Introduction

Heart disease poses a significant health concern on a global scale, particularly in the United States, affecting individuals of diverse backgrounds, both men and women. It stands as the leading cause of mortality in the country, with a shocking statistic indicating that a life is lost to cardiovascular disease approximately every 33 seconds (National Center for Health Statistics, 2023). This condition's impact is profound, as evidenced by the staggering figure of 695,000 deaths attributed to heart disease in 2021 alone, accounting for roughly one in every five deaths (National Center for Health Statistics, 2021). Moreover, the economic ramifications associated with heart disease are substantial, with an estimated annual cost of $239.9 billion. This encompasses expenses related to healthcare services, medications, and the economic productivity lost due to premature death. These statistics underscore the pressing need for effective strategies in prevention, diagnosis, and management to address the impact of heart disease on individuals and society.

This report aims to utilize exploratory data analysis tools to identify the leading causes or indicators of heart disease, employing a 2015 dataset sourced from the Behavioral Risk Factor Surveillance System (BRFSS). Various tools, such as summary statistics and visualizations, will be employed to assess the significance of these variables in determining the likelihood of developing heart disease. It is important to note that this analysis represents a simplified approach and acknowledges certain limitations, thereby recommending the utilization of additional tools discussed in the conclusion section.

To select feature variables for predictor of heart disease, a literature review was conducted. The literature highlighted numerous indicators or factors that can serve as predictors of heart disease, enabling a proactive approach to addressing this health issue. Our first variable found is high blood pressure. Scholars have observed a continuous relationship between increasing blood pressure and cardiovascular disease, emphasizing the importance of monitoring and managing hypertension (Klag, Whelton, & Randall, 1996). Hypertension, a well-established risk factor, can lead to complications such as stroke, ischemic heart disease, and renal dysfunction (Escober, 2002). Furthermore, the National Health and Nutrition Examination Survey (NHANES) reports a higher prevalence of hypertension in men compared to women, with rates of 30.5% and 28.5% respectively (Guo, He, Zhang et al., 2010). Age is the second variable of interest. Age-related changes in the heart and blood vessels can significantly elevate an individual's risk of heart disease, emphasizing the need for comprehensive prevention efforts (Williams et al., 2006). Diabetes is another indicator of risk of heart disease. Individuals with diabetes are more prone to developing heart disease at an earlier stage due to the adverse effects of high blood glucose on heart function (acadmin, 2019). Body Mass Index (BMI), a measure of body fat relative to height, serves as another valuable indicator, with higher BMI values associated with an increased risk of heart problems (acadmin, 2019).

In addition to physical factors, One’s lifestyle plays a role in indicating the presence of heart disease. Smoking history, inadequate consumption of fruits and vegetables, and a sedentary lifestyle further contribute to the risk of heart disease (acadmin, 2019; Nystoriak & Bhatnagar, 2018). Mental health also plays a pivotal role in overall well-being, encompassing psychological, emotional, and social aspects. Prolonged periods of depression, anxiety, and stress can lead to heightened cardiac reactivity, resulting in increased heart rate, elevated blood pressure, and reduced blood flow to the heart, thus raising the risk of heart disease (Sowden & Huffman, 2009). Individuals with mental health disorders often possess inadequate coping mechanisms, which may manifest in behaviors such as smoking or a sedentary lifestyle, further exacerbating the risk of heart disease (Del Gaizo, Elhai, & Weaver, 2011).

Given the multifaceted nature of heart disease and its immense impact on individuals and society, it is crucial to delve deeper into the exploration of these indicators and risk factors. By analyzing the BRFSS dataset using various data analysis tools, this report aims to shed light on the significant contributors to heart disease and provide valuable insights for developing effective prevention and management strategies for the community.

Data Analysis (*Section 2 to 4*)

