**COMP421X Homework 06**

**Prepared By: Burak Özbek**

**Student Id: 0073811**

**HW06:** **Principal Component Analysis and Clustering**

**Question 1:** You are given a dataset (data\_set.csv) describing several gene expressions of patients having different types of tumor. The original dataset contains 20532 gene expressions for each patient, but for simplicity, we have preprocessed and cleansed the dataset and reduced this number to 209 gene expressions, given under columns G1,G2,…,G209. The last column presents tumor types (1:BRCA, 2:COAD, 3:KIRC, 4:LUAD, 5:PRAD).

**Note**: scale your X matrix using the code **X <- scale(data[,-210])**

**Answer:** Data set implementation is like this:

# read data into memory

data\_set <- read.csv("data\_set.csv", header = TRUE)

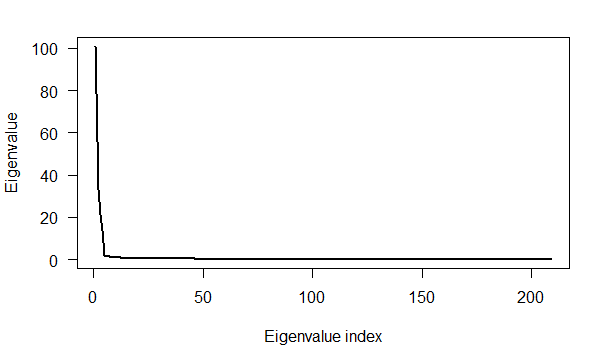
# get X and y values

X <- as.matrix(subset(scale(data\_set), select = -c(Type) ))

y <- data\_set$Type

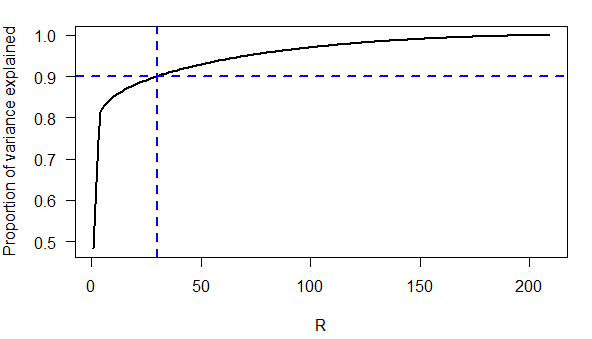
**Question 2:** Implement a principal component analysis on the variables of the dataset. Plot the scree graph.

**Answer:** First graph is like this:



**Question 3:** Calculate the proportion of variance explained and plot it. Show on the plot the number of principle components required to explain 90% of variance.

**Answer:** Second graph is like this:



**Question 4:** Calculate the first three principal components for each data point. Note: your Z matrix should be a 801 by 3 matrix.

**Answer:** The first 4 rows of calculated Z matrix are shown as below:

[,1] [,2] [,3]

[1,] -6.591802 -11.758849 -0.3534791

[2,] -4.271631 3.810821 -0.9956870

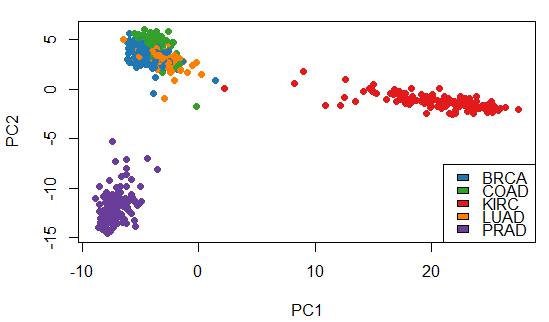
[3,] -7.409462 -5.324476 0.4347877

[4,] -8.460449 -11.654211 0.2815669

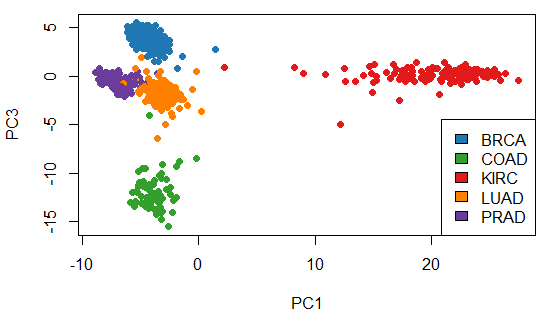
**Question 5:** In this question, you will plot three scatter plots, one for each pair of principal components computed in question 4. Plot PC1 vs. PC2, PC1 vs. PC3 and PC2 vs. PC3 in three separate scatter plots.

**Answer:** Three scatter plots is shown as below:

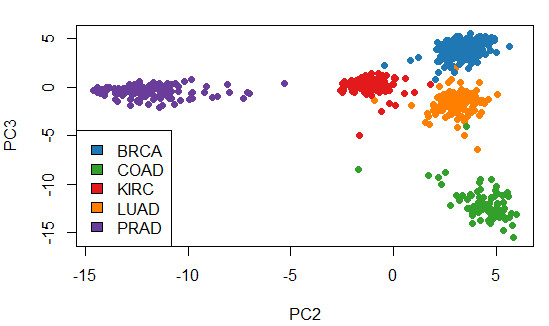
PC1 vs PC2:



PC1 vs PC3:



PC2 vs PC3:



**Question 6:** Describe what you observe in the two-dimensional plots you derived in question 5. Which tumor type can be separated using PC1? Which can be separated using PC2? Which principal components can be used for separating the five tumor types?

**Answer:** When we look at this three graph, it is clear to see that PC1 is seperates “KIRC” tumor type(The red coloured one). PC2 is seperates “PRAD” tumor type. Among this three principal components, PC2 and PC3 looks better to seperate 5 tumor types clearly.

**Question 7:** Apply the k-means clustering algorithm with k=5 clusters to the Z matrix obtained in question 4 using PC1 and PC3 only (*use set.seed(421) and 10 iterations of the k-means algorithm*).

**Answer:** Cendroids are like this:

[,1] [,2]

[1,] -3.081224 -1.6447591

[2,] 20.954489 0.1624742

[3,] -3.889589 -12.0873003

[4,] -7.138182 -0.4456767

[5,] -4.358387 4.0199409

**Question 8:** Draw the clustering results of your k-means algorithm.

**Answer:** The graph of clustring results is like this:

