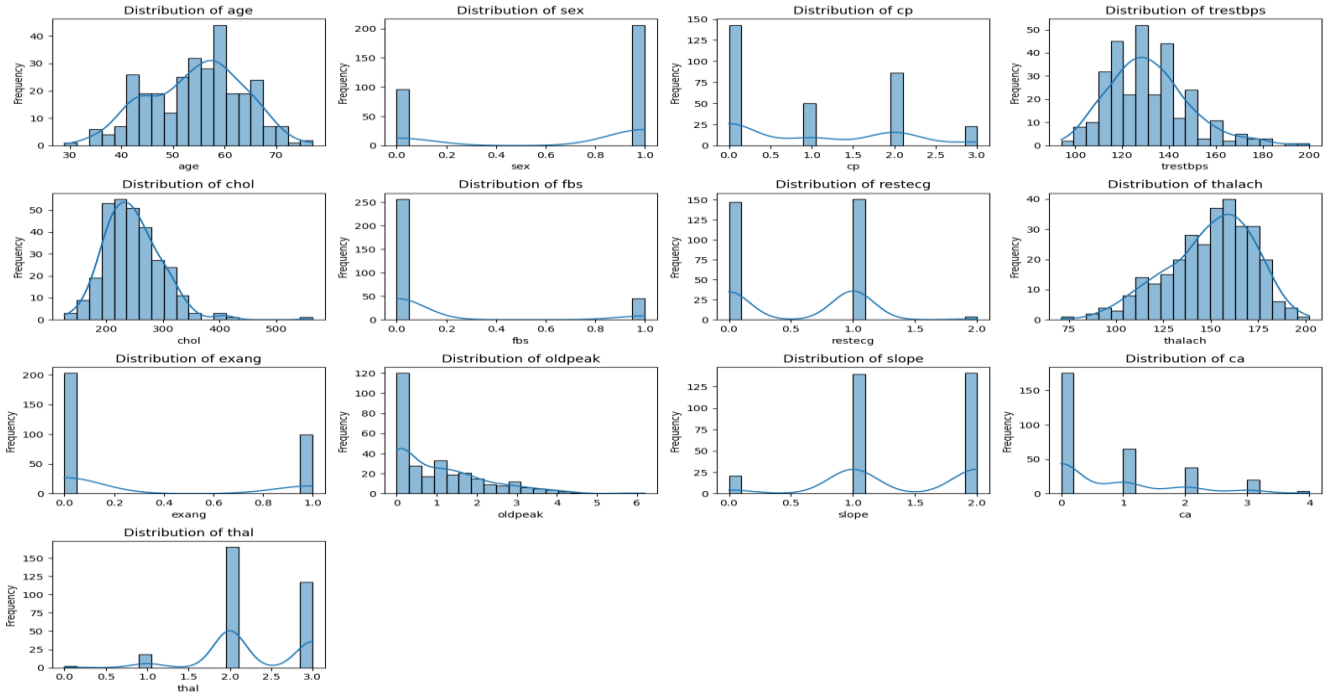
GitHub Repo Link: https://github.com/rameshpc9/clustering-and-fitting

Student ID: 23038360

Dataset Link: https://www.kaggle.com/datasets/reenapinto/heart-disease-analysis

Heart Disease Analysis

This project focuses on the use of clustering techniques and dimensionality reduction methods to analyze heart disease data sets. Using unsupervised learning methods, the aim is to reveal patterns and clusters in the data, which helps to identify potential risk factors for cardiovascular disease heart disease databases with physiological and clinical characteristics like cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal provide valuable insights into the understanding heart disease dataset. We found no null values in dataset in preprocessing but found duplicates so remove all the duplicates from the dataset.