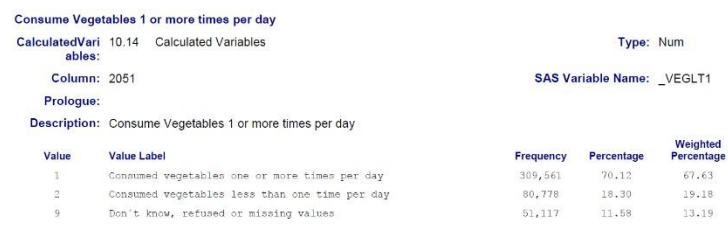
To begin the analysis on indicators of heart disease, the 2015 BRFSS dataset was used which contained 330 variables on 441456 respondents. The BRFSS is the US's premier system of telephone surveys on gathering state-level data about health- related risk behaviors, chronic health conditions and utilization of preventive services among US residents. The dataset was reduced to 14 variables to only show the variables of interest. However, the dataset had some missing values as can be seen in the number of nonnulls on some columns such as TOLDHI2 in the diagram below. These rows with the missing data were dropped.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 441456 entries, 0 to 441455
Data columns (total 14 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   _MICHD    437514 non-null  float64
 1   _RFHYPE5  441456 non-null  float64
 2   TOLDHI2   382302 non-null  float64
 3   _BMI5     405058 non-null  float64
 4   SMOKE100  427201 non-null  float64
 5   CVDSTRK3  441456 non-null  float64
 6   DIABETE3  441449 non-null  float64
 7   _TOTINDA  441456 non-null  float64
 8   _VEGLT1   441456 non-null  float64
 9   _RFDRHV5  441456 non-null  float64
 10  MENTHLTH  441456 non-null  float64
 11  PHYSHLTH  441455 non-null  float64
 12  SEX       441456 non-null  float64
 13  _AGEG5YR  441456 non-null  float64
dtypes: float64(14)
memory usage: 47.2 MB
```

This then left the dataset with 343,612 respondents. Data imputation methods could have been used for missing data however their use in the medical field is limited because these methods fail to adequately address the complexities of missing data in healthcare applications.

The dataset was made up of survey data, as such it was further cleaned to ensure it makes sense. This was done by first referring to the codebook which stated what the selected variable meant and entailed. This codebook can be found by clicking the link below:
https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf.

A snippet of some selected variables from the codebook have been provided below. We can see the variables have values which are numbers that represent a label to it. To make sense out of this, for example if we look at the first figure with variable name _VEGLT1, the value 1 was left as it is and 2 was changed to 0 to represent no vegetables taken in a day. Respondents who had value 9 were dropped.



Consume Vegetables 1 or more times per day

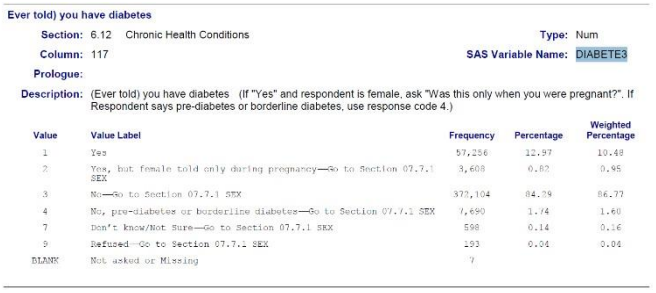
CalculatedVariables: 10.14 Calculated Variables — Type: Num

Column: 2051 — SAS Variable Name: _VEGLT1

Prologue:

Description: Consume Vegetables 1 or more times per day

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Consumed vegetables one or more times per day	309,561	70.12	67.63
2	Consumed vegetables less than one time per day	80,778	18.30	19.18
9	Don't know, refused or missing values	51,117	11.58	13.19



Ever told) you have diabetes

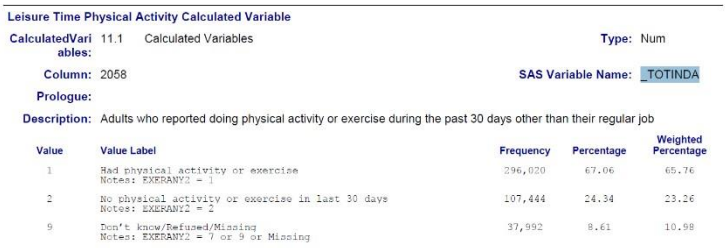
Section: 6.12 Chronic Health Conditions — Type: Num

Column: 117 — SAS Variable Name: DIABETE3

Prologue:

Description: (Ever told) you have diabetes (If "Yes" and respondent is female, ask "Was this only when you were pregnant?". If Respondent says pre-diabetes or borderline diabetes, use response code 4.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	57,256	12.97	10.48
2	Yes, but female told only during pregnancy—Go to Section 07.7.1 SEX	3,608	0.82	0.95
3	No—Go to Section 07.7.1 SEX	372,104	84.29	86.77
4	No, pre-diabetes or borderline diabetes—Go to Section 07.7.1 SEX	7,690	1.74	1.60
7	Don't know/Not Sure—Go to Section 07.7.1 SEX	598	0.14	0.16
9	Refused—Go to Section 07.7.1 SEX	193	0.04	0.04
BLANK	Not asked or Missing	7		



Leisure Time Physical Activity Calculated Variable

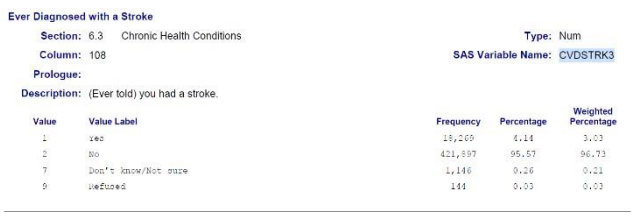
CalculatedVariables: 11.1 Calculated Variables — Type: Num

Column: 2058 — SAS Variable Name: _TOTINDA

Prologue:

Description: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Had physical activity or exercise Notes: EXERANY2 = 1	296,020	67.06	65.76
2	No physical activity or exercise in last 30 days Notes: EXERANY2 = 2	107,444	24.34	23.26
9	Don't know/Refused/Missing Notes: EXERANY2 = 7 or 9 or Missing	37,992	8.61	10.98



Ever Diagnosed with a Stroke

Section: 6.3 Chronic Health Conditions — Type: Num

Column: 108 — SAS Variable Name: CVDSTRK3

Prologue:

Description: (Ever told) you had a stroke.

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	18,269	4.14	3.03
2	No	421,897	95.57	96.73
7	Don't know/Not sure	1,146	0.26	0.21
9	Refused	144	0.03	0.03

A similar procedure was done to the other variables in the dataset. Note most of the variables were categorical except for a few such as BMI, Mental health and physical days and Age which was in ranges. The variables were then renamed to make it easier to understand and work with. A summary of all the variables has been provided in the descriptive table below:

Attribute	Description

Age	*Value 1 to 13 with increments of 5 yrs ie 1 represents ages 18-24, 2: 25-29, 3: 30-34*
HeartDisease	*1: has heart disease , 0: no heart disease*
Smoker	*0: Non smoker 1: Smoker*
Mental Health days (coded as 'MentHlth')	*Days in the past 30 days when mental health was not good ie 0 represents none 2,3,4 represents the days when a respondent's mental health was not good*
Physical Health days (coded as 'PhysH')	*Days in the past 30 days when physical health was not good due to various factors such as physical illness or injury. Similar to mental health days*
Stroke	*1: Had a stroke, 0: No Stroke*
Sex	*1: Male, 0: Female*
Diabetes	*0 is for no diabetes or only during pregnancy, 1 is for pre-diabetes or borderline diabetes, 2 is for yes diabetes*
High Blood Pressure (coded as HiBP)	*0: No blood pressure, 1: high blood pressure*
High Cholesterol levels (coded as HiChol)	*0: No high Cholesterol 1: High Cholesterol*
In take of Vegetables (coded as Veggies)	*0: this means no vegetables consumed per day. 1 will mean consumed 1 or more pieces of vegetable per day*
Exercise	*0: No physical activity/exercise 1: there is physical activity*
Alcohol (heavy intake of alcohol ie more than 14 drinks per week for men and more than 7 for women)	*0: No heavy intake of alcohol 1: heavy intake of alcohol*
BMI	*Body Mass Index from 1 to 100*

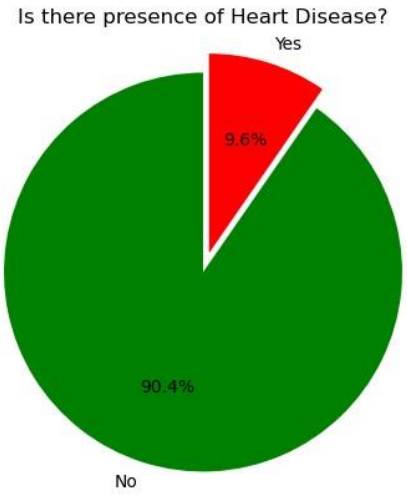
The final dataset used for analysis includes 14 variables with 297560 observations. As can be seen in the table below, there are no missing values which ensures the integrity and avoidance of biases in the dataset. This is also beneficial as it ensures completeness, integrity, and consistency of the data. Furthermore, it provides valid and reliable results when visualisations or statistical analysis is made.

The image below gives us a picture of what our dataset consists from column names, data types in each column and number of non-missing values.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 297560 entries, 0 to 441455
Data columns (total 14 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   HeartDisease  297560 non-null  int32
 1   HiBP          297560 non-null  int32
 2   HiChol        297560 non-null  int32
 3   BMI           297560 non-null  float64
 4   Smoker        297560 non-null  int32
 5   Stroke        297560 non-null  int32
 6   Diabetes      297560 non-null  int32
 7   Exercise      297560 non-null  int32
 8   Veggie        297560 non-null  int32
 9   Alcohol       297560 non-null  int32
 10  MentHlth      297560 non-null  int32
 11  PhysH         297560 non-null  int32
 12  Sex           297560 non-null  int32
 13  Age           297560 non-null  int32
dtypes: float64(1), int32(13)
memory usage: 19.3 MB
```

