Anushka Sinha

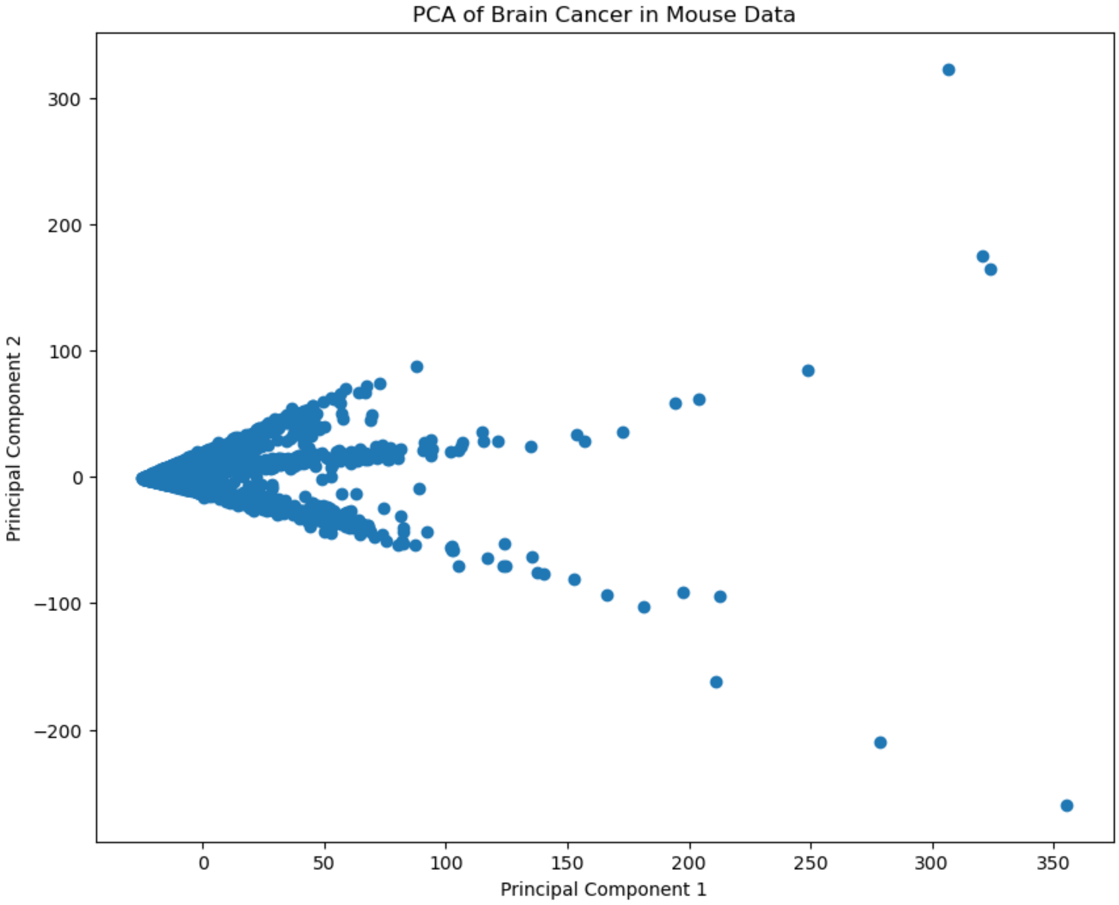
02718 HW1

Sept 27, 2023

1. **The first thing you decide to look at is whether the cell's gene expression can be used to separate the cells into the different cell types.**
   a. **By looking at the brain_metadata file, how many different cell types do you observe?**

      Answer: There are 7 different cell types.

