Course Code: BT 4019

Course Title: Statistical Methods in Bioinformatics

Take-home Assignment - Report

Name : Nimna Alupotha Gamage (NIMNA A. G. T.)

Index No.: S14682

Reg. No. : 2019s17241

Table of Contents

Question 1: Breast cancer is the most common malignancy among women. It occurs due to the abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor can be can be benign (not cancerous), or malignant (cancerous). MRI, mammogram and biopsy are commonly used tests to diagnose breast cancer.

You are given a dataset, breast-cancer-data.csv and variable information are as follows.

Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign) 3-32)
2. Ten real-valued features are computed for each cell nucleus:

   a. radius (mean of distances from center to points on the perimeter)

   b. texture (standard deviation of gray-scale values)

   c. perimeter

   d. area

   e. smoothness (local variation in radius lengths)

   f. compactness (perimeter^2 / area - 1.0)

   g. concavity (severity of concave portions of the contour)

   h. concave points (number of concave portions of the contour)

   i. symmetry

   j. fractal dimension (“coastline approximation” - 1)

Load the data set

The "breast-cancer-data.csv" file is loaded using the `read.csv()` function. The function `read.csv()` reads a file and creates a data frame from it. The `header` attribute was set as 'True' to indicate that the file contains the names of the variables as its first line.

Code;

```
data_breast_cancer = read.csv("breast-cancer-data.csv", header = T)
```

Carry out a principal component analysis and identify the important principal components.

Extract numeric variables

The "breast-cancer-data.csv" file contains 32 columns/variables and all the variables except the first two variables are real-valued features. All the rows of those 30 numeric variables are extracted.

Code;

```
data_breast_cancer_numeric = data_breast_cancer[, 3:32]
```

In principal component analysis (PCA), it is required to extract only the numeric variables for the analysis due to some reasons such as,

- A number of mathematical processes, including computing means, variances, covariances, and eigenvalues, are involved in PCA. Only numerical variables can be used in those calculations to yield meaningful and valid results.
- The data standardization is a meaningful operation only for numeric variables.

Standardize data

In this step, the data is standardized using the `scale()` function. The scale of the variables can affect PCA. Variables with larger scales will predominate the first principal components if they are assessed in different units or have distinct ranges. Therefore, standardizing the variables before using PCA is crucial (subtract the mean and divide by the standard deviation).

Code;

```
scaled_data = scale(data_breast_cancer_numeric)
```

Perform PCA

The PCA can be performed using both `prcomp()` and `princomp()` functions. `prcomp` returns the results as an object of class `prcomp` whereas `princomp` returns the results as an object of class `princomp.`

Code;

```
Pca_pr_breast_cancer = prcomp(scaled_data)

pca_prin_breast_cancer = princomp(scaled_data)
```

As the result of PCA, it generates principle components (PCs). The generated number of PCs are equal to the number of original numeric variables. In this case, it generates 30 PCs. Principal components are new variables created by combining the initial variables in a linear fashion.

1.How many principal components would you choose? Justify your answer.

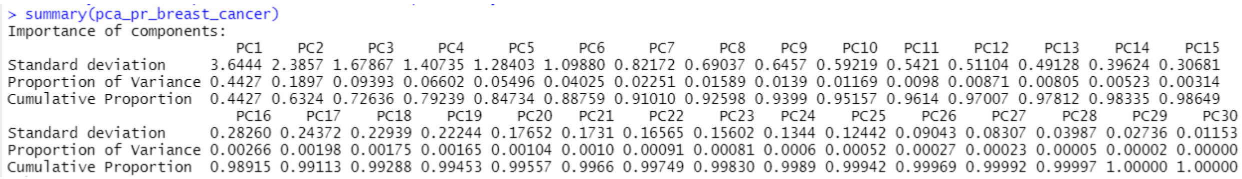
Prcomp() method

`summary()` function is used to get the standard deviation, variance and the cumulative proportion explained by each PC.

Code;

```
summary(pca_pr_breast_cancer)
```

Output;



```
> summary(pca_pr_breast_cancer)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8    PC9    PC10   PC11    PC12    PC13    PC14    PC15
Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172 0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.39624 0.30681
Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.00523 0.00314
Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010 0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.98335 0.98649
                          PC16    PC17    PC18    PC19    PC20   PC21    PC22    PC23   PC24    PC25    PC26    PC27    PC28    PC29    PC30
Standard deviation     0.28260 0.24372 0.22939 0.22244 0.17652 0.1731 0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987 0.02736 0.01153
Proportion of Variance 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005 0.00002 0.00000
Cumulative Proportion  0.98915 0.99113 0.99288 0.99453 0.99557 0.9966 0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997 1.00000 1.00000
```

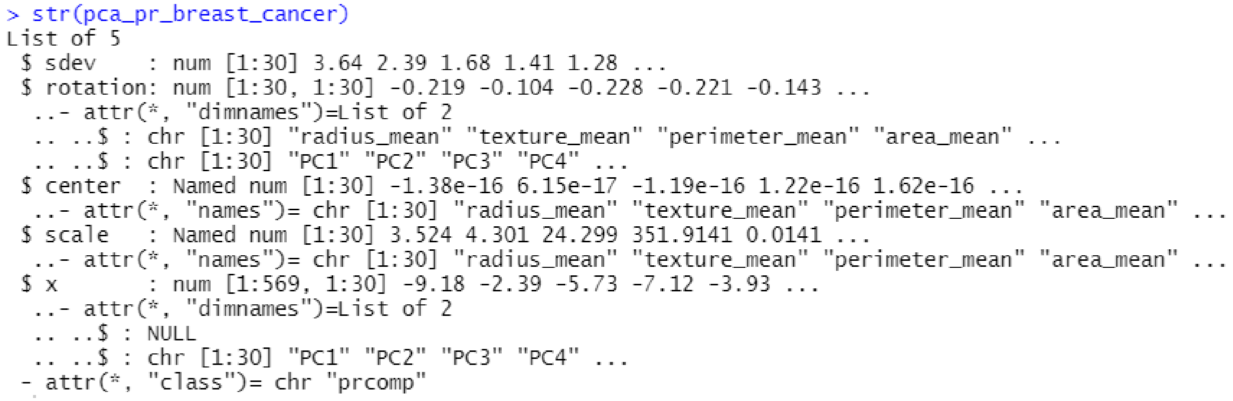
Answer: According to the summary, the first PC explains 44.27% variance, the second PC explains 18.97% variance and the third PC explains 9.393% variance. But the forth and the other PCs explain less variance compared to the first 3 PCs. Therefore, first 3 PCs can be selected as they explain 72.636% cumulative proportion of the total variance.

`Str()` function displays the structure of the PCA output.

Code;

`str(pca_pr_breast_cancer)`

output;



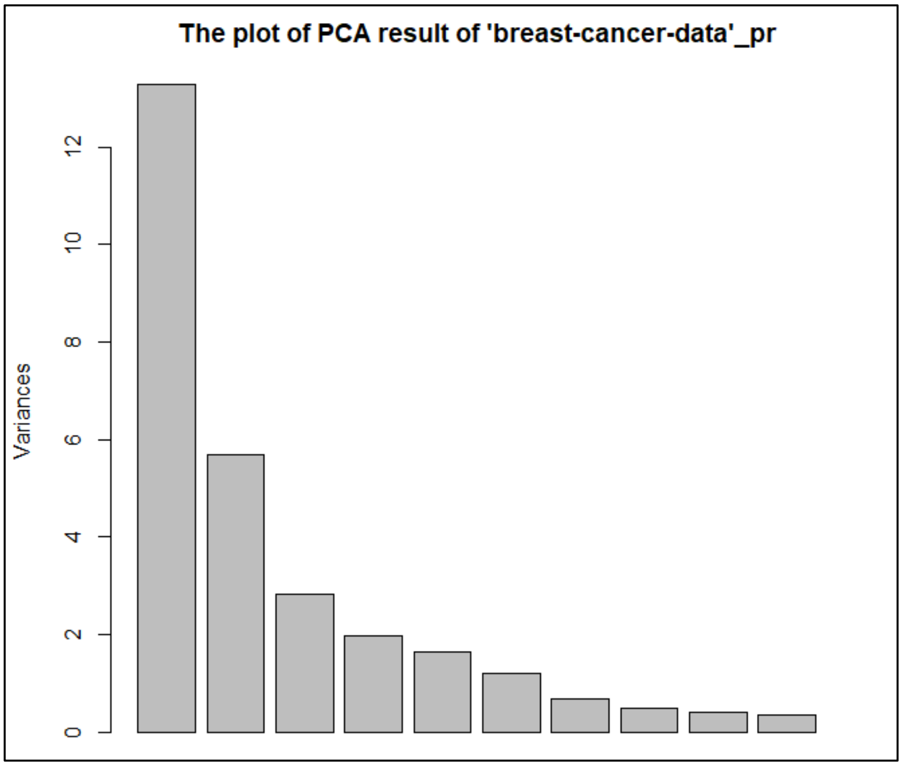
```
> str(pca_pr_breast_cancer)
List of 5
 $ sdev    : num [1:30] 3.64 2.39 1.68 1.41 1.28 ...
 $ rotation: num [1:30, 1:30] -0.219 -0.104 -0.228 -0.221 -0.143 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_mean" ...
  .. ..$ : chr [1:30] "PC1" "PC2" "PC3" "PC4" ...
 $ center  : Named num [1:30] -1.38e-16 6.15e-17 -1.19e-16 1.22e-16 1.62e-16 ...
  ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_mean" ...
 $ scale   : Named num [1:30] 3.524 4.301 24.299 351.9141 0.0141 ...
  ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_mean" ...
 $ x       : num [1:569, 1:30] -9.18 -2.39 -5.73 -7.12 -3.93 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:30] "PC1" "PC2" "PC3" "PC4" ...
 - attr(*, "class")= chr "prcomp"
```

`plot()` function shows the bar chart of PCA result. The `main` attribute specifies the title of the plot. It plots the variances against the principal components.

