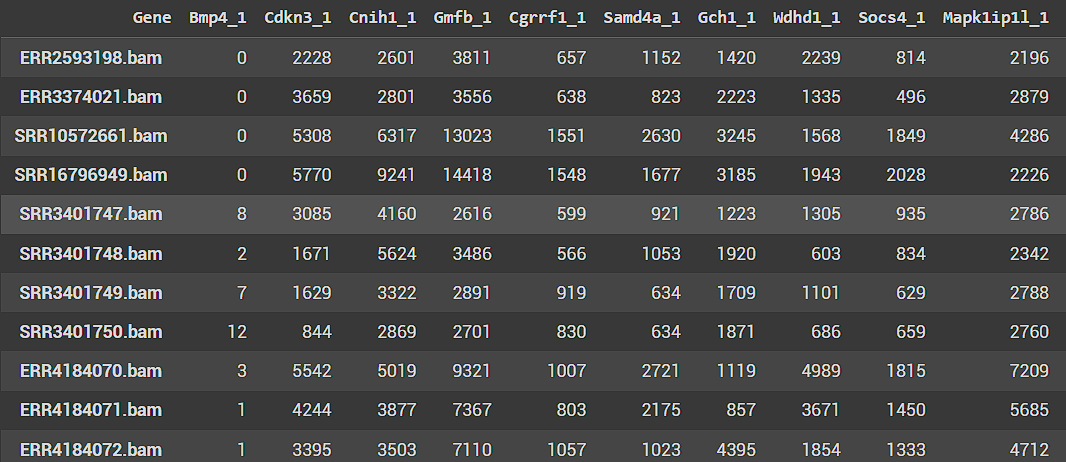
**Dataset**

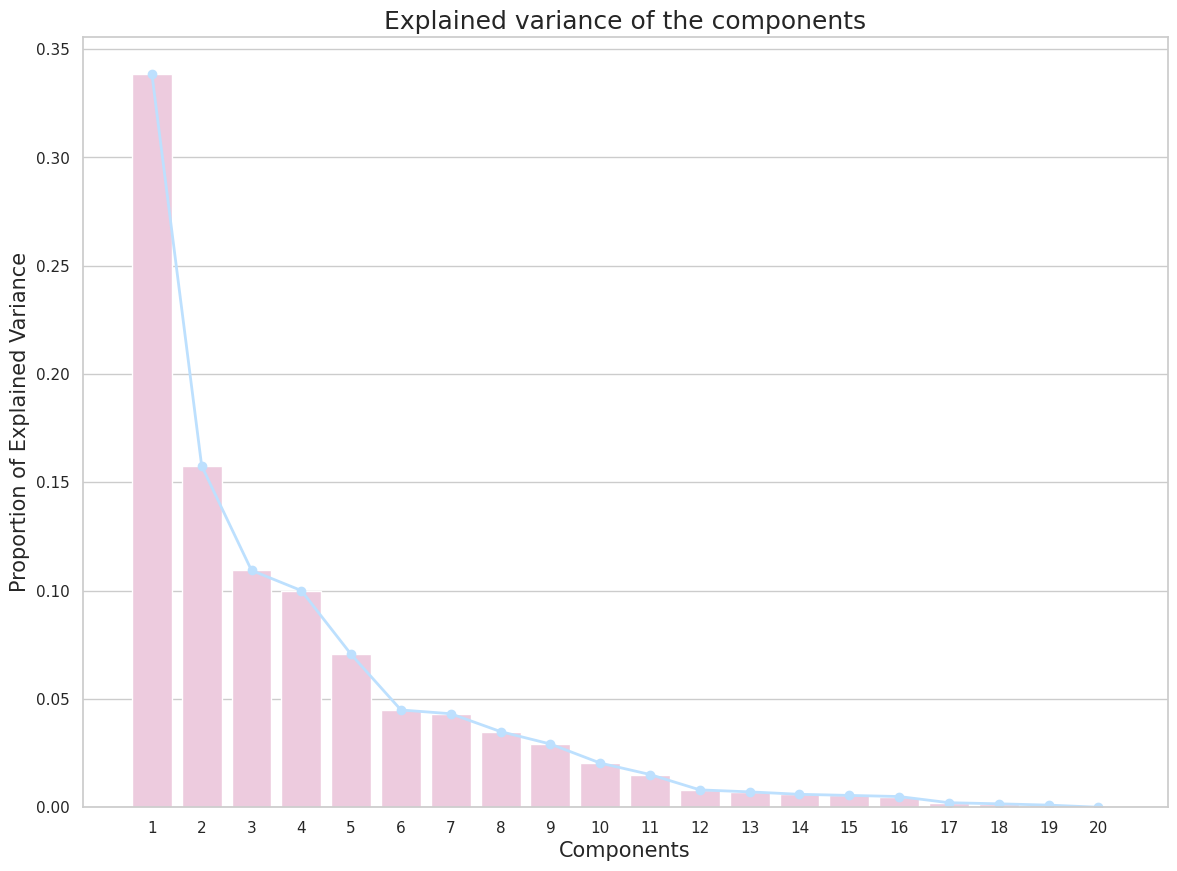
This dataset contains the expression values of 21487 genes measured across 20 different tissue/cell lines of Chinese Hamsters.



