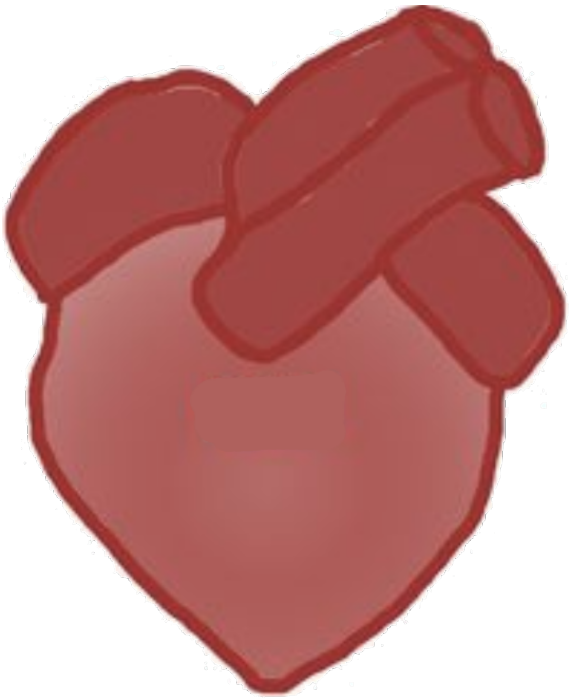
CHAMPLAIN COLLEGE DAT-310

Heart Disease Clustering

Eric Cacciavillani

February 5, 2019



.

1 About the data

1.1 Details

The given dataset contains patients with heart diseases. The original dataset has 76 attributes. But for class simplicity we are to focus on these given 11 features:

- **age**: Patient's age in years.
- **sex**: Patient's biological sex.
- **cp**: Patient's type of chest pain.
- **trestbps**: Patient's resting blood pressure.
- **chol**: Serum cholestoral in mg/dl.
- **fbs**: Patient's blood sugar is at fasting levels.
- **restecg**: Electrocardiographic results.
- **thalach**: Maximum heart rate achieved.
- **exang**: Exercise induced angina.
- **oldpeak**: ST depression induced by exercise relative to rest.
- **slope**: The slope of the peak exercise ST segment.

1.2 Dataset's Origin

The original dataset is from the UCI; which is a repository for datasets for developing and working with machine learning algorithms. However, the original distributors of the datasets were:

- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

2 Data Analysis

2.1 Correlation of Features

Looking bellow at **Figure 2.1** shows the correlation between features. High levels of correlation between features can cause those features to have an overall higher impact than intended when generating models. Some features in this dataset with high levels of correlation seems to be 'oldpeak' and 'slope'.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
age	1	-0.0975423	0.104139	0.284946	0.20895	0.11853	0.148868	-0.393806	0.0916608	0.203805	0.16177
sex	-0.0975423	1	0.0100839	-0.0644559	-0.199915	0.0478621	0.0216474	-0.0486633	0.146201	0.102173	0.0375329
cp	0.104139	0.0100839	1	-0.0360772	0.0723189	-0.039975	0.0675052	-0.334422	0.38406	0.202277	0.15205
trestbps	0.284946	-0.0644559	-0.0360772	1	0.13012	0.17534	0.14656	-0.0453509	0.0647625	0.189171	0.117382
chol	0.20895	-0.199915	0.0723189	0.13012	1	0.00984102	0.171043	-0.00343183	0.0613104	0.046564	-0.00406183
fbs	0.11853	0.0478621	-0.039975	0.17534	0.00984102	1	0.0695645	-0.00785415	0.0256651	0.00574722	0.0598942
restecg	0.148868	0.0216474	0.0675052	0.14656	0.171043	0.0695645	1	-0.0833894	0.084867	0.114133	0.133946
thalach	-0.393806	-0.0486633	-0.334422	-0.0453509	-0.00343183	-0.00785415	-0.0833894	1	-0.378103	-0.343085	-0.385601
exang	0.0916608	0.146201	0.38406	0.0647625	0.0613104	0.0256651	0.084867	-0.378103	1	0.288223	0.257748
oldpeak	0.203805	0.102173	0.202277	0.189171	0.046564	0.00574722	0.114133	-0.343085	0.288223	1	0.577537
slope	0.16177	0.0375329	0.15205	0.117382	-0.00406183	0.0598942	0.133946	-0.385601	0.257748	0.577537	1

Figure 2.1: Feature correlation map.

Looking bellow at **Figure 2.2** shows the correlation averages for features. Which highlights that oldpeak, slope, and exang have the highest average correlations.

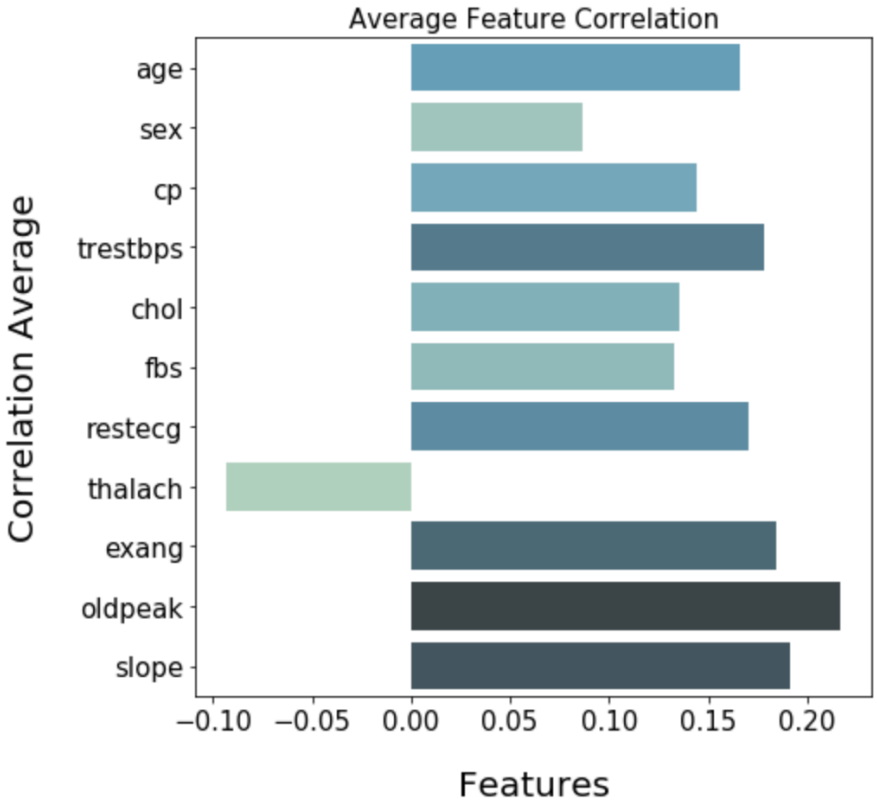


Figure 2.2: Feature correlation means ranking.

2.2 Graph independent features

Listing off some graphs that seemed interesting. To see more inspect the ipython notebook.

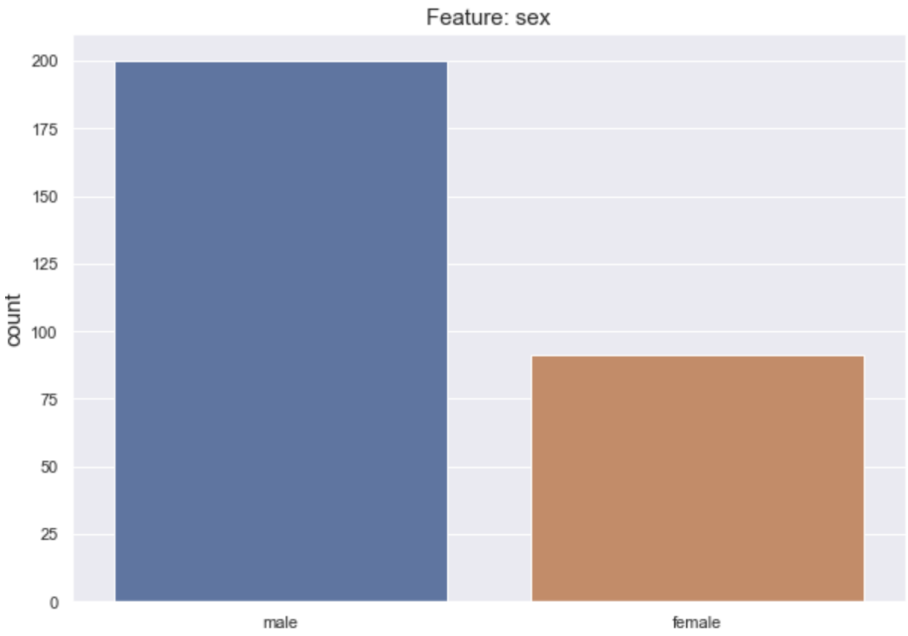


Figure 2.3: Around twice as many males as females in this dataset.

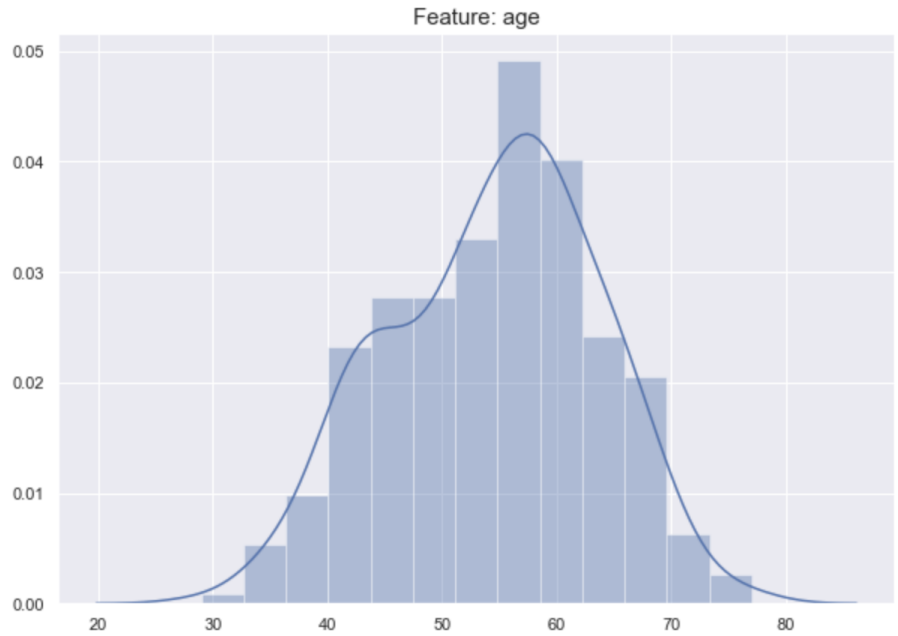


Figure 2.4: Slightly skewed distribution. Most common age is between 51 years to 60 years.

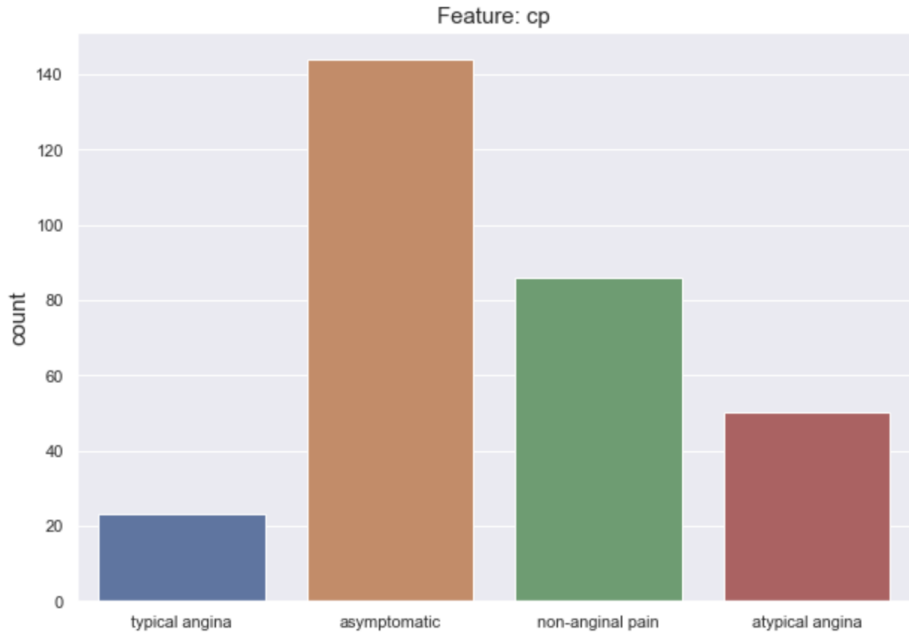


Figure 2.5: Asymptomatic is the most frequent type of chest pain for this dataset.

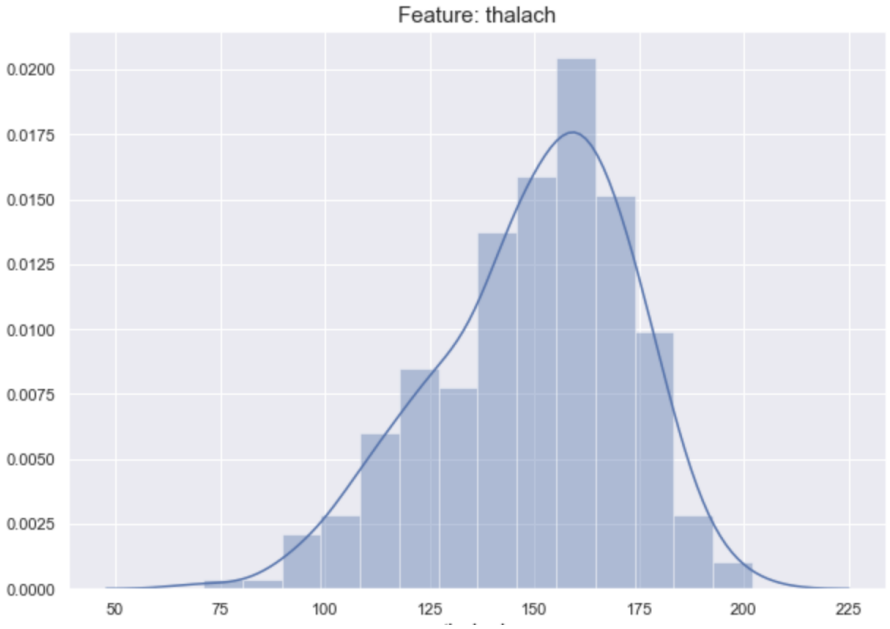


Figure 2.6: Clean distribution between the values of max heart rate pf 100bpm to 200bpm.

3 Data transformation

3.1 Dimension reduction

Removing features 'slope' and 'oldpeak' due to high feature correlation (**Figure 2.1 and Figure 2.2**) and strange internal distributions.

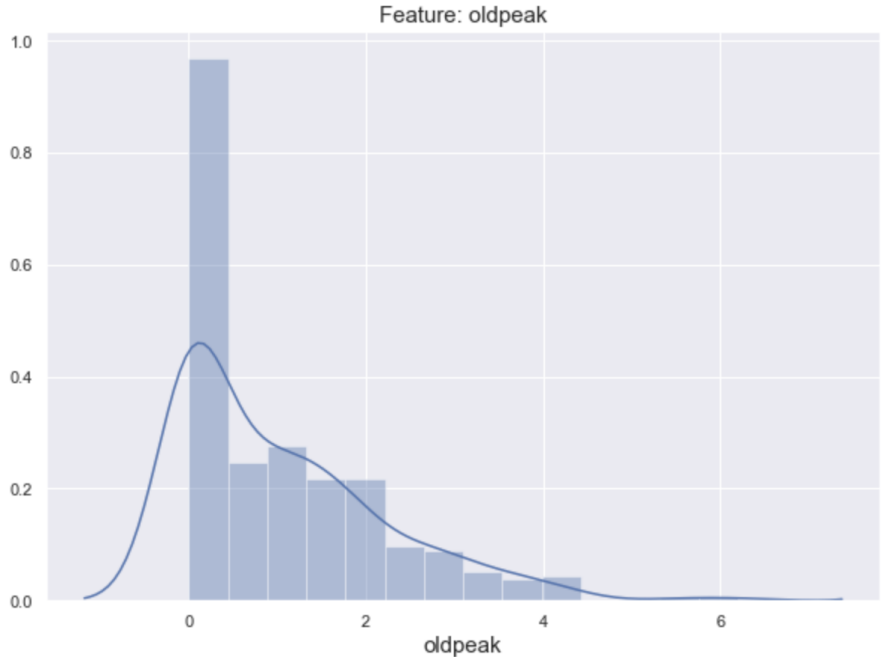


Figure 3.1: Ample amounts of values near zero make it impossible to make 'oldpeak' properly distributed.

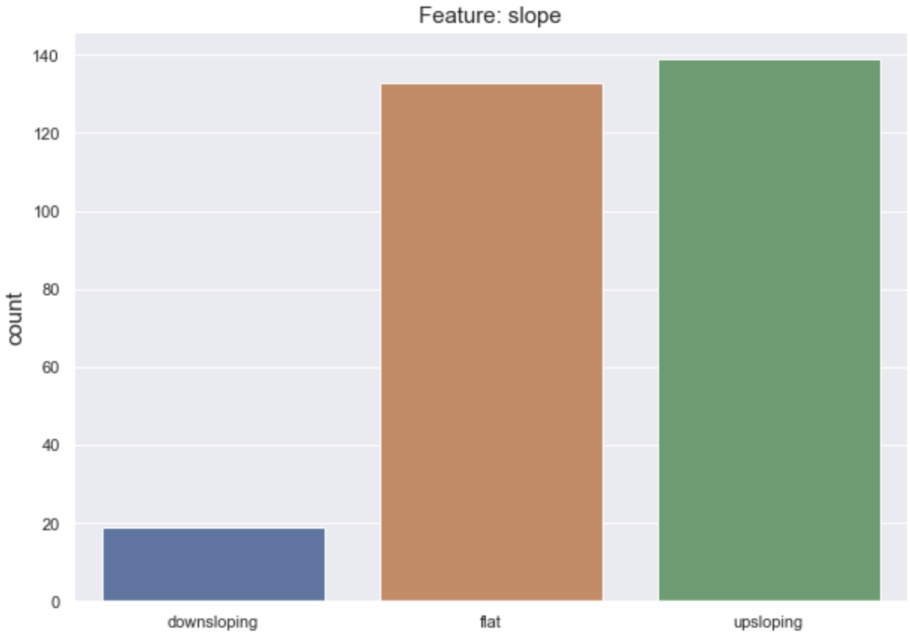


Figure 3.2: Feature 'slope' doesn't have enough variation within itself to make a strong enough impact for modeling data.

3.2 Fixing numerical features

After removing outliers for each numerical feature and then centering each using np.log we get a much cleaner and more centered distribution for each feature.