Code;

```
plot(pca_pr_breast_cancer, main = “The plot of PCA result of ‘breast-cancer-data’_pr”)
```

Output;



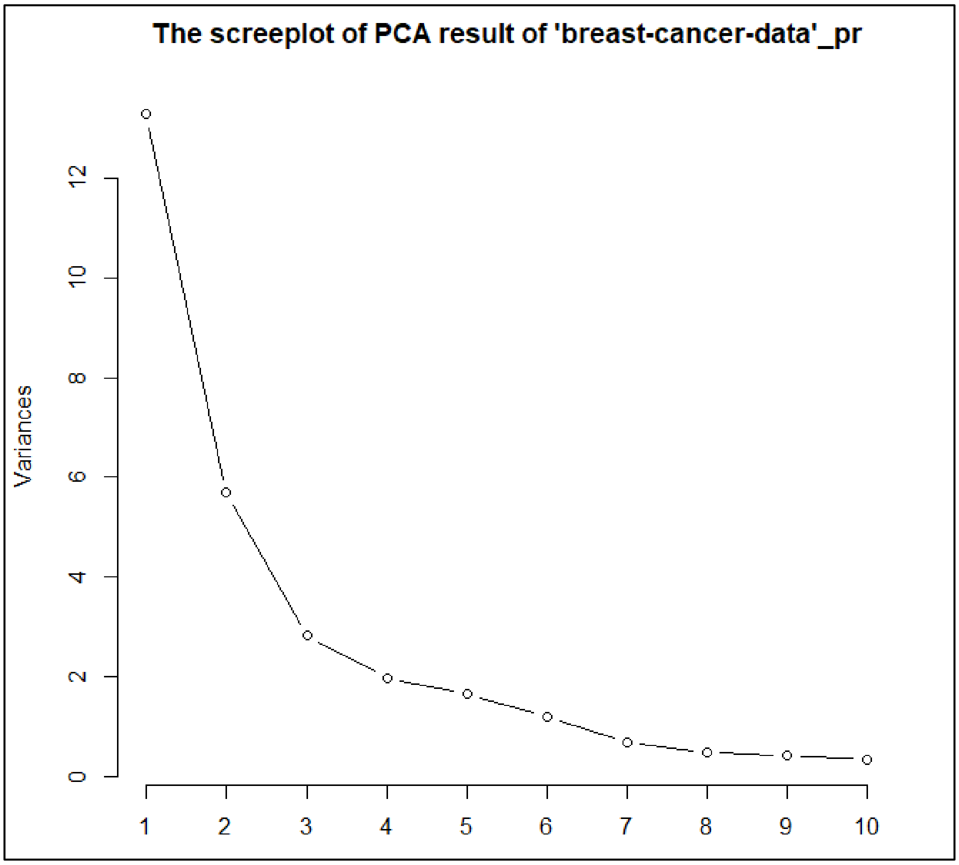
Answer: According to the above bar chart, it shows a drastic change of variances within first 3 PCs. After the third PC the bars of PCs do not show a significant difference. It means the variance explained by those PCs are lesser. Therefore, first 3 PCs can be selected as they visualize drastic change of variance in the above plot.

`Screeplot()` function displays the scree plot/ elbow plot of the PCA output. The `main` attribute specifies the title of the plot. The `type` attribute specifies the type of the plot. Following code results in a line chart because the type is specified as `“lines”.` It plots the variances against the principal components. The elbow point shows the number of PCs with significant variance.

Code;

```
screeplot(pca_pr_breast_cancer, main = “The screeplot of PCA result of
‘breast-cancer-data’_pr”, type = “lines”)
```

Output;



Answer: According to the above scree plot, it shows a drastic change of variance up to 3 PCs. After the third PC the graph looks like flattens. Therefore, the **elbow point** is the **third PC**. It indicates that the first 3 PCs explain a significant variance whereas other PCs do not explain a significant variance. **Three PCs** can be selected according to the above scree plot.

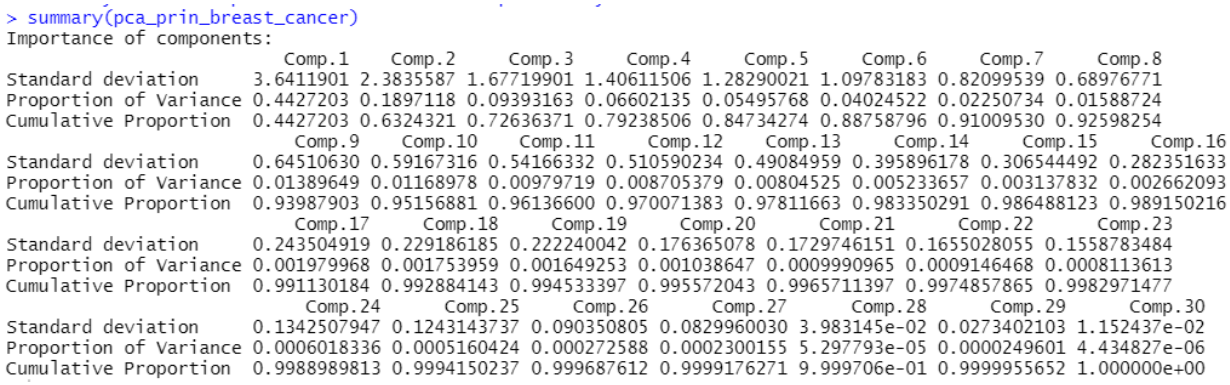
According to the summary, bar chart and the scree plot, **3 PCs would be chosen** as they explain **higher (72.636%) cumulative proportion** of the total variance.

Princomp() method

Code;

```
summary(pca_prin_breast_cancer)
```

Output;

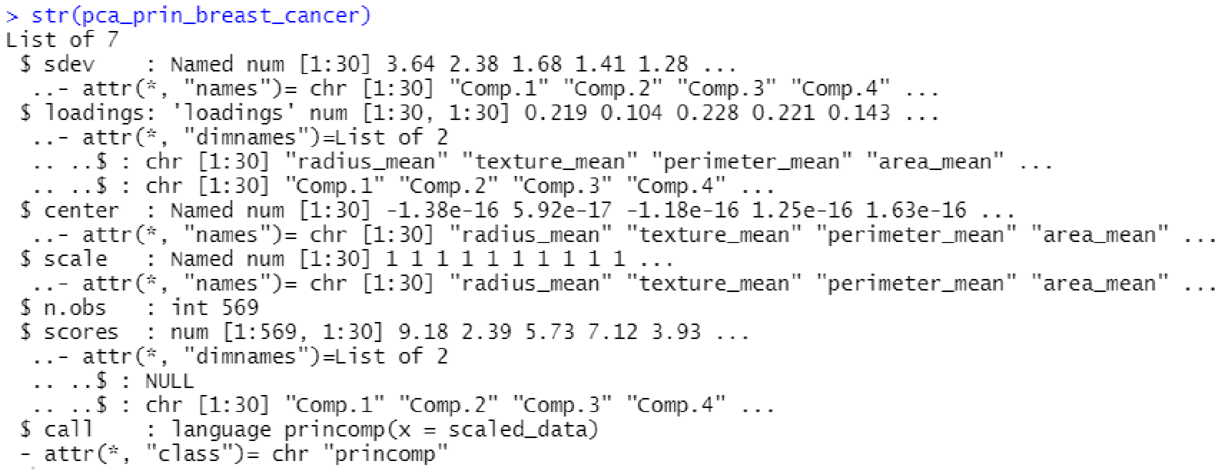


```
> summary(pca_prin_breast_cancer)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     3.6411901 2.3835587 1.67719901 1.40611506 1.28290021 1.09783183 0.82099539 0.68976771
Proportion of Variance 0.4427203 0.1897118 0.09393163 0.06602135 0.05495768 0.04024522 0.02250734 0.01588724
Cumulative Proportion  0.4427203 0.6324321 0.72636371 0.79238506 0.84734274 0.88758796 0.91009530 0.92598254
                           Comp.9    Comp.10    Comp.11     Comp.12    Comp.13     Comp.14     Comp.15     Comp.16
Standard deviation     0.64510630 0.59167316 0.54166332 0.510590234 0.49084959 0.395896178 0.306544492 0.282351633
Proportion of Variance 0.01389649 0.01168978 0.00979719 0.008705379 0.00804525 0.005233657 0.003137832 0.002662093
Cumulative Proportion  0.93987903 0.95156881 0.96136600 0.970071383 0.97811663 0.983350291 0.986488123 0.989150216
                           Comp.17     Comp.18     Comp.19     Comp.20      Comp.21      Comp.22      Comp.23
Standard deviation     0.243504919 0.229186185 0.222240042 0.176365078 0.1729746151 0.1655028055 0.1558783484
Proportion of Variance 0.001979968 0.001753959 0.001649253 0.001038647 0.0009990965 0.0009146468 0.0008113613
Cumulative Proportion  0.991130184 0.992884143 0.994533397 0.995572043 0.9965711397 0.9974857865 0.9982971477
                            Comp.24      Comp.25     Comp.26      Comp.27      Comp.28      Comp.29      Comp.30
Standard deviation     0.1342507947 0.1243143737 0.090350805 0.0829960030 3.983145e-02 0.0273402103 1.152437e-02
Proportion of Variance 0.0006018336 0.0005160424 0.000272588 0.0002300155 5.297793e-05 0.0000249601 4.434827e-06
Cumulative Proportion  0.9988989813 0.9994150237 0.999687612 0.9999176271 9.999706e-01 0.9999955652 1.000000e+00
```

