Bisma Nazir
L21-7310
Assignment 3
Clustering
Data Mining Fall 2022

1. **Pre-Process the dataset using different techniques like normalization, discretization, concept hierarchy, and correlation. Explain which pre-processing steps are performed on the data and why?**

Clustering:

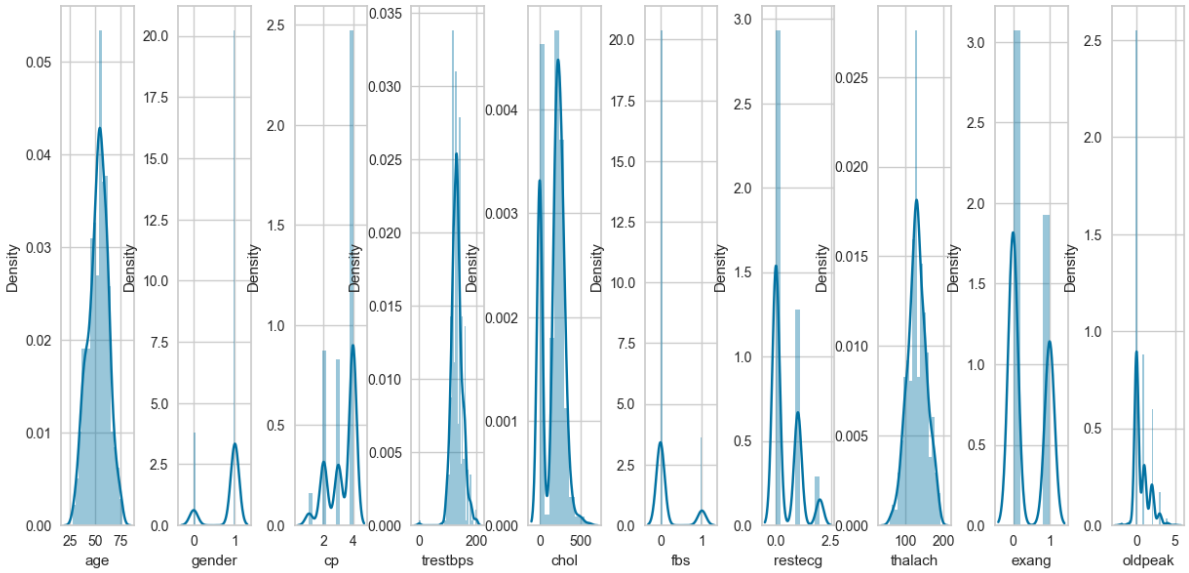
Clustering is used to group data points having similar features. It is used for segmentaion i.e if we want to segment a picture according to colors. There are many ways of clustering types out of which some are supervised clustering and some are unsupervised clustering.

We have a heartdisease data on which we want to apply Kmeans, DBSCAN and Heirarchical clustering to cluster people according to the heartdisease depending upon different conditions like fasting blood sugar, exercise etc.

First step to perform any algorithm on dataset is preprocessing data. On heartdisease data we will preprocess data to apply Kmeans and other clustering algorithms.

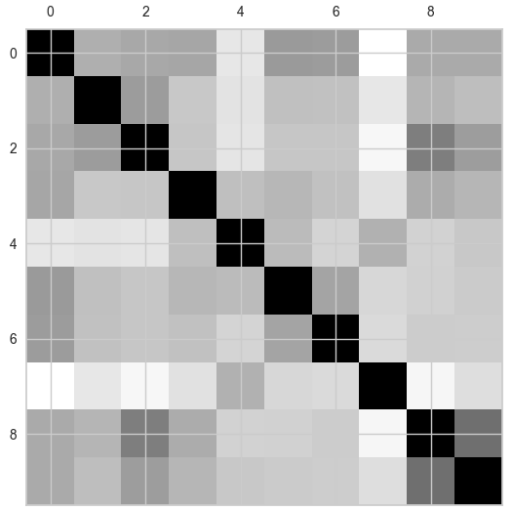
Preprocessing:

- **EDA:** We will perform exploratory data analysis on data to deeply analyse and understand dataset. Distance plot to see division of data.

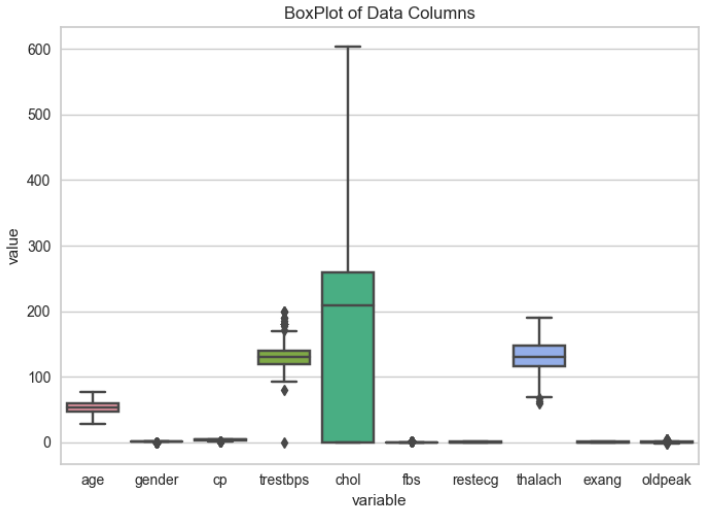


- **Missing Values:** There are many missing values in almost all the columns. We will remove the features that have 50% or more missing values, so we will remove ca, thal and slope from data as they have major portion of missing values.
- **Replace Missing Values by Mean:** We will first remove null values by applying NaN to replace "?" which shows missing value in dataset. Then we will fill missing values by mean of the column.

- **Correlation:** We will see correlation of the data because features that have high correlation among them provide the same information. We set a threshold of 0.5. attributes having 0.5 or more correlation will be removed. In our dataset no features are correlated so we will not remove any feature in this phase.