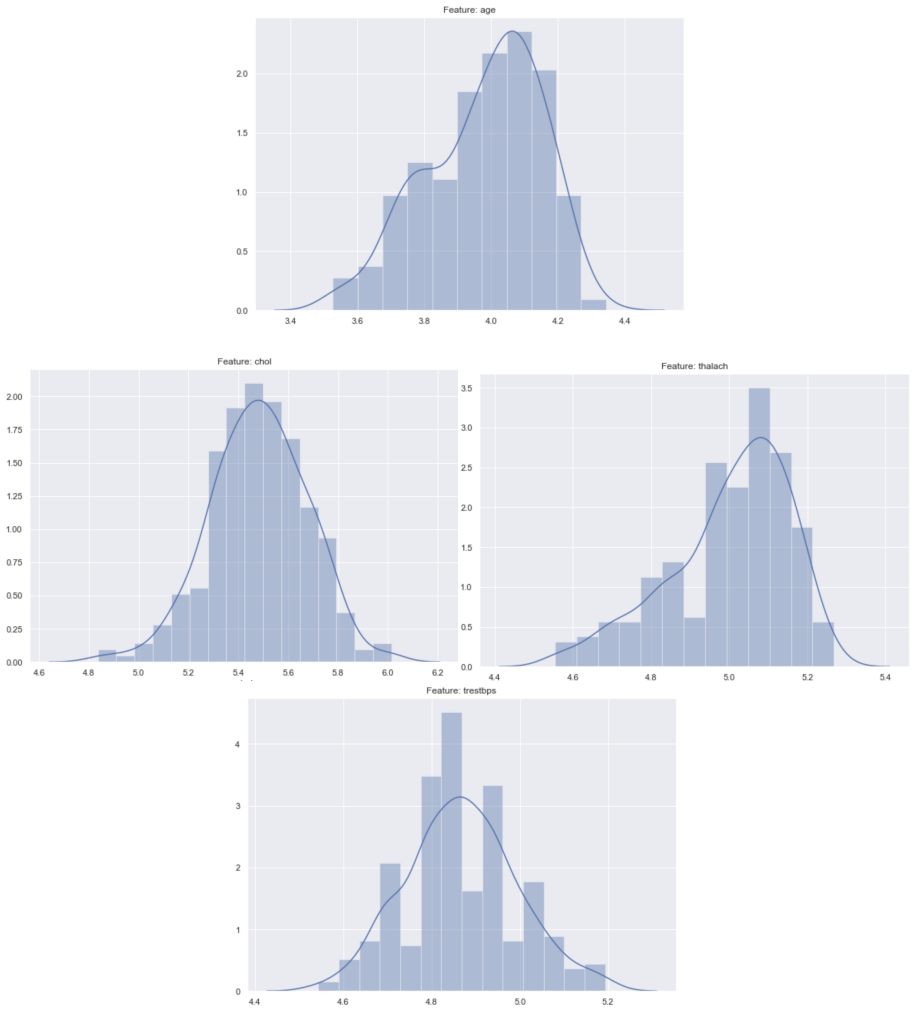
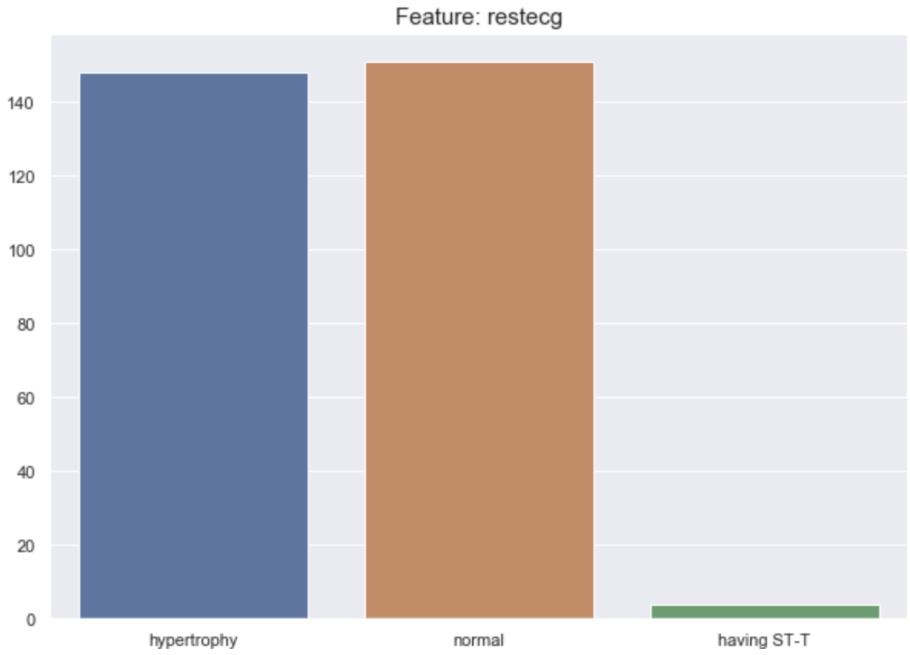


Figure 3.3: Centered and outliers removed

3.3 Remove feature values

Remove 'having ST-T' from feature restecg due to only one patient with a ST-T.



3.4 Re-Configure the data

Now we need to re-configure the data to ensure our data is properly clusterizable. First, we need to apply one hot encoding to our categorical features so that their encoded values aren't applied as having quantifiable value. Second, we need to properly scale our data so features with with arbitrarily bigger values so that each feature is interrupted with very similar weights. Next, apply PCA to the data to de-correlate the relationships between features. Finally, re-scale the data down again. The data is now ready to be clustered.

4 Clustering

4.1 Find best k value for KMeans

Generating multiple kmeans models with different 'k' values for evaluation. A lower 'k' is better for simplicity; but is susceptible to having higher variation(inertia) between the points. Thus, we want to take the area where the inertia first dips also know as the 'elbow' of the graph.

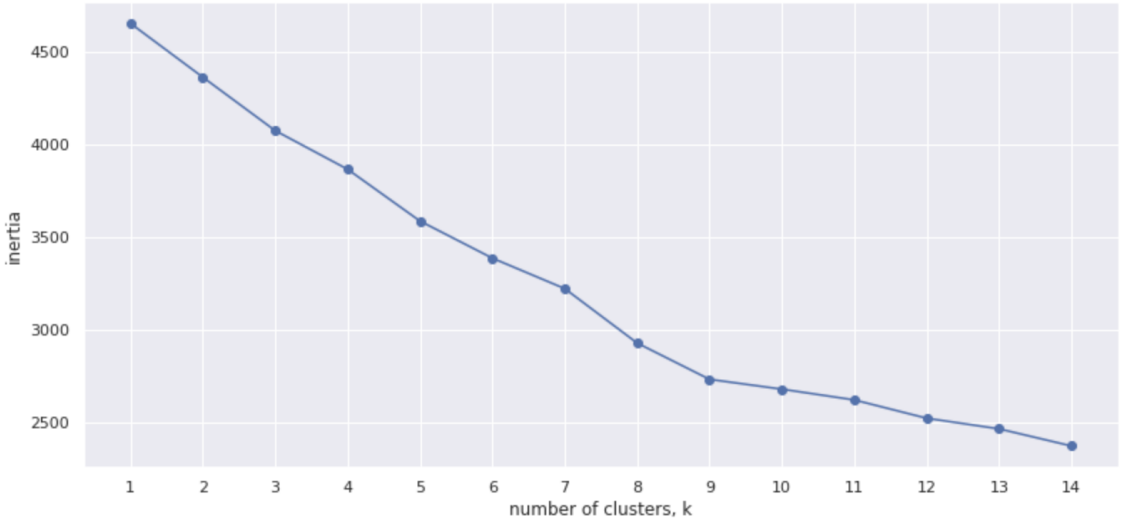


Figure 4.1: Best 'k' value seems to be 8 using the elbow method.

4.2 Confirm k value by looking at Hierarchical clustering

Looking at another graph to confirm that our given elbow found 'k' of 8 is within the bounds of reason for a good clustering model.

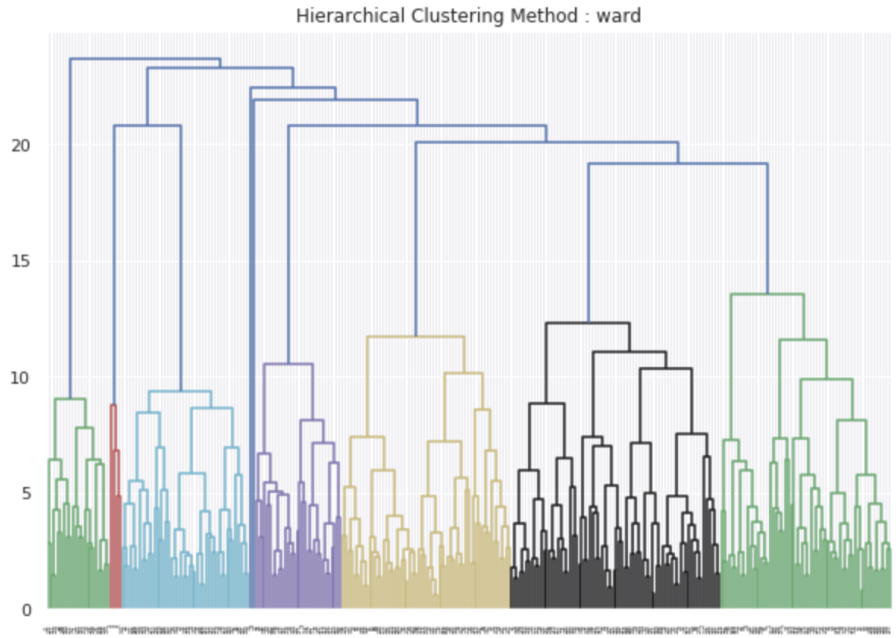


Figure 4.2: 'k' seems to be between 7-8 in this case.

