MULTIVARIATE ANALYSIS

Breast Cancer Analysis: Malignant vs. Benign Tumors

Introduction

One out of every eight women in the United States is diagnosed with breast cancer. According to the World Health Organization (WHO), 2,3,000,000 women were diagnosed with breast cancer in 2020, resulting in 685,00 fatalities. It is a disease characterized by abnormal breast cell proliferation primarily caused by DNA mutations. Tumors caused by breast cancer are classified as malignant or benign. This classification is utilized to analyze breast tumors, masses, and other breast tissue abnormalities. Cancer classified as benign is generally non-lethal and has a higher survival rate, whereas cancer classified as malignant is lethal. A malignant tumor can develop rapidly, invading the lymph system and invading other healthy tissues in the surrounding area, producing catastrophic effects; in contrast, a benign tumor cannot grow beyond a certain size and remains contained within its mass. Early cancer detection guarantees effective treatment and increases the likelihood of survival.

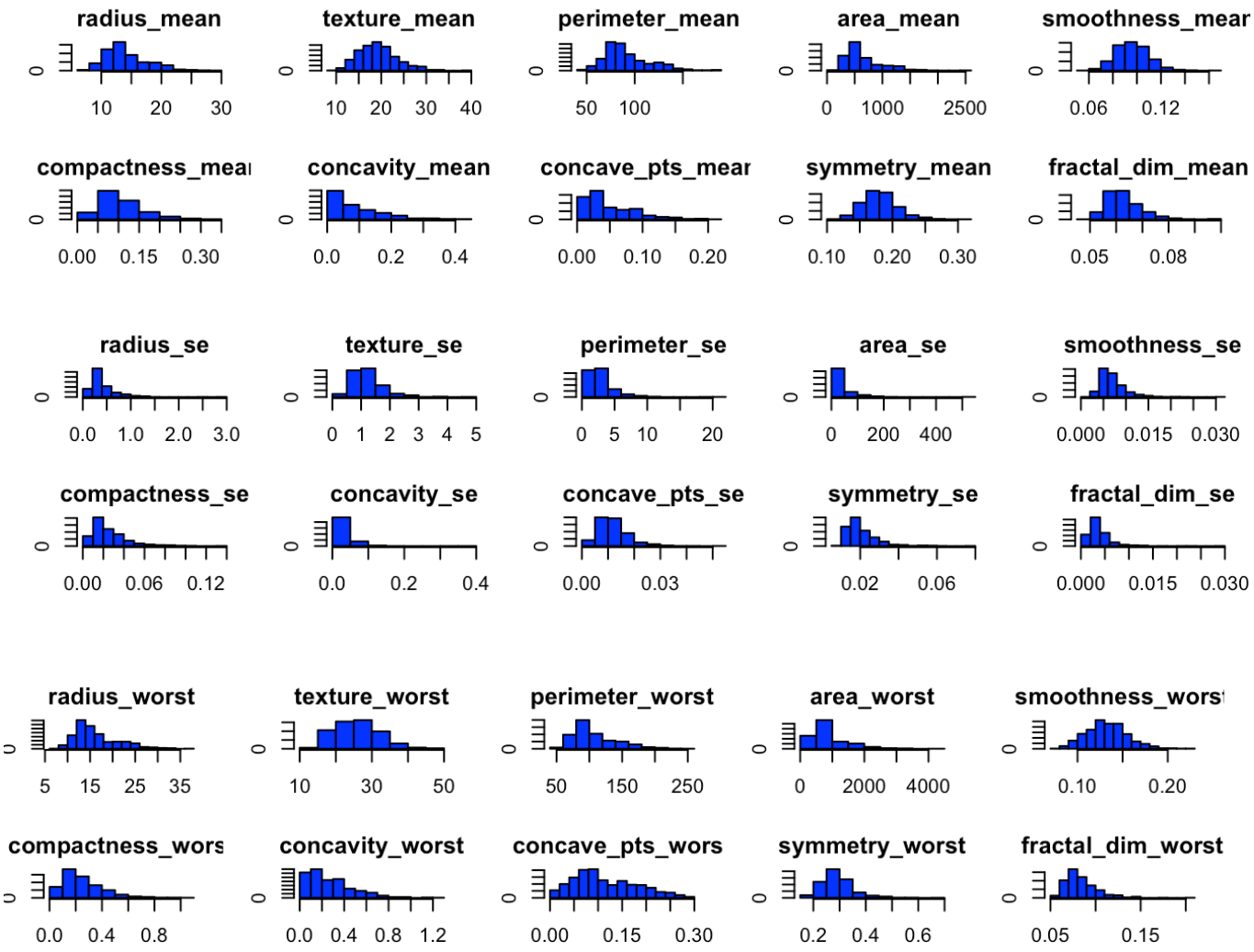
Variables in the dataset:

The WBCD (Wisconsin Breast Cancer Database) dataset contains information about breast cancer tumor samples. It includes measurements from fine needle aspirates of 569 breast mass samples, with 212 malignant and 357 benign diagnoses. The variables in the WBCD dataset are:

- **id:** unique ID number for each sample
- **diagnosis**: categorical variable with values "M" (for malignant) and "B" (for benign)
- **radius_mean**: mean of distances from center to points on the perimeter of the tumor
- **texture_mean**: standard deviation of gray-scale values in the image
- **perimeter_mean**: perimeter of the tumor
- **area_mean:** area of the tumor
- **smoothness_mean:** local variation in radius lengths
- **compactness_mean:** perimeter^2 / area - 1.0
- **concavity_mean:** severity of concave portions of the contour
- **concave_pts_mean:** number of concave portions of the contour
- **symmetry_mean:** symmetry of the tumor
- **fractal_dim_mean:** "coastline approximation" - 1
- **radius_se:** standard error of the mean of distances from center to points on the perimeter
- **texture_se**: standard error of gray-scale values in the image
- **perimeter_se:** standard error of the perimeter
- **area_se:** standard error of the area
- **smoothness_se:** standard error of local variation in radius lengths
- **compactness_se:** standard error of perimeter^2 / area - 1.0
- **concavity_se:** standard error for severity of concave portions of the contour
- **concave_pts_se:** standard error for number of concave portions of the contour
- **symmetry_se:** standard error for symmetry of the tumor
- **fractal_dim_se:** standard error for "coastline approximation" - 1
- **radius_worst:** "worst" or largest mean value for mean of distances from center to points on the perimeter
- **texture_worst:** "worst" or largest mean value for standard deviation of gray-scale values in the image
- **perimeter_worst:** "worst" or largest mean value for perimeter of the tumor
- **area_worst:** "worst" or largest mean value for area of the tumor
- **smoothness_worst:** "worst" or largest mean value for local variation in radius lengths
- **compactness_worst:** "worst" or largest mean value for perimeter^2 / area - 1.0
- **concavity_worst**: "worst" or largest mean value for severity of concave portions of the contour
- **concave_pts_worst:** "worst" or largest mean value for number of concave portions of the contour
- **symmetry_worst:** "worst" or largest mean value for symmetry of the tumor
- **fractal_dim_worst:** "worst" or largest mean value for "coastline approximation" - 1.

