**IDM challenge 3 by Maaz Adnan**

**Business Understanding:**

Trying to cluster patients with same symptoms together

**Data Understanding:**

Normalized binary data

**Data Preparation:**

Removing data with low variance, and dropping highly correlated columns

**Model:**

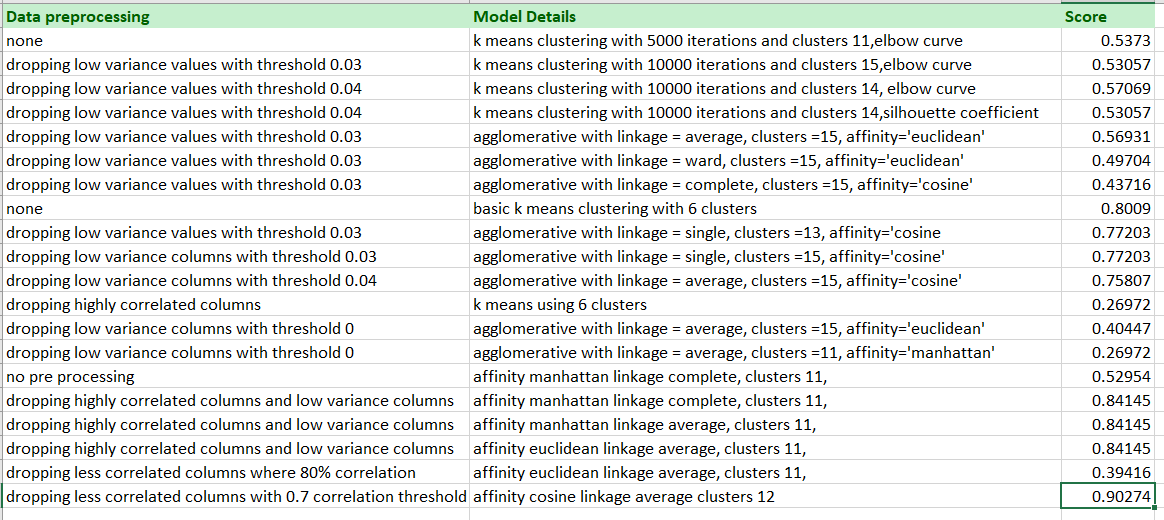
Clustering algorithms, k means and agglomerative

**Evaluation:**

Using Rand index

**Report**

* Started off with basic clustering simply using k means on the original data
* Drew elbow curve to figure out optimum number of clusters it was between 5 and 6 as both gave similar results
* used variance filter to filter out columns with 0 variance, this improved the score slightly
* tried different variance filter thresholds to test my answers, anything above 0.04 and the score dropped
* similarly dropping highly correlated columns, columns with 96% correlation and above were dropped
* reduced the value for correlated columns but score dropped significantly
* then used agglomerative clustering with different linkages, affinities and clusters to check best score
* used silhouette score and dendrogram to check for best number of cluster,11 was the optimal number of clusters
* “average” linkage worked better than other linkages and “manhattan or Euclidean” affinity worked best
* Silhouette coefficient score was essentially very inconsistent as entries with higher score on kaggle had a low silhouette coefficient score
* Dendrogram helped determine decent number of clusters for the agglomerative algorithm
* An interesting finding was that between the correlation threshold of 0.69 and 0.7 only 2 columns dropped but it increased the score massively from 0.43709 to 0.902

****

**The overall algorithm that worked best was agglomerative clustering with 12 clusters and affinity cosine with linkage average. Correlation threshold =0.70 and var threshold =0**

**Data pre pre processing and choosing different clusters was the main issue as small changes lead to massive drops in score**