**STAT 456 Final Report**

Emily Zhang

**Introduction: Background and Problem Statement**

As a major cause of death worldwide, heart disease can be influenced by several risk factors, such as behavioral, genetical and environmental variables. However, a lot of people still seem to be unaware of the significant impact of heart disease. Thus, I found a heart disease data on Kaggle in order to understand and predict the association between heart disease and some potential factors, which includes age, sex, cp (chest pain type), trestbps (resting blood pressure), chol (cholesterol levels), fbs (fasting blood sugar), restecg (resting electrocardiographic results), thalach (maximum heart rate achieved), exang (exercise-induced angina), oldpeak (ST depression) and slope (slope of the peak exercise ST segment). Two of the main research questions I raised in the preliminary report are:

1. Which factors contribute significantly to heart disease?
2. How can participants be categorized using health indicators, and what characteristics distinguish these clusters?

My goal in this project is to illustrate the relationship between heart disease and individual’s physical condition using two multivariate methods: Principal Components Analysis and K-Mean Clustering.

**Body-Method choice**

For this project, I decided to use Principal Component Analysis (PCA) and K-means clustering as the methods since they can handle the multivariate data efficiently. For my first research question (Which factors are most predictive of heart disease), Principal Component Analysis is suitable since it will help me to simplify the complexity in the data by reducing the number of dimensions. Thus, the large dataset will be transformed into a smaller one so I can focus on the principal component that explains most of the variance in the dataset, and then find the most significant factors that contribute to heart disease. On the other hand, K-means Clustering method is appropriate for my second research question since it is able to categorize participants into number of clusters based on the similarity in their physical condition. Thus, we can understand different risk profile among patients.

**Body-Data analysis and preliminary interpretation**

In order to prepare for the Principal Component Analysis, I preprocessed the data first by standardizing numerical values and encoding categorical variables, which is essential for correctly analyzing our data. Then, we can start the Principal Component Analysis.

**A close-up of numbers

Description automatically generated**

**A graph of a number of gray rectangular objects

Description automatically generated**

These pictures show the PCA results, from which we can see the value of standard deviations, proportion of variance, cumulative proportion, and a bar plot of variances. It is clear that the first component captures about 35.97% of the variance, which is significant amount of information. And the first two components together explain almost half the variance, which makes it a good summary. Moreover, the PCA suggests that a reduced number of dimensions could effectively summarize the original. This reduction can simplify my further analysis of by focusing on these principal components.

