Big Data Fundamentals (CS989)

ANALYSIS OF THE KEY INDICTORS OF HEART DISEASE

Table of Contents

List of Figures

Chapter 1: Identification of key challenge

Heart disease, or cardiovascular disease is an umbrella term for a wide array of complications that can occur within the cardiovascular system. The cardiovascular system consists of both the heart and its blood vessels. Heart disease can be divided into 4 conditions: Coronary heart disease, Cerebrovascular disease, peripheral artery disease and aortic atherosclerosis (Olvera et al, 2022). Heart diseases are the leading cause of death worldwide, with an estimated 17.9 million deaths each year. According to the CDC (Centers for Disease Control and Prevention), 47% of Americans have at least 1 out of 3 risk factors for disease; high blood pressure, smoking and high cholesterol. Other factors which are causing the rapid increase in heart disease is obesity, type 2 diabetes, physical inactivity, smoking and an excess consumption of alcohol (Walden et al, 2011). As well as physical health, an individual's mental health is also shown to have an effect on their risk of developing heart disease although this link has not yet been clearly established (Hert et al, 2018). In the last decade there has advances in the treatment of skin cancer and these systemic therapies have been associated with cardiac toxicities leading to an increased risk of heart disease (Wang et al, 2022). Finally, patients with kidney disease have an increased stress put on the heart due to the kidneys lack of function, also resulting in an increased risk (Said & Hernandez, 2014).

Many heart diseases can be prevented by detecting and addressing behavioural risk factors early on. This early detection of these factors can be the difference between life and death.

This report examines a 2020 census data in an effort to identify patterns from the data which can enable identification of individuals most at risk of heart disease. The data exploration showcased below will add validity to the evidence that lifestyle factors are a key contributing factor to being predisposed to heart disease.

Chapter 2: Introduction to the dataset

This data was obtained from the freely available Kaggle website. It originally came from the CDC and is an important part of the Behavioural Risk Factor Surveillance System (BRFSS). This survey gathers data on residents in the US regarding their health-related risk behaviours and their chronic health conditions. The BRFSS collects data in all of the 50 states and conducts over 400,000 interviews every year, it the largest continuously conducted health survey system in the world.

The original survey data collected from CDC contained answers from 400k adults, the dataset was made smaller and now contains 320K rows and 18 columns. The survey was conducted over the phone and asked US residents on their current health status. The features that were collected from each patient are shown in Table 1.

No.	Feature	Description
1	HeartDisease	Respondents ever reported having coronary heart disease or myocardial infarction
2	BMI	Body mass index (BMI)
3	Smoking	Has the respondent smoked at least 100 cigarettes in their entire life
4	AlcoholDrinking	Is respondent a heavy drinker (Men – more than 14 drinks a week, Women- more than 7 drinks a week)
5	Stroke	Has respondent ever had a stroke
6	PhysicalHealth	How many days during the past 30, was physical health not good?
7	MentalHealth	How many days during the past 30, was mental health not good?
8	DiffWalking	Does respondent have difficulty walking or climbing stairs?
9	Sex	It's respondent male or female
10	AgeCategory	14 level age category
11	Race	Race/ethnicity of respondent
12	Diabetic	Has/does respondent have diabetes
13	PhysicalActivity	Has respondent done any physical activity during past 30 days?
14	GenHealth	What is general health of respondent?
15	SleepTime	How many hours of sleep does respondent get in a 24 hour period?
16	Asthma	Does/Has respondent ever had asthma
17	KidneyDisease	Has respondent ever had kidney disease?
18	SkinCancer	Has respondent ever had skin cancer?

Table 2.1: Features of patients collected in survey, most were yes or no questions.

Chapter 3: Summary of dataset

As described before the dataset consists of 320K rows and 18 columns.

General analysis of the data is performed by describing each figure produced giving an idea of what factors increase the risk of heart disease

Analysis began with getting an overview how of equal the dataset was in terms of heart disease. Figure 1 shows the distribution of respondents affected by heart disease vs unaffected. It is clearly obvious that the dataset is unbalanced, displaying 292422 respondents who don't have heart disease and only 27373 respondents who do. Further on, this dataset will be balanced for better, more accurate comparisons.

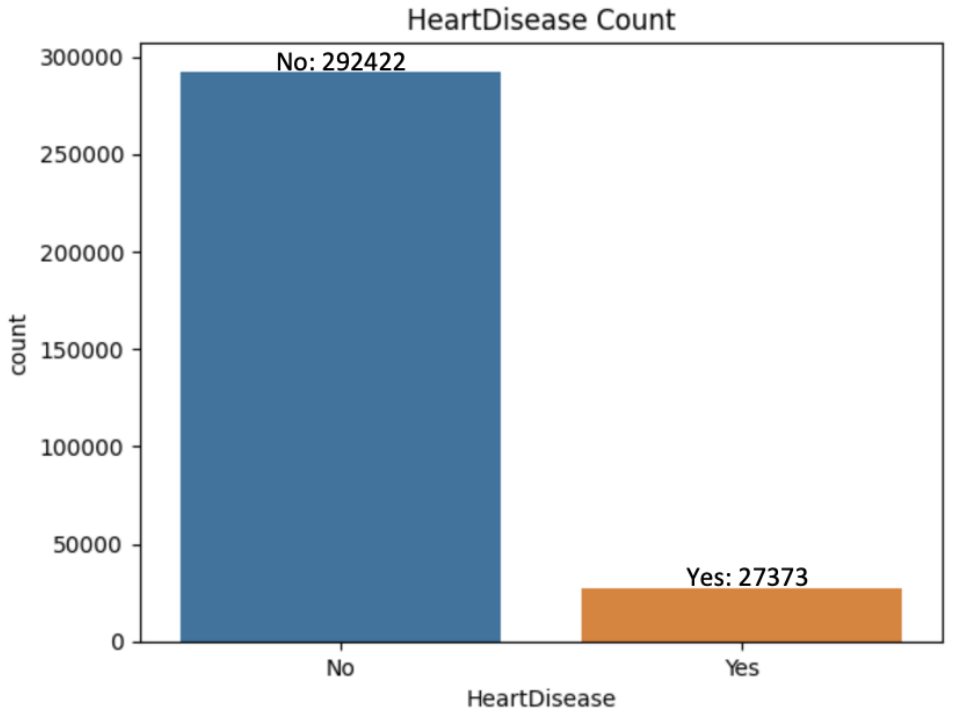

Figure 1:Heart Disease Count

The gender composition of the dataset was then investigated. First by looking at the whole dataset and then only people who were known to have suffered from heart disease. Figure 3.2 shows that the dataset was considerably balanced, with 52% of respondents' female and 48% respondents' male (figure 2 (i)). However, despite the dataset being made up of mostly females, more males suffered from heart disease than females, shown in figure 2(ii). This is unsurprising as men generally are more likely to develop heart disease than women (Bots et al, 2017).

