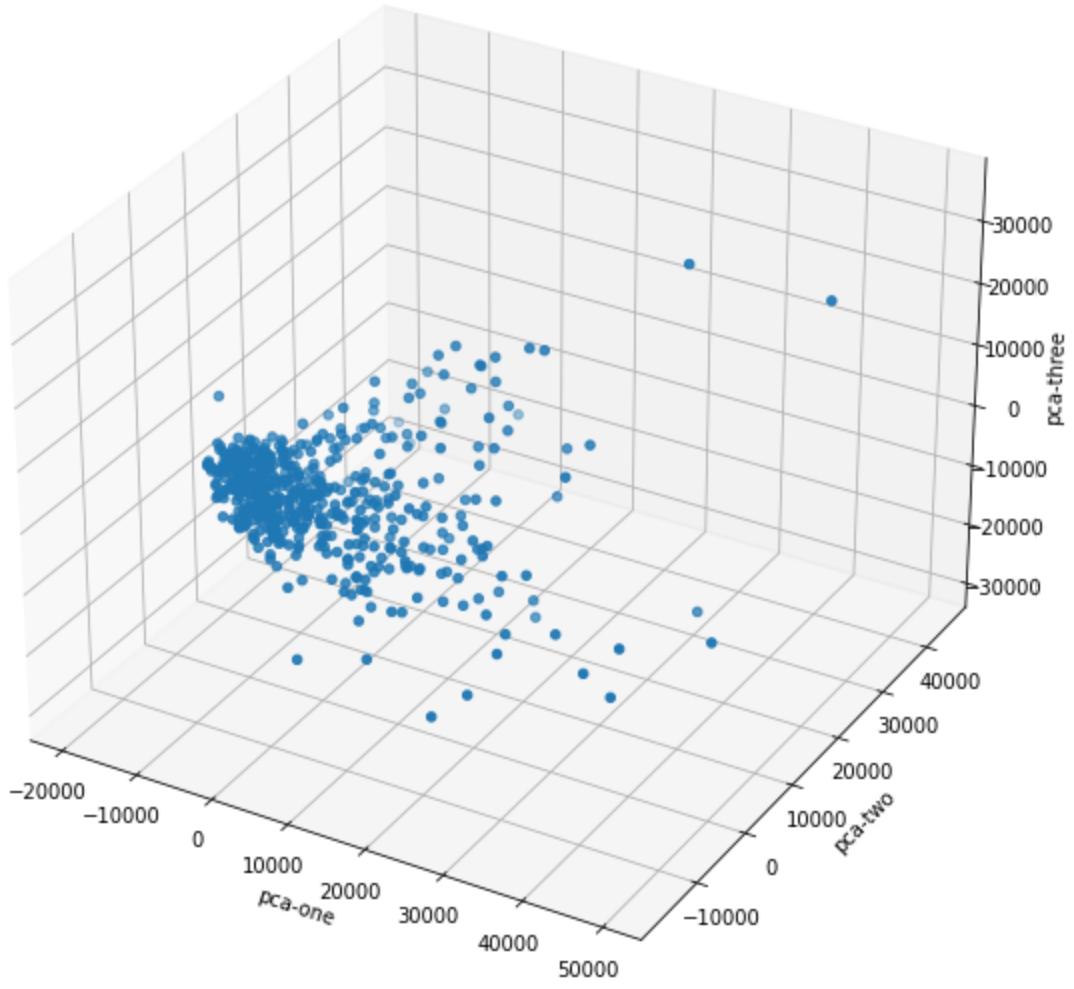
Homework 2

Class	33750
Created	@Oct 14, 2020 7:27 PM
Type	Write-up

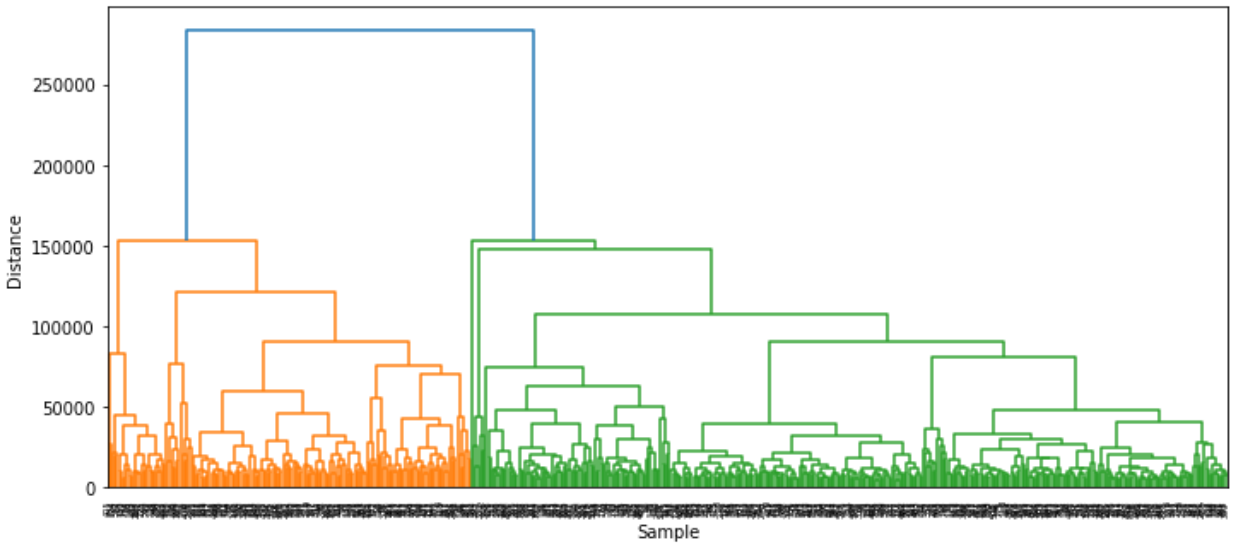
For the RNAseq data, there are three types of files available for each sample — read counts, FPKM, and FPKM-UQ. FPKM are normalized reads. There is some controversy over using FPKM values for expression analysis; however, for the sake of this exploratory assignment on clustering, I decided to proceed with this data.

I concatenated all the files into one dataframe. Then, I preprocessed the data by removing genes with a value of 0.0 across all samples. In addition, I removed genes which were expressed in less than or equal to 100 (arbitrarily chosen) of the samples. The data went from 546 samples and 60,483 genes to 546 samples and 38,364 genes.

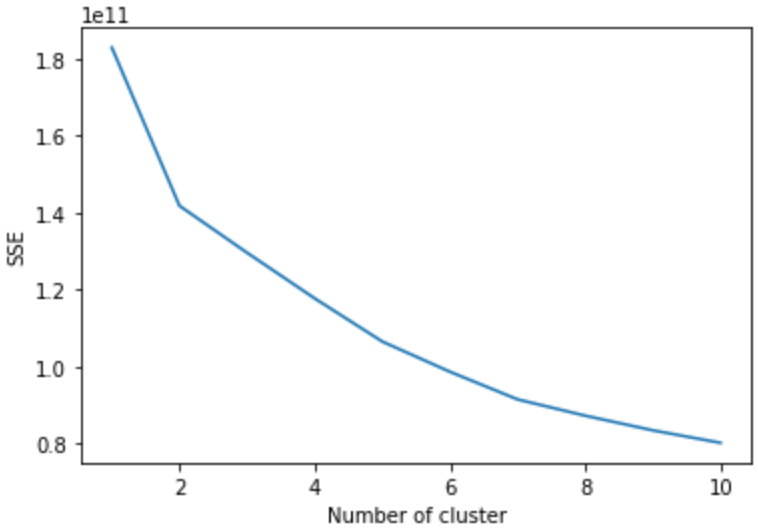