Questions for Analysis:

- Are there any missing values or outliers in the data?
- Is the response variable distributed in a balanced manner?
- Can we reduce the dimensionality of the data while preserving the important information?
- What are the most important variables or features in the data?
- Can we identify any relationships or correlations between the factors?
- Can we identify any natural groupings or clusters in the data?
- Can we find a linear combination of variables that can effectively discriminate between different groups or classes in the data?
- How accurate is the classification based on this linear combination of variables?



Some features appear to be normally distributed, such as radius_mean, texture_mean, smoothness_mean, and symmetry_mean. These features have approximately symmetric bell-shaped histograms. Some features are skewed, such as concavity_mean, concave_pts_mean, concavity_worst, and concave_pts_worst. These features have histograms that are not bell-shaped and have a longer tail on one side.

Target	Predictor	Correlation
diagnosis	radius_mean	1.000000000
diagnosis	perimeter_mean	0.997855281
diagnosis	area_mean	0.987357170
diagnosis	radius_worst	0.969538973
diagnosis	perimeter_worst	0.965136514
diagnosis	area_worst	0.941082460
diagnosis	concave_pts_mean	0.822528522
diagnosis	concave_pts_worst	0.744214198
diagnosis	area_se	0.735863663
diagnosis	radius_se	0.679090388
diagnosis	concavity_mean	0.676763550
diagnosis	perimeter_se	0.674171616
diagnosis	concavity_worst	0.526911462
diagnosis	compactness_mean	0.506123578
diagnosis	compactness_worst	0.413462823
diagnosis	concave_pts_se	0.376168956
diagnosis	texture_mean	0.323781891
diagnosis	texture_worst	0.297007644
diagnosis	compactness_se	0.205999980
diagnosis	concavity_se	0.194203623
diagnosis	smoothness_mean	0.170581187
diagnosis	symmetry_worst	0.163953335
diagnosis	symmetry_mean	0.147741242
diagnosis	smoothness_worst	0.119616140
diagnosis	fractal_dim_worst	0.007065886
diagnosis	fractal_dim_se	-0.042641269
diagnosis	texture_se	-0.097317443
diagnosis	symmetry_se	-0.104320881
diagnosis	smoothness_se	-0.222600125
diagnosis	fractal_dim_mean	-0.311630826

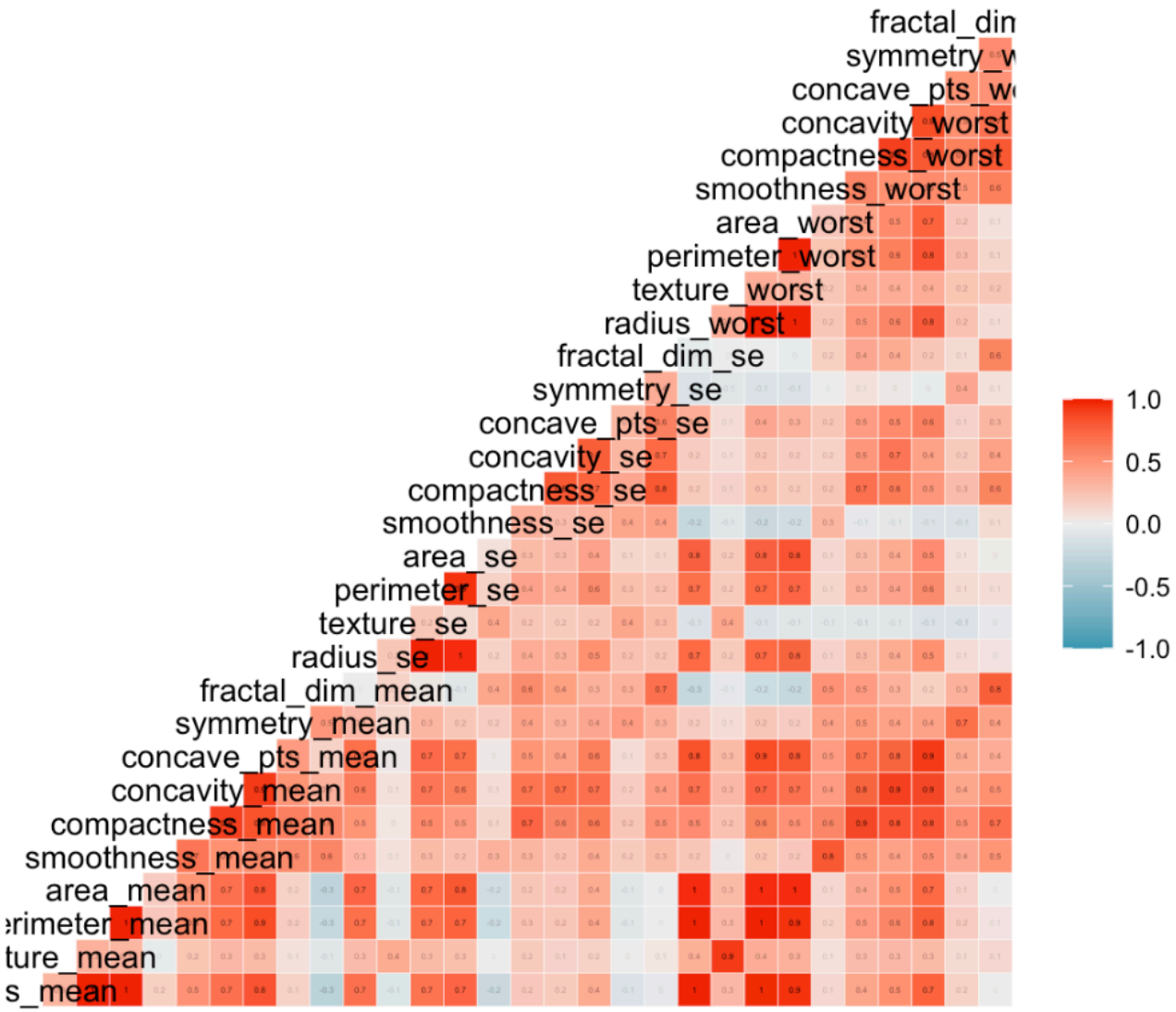
The output shows the correlation values between the target variable, which is the diagnosis of the tumor (M=malignant or B=benign), and each of the predictor variables in the dataset.