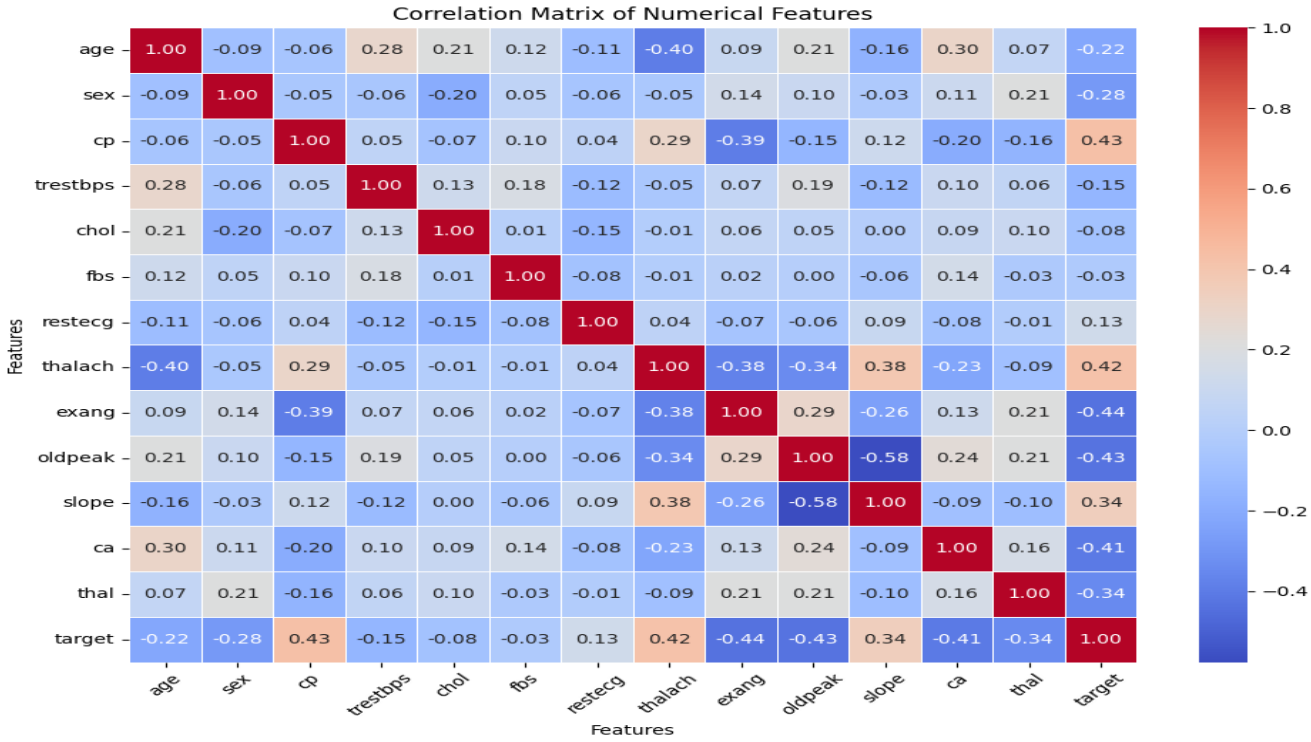
Examining the histogram of our data set, notable features emerge in several cases.

Factors such as "trestbps" (resting blood pressure), "chol" (serum cholesterol level), and "thalach" (maximum heart rate reached) refer more data to a particular value, indicating variation which can occur in measurement in individuals In contrast, factors such as "age" and "thalassaemia level) show a more balanced distribution, indicating a more homogeneous distribution in the data set.These findings suggest potential heterogeneity of health indicators in the dataset, with somefeatures exhibiting if there are significantly lower or higher values than others

