Anirudh Rao
BE21B004

BT3041 – Analysis and Interpretation of Biological Data

Assignment 3

Report

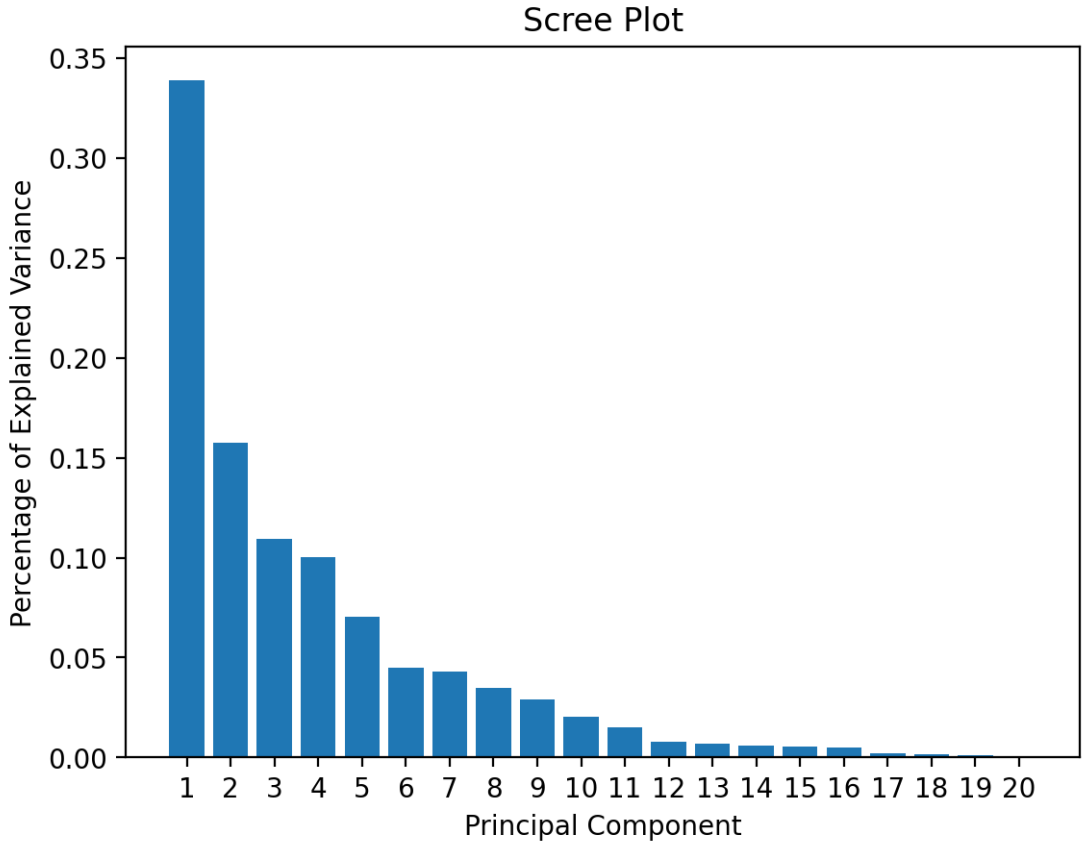
The link to the Python notebook containing the code to generate the plots and other information in this report can be found here –

https://drive.google.com/file/d/1gWVZYcJFhuqsCIAlJsyJ4rp-UuXakemT/view?usp=sharing

- The given dataset contains the gene expression profiles of 20 cell lines or tissues derived from the Chinese hamster.
- The profile contains information about 21487 genes.
- In order to reduce the dimensions of the dataset from 21487 to 20, we perform principal component analysis (PCA).
- Before performing PCA, we first scale the data to ensure PCA is implemented properly.