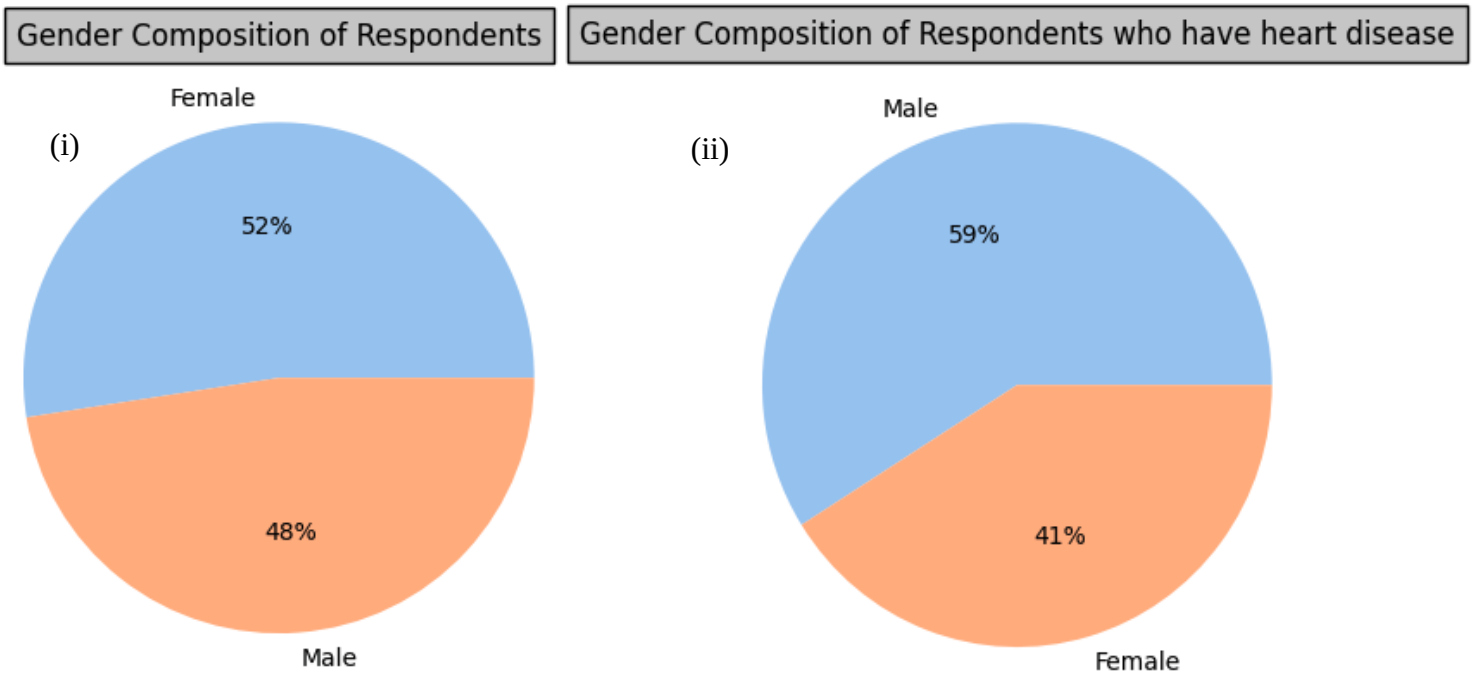

Figure 2: Gender split in the whole dataset vs in respondents who have heart disease.

Continuing to only look at respondents with heart disease, their general health was investigated to identify any correlation with heart disease and poor general health, as these factors are usually related. Figure 3 shows the distribution of responses. Although there are very few people with excellent health which is expected (5%), surprisingly most of the respondents have overall good general health within this data (35%). Only 14% said their general health was poor.

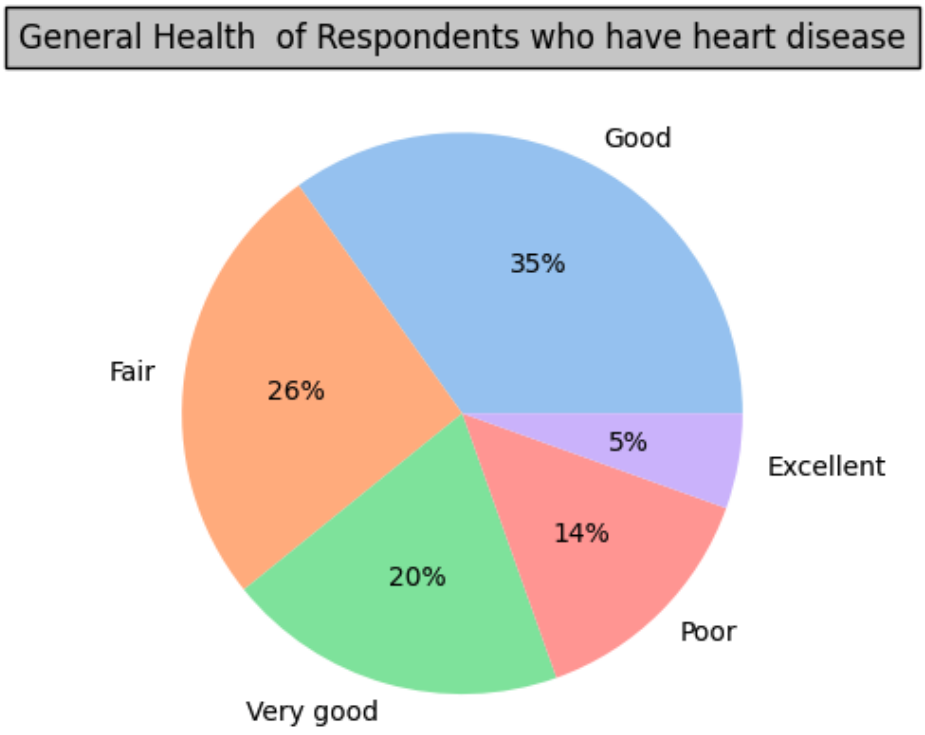

Figure 3: General Health of respondents with Heart disease

Race was the next factor investigated. Using only the respondents with heart disease (27,373), a bar chart was plotted against race, shown on Figure 4. First observations of this graph suggest that respondents being white puts them at a much higher risk of heart disease. However, a pie chart was produced to identity the race composition of the whole dataset. This is displayed on figure 5 and clearly shows that the race distribution of this dataset is unbalanced, with most respondents being white (77%) and the other 5 all below 10%. Therefore, it is expected that most people with heart disease would be white compared to other races. Due to this unequal composition of race, no further analysis will be performed using this feature.