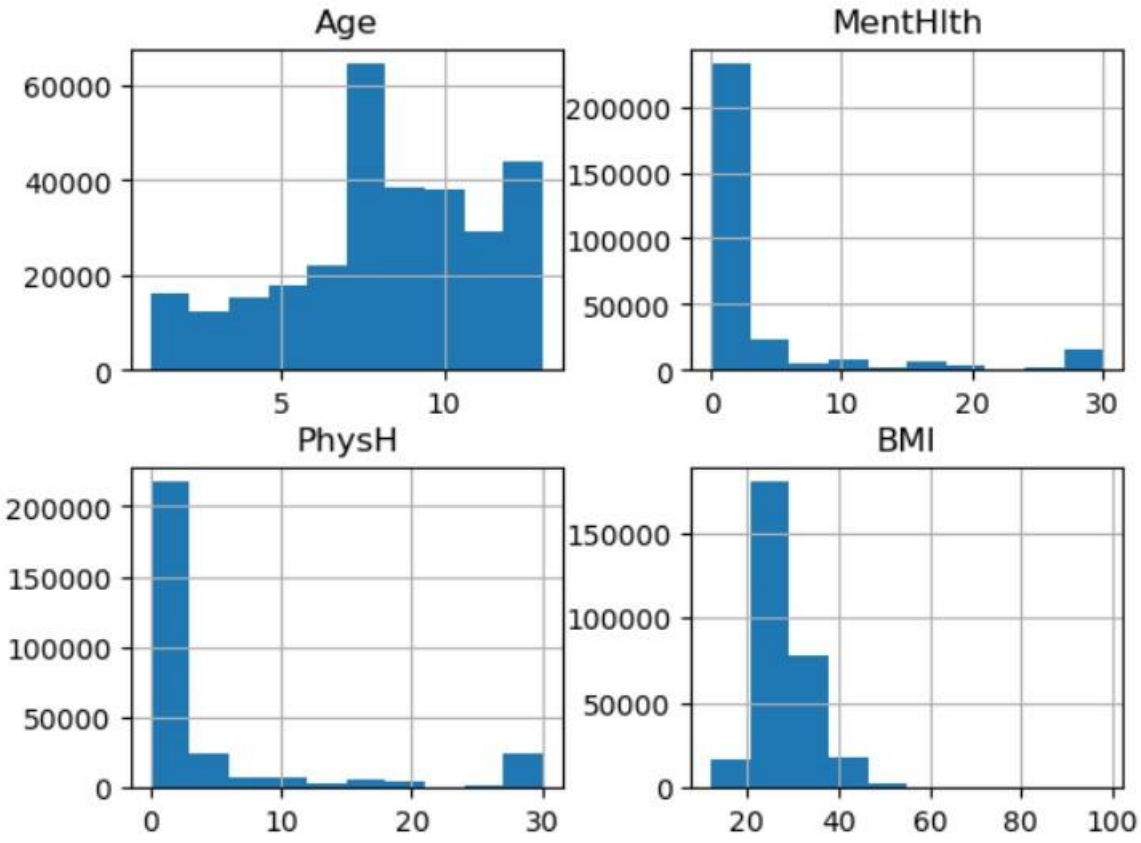
The next section is going to explore the distribution of the feature variable and the other variables with more than two values (ie non-categorical) in the dataset.

The feature variable in the dataset is HeartDisease which is categorical. The pie chart below shows the proportion of respondents who have heart disease vs those who do not.