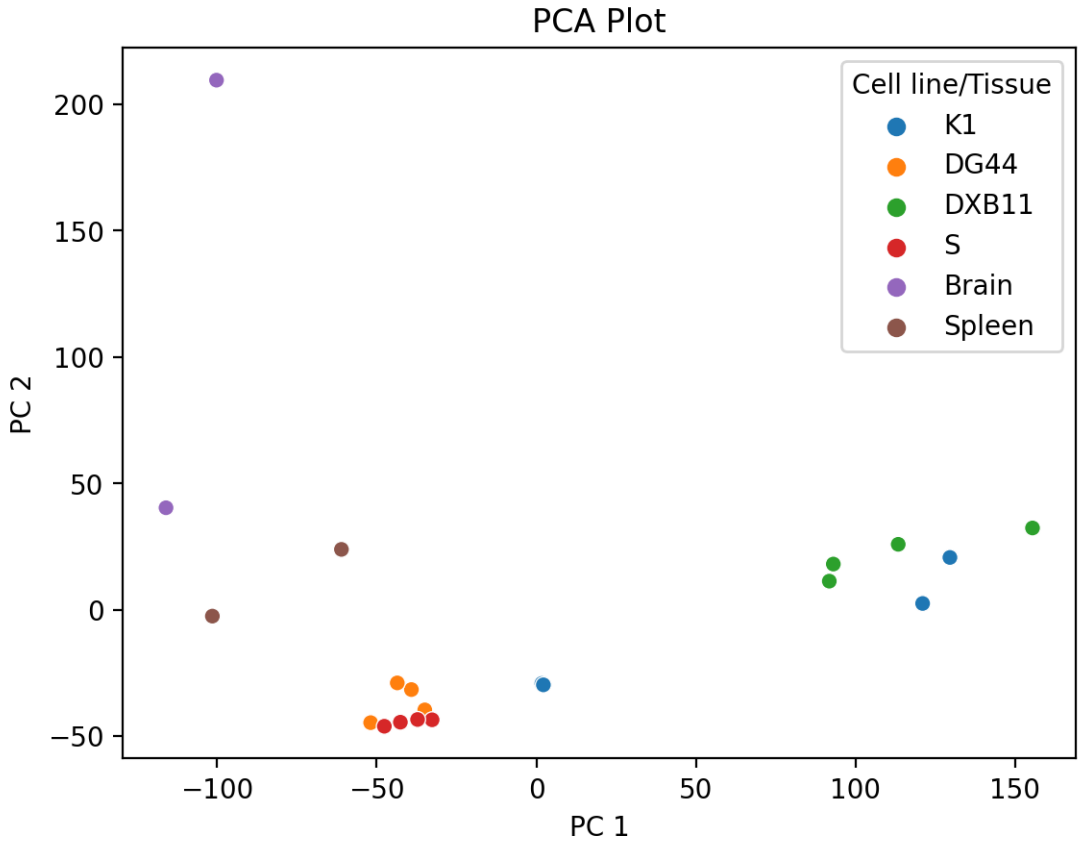
Percentage of Explained Variance

After performing PCA and obtaining the 20 principal components (PCs), we are interested in the percentage of variance explained by each of the 20 PCs. We show this information using a **scree plot**. PC 1 explains 33.86% of the variance while PC 20 explains a negligible amount.



Identifying Clusters

Using the 20 PCs identified, we transform the dataset to have only 20 features given by each of the PCs. We then use a scatter plot to identify clusters in the 20 samples using the first two PCs (which explain the most variance). The samples are coloured based on the cell line or tissue they are derived from.



From this we can clearly observe the following:

- Samples from the DG44 and S cell lines cluster together.
- Samples from the DXB11 and K1 cell lines cluster together.
- Samples from the spleen are close together.
- Samples from the brain are quite separated from the other samples.

Genes with the Greatest Contribution

We then identify the top 10 genes that contribute to PC 1 and PC 2. We decide this based on the absolute values of the coefficients of the genes' contribution to the particular principal component.

	PC 1
Gene	
Rassf3_1	0.011930
Szrd1_1	0.011926
Mlh1_1	0.011879
Pigw_1	0.011871
Adpgk_1	0.011862
Cep135_2	0.011855
LOC100773300_2	0.011848
Styx_1	0.011844
Tor3a_1	0.011830
Zdhhc3_1	0.011819

	PC 2
Gene	
Trim36_1	0.017212
Dapk1_1	0.017193
Chd3_1	0.016973
Ano7_1	0.016914
Bhmg1_1	0.016914
Foxred2_1	0.016908
Smpd3_1	0.016865
Fam189b_1	0.016857
CUNH2orf88_1	0.016839
Mmp24_1	0.016829

- The gene Rassf3 contributes the most to PC 1.
- The gene Trim36 contributes the most to PC 2.