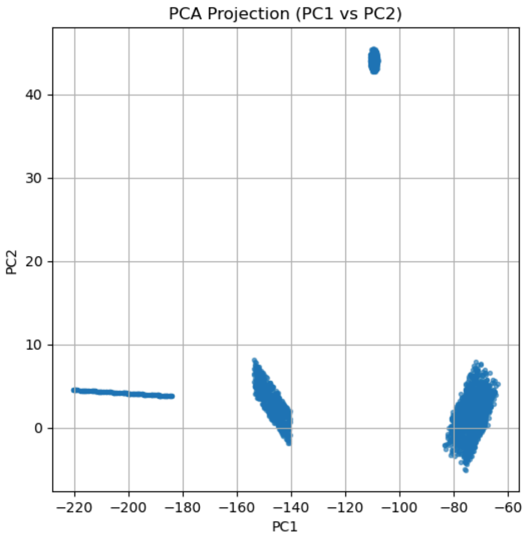
Joyce Goh

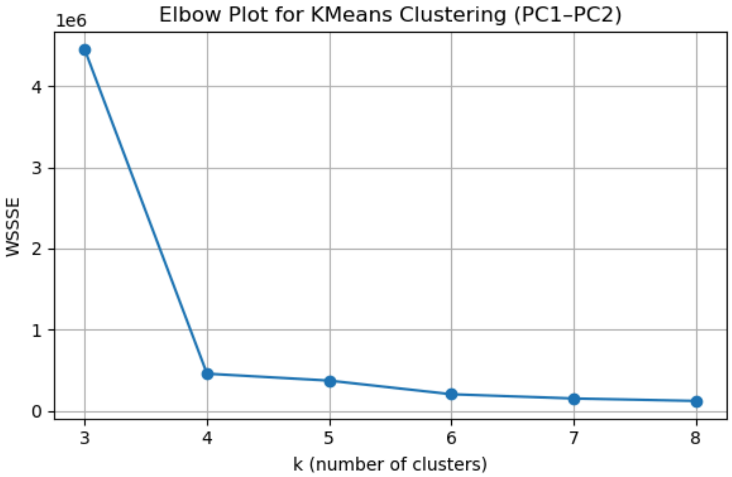
Assignment 4 - Details and Results

1) Basic EDA and general PCA
   a) The output of the cumulative variance is: [0.7947434 0.9912457 0.99495722 0.99842398 0.99992542 1.   ]
      Since PC1 and PC2 account for 99.1% of the variance, this indicates that most of the shapes are 2 dimensional



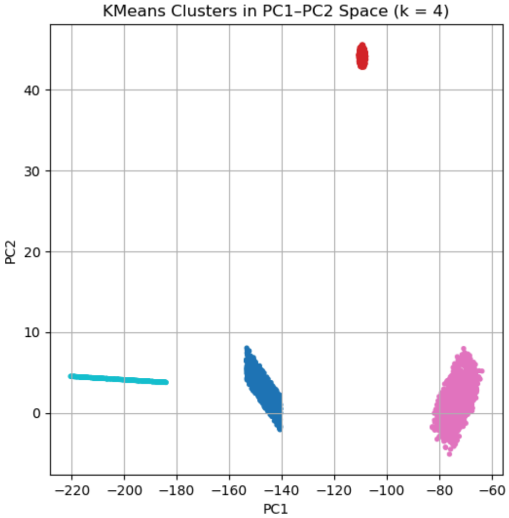
      This is the original PCA, with all the points. From this graph, we can see that there are 4 clear clusters, but we will investigate further.

2) K-means clustering algorithm
   a) Based on the above PCA, it seems like k = 4 will be the ideal k value. However, we test k = 3 to 8 and plot an elbow plot to determine the ideal value of k.

