Assignment 1

General information

Please give short (2-3 sentences) interpretations / explanations to your answers, not only the program code and outputs. Be concise and focused (less could be more ;)).

Grades will be distributed with the following rule: from the points you earn, you get 100% if you submit until the due date (**2025-03-09 21:00 CET**), 50% within 24 hours past due date, and 0% after that.

Gene expression data

From the ISLR website, we can download a gene expression data set (`Ch10Ex11.csv`) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

We would have no chance to estimate any model on these 1,000 features. However, we could reduce the dimensionality with PCA. Then, we could look at the relation of the first few principal components (that captures part of the variance of all 1,000 features) and the outcome.

Note that as we only have 40 observations we can only define a subspace with at most 39 dimensions. Think of it like trying to define a volume in 3D space using only 2 points - you can only define a line, not a full 3D space. Technically, the sklearn PCA implementation will result in 40 components, but the last one will have a variance of 0.

In [1]:
```
# Load the data
import pandas as pd

# Read the CSV file
url = 'https://www.statlearning.com/s/Ch10Ex11.csv'
genes = pd.read_csv(url, header=None)

# Transpose the dataframe and convert to pandas DataFrame
genes = genes.T
genes = pd.DataFrame(genes)
print('Dimensions of genes dataframe:', genes.shape)

# Define health_status
health_status = ['healthy'] * 20 + ['diseased'] * 20
```

```
Dimensions of genes dataframe: (40, 1000)
```

Tasks

1. Plot a histogram of the variances of each feature. Can we consider them as similarly-scaled? Would you recommend standardizing the features before applying PCA? Why or why not? (*Hint*: If the variances differ at most by an order of magnitude, I would consider them as similarly-scaled.)

2. Run PCA on the (full) dataset. How much components can explain at least 90% of the total variance? How much variance is explained by the first two principal components?

3. Plot the first two principal component scores (i.e., the projections onto the first two principal components). Color the data points by their health status. Are these components informative of patient health status? Which principal component appears more important for separating healthy from diseased patients?

4. Identify the genes that contribute most significantly to the principal component that best separates healthy and diseased patients. (*Hint*: Look at the loadings of the principal component.)