Assignment

In this assignment you will have to make a PCA package in R or Python. The input to PCA function would be two files, an intensity file and a metadata file. The formats for them can be seen in the attached files named gene_data.csv and meta.csv. The package should have appropriate error handling if the input files are not of the desired format. The output of the PCA function in the package should be an interactive PCA plot of PC1 vs PC2 as shown in this image. To test out the package you need to host it behind opencpu. The final part of the assignment is to evaluate what PCA tells us about the given dataset (to understand about the data read the following paragraph), that is, are the different timepoints in the metadata file differentiating when seen on the PCA plot or not? What does that mean for the dataset given?

To understand the data let us first view the gene data. Gene data contains 32 columns of
which the gene names have been provided in a column called symbol. Corresponding to
each gene (a row) we see 30 different values. These values correspond to different samples
corresponding to the column they are in. Now, let us look at the metadata which was in
meta.csv. Once you view this data you will see a column for the sIdx. This column
corresponds to the sample names in the gene data. The next column you will see is the Time
column which corresponds to the time at which this sample was taken. Do note that there
are multiple samples for each time point.

You can refer to the following blogpost for more information about PCA.