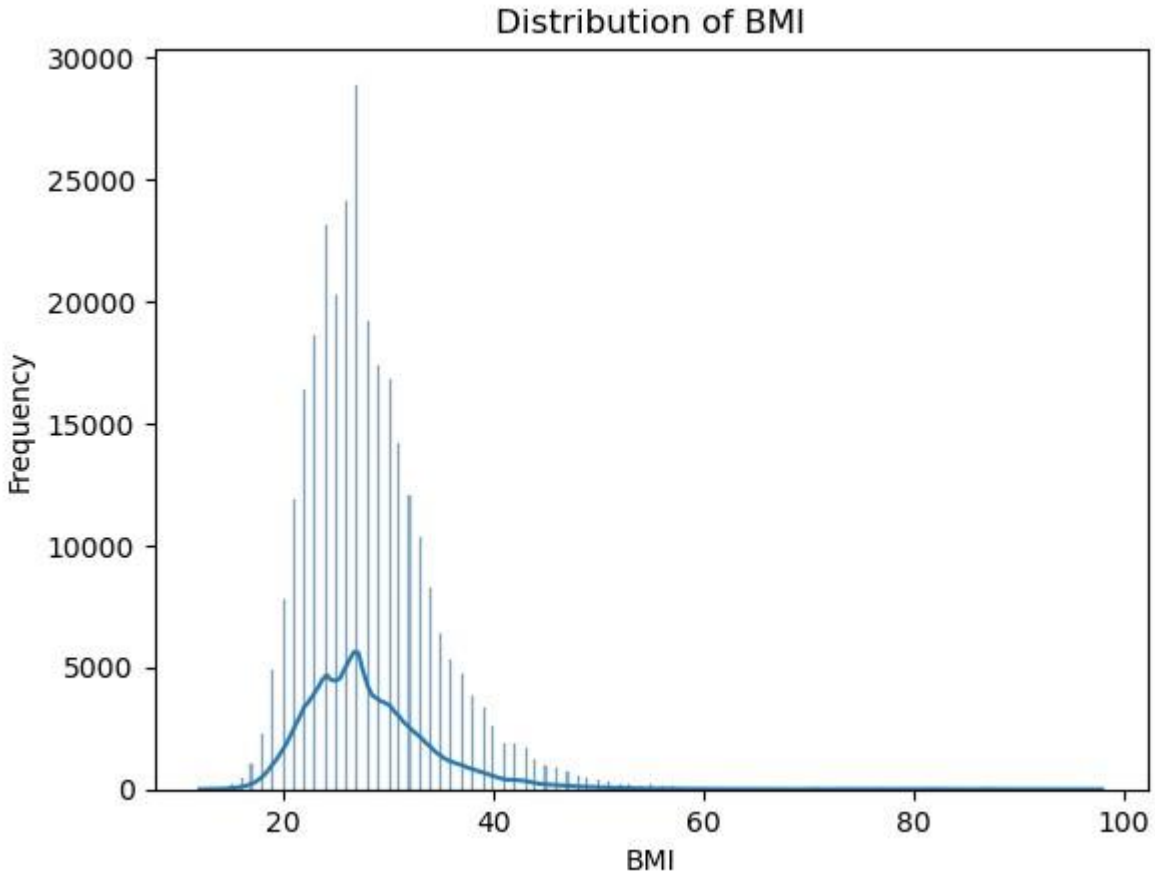
We have a disproportionate dataset as majority of the respondents do not have heart disease. This would be disadvantageous if we would run machine learning models on this dataset as there will be an issue of class imbalance. This may cause the model to become biased towards predicting majority class (no heart disease) due to the abundance of instances representing that class. However, there are various ways to fix this problem.

The distribution of the non-categorical variables can be seen below:



From the above, BMI displays a normal distribution while the distribution of the others cannot be clearly classified. Majority of respondents fall within a BMI of between 20 to 32 which is a healthy score. The most common value in 'Age' variable is bin 7 which represents ages 50 to 54. The Age variable also seems to have a slightly fair distribution of the age ranges and there is no outlier. The 'MentHlth' and 'PhysH' variable which represents mental health and physical health days show that majority of respondents fell in the 0 range. This meant most of them did not experience mental health issues or physical illness or injury. Additionally, there are no outliers as there are no individual bars that are far isolated from the dataset. Due to the disproportionate classification of the dataset as stated earlier on, it makes sense that majority of the dataset fell in the 0 bin for physical and mental health.

The graph below shows the normal distribution of the BMI clearly as opposed to the figure above.



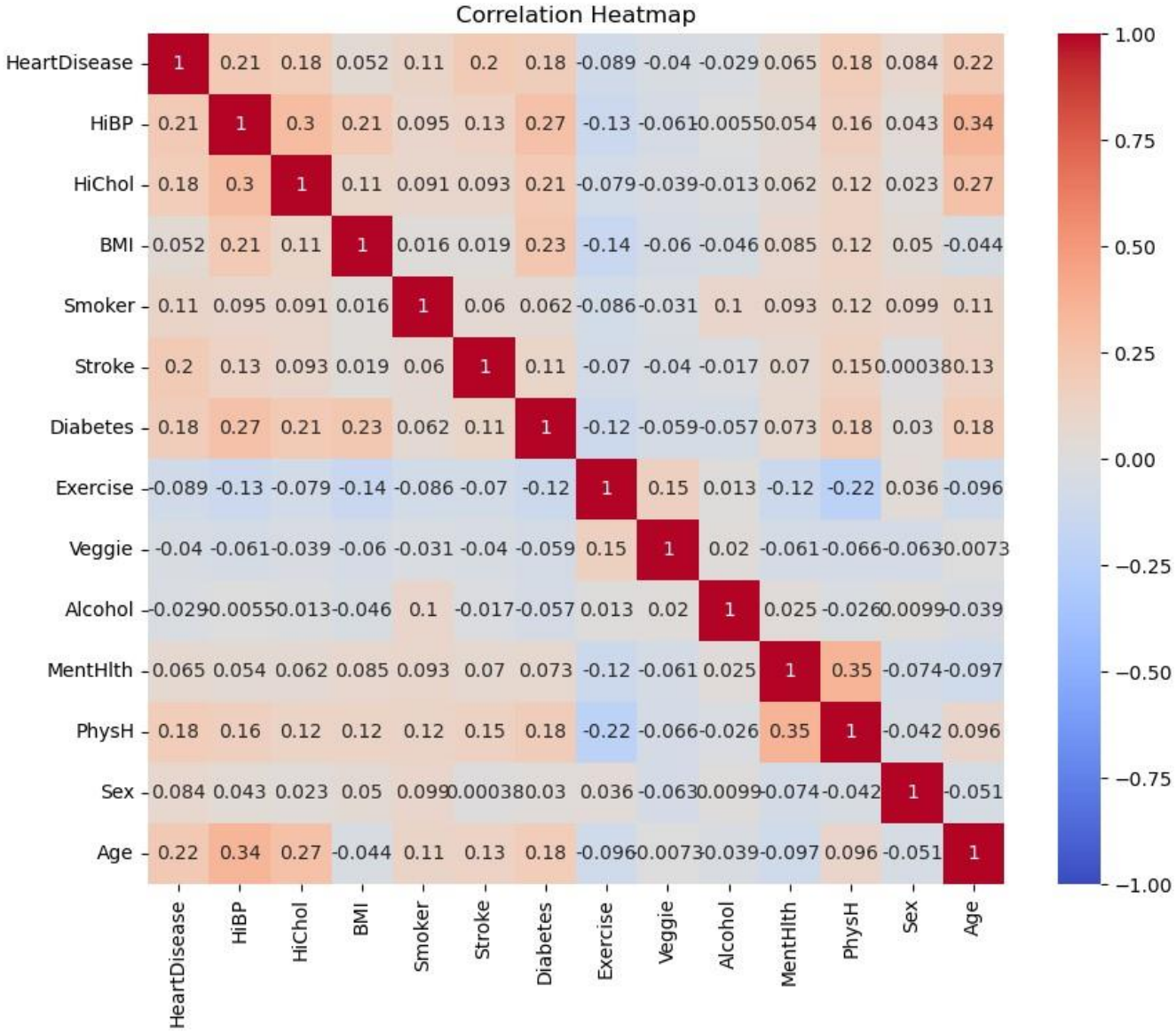
The summary statistics of all the variables is attached below:

	HeartDisease	HiBP	HiChol	BMI	Smoker	Stroke	Diabetes	Exercise	Veggie	Alcohol	MentHlth	PhysH	Sex	Age
count	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000	297560.000000
mean	0.096478	0.433714	0.425165	28.265056	0.439894	0.042395	0.300417	0.750151	0.808197	0.053435	3.178556	4.350410	0.428129	8.160018
std	0.295246	0.495588	0.494369	6.591770	0.496375	0.201489	0.701432	0.432926	0.393720	0.224899	7.442738	8.848133	0.494808	3.106151
min	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	0.000000	0.000000	0.000000	24.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	6.000000
50%	0.000000	0.000000	0.000000	27.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	9.000000
75%	0.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	2.000000	3.000000	1.000000	10.000000
max	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	2.000000	1.000000	1.000000	1.000000	30.000000	30.000000	1.000000	13.000000

Most of the variables are categorical so will not make sense, however the summary statistics compliments the histograms above. It gives clear indication of measures such as min, max and mean. The mean BMI is 28 which is represented by the peak on the histogram.

The next section of this report will now explore the relationship between the variables in the dataset. More importantly, this will help us select feature variables that help predict the risk of heart disease by focusing on variables that have a significant relationship ie high correlation value with HeartDisease

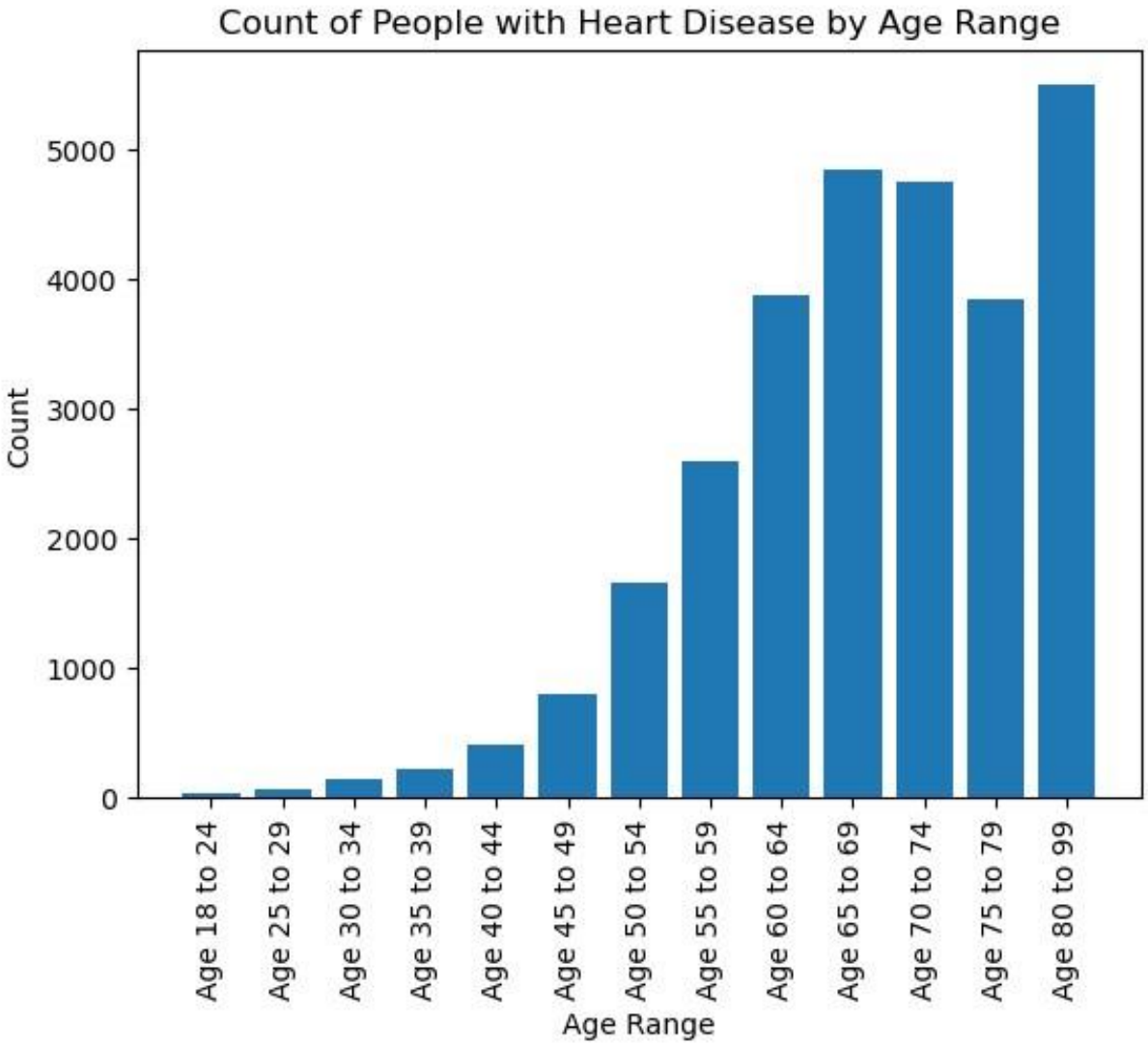
A correlation matrix has been employed to establish the numeric measure of the strength and direction of the linear relationship between the variables. This was summarised in the heat correlation map as can be seen below:



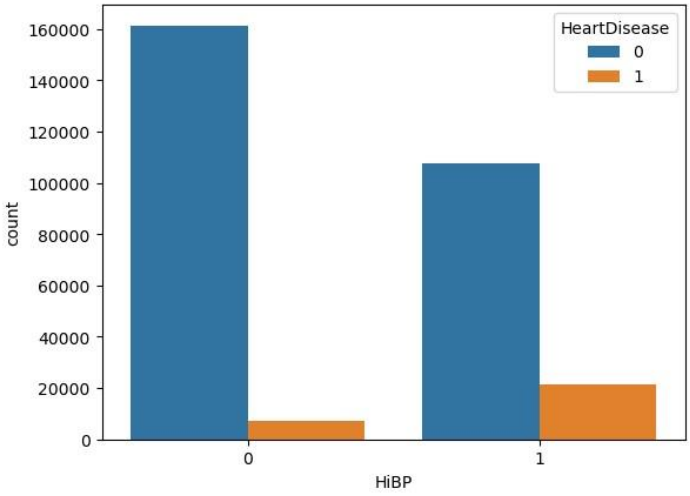
From the heat map, physical health ('Physh') and mental health days ('MentHlth') days have a moderate positive correlation of 0.35. Among all the variables, the following have a significant relationship with Heart disease; Age with 0.22 ,High BP (hiBP) with 0.21, stroke with 0.2, diabetes, high cholesterol levels, physical health are all 0.18 and smoker with 0.11. Age has a moderate positive correlation of 0.34 with high blood pressure which is in line with research in section 1. Diabetes is positively correlated with high blood pressure, high cholesterol levels and BMI by 0.27,0,21 and 0.23. Diabetes can be explored further as another target variable with the features it is significantly correlated with.

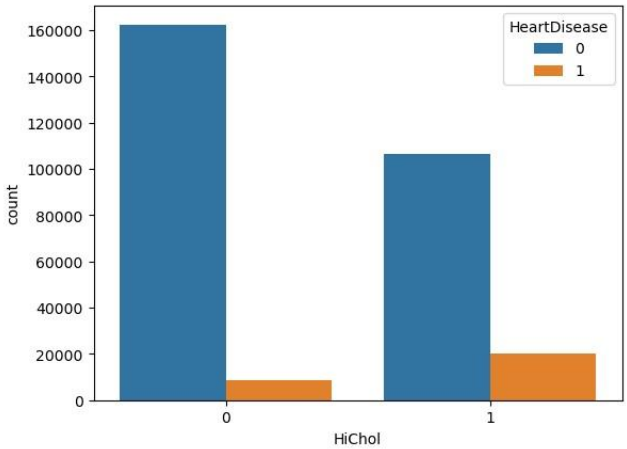
The next section will focus on visualizing some relationships outlined from the analysis of the correlation heatmap. A lot of focus will be on exploring the relationship between heart disease and some selected variables.

Age was a significant factor in predicting the risk of heart disease. This is in line with the research covered in section 1 of the report as it stated with aging brings major changes in the heart and blood vessel. A visualization of this relationship is shown below. There is an upward trend of age and heart disease presence ie as age increases, higher presence of heart disease.

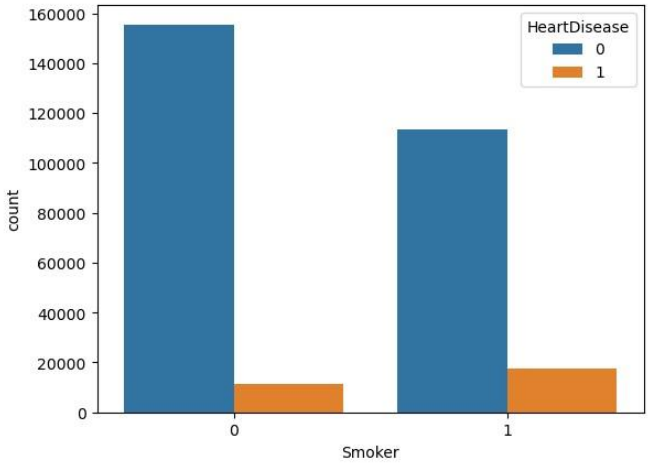


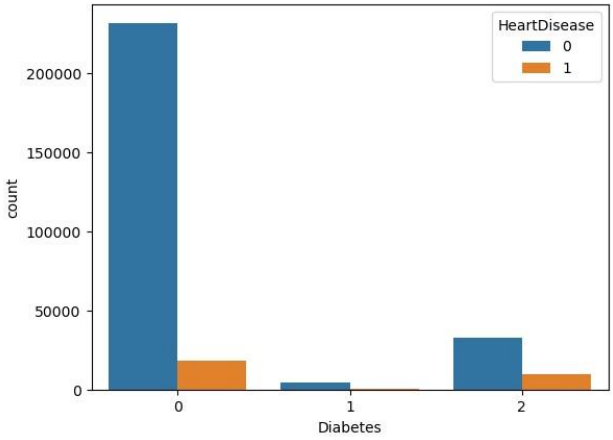
A visualization of the highly correlated variables with heart disease can be seen below. Afterwards, we will also explore and visualize some interesting variables with heart disease though having a weak correlation.





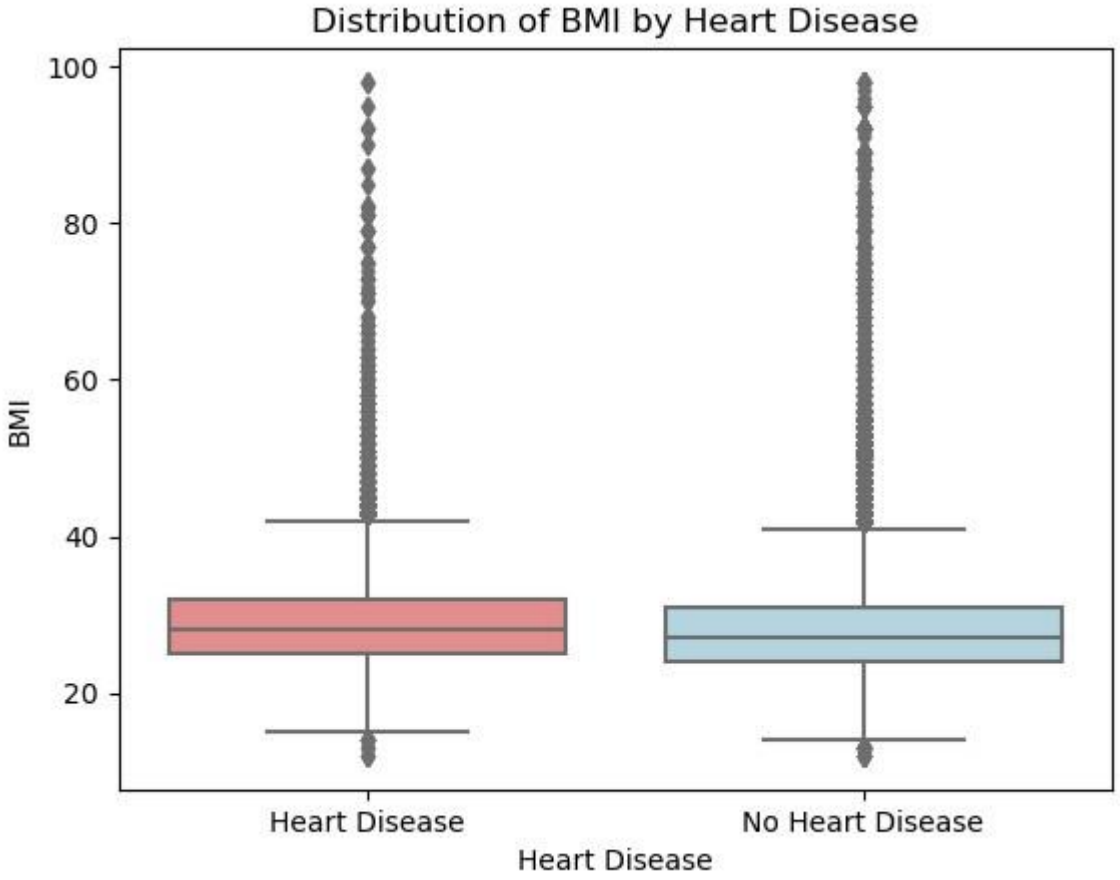
In the left graph, we can see the presence of high blood pressure is likely to put one at a high risk of heart disease as opposed to not having high pressure. Compare the orange bars which represent the presence of heart disease in either presence of high BP or without BP. To the right, High cholesterol levels can be associated with higher risk of heart disease. You interpret this similarly to the left graph for HiBP.