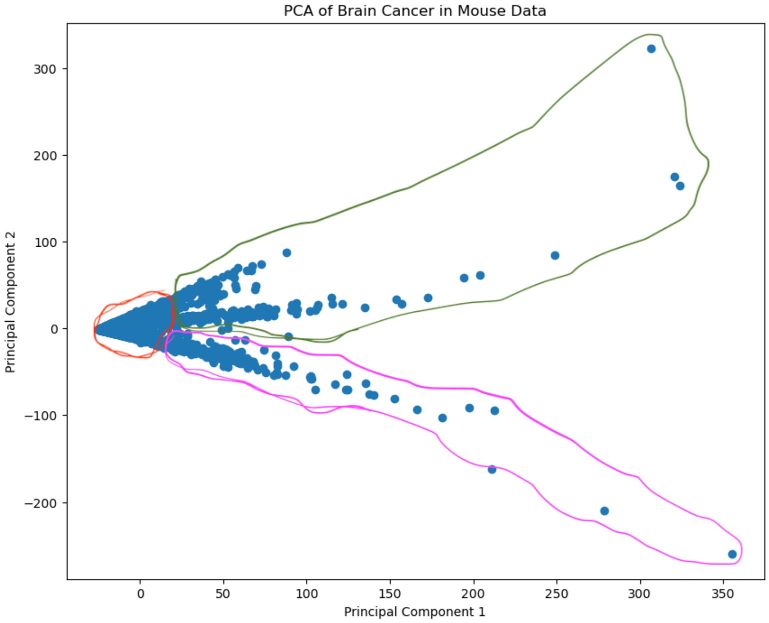
   b. **Provide the PCA plot.**



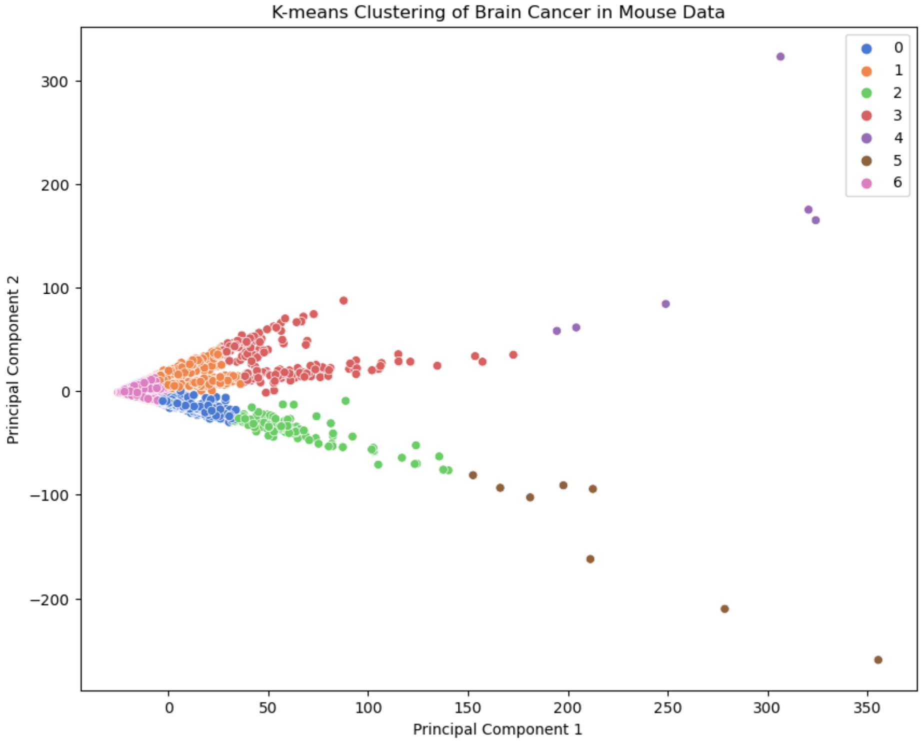
   c. **We will now perform k-means clustering on our gene expression data. By looking at your PCA plot, hypothesize how many clusters (K) of the cell types**

are present in the dataset and annotate your plot from question 1b to demonstrate these K clusters.

Answer: I hypothesize that there will be three different clusters based on the proximity or grouping of the data points on the PCA plot, therefore K = 3.



d. Provide the K-means plot.