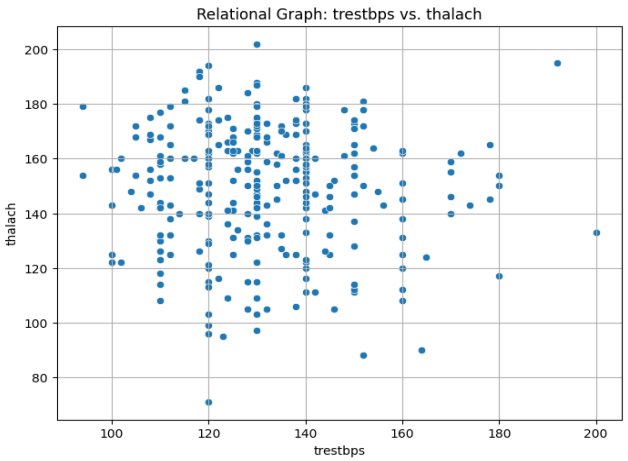
Statistical Moments

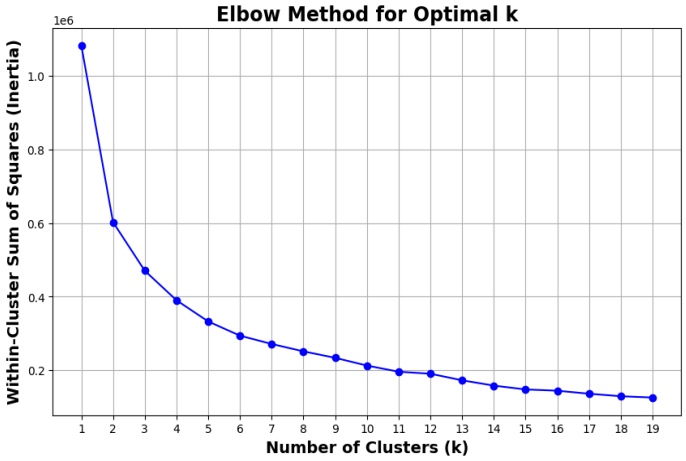
	Mean	Median	Standard Deviation	Skewness	Kurtosis
age	54.420529801324506	55.5	9.047969746247457	-0.20374327959596905	-0.5275122997069857
sex	0.6821192052980133	1.0	0.46642573806726434	-0.7861201403379438	-1.3912730609583188
cp	0.9635761589403974	1.0	1.0320436419542316	0.49302157403846975	-1.1837292153374055
trestbps	131.60264900662253	130.0	17.56339423003756	0.7165414326647316	0.9229963552001497
chol	246.5	240.5	51.75348865574056	1.147332413980798	4.542591352463679
fbs	0.1490066225165563	0.0	0.3566860293648133	1.9812008559042784	1.93794719680965 5
restecg	0.5264900662251656	1.0	0.5260271694099752	0.169466604898877	-1.3594641004045322
thalach	149.56953642384107	152.5	22.903527251969845	-0.5326712468229613	-0.062186318831145115
exang	0.32781456953642385	0.0	0.47019596400976954	0.7372812469727277	-1.4661703181043508
oldpeak	1.0430463576158941	0.8	1.1614522890634562	1.266172720910219	1.5678764941867342
slope	1.3973509933774835	1.0	0.6162739844441468	-0.5032467424064804	-0.6299346735058267
ca	0.7185430463576159	0.0	1.0067482586428396	1.2957384547781445	0.781003295709771
thal	2.314569536423841	2.0	0.6130255397814752	-0.48123240915755733	0.2958547127549438
target	0.543046357615894	1.0	0.49897035961141234	-0.17369101189529748	-1.9830082305694372

Our statistical analyzes showed that some items exhibit negative values of skewness and kurtosis, indicating any left-skewed platykurtic distribution This showed long tails on the left and flat top long tails compared to the normal distribution. For example, the attributes 'age' and 'sex' exhibit such a trend, providing insight into the shape and consistency of the dataset distribution.