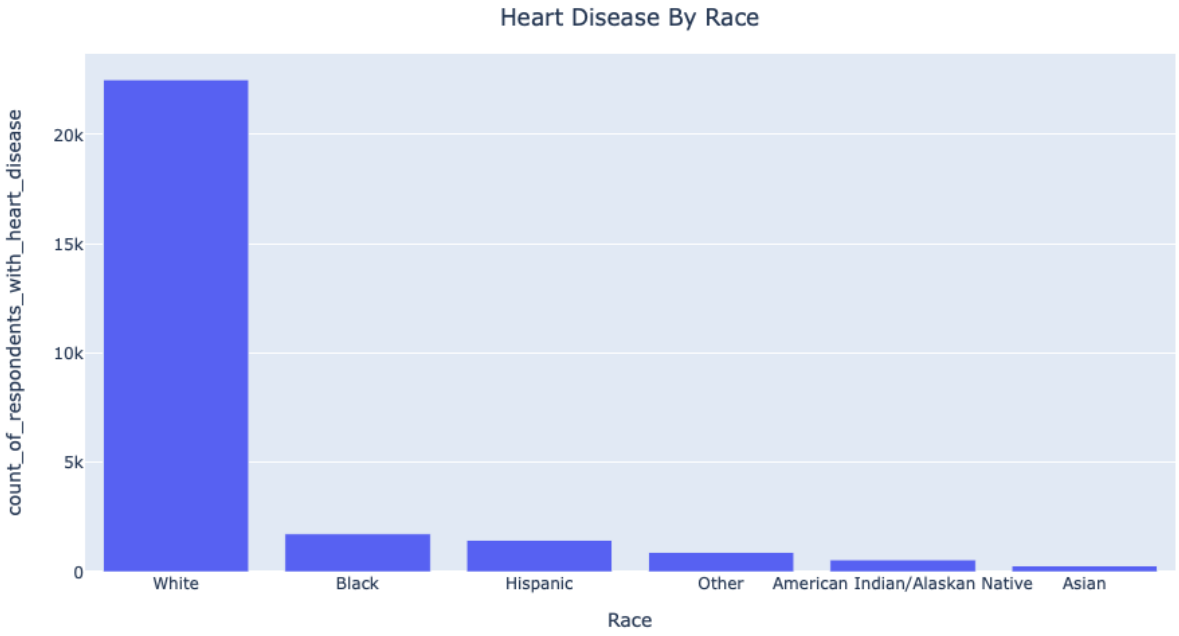

Figure 4: Race composition in respondents with heart disease

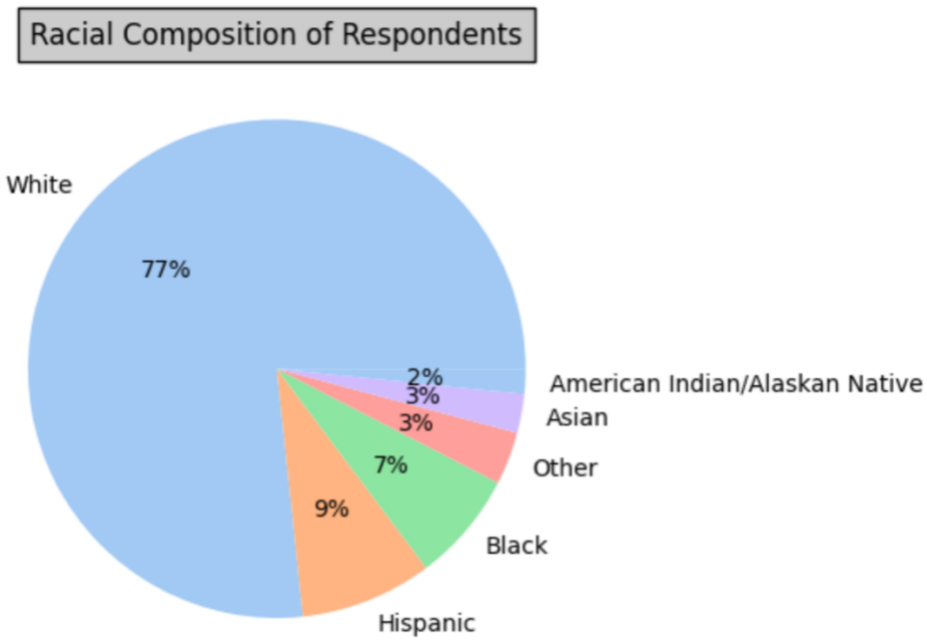

Figure 5 :Race composition of respondents in whole dataset.

Next steps in the analysis of this dataset were to compare gender vs age group in both respondents with heart disease and without. As described before, the heart disease to no heart disease ratio is unbalanced. So, to balance this, a sample size of 27373 people from the unaffected respondents was taken, this sample is the equal to the number of people with heart disease. However, the gender composition of this new sample of unaffected respondents had to be balanced as well, so 16139 of each gender was taken from the unaffected heart disease respondents and concatenated to form the sample. The age composition of this new formed sample was then inspected (Figure 6) and appeared to be relatively balanced.

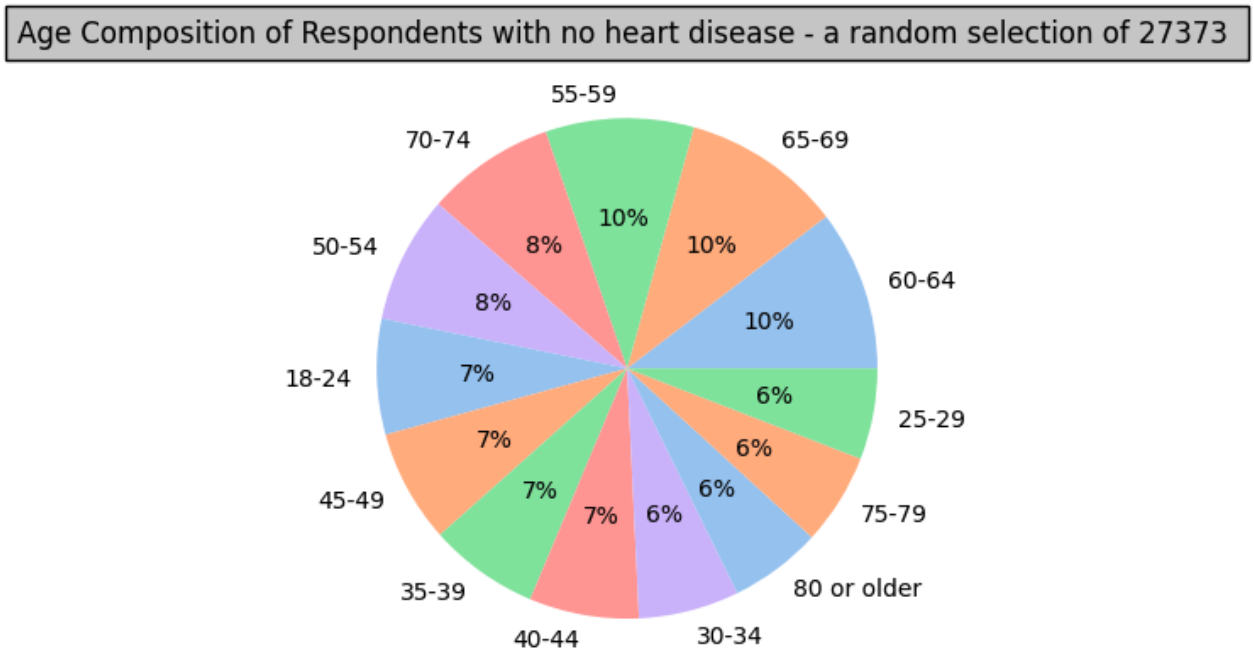

Figure 6 :Age Composition of a random selection of respondents with no heart disease (27,373).

After having 2 balanced samples of heart disease and no heart disease, bar charts were produced describing the age and gender in both respondents with heart disease and without. Figure 7 shows heart disease by gender and age group. From this, there is an obvious increase in males who have heart disease across all ages compared to women, this is expected as it's been shown earlier that males are more likely to suffer from this condition. There is also a noticeable difference in the age groups, there are very few people in the age groups 18-24, 25-28 and 30-34 with heart disease. In specifically 18-24 year olds, only had 130 people out of the 27373 people with heart disease. After the age of 40, heart disease cases start to rise gradually up to ages 70-74 for both genders. This rises from 486 people in the 40-44 age range to 4847 in the 70-74 age range. This data perfectly describes the effect of age on risk of heart disease, it is well known that increasing age increases a persons risk of developing heart disease. This is due to the changes in the heart and blood vessels as you get older (Greiser et al, 2005). After age 74, the number of people with heart disease drops slightly. However, at the of 80 and over, there is a significant increase in respondents with heart disease patients in females particularly. This observation matched up studies describing that women are more likely over 80 are more predisposed to heart disease then men of the same specific age group (Rodgers et al, 2019).