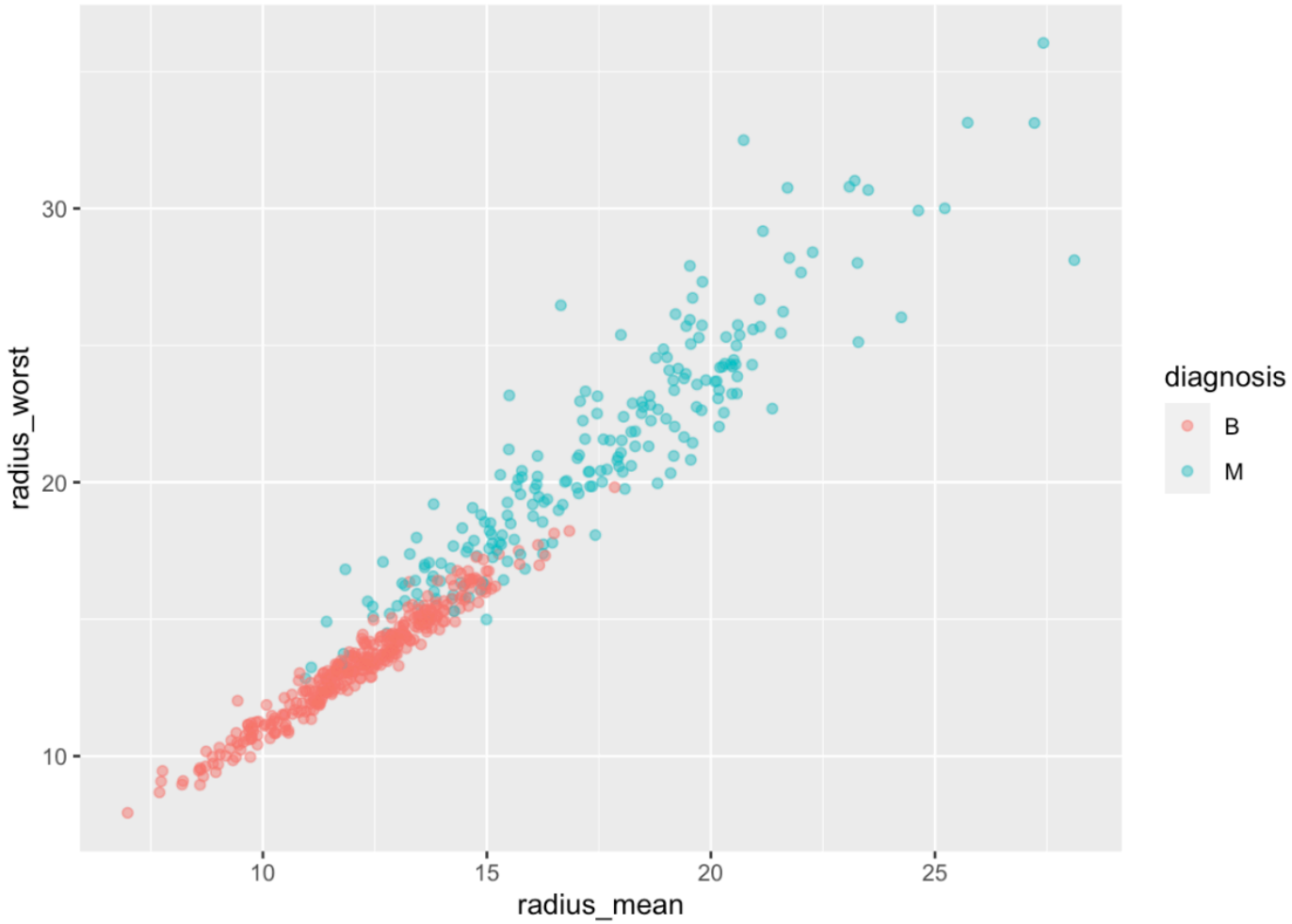
#Based on the output, we can infer that the size-related variables such as radius, perimeter,

and area, are strongly positively correlated with the tumor being malignant. This suggests that larger tumor sizes are more likely to be malignant. Additionally, other variables such as concave points mean and worst, concavity mean and worst, and compactness mean and worst also have a positive correlation with malignancy. On the other hand, variables such as fractal dimension mean and worst, smoothness mean, worst, and se, symmetry mean, worst, and se, and texture mean, worst, and se have a negative correlation with malignancy. This suggests that tumors with lower values for these variables are more likely to be malignant. It is important to note that correlation does not imply causation, and other factors not included in the dataset may also play a role in determining the malignancy of a tumor.

