Aastha Bhatt May 3, 2023

Title:
Heart Disease Patients Characteristics in 1988

Abstract:

Cases of heart diseases are increasing mostly due to the lifestyle of the people today. “One in five people die”in the United States due to a heart disease, speaking in terms of time, “one person dies every 34 seconds” from heart disease (CDC, 2022). Thus, it is important to know the characteristics of heart disease patients which will inform us the people who are at risk. Also, we will be able to diagnose heart disease early allowing the person to get the right medications and treatments. In order to diagnose quickly, people need to be aware of their symptoms and what they indicate towards.

This dataset provides 1025 records of patients mostly from Cleveland. The dataset provides the age, gender, and the other 11 characteristics. The characteristics are the following: type of chest pain, resting blood pressure, cholesterol, whether the fasting blood sugar is higher than 120, resting electrocardiographic, maximum heart rate, exercise induced angina, ST depression, slope of ST depression, number of major vessels, and the range of thalassemia.

For this dataset, the best k value for clustering is 7. Cluster 5, consisted of older people and those with very high cholesterol, double the healthy level. Cluster 6 consisted patients who had a median of 124 blood pressure which is slighlty higher than the healthy blood pressure range and the cholestrol level was about 190 which is in the healthy range.

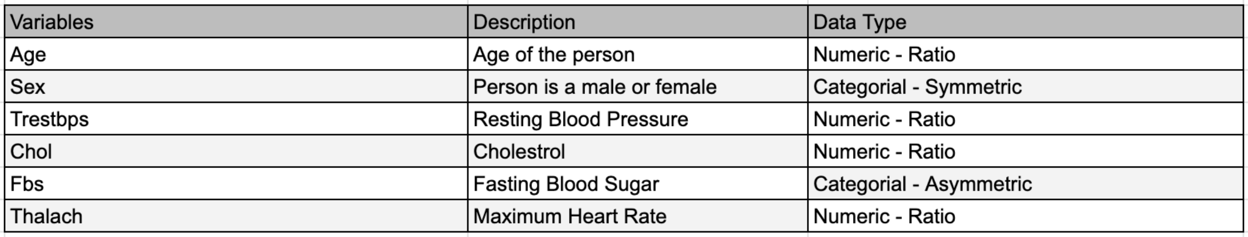
Introduction:

The dataset discussed in this report is the characteristics and symptoms patients had. This report will discuss the k-means clustering method to cluster the data with the characteristics such as age, blood pressure, and etc. The purpose of trying to cluster the data is to understand the common characteristics of heart patients and based on that bring awareness to people that if they have them, then they need to change their lifestyle and seek a doctor. If they are showing multiple characteristics that likely cause heart disease then seeking the doctor early means, early on diagnosis which leads to sooner treatment.

The report will first discuss the source of the data and describe the variables of the dataset. After that, describe the methods and the steps taken to produce the results and findings. After discussing this, the results of the tests and the code ran will be shown through tables and graphs. In the conclusion section, the findings, limitations, and future recommendations on further analysis will be discussed.

Data and Data Preparation:

I received this dataset from Kaggle, and was uploaded three years ago by Chernegs in 2019. This dataset was donated in 1988 to University of California Irvine Machine Learning. The data in the dataset mostly comes from patients in Cleveland. There were a lot of variables, but I trimmed down the number of variables which are characteristics of patients to six. The other characterstics were specific such as the slope of ST depression, type of chest pain, range of thalassemia, and more as listed in the abstract section of the paper. I will discuss the six variables in depth and the data type below. The attribute “age” is a numerical- ratio data type that lists the age of the patient. The sex attribute, is a categorical - symmetric data type, with a range of two values which are Male and Female. The third variable is “trestbps” which tells the resting blood pressure of the patient, this is a numerical-ratio data type. The fourth variable is “chol”, this tells the cholesterol level of the patient, and is a numerical-ratio data type. The fifth attribute is a categorical - asymmetric data type, this indicates whether the patients fasting blood sugar is greater than 120 or not. The last attribute is “thalach” which informs us the maximum heart rate of the patient, and this is a numerical- ratio data type. Below is a table, that organizes the variables, description, and the data type.