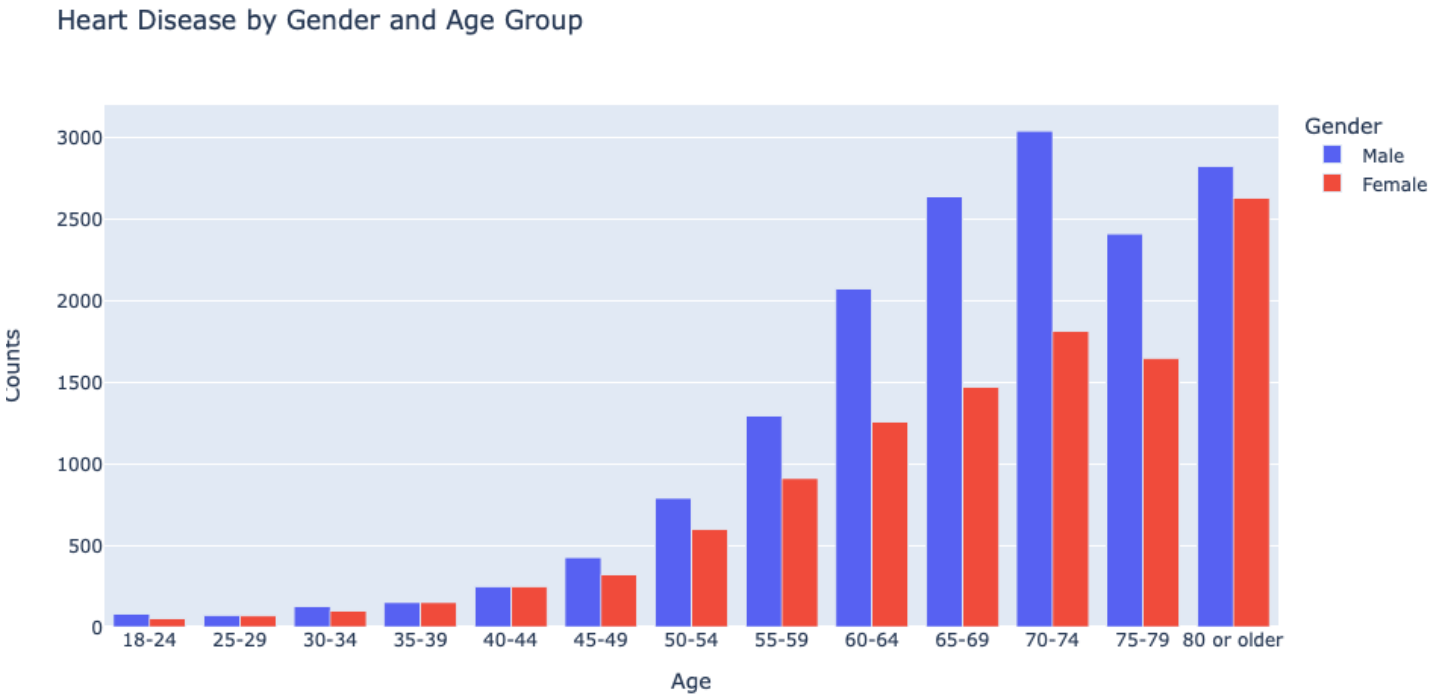

Figure 7: Heart Disease respondents by gender and age

Figure 8 shows the random selection of 27,373 respondents without heart disease against their age and gender. From direct observation, all age groups show somewhat equal proportions of no heart disease counts. This further confirms the impact age has on heart disease when comparing it to figure 7 above.

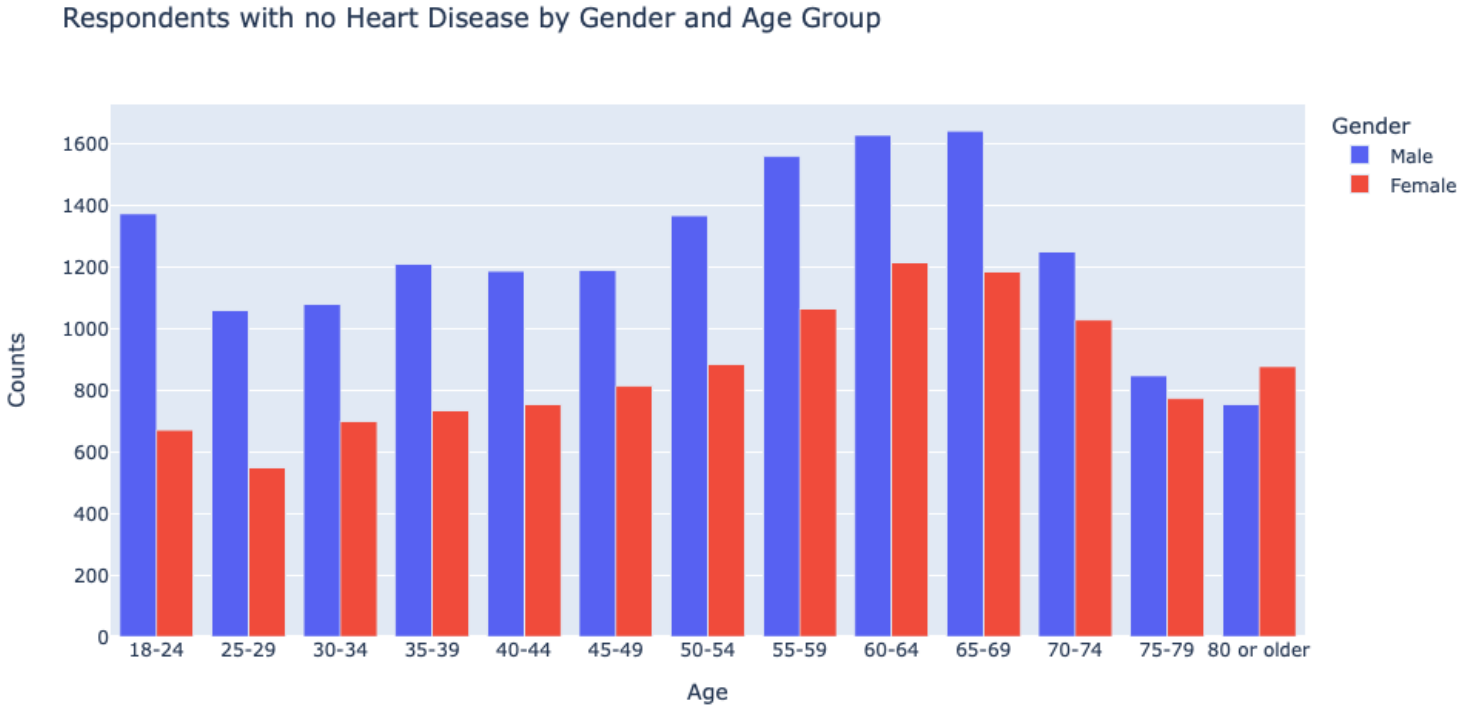

Figure 8:Respondents with no heart disease by gender and age

Using the same two balanced samples, lifestyle habits were plotted along with gender. Figure 9 shows the respondents that have heart disease by lifestyle habits: smoking, alcohol drinking, difficulty walking and physical activity. While figure 10 shows the respondents that don't have heart disease by lifestyle habits.