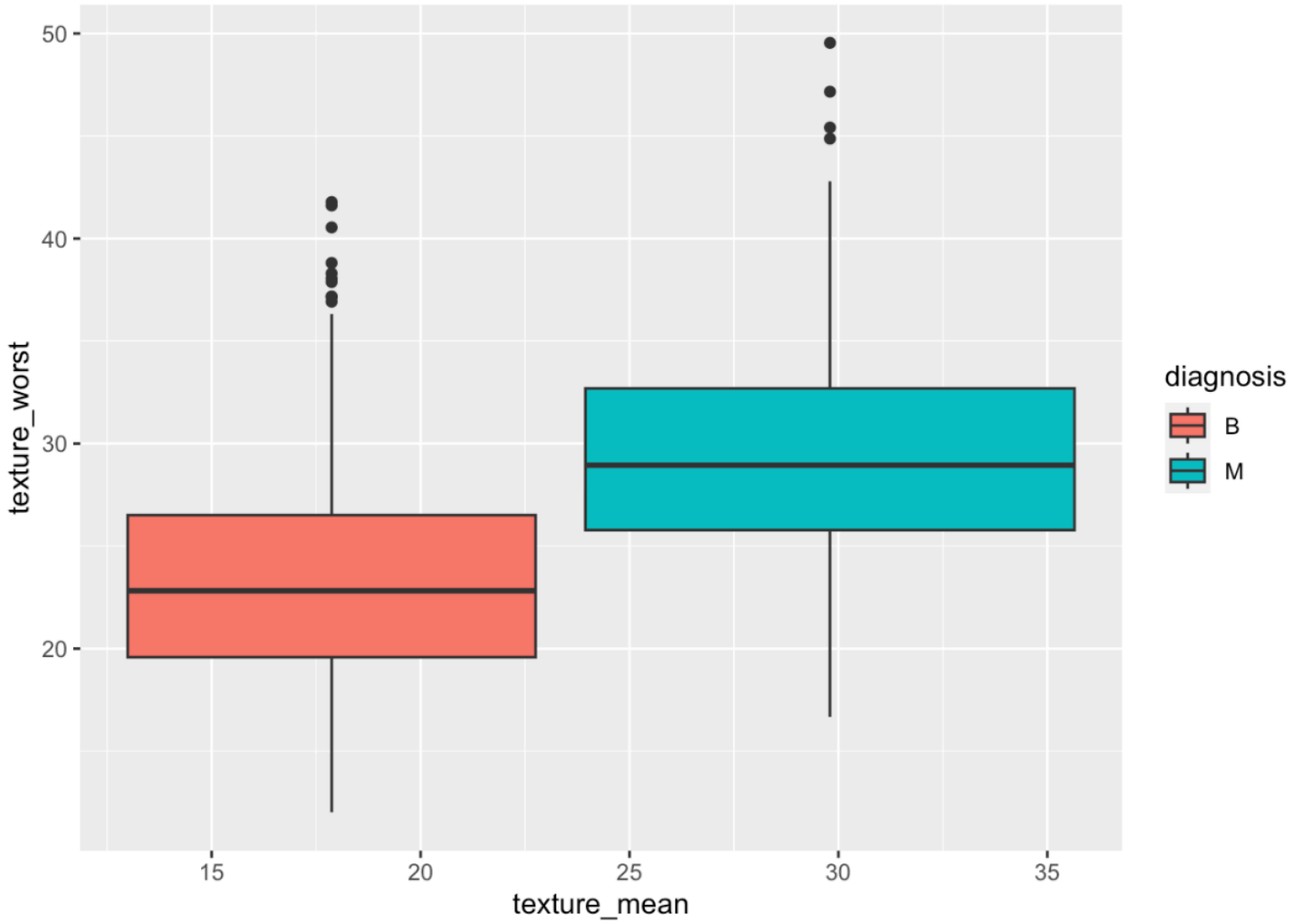
```
##                  row               col       cor
## 1     perimeter_mean       radius_mean 0.9978553
## 2          area_mean       radius_mean 0.9873572
## 3       radius_worst       radius_mean 0.9695390
## 4    perimeter_worst       radius_mean 0.9651365
## 5         area_worst       radius_mean 0.9410825
## 6      texture_worst      texture_mean 0.9120446
## 7        radius_mean    perimeter_mean 0.9978553
## 8          area_mean    perimeter_mean 0.9865068
## 9       radius_worst    perimeter_mean 0.9694764
## 10   perimeter_worst    perimeter_mean 0.9703869
## 11        area_worst    perimeter_mean 0.9415498
## 12       radius_mean         area_mean 0.9873572
## 13    perimeter_mean         area_mean 0.9865068
## 14      radius_worst         area_mean 0.9627461
## 15   perimeter_worst         area_mean 0.9591196
## 16        area_worst         area_mean 0.9592133
## 17  concave_pts_mean    concavity_mean 0.9213910
## 18    concavity_mean  concave_pts_mean 0.9213910
## 19 concave_pts_worst  concave_pts_mean 0.9101553
## 20      perimeter_se         radius_se 0.9727937
## 21           area_se         radius_se 0.9518301
## 22         radius_se      perimeter_se 0.9727937
## 23           area_se      perimeter_se 0.9376554
## 24         radius_se           area_se 0.9518301
## 25      perimeter_se           area_se 0.9376554
## 26       radius_mean      radius_worst 0.9695390
## 27    perimeter_mean      radius_worst 0.9694764
## 28         area_mean      radius_worst 0.9627461
## 29   perimeter_worst      radius_worst 0.9937079
## 30        area_worst      radius_worst 0.9840146
## 31      texture_mean     texture_worst 0.9120446
## 32       radius_mean   perimeter_worst 0.9651365
## 33    perimeter_mean   perimeter_worst 0.9703869
## 34         area_mean   perimeter_worst 0.9591196
## 35      radius_worst   perimeter_worst 0.9937079
## 36        area_worst   perimeter_worst 0.9775781
## 37       radius_mean        area_worst 0.9410825
## 38    perimeter_mean        area_worst 0.9415498
## 39         area_mean        area_worst 0.9592133
## 40      radius_worst        area_worst 0.9840146
## 41   perimeter_worst        area_worst 0.9775781
## 42  concave_pts_mean concave_pts_worst 0.9101553
```

The same has been represented in the correlation plot below:


fractal_din
symmetry_w
concave_pts_w
concavity_worst
compactness_worst
smoothness_worst
area_worst
perimeter_worst
texture_worst
radius_worst
fractal_dim_se
symmetry_se
concave_pts_se
concavity_se
compactness_se
smoothness_se
area_se
perimeter_se
texture_se
radius_se
fractal_dim_mean
symmetry_mean
concave_pts_mean
concavity_mean
compactness_mean
smoothness_mean
area_mean
erimeter_mean
ture_mean
s_mean
1.0
0.5
0.0
-0.5
-1.0

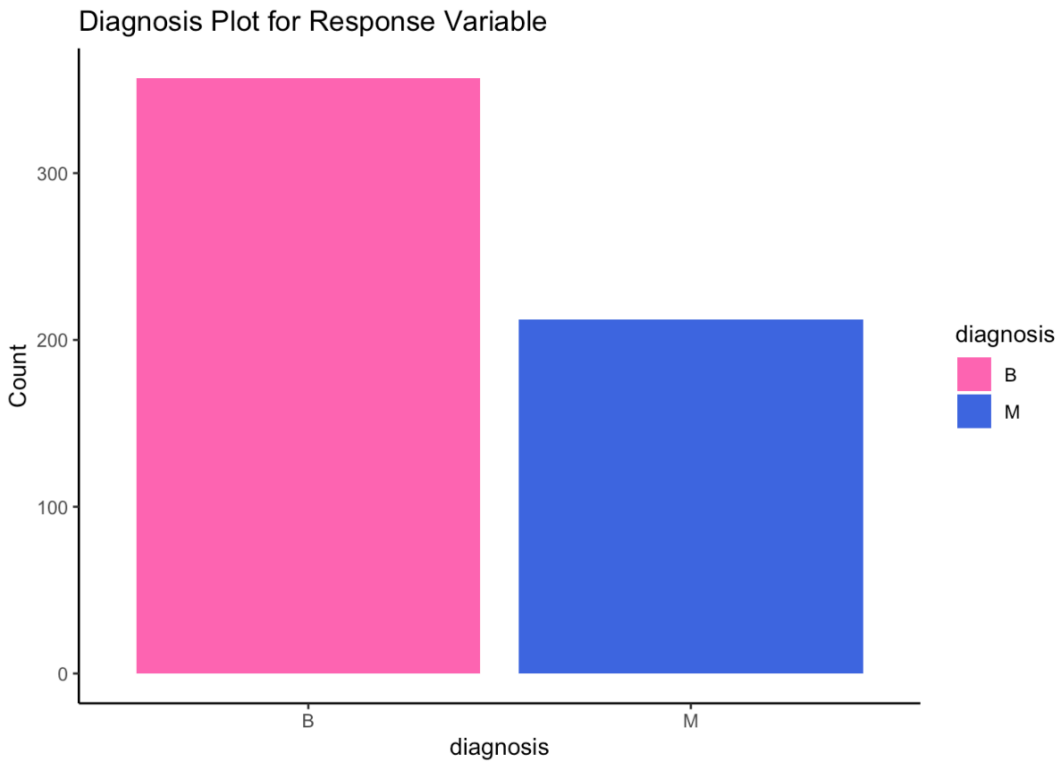


This gives us some helpful information. First, it tells us that the average radius for benign tumors is lower than the average radius for malignant tumors.Secondly, it shows us that there is some overlap where we could potentially misdiagnose the tumors if these were the only features measured.