- **Outliers:** We will draw boxplot to see if there are any outliers in our data. We found very less outliers in trestbps and as our data is very small so we will not remove the outliers and move on with the same data.
  Box Plot



- **Mixed Datatype:** We cant apply clustering on mixed datatype. We have to get the data in same space in order to perform clustering. For example if we have binary, discrete and continuous data we will change the dataset to either make all the attributes continuous type or discrete. For reference please visit [1]. So the features that are discrete we will use **one hot encoding** on those features to get them in same space.
- **Normalization:** We will normalize data as we use Euclidean Distance for clustering algorithms and Euclidiean distance is really affected by values, so we need our data to be within some scale. For scaling and normalization we will do minmax normalization. We will not apply normalization on binary data.

- **Feature Reduction:** Class is not used in clustering so we will remove that feature.

2. **Formulate a question for which you wish to cluster the dataset. Select a subset of attributes that can be beneficial for your task.**
   Question: We want to Cluster data according to the heartdisease based on some conditions i.e whats their age, cholesterol level, fasting blood sugar etc.
   Feature Engineering: Heart disease has nothing to do with gender so we will remove gender now we are left with 9 features.
   PCA: We will also apply PCA to reduce data dimensions. So we will reduce dimensions to 3 visuals become clear.
3. Perform **K-MEANS** clustering on the given dataset.
   a. **Run K-means multiple times for each K. Report your findings (error in each clustering, the time required)**

- We apply elbow method in range of 2-15 to see which K is best to perform clustering. Its giving K=7 as optimized K in terms of time which 0.14 seconds and score which is 402.96.

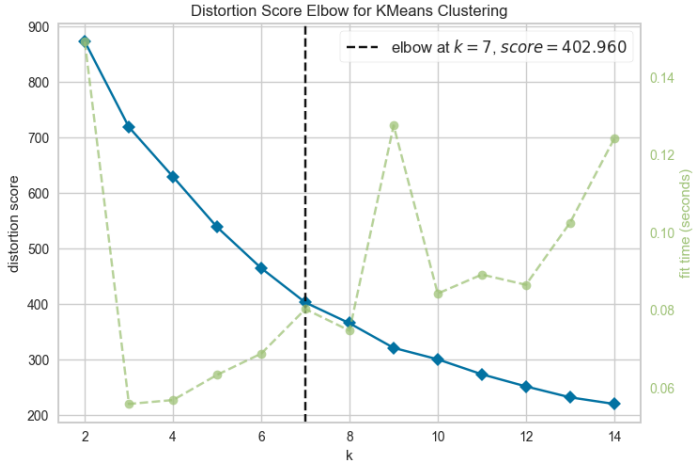


Figure 1: Optimal K using Elbow method

- We find silhouette score to see quality of clusters. Which is approximately 0.42 for K=7.

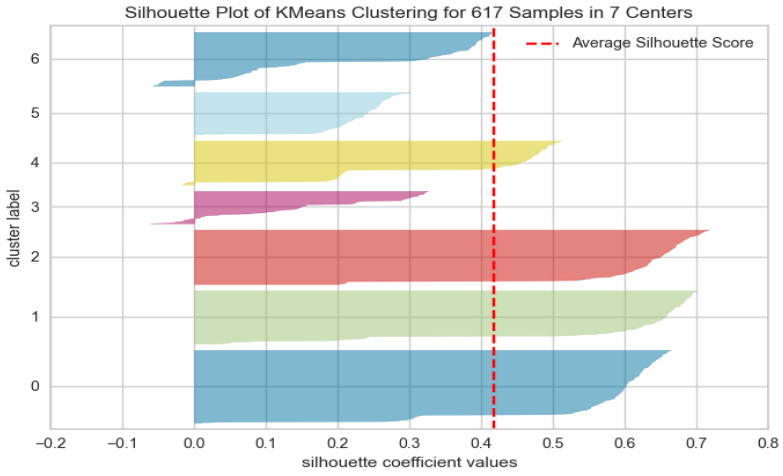


Figure 2: Silhoutte Score for K=7

- Then we apply Kmeans for K=10 and save lables.
- We apply PCA to apply Kmeans on another subset of attributes. We reduce the attributes to 3. This is the projection of PCA data.

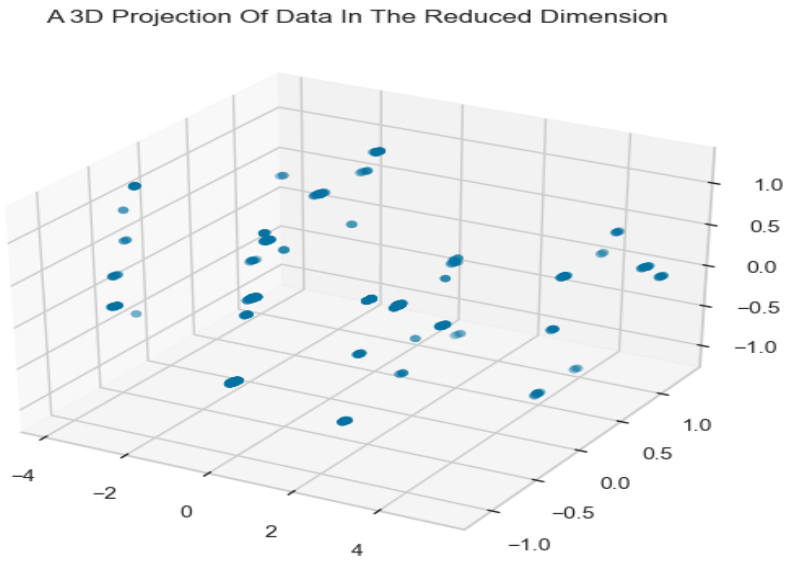


Figure 3: PCA projection of Data

- Then we apply Elbow method to see optimized K on PCA reduced data. Which also gives same value of k=7 with score 171.655 decreased and time also decreased to perform the algo as 0.1second. So, results show that PCA is giving good results so we will use this method as our next process.

