The objective of the project:

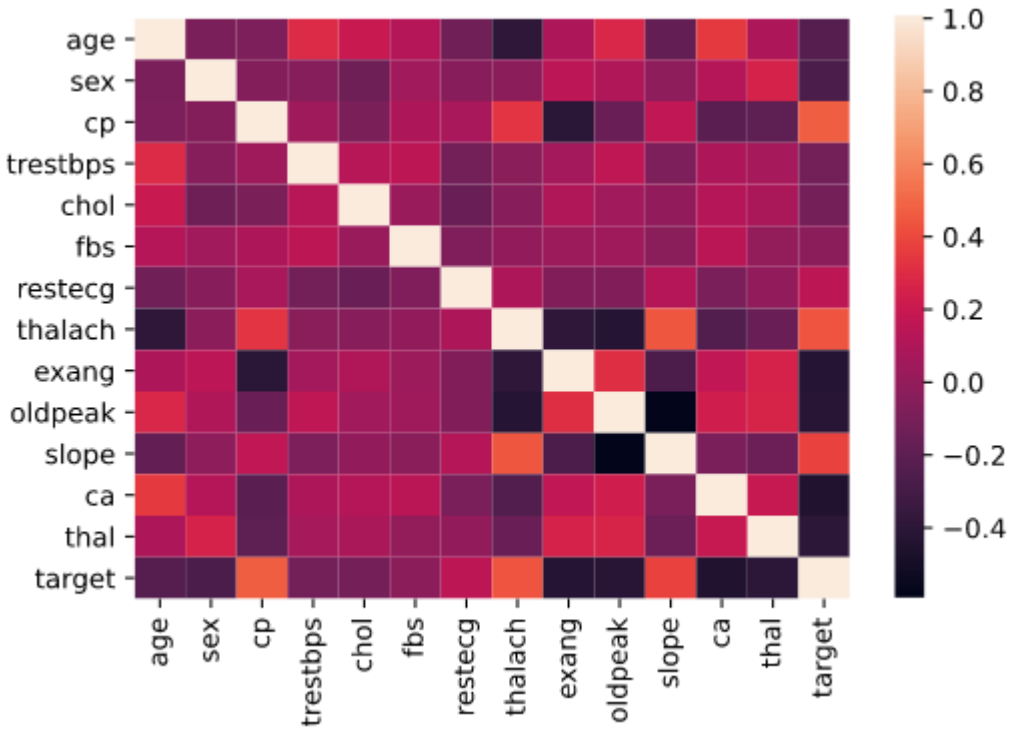
Compare the clustering algorithm to the classification algorithm. The clustering used within this project is K-Means and the classifications algorithm is LDA. The data we choose for this project is labeled heart disease data. The choice for this particular data was not a real concern as long as it was labeled.

Data link : “https://www.kaggle.com/ronitf/heart-disease-uci”

