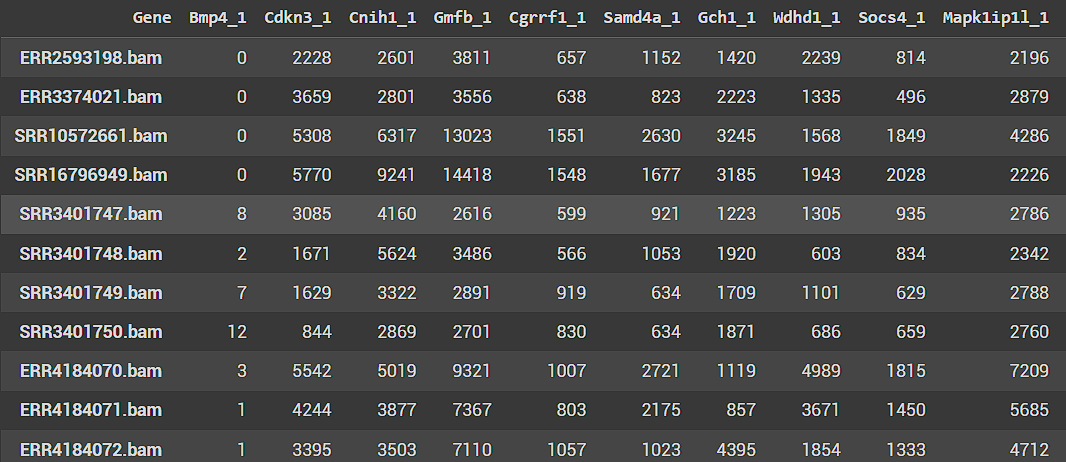
**Dataset**

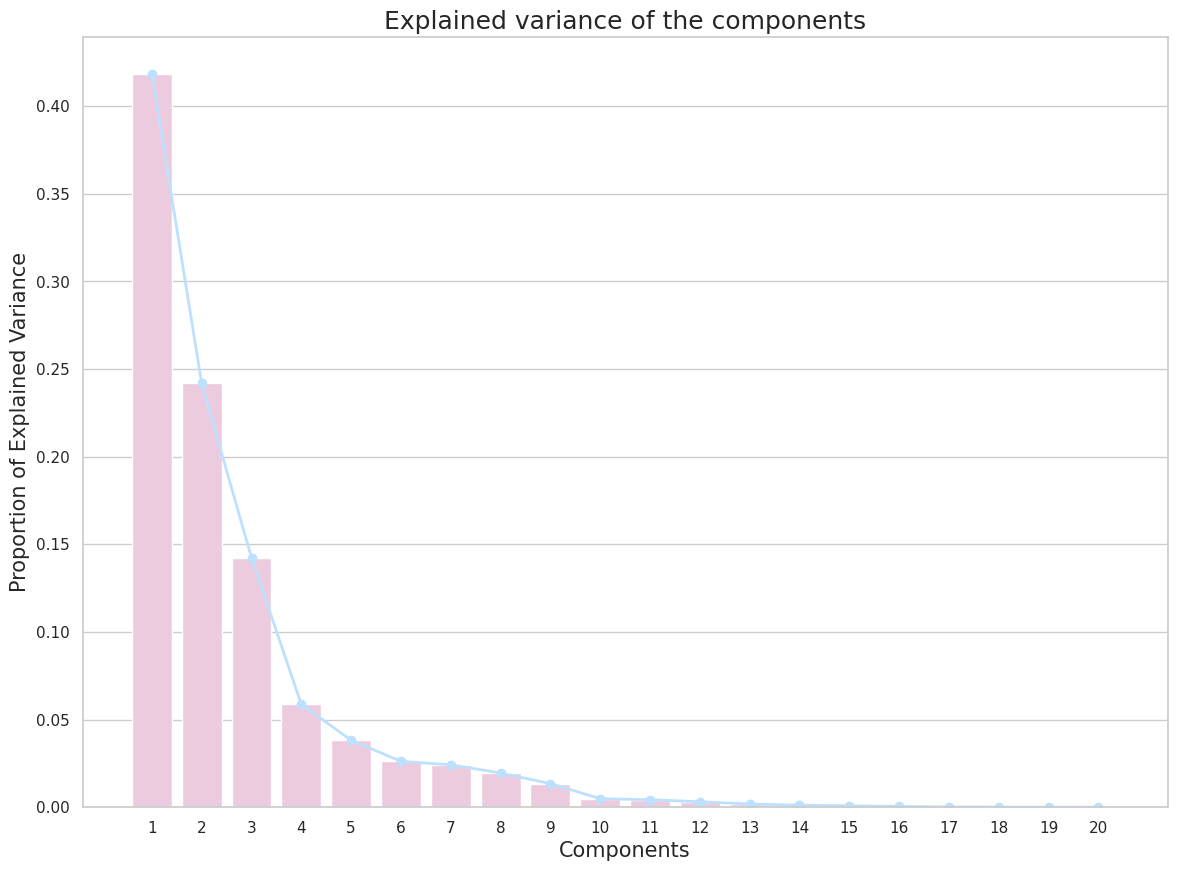
This dataset contains the expression values of 21487 genes measured across 20 different tissue/cell lines of Chinese Hamsters.