This shows us that, on average, benign tumors have lower values of both texture mean and texture worst measurements. It also shows us that some samples would cause an error in classification if we did not know their diagnosis in advance. So, we need to dig a little deeper and see if we can get a better delineation of classes (benign or malignant) through principal component analysis. Most histograms present very asymmetric behavior with similar to exponential distribution. Some predictor look like exponential distribuition as radius_se, perimeter_se, area_se, concavity_se and fractal_dimension_se. There is no true outliers, the outliers at box-plot is due the kind of distribution. There is no true outliers, the outliers at box-plot is due the kind of distribution. There is no missingvalues or NAs. Identified 21 pairs of highly correlated predictors, r> 0.9, this was due to the choice of predictors that are associated, measures things related: radius, perimeter and area.

There are 14 predictors related with the response, Diagnosis, with r>=0.6, which is good.



The response variable looks slightly unbalanced.It may indicate that there is some bias in the data or that the study was not designed to produce a balanced sample. This could potentially affect the results of any statistical analyses performed on the data, as the sample may not be representative of the population of interest.

Principal Component Analysis (PCA):

PCA is an unsupervised learning method which serves as an dimension reduction method as well as an exploratory data analysis tool.

Since our dataset is quite high dimensional, it is reasonable and useful to apply PCA and try to reduce dimensionality. Moreover, PCA also gives an idea about whether the data is suitable for clustering.

Before applying PCA, data was scaled and standardized. Variables with very large/low variance can dominate others and appear like more important in explaning the variance, when it is not.

In R, prcomp function was used for principal component analysis. This function uses correlation matrix and the outputs returned by this function are as follow:

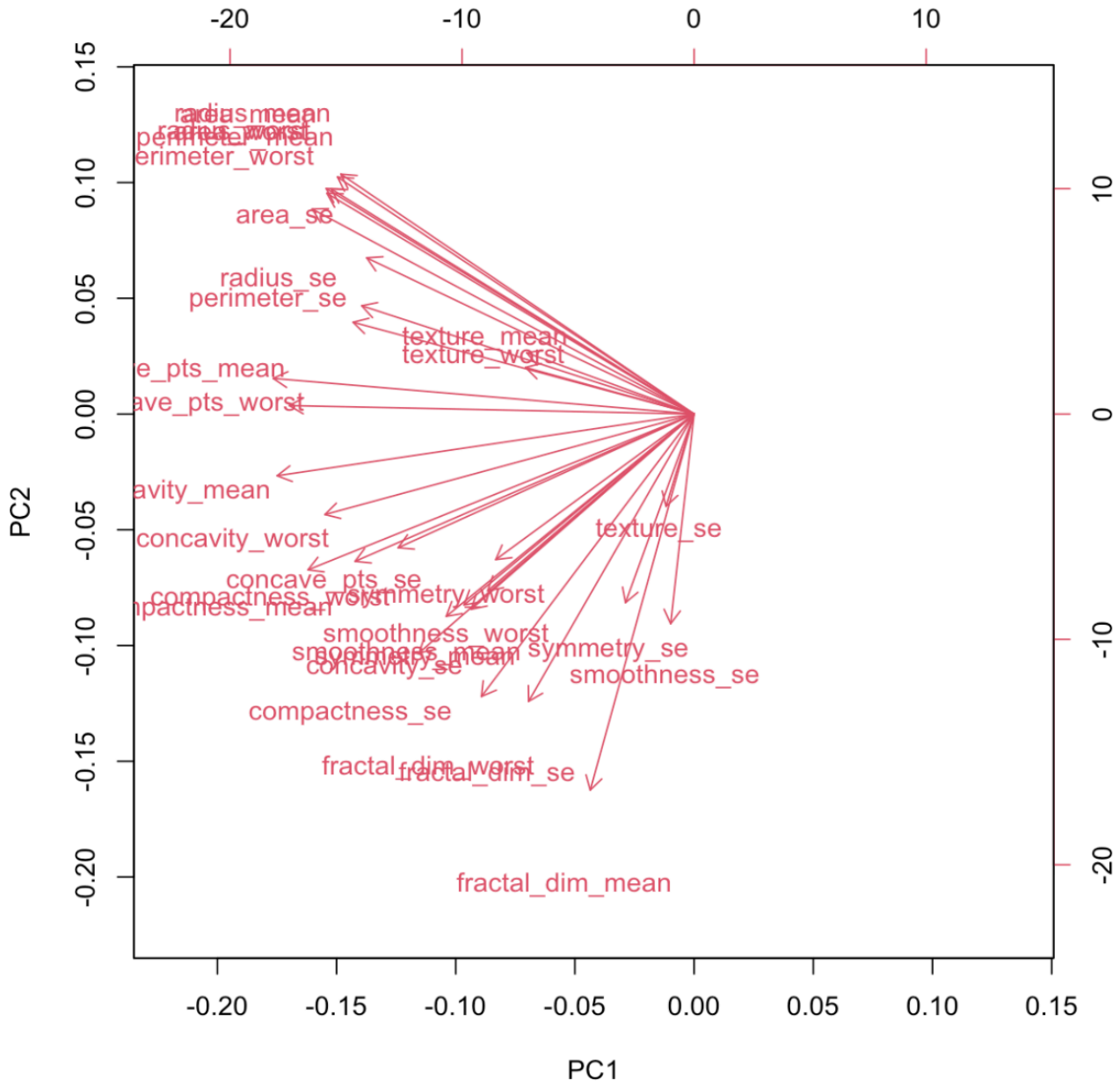
- Sdev - the standard deviations of PCs

- Rotation – loadings (eigenvectors)

- Center - variable mean

- Scale - variable standard deviations

- X – scores

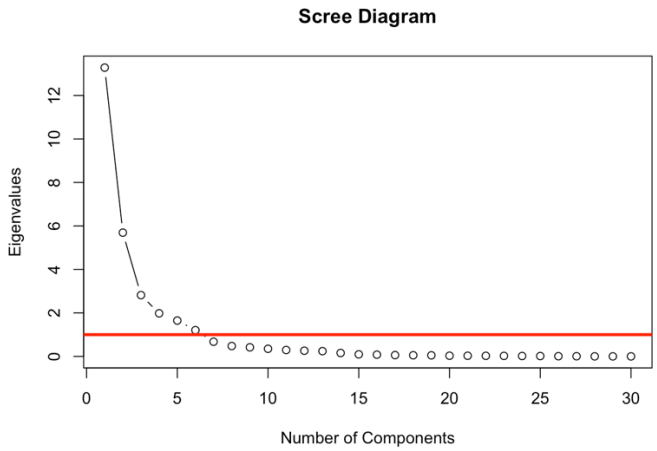


Biplots bring together loadings and scores. Variables are represented by the vectors . the biplot can provide insights into the relationships between the observations and the variables in the dataset. Points that are close together in the biplot are similar, and variables that are close to each other are strongly correlated. The direction of an arrow indicates the direction of increasing values for that variable. Therefore, if two arrows are pointing in roughly the same direction, they are positively correlated, and if they are pointing in roughly opposite directions, they are negatively correlated. In other words, influence that a vector has on a

principal component is large when it is further away the PC's origin.

