Analytical Interpretation of Biological Data

BE21B037 – Assignment 3

1 Question

Download the gene expression data set from this website This dataset contains the expression values of 21487 genes measured across 20 different tissue/cell lines of Chinese Hamsters. Perform Principal Component Analysis (PCA) on this dataset to reduce the dimensions of number of genes. Do scale the data with zero mean for each row, i.e. each gene, before performing PCA. Report the percentage variance captured in each of the 20 principal components using a scree plot. Then using the first two principal components, plot the PCs and look for clustering of groups, if any. Also, identify the top contributor genes for the first two PCs.

2 Data Provided

The Data provided includes two csv files, one with the gene-expression values and another with the cells metadata which includes the location of each sample's cell line/tissue. We extract both the dataset and the metadata, and create the workable dataset with the updated cell names.

The dataset:

	Gene	Genes	Bmp4_1	Cdkn3_1	Cnih1_1	Gmfb_1	Cgrrf1_1	Samd4a_1	Gch1_1	Wdhd1_1	Socs4_1	...	LOC113839138	LOC113839140	LOC113839142	LOC10076415
0	K1	0	2228	2601	3811	657	1152	1420	2239	814	...	0	0	0	
1	K1	0	3659	2801	3556	638	823	2223	1335	496	...	0	0	0	
2	K1	0	5308	6317	13023	1551	2630	3245	1568	1849	...	0	1	0	
3	K1	0	5770	9241	14418	1548	1677	3185	1943	2028	...	0	2	25	
4	DG44	8	3085	4160	2616	599	921	1223	1305	935	...	0	0	0	
5	DG44	2	1671	5624	3486	566	1053	1920	603	834	...	0	0	0	
6	DG44	7	1629	3322	2891	919	634	1709	1101	629	...	0	0	0	
7	DG44	12	844	2869	2701	830	634	1871	686	659	...	0	0	2	
8	DXB11	3	5542	5019	9321	1007	2721	1119	4989	1815	...	0	0	0	
9	DXB11	1	4244	3877	7367	803	2175	857	3671	1450	...	0	0	0	
10	DXB11	1	3395	3503	7110	1057	1023	4395	1854	1333	...	3	4	1	
11	DXB11	2	3823	3630	7867	1184	1132	4863	2005	1451	...	0	1	0	
12	S	0	2287	2842	3540	397	481	1759	781	521	...	0	0	0	
13	S	0	2004	2278	2594	401	357	1290	619	400	...	0	0	0	
14	S	1	2107	2534	2754	391	342	1317	553	390	...	0	0	0	
15	S	2	1603	2365	3224	372	433	1515	622	422	...	0	0	0	
16	Brain	83	88	508	718	206	595	53	184	327	...	0	0	0	
17	Brain	79	72	765	2839	515	880	55	80	161	...	0	0	0	
18	Spleen	137	5013	796	1361	354	330	767	1792	729	...	0	0	0	
19	Spleen	165	132	501	1544	336	801	270	207	787	...	0	0	0	

20 rows × 21488 columns

The above shown data was pre-processed as follows:

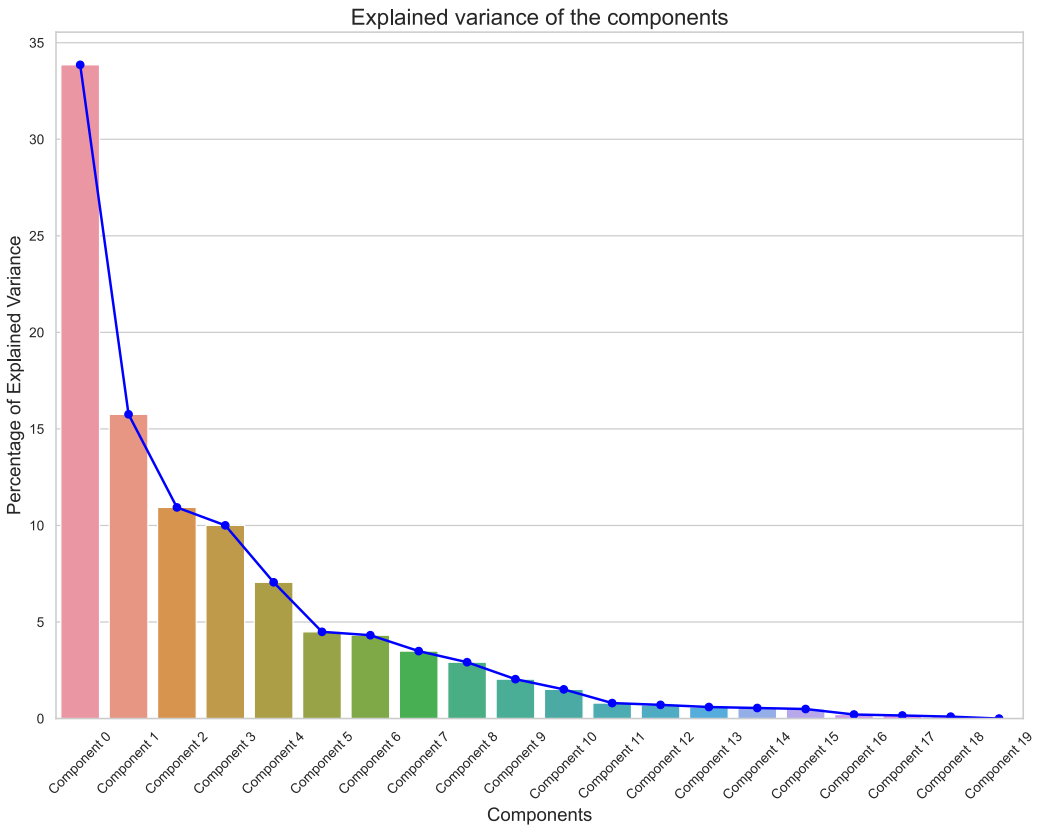
1. Transpose dataset.csv

2. Merge the two datasets by replacing gene name with their sample tissue/cell line

3 Scree Plot

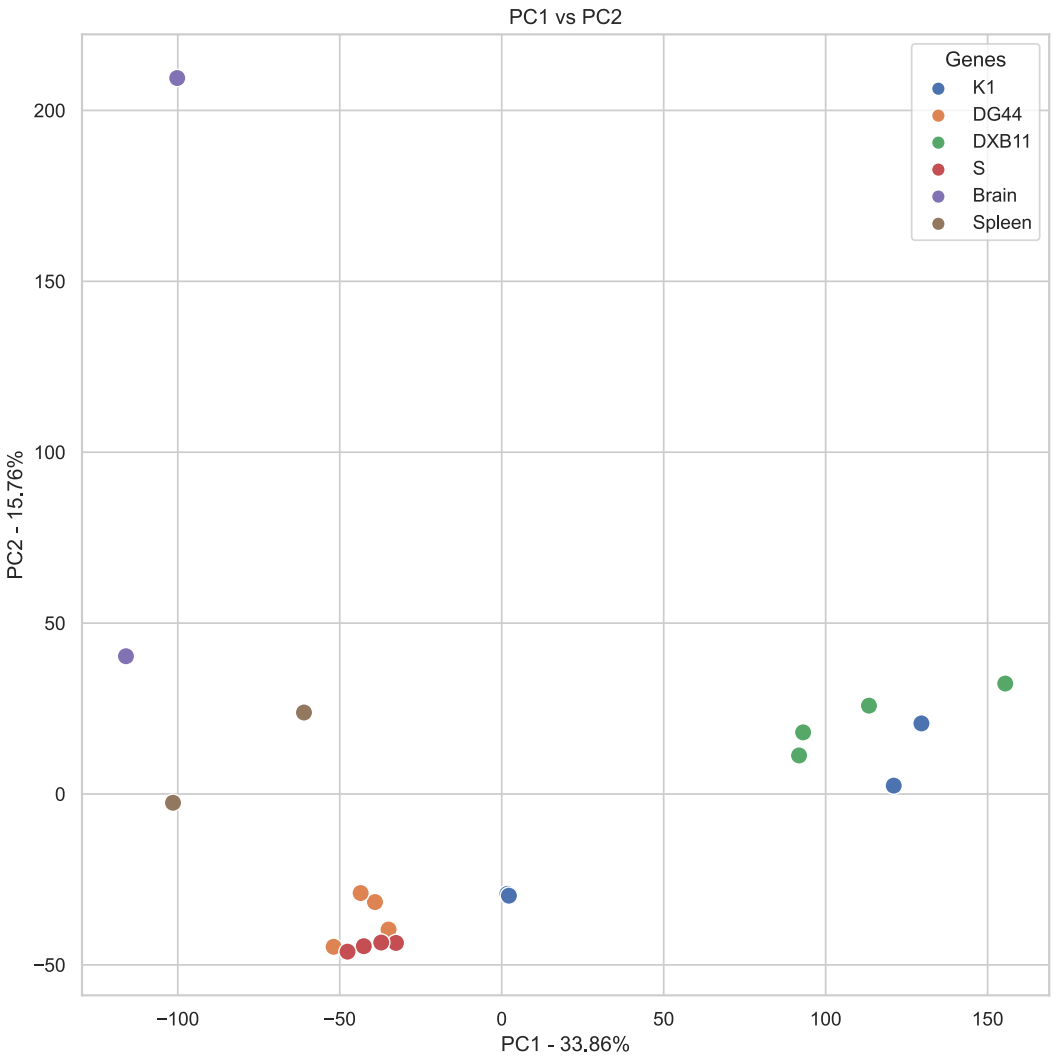
The scree plot then observed is plotted along with the histogram plot.

Assumption: The question only specify details about the mean being zero but not about the standard deviation, which we will set as 1, so as to normalize the data.



We now plot the top two principal components

4 PC1 vs PC2



We observe that the DG44 & the S cell lines are clustered together so are the DXB11 & K1 cell lines. One Sample from the brain tissue is an outlier in this scatter plot

The top contributors for the first two Principal components are:

Gene	PC1 (gene: contribution factor)	PC2 (gene: contribution factor)
1st most	Rassf3_1: 0.01193	Trim36_1: 0.017212
2nd most	Szrd1_1: 0.011926	Dapk1_1: 0.017193
3rd most	Mlh1_1: 0.011879	Chd3_1: 0.016973
4th most	Pigw_1: 0.011871	Ano7_1: 0.016914
5th most	Adpgk_1: 0.011862	Bhmg1_1: 0.016914

Table 1: Top contributing genes for PC1 and PC2

5 Appendix

The code to recreate all the above plots and tables can be found here.

* **