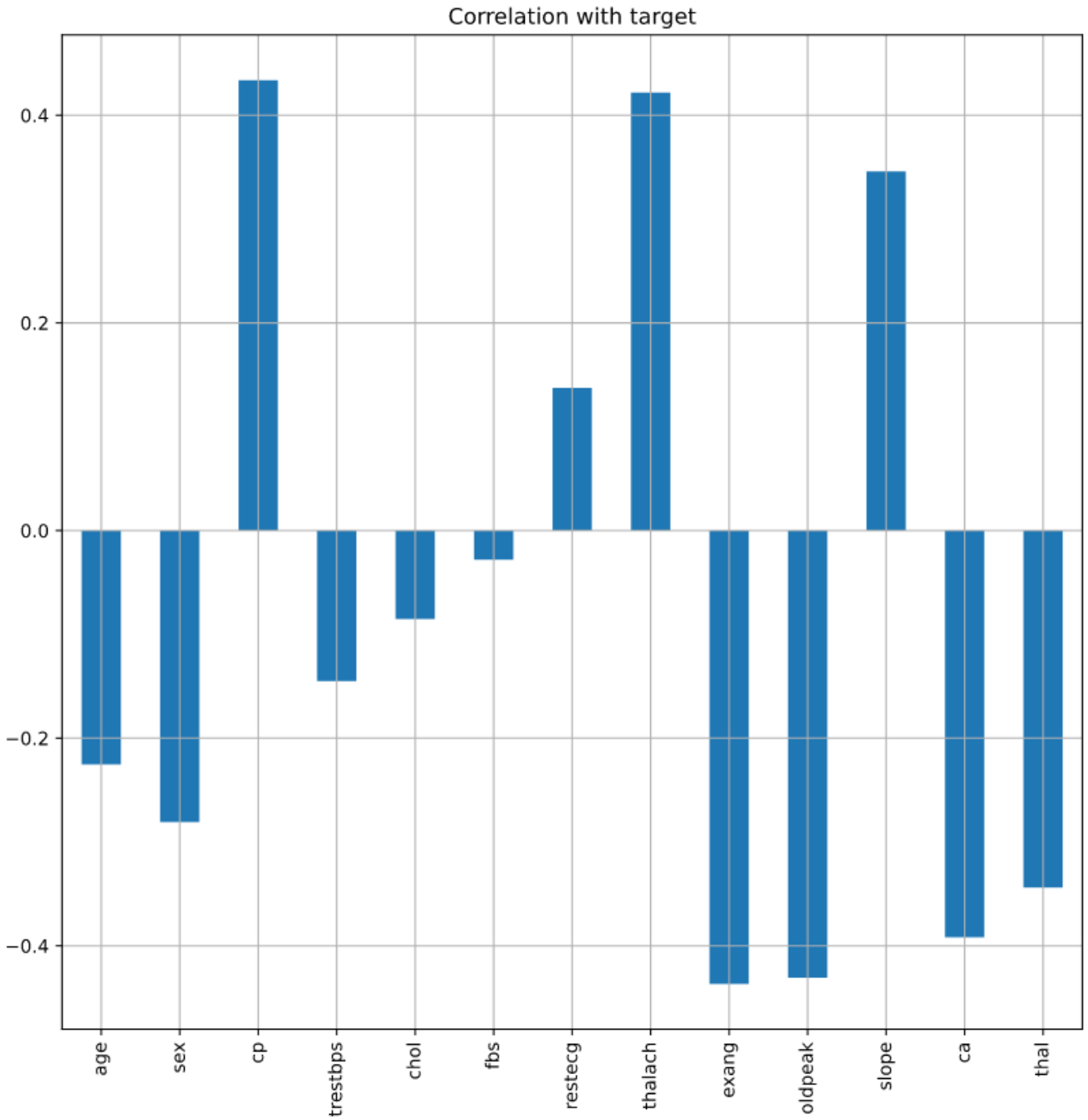
Exploring The Data

While exploring the data we found there were no null values and there was only one duplicate value which was removed.

Data Correlation



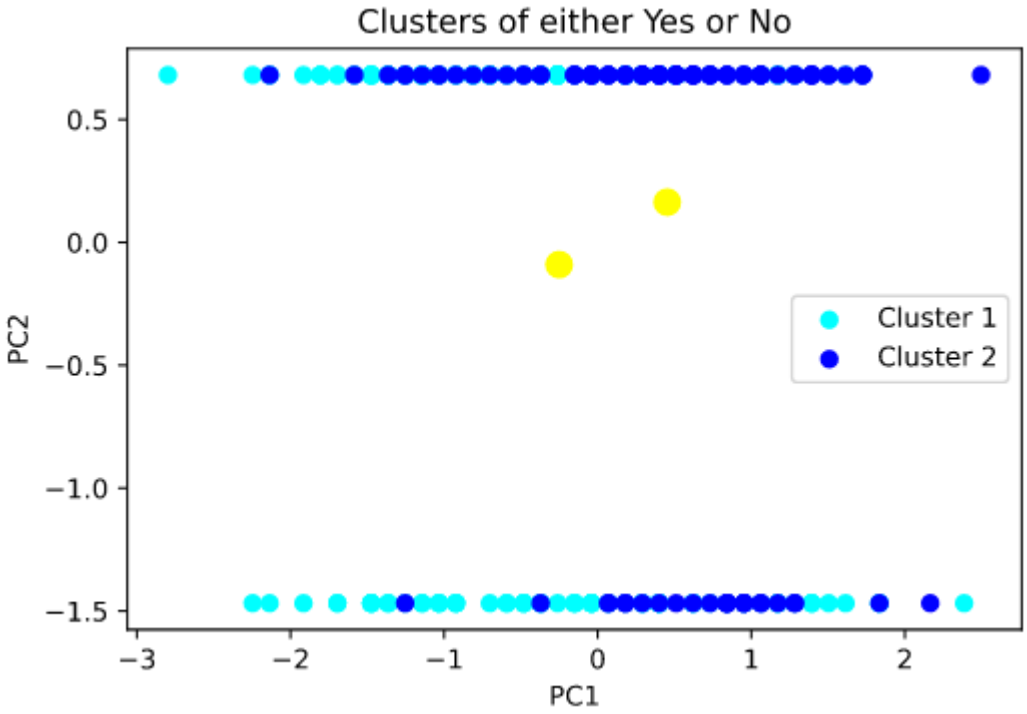
In this heat map there is a positive relation between CP , Thalach and Slope, and a negative relation between target and exang , oldpeak , ca and thal. This relations are also seen better in the next graph



The upwards bars represent positive relation with the target while the downward bars represent negative data, We could have removed strong correlating data but we choose not to remove this data because we wanted to try the k-means algorithm on a high dimensionality and compare it to when we use a dimensionality reduction algorithm such as PCA.

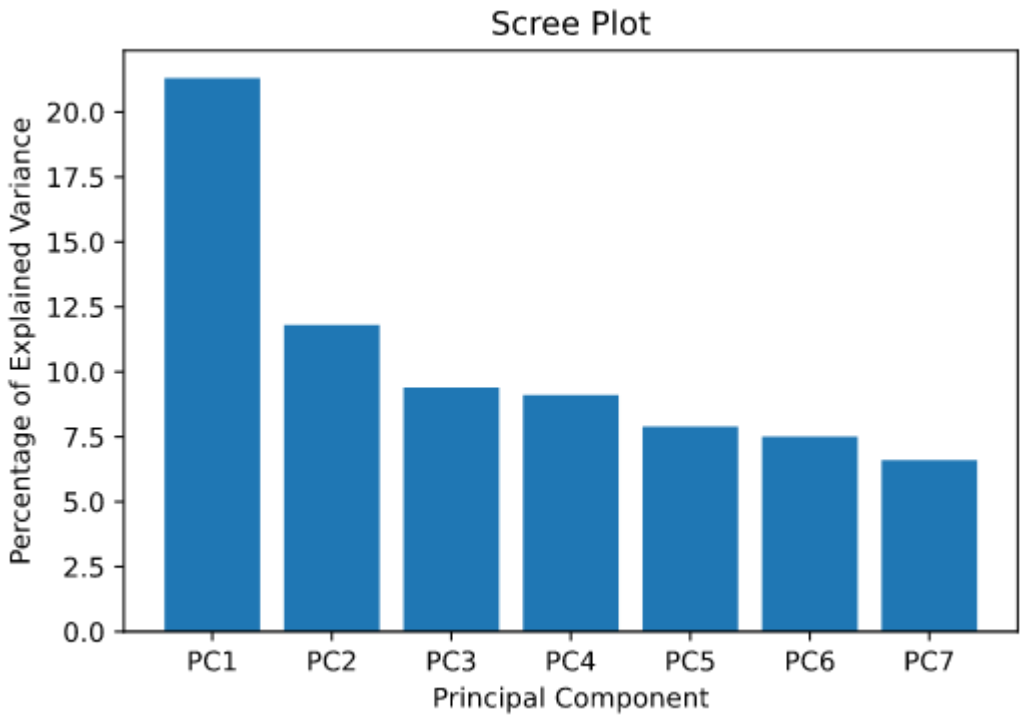
Normal K-Means

After standardizing the data and removing the label (so we can cluster the data) so that our results are not biased, we use the normal k-means algorithm on the data. Keep in mind that the data is in high dimension so our graph may not be accurate , also euclidean geometry tends to not be accurate as the dimensions get higher.



