Assignment 3

Question. Download the gene expression data set from this website: https://drive.google.com/drive/folders/1M3Q5TGttSiK_TDTezNN8vOMROOK_hMeE?usp=sharing. 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 a 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 contributing genes for the first two PCs.
Please provide your analyses results in a report form, specifically answering each of the above questions with relevant figures, etc. Also, state any assumptions made clearly in the report.
Attach the Google Drive link to your software codes (MATLAB/Python) used for performing calculations with the report.

Solution. I will be using the Python programming language to code for each of the questions. Firstly, I have loaded the data for gene expression. I have also loaded the metadata that contains the cell line/tissue associated with each gene.
The initial part of the code is as follows:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

sns.set(style='darkgrid', context='paper', rc={'figure.figsize':(1,10)})

from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

full_data = pd.read_csv("Dataset.csv")
data = pd.read_csv("Dataset.csv")
metadata = pd.read_csv("Metadata_bt.csv")
```

Listing 1. Loading the data

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables.PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models.
The main goal of principal component analysis (PCA) is to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables without any prior knowledge of the target variables.
In PCA, we choose the principal components based on their explained variance. PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.
Explained variance is the term that gives us an idea of the amount of total variance that has been retained by selecting the principal components instead of the original feature space.
In order to perform the Principal Component Analysis, I need to drop the gene column from the data. This is because the PCA can be performed only on the numerical representation of data, and the gene column's purpose is just to associate with the metadata.

After dropping the gene column from the data, I created a PCA instance and fit in on the gene expression data. The code to perform the above is given below:

```
data.drop('Gene', axis='columns', inplace=True)
scaled_data = StandardScaler().fit_transform(data.T)

pca = PCA(n_components=20)
pca.fit(scaled_data)
print(end="")
```

Listing 2. Code to fit PCA on data

We now know the percentage variance that has been captured by each principal component. It can be expressed in the form of a bar graph as well as a screeplot. The code for the bar graph and the screeplot is shown below, along with the plots generated.

```
PC_values = np.arange(pca.n_components_) + 1
plt.figure(figsize=(10,5))
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', color='green')
plt.title('Scree Plot for explained variance of the components', size=20)
plt.xlabel('Principal Components', size=15)
plt.ylabel('Proportion of Variance Explained', size=15)
plt.show()