For the clustering, I used multiple algorithms to explore the data. First I plotted the samples using the gene expression as the features and visualized the data using a **PCA** plot with three PCs.



By eye, I could see two or three potential clusters. To test whether this is true, I then performed **Ward Hierarchical Clustering,** which does not require the number of clusters as an argument.

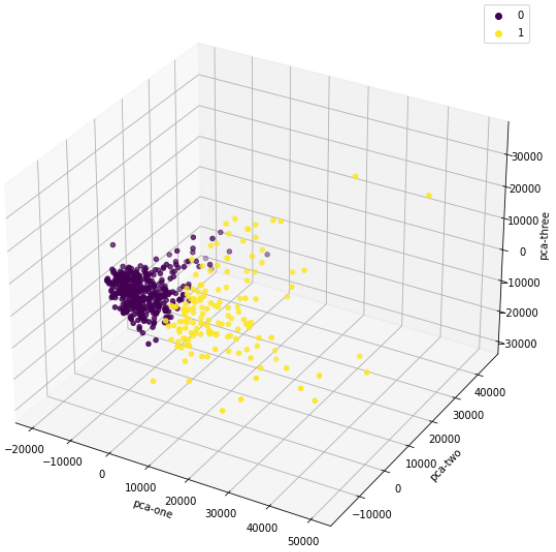


Here, I can see that the hierarchical clustering algorithm found two clusters from our gene expression data. In order to validate this clustering, I performed KMeans with a range of clusters (1 to 10) and evaluated the performance with multiple metrics. I used an elbow plot of the sum of squared errors to determine at which *k* we start to have diminishing returns.