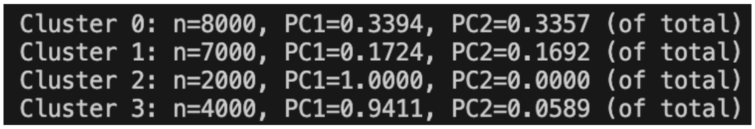


Based on the above elbow plot generated, k = 4 is indeed the ideal k value for the clustering algorithm.

We re-plot the PCA, using k-means with k = 4. Below is the plot obtained:

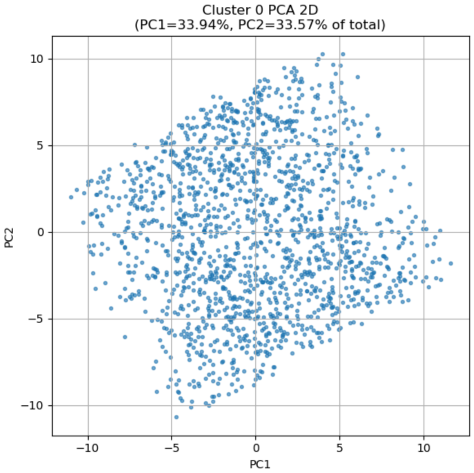


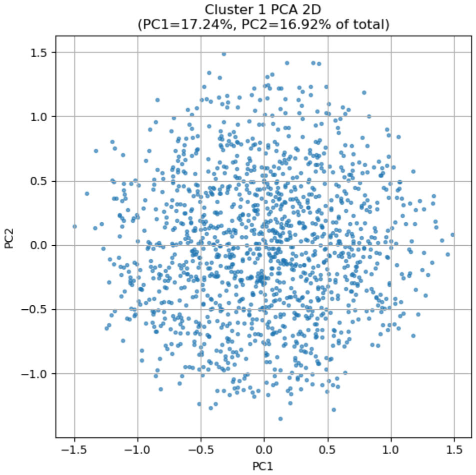
3) Next, we split the 6D data into each of the 4 clusters using a cluster ID, and plot a PCA in 2D and 3D (for applicable clusters) for each of the 4 clusters
   a) Below is the results of the 2D PCA for each cluster, along with the fraction of the total variance explained for PC1 and PC2:

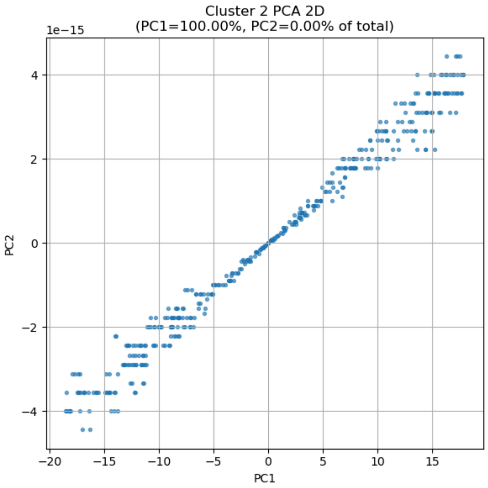


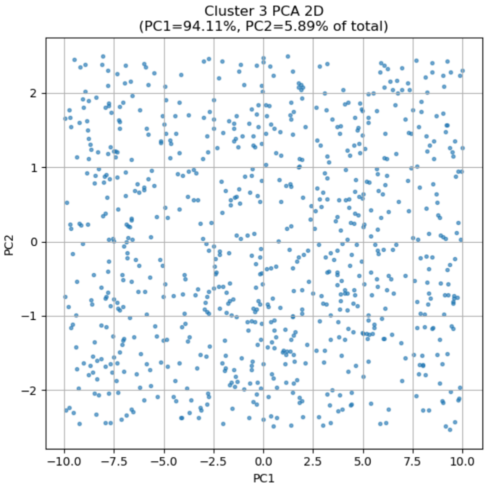
```
Cluster 0: n=8000, PC1=0.3394, PC2=0.3357 (of total)
Cluster 1: n=7000, PC1=0.1724, PC2=0.1692 (of total)
Cluster 2: n=2000, PC1=1.0000, PC2=0.0000 (of total)
Cluster 3: n=4000, PC1=0.9411, PC2=0.0589 (of total)
```