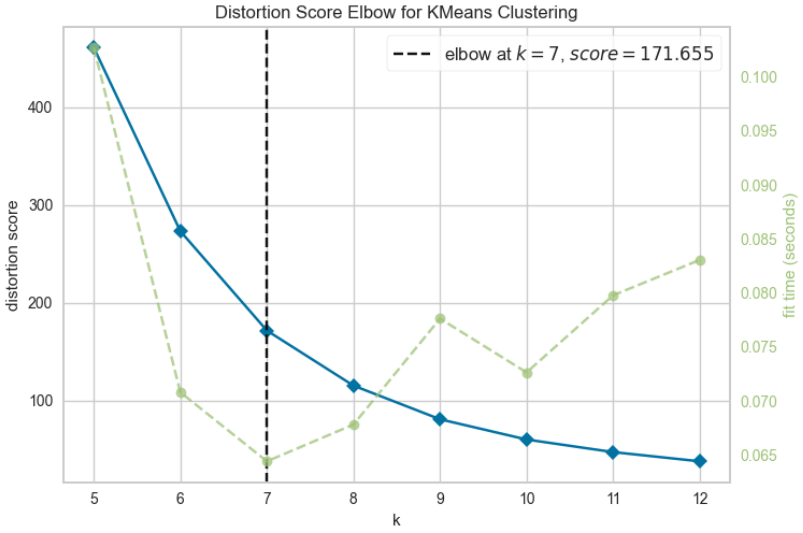


Figure 3: value of K using Elbow Method on PCA reduced data.

- Then we apply clusterin with k=8 just to see results and compare with other values of K.

- Then we calculate silhoutte coefficient for k in range of 5-13 for PCA reduced data. Minimum SSE is 34.3500469022975 which is distance between the datapoints of dataset.
- After that we calculate SSE for PCA reduced data for K in range of 3-13. This is trend of silhouette coefficient.

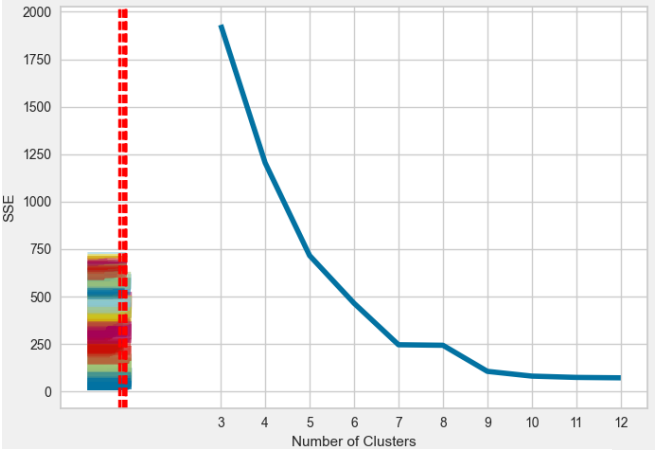


Figure 5: SSE for K in range of 3-13

- We visualized features of PCA in 2d plan.

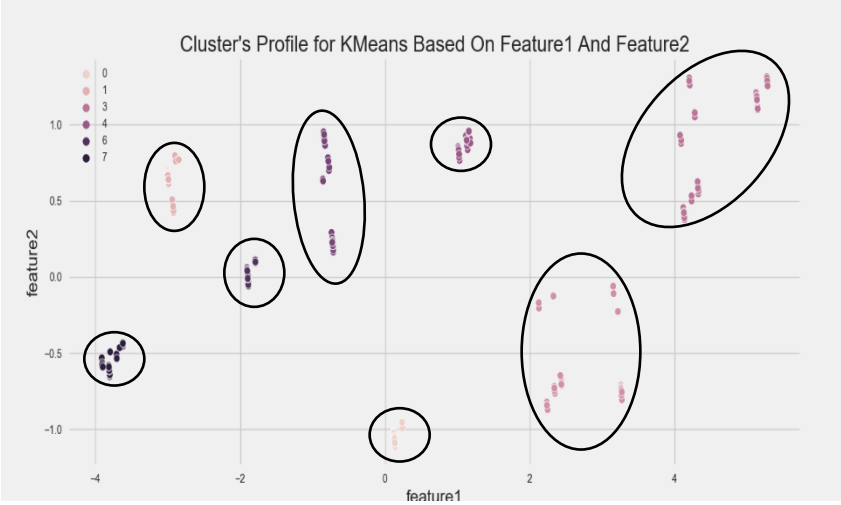


Figure 6: Clusters in 2d plan.

- Clusters in 3d plan

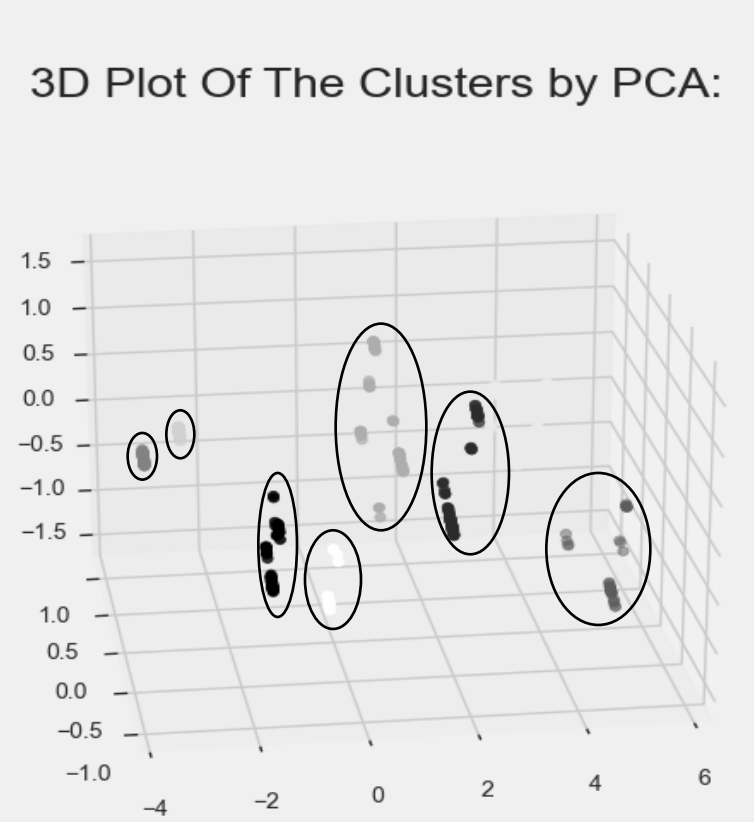


Figure 7: Clusters in 3d plan

b. Run K-means with different K.