Code;

```
str(pca_prin_breast_cancer)
```

Output;

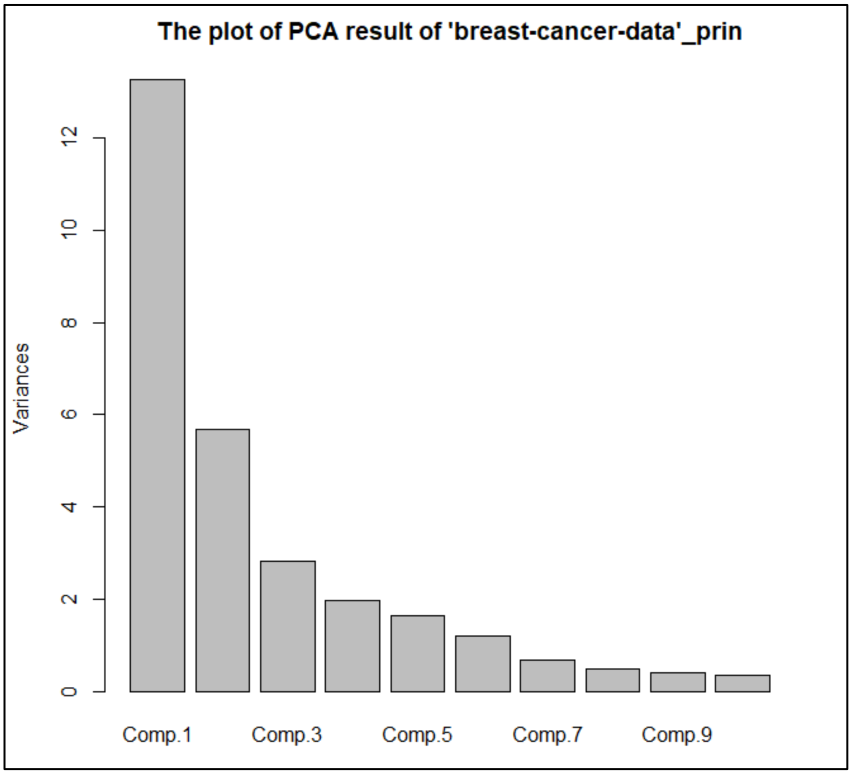


```
> str(pca_prin_breast_cancer)
List of 7
 $ sdev    : Named num [1:30] 3.64 2.38 1.68 1.41 1.28 ...
  ..- attr(*, "names")= chr [1:30] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...
 $ loadings: 'loadings' num [1:30, 1:30] 0.219 0.104 0.228 0.221 0.143 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_mean" ...
  .. ..$ : chr [1:30] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...
 $ center  : Named num [1:30] -1.38e-16 5.92e-17 -1.18e-16 1.25e-16 1.63e-16 ...
  ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_mean" ...
 $ scale   : Named num [1:30] 1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_mean" ...
 $ n.obs   : int 569
 $ scores  : num [1:569, 1:30] 9.18 2.39 5.73 7.12 3.93 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:30] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...
 $ call    : language princomp(x = scaled_data)
 - attr(*, "class")= chr "princomp"
```

Code;

```
plot(pca_prin_breast_cancer, main = “The plot of PCA result of ‘breast-cancer-data’_prin”)
```

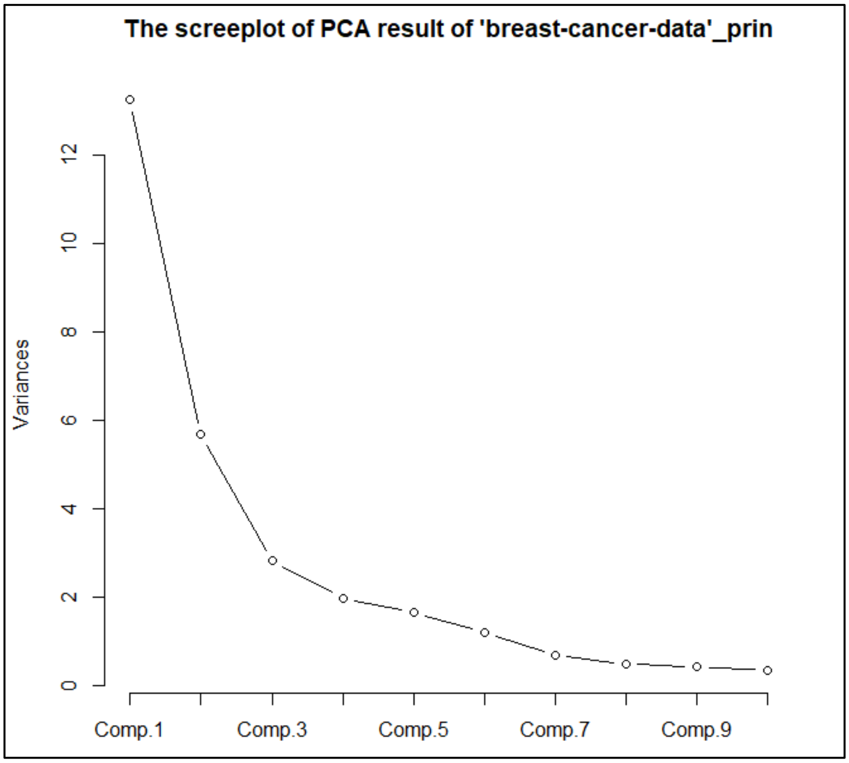
Output;



Code;

```
screeplot(pca_prin_breast_cancer, main = “The screeplot of PCA result of
‘breast-cancer-data’_prin”, type = “lines”)
```

Output;



2.Select the first two principal components and identify whether there are any potential groupings through a suitable visualization technique.

Two visualization techniques are used to plot the data between PC1 and PC2. Those two techniques are bi plot and scatter plot.