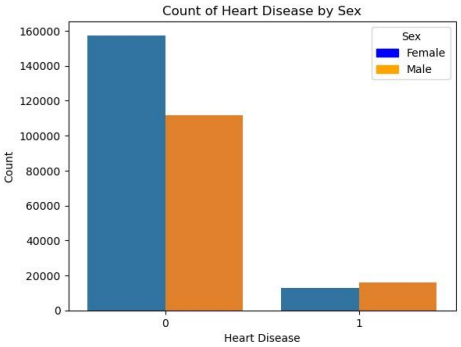
The graph on left shows a smoker is more at risk of getting heart disease as opposed to a non-smoker. Also, one with diabetes is more at risk of getting heart disease than one without as can be seen on graph to the right.

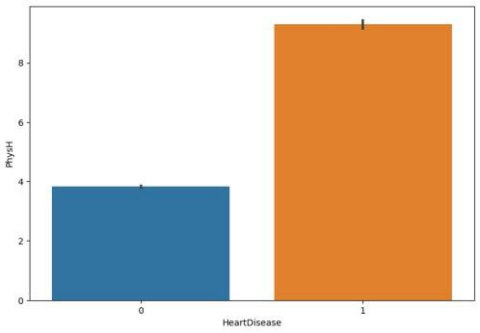
The literature review provided in section 1 has some interesting facts about sex and BMI with heart disease. Let us see if the data is in line with the research results.

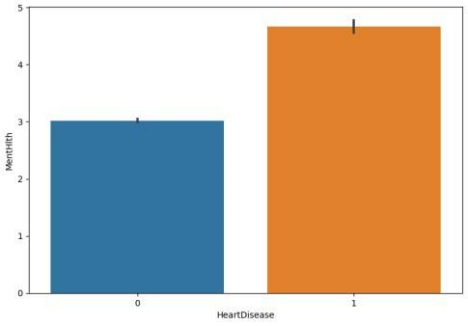


The graph above is in line with research findings as higher BMI values is associated with higher risk to heart problems. The box plot with heart disease shows a higher interquartile range and has a slightly higher mean.

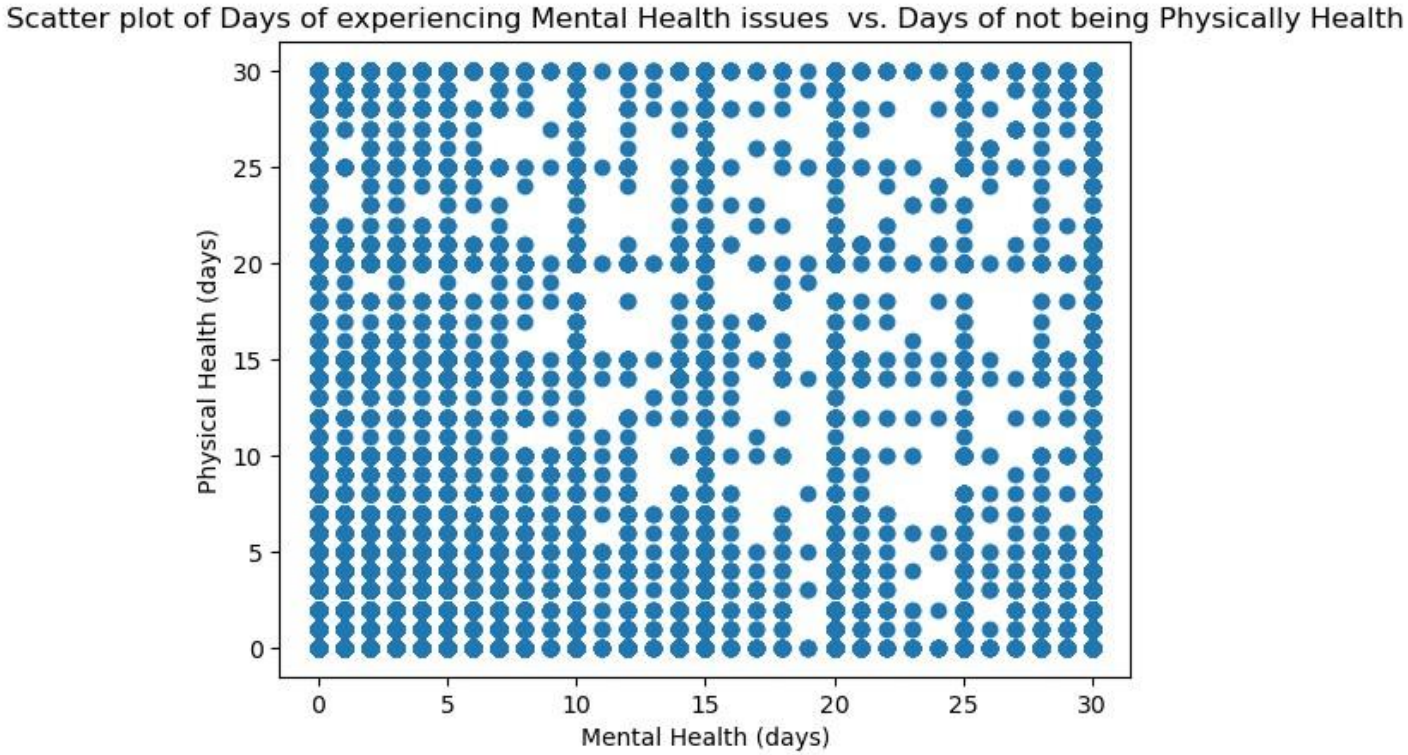
Males are at higher risk of having heart disease compared to females as can be seen in left most graph below. Additionally, one is at high risk of heart disease the more days they are physically unfit and more days they are mentally ill. This too is in line with the literature findings in section 1 of the report.