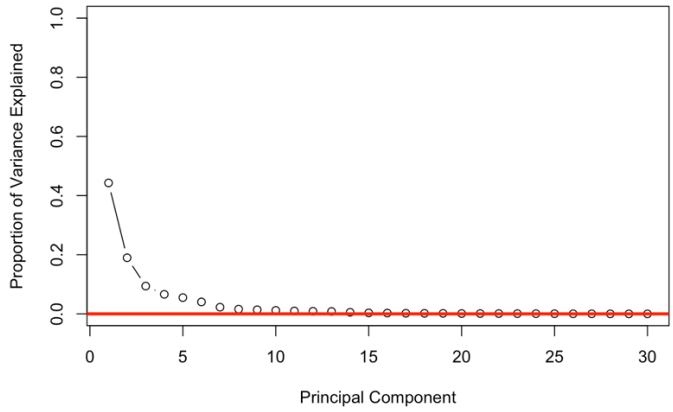
```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172
## Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251
## Cumulative Proportion  0.4427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

The output above shows the importance of each principal component in decreasing order of importance. The standard deviation of each principal component shows how much of the variance in the data is explained by that component. The proportion of variance shows the fraction of the total variance in the data explained by each component. The cumulative proportion shows the total fraction of the variance explained by each component up to that point. The output shows that the first principal component (PC1) is the most important, with a standard deviation of 3.6444 and explaining 44.27% of the total variance. The second component (PC2) is the second most important, with a standard deviation of 2.3857 and explaining 18.97% of the variance. The cumulative proportion shows that the first two components explain 63.24% of the total variance. As we move down the list of components, each subsequent component explains less variance than the previous one. The final components explain very little variance, with the 30th component explaining only 0.00000% of the total variance.



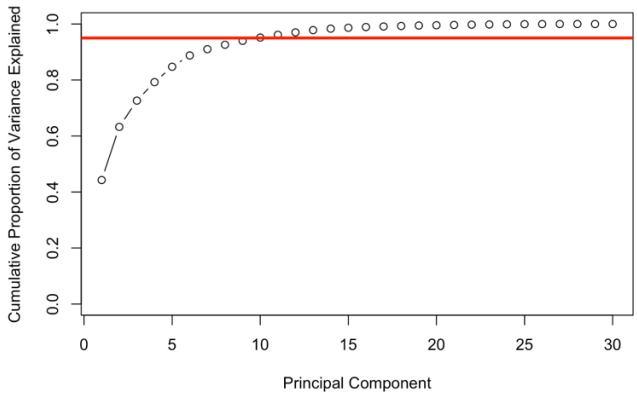
The eigenvalues represent the amount of variance explained by each principal component. In general, we want to retain the principal components with high eigenvalues and drop the ones with low eigenvalues.

The scree plot helps in identifying the "elbow" or the point after which the eigenvalues start to level off. This point indicates the number of principal components that should be retained for further analysis.