Variables	Description	Data Type
Age	Age of the person	Numeric - Ratio
Sex	Person is a male or female	Categorial - Symmetric
Trestbps	Resting Blood Pressure	Numeric - Ratio
Chol	Cholestrol	Numeric - Ratio
Fbs	Fasting Blood Sugar	Categorial - Asymmetric
Thalach	Maximum Heart Rate	Numeric - Ratio

There was no missing data so no preprocessing steps were done in that aspect.

Methods:

After importing the data, I did an exploratory analysis by running a summary of the data, part of this can be seen in Table 1. Then I looked at the number of patients for each cholesterol level visually through a dotplot as seen in Figure 1. This exploratory analysis helped me understand the variables in the dataset. Then I took out the following variables because they were very specific: exercise induced angina, ST depression, slope of ST depression, number of major vessels, and the range of thalassemia.

After cleaning up the data, I normalized all attributes on a scale between 0 and 1 using the rescale function. Then I ran k-means with different k values. The first four I ran were 2, 3, 4 and 5. Then, I evaluated the clusters by first seeing up a distance matrix for cluster statistics. Then I calculated the cluster statistics for k=2,3,4,5. Then, I evaluated the between and within cluster distances. Afterwards, I computed a silhouette plot for the above k values. To find the best number of clusters, I used the elbow method kmeans sse as seen in Figure 2 . Based on the plot, I decided to run the kmeans for k= 6,7,8. From running these kmeans I was able to obtain the average between, average within, within cluster sum of squares by cluster, and the average silhouette width for k=2,3,4,5,6,7,8 as seen in Table 2. Based on Table 2, the cluster statistics and the plot from SSE, I decided to chose k=7. Using the pairs function, I visualized the relationship between the variables as seen in Figure 3. Then I turned the k=7 cluster to a factor. Afterwards I explored the distribution of variables among each cluster using a box plot as seen in Figure 4 through 7.

Results:

Table 1 shows the results after running the summary of the six variables used to cluster the data. The first attribute is age, which we can see that the average is about 54, median is 56, and 3rd quartile is 66. This shows that the age of heart diesease patients is around 55. The average for gender is 0.69. In the dataset, if the patient was a male it was indicated by 1 and females were indicated by 0. So, the higher the average, that means there were male patients in the dataset. A .69 shows that there were a little more male patients than females. The third attribute is resting blood pressure and the average value is 131.6, this is higher than 120 which is the standard for measuring healthy resting blood pressure. Anything higher than 120 indicates high blood pressure. Thus, a lot of the patients had high blood pressure. The fourth variable is cholestrol and the average for this was 246. Healthy cholestrol is less than 200, meaning that the average of the patients had high cholesterol. Fasting blood sugar variable had an average of 0.1493. This variable had two ranges, if the patients fasting blood sugar was more than 120 then it was indicated by 1 and if not then 0. The average is lower than .5, showing that the average of patients had less than 120 for their fasting blood sugar levels. The last variable is the maximum heart rate whose average is 149.1, but based on the min, 1st quartile, median, 3rd quartile, and max, the range for maximum heart rate seems to vary a bit. Another part of the exploratory analysis was to visualize the number of patients per cholestrol level which can be seen in Figure 1. Figure 1 shows that there were more patients with high cholestrol, infact as the cholesterol level increased on the graph, the number of patients with that level also increased.

Table 2 shows the cluster statistics of of the various k-mean values. The first k-mean is k=2, and the average between was 93.69, the average within was 53.49, within cluster sum of squares by cluster was 46.10% and the average silhouette width was 0.4. This shows that there was an average distance between the two clusters of 93.69 and the distance within the cluster itself was 53.59. The second k-mean ran is k=3, and the average between was 86.96, the average within was 46.966, the within cluster sum of squares by cluster was 58.2% and the average silhouette width was 0.3. This shows that there was an average distance between the three clusters of 86.96 and the distance within the cluster itself was 46.96. The third k-mean ran is k=4, and the average between was 82.81, the average within was 42.99, the within cluster sum of squares by cluster was 64.50% and the average silhouette width was 0.26. This shows that there was an average distance between the four clusters of 82.81 and the distance within the cluster itself was 42.99. The fourth k-mean ran is k=5, and the average between was 81.21, the average within was 40.20, the within cluster sum of squares by cluster was 68% and the average silhouette width was 0.25. This shows that there was an average distance between the four clusters of 81.21 and the distance within the cluster itself was 40.20. After running these k-values, I computed the SSE plot with the elbow method as seen in Figure 2, to find the best k-value to use for cluster. The elbow point of the plot is about 7. Thus, k=7 was ran. The average between was 79.54, the average within was 35.9, the within cluster sum of squares by cluster was 76.3% and the average silhouette width was 0.27. This shows that there was an average distance between the seven clusters of 79.54, and the distance within the cluster itself was 35.9.

The SSE plot showed that the elbow could be around 7, thus k=6 and k=8 were ran. For k-means, k=6, and the average between was 79.44, the average within was 38.50, the within cluster sum of squares by cluster was 70.5% and the average silhouette width was 0.23. This shows that there was an average distance between the six clusters of 79.44, and the distance within the cluster itself was 38.50. Lastly k=8 was run, and the average between was 78.94, the average within was 33.84, the within cluster sum of squares by cluster was 78.8% and the average silhouette width was 0.29. This shows that there was an average distance between the eight clusters of 78.94, and the distance within the cluster itself was 33.84.

Figure 3, helps understand the relationship between the variables with the clustering of k=7 through a scatterplot. The scatterplot for the clusters is more spread out for age and the thalach variable. The age and the maximum heart rate have a correlation, as the older the person gets, there is a slight decrease in the maximum heart rate. The relationship between other variables is more cluttered, making it hard to see a pattern or relationship between the variables. For example, the cholesterol levels and the resting blood pressure is cluttered, and the data points in the cluster vary blood pressure ranging from 100 to 180.

Figure 4 through 7 show the distribution of the variables among each cluster. The median for the age among the clusters were the following: cluster 1 - 58 years old, cluster 2- 54 years old, cluster 3 - 52 years old, cluster 4- 59 years old, cluster 5 - 63 years old, cluster 6- 56 years old, and cluster 7- 59 years old. The range of the box whiskers plot shows that the age frame for the clusters were around the same, consisted mostly of patients in their late 50's. The median for the resting blood pressure among the clusters were the following: cluster 1 - 135, cluster 2- 129, cluster 3 - 122, cluster 4- 125, cluster 5 - 138, cluster 6- 124, and cluster 7- 150. The range of the box whiskers plot shows that the clusters included the blood pressure ranges, which for the most part were 120-150, and just a few outliers. The median for the cholesterol levels among the clusters were the following: cluster 1 - 302, cluster 2- 252, cluster 3 - 202, cluster 4- 230, cluster 5 - 403, cluster 6- 190, and cluster 7- 240. The box whisker plot for all the clusters were in the range of 200-300 cholesterol levels except for cluster 5. Cluster 5 consisted of cholesterol levels from 400-425 which is way above the rest of the cholesterol levels for the other clusters. The median for the maximum heart rate among the clusters were the following: cluster 1 - 153, cluster 2- 156, cluster 3 - 160, cluster 4- 111, cluster 5 - 152, cluster 6- 132, and cluster 7- 154. The maximum heart rate seems to be in the same range among the clusters except for cluster 4, whose max heart rate is below the other clusters.

Table 1: Summary of the Variables

	Age	Sex	Resting BP	Cholestrol	Fasting Blood Sugar	Max Heart Rate
Min	29	0	94	126	0	71
1st Quartile	48	0	120	211	0	132
Median	56	1	130	240	0	152
Mean	54.43	0.6956	131.6	246	0.1493	149.1
3rd Quartile	61	1	140	275	0	166
Max	77	1	200	564	1	202

Figure 1: Number of Patients per Cholestrol Level

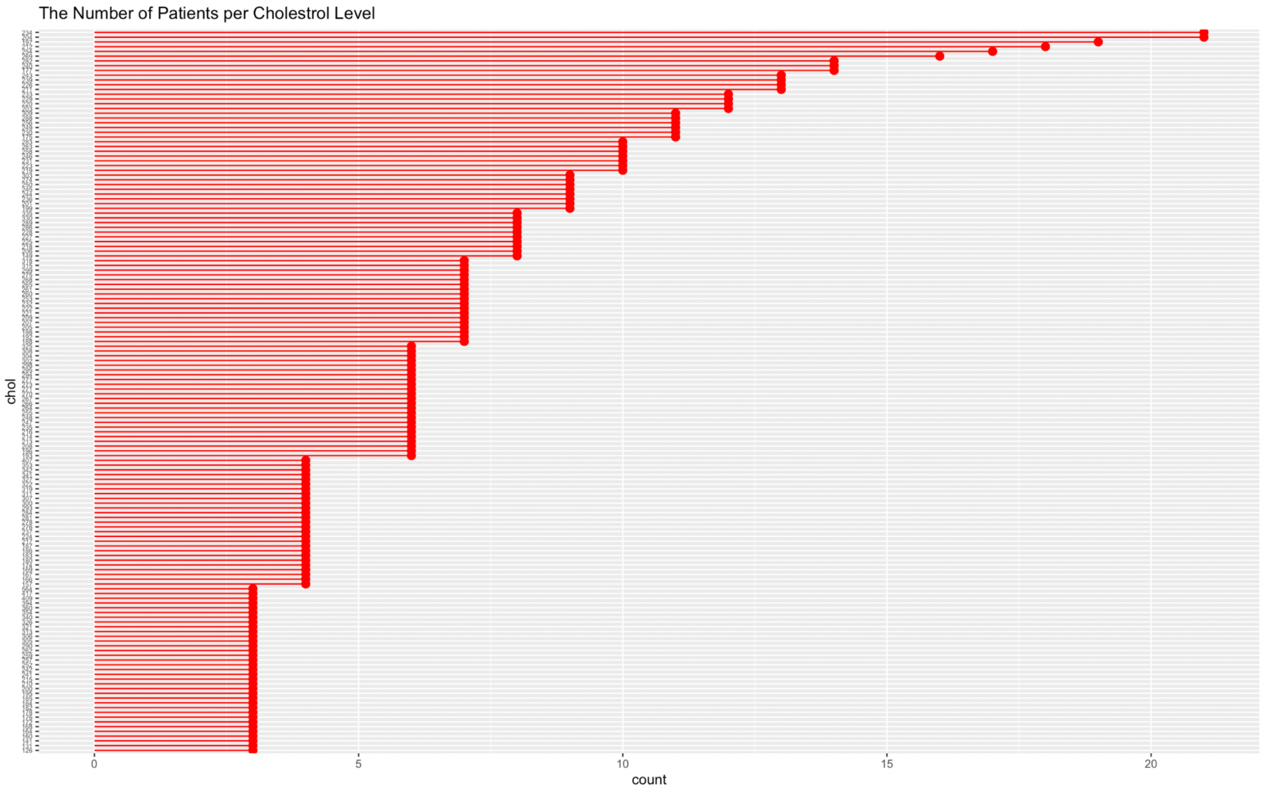
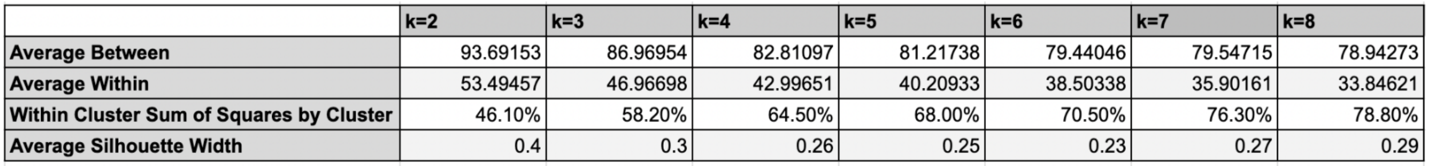


Table 2: Cluster Statistics



	k=2	k=3	k=4	k=5	k=6	k=7	k=8
Average Between	93.69153	86.96954	82.81097	81.21738	79.44046	79.54715	78.94273
Average Within	53.49457	46.96698	42.99651	40.20933	38.50338	35.90161	33.84621
Within Cluster Sum of Squares by Cluster	46.10%	58.20%	64.50%	68.00%	70.50%	76.30%	78.80%
Average Silhouette Width	0.4	0.3	0.26	0.25	0.23	0.27	0.29

Figure 2: SSE plot with Elbow Method

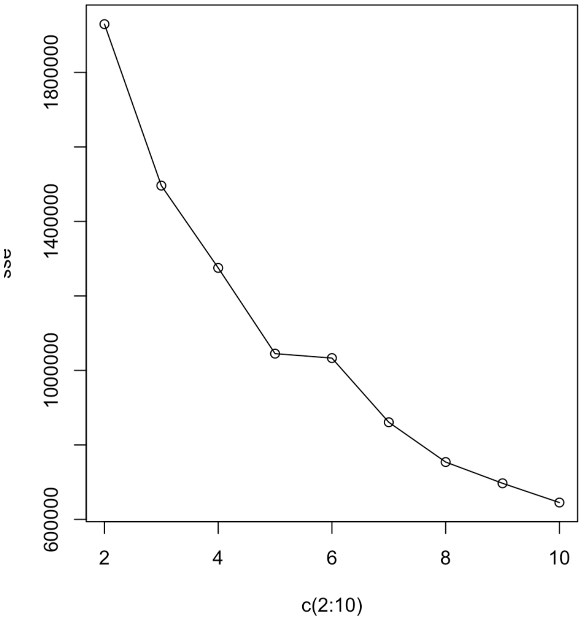


Figure 3: Scatterplot when k=7, using Pairs Function

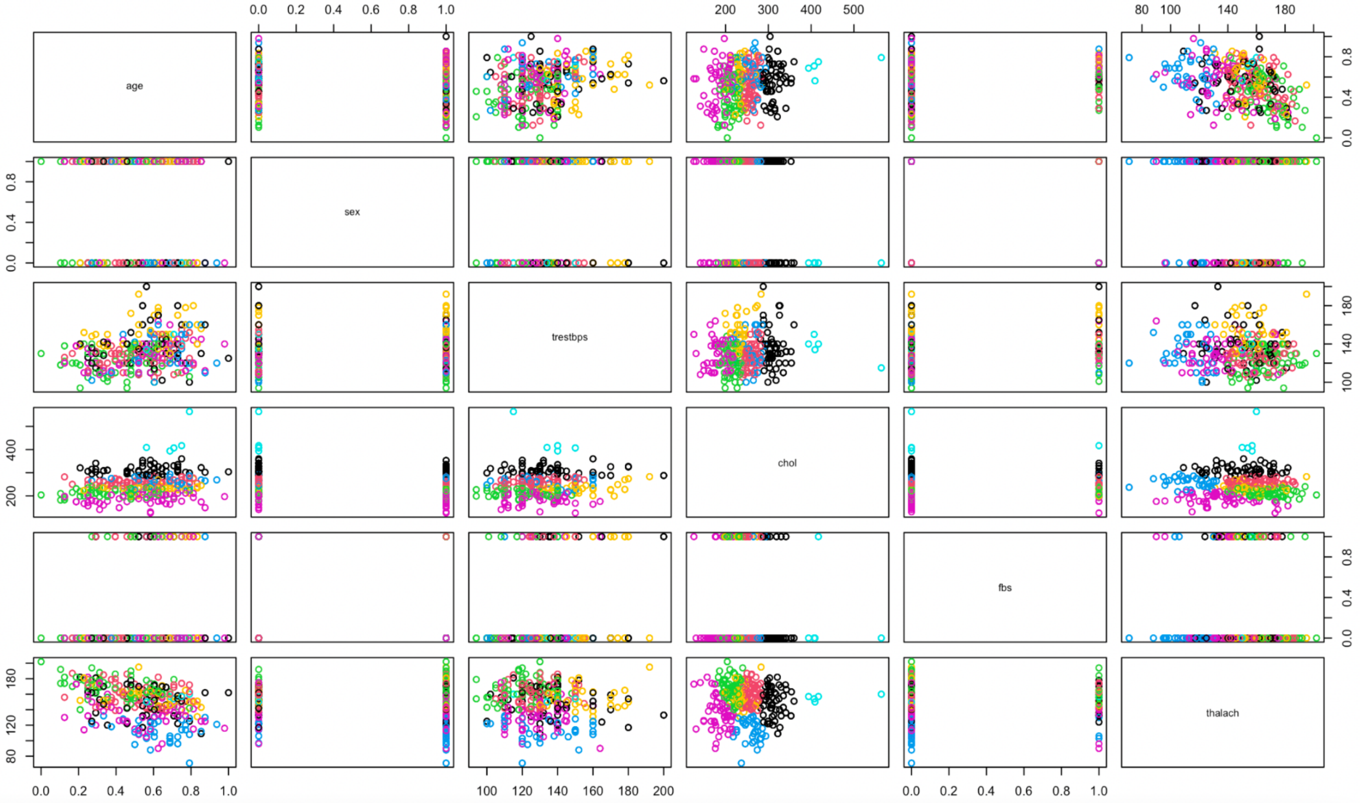


Figure 4: Box Plot: Distribution of Age in Each Cluster

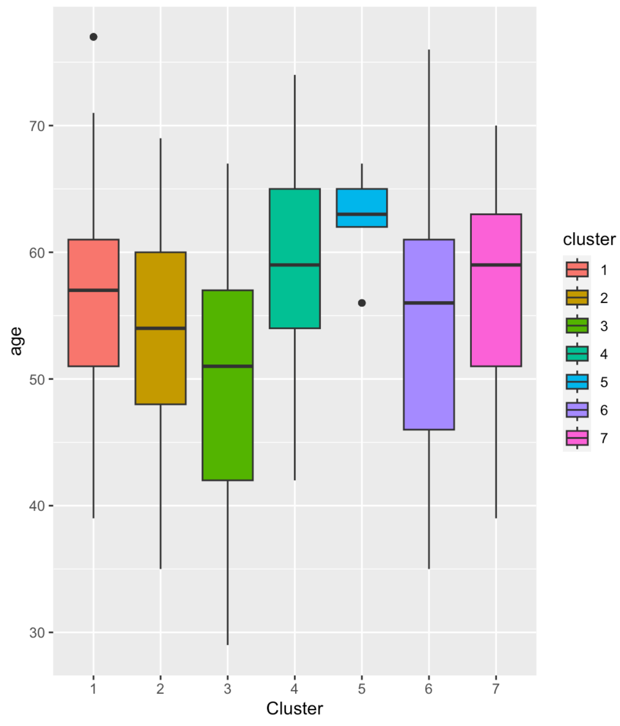


Figure 5: Box Plot: Distribution of Resting Blood Pressure in Each Cluster

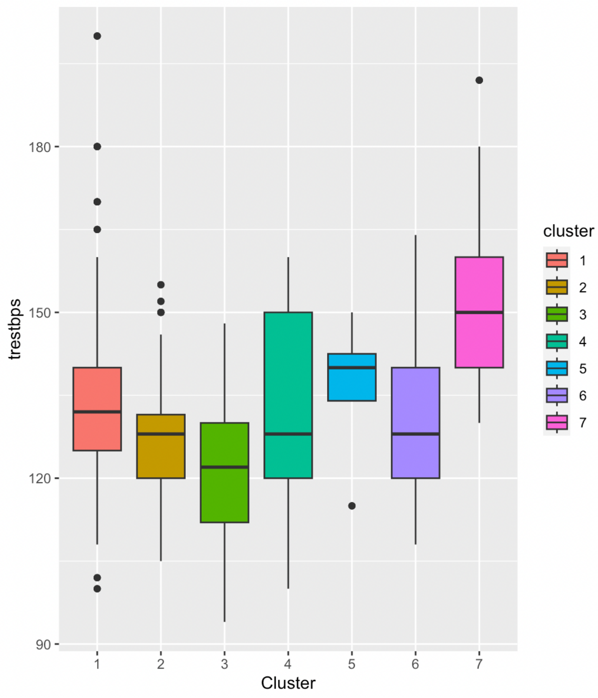


Figure 6: Box Plot: Distribution of Cholesterol in Each Cluster

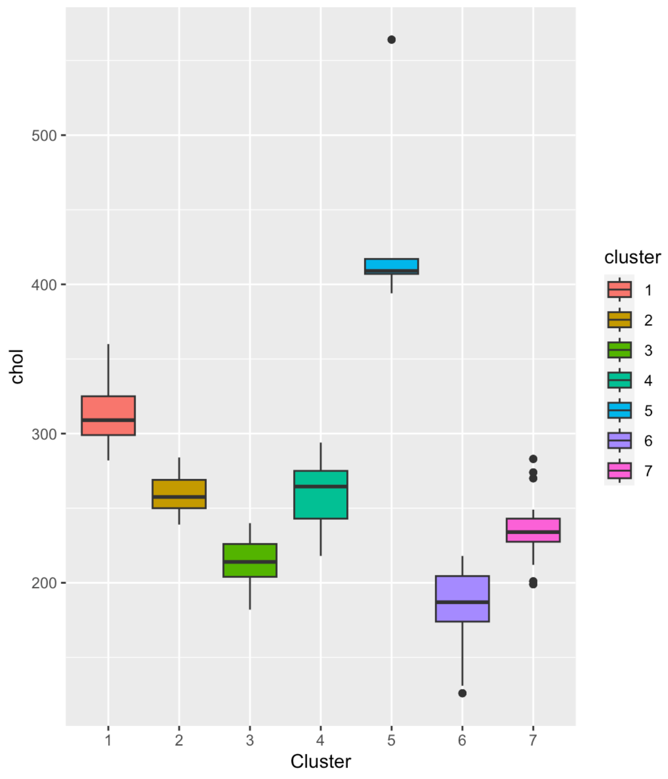
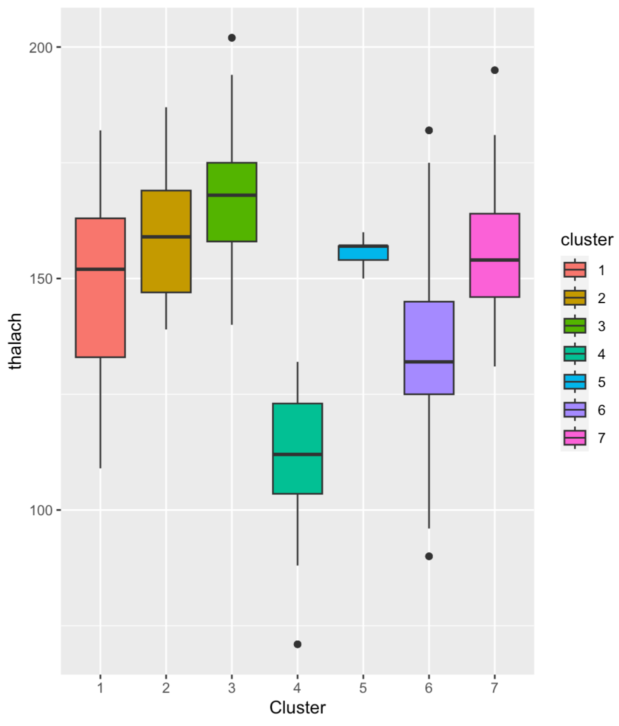


Figure 7: Box Plot: Distribution of Maximum Heart Rate in Each Cluster



Conclusion and Discussion:

After running different k values, I decided to use k=7 based on the betweeness, withinness, percentage of within cluster sum of squares by cluster, the silhouette width and the SSE plot with the elbow plot. The betweeness should be a high number as the more separate the clusters are the better clustering, and also the withiness of the clusters should be as minimal as possible because that means the more compact the cluster is. The silhouette width is better when closer to 1 and the higher the within cluster sum of squares percentage, the better. A good balance of of the above variables was cluster 7. Some clusters such as cluster 2 had a very nice betweenness of 93.69 but the withness was 53.49 which is not good because that means the cluster among itself was not compact. As for the comparison of the clusters, cluster 5 had a higher age median than the rest of the clusters, meaning there were more elderly people in that cluster. Also, the cholesterol level for this cluster was the worst as the range was from 400-425. This is double the healthy cholestrol level. The rest of the clusters were slightly better as the range of the cholesterol was 200-300. Cluster 6 had more “healthy patients" compared to the rest as the resting blood presure was 124 which is four more than the healthy blood pressure level of 120. The cholesterol level was also 190 which is healhty as it is less than 200.

One of the limitations of this analysis is that this data was donated in 1988, which is a while ago compared to today, 2023. Especially since the lifestyle of the people have changed which is one of the reason for the increase in the number of heart disease patients. Another limitation to this is that it does not have ethnicity, and does not include different regions. The patients were mostly from Cleveland which is not a sample of the worldwide patients and their characterstics. This being said, for further analysis on characteristics of heart disease patient, different regions can be looked at. In conclusion, the number of heart disease patients are growing. People becoming aware of the characteristics and what they indicate could encourage them to change their lifestyle and seek the doctor.

References:

American Heart Association. (2023, April 28). *Understanding blood pressure readings*. Heart Attack and Stroke Symptoms. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Centers for Disease Control (CDC). (2020). Heart disease facts| cdc. Gov.

Chernegs (2022). *Heart Disease Celeveland UCI.* Kaggle. https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci

Mayo Foundation for Medical Education and Research. (2021, November 17). *Thalassemia*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/thalassemia/symptoms-causes/syc-20354995

Zach. (2020, August 11). *How to create and interpret pairs plots in R*. Statistics. Simplified. Statology. https://www.statology.org/pairs-plots-r/