercise during the past 30 days other than their regular jobs. (Yes/No)
14. GenHealth: General health of respondents.
15. SleepTime: How many hours of sleep do you get in a 24 period?
16. Asthma: (Ever told) (you had) athsma? (Yes/No)
17. KidneyDisease: (Ever told) (you had) kidney disease? (Yes/No)
18. SkinCancer: (Ever told) (you had) skin cancer? (Yes/No)

# Implementation Process

I began my project by importing all the libraries I would require into my Jupiter notebook. I imported pandas, numpy, csv, seaborn and matplotlib.pyplot. My next step was importing the csv file, and then using the .head() function to view the dataset in my Jupiter notebook.

I wanted to explore the dataset a bit more so I used the .dtypes() function to view the data types I had in my dataset. This resulted in two data types, object and float64. After this, I wanted to view the shape of the dataset so I used the function .shape(), and this told me that my dataset had 18 columns and 319,795 rows. I called on the describe function to ignore missing values and exclude categorical variables, and then I demonstrated I can include the categorical variables again by editing the function.

Once I had created a pandas dataframe, I wanted to experiment with various codes. I started by accessing the columns. I began by returning one column, ‘Race’. Then I returned three columns all at once, ‘Sex’, ‘PhysicalHealth’ and ‘MentalHealth’. Next, I returned rows by slicing and finally by location.

Using multiple codes, I wanted to explore the dataset and see how I could clean the dataframe. I began by dropping any duplicates in my subet, [‘BMI’,’Smoking’], to find any rows that were the same. However, there were no rows that shared the same values as ‘BMI’ contained only numerical values, and ‘Smoking’ contained categorical values. I checked for any missing values but there were none. Similarly, I attempted to drop any rows in the subset that shared the same values but there were none. I grouped the columns ‘Diabetic’ and ‘Sex’ together and I aggregated the data with mean and median. I then reset the index so I could visually compare the two dataframes.

During my analysis of the dataset, I used pandas iteration function, iterrows(), to iterate over my pandas Data frame rows in the form of index pair. I used the code .info() to recall the data types I have in my data frame. I then separated my categorical and numerical variables

The most comprehensive section of my project is the visualization section. I began by creating a pie chart to view the percentage of people with heart disease from the sample. I then moved on to comparing men and women with regards to heart disease. From this I was able to calculate the likelihood of men having heart disease compared to women. Thereafter, I wanted to look at the relationship between heart disease and smoking. To accomplish this, I first needed to concatenate the two columns, then I used the code .value\_counts() to count the number of times of the four values were repeated. Using these figures, I was able to create a stacked bar chart where I could distinctly view the relationship. Afterwards, I calculated the percentage of each category to further aid my stacked bar chart. I repeated these exact steps in calculating the relationship between heart disease and drinking alcohol, then I was able to compare the two stacked bar charts. Following this, I created a pairplot to show the pairwise relationships in the dataset. This plotted each numerical feature against one another as a matrix. The final stage of my visualization was creating a heatmap to show the correlation of each numerical feature to one another.

# Results

The first visualization aid I created was a pie chart. This diagram displayed the percentage of people with heart disease from the sample.

Percentage of people with heart disease from the sample.


From this pie chart I can see that only 8.56% of the sample suffer with heart disease, thankfully 91.44% do not.

My next diagram is shows which sex, from the sample, has a higher report of heart disease.

Chart, bar chart

Description automatically generated

This visualisation support clearly shows that heart disease is more commonly present in men, than it is in women. After seeing these results, I decided to calculate what is the probability of each sex having heart disease, and then the likelihood of men having heart disease compared to women. I calculated that men are 1.6 more times likely to have heart disease than women.

I analysed the relationship between heart disease and smoking using a stacked bar chart.

Chart, bar chart

Description automatically generated

At first glance at the stacked bar chart, it appears that people who smoke are more likely to have heart disease. I decided to take this analysis one step further by calculating the percentages. From the sample, 91.67% of people who do not have heart disease smoke, while 8.33% do smoke. Similarly, 91.09% of people who do have heart disease do smoke, while 8.91% do not smoke.

Subsequently, I created a stacked bar chart to look at the relationship between heart disease and drinking alcohol.Chart, bar chart

Description automatically generatedLooking at this diagram, it appears that the majority of the people from the sample do not have heart disease, nor do they drink alcohol. I decided to find the percentages of these variables. From the sample, only 5.24% of people who do not suffer from heart disease drink alcohol. Furthermore, only 8.8% of people who do suffer from heart disease drink alcohol.

To view the pairwise relationships in the dataset, I created a pairplot. This plotted the four numerical variables, BMI, physical health, mental health and sleep time, against one another as a matrix.

A picture containing chart

Description automatically generated

From this pairplot, I can see the greatest correlation is between mental health and physical health. To confirm this, I created a heatmap to show the correlation of each numerical value to one another.

Chart, treemap chart

Description automatically generated

This heatmap confirms that the strongest correlation is between mental health and physical health, with a value of 0.29.

# Insights

* As heart disease is the leading cause of death in many countries, only 8.56% of the sample suffer from heart disease. This is because very survive a heart attack, hence many do not live with heart disease.
* Men are 1.6 times more likely to suffer with heart disease. This may be due to the possibility than men’s ability to cope with stressful event are perhaps less adaptive physiologically and psychologically. This all contributes to a higher risk of heart disease.
* Smoking will not necessarily cause heart disease, but 91.09% of people from the sample who do have heart disease do smoke.
* Drinking alcohol is even less likely to cause heart disease than smoking, however 8.8% of people who do suffer from heart disease drink alcohol.
* Between the numerical values, physical health and mental health have the strongest correlation. This suggests that they go hand-in-hand, as they support each other.
* In terms of machine learning, I could use classification methods to see how smoking and drinking alcohol will affect the chance of someone having heart disease in the future.

# References

* Centers for Disease Control and Prevention. 2022. *Heart Disease*. [online] Available at: <https://www.cdc.gov/heartdisease/index.htm> [Accessed 16 June 2022].
* Uslu, Ç., 2022. *What is Kaggle?*. [online] datacamp.com. Available at: <https://www.datacamp.com/blog/what-is-kaggle> [Accessed 16 June 2022].