PC_values = np.arange(pca.n_components_) + 1
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', linewidth=2, color='
    green')
plt.title('Scree Plot for explained variance of the components', size=20)
plt.xlabel('Principal Components', size=15)
plt.ylabel('Proportion of Variance Explained', size=15)
plt.show()
```

Listing 3. Code to plot the explained variance

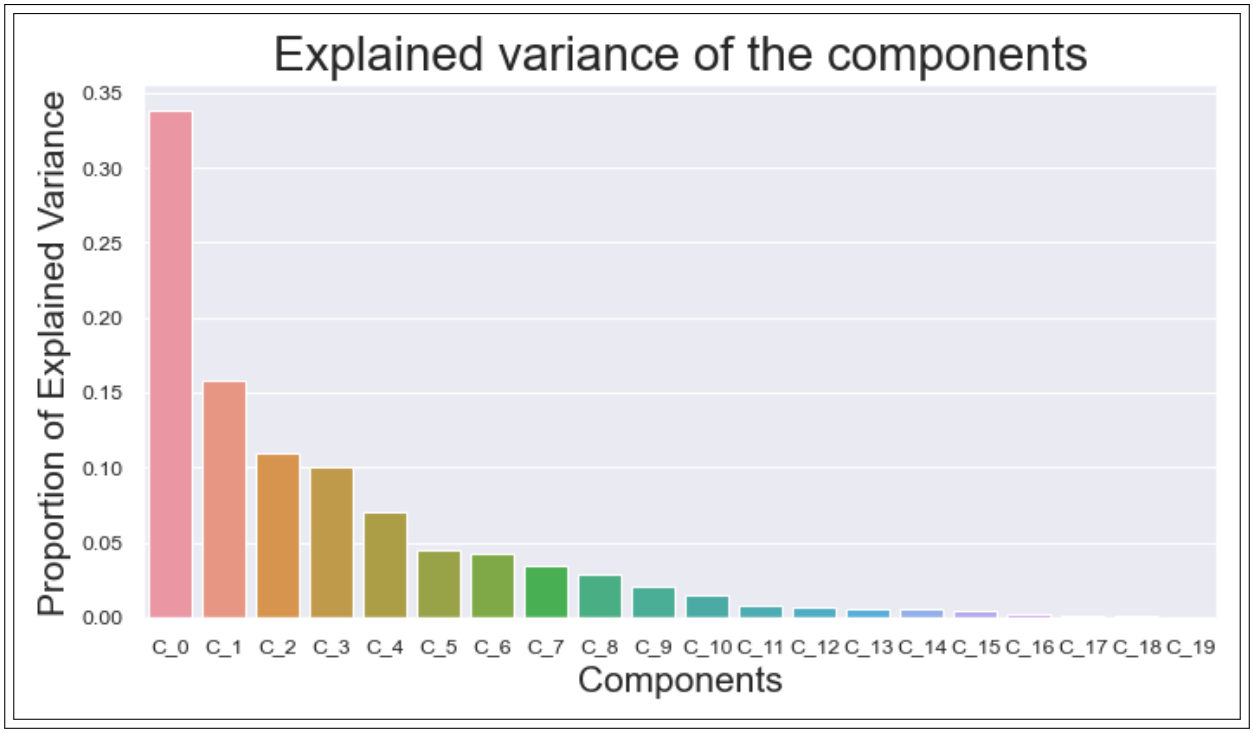

Figure 1. Barplot for proportion of explained variance

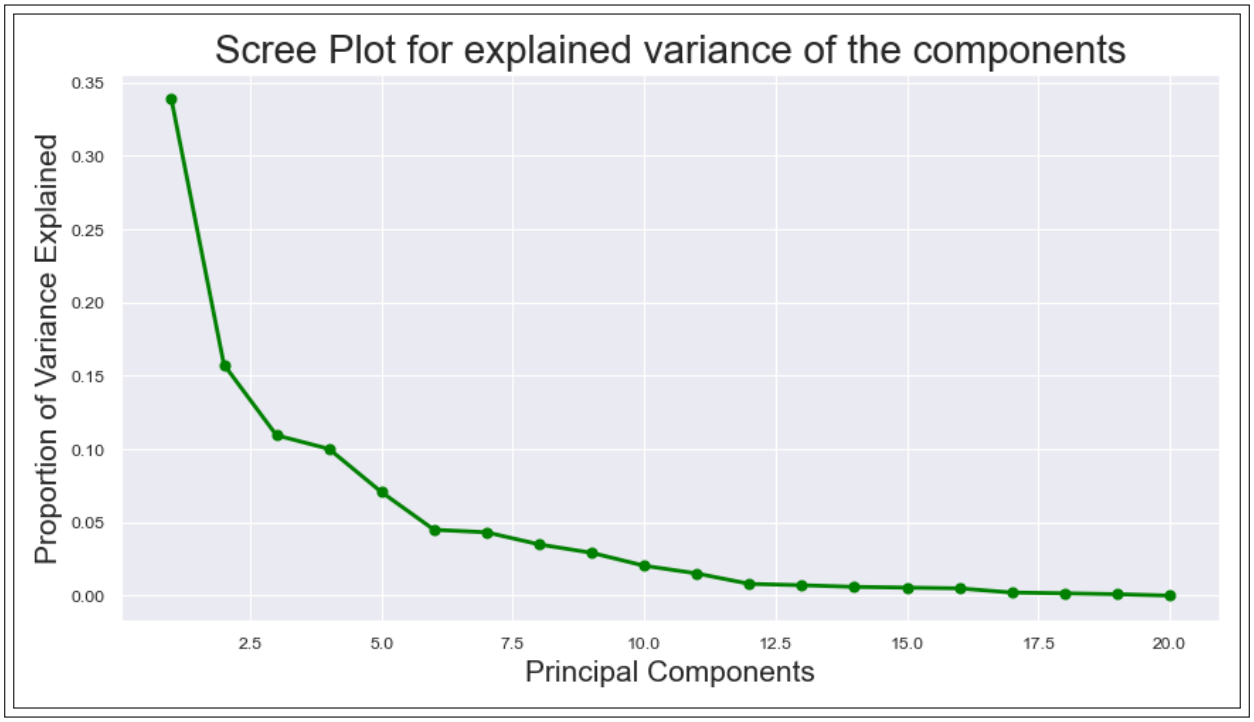

FIGURE 2. Screeplot for proportion of explained variance

% Variance of explained variance			
PC number	**% of explained variance**	**PC number**	**% of explained variance**
PC_1	33.85575081185059	PC_11	1.5116312842837301
PC_2	15.755995736369505	PC_12	0.8022155390154433
PC_3	10.935363385048383	PC_13	0.7095441010041783
PC_4	10.005365969506357	PC_14	0.59773536586614166
PC_5	7.056815408187435	PC_15	0.5467895697413557
PC_6	4.491091689448546	PC_16	0.494633597676278
PC_7	4.317314391092499	PC_17	0.20950431178734566
PC_8	3.4931966373824355	PC_18	0.15894167243431017
PC_9	2.9188019165569037	PC_19	0.09911556080759519
PC_10	2.04019305194095	PC_20	4.307883021936279e-31

TABLE 1. % Variance captured by PCs

The percentage variance captured in each of the 20 principal components using the screeplot is shown in the table below:
Now, we have obtained the 20 principal components for gene expression using PCA. I create a dataframe out of this transformed data to make the plotting task easy. This means that the new dataframe will have 20 components. I added another column for **Cell line/tissue**, which contains the gene-to-tissue correlation as given in the metadata. This helps us group the classes that are plotted on the PCA plot.

The observations from the screeplot for the explained variance of the principal components are as follows:

- The y-axis represents the proportion of variance explained by each PC. A higher value indicates that the PC captures a larger portion of the total variance in the data.
- The initial, steeper decline in the curve suggests that the first few principal components capture the most significant sources of variation in the data. These components likely represent the most important underlying patterns in gene expression levels.
- There's no clear **elbow** point where the curve plateaus definitely. However, the rate of decrease in variance explained slows down around PC5 or PC6. This suggests that these components might still explain a meaningful portion of the variance, while subsequent components likely capture progressively smaller amounts.
- Now, if we were to keep the threshold for proportion of variance explained as 0.05, which is 5%, we can see that the first 5 components are sufficient to capture most of the proportion of variance observed across the entire gene expression.
- The sum of the proportion of variance captured by the first 5 components comes up to 0.7760929131096226, which is about 77% of the explained variance. Similarly, the sum of the proportion of variance captured by the first 8 components comes up to 0.8991089402888574, which is about 90%. Depending on the threshold of total variance that we chose to represent, we can take as many components and go ahead.
- The decision of how many PCs to retain depends on the specific goals of the analysis and a balance between capturing variance and avoiding overfitting. In the screeplot, though there is no strict rule, the thumb rule is to choose the number of PCs where the curve starts to flatten out significantly, which here could be around PC5 or PC6.

Now, I plot the PCA of the first two principal components. The code for the same is given, followed by the plot:

```
fig, ax = plt.subplots(figsize=(10,6))

sns.scatterplot(data = transformed_df, x='C_1', y='C_2', hue = 'Cell line/
    Tissue', s=200)

ax.set_xlabel(f"PC1 - {pca.explained_variance_ratio_[0]*100:.2f}%")
ax.set_ylabel(f"PC2 - {pca.explained_variance_ratio_[1]*100:.2f}%")
ax.set_title("PCA Score Plot")
plt.show()

# The first two PCs capture about 50\% of the explained variance.
```

Listing 4. Code to plot the PCA of two components

Some of the observations from the PCA plot are as follows:

- We can see roughly two clusters in the graphical plot. One is a slightly dense cluster, consisting of genes corresponding to S and CG44 cell lines/tissues. The other is a slightly sparser cluster, consisting of genes corresponding to K1 and DXB11. This suggests that the gene expression patterns in the cell lines of S and CG44 are quite similar. The same observation is true for the cell lines K1 and DXB11.
- The gene expression values of S and CG44 cell lines are quite similar and less spread out with respect to the principal components PC1 and PC2.
- The variance in the gene expression of brain tissue is quite high with respect to PC2.
- The variance in the gene expression of K1 tissue is quite high with respect to PC1.
- One trivial observation is that the spread of the data points is larger along the PC1 axis as compared to PC2. This indicates that PC1 captures a larger proportion of variance in gene expression levels across samples. This just serves as confirmation for the theory behind PCA, where the higher components capture greater variance in data.

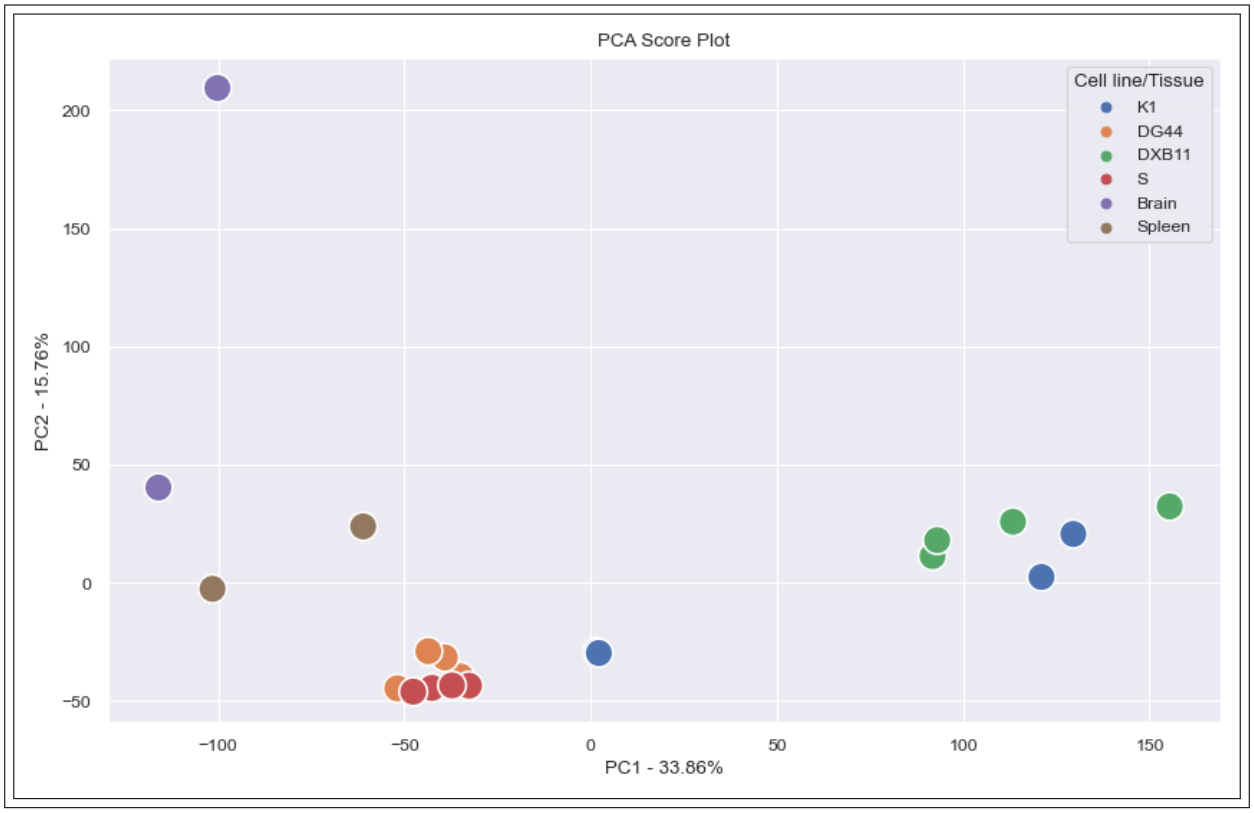

FIGURE 3. PCA plot for first two components

Top Contributors in PC1		**Top Contributors in PC2**	
Gene	**Coefficient in component**	**Gene**	**Coefficient in component**
*Rassf*3_1	0.011930	*Trim*36_1	0.017212
*Szrd*1_1	0.011926	*Dapk*1_1	0.017193
*Mlh*1_1	0.011879	*Chd*3_1	0.016973
*Pigw*_1	0.011871	*Ano*7_1	0.016914
*Adpgk*_1	0.011862	*Bhmg*1_1	0.016914

TABLE 2. Top Contributing genes

Top contributing genes for Principal Components

PCA helps identify the most significant underlying patterns in the dataset. By analyzing the loadings (contributions) of the genes on each PC, we can determine the genes with the strongest influence on the variation captured by that PC.

By examining the genes with the highest absolute loadings on PCs, we can identify the genes that are most responsible for the overall differences observed in gene expression patterns across the samples.

Here, we focus on the top contributing genes for the first two principal components based on their absolute loading values. Below are the tables for the top contributing genes for PC1 and PC2.

The table lists the top genes that contribute most to the variation captured by PCs in the PCA of the gene expression data. The value of the **Coefficient** column represents the loading of each gene. Genes with higher absolute loading values (positive or negative) contribute more to the variation along the PC.

Here are some observations for the top contributing genes of PC1 from the table:

- The gene with the highest absolute loading on PC1 is $Rassf3\_1$
- Other potentially important contributors to PC1 include $Szrd1\_1$, $Mlh11\_1$, $Pigw1\_1$, and $Adpgk\_1$

Here are some observations for the top contributing genes of PC2 from the table:

- The gene with the highest absolute loading on PC2 is $Trim36\_1$
- Other potentially important contributors to PC1 include $Dapk1\_1$, $Chd3\_1$, $Ano7\_1$, and $Bhmg1\_1$

Drive link to codes