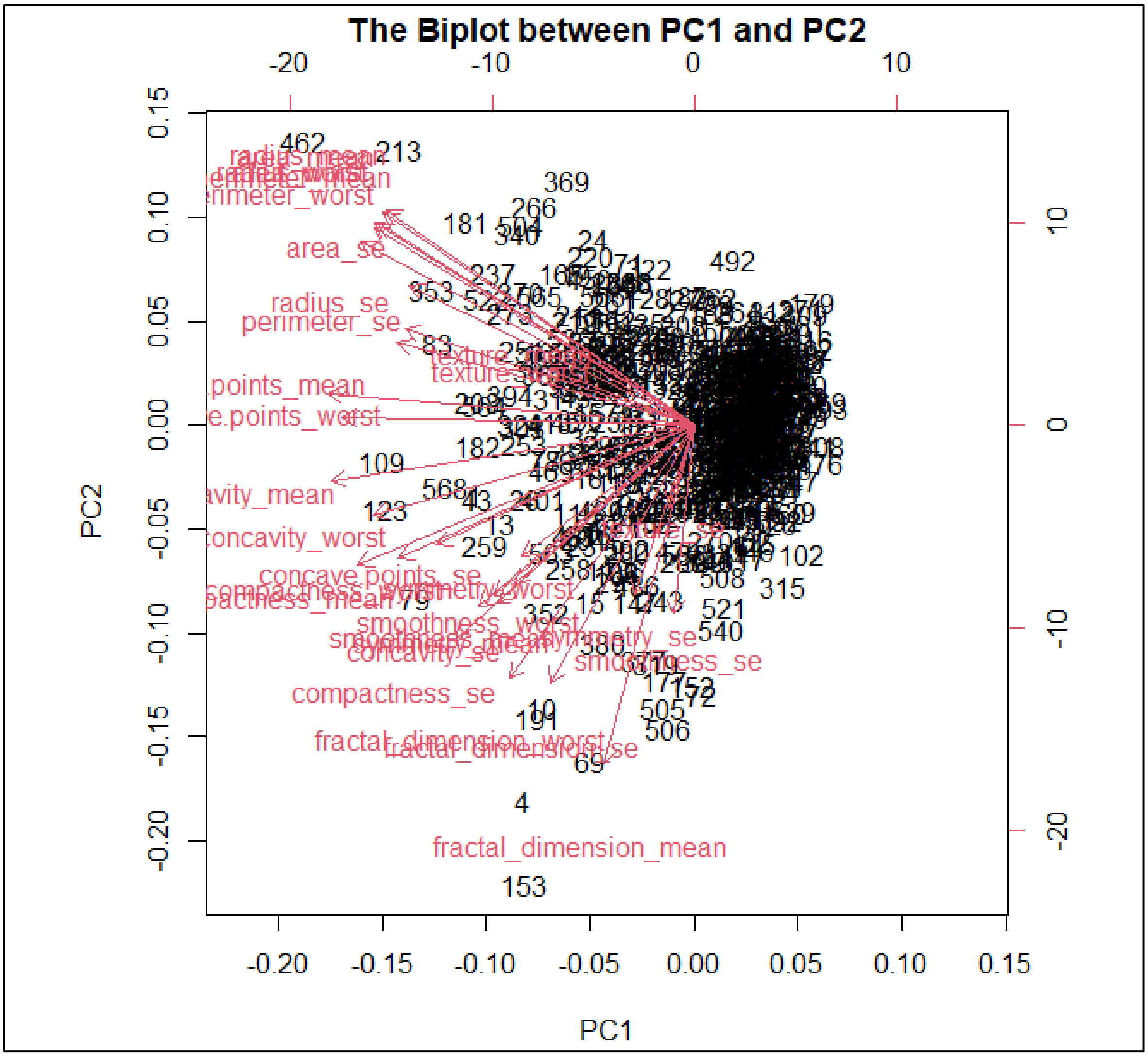
Using Bi plot

The `biplot()` function is used to plot the PCA result. The `main` attribute specifies the title of the biplot.

Code;

```
###without labels
biplot(pca_pr_breast_cancer, main = "The Biplot between PC1 and PC2")
```

Output;



Answer: The above biplot does not visualize clear distinct clusters or groupings. Some data points cluster together to one location and some points are spread out. Clear groupings cannot be identified. It may be due to several reasons such as,

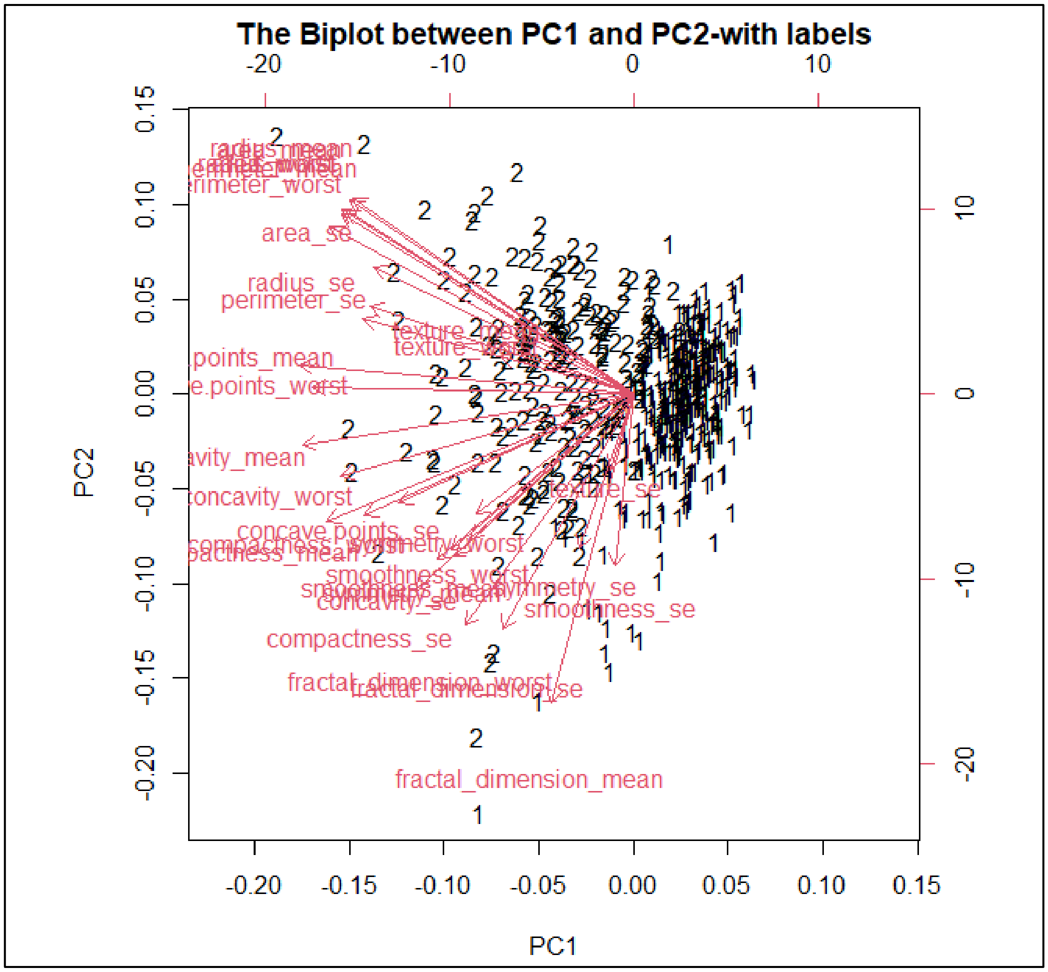
- The complexity of the data set. The linear combinations of variables generated by PCA might not be able to fully reflect the underlying patterns in the original dataset due to the complexity.
- There may not be a robust underlying structure to the variables that can be described by linear combinations.
- Noise or the outliers in the data set.

The following code is used to label the data points according to the categorical variable of the data set. In this case, the bi plot is generated with the labeled data points. Labeling is done using the categorical variable “diagnosis”. That column is extracted, converted to numeric and then used as labels.

Code;

```
###with labels
fac = as.factor(paste(data_breast_cancer[,2]))
biplot(pca_pr_breast_cancer, main = "The Biplot between PC1 and PC2-with labels", xlabs = as.numeric(fac))
```

Output;



Answer: The above plot shows the labeled data points. Data points are divided into 2 clusters as ‘1’ and ‘2’. Cluster 1 is gathered into one point whereas cluster 2 has spread out. Therefore, it is hard to identify 2 clusters without labels.