- We Run K in range of 3-13. Then we separately run for K=8 and K=10. We observed that K=7 gives best results both in term of time and distortion score.

c. Report the K that gives the best result for each subset of attributes.

- We apply elbow method in range of 2-15 to see which K is best to perform clustering. Its giving K=7 as optimized K in terms of time which 0.14 seconds and score which is 402.96 with 9 features as age, trestbps, chol, thalach, oldpeak, cp, fbs, restecg, exang.
- We also applied PCA to rduce the dimension of data to 3. Then we apply elbow method which gives best result at K=7 with distortion score as 117 and time 0.1 second.

d. Indicate the number of iterations to convergence for different runs.

For K=7 we iterated K for 100 times and the convergence is on after 15 iterations with minimum SSE of 42.

e. Examine the quality of clusters and also of clusterings. Report the errors: within-cluster sum of squared error and between-cluster sum of the square error for each run of K-mean.

- We observed clustering with different k values and the best value is found to be k=7 as it has lowest SSE of 42 and highest silhouette score. Kmeans took 0.03795909881591797 miliseconds to perform on this dataset.

4. Also, Cluster the dataset using **Hierarchical Clustering (single link, complete link, average link) and DBSCAN.**
    a. **Run for different values of the number of clusters. Include the dendrogram(for heirarchical) and time taken by each clustering in your report.**

Heirarchical Clustering:

- **Complete Linkage:** We applied a loop between 3-15 and the best cluster is found to be k=7. The 3d plot for clusters is given below.

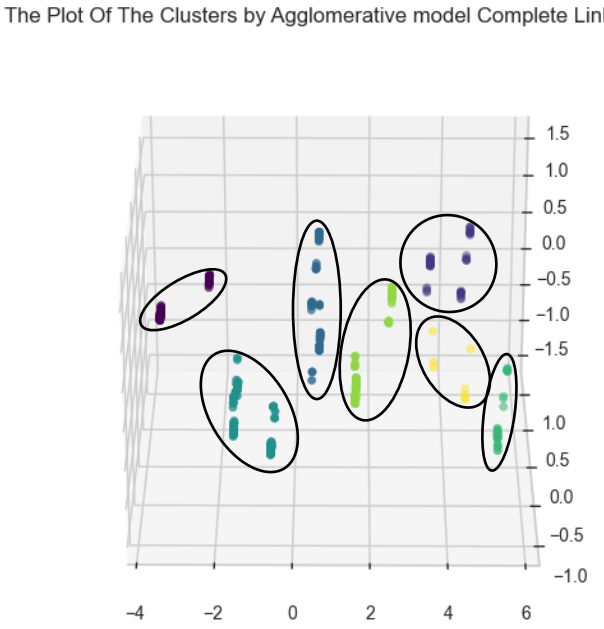


Figure 8: Clustering in 3d plan with complete linkage

    - Complete Linkage in 2d plan:

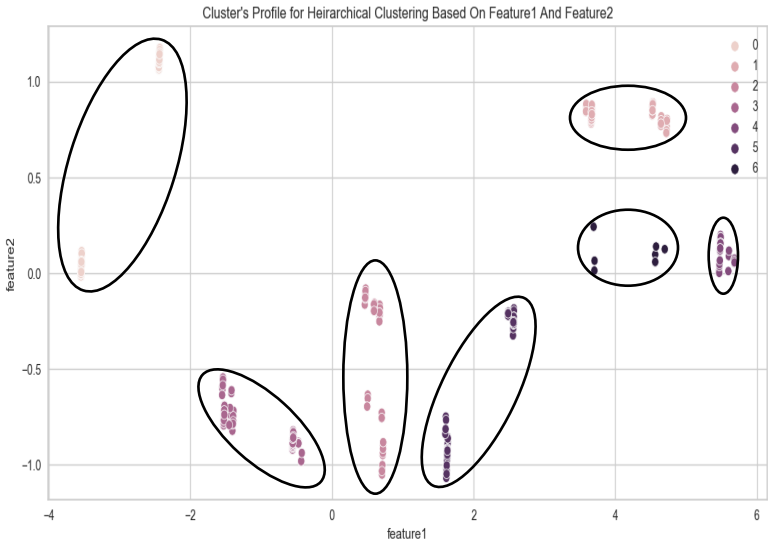


Figure 9: Clustering with Complete Linkage in 2d plan.

- Dendrogram of complete Linkage

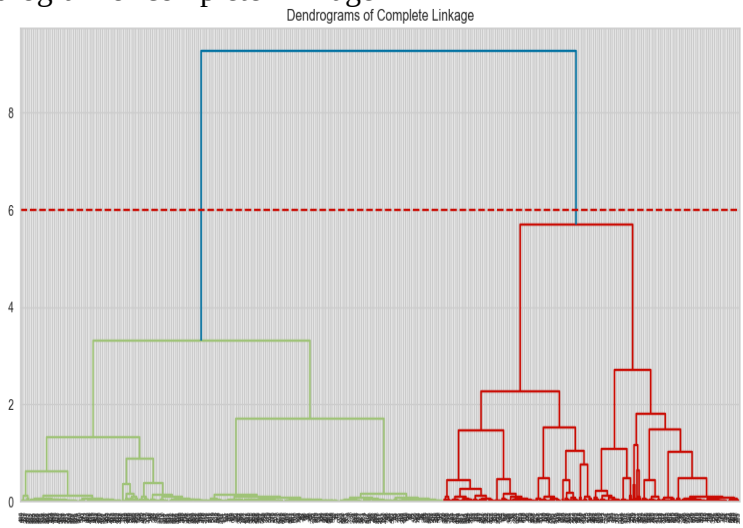


Figure 10: Dendrogram of Complete Linkage.