A diagram of numbers and points

Description automatically generated

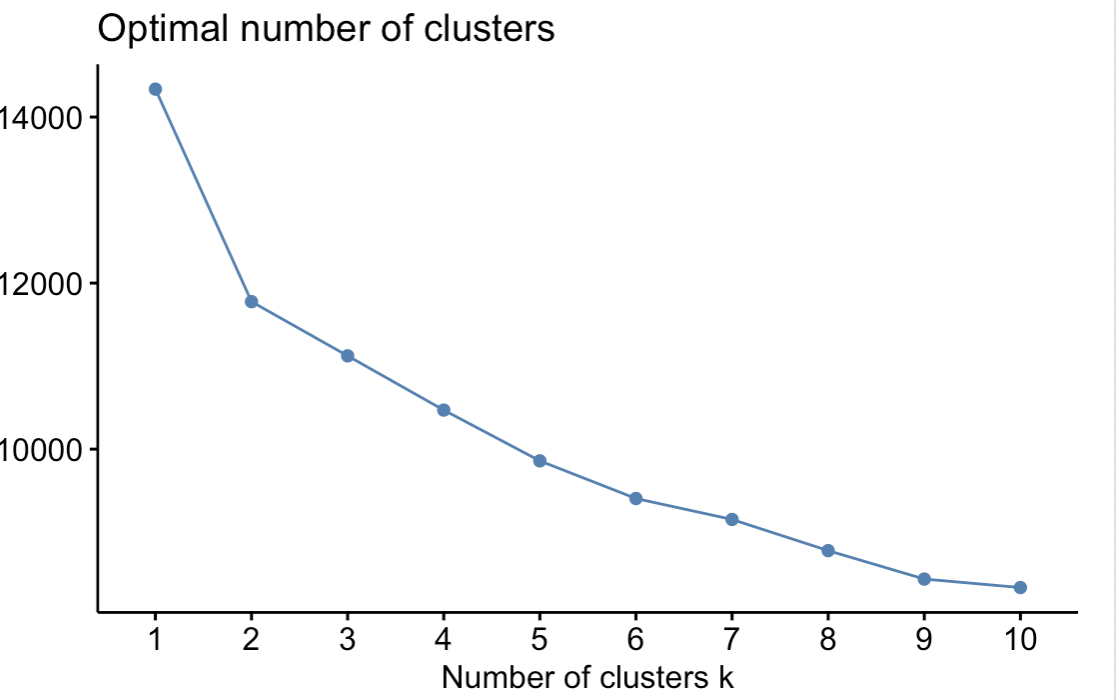
**A diagram of numbers and points

Description automatically generated with medium confidence**

Next, I generated a biplot from the PCA of the heart dataset and each point in the plot shows an observation from the dataset. The X axis that also imply the first component captures 36.1% of the variance in the dataset and the Y axis that also imply the second component captures 21.6% of the variance. The arrow in the plot shows the correlation of the variables and principal components. Then, I used k=2 and 4 to generate the K-means clustering plot with X-axis corresponds to the first principal component, and the y axis corresponds to the second principal component. Different colors represent different clusters based on their similarities. A graph of a cluster of colored triangles

Description automatically generated with medium confidenceA graph showing a cluster of numbers

Description automatically generated with medium confidence

****

**Body-Interpretation**

In the data analysis part, I have provided principal component table, bar plot, biplot, and k-clustering plot. In the principal component table, standard deviations, proportion of variance, and cumulative proportion are included, from which I realize that the first principal component and the second principal component explain for most of the total variance in the dataset. This result is also clear in the bar plot. The PCA biplot is a plot which represent both the observation and variable pf a matrix of multivariate data on the same plot. Arrows in the biplot shows different physical condition indicators, and their direction and length can indicate how did variables affect the principal components. The arrow for ‘thalach’ points downward to the left, which represents a significant negative relationship with the second principal component. Thus, we can conclude that heart disease is associated with lower level of ‘thalach’. The arrow for ‘age’, ‘trestbps’ and ‘chol’ points downward to the right, which suggests that higher value in these variables is correlate with higher risk of heart disease. As the age, blood pressures, and cholesterol level increase, individuals may have higher risk of heart disease. In the following K-means clustering plot, I believe choosing cluster for two is more direct and clearer. The red cluster includes characteristics as older age, higher blood pressure, higher cholesterol levels and serious ST depression since the points are higher scores on the first principal component and lower scores on the second principal component. Meanwhile, the green cluster includes lower scores on the first principal component and higher on the second, suggesting younger age, lower blood pressure, lower cholesterol levels and less ST depression.

**Conclusion**

The data analysis and the interpretation above should have answered my research questions. Let me restate my research questions:

The first one is which factors contribute significantly to heart disease and the second one is how can participants be categorized using health indicators and what characteristics distinguish these clusters? For the first question, we can get the factors that contribute significantly to heart disease in the PCA biplot. The most outstanding variable I choose are age, resting blood pressure (trestbps), maximum heart rate (thalach) and ST depression (oldpeak). Age shows a positive association with principal component one, suggesting that elder people have a higher rate of getting heart disease. Resting blood pressure shows a positive relationship with principal component one, suggesting that higher blood pressure is crucial in heart disease risk. The level of cholesterol is the same as the resting blood pressure. ST depression shows a positive correlation with the second principal component, suggesting that higher ST depression is more likely to cause heart disease. Maximum heart rate shows a negative correlation with the second principal component, suggesting that higher heart rates are correlate with healthier heart status. For my second research question, K-means clustering plot can help us to illustrate the answer. The red cluster includes participants who have higher scores on the first principal component and lower scores on the second principal component. These people have higher ages, higher resting blood pressure and greater ST depression, which represents a higher risk for heart disease. Meanwhile, the green cluster includes lower scores on the first principal component and higher on the second, suggesting younger age, lower blood pressure, lower cholesterol levels and less ST depression. Thus, they are less likely to suffer from heart disease.

**Summary**

The project’s main goal is to explore and understand the variables that contribute to heart disease and to categorize patients into distinct groups based on heart health indicators using Principal Component Analysis (PCA) and K-means clustering. The PCA reduces the dimensionality of the dataset, representing the most significant factors impacting heart disease, while K-means clustering utilizes these factors to categorize patients into meaningful groups. After doing data analysis, we able to conclude that people who have higher ages, higher resting blood pressure and greater ST depression are more likely to suffer from heart disease.

**Appendix**

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated