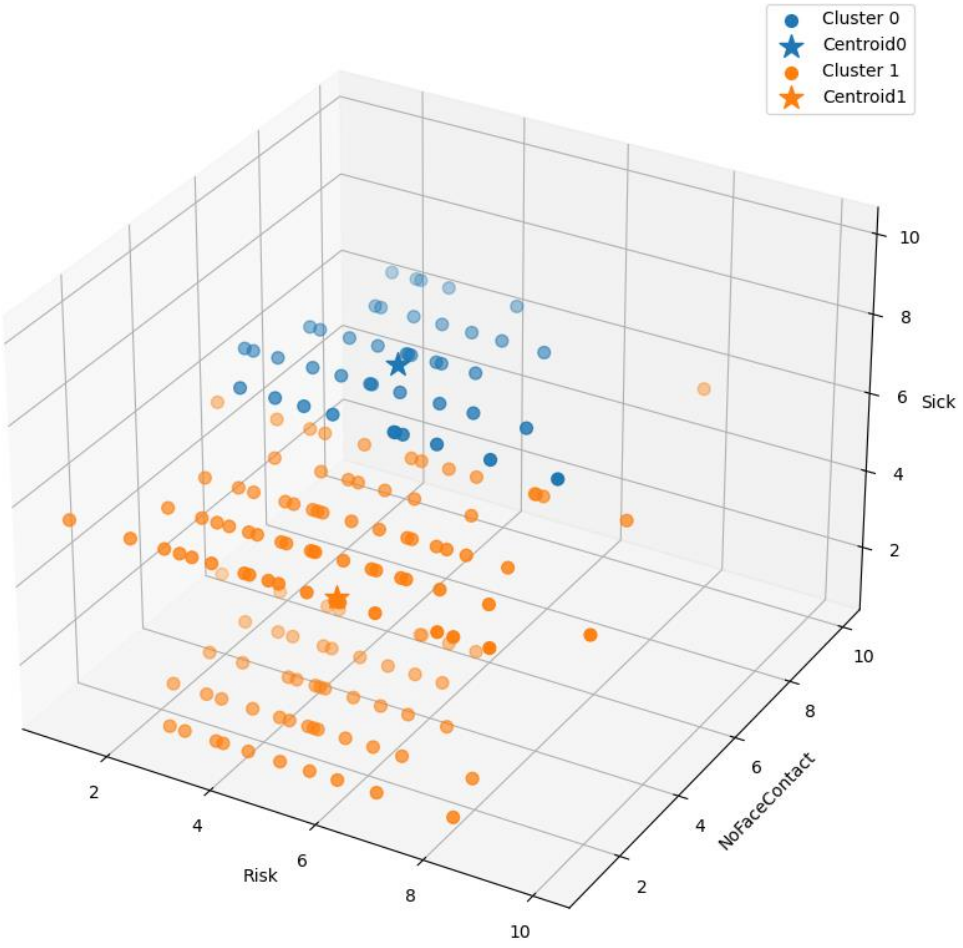
Assignment1

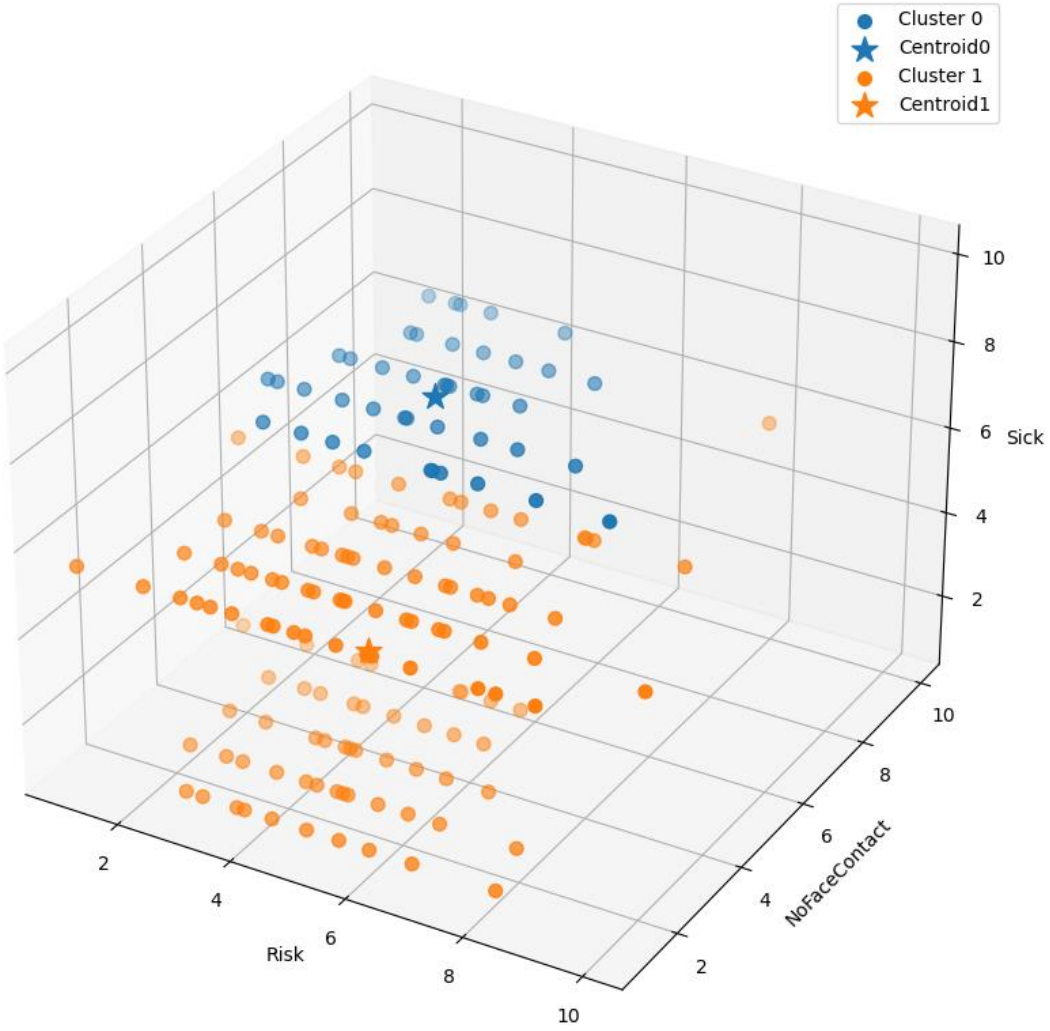
1.K-Means Clustering with different number of clusters:

a. Here I have written KMeans clustering algorithm code from scratch and clustered the data based on three features namely '**Risk', 'NoFaceContact' and 'Sick'**. In the 3D scatter plot they are visualised in three dimensions. Data points in same cluster are marked with similar color and centroids for them are highlighted. Number of clusters is **2**.
Below is the plot

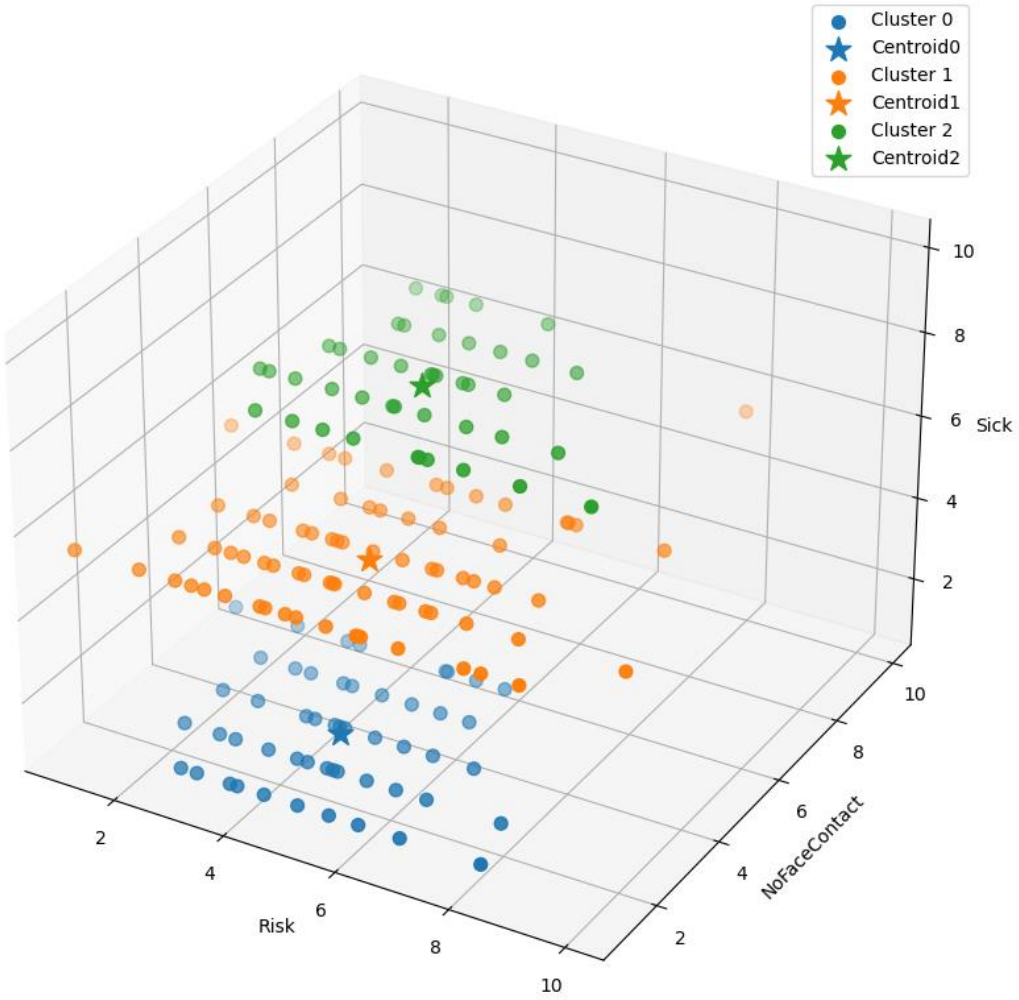


b. We are going to test it with different number of cluster from 2 to 10.

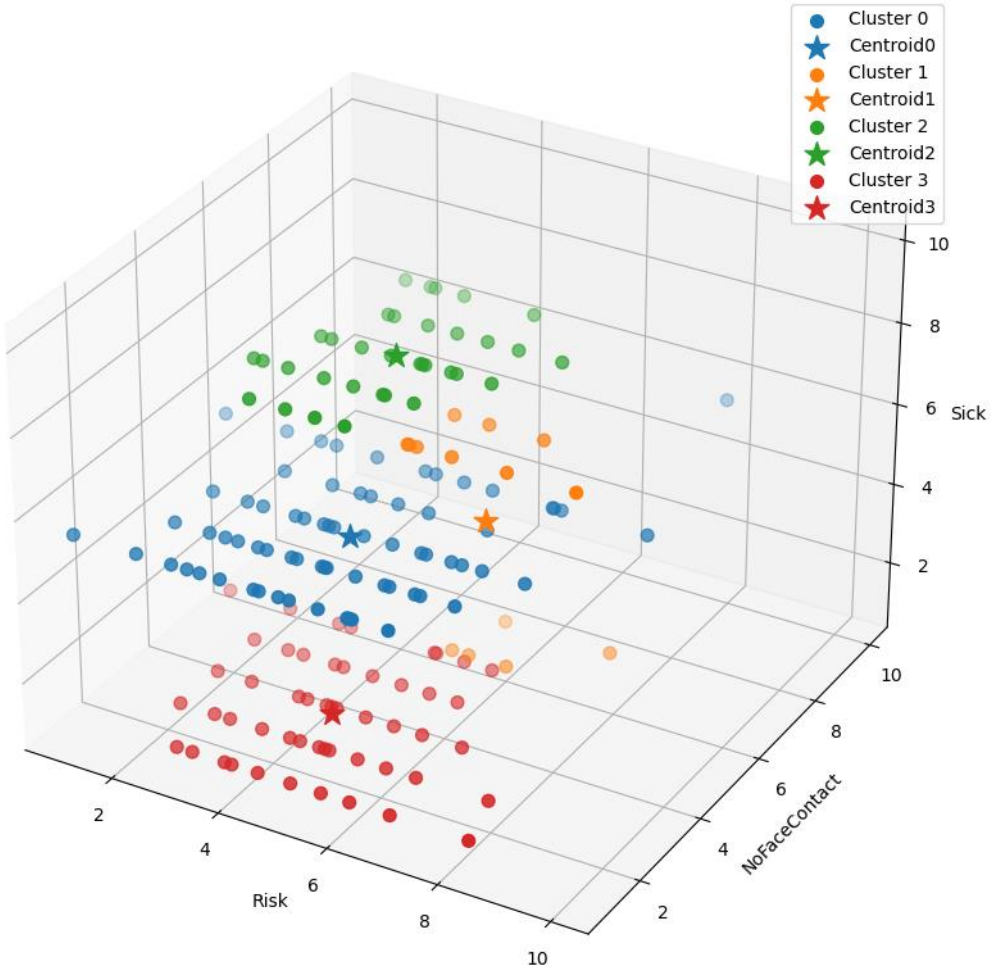
2 Clusters:


Cluster 0
Centroid0
Cluster 1
Centroid1
Risk
NoFaceContact
Sick
2
4
6
8
10

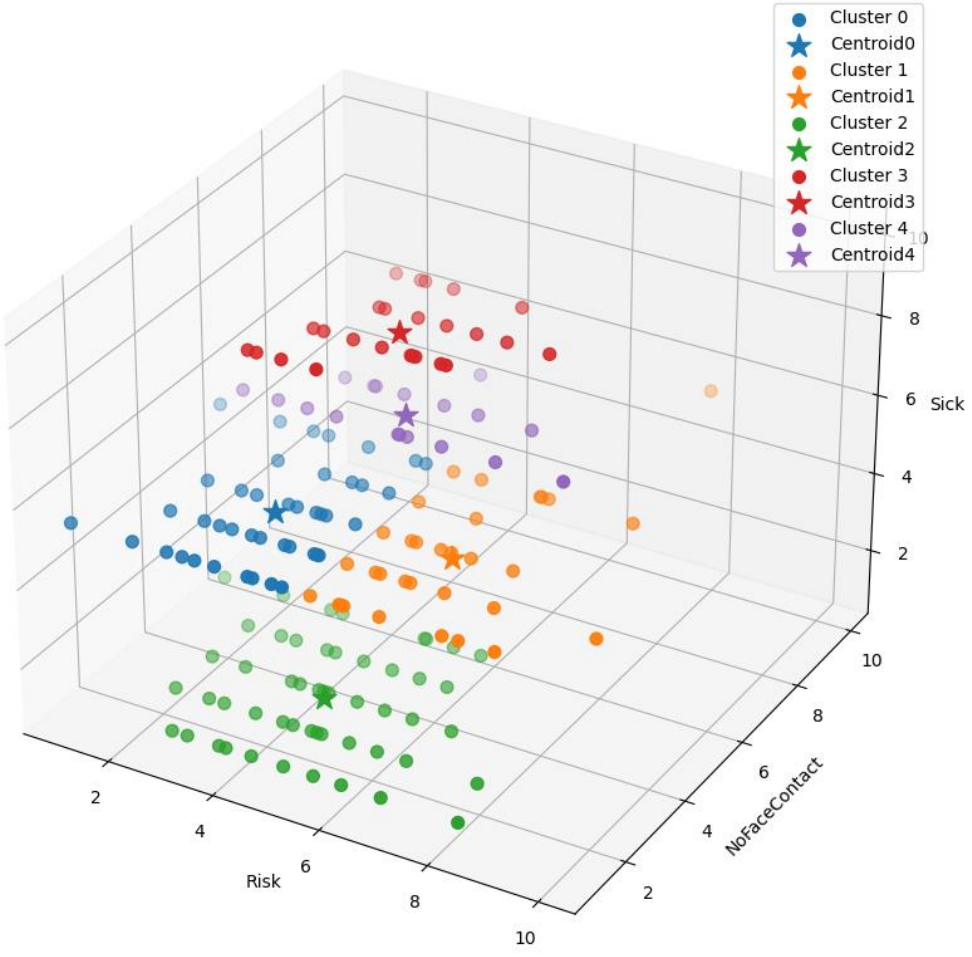
3 clusters:


Cluster 0
Centroid0
Cluster 1
Centroid1
Cluster 2
Centroid2
Risk
NoFaceContact
Sick
2
4
6
8
10

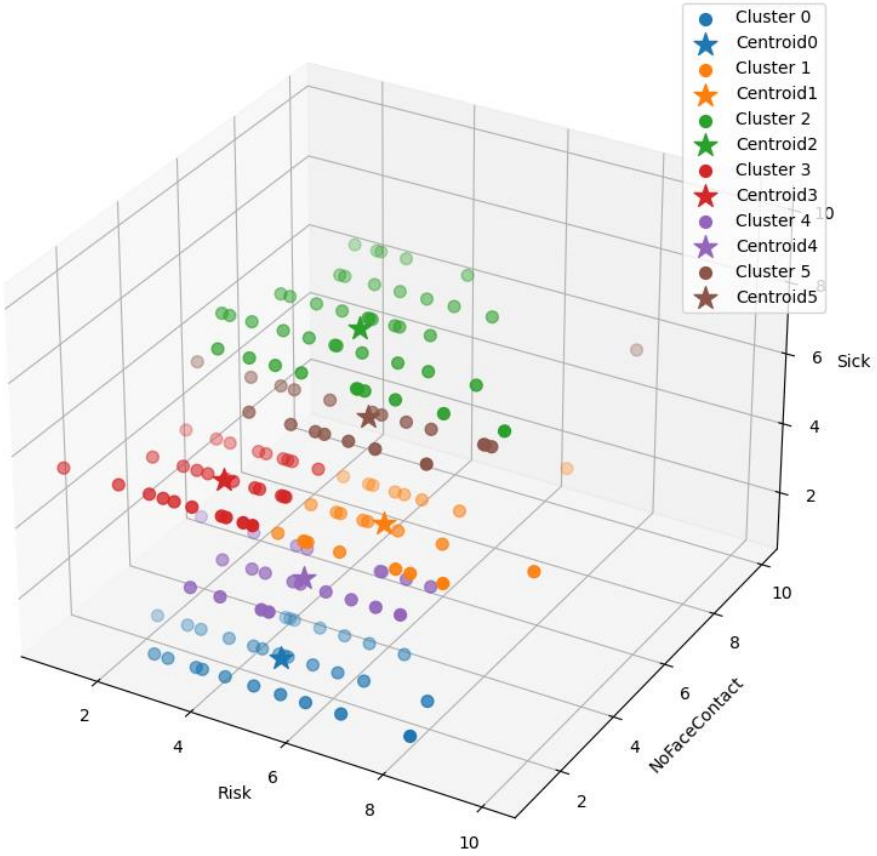
4 clusters:


Cluster 0
Centroid0
Cluster 1
Centroid1
Cluster 2
Centroid2
Cluster 3
Centroid3
Risk
NoFaceContact
Sick
2
4
6
8
10

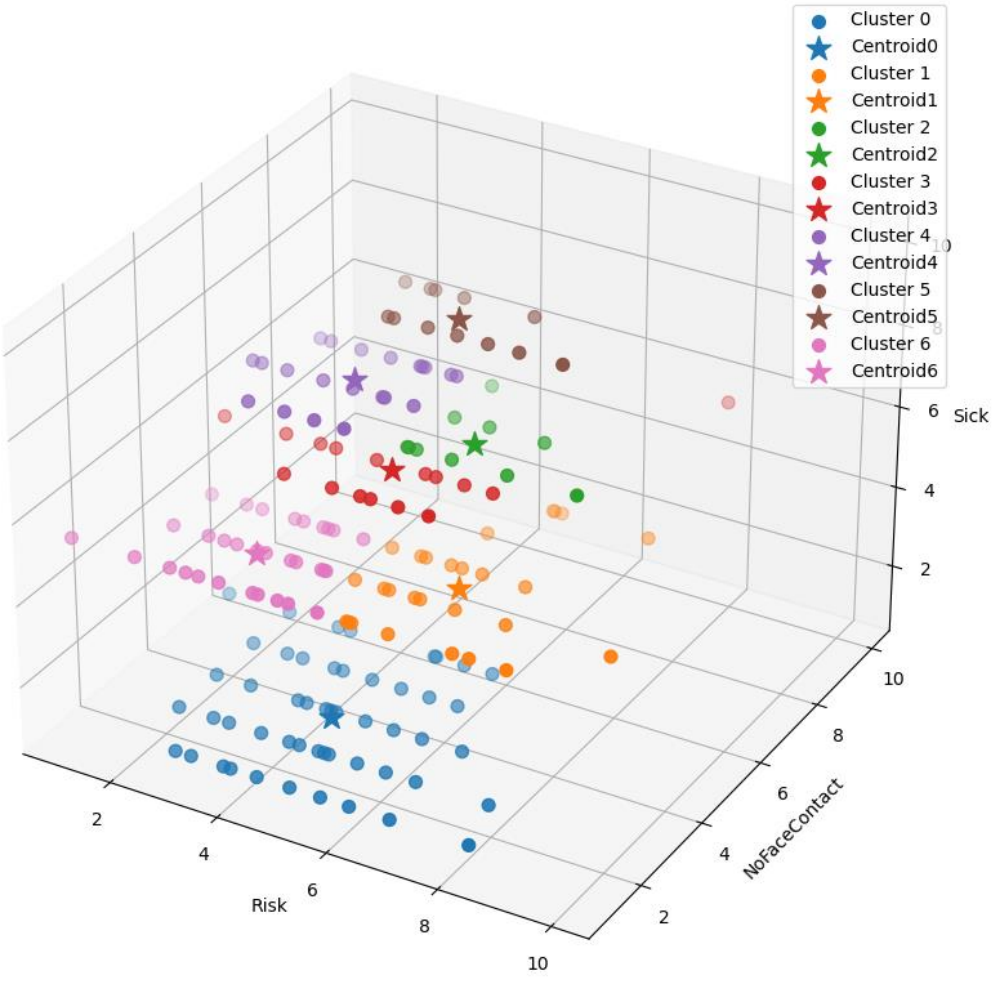
5 clusters:


Cluster 0
Centroid0
Cluster 1
Centroid1
Cluster 2
Centroid2
Cluster 3
Centroid3
Cluster 4
Centroid4
Sick
Risk
NoFaceContact
2
4
6
8
10

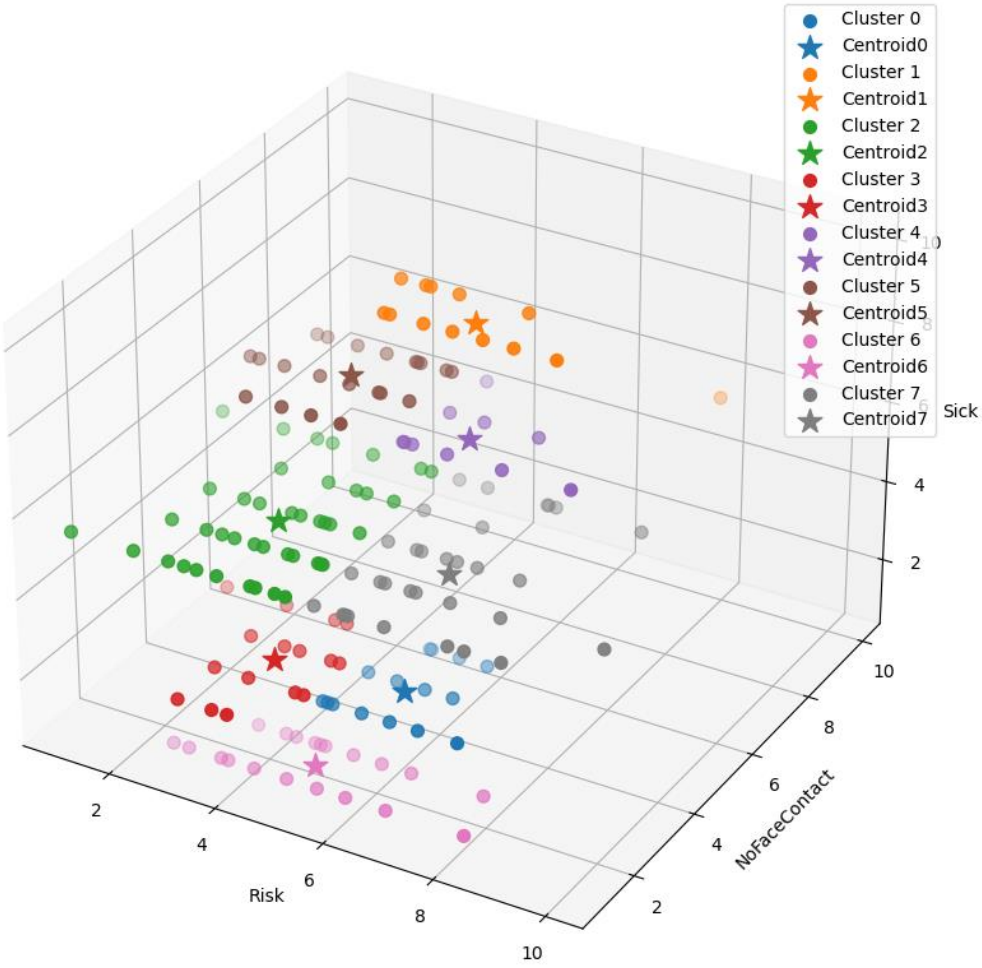
6 clusters:


Cluster 0
Centroid0
Cluster 1
Centroid1
Cluster 2
Centroid2
Cluster 3
Centroid3
Cluster 4
Centroid4
Cluster 5
Centroid5
Risk
NoFaceContact
Sick
2
4
6
8
10

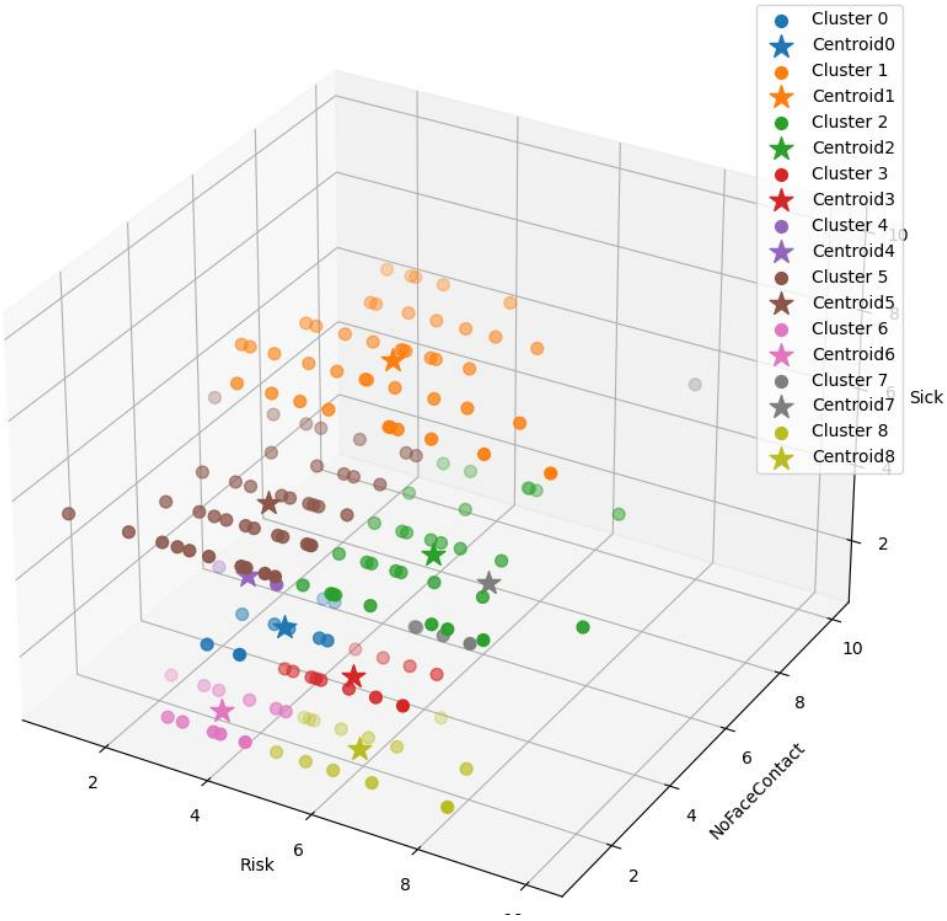
7 clusters:


Cluster 0
Centroid0
Cluster 1
Centroid1
Cluster 2
Centroid2
Cluster 3
Centroid3
Cluster 4
Centroid4
Cluster 5
Centroid5
Cluster 6
Centroid6
Risk
NoFaceContact
Sick
2
4
6
8
10

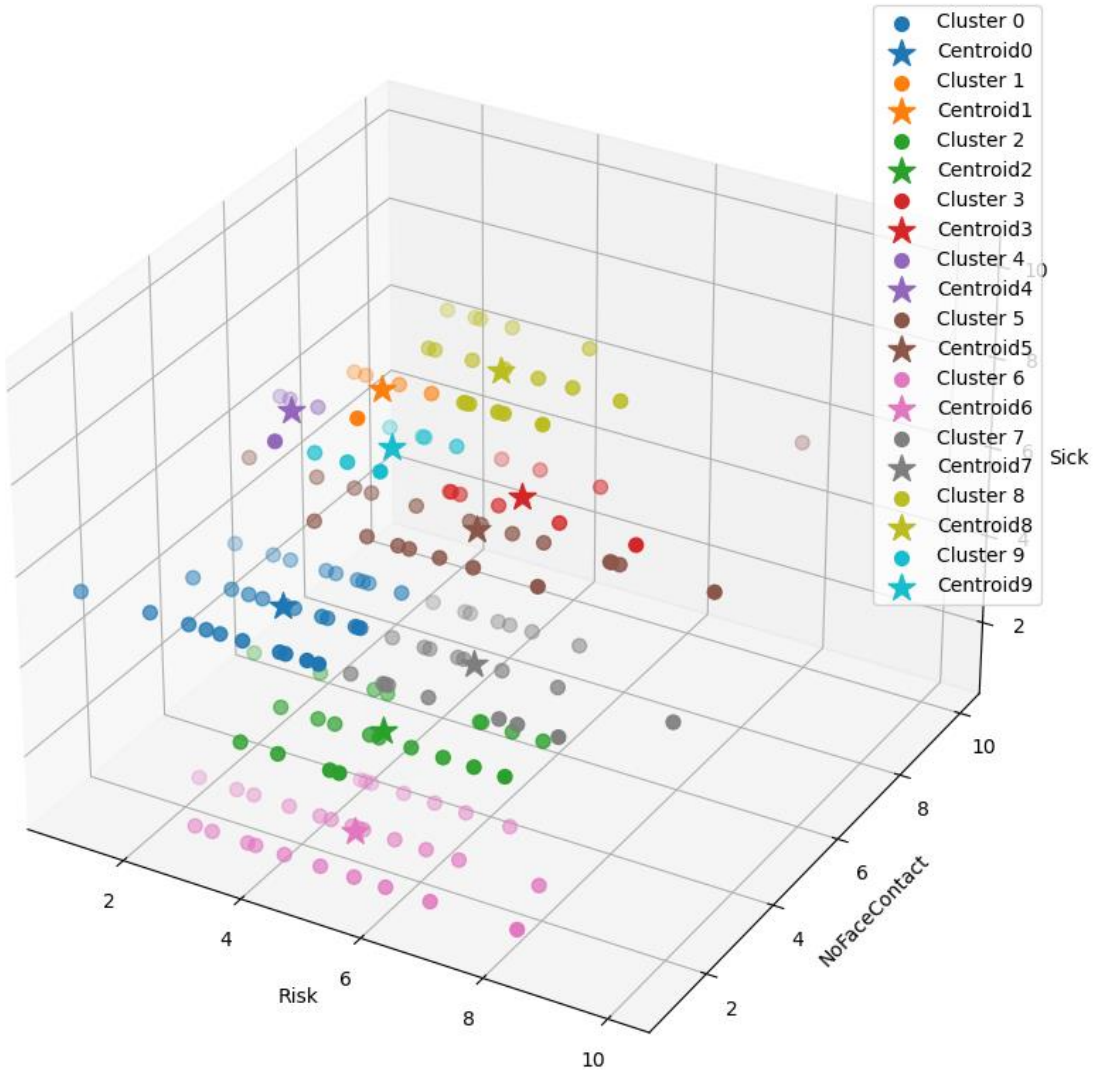
8 clusters:


Cluster 0
Centroid0
Cluster 1
Centroid1
Cluster 2
Centroid2
Cluster 3
Centroid3
Cluster 4
Centroid4
Cluster 5
Centroid5
Cluster 6
Centroid6
Cluster 7
Centroid7
Sick
Risk
NoFaceContact
2
4
6
8
10

9 clusters :


Cluster 0
Centroid0
Cluster 1
Centroid1
Cluster 2
Centroid2
Cluster 3
Centroid3
Cluster 4
Centroid4
Cluster 5
Centroid5
Cluster 6
Centroid6
Cluster 7
Centroid7
Cluster 8
Centroid8
Risk
NoFaceContact
Sick

10 clusters:



Best number of clusters seem to be **8** according to the plots we got, as the clusters are evenly spaced out in this plot.

c.

Below are the dunn index values from 2 to 10 clusters.

Dunn index for 2 clusters is 0.8813019488371086

Dunn index for 3 clusters is 0.9817567567567569

Dunn index for 4 clusters is 0.11930494867496924

Dunn index for 5 clusters is 0.08047789437246464

Dunn index for 6 clusters is 0.12375876628655316

Dunn index for 7 clusters is 0.32203389830508555

Dunn index for 8 clusters is 0.22403153011819388

Dunn index for 9 clusters is 0.5428571428571435

Dunn index for 10 clusters is 0.477401129943503

The best number of clusters as per dunn index measure is 3 as it has higher dunn index value. In 1b we thought of 8 as best number of clusters but it our observation is not correct after calculating dunn index values.

2. K-means clustering with different features

a.

In 1c, we found out the best number of clusters as 3. Now here we are adding another feature HndWshFreq and clustering.

Dunn index for 3 clusters using Risk, NoFaceContact, Sick, HndWshFreq is 0.9775969619675613

The clustering results were the same after adding another feature too.

b.

We are adding another feature HndWshQual and clustering.

Dunn index for 3 clusters using Risk, NoFaceContact, Sick, HndWshFreq, HndWshQual features is 0.484055350246581

The clustering results decreased as dunn index decreased after adding this feature.

3. Fuzzy C-means clustering

a.

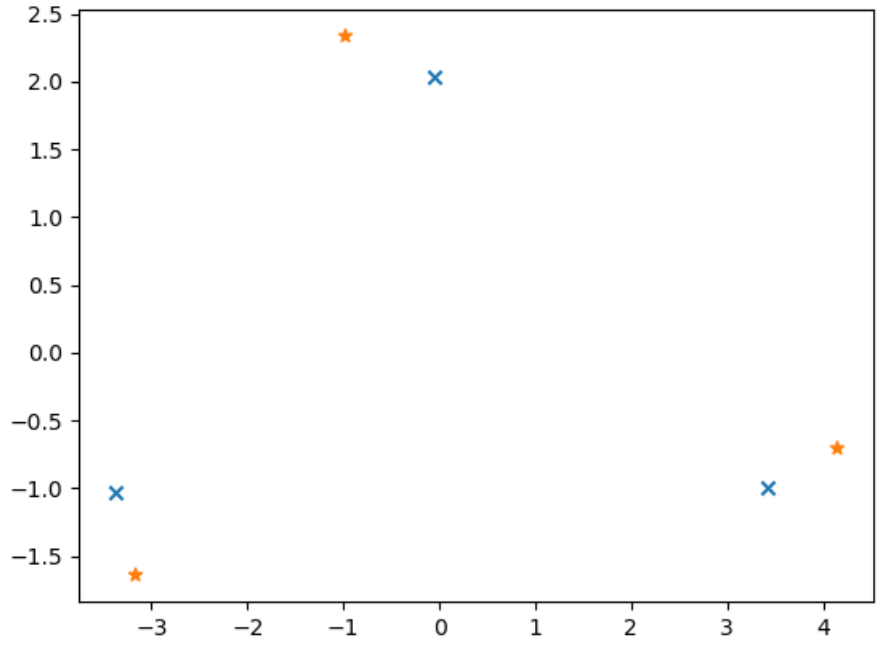
As we are going to use 4 features, we are converting them to 2 dimensional way to easily visualize using sklearn.decomposition module PCA. Below we have plotted centroids of fuzzy c and kmeans.

Fuzzy C centroids

	0	1	2
Risk	4.638221	5.805574	5.019491
NoFaceContact	2.83898	2.753346	2.736405
Sick	2.035797	8.734541	5.612728
HndWshFreq	5.132324	4.813682	8.00712

K Means Centroids

	0	1	2
Risk	4.734995	4.762666	6.217098
NoFaceContact	2.707353	2.855769	2.751786
Sick	3.064706	4.538462	9.935714
HndWshFreq	3.594118	7.884615	5.821429



The clusters are almost similar as we can see in the plot the centroids for kmeans and fuzzy c means are almost similar and are evenly spaced.

b.

Dunn index for fuzzy cmeans using 3 clusters and Risk, NoFaceContact, Sick, HndWshFreq features is 0.54678934.

Both fuzzy c means and kmeans have similar results but for fuzzy c means the best clusters when we change the epsilon value to minimum but we have to trade of performance for it.

c.

We ran fuzzy c clustering over various and below are the dunn index values

unn index for fuzzy cmeans using 3 clusters and Risk, NoFaceContact, Sick, HndWshFreq,Vaccin features is 0.1525174941322433

Dunn index for fuzzy cmeans using 3 clusters and Risk, NoFaceContact, Sick, HndWshQual,HndWshFreq features is 0.07540175907384514

Dunn index for fuzzy cmeans using 3 clusters and Risk, NoFaceContact, Sick, SociDist,HndWshFreq features is 0.05708862075526551

Dunn index for fuzzy cmeans using 3 clusters and Risk, NoFaceContact, Sick, PersnDist,HndWshFreq features is 0.33022503181807467

Dunn index for fuzzy cmeans using 3 clusters and Risk, NoFaceContact, Sick, HandSanit,HndWshFreq features is 0.18194736653148122

Dunn index for fuzzy cmeans using 3 clusters and Risk, NoFaceContact, Sick, Complications,HndWshFreq features is 0.09315240310796649

No other features improved the clustering results.