Using Scatter plot

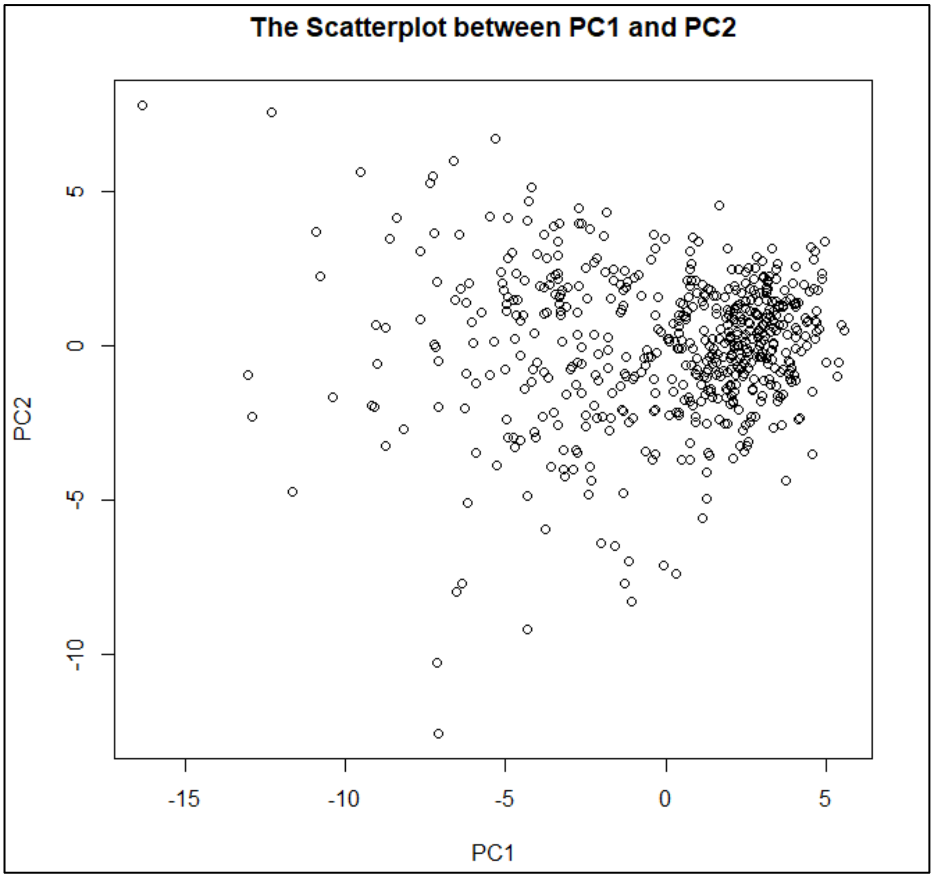
First two PCs are extracted using the element ‘x’ as the PCA result has taken from the prcomp() method. The element ‘scores’ has to be used if the PCA result is from princomp() method. Extracted PC1 and PC2 are used to generate the scatter plot using the plot() function. The main attribute specifies the title of the plot. The x-axis of the plot is labeled as ‘PC1’ and y-axis is labeled as ‘PC2’.

Code;

```
#Select the first two principal components
#Extract PC1
pc1 = pca_pr_breast_cancer$x[,1]
```

```
#Extract PC2
pc2 = pca_pr_breast_cancer$x[,2]

###without labels
plot(pc1, pc2, main = "The Scatterplot between PC1 and PC2", xlab = "PC1", yl
ab = "PC2")
```

Output;



Answer: The above scatterplot does not visualize clear distinct clusters or groupings. Some data points cluster together in one location and some points are spread out. Clear groupings cannot be identified. It may be due to several reasons such as,

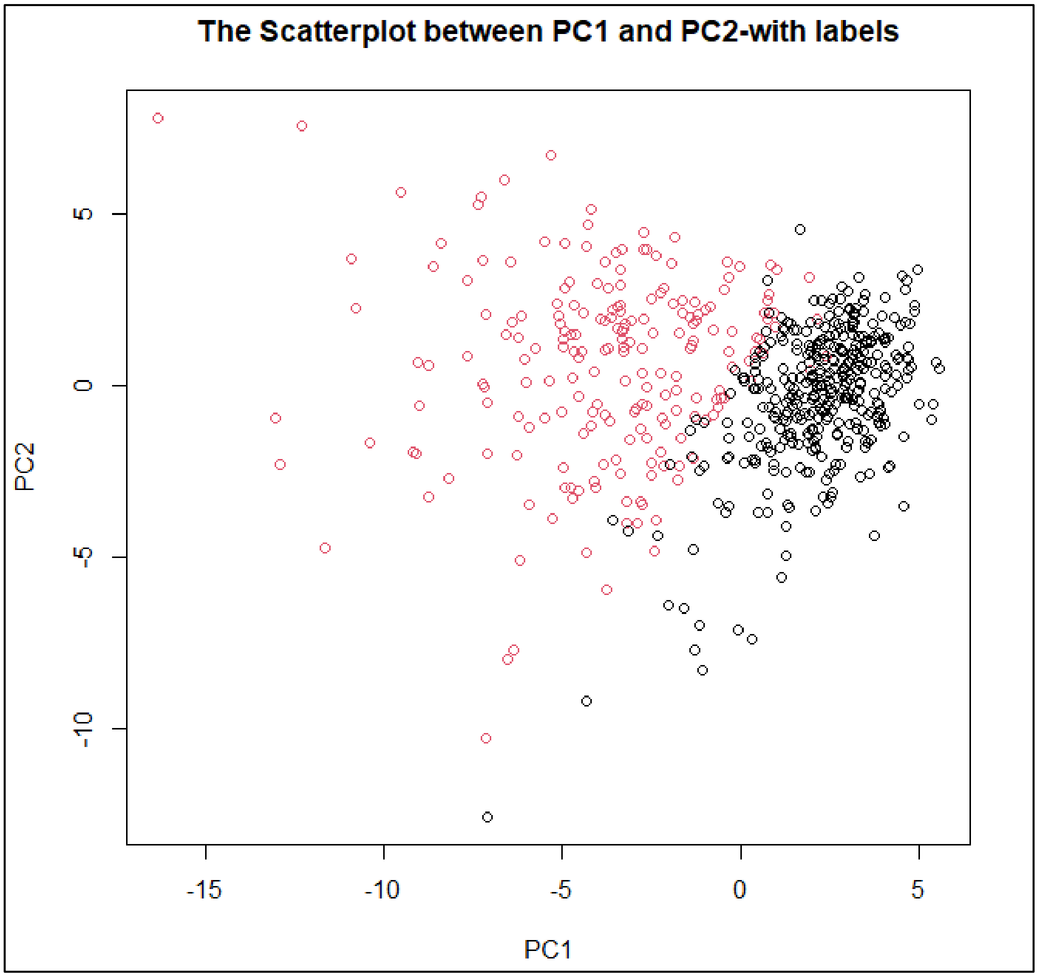
- The complexity of the data set. The linear combinations of variables generated by PCA might not be able to fully reflect the underlying patterns in the original dataset due to the complexity.
- There may not be a robust underlying structure to the variables that can be described by linear combinations.
- Noise or the outliers in the data set.

The following code is used to label the data points according to the categorical variable of the data set. The scatter plot is generated with the labeled data points. Labeling is done using the categorical variable “diagnosis”. That column is extracted and then used as labels.

Code;

```
###with labels
#plot between PC1 and PC2
plot(pc1, pc2, main = "The Scatterplot between PC1 and PC2-with labels", xlab
= "PC1", ylab = "PC2")

#label the points
grps = as.factor(data_breast_cancer$diagnosis)
points(pc1, pc2, col=grps)
```

Output;



Answer: The above scatter plot shows labeled the data points with colors. Data points of each cluster are visualized in ‘black’ and ‘red’. Black color data points are gathered into one point whereas red color data points have spread out. Therefore, it is hard to identify 2 clusters without labels.

How many groups can you identify?

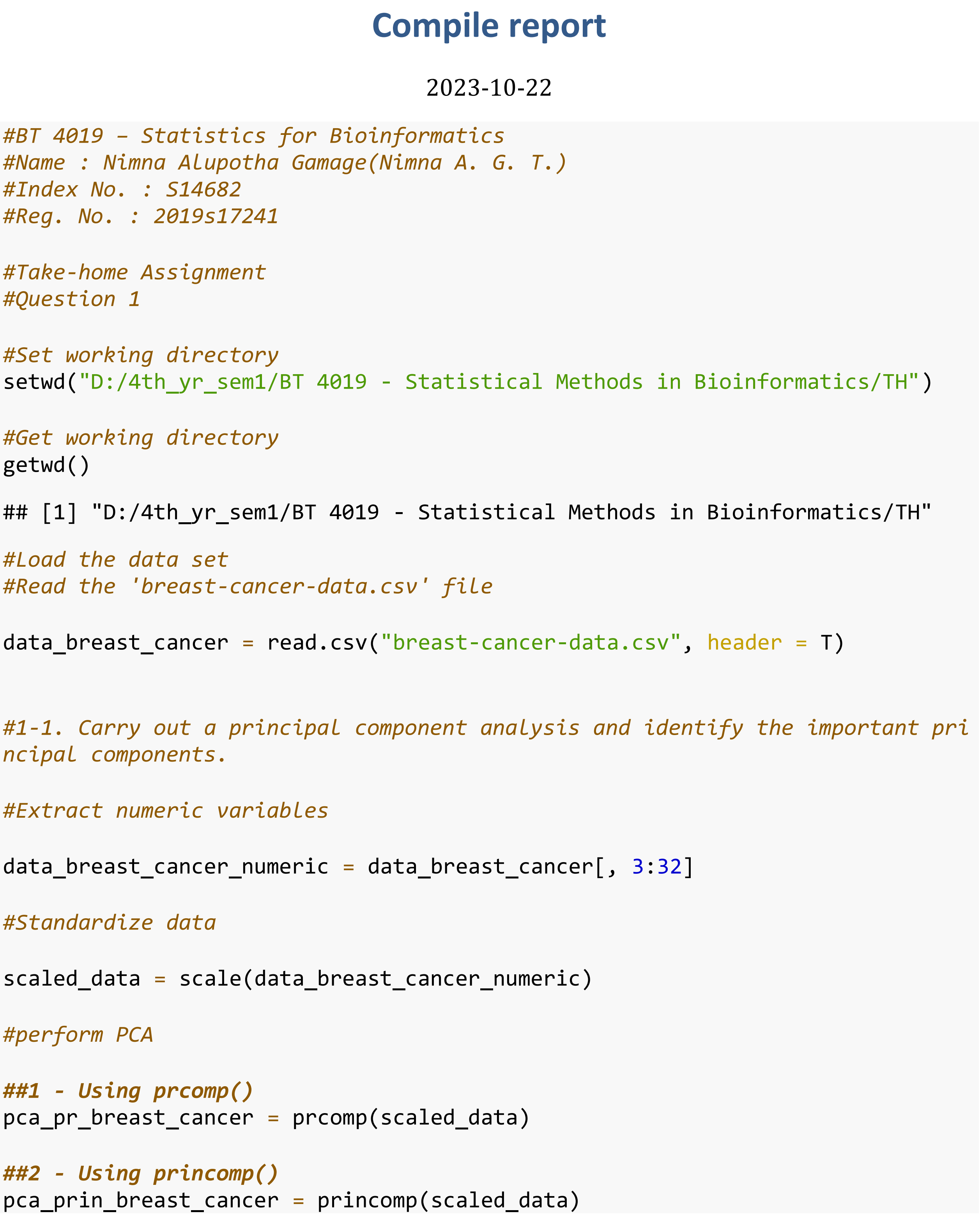
Without labels: If the data points are not labeled, all the points appear as one cluster. Hard to distinguish between distinct clusters due to noise of the data.