The number of respondents smoking with heart disease (16,037) is much higher compared to respondents who don't have heart disease (11,092). These findings are unsurprising as smoking has been associated with a 2-to-4-fold increased risk of heart disease (Lakier, 1992). There were also more males who smoke in both heart disease and non-heart disease patients. This is also expected as males were more prone to heart disease.
A similar trend was seen for the difficulty walking feature. The number of respondents who have difficulty walking was much higher in the heart disease sample at 10,028 compared to respondents in the non -heart diseased sample at only 3,107.
Physical activity was slightly higher in patients with who don't have heart disease (21,745) compared to patients who do (17,489). The trends seen in these two factors are predictable as not doing enough physical activity can lead to an increased risk of heart disease (Kubota et al, 2017).
The final lifestyle compared was alcohol drinking, there was no significant difference seen between respondents who have heart disease vs respondents who don't.

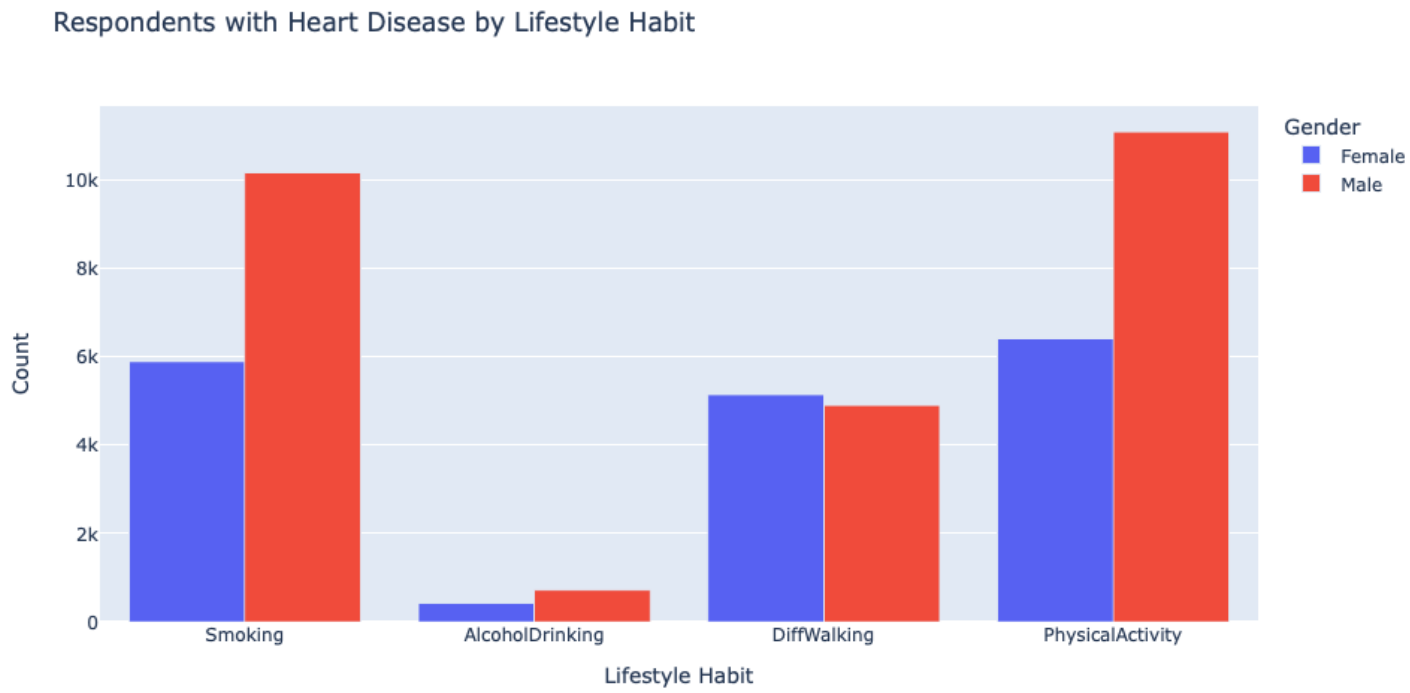

Figure 9: Respondents with heart disease by life style habit

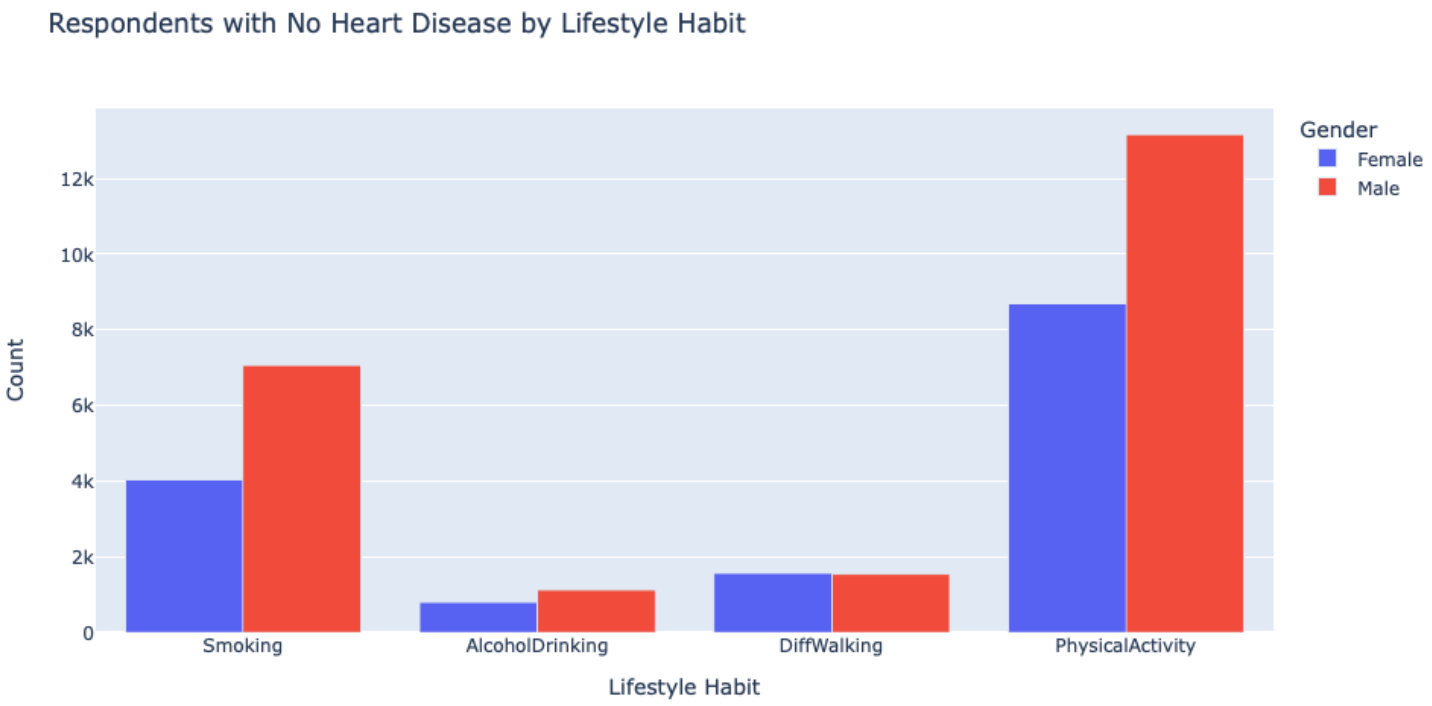

Figure 10: Respondents with no heart disease by life style habit

Chapter 4: Unsupervised Machine Learning

The aim of unsupervised machine learning is to use algorithms to find hidden patterns in data. Unlike supervised, unsupervised involves no training and the data is passed to the algorithm unlabelled. Using mathematical computation, it discovers interesting patterns within the data. These patterns typically form clusters. In this case the data passed to the algorithm were the lifestyle factors in the dataset. These were smoking, alcohol drinking, difficulty walking, physical and mental health.