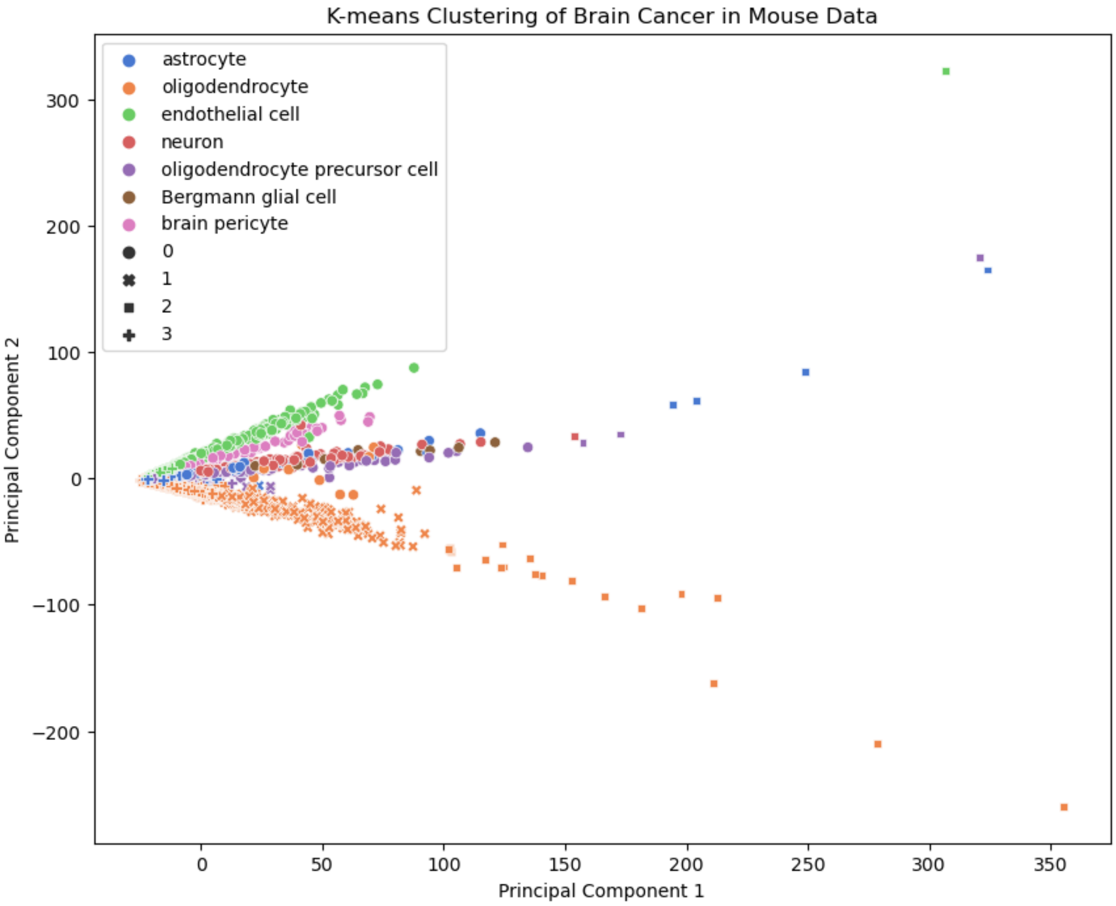


2. **Although we were able to perform clustering on our data, we want to evaluate our clusters.**
    a. **To start, let's perform some internal evaluations. Compute the Silhouette coefficient and explain what the value you got means in regard to the clustering.**

    Answer: The silhouette score for this K-means clustering is 0.51. Since a Silhouette score closer to 1 indicates perfect performance of a clustering method, our score demonstrates that the k-means clustering did a reasonable job at clustering our data points. This means that the data points in the same cluster tend to be more similar than the data points in other clusters. It also indicates that intra-cluster distance is slightly higher than the inter-cluster distance. However, there is still room for more meaningful separation of data points since there is still some overlap and ambiguity with clustering of data points.

    b. **Annotate each cell by color-coding it in the PCA plot to represent what the true "cell_ontology_class" of the cell is according to the metadata file, and sue a different annotation method (e.g. different symbols) on top of this plot**

to represent what cluster that cell represents.



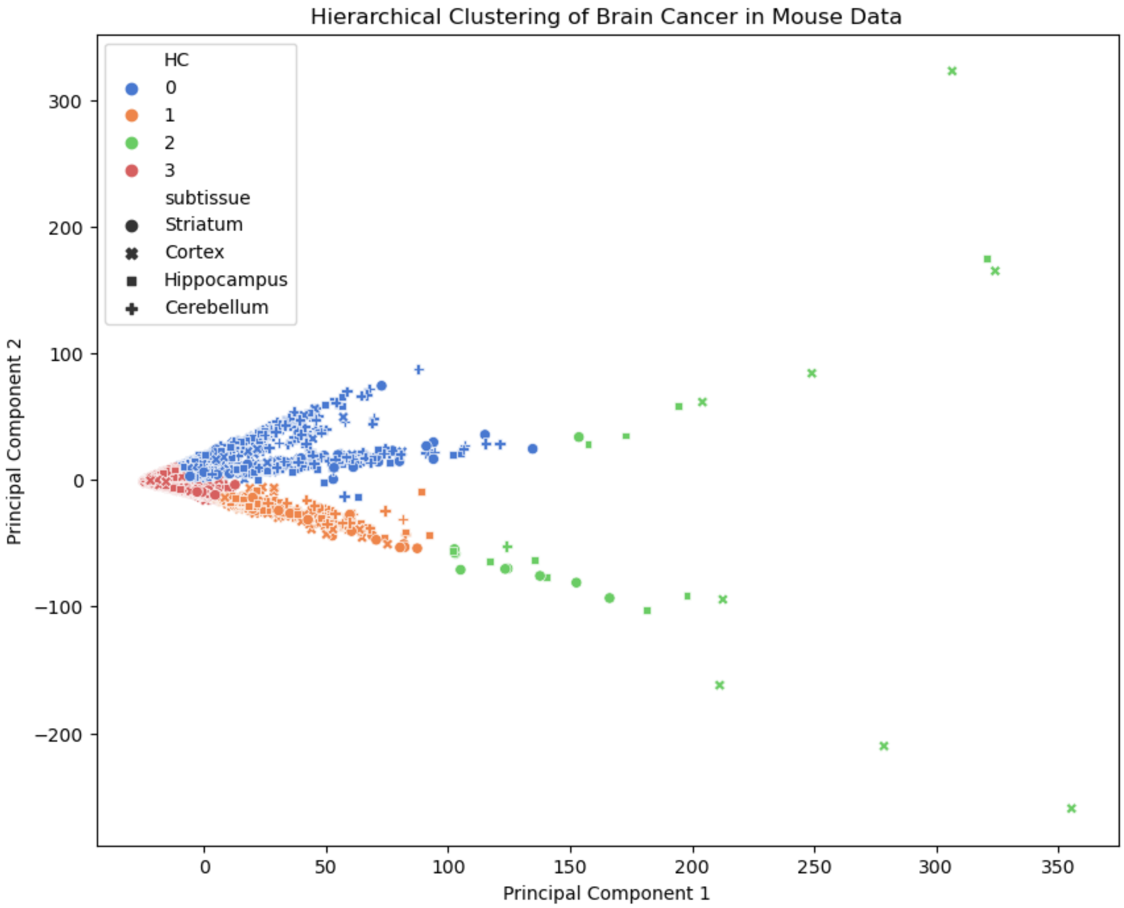
c. **How well was K-means clustering in separating the cells based on the "true" cell types? Answer this by providing the F-measure of the clustering and what the value means in relation to the clusters.**