- **Single Linkage:** Time taken to perform Single linkage with k=7 is 0.007350444793701172 miliseconds.

- **Single Linkage in 3d Plan**:

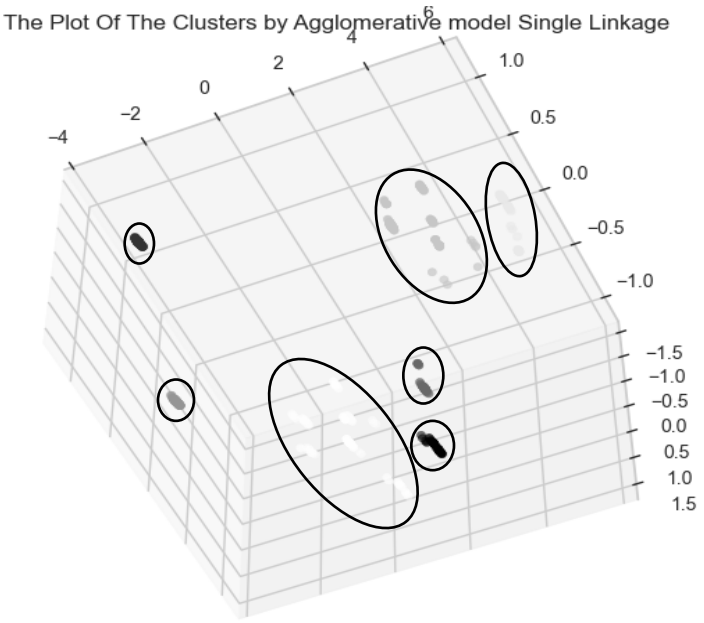


Figure 11: Clusters using single linkage.

- **Single Linkage Dendrogram**:

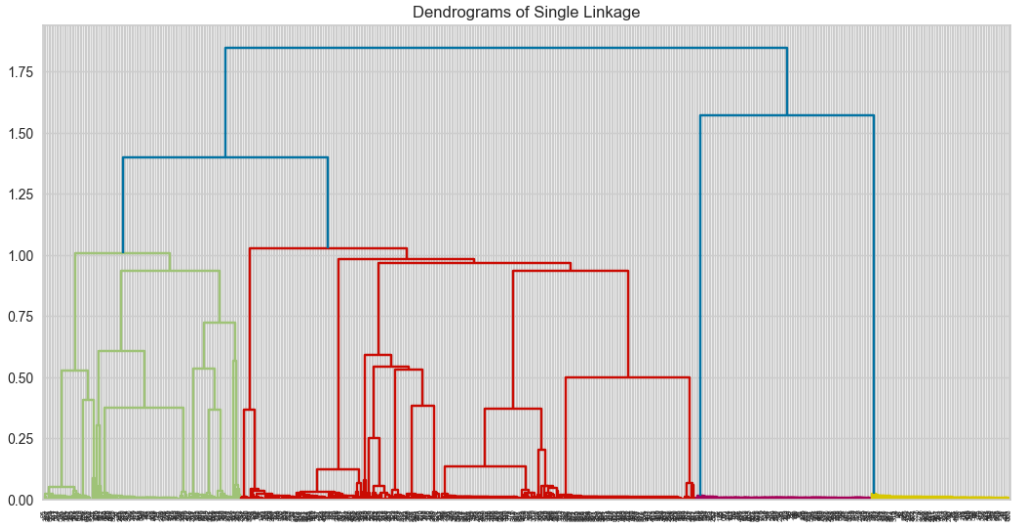


Figure 12: Dendrogram for single Linkage

- **Average Linkage:** Time taken to perform average linkage with k=7 is 0.010043859481811523 miliseconds.
  - Average Linkage in 3d Plan:

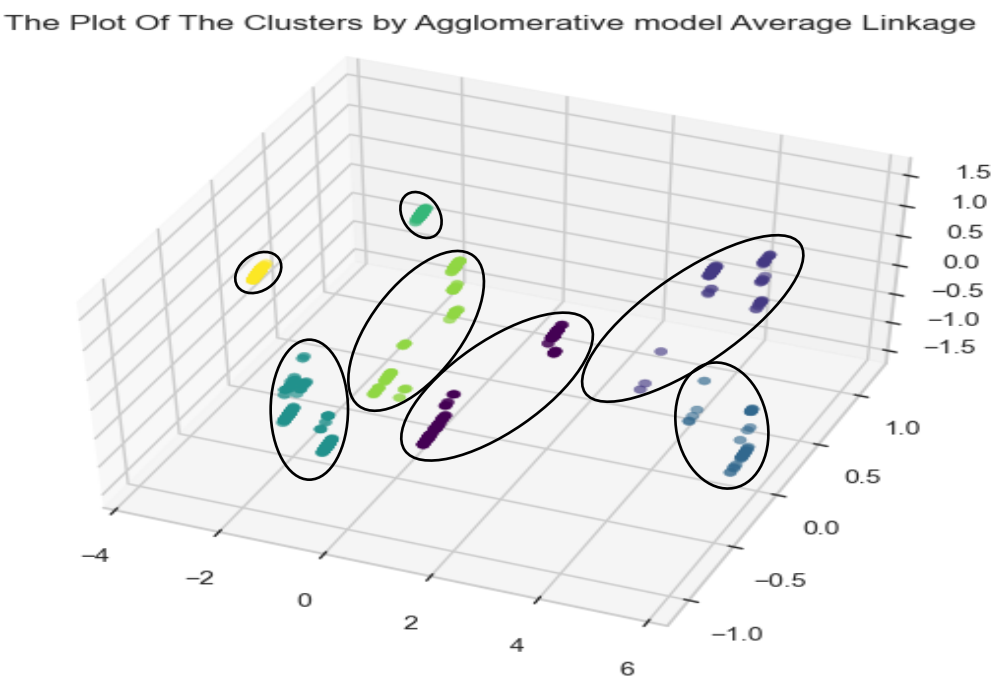


Figure 13:Clustering using Avergae linkage in 3d plan.