Agglomerative Clustering

Hierarchical clustering works by grouping data into a tree of clusters. There are two types known as agglomerative and divisive. Agglomerative was used for the analysis of this data, it uses a bottom-up approach, each data point starts in its own cluster and then groups are formed by merging similar clusters together located nearby Euclidean space. A dendrogram captures this form of aggregation into clusters. It therefore enables the identification of optimal number of clusters to use for the agglomerative clustering (Figure 11).

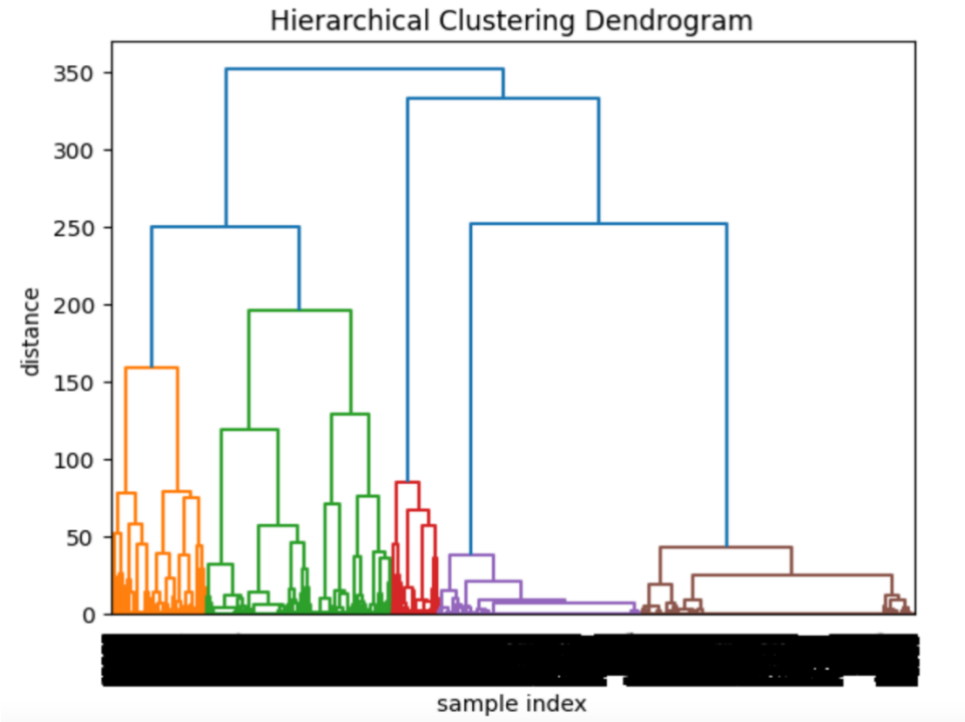

Figure 11: Dendrogram produced for agglomerative clustering (5 clusters identified). Ward linkage* used and Euclidean distance.

As can be viewed from the dendrogram above, many of the datapoints seem to cluster well whereas the datapoints on the right are more spaced out, indicating greater dissimilarity.
The agglomerative hierarchical algorithm was then executed with the clustering parameter set at 5 as identified by the dendrogram.
Agglomerative clustering is optimal with data that is not high dimensional. Therefore, the 5 lifestyle features were reduced to 3 using the Principal Component Analysis (PCA) algorithm. This also enables easier visualisation.

*See Appendix

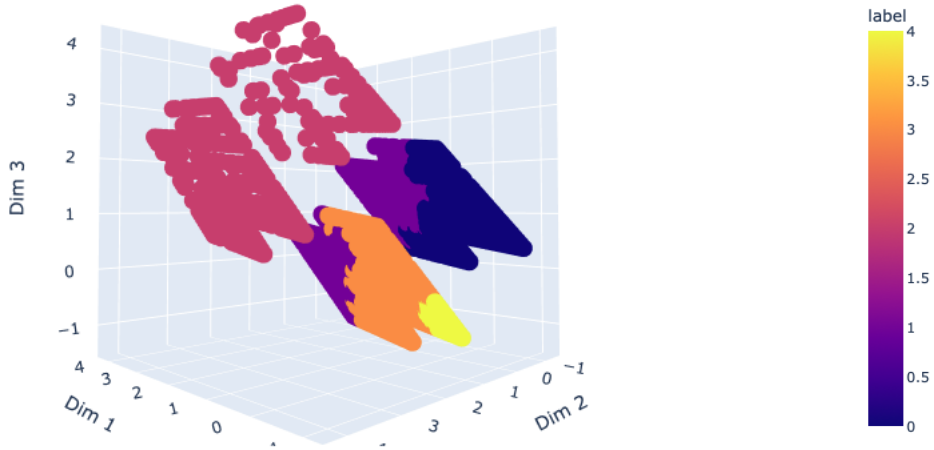

Figure 12: 3D scatter plot of clusters determined using agglomerative clustering.

The proximity of data points within the 5 clusters is not easily appreciated in Figure 12 above due to the volume of the data at 54000 rows. However, silhouette score* of **0.61** is indicative of clusters that are distinctively grouped.

K-Means, another unsupervised machine learning was then also used to show distribution of clusters and was compared with figure 4.2. K-Means is known to work better on large datasets compared to agglomerative clustering. First to determine the optimal number of clusters for k means, an elbow graph was produced (Figure 13).

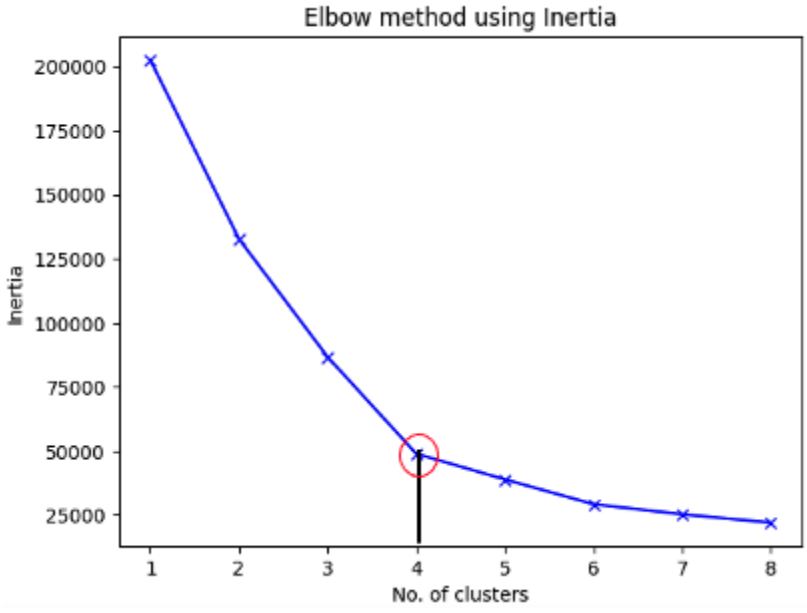

Figure 13: Elbow graph produced to determine optimal number of clusters for K-Means.