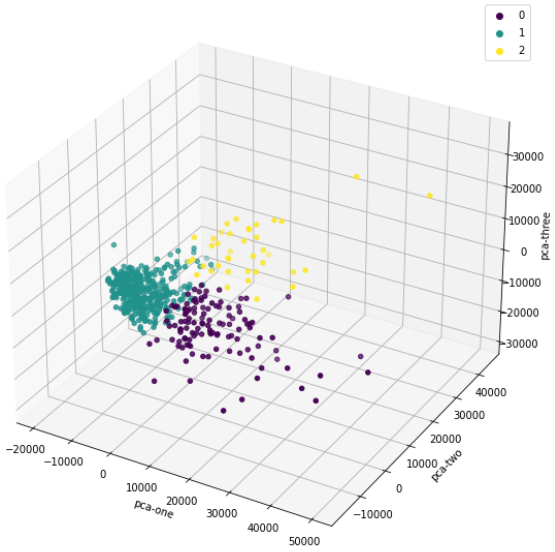


Here, we see that $k = 2$ is where there is an elbow in the plot. Finally, I plotted and labeled clusters for several values of k to visualize what the results look like

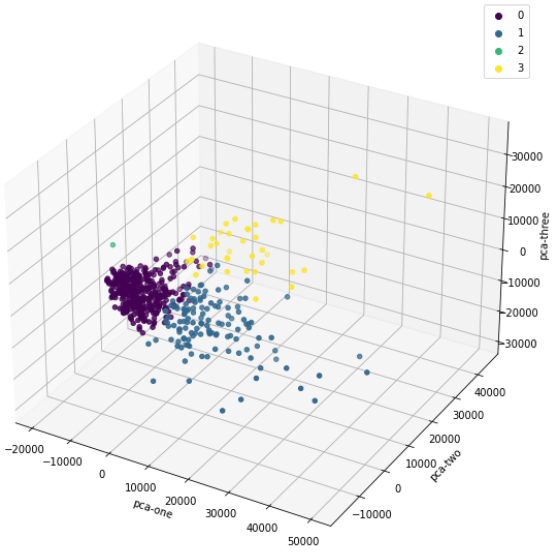
and evaluated each **KMeans** clustering with the following metrics: Silhouette score and Davies Bouldin Score. The results for $k = 2, 3, 4, 5$ are below:

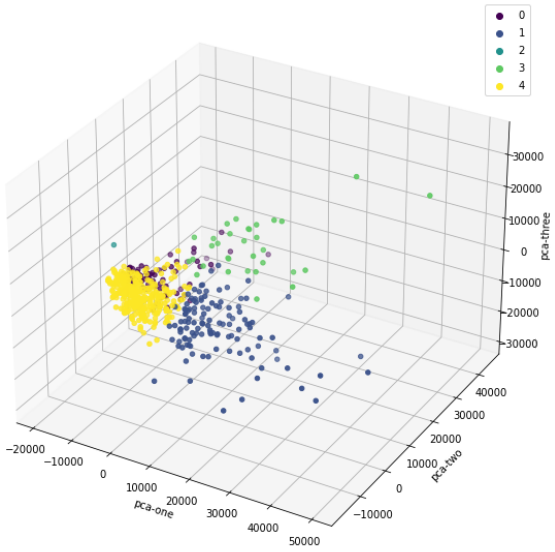



```
Silhouette Coefficient: 0.311
Davies Bouldin Score: 1.532
```




```
Silhouette Coefficient: 0.311
Davies Bouldin Score: 1.527
```





```
Silhouette Coefficient: 0.308
Davies Bouldin Score: 1.155
```

```
Silhouette Coefficient: 0.240
Davies Bouldin Score: 1.323
```

▼ Definitions:

Silhouette Coefficient: A higher Silhouette Coefficient score relates to a model with better defined clusters.
- s = (b - a)/(max(a, b))
- a: The mean distance between a sample and all other points in the same class.
- b: The mean distance between a sample and all other points in the next nearest cluster.

Davies Bouldin Score: A lower Davies-Bouldin index relates to a model with better separation between the clusters. This index signifies the average ‘similarity’ between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.

Interestingly, when analyzing the clusters in more detail we see that 2 clusters is not our only potential option. Two and three clusters both appear viable, both visually and using our evaluation metrics, and our plot with four clusters seems to show a potential outlier, where cluster 2 only has one datapoint.

Although much of our cursory analysis prior to KMeans seemed to indicate that the optimal number of clusters is 2, I think that based on our more detailed KMeans analysis, there is a potential for three clusters of samples based on the gene expression. This means that there are possibly three subtypes of HNSC present in the population from TCGA. We come to this conclusion because there are three clusters of samples based on the gene expression, indicating there are three gene expression profiles for HNSC.