4.3 Visualize k=8 kmeans model

Data points are very spread out even amongst their own clusters; overlapping between clusters is quite apparent. Does seem to be three semi-distinct clusters. Clusters: [1,3,7].

Figure 4.3: Clusters 4,5,6 are dropped for only having one patient.

4.4 Compare cluster profiles

4.4.1 Mean of clustered profiles

These profiles are generated by creating a subset of data associated with cluster label; then removing outliers of this subset by finding any zscore value at or between 2 and -2. Finally, we get averages of each feature and then move those given averages back to known feature data.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
Cluster:0	50	male	atypical angina	124	244	false	normal	163	no
Cluster:1	55	male	asymptomatic	128	242	false	hypertrophy	152	no
Cluster:2	56	male	typical angina	140	239	false	hypertrophy	158	no
Cluster:3	52	male	non-anginal pain	130	239	false	normal	158	no
Cluster:7	56	male	asymptomatic	130	249	false	hypertrophy	134	yes

Figure 4.4: Only varying feature is chest pain.

4.4.2 Kmeans clustering decision tree visualization

Modeled a decision tree that was fitted to outlier reduced clusters.

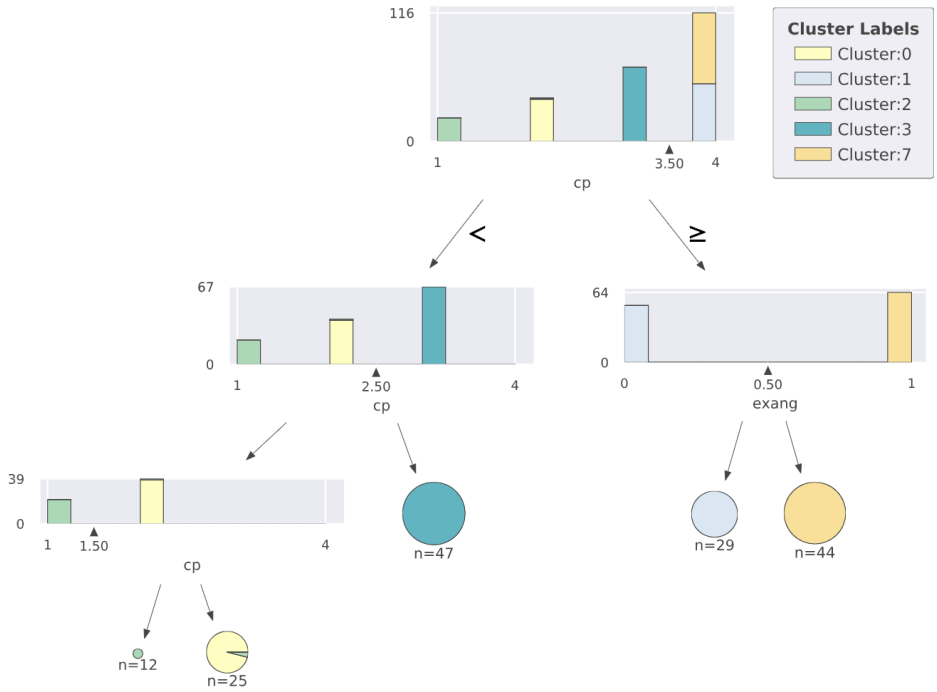


Figure 4.5: The descion tree is showing that the current kmeans model is entirely clustering around chest pain and exercise induced angina.

5 CONCLUSION

This data is not very clusterizable due each heart disease patient having attributes overlapping with one another. Thus, making it very difficult to find any meaningful distinction between the clusters. The clusters that were found found were very spread apart and overlapped with one another (**Figure 4.3**). Furthermore, any features that did help properly cluster were limited. Ex: **Figure 4.5** shows how only 2 features were needed to cluster nearly the entire dataset; further highlighting the fact that the patients' data overlapped heavily across the rest of the given features.