Contrary to the data from the dendrogram, the elbow graph appears to show 4 as the optimal number of clusters. This number of clusters was then passed to the K-Means algorithm and the results shown on a scatterplot (Figure 14).

*See Appendix

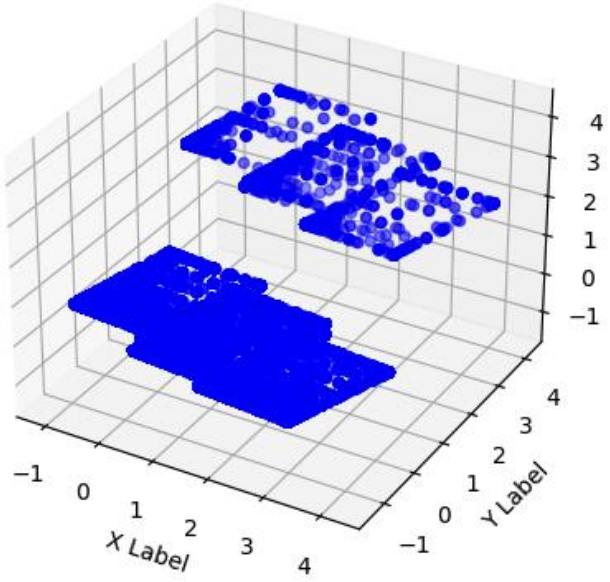

Figure 14: 3D scatter plot of clusters determined using K-Means algorithm.

Again, all 4 clusters cannot be fully visualised due to volume of the data and consequently poorer appreciation of proximity of data points. Silhouette score for K-Means was the also same as for agglomerative clustering at **0.61**, suggesting appropriate clustering.
Both K-Means and agglomerative clustering outputs showed similarities among the lifestyle features of the dataset. Next steps would include using these scatter plots and marking on them which of these data points have the target variable with heart disease or no heart disease.

Chapter 5: Supervised Machine Learning

The thrust behind supervised machine learning is the use of training algorithms to learn the relationships and patterns within the data. For this dataset, the aim was to use these algorithms to predict whether the respondent had heart disease based on their underlying conditions. The features used were Asthma, Kidney Disease, Skin Cancer, Diabetic, Stroke and also including their age and gender. Logistic regression methods and decision tree were executed with training/test set of 75% and 25 % respectively.

Logistic regression

Logistic regression was chosen as it's easy to interpret and efficient to train. It's also commonly used when the target variable, In this case, heart disease is a categorical in nature. Logistic regression predicts a binary value (0,1) based on the inputs into the model (in this case underlying conditions). All features first were converted into a numerical representation using the LabelEncoder from sklearn. A confusion matrix was output to show the performance of the model. This is shown on figure 15. 72% of the test set that was placed in the right category. True positive being 4607 records and true negatives being 5297 records. This is known as the models accuracy. The exact performance evaluation is shown by the classification report in table 5.1. Precision is the number of items that were positive, and negatives correctly identified. This was 75% of no heart disease patients and 71% of heart disease patients. Whereas Recall, is the number of items which were positive and identified as positive. This was 68% of no heart disease patients of 77% of heart disease patients. Finally, the F1 value is an average of precision and recall.

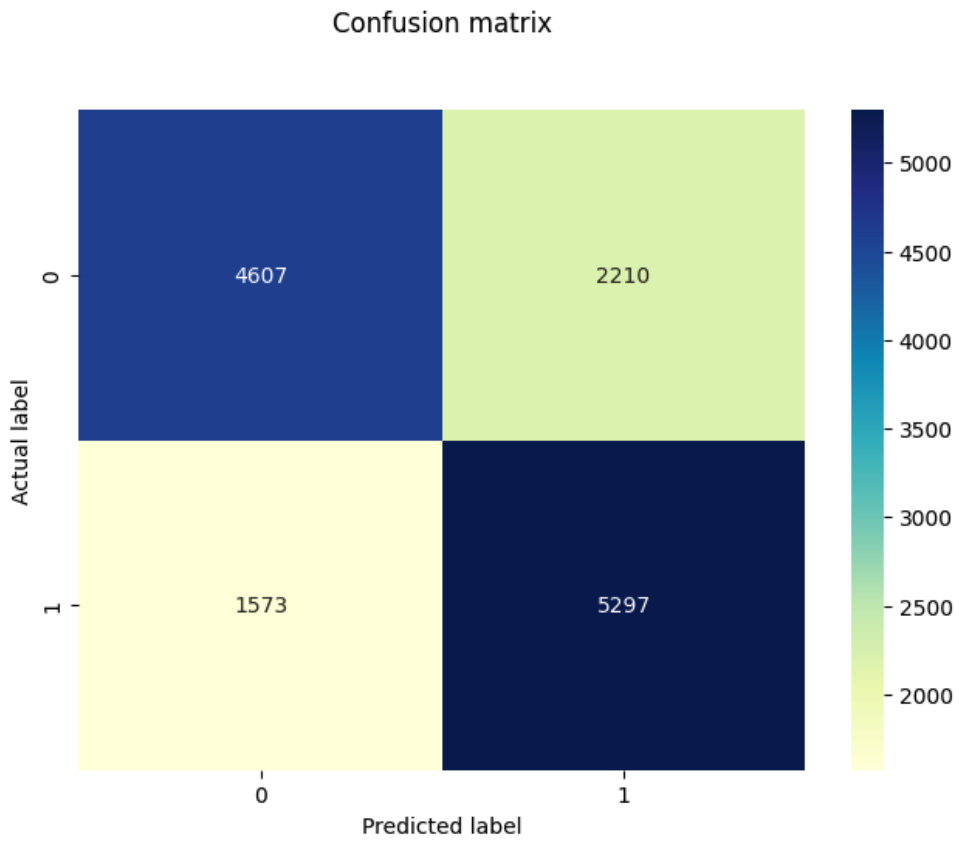

Figure 15: Heat map of confusion matrix, made using Logistic regression

	Precision	Recall	f1-score	Support
No heart disease	0.75	0.68	0.71	6817
Heart disease	0.71	0.77	0.74	6870
Accuracy				72%

Table 5.1: Classification report from logistic regression

Decision tree, another type of supervised machine learning algorithm was then used in order to compare with the results from logistic regression. The decision tree classifier is able to handle categorical values such as the values in the heart disease column. The decision tree classifier works by making predictions based on how a previous set of questions were answered. It was applied to the data, again using 25% of the data to test and a confusion matrix was generated shown in figure 16.