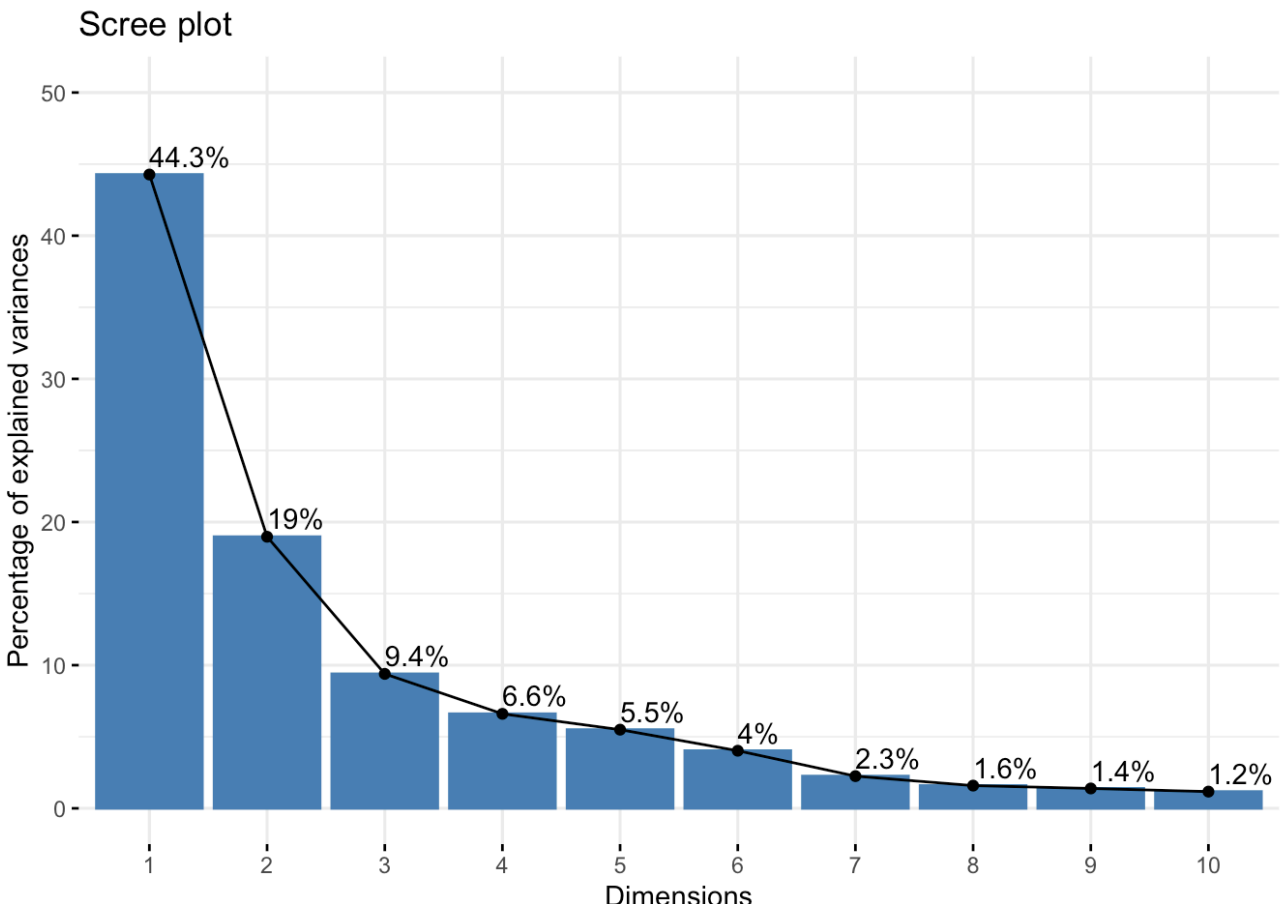


Inference from scree plot can be made by looking at the "elbow" point in the plot, which is the point where the rate of decrease in the proportion of variance explained starts to slow down. This point is generally considered as a reasonable estimate of the number of components to retain. Typically, we retain components that are above the elbow point, as these components capture a significant amount of information

from the data.



Inference from the scree plot with the number of components vs. the cumulative proportion of variance explained can be made by looking at the steepness of the curve. The steeper the curve, the more the variance is explained by each additional component. When the curve starts to flatten, it indicates that additional components will not explain much more of the variance. Therefore, the point at which the curve flattens is a good indication of the number of components to retain.



From all the scree plots, we can infer that adding any more than 7 components does not add much. Value to the analysis.

Cluster Analysis:

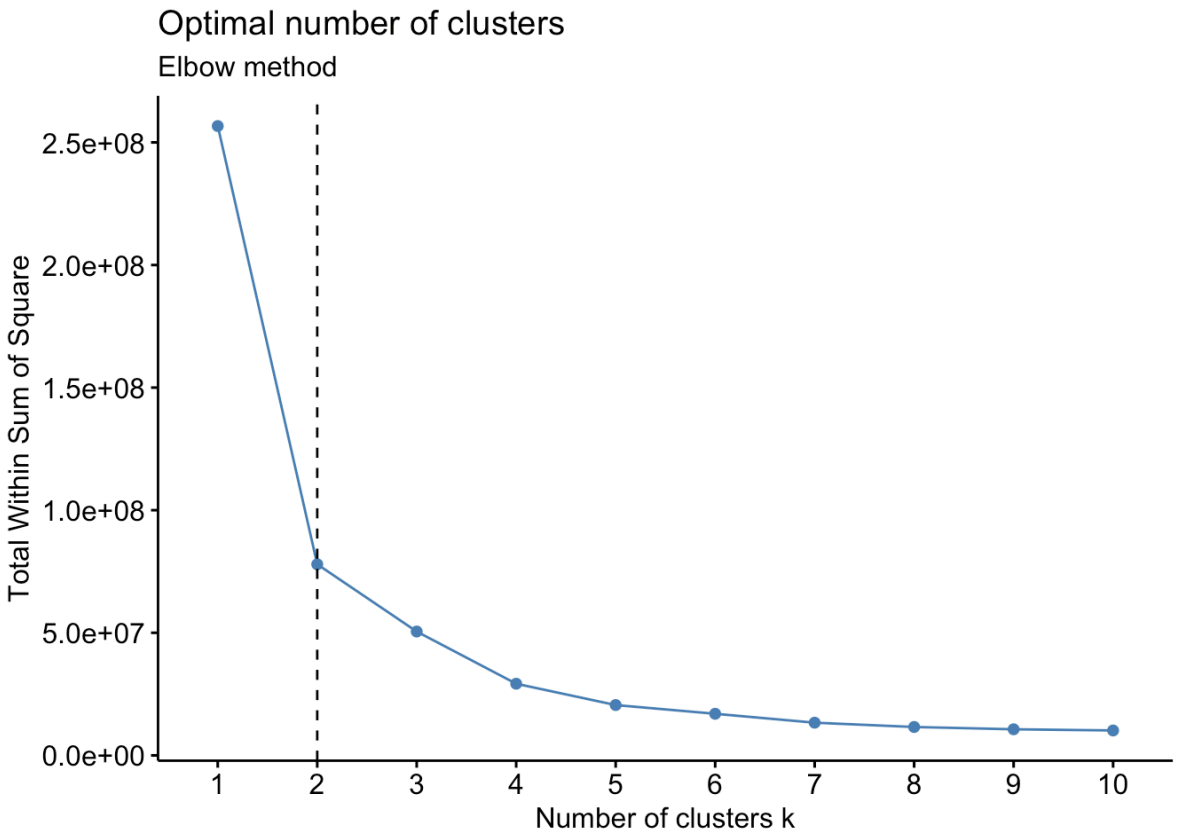
Clustering can be performed on the WBCD dataset from UCI to identify different groups or clusters of similar cases based on their characteristics or features. The dataset contains measurements of various features of breast cancer cells, such as their size, shape, and texture. Clustering analysis can be used to identify groups of cells that have similar features and could potentially be used to diagnose and treat different types of breast cancer. Clustering can also be used for exploratory data analysis to gain insights into the data and identify any patterns or anomalies. In addition, clustering can be used as a preprocessing step for other analyses such as