- Average Linkage Dendrogram

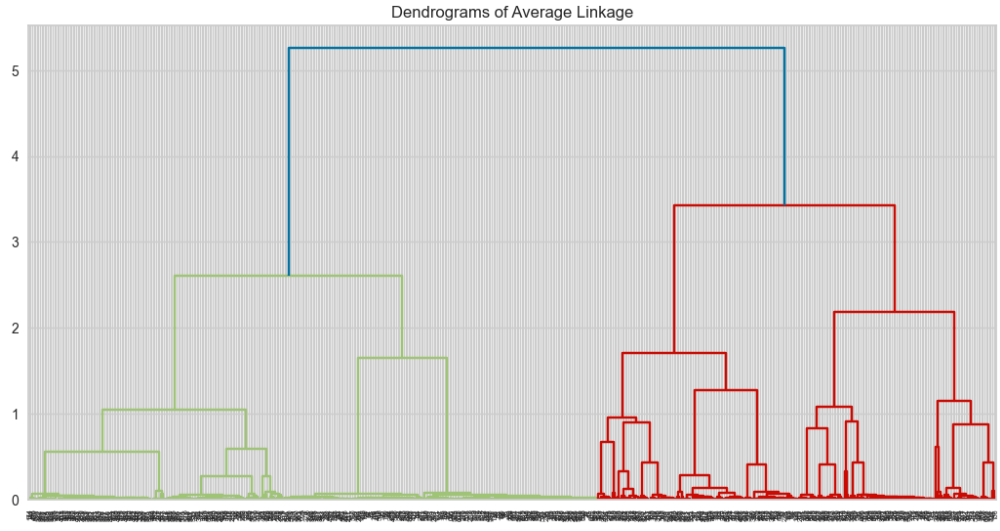


Figure 14: Dendrogram for average linkage.

- **DBSCAN**

DBSCAN is a density based clustering algorithm which separates the the data points in low density from the data points in high density. It covers the lacking points of Kmean clustering. Here we

don’t need to provide number of clusters we need to provide eps size and minimum number of points in eps circle. In this way its less prone to noise and outlier points.

DBSCAN in 3d plan: with eps size 0.1 and min-samples=5:

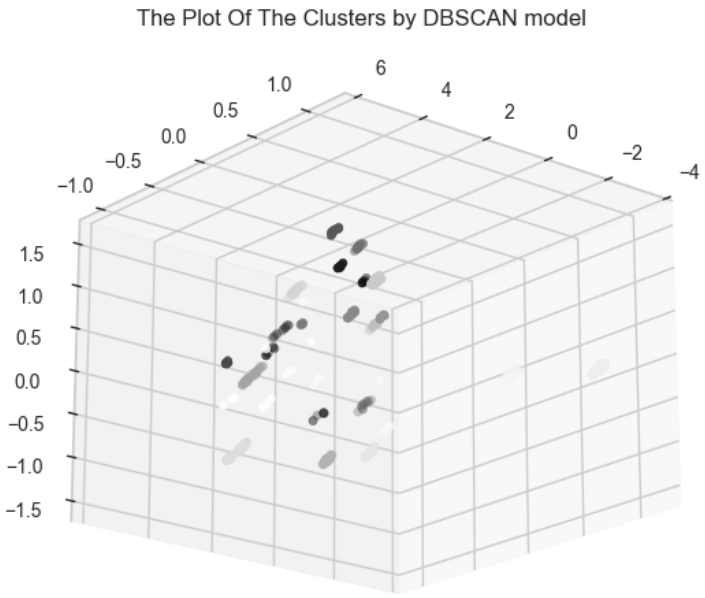


Figure 15: DBSCAN clustering

DBSCAN in 2d plan: when applying DBSCAN on dataset with eps 0.1 and min samples as 5 it automatically makes 21 samples

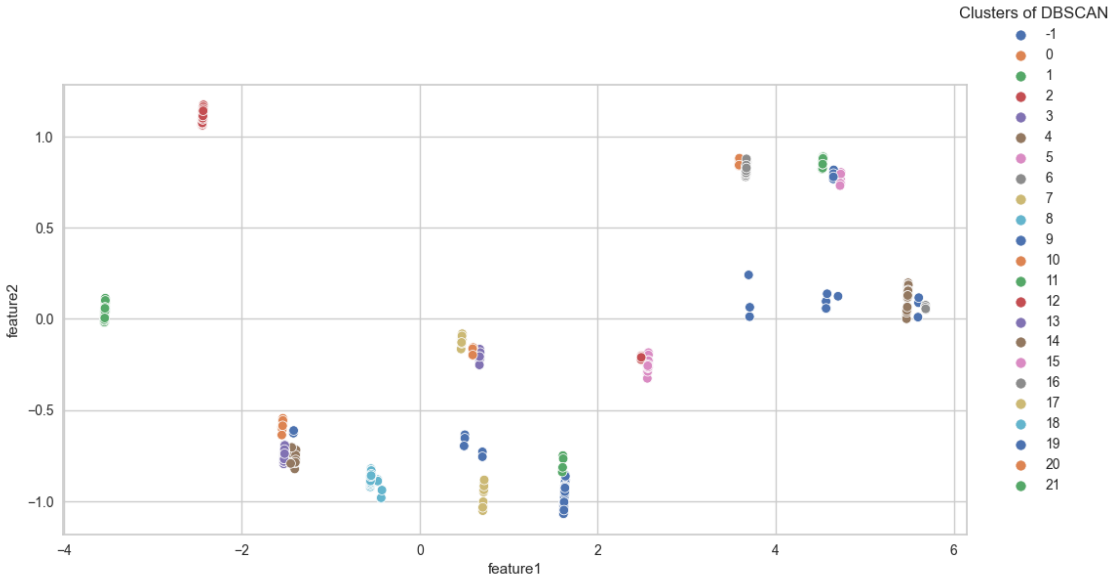


Figure 16: DBSCAN clustering in 2d Plan.

If a part of 2d plan is zoomed it, the clusters looks like this:
Noise point: cluster -1 shows the noise points.

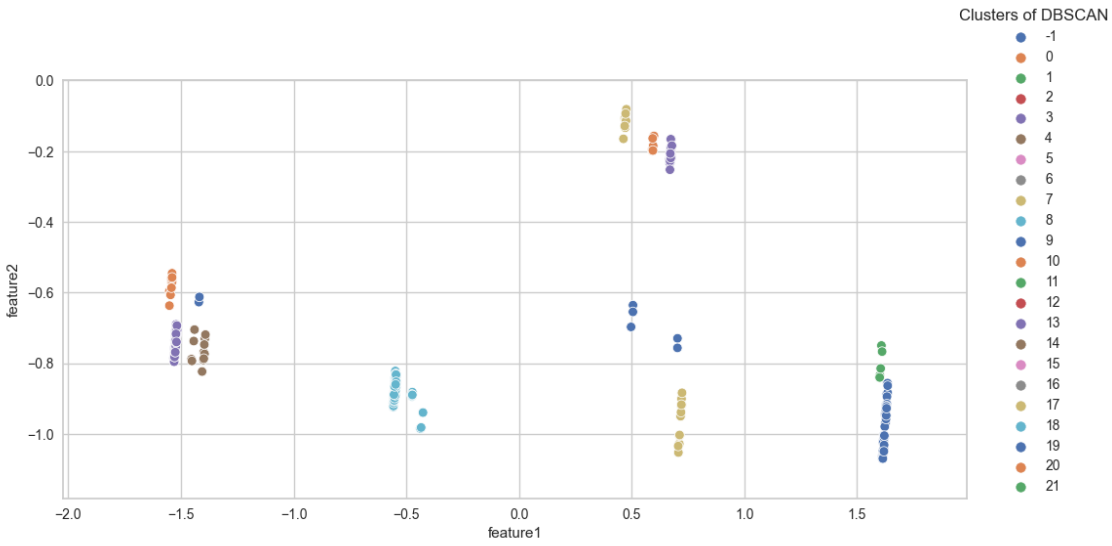


Figure 17: DBSCAN clusterin portion in 2d plan.

5. Draw different plots to visualize the clustering results. (include plots in your report).

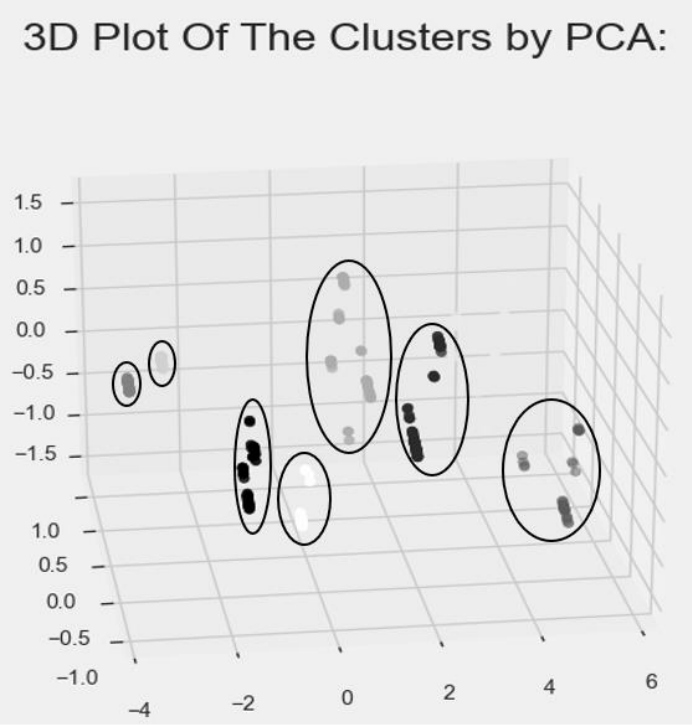
- Box Plot to see outliers.
- Distance plot to see division of data.
- Heatmap for correlation.
- Clustering in 2d and 3d plans are shows.
- Silhoutte score is visualized along with the number of K.
- SSE is visualized along with K.
- Elbow method is shown to pick best value of K.
- Dendrograms are shown for all types of hierarchical clustering algorithms.

6. Compare the clustering results of the K-means, Hierarchical, and DBSCAN in terms of time and quality of clustering.

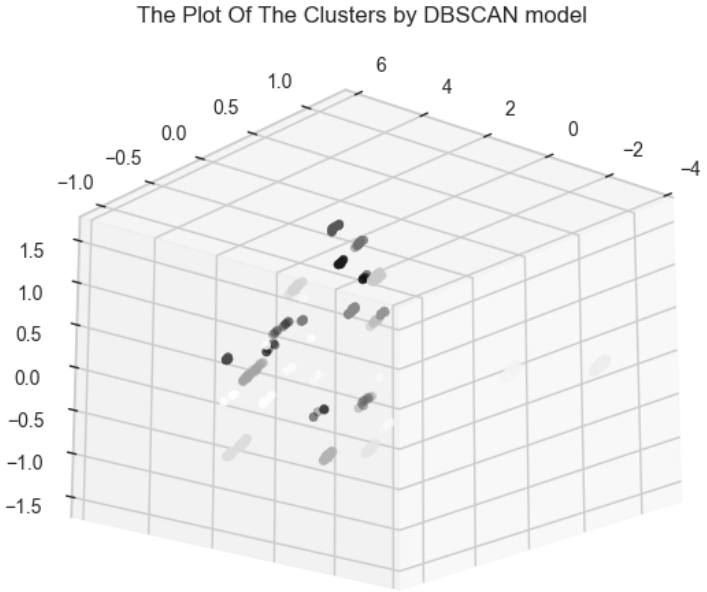
We take k=7 and run all algorithms to compare the results which are given below.

- ***Kmeans:***
  - Time taken to perform Kmeans: 0.15954065322875977 miliseconds
  - Silhoutte score for Kmeans: 0.76

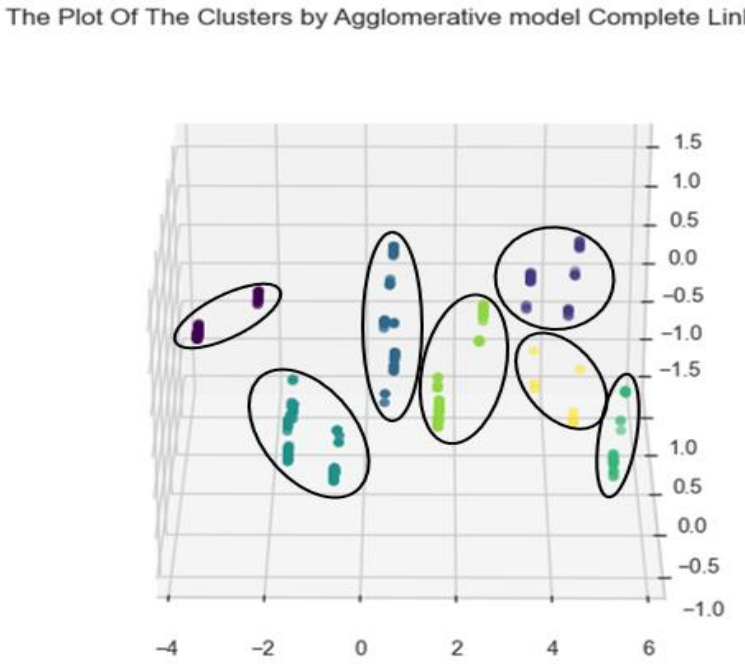
- SSE for Kmeans: 132.27185631133474



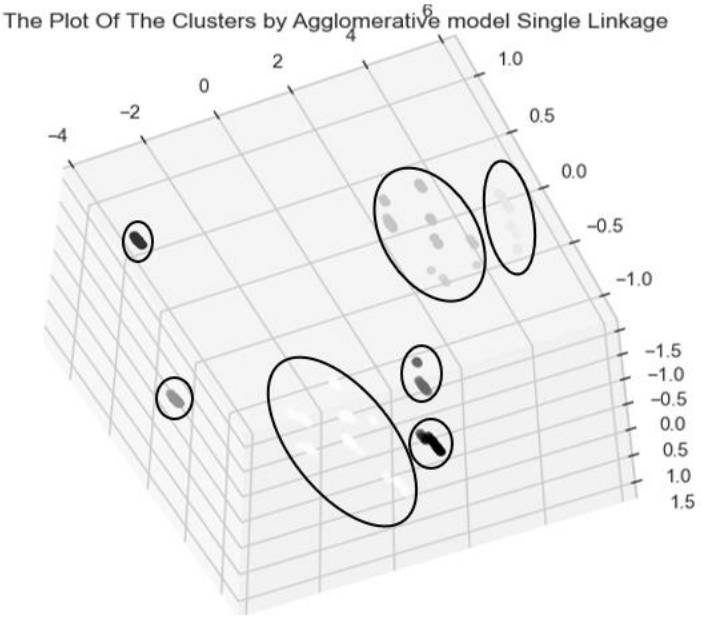
- ***DBSCAN*:**
  - DBSCAN with eps=0.1 and min points=5
  - Time taken to perform DBSCAN: 0.03637409210205078 miliseconds
  - Silhoutte score for DBSCAN: 0.89



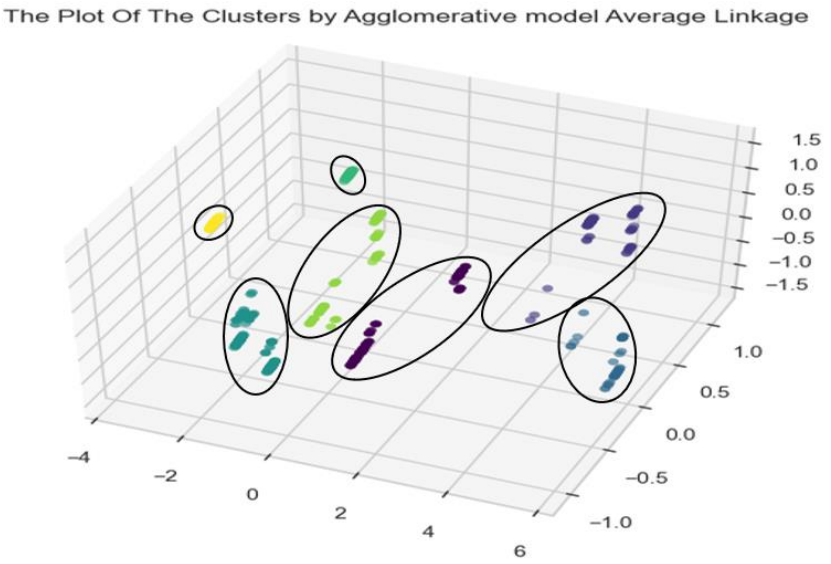
- ***Complete Linkage:***
    - Silhoutte score for Heirarchical clustering Complete Linkage: 0.65
    - Time taken to perform complete average: 0.027927398681640625 miliseconds



- ***Single Linkage:***
    - Silhoutte score for Heirarchical clustering Single Linkage: 0.68
    - Time taken to perform single linkage: 0.016986608505249023 miliseconds



- **Average Linkage:**
    - Time taken to perform average linkage: 0.010043859481811523 miliseconds.
    - Silhoutte score for Heirarchical clustering average Linkage: 0.76.



References:

1. **Clustering cant be applied on mixed datatype:** https://www.tomasbeuzen.com/post/clustering-mixed-data/
2. **Why do we need normalization in clustering:** Virmani, D., Taneja, S., & Malhotra, G. (2015). Normalization based K means Clustering Algorithm. arXiv preprint arXiv:1503.00900.