With labels: If the data points are labeled using the ‘diagnosis’ variable, **2 clusters/ groups** can be identified. Data points of one cluster are gathered and the data points of the other cluster are spread out.

What are they?

According to the data set (categorical variable “diagnosis”), the two groups are ‘**malignant**’(**M**) and ‘**benign**’ (**B**).

The complete code and the outputs:



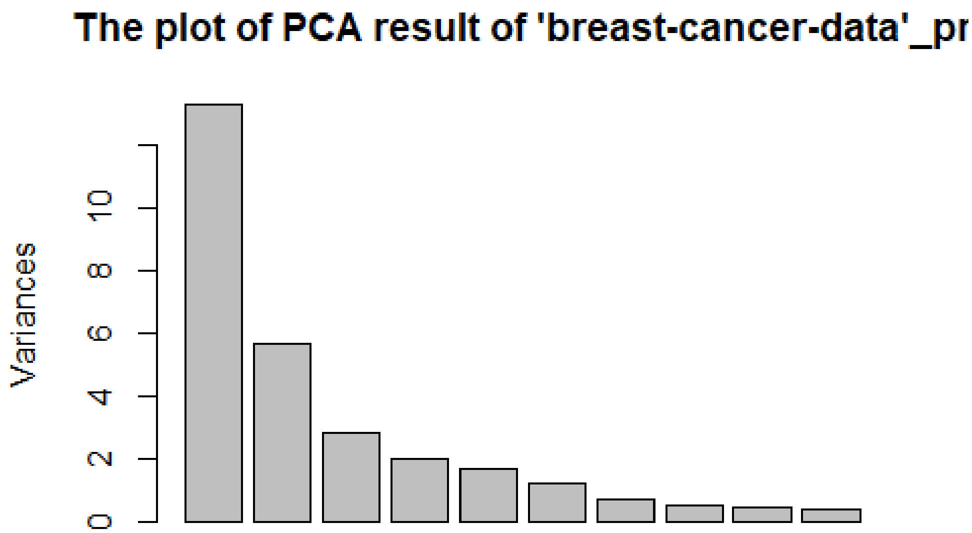
Compile report

2023-10-22

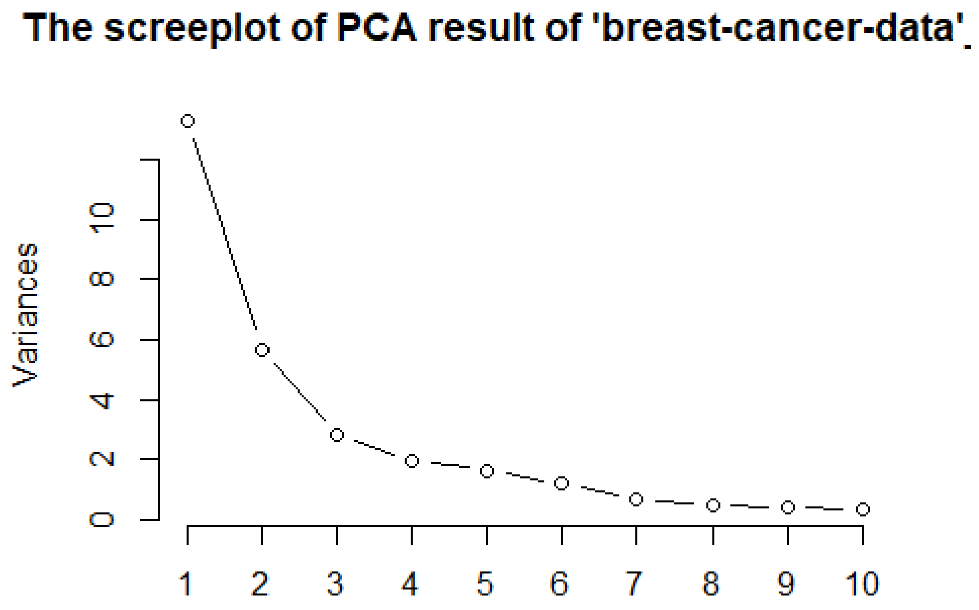
```
#BT 4019 – Statistics for Bioinformatics
#Name : Nimna Alupotha Gamage(Nimna A. G. T.)
#Index No. : S14682
#Reg. No. : 2019s17241

#Take-home Assignment
#Question 1

#Set working directory
setwd("D:/4th_yr_sem1/BT 4019 - Statistical Methods in Bioinformatics/TH")

#Get working directory
getwd()

## [1] "D:/4th_yr_sem1/BT 4019 - Statistical Methods in Bioinformatics/TH"

#Load the data set
#Read the 'breast-cancer-data.csv' file

data_breast_cancer = read.csv("breast-cancer-data.csv", header = T)


#1-1. Carry out a principal component analysis and identify the important pri
ncipal components.

#Extract numeric variables

data_breast_cancer_numeric = data_breast_cancer[, 3:32]

#Standardize data

scaled_data = scale(data_breast_cancer_numeric)

#perform PCA

##1 - Using prcomp()
pca_pr_breast_cancer = prcomp(scaled_data)

##2 - Using princomp()
pca_prin_breast_cancer = princomp(scaled_data)
```

```
#1-2. How many principal components would you choose? Justify your answer.

##1 - Using prcomp()

#summary of the sample-shows the variance explained by each PC
summary(pca_pr_breast_cancer)

## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     P
C7
## Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.821
72
## Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.022
51
## Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.910
10
##                             PC8    PC9    PC10   PC11    PC12    PC13    PC
14
## Standard deviation     0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.396
24
## Proportion of Variance 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.005
23
## Cumulative Proportion  0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.983
35
##                            PC15    PC16    PC17    PC18    PC19    PC20   P
C21
## Standard deviation     0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1
731
## Proportion of Variance 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0
010
## Cumulative Proportion  0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9
966
##                            PC22    PC23   PC24    PC25    PC26    PC27    P
C28
## Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03
987
## Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00
005
## Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99
997
##                            PC29    PC30
## Standard deviation     0.02736 0.01153
## Proportion of Variance 0.00002 0.00000
## Cumulative Proportion  1.00000 1.00000
```

```
#structure of the sample
str(pca_pr_breast_cancer)

## List of 5
##  $ sdev    : num [1:30] 3.64 2.39 1.68 1.41 1.28 ...
##  $ rotation: num [1:30, 1:30] -0.219 -0.104 -0.228 -0.221 -0.143 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_
mean" ...
##   .. ..$ : chr [1:30] "PC1" "PC2" "PC3" "PC4" ...
##  $ center  : Named num [1:30] -1.38e-16 6.15e-17 -1.19e-16 1.22e-16 1.62e-
16 ...
##   ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter
_mean" "area_mean" ...
##  $ scale   : Named num [1:30] 3.524 4.301 24.299 351.9141 0.0141 ...
##   ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter
_mean" "area_mean" ...
##  $ x       : num [1:569, 1:30] -9.18 -2.39 -5.73 -7.12 -3.93 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:30] "PC1" "PC2" "PC3" "PC4" ...
##  - attr(*, "class")= chr "prcomp"

#plot the data - Bar chart
plot(pca_pr_breast_cancer, main = "The plot of PCA result of 'breast-cancer-d
ata'_pr")
```



```
#Scree plot
screeplot(pca_pr_breast_cancer, main = "The screeplot of PCA result of 'breast-cancer-data'_pr", type = "lines")
```