classification or regression. Overall, clustering can be a useful tool for analyzing and understanding the WBCD dataset.

K-Means Clustering :

It has been already confirmed by PCA algorithm that the data is separable. Therefore, in order to further exploit this discovery K-means clustering was performed. K-means is a simple and effective clustering algorithm with one drawback which is the necessity to specify the number of clusters beforehand. Once the number of clusters k has been decided, algorithm assigns each data points to one of those clusters using Euclidean distance. There are several techniques which give idea about the optimal number of clusters for data at hand. Of those techniques Elbow method has been employed in order to decide the number of clusters.

Elbow Method:



Elbow method looks at total within-cluster sum of squares for different numbers of clusters and chooses the number after which adding another cluster does not improve the total within-cluster sum of squares. In other words, withincluster variation is minimized. In this analysis, total wss was compared for varying numbers of clusters within the range (0,10). The number of clusters chosen by Elbow method for this data is 2.

K-Means Algorithm

K-means clustering algorithm works as follows:

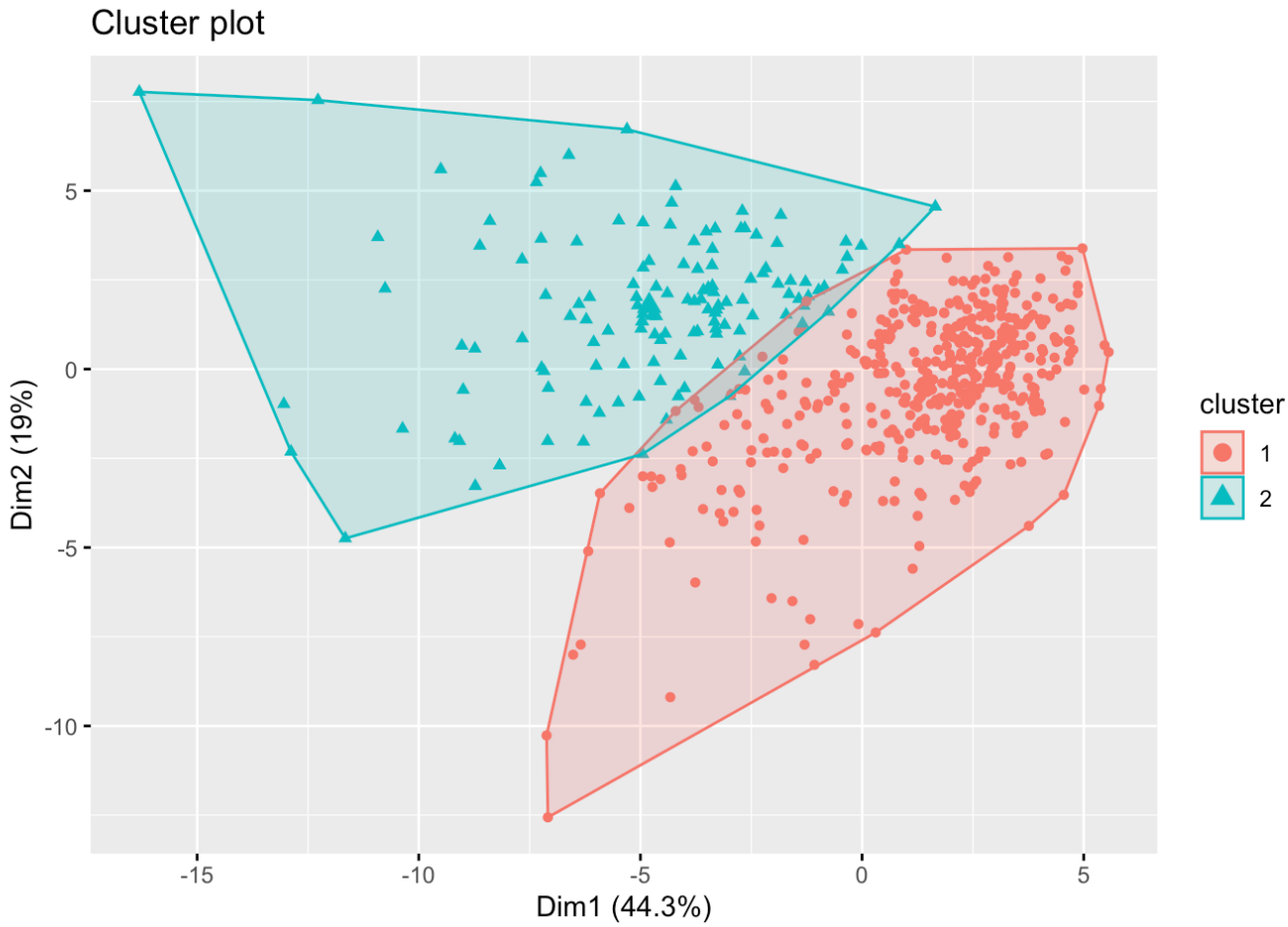
Step 1: Randomly assign each data point to the clusters from 1 to K.

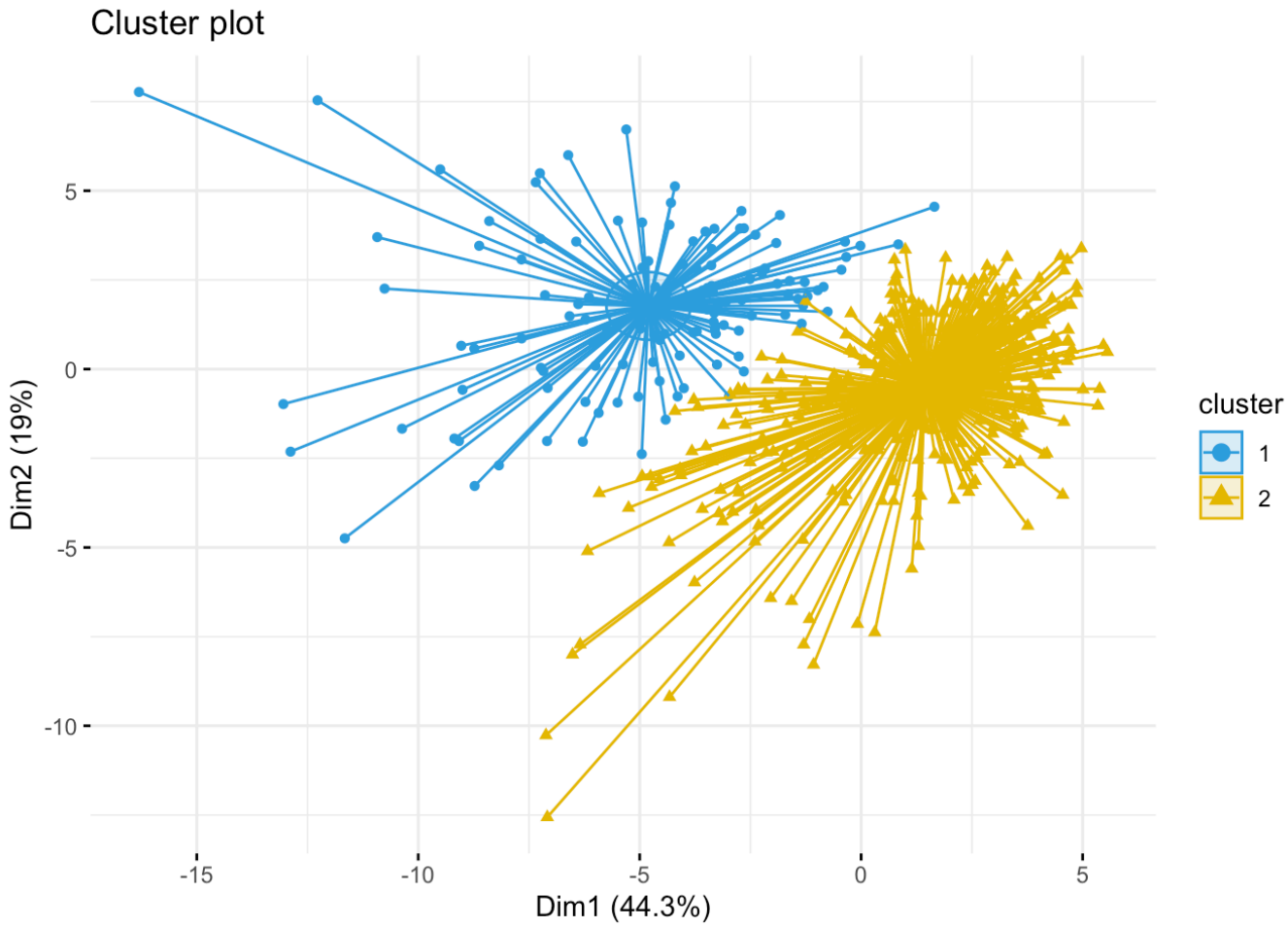
Step 2: For each cluster k cluster centroids are calculated. Centroids are the vectors containing the means of the observations for each feature.

Step 3: data points then are assigned to the cluster whose centroid is the closest to that data point in terms of Euclidean distance.

Step 4: Step 2 and 3 are repeated until cluster assignments stop changing. 9

Elbow