Answer: The F-measure obtained is 0.0128 which is a very low score and indicates poor performance of the clustering method in classifying the data points to its true classes. It indicates that the clustering obtained has very low precision and recall. The K-means method did not perform too well in separating the cells based on the "true" cell classes. For e.g. there are two different clusters (2 and 5) that only contain data points that belong to the "oligodendrocyte" ontology class. It shows that the K-means clustering algorithm was not able to put these data points in the same cluster even though they all belong to the same ontology class.

d. **What might you recommend based on your analysis as next steps to improve your clustering (give 2 recommendations)?**

Answer: (1) Trying different values of k and alternative clustering algorithms → we can use elbow method to find the optimal value of K for this dataset. We can then use that optimal K values to cluster our data points. We can also try experimenting with other clustering algorithms to see if other algorithms perform better. (2) Using different metrics: We can also use K-means clustering with different metrics and evaluate which metric provides us with the most well-defined clusters. Different metrics to try are manhattan distance and mahalanobis.

3. **After performing that analysis, you chose to look further into the other most significant variable present in the metadata, the origin where the cell was cultured from (see the "subtissue" column in the metadata file). Choose another clustering technique besides K-means (i.e., Hierarchical clustering, Gaussian Mixture Models, Density Bases) you learned in class, cluster the gene expression data and provide the plot. Afterwards provide a brief description of your results addressing the final criteria: (1) what clustering technique did you choose, (2) how well did the data cluster using that technique (using two evaluation methods to justify your answer) and (3) what conclusions (express in both technical and laymen terms) might you make from this analysis that can be used to further explore glioblastomas in mice.**



Answer: (1) I chose Hierarchical Clustering. (2) I used Silhouette Coefficient and Davies Bouldin Score to evaluate my clustering. It obtained a silhouette score of 0.46 and a davies bouldin score of 0.94. Based on the silhouette score, I can say that the Hierarchical clustering method did a decent job at clustering data points into four clusters. A silhouette score of 0.46 indicates a reasonable intra-cluster distance which implies that relatively similar data points are present in the same cluster. It shows clusters are reasonably well separated. This score also shows a vast room for improvement in the quality of clusters obtained. Trying out different linkages and distance methods could yield a better silhouette score. Based on the Davies bouldin score, I can say that the clusters obtained through the Hierarchical clustering method are decent but show some overlap and ambiguity with the neighboring clusters. It suggests that cluster assignment is not very clear cut for some data points which is causing this overlap and ambiguity while also maintaining reasonably well-defined clusters for this dataset. This score also shows that there is room for improvement in the quality of clusters obtained. (3) We implemented k-means on this dataset to create clusters for gene expression with the prior knowledge of 7 cell ontology classes. After evaluating the clusters using the silhouette coefficient, we concluded that although k-means provide us with decently separated clusters, there is still room for improvement in the quality of clusters. Such improvement could be achieved by trying out different values for k and by implementing other clustering algorithms. We can also try collecting data points of gene expression based on subtissues separately and then cluster for seven different cell types. For e.g. one dataset containing gene expression observed only in striatum can then be clustered for the seven different cell types. Or, we can try to collect data points of gene expression for different cell ontology classes separately and then cluster for four different subtissues. We could also try using different dimensionality reduction methods other than PCA to get better feature selection for clustering methods.