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

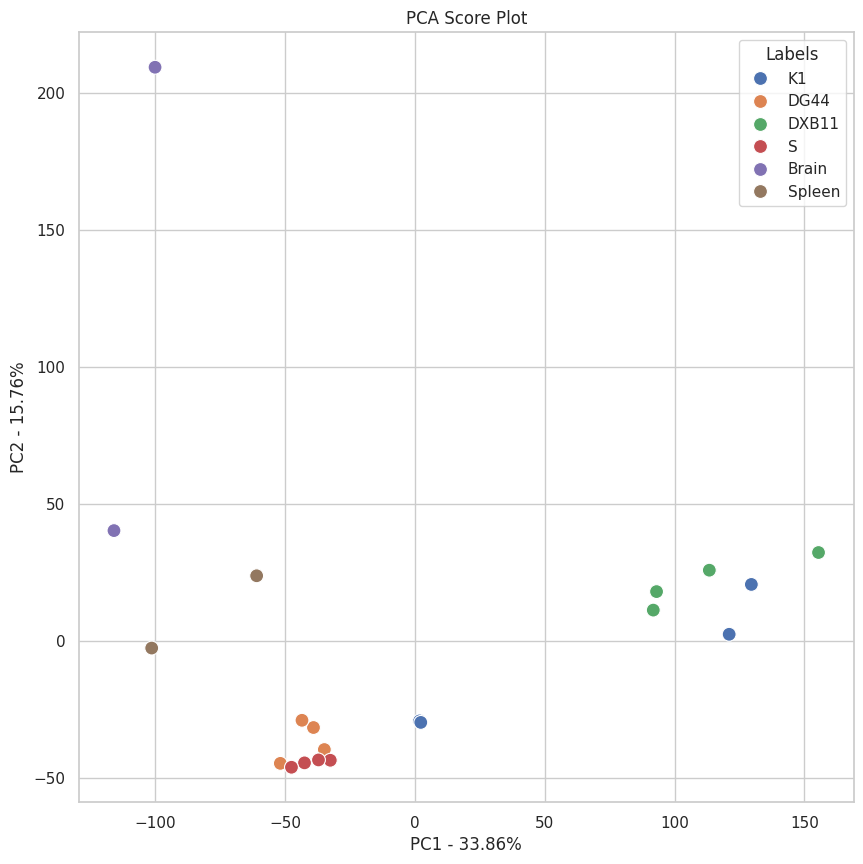
**Performing PCA**

Now performing PCA on the above-mentioned dataset 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 33.86% and 15.76% respectively.

Further looking at the Score plot for PC1 vs PC2,



we can see that roughly 4 clusters can be formed. Cell line groups clustering together include **DG44 & S**, **Brain & Spleen,** and **K1 & DXB11.**

To see which genes contributed the most to PC1 and PC2, an eigenanalysis will provide a set of unit vectors (eigenvectors) and a corresponding vector of values (eigenvalues). The unit vectors define orthogonal directions in the same feature space as before, but optimized to capture the maximum amount of the variance of the input data in the fewest number of vectors. The eigenvalues tell you how much of the original variance is captured in the direction of the corresponding eigenvector. Therefore, a larger eigenvalue means a higher contribution.

Top 5 genes with the highest contribution to PC1:

1. LOC100770091\_2
2. Spata21\_1
3. Lrrfip2\_1
4. Myo19\_1
5. Bbs4\_1

Top 5 genes with the highest contribution to PC2:

1. LOC100761590\_1
2. LOC100766062\_1
3. LOC100755925\_1
4. Fbxo46\_1
5. Ppp1r7\_1

**Final Analysis**

Principal Component Analysis (PCA) is a powerful tool used in exploratory data analysis and for making predictive models. It is a method used to bring out strong patterns in a dataset by suppressing variations.

When applied to the gene expression data, PCA helped us understand the following:

1. **Variation in Gene Expression**: The principal components represent the largest sources of variation in the gene expression data. The first principal component represents the direction in the high-dimensional space along which the samples vary the most. We were thus able to see the trend of gene expression levels in a lower dimensional space.
2. **Sample Grouping**: The PCA plot can show whether samples within the same group have similar gene expression profiles. Samples from the same group would cluster together in the PCA score plot.
3. **Outlier Detection**: Outliers, or samples that have unique gene expression profiles, will appear as isolated points in the PCA score plot.

Thus, PCA allows us to visualize the trends of the higher dimensional data in a low dimensional space allowing us to deduce important insights from it.

**Code**

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

<https://drive.google.com/drive/folders/1vuBeYWczmkf7Bh3ZE-sH7JkYWfYS6QwS?usp=drive_link>

**The End**