And the PCA graphs (in 2D) for each cluster:

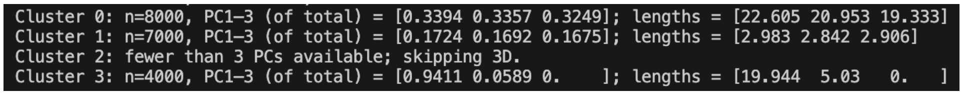






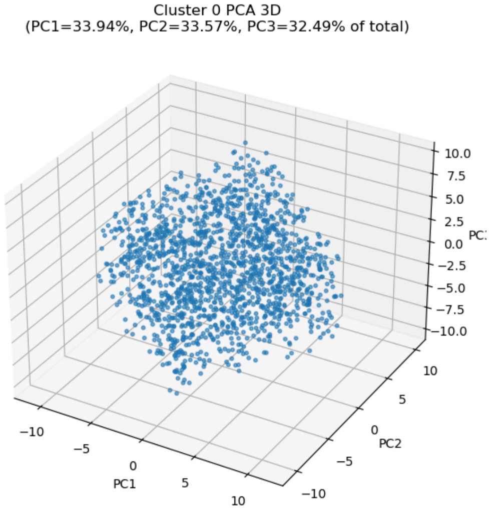


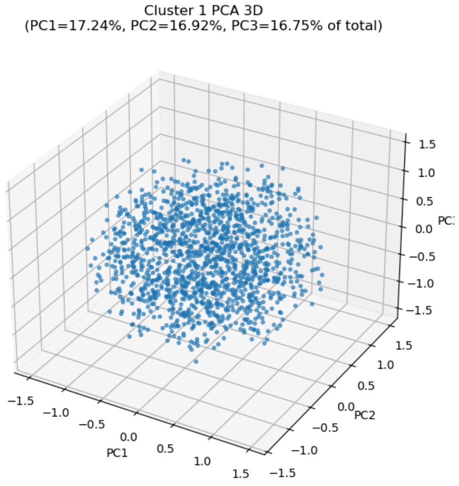
b) Below is the results of the 3D PCA for each cluster, along with the fraction of the total variance explained for PC1-3:

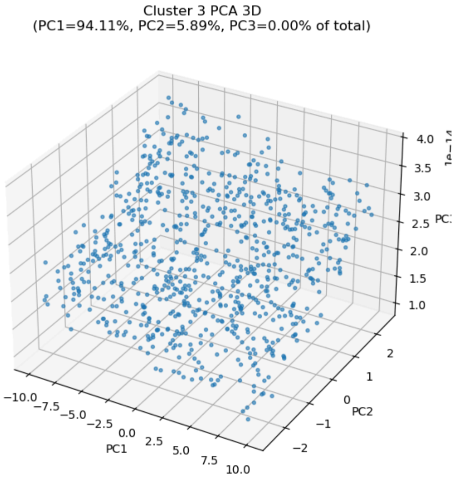


```
Cluster 0: n=8000, PC1–3 (of total) = [0.3394 0.3357 0.3249]; lengths = [22.605 20.953 19.333]
Cluster 1: n=7000, PC1–3 (of total) = [0.1724 0.1692 0.1675]; lengths = [2.983 2.842 2.906]
Cluster 2: fewer than 3 PCs available; skipping 3D.
Cluster 3: n=4000, PC1–3 (of total) = [0.9411 0.0589 0.    ]; lengths = [19.944  5.03   0.   ]
```

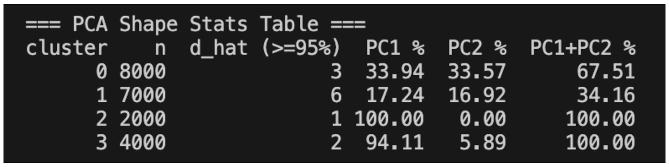
And the PCA graphs (in 3D) for each cluster:







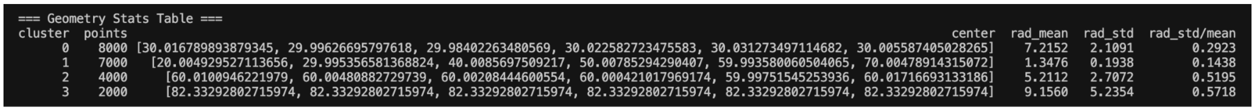
c) Here we can see the number of PCs that explain >95% of the variance for each cluster



```
=== PCA Shape Stats Table ===
cluster    n  d_hat (>=95%)  PC1 %  PC2 %  PC1+PC2 %
      0 8000              3  33.94  33.57      67.51
      1 7000              6  17.24  16.92      34.16
      2 2000              1 100.00   0.00     100.00
      3 4000              2  94.11   5.89     100.00
```

d) From these results, we can deduce that:
   - Cluster 0: 3D cube
   - Cluster 1: 6D sphere
   - Cluster 2: 1D rod
   - Cluster 3: Nearly 1D rod

4) Per-cluster geometry core statistics
   a) I then found the center of each cluster (in 6D), the average radius, the radius standard deviation, and the ratio of the average radius/radius to determine where in the 6D realm each cluster resided, along with the distribution of their data points. Below is the table of results:



```
=== Geometry Stats Table ===
cluster  points                                                                                                                    center  rad_mean  rad_std  rad_std/mean
      0    8000 [30.016789893879345, 29.99626695797618, 29.98402263480569, 30.022582723475583, 30.031273497114682, 30.005587405028265]    7.2152   2.1091        0.2923
      1    7000   [20.004929527113656, 29.995356581368824, 40.0085697509217, 50.00785294290407, 59.993580060504065, 70.00478914315072]    1.3476   0.1938        0.1438
      2    4000     [60.0100946221979, 60.00480882729739, 60.00208444600554, 60.000421017969174, 59.99751545253936, 60.01716693133186]    5.2112   2.7072        0.5195
      3    2000     [82.33292802715974, 82.33292802715974, 82.33292802715974, 82.33292802715974, 82.33292802715974, 82.33292802715974]    9.1560   5.2354        0.5718
```

5) Conclusion
   a) Cluster 0 is a 3D cube, centered at [30, 30, 30, 30, 30, 30]. Since it is a 3D object in a 6D space, the object lacks a unique orientation. It has 8000 points, with a mean radius of 7.2, and a radius standard deviation of 2.11, indicating a moderately compact distribution with a small spread around the center.
   b) Cluster 1 is a 6D sphere, centered at [20, 30, 40, 50, 60, 70]. Since it is a 6D object in a 6D space, there is no orientation. It has 7000 points, with a mean

radius of 1.35, and a radius standard deviation of 0.19. This signifies a very tight and uniform distribution of points around the center.

c) Cluster 2 is a 1D rod, centered at [60, 60, 60, 60, 60, 60], and is oriented along its first principal component, which explains 100% of the variance. It consists of 4,000 points with a mean radius of 5.21 and radius standard deviation of 2.70. The relatively large variance around the mean radius indicates that points are somewhat dispersed along the rod, consistent with its linear, elongated geometry visible in the 2D PCA projection.

d) Cluster 3 forms an elongated 1D line with a slight width, centered at [82, 82, 82, 82, 82, 82]. It is strongly oriented along PC1, which explains 94% of the variance, while PC2 contributes 6%, giving it a subtle planar spread. The cluster contains 2,000 points, with a mean radius of 9.16 and radius standard deviation of 5.23, indicating a more diffuse distribution.