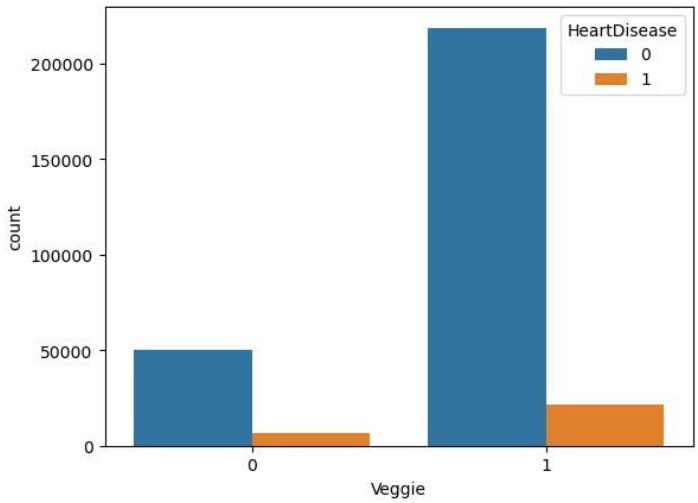




I tried to explore whether there exists a relationship between physical days of being unfit and mentally ill days. I did not standardize as both variables have the same scale which is days ranging from 0 to 30. As can be seen in scatter plot below, there is no relationship that can be observed.



Lifestyle has a role to play on increasing risks of heart disease. Let us examine whether eating vegetables does make a difference.



The results are quite odd as it shows not eating vegetables results in less risk of heart disease. There could be an error in the coding of this and therefore should be explored further. The correlation heatmap however showed a negative weak relationship between heart disease and eating vegetables as such the results of the graph are supposed to be the other way round.

CONCLUSION

The analysis of indicators of heart disease yielded several important findings. Age emerged as a significant factor, with a positive correlation observed between age and the risk of heart disease. This aligns with existing research that highlights the impact of aging on cardiovascular health. Additionally, high blood pressure, stroke, diabetes, high cholesterol levels, and smoking were identified as significant risk factors for heart disease. These findings underscore the importance of managing these conditions to mitigate the risk of developing heart disease. Diabetes also showed moderately high correlation numbers with some other variables in the dataset such as high cholesterol, BMI and BP. As such it can be made a target variable and further analysis can be done.

The results also highlighted the role of lifestyle factors in heart disease. Males were found to be at a higher risk compared to females, suggesting the need for targeted interventions for men. Furthermore, higher BMI values were associated with an increased risk of heart problems, emphasizing the importance of maintaining a healthy weight. It is essential to promote healthy lifestyle choices, such as regular physical activity and a balanced diet, to reduce the risk of heart disease.

While the analysis provided valuable insights into the relationships between various factors and heart disease, it is crucial to note that correlations and trends alone do not establish causation. To gain a more comprehensive understanding, further analysis using advanced tools, such as linear regression or machine learning models, should be considered. These methods can help identify the most influential factors and develop predictive models for assessing the risk of heart disease.

In summary, the findings highlight the significance of age, high blood pressure, stroke, diabetes, high cholesterol levels, smoking, gender, and BMI as indicators of heart disease. These insights can inform public health strategies, clinical interventions, and personalized approaches to prevent, diagnose, and manage heart disease. Continued research and exploration of additional analytical techniques will contribute to a deeper understanding of the complex interplay between these factors and the development of effective preventive measures for heart disease.

REFERENCES

Acadmin,2019. https://www.apolloclinic.com/blog/5-indicators-to-predict-your-heart-health/

Del Gaizo AL, Elhai JD, Weaver TL. Posttraumatic stress disorder, poor physical health and substance use behaviors in a national trauma-exposed sample. *Psychiatry Res* 2011;188(3):390–5.

Escobar E, Hypertension and coronary heart disease, J. Hum. Hypertens. 16 (1) (2002) S61–S63.

Guo F, He D, Zhang W, Walton RG. Trends in prevalence, awareness, management, and control of hypertension among United States adults, 1999 to 2010

JAMA,2016. Throughout life, heart attacks are twice ass common in men than women. https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-menthan-women

Klag MJ, Whelton PK, Randall BL, et al. Blood pressure and end-stage renal disease in men. N Engl J Med. 1996;334(1):13–8. [PubMed] [Google Scholar]

Nystoriak MA, Bhatnagar A. Cardiovascular Effects and Benefits of Exercise. Front Cardiovasc Med. 2018 Sep 28;5:135. doi: 10.3389/fcvm.2018.00135. PMID: 30324108; PMCID: PMC6172294.

R.G. Williams, G.D. Pearson, R.J. Barst, J.S. Child, P. Del Nido, W.M. Gersony, K.S. Kuehl, M.J. Landzberg, M. Myerson, S.R. Neish, et al., Report of the national heart, lung, and blood institute working group on research in adult congenital heart disease, J. Am. College Cardiol. 47 (4) (2006) 701–7

R. Huxley, F. Barzi, M. Woodward, Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies, Bmj 332 (7533) (2006) 73–78.

Sowden GL, Huffman JC. The impact of mental illness on cardiac outcomes: a review for the cardiologist. *Int J Cardiol* 2009;132(1):30–7.