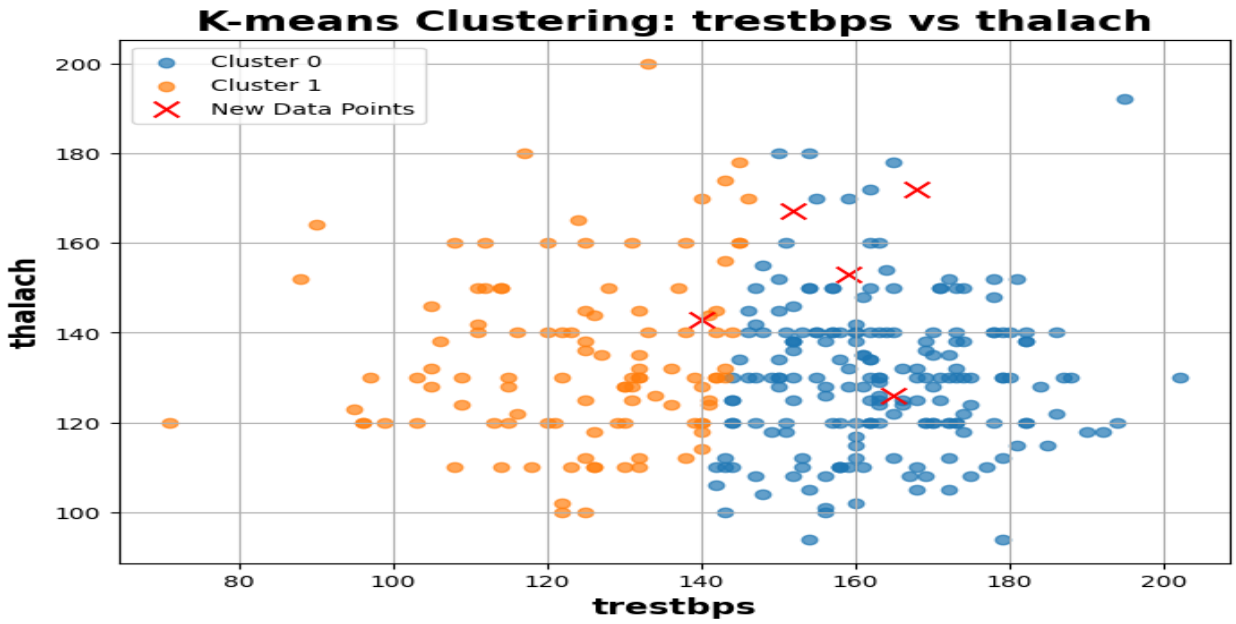
- **Thalach and Exang (Correlation: 0.42)**: Positive correlation indicates that as maximum heart rate increases, the likelihood of exercise-induced angina also increases.
- **Thal and Target (Correlation: -0.34)**: Negative correlation shows that higher scores for heart defects (Thal) are associated with a lower probability of heart disease.
- **Ca and Thal (Correlation: 0.16)**: Weak positive correlation suggests that an increase in the number of major vessels detected by fluoroscopy correlates with higher Thal scores, indicating more severe defects.
- **Slope and Thal (Correlation: -0.10)**: Weak negative correlation implies that a steeper peak exercise segment slope might be associated with fewer heart defects.





With clustering, we clustered similar individuals together based on their physiological and clinical attributes. For instance, leveraging the elbow method, we identified the optimal number of clusters for our heart disease dataset. Subsequently, algorithms like K-means facilitated the assignment of new observations to these clusters. This enabled us to predict which cluster a new individual belongs to base on their features, offering insights into their similarity to existing heart disease risk profiles.

With the variables 'treetops' (resting blood pressure) and 'thali' (maximum heart rate achieved), a K-means clustering analysis was run. K-means clustering is a method of partitioning the data into clusters to minimize within-cluster variance based on similarity. In this plot, the data is clustered into two clusters, depicted by different colors to bring out each cluster, grouped around its closest centroid, which reflects different patterns: resting blood pressure and maximum heart rate reached.



Predicted Cluster Labels for New Data:

Data_Point-1 = Cluster 0 Data_Point-2 = Cluster 1 Data_Point-3 = Cluster 0 Data_Point-4= Cluster 0
Data_Point5: Cluster 0

With fitting predictions, it will mean predicting the levels of the risk of heart disease from a set of various physiological and clinical attributes. We utilize linear regression for the predictive purpose of the risk levels of heart disease. With the model fitted to our data set, we made predictions on the risk of getting heart disease, given the attribute value of an individual. These would, therefore, provide us with quite helpful information about how the variable values determine the probability of obtaining heart disease.