# Heart Disease

Ricardo Aguila
Casey Hernandez
Isaiah Stokes
Joshua White

December 8, 2022

## Executive Summary

In this project, we utilized an n=303 data set containing 11 variables about people with known heart issues and grouped them in a way that will be useful for healthcare professionals to determine ideal treatment procedures. By utilizing clustering algorithms, we found that these patients were best separated into six distinct clusters. After an analysis of these clusters, we decided to label them as such: Healthy with Idiopathic Chest Pain, Low Fitness, High Blood Pressure and Cholesterol, Older with Low Fitness and Hypertension, Exercise Heart Pain, and Young with Unhealthy Lifestyle.

# Introduction

The purpose of this project was to look at a data set of patients with known heart issues and see if they could be grouped based on having similar qualities. The data set included 11 variables which were a mix of categorical and numerical variables, which were: gender of the patient (sex, categorical), age of the patient, ranging from 29 to 77 (age, numerical), type of chest pain (cp, categorical), resting blood pressure (trestbps, numerical), amount of total cholesterol in blood of patient (chol, numerical), fasting blood pressure (fbs, categorical), resting electrocardiographic rate (restecg, categorical), maximum heart rate achieved (thlach, numerical), exercise-induced angina (exang, categorical), ST Peak (numerical), and slope (categorical). The first step when analyzing the data was to convert the categorical variables, **sex, cp, fbs, restecg, exang,** and **slope** into dummy variables. Since the data was not scaled, we scaled the numerical variables, **age, trestbps, chol, thalach,** and **oldpeak.**

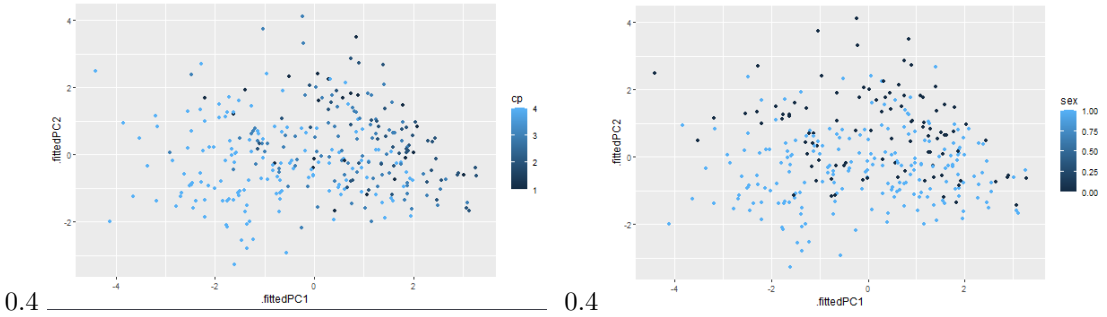
# Clustering Method

One of the first ideas that we implemented was trying to cluster the data with principal component analysis (PCA). To start this process we will be removing the ID variable because it is likely to just create noise in the data. Then we looked at the correlation between all the variables.

```
             [,1]        [,2]        [,3]        [,4]         [,5]         [,6]        [,7]         [,8]        [,9]        [,10]        [,11]
 [1,]  1.00000000 -0.09754228  0.10413895  0.28494592  0.208950270  0.118530242  0.14886759 -0.393805806  0.09166077  0.203805481  0.161769559
 [2,] -0.09754228  1.00000000  0.01008389 -0.06445590 -0.199914683  0.047862121  0.02164735 -0.048663296  0.14620149  0.102172644  0.037532882
 [3,]  0.10413895  0.01008389  1.00000000 -0.03607725  0.072318884 -0.039974990  0.06750523 -0.334421706  0.38405953  0.202276541  0.152050417
 [4,]  0.28494592 -0.06445590 -0.03607725  1.00000000  0.130120108  0.175340227  0.14656039 -0.045350879  0.06476246  0.189170971  0.117381584
 [5,]  0.20895027 -0.19991468  0.07231888  0.13012011  1.000000000  0.009841023  0.17104253 -0.003431832  0.06131038  0.046563989 -0.004061834
 [6,]  0.11853024  0.04786212 -0.03997499  0.17534023  0.009841023  1.000000000  0.06956450 -0.007854147  0.02566515  0.005747223  0.059894178
 [7,]  0.14886759  0.02164735  0.06750523  0.14656039  0.171042532  0.069564497  1.00000000 -0.083389433  0.08486695  0.114132813  0.133945690
 [8,] -0.39380581 -0.04866330 -0.33442171 -0.04535088 -0.003431832 -0.007854147 -0.08338943  1.000000000 -0.37810342 -0.343085392 -0.385601162
 [9,]  0.09166077  0.14620149  0.38405953  0.06476246  0.061310377  0.025665147  0.08486695 -0.378103424  1.00000000  0.288222808  0.257748369
[10,]  0.20380548  0.10217264  0.20227654  0.18917097  0.046563989  0.005747223  0.11413281 -0.343085392  0.28822281  1.000000000  0.577536817
[11,]  0.16176956  0.03753288  0.15205042  0.11738158 -0.004061834  0.059894178  0.13394569 -0.385601162  0.25774837  0.577536817  1.000000000
```

Figure 1: The correlation between all variables in the data set.

From the figure above it should be noted that the correlation between all the variables is relatively low with the highest magnitude only being .57. With just this information it seems that PCA will not be a good fit for our data set because of the lack of correlation and the number of variables. However, we will continue the process to see if we can extract any interesting information from the principal components. A summary of the principal components shows that the first component only explains 22% of the variance and the second component explains 13%. Out of the 11 components at least 9 of them would be needed to get above 90% cumulative proportion of variance explained. Even with these numbers, we decided to create some plots with the first two components.




0.4 0.4

The graphs do not give the clearest division between the types of chest pain, but we can see that chest pain four is more common on the left side of the graph while chest pain one is more common on the right side of the graph. For the other graph of Pc1 vs Pc2 now the colored variable represents male or female and

this graph more accurately represents why PCA is not a great choice for this data set because there is no clear trend present in the graph.

Next, a Scree Plot (shown below) was constructed, and the elbow method was utilized to determine the appropriate number of clusters.

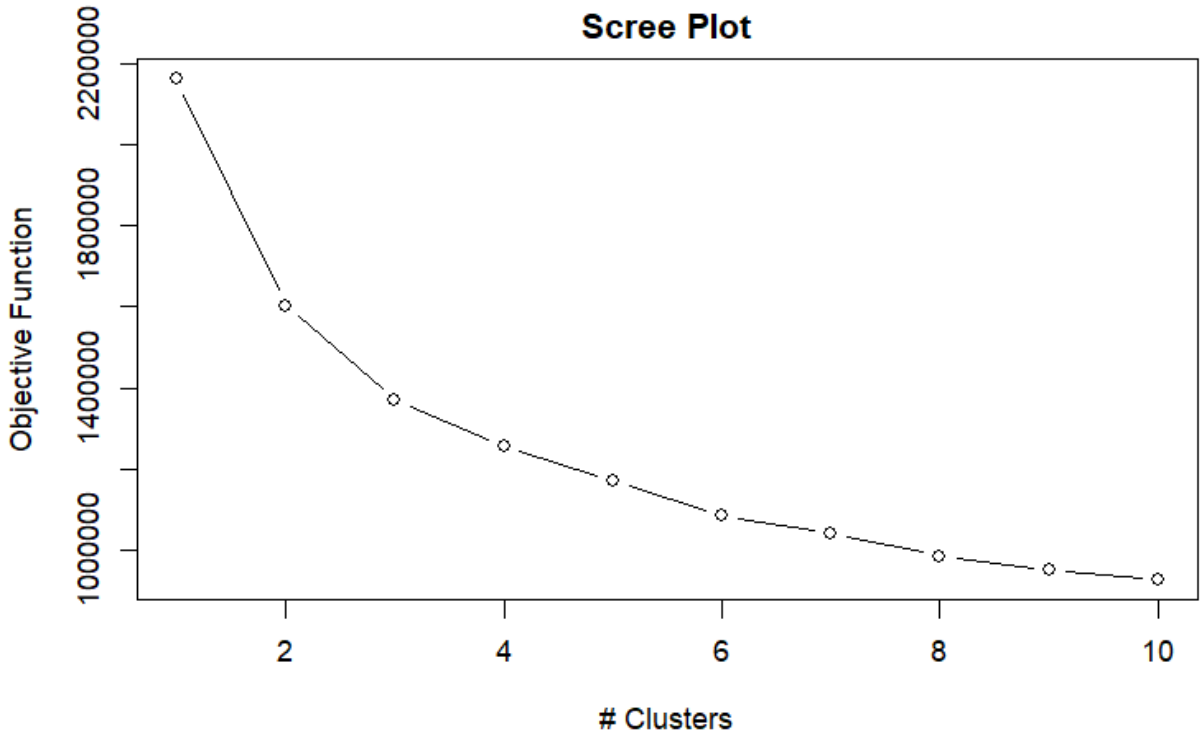



Figure 2: A Scree plot showing the objective function of each k # of clusters.

Observing the Scree plot, we determined that the notable number of clusters were three and six, so from here we began examining both possibilities.

## K = 3

| Cluster | Frequency | Age | Sex | Chest Pain Type | Resting HR | Cholesterol | Fasting Blood Sugar | Rest ECG | Max HR | Exercise Pain | ST Peak | ST Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 81 | 0.4035416 | 1 | 3 | 0.3908086 | 0.45753297 | 0 | 2 | 0.4505494 | 0 | -0.3445925 | 1 |
| 2 | 110 | 0.4492983 | 1 | 4 | 0.2314731 | 0.06509787 | 0 | 2 | -0.8913257 | 1 | 0.7527786 | 2 |
| 3 | 112 | -0.7331222 | 1 | 4 | -0.5099779 | -0.39482979 | 0 | 2 | 0.5495654 | 0 | -0.4901219 | 1 |

Table 1: The table above illustrates the clusters generated when the algorithm is set to create k=3 distinct groups.

Immediately, we found k=3 to be an inadequate form of clustering considering the poor range of total results. There isn't a single group with majority female sex or a group that doesn't have an abnormal resting ECG. Without exploring much further, we decided k=6 would be more appropriate.

## K = 6

| Cluster | Frequency | Age | Sex | Chest Pain Type | Resting HR | Cholesterol | Fasting Blood Sugar | Rest ECG | Max HR | Exercise Pain | ST Peak | ST Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 56 | -0.5385216 | 1 | 2 | -0.58201632 | -0.424489997 | 0 | 0 | 0.6323133 | 0 | -0.74004418 | 1 |
| 2 | 43 | -0.4190643 | 1 | 3 | -0.22814818 | -0.568090974 | 0 | 0 | 0.2418480 | 0 | 0.09808532 | 2 |
| 3 | 44 | 0.6554833 | 0 | 3 | 0.25910775 | 1.041840590 | 0 | 2 | 0.2039548 | 0 | -0.48823433 | 1 |
| 4 | 46 | 0.8629802 | 1 | 4 | 1.01813543 | 0.089900230 | 0 | 2 | -0.3924295 | 0 | 1.06307436 | 2 |
| 5 | 68 | 0.2931066 | 1 | 4 | -0.22218257 | -0.032415283 | 0 | 0 | -1.1155857 | 1 | 0.56877982 | 2 |
| 6 | 46 | -0.8759266 | 1 | 4 | -0.01572327 | 0.009286836 | 0 | 2 | 0.8506209 | 0 | -0.62763765 | 1 |

Table 2: The table above illustrates the clusters generated when the algorithm is set to create k=6 distinct groups.

Our decision to move onto k = 6 clusters proved much more promising and aligns better with the stated goal of this project. Each cluster has now represented something that meaningfully sets it apart from the others. These differences are explored in detail in the next section. It is of note that no group has a maximum heart rate that could be considered too high for their age group, so for the analysis below, high maximum heart rate is considered a good thing. The same is true for resting blood pressure, in that no group has a blood pressure that is too low to be considered healthy, so low blood pressure is a good thing.

Now revisiting the PCA using these six cluster to see if there are any interesting trends that can be seen.

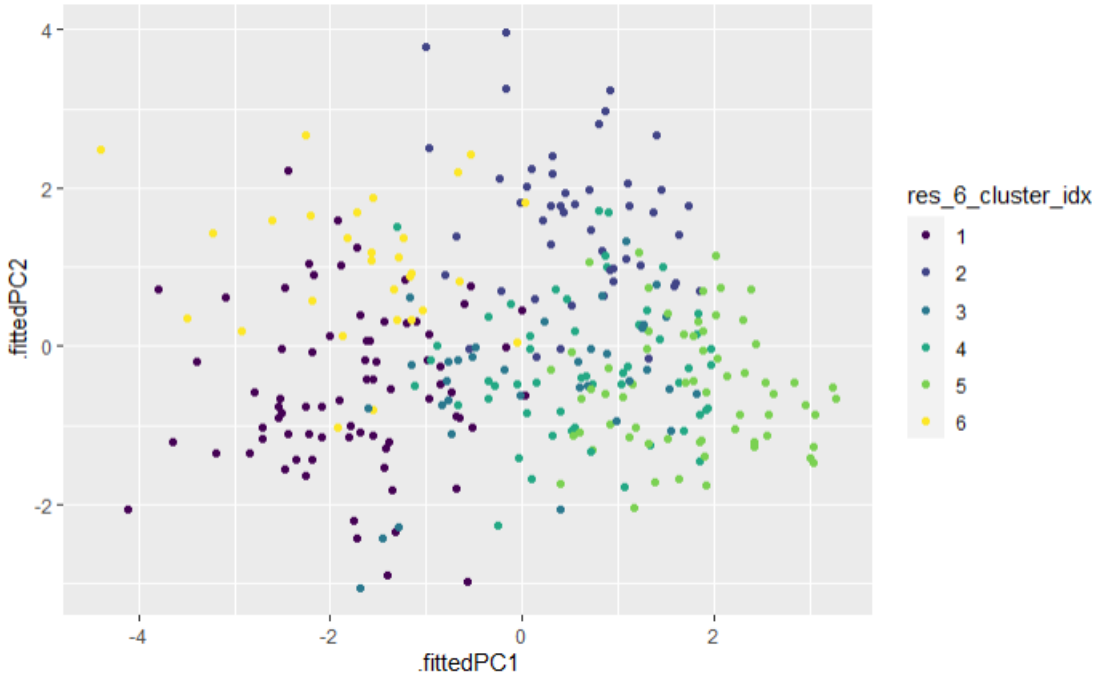



Figure 3: PC1 vs PC2 colored by cluster.

With these six cluster now there is a much more clear divide between all the groups. Where cluster 1 tends more towards the bottom left, cluster two is more centered, cluster three tends towards the middle right, cluster five is far right bottom of the graph, while cluster 6 tends towards the upper left side of the graph.