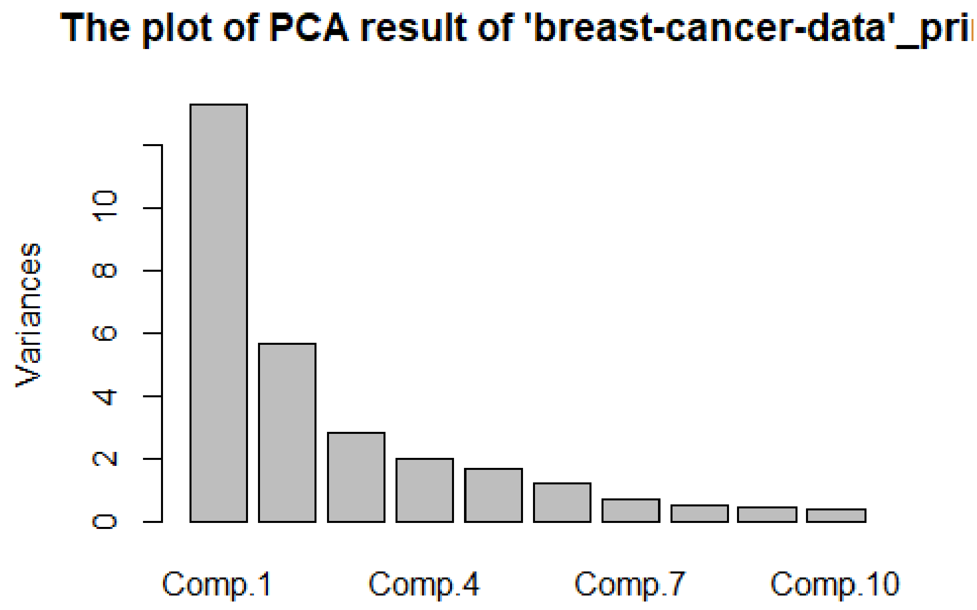
The screeplot of PCA result of 'breast-cancer-data'

```
##2 - Using princomp()

#summary of the sample-shows the variance explained by each PC
summary(pca_prin_breast_cancer)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.
5
## Standard deviation     3.6411901 2.3835587 1.67719901 1.40611506 1.2829002
1
## Proportion of Variance 0.4427203 0.1897118 0.09393163 0.06602135 0.0549576
8
## Cumulative Proportion  0.4427203 0.6324321 0.72636371 0.79238506 0.8473427
4
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp
.10
## Standard deviation     1.09783183 0.82099539 0.68976771 0.64510630 0.59167
316
## Proportion of Variance 0.04024522 0.02250734 0.01588724 0.01389649 0.01168
978
## Cumulative Proportion  0.88758796 0.91009530 0.92598254 0.93987903 0.95156
881
##                           Comp.11     Comp.12    Comp.13     Comp.14
## Standard deviation     0.54166332 0.510590234 0.49084959 0.395896178
## Proportion of Variance 0.00979719 0.008705379 0.00804525 0.005233657
## Cumulative Proportion  0.96136600 0.970071383 0.97811663 0.983350291
##                            Comp.15     Comp.16     Comp.17     Comp.18
## Standard deviation     0.306544492 0.282351633 0.243504919 0.229186185
## Proportion of Variance 0.003137832 0.002662093 0.001979968 0.001753959
## Cumulative Proportion  0.986488123 0.989150216 0.991130184 0.992884143
##                            Comp.19     Comp.20      Comp.21      Comp.22
## Standard deviation     0.222240042 0.176365078 0.1729746151 0.1655028055
## Proportion of Variance 0.001649253 0.001038647 0.0009990965 0.0009146468
## Cumulative Proportion  0.994533397 0.995572043 0.9965711397 0.9974857865
##                             Comp.23      Comp.24      Comp.25     Comp.26
## Standard deviation     0.1558783484 0.1342507947 0.1243143737 0.090350805
## Proportion of Variance 0.0008113613 0.0006018336 0.0005160424 0.000272588
## Cumulative Proportion  0.9982971477 0.9988989813 0.9994150237 0.999687612
##                             Comp.27      Comp.28      Comp.29      Comp.30
## Standard deviation     0.0829960030 3.983145e-02 0.0273402103 1.152437e-02
## Proportion of Variance 0.0002300155 5.297793e-05 0.0000249601 4.434827e-06
## Cumulative Proportion  0.9999176271 9.999706e-01 0.9999955652 1.000000e+00
```

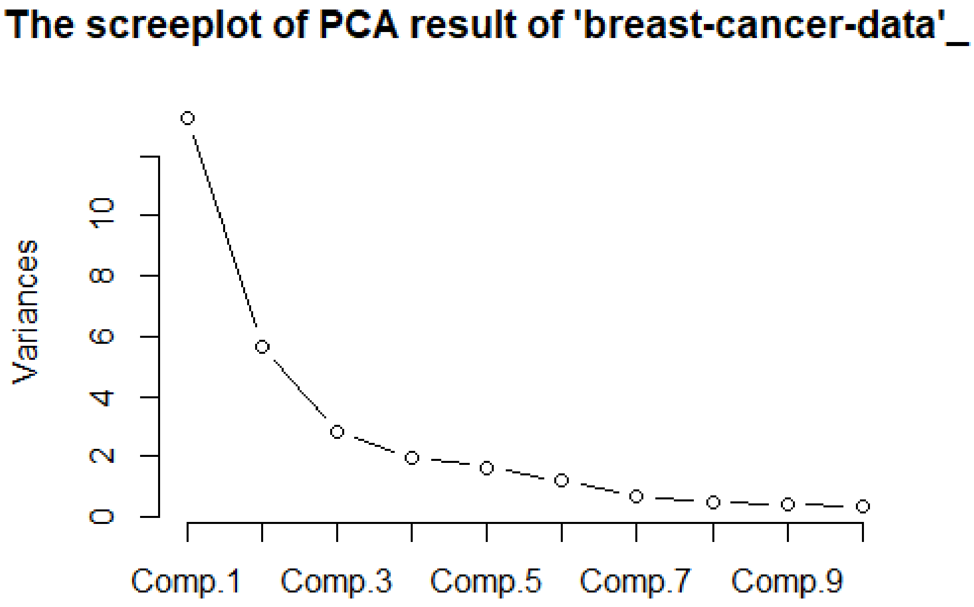
```
#structure of the sample
str(pca_prin_breast_cancer)
```

```
## List of 7
##  $ sdev    : Named num [1:30] 3.64 2.38 1.68 1.41 1.28 ...
##   ..- attr(*, "names")= chr [1:30] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...
##  $ loadings: 'loadings' num [1:30, 1:30] 0.219 0.104 0.228 0.221 0.143 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:30] "radius_mean" "texture_mean" "perimeter_mean" "area_
mean" ...
##   .. ..$ : chr [1:30] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...
##  $ center  : Named num [1:30] -1.38e-16 5.92e-17 -1.18e-16 1.25e-16 1.63e-
16 ...
##   ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter
_mean" "area_mean" ...
##  $ scale   : Named num [1:30] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "names")= chr [1:30] "radius_mean" "texture_mean" "perimeter
_mean" "area_mean" ...
##  $ n.obs   : int 569
##  $ scores  : num [1:569, 1:30] 9.18 2.39 5.73 7.12 3.93 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:30] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...
##  $ call    : language princomp(x = scaled_data)
##  - attr(*, "class")= chr "princomp"
```

```
#plot the data - Bar chart
plot(pca_prin_breast_cancer, main = "The plot of PCA result of 'breast-cancer
-data'_prin")
```

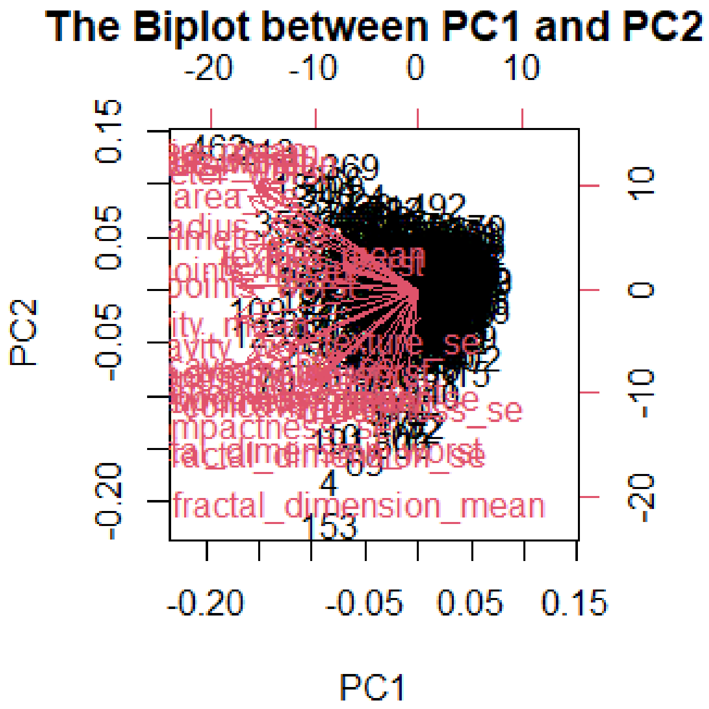


```
#Scree plot
screeplot(pca_prin_breast_cancer, main = "The screeplot of PCA result of 'breast-cancer-data'_prin", type = "lines")
```