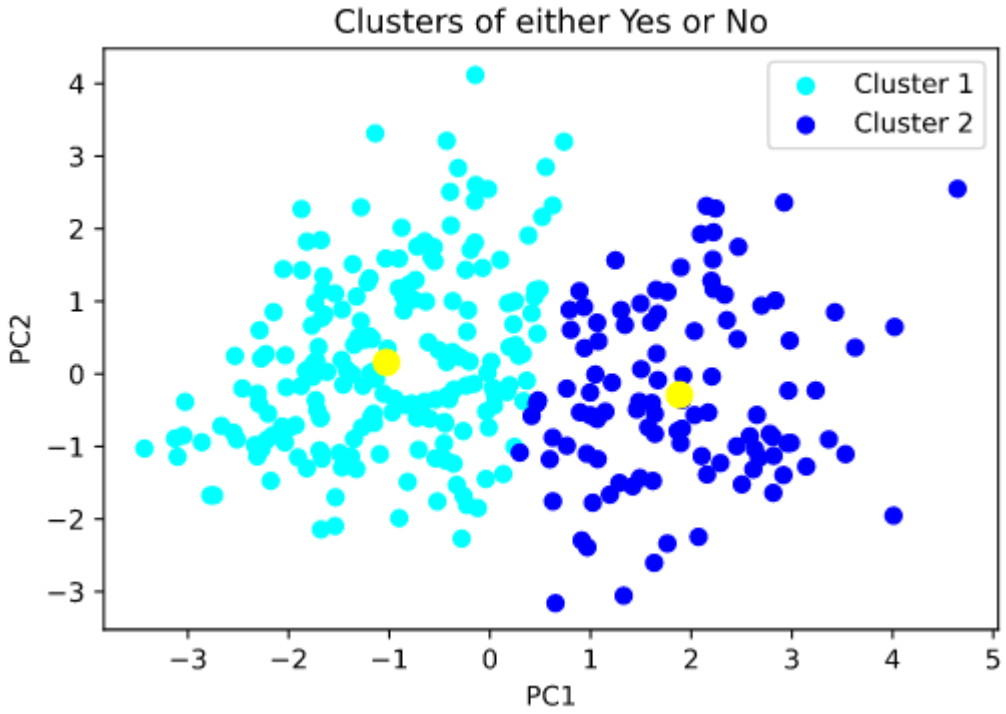
After we run the algorithm and compare the output to the label column we calculated a 81.5% accuracy which is relatively high considering the data.

Using PCA for dimensionality reduction



We utilized the PCA algorithm to reduce the dimensionality, and chose the first 7 PCs to visualize. And then to perform the K-means algorithm we chose PC1 & PC2 with a total of 33% (which isn't the best case, but because we want to visualize K-means on only 2 dimensions).

K-means after PCA:



After applying PCA to the data and then performing K-means on the reduced data, we can see that the clusters forms are more of what would we expect. With only 2 clusteres, one of them is for positive outcome, and the other is for negative outcome. After comparing this data to the labels (target), we discovered an accuracy of 80.5%.

LDA classification algorithm:

Linear discriminant analysis, we chose this model because we are the most familar with, and it's an overall good classifying algorithm.

After we split the data into traing set and test set (training set 0.75, test set 0.25), the accuracy from this algorithm, on the same dataset, came out to be 86.8%.

Conclusion:

Discoveries:

1. The benefit gained out of using a dimensionality reduction algorithm such as PCA allowed us to reduce computational power, have better visualization of the dataset, with only a minor cost in accuracy (81.5% - 80.5% = only 1% lost).
2. Although a classifying algorithm such as LDA has a better accuracy percentage overall, the difference between a classifying algorithm and a clustering algorithm such as K-means has only a cost of accuracy (86.8% - 81% = 5.8% lost). It's a relatively low cost for not having a labeled data.