Each row consists of gene expression levels of 21487 genes whereas each column consists of the 20 different tissue/cell line samples.

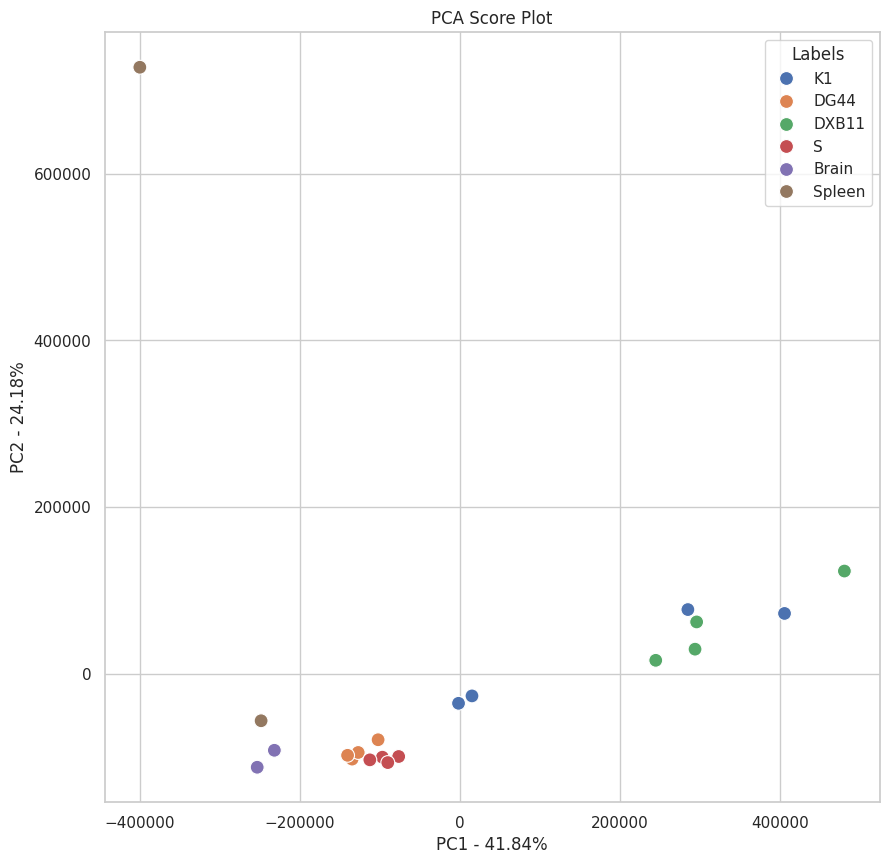
**Performing PCA**

Now performing PCA on the above-mentioned dataset (Data was scaled with zero mean before performing the PCA) reduces the total number of components to 20 (i.e. min(no. of rows, no. of cols)). The following scree plot shows the percentage variance of each Principal component.



The top 2 components have a variance of 41.84% and 24.18% respectively.

Further looking at the Score plot for PC1 vs PC2,



we can see that roughly 5 clusters can be formed. Cell line groups clustering together include **DG44 & S**, **Brain & Spleen** and **K1 & DXB11.**

**The**

**Final Analysis**

After comparing the overall results, we can see that clustering with the ‘Manhattan’ distance metric has given much better results compared to the ‘Euclidean’ distance metric. We can also say that the expression values for the genes with RMS and EWS tumour cell types are quite similar and thus end up in the same cluster. Hence, other types of clustering methods like DBSCAN should be used which excel in grouping together points that are packed closely together (points with many nearby neighbours), and marking points that lie alone in low-density regions as outliers.

For the noticeable difference in the cell line sample expression values and tumour samples’ expression values, we can suggest something like k-means clustering as it is a much simpler algorithm which excels at binary/categorical classifications and also takes the number of clusters as an input.

**Code**

The link for the code and all the images used can be found here:

<https://drive.google.com/drive/folders/1ySFPD5L_9d6sR1o7ktTNgQTFef90KIZJ?usp=sharing>

**The End**