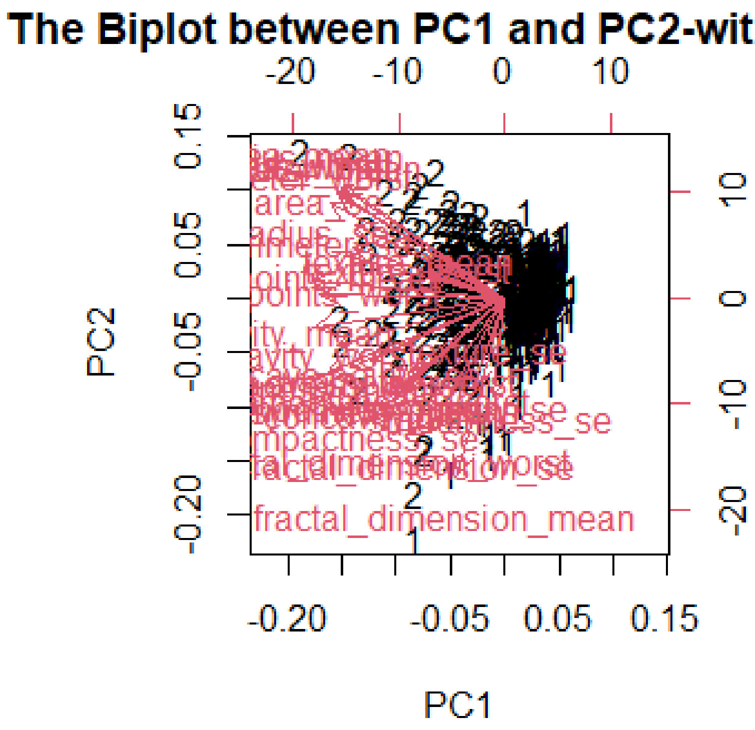


The screeplot of PCA result of 'breast-cancer-data'_

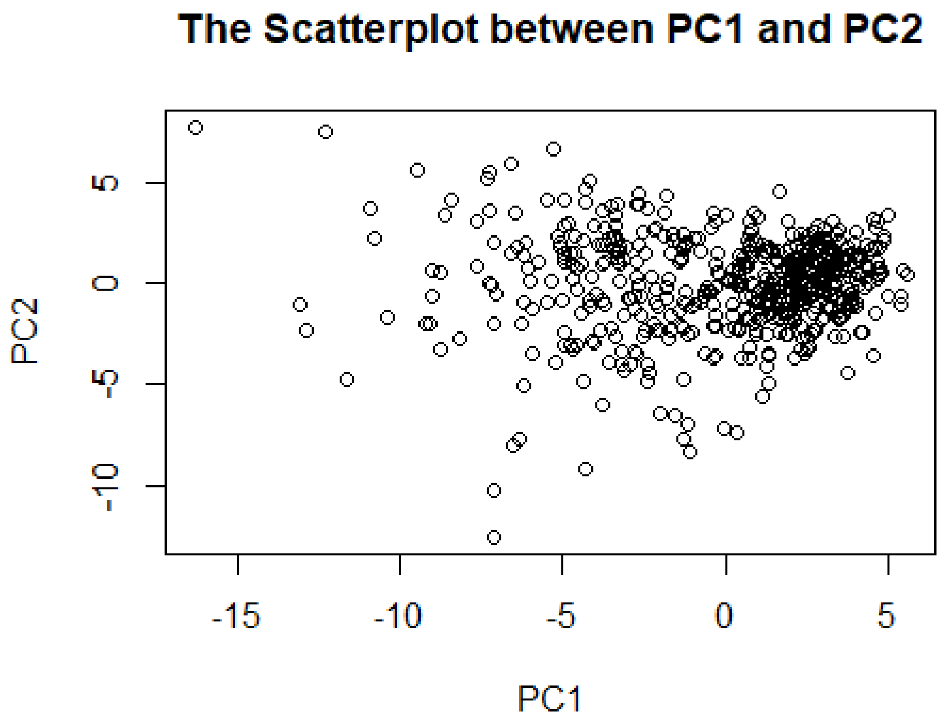
```
#2-1. Select the first two principal components and identify whether there ar
e any potential
#groupings through a suitable visualization technique.
#2-2. How many groups can you identify?
#2-3. What are they?

##1 - Using Bi plot

###without labels
biplot(pca_pr_breast_cancer, main = "The Biplot between PC1 and PC2")
```



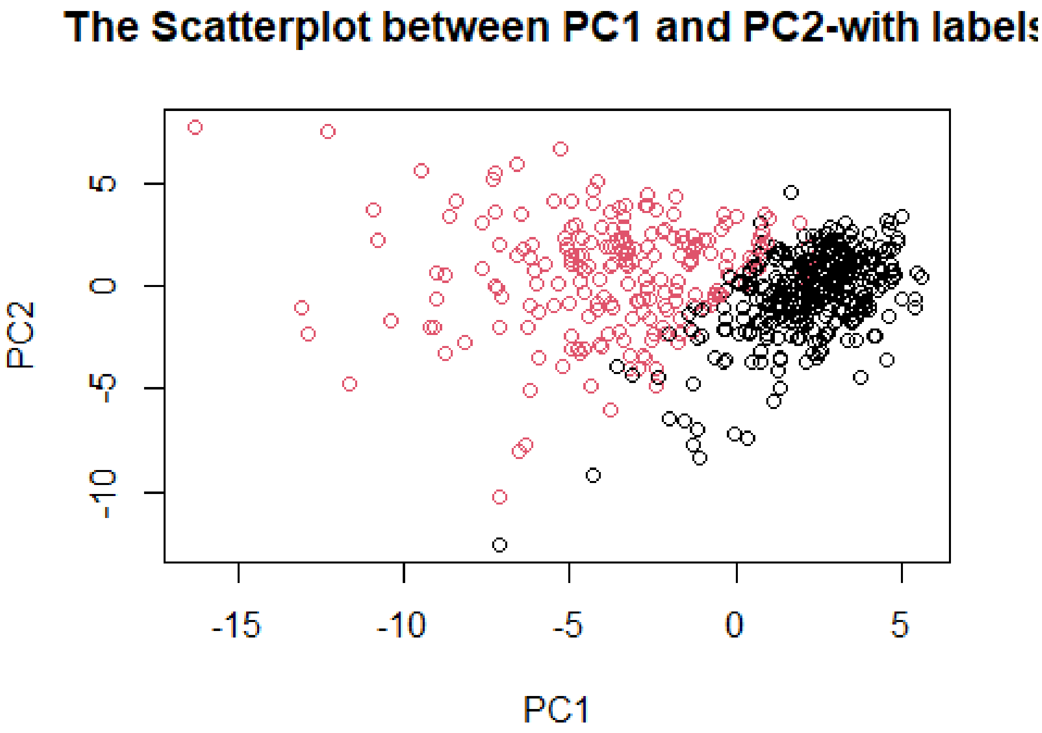
```
###with labels
fac = as.factor(paste(data_breast_cancer[,2]))
biplot(pca_pr_breast_cancer, main = "The Biplot between PC1 and PC2-with labe
ls", xlabs = as.numeric(fac))
```



```
##2 - Using Scatter plot

#Select the first two principal components
#Extract PC1
pc1 = pca_pr_breast_cancer$x[,1]

#Extract PC2
pc2 = pca_pr_breast_cancer$x[,2]

###without labels
plot(pc1, pc2, main = "The Scatterplot between PC1 and PC2", xlab = "PC1", yl
ab = "PC2")
```



```
###with labels
#plot between PC1 and PC2
plot(pc1, pc2, main = "The Scatterplot between PC1 and PC2-with labels", xlab
= "PC1", ylab = "PC2")

#label the points
grps = as.factor(data_breast_cancer$diagnosis)
points(pc1, pc2, col=grps)
```


The Scatterplot between PC1 and PC2-with labels
5
0
-5
-10
PC2
-15
-10
-5
0
5
PC1