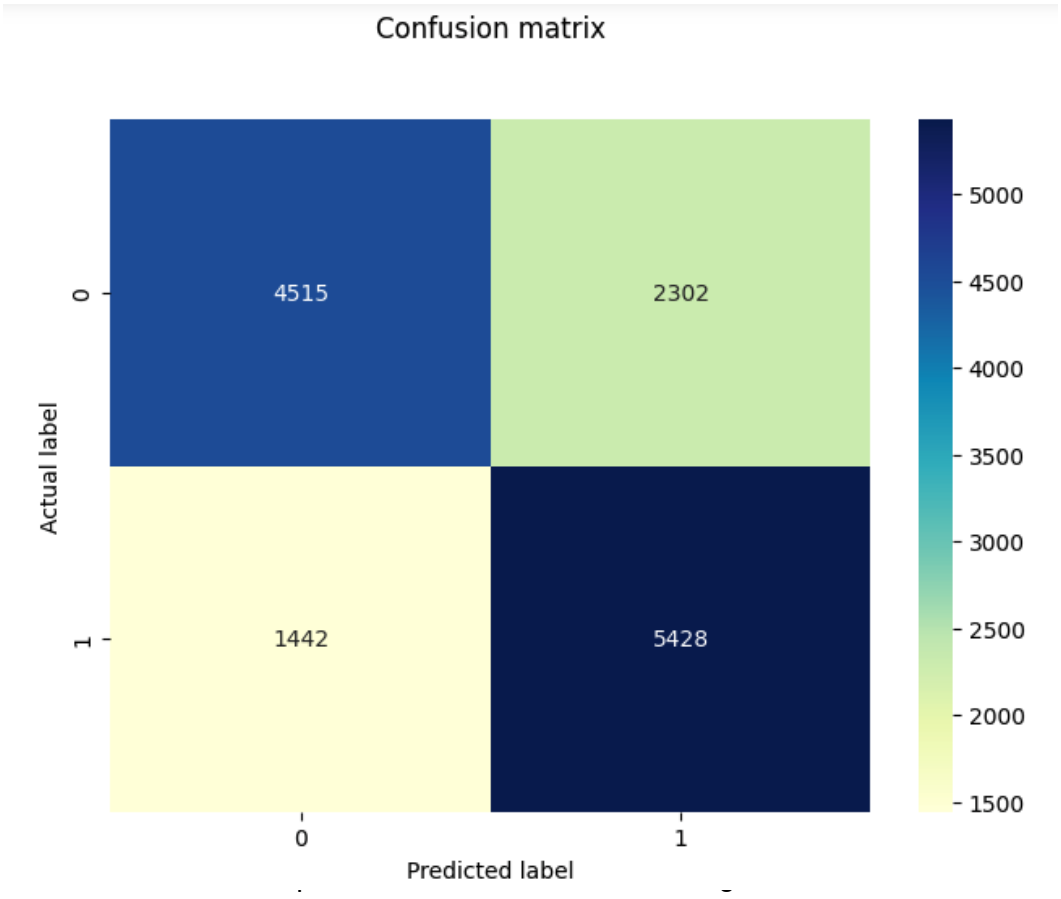

Figure 16 : Heat map of confusion matrix, made using Decision Tree

	Precision	Recall	f1-score	Support
No heart disease	0.75	0.68	0.70	6817
Heart disease	0.71	0.77	0.73	6870
Accuracy				73%

Table 5.2: Classification report from logistic regression

By comparing, it can be seen that both methods performed very similarly. The decision tree performed better in predicting the true negative values showing more values in this category (5,428) than the logistic regression (5,297). There were also fewer false positives using the decision tree classifier (1,442) compared to logistic regression (1573). However, decision tree performed worse for the true positive and false negative categories. There were fewer true positive values and more false negatives. The classification report confirms the highly similar results with both decision tree and logistic regression having the same F1 value for both heart disease and no heart disease. Neither one performed more favourably.

Chapter 6: Reflections and Conclusions

After the analysis of this data, it can be confirmed that there are many contributing factors of heart disease. Exploratory data analysis of this dataset showed that:

- People older than 50 tend to have heart disease than those younger than 50.
- Smoking habits and a difficulty in walking contribute to a higher risk of heart disease
- Males have a higher chance of heart disease than females
- Alcohol drinking does not contribute to risk of heart disease
- General health didn't to seem affect risk of heart disease

With the benefit of hindsight, it may have been better to have chosen a different dataset with more balance of features. Not only was heart disease vs no heart disease counts unbalanced but also the race composition, preventing any solid conclusions from this feature.

Supervised machine learning methods performed well to correctly categorise respondents with heart disease and without, suggesting that these underlying conditions do have an impact on a persons risk. The decision tree classifier was the better choice for a second supervised method compared to linear regression. Linear regression is ideally used when the dependent variable is numerical rather than categorical, unlike the data used here.

Unsupervised machine learning provided strong indication that there were relationships within the data. This was shown up in well-formed clusters (denoted by the silouttes scores). As Hierarchical clustering is known to be sub optimal with large datasets, Kmeans was also executed to confirm performance of the clustering. The target variable (in this case heart disease/no heart disease) is typically unused in clustering algorithms. Hence, further analysis is required to gauge how the clusters formed relate to the central question in this case being: Do the clusters bear a strong relationship to the cases(records)of heart disease/no heart disease?

The dataset used was replete with a rich set of features (relating to underlying condition/ lifestyle factors). This renders it ideal for supervised machine learning as the training algorithm is given a rich set of features to learn relationship and connect it with the target variable (heart disease/no heart disease).

Appendix

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

Python: 2.7.16
jupyter_client: 7.3.5
jupyter_core: 4.11.1
jupyter_server: 1.19.1
jupyterlab: 3.4.7

Packages used:

- Pandas
- Plot.ly
- Seaborn
- matplotlib
- Sklearn – metrics/cluster/PCA/AgglomerativeClustering/make_blobs/KMeans/LabelEncoder
- Sklearn.preprocessing – scale
- Sklearn.linear_model – logistic regression/ decision tree
- Sys
- Numpy
- Scipy- dendrogram/sch

Ward linkage - minimising the sum of squared distances between datapoints within clusters.

Silhouette score – Measure of how similar the data point is to its own cluster compared to the other clusters. The closer to 1, the better.

References

Bots, S.H., Peters, S.A.E. and Woodward, M. (2017). Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. *BMJ Global Health*, [online] 2(2), p.e000298. doi:10.1136/bmjgh-2017-000298.

Greiser, K.H., Kluttig, A., Schumann, B., Kors, J.A., Swenne, C.A., Kuss, O., Werdan, K. and Haerting, J. (2005). Cardiovascular disease, risk factors and heart rate variability in the elderly general population: Design and objectives of the CARdiovascular disease, Living and Ageing in Halle (CARLA) Study. *BMC Cardiovascular Disorders*, 5(1). doi:10.1186/1471-2261-5-33.

Lakier, J.B. (1992). Smoking and cardiovascular disease. *The American Journal of Medicine*, 93(1), pp.S8–S12. doi:10.1016/0002-9343(92)90620-q.

KUBOTA, Y., EVENSON, K.R., MACLEHOSE, R.F., ROETKER, N.S., JOSHU, C.E. and FOLSOM, A.R. (2017). Physical Activity and Lifetime Risk of Cardiovascular Disease and Cancer. *Medicine & Science in Sports & Exercise*, [online] 49(8), pp.1599–1605. doi:10.1249/mss.0000000000001274.

Edgardo Olvera Lopez and Jan, A. (2019). *Cardiovascular Disease*. [online] Nih.gov.

Walden, R. and Tomlinson, B. (2011). *Cardiovascular Disease*. 2nd ed. PubMed.

Wang, C.Y., Zoungas, S., Voskoboynik, M. and Mar, V. (2022). Cardiovascular disease and malignant melanoma. *Melanoma Research*, [online] 32(3), pp.135–141. doi:10.1097/CMR.0000000000000817.

De Hert, M., Detraux, J. and Vancampfort, D. (2018). The intriguing relationship between coronary heart disease and mental disorders. *Dialogues in Clinical Neuroscience*, [online] 20(1), pp.31–40. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6016051/.

Said, S. and Hernandez, G.T. (2014). The link between chronic kidney disease and cardiovascular disease. *Journal of nephropathology*, [online] 3(3), pp.99–104. doi:10.12860/jnp.2014.19.