## Detailed Comparison of Groups

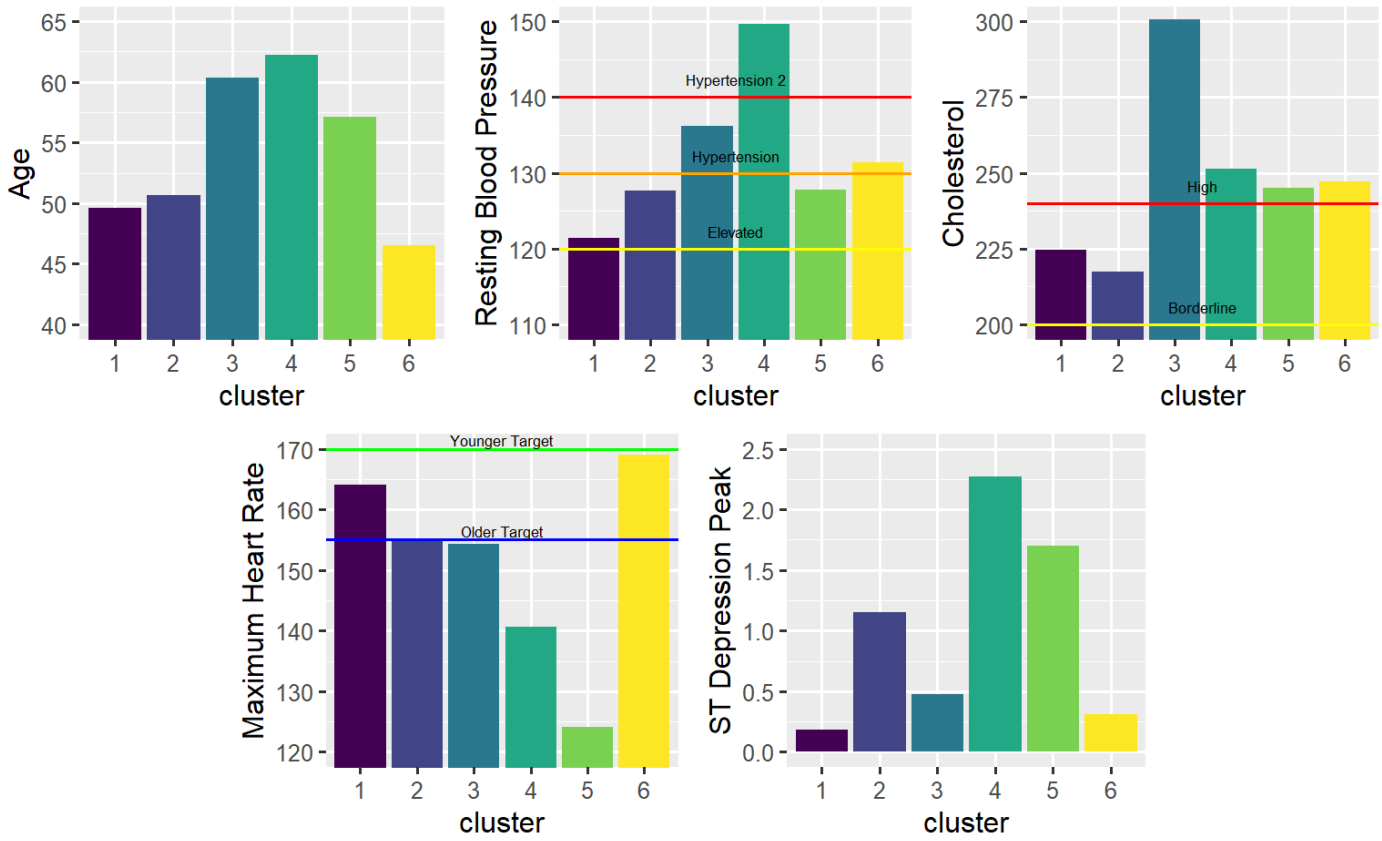



Figure 4: Bar plots showing each the mean of each cluster for all numeric variables with horizontal lines indicating target or threat ranges.

1. Younger, majority male, Chest Pain Type 2. This group has a healthy resting blood pressure and cholesterol, normal ECG, and minimal ST depression. This group can be called the "Healthy with Idiopathic Chest Pain Group". This means that the chest pain is present but cannot be explained by diagnostics. These findings may be useful since this group with chest pain type 2 is considered healthy by all diagnostics.

2. Younger, majority male, Chest Pain Type 3. This group has moderately elevated resting blood pressure and low cholesterol with normal resting ECG. The maximum heart rate is lower for this group considering their age, and they also exhibit moderate ST depression. Considering the normal resting ECG, lower maximum heart rate, and moderate ST depression, this group may be called the "Low Fitness Group". This information suggests that the heart pain may be caused by, or may be causing, low fitness. However, considering the heart pain is not caused by exercise, the former may be more probable.

3. Moderately older, majority female, Chest Pain Type 3. The resting blood pressure of this group is consistent with hypertension and the cholesterol is extremely high. Also, they exhibit abnormal resting ECG while having a normal maximum heart rate for their age and normal exercise diagnostics. By all metrics, this group can be called the "High Blood Pressure and Cholesterol Group".

4. Oldest group, majority male, Chest Pain Type 4. This group has resting blood pressure consistent with hypertenstion type 2, fairly high cholesterol, abnormal rest ECG, lower maximum heart rate, and

extremely high ST depression from exercise. This group can be called the "Older with Low Fitness and Hypertension".

5. Moderately older, majority male, Chest Pain Type 4. This group has elevated resting blood pressure, fairly high cholesterol, normal rest ECG, extremely limited maximum heart rate, and exhibiting elevated ST depression. This is the only group with exercise chest pain. This group can be called the "Exercise Heart Pain" group since their diagnostics are fairly normal when they are not exercising.

6. Youngest group, majority male, Chest Pain Type 4. This group has resting blood pressure consistent with hypertension, fairly high cholesterol, abnormal rest ECG, and a very high maximum heart rate with minimal ST Depression. This group is similar to group 1, but they have a different type of chest pain and an abnormal rest ECG. This group is the "Young with Unhealthy Lifestyle" group.

# Conclusion

Toward completing the goal of this project, we have separated the patients into six groups and labeled them as follows: Healthy with Idiopathic Chest Pain, Low Fitness, High Blood Pressure and Cholesterol, Older with Low Fitness and Hypertension, Exercise Heart Pain, and Young with Heart Arrhythmia. We believe that these groups can be helpful as an early diagnostic tool to assist healthcare professionals in quickly finding the right treatment option for their patients.

There are a few limiting factors for this assignment relating to the data set. Like many social surveys, this one suffers from a small percentage of female patients, comprising only a third of the data set. In its current state, sex is not a meaningful factor, so this clustering may be improved by removing sex or by performing a separate clustering for both male and female. This data comes from the VA (Veteran's Affairs), which receives more male patients than female, which is one possible explanation. This clustering also fails to make use of the fasting blood sugar variable, comprising about 15% of patients. It is either the case that the profile for a patient with high fasting blood sugar (indicating diabetes) is not consistent with any other variables, or there simply aren't enough patients for this to be a significant factor in clustering. If results are needed that include diabetes as a factor, then either a separate clustering or a higher number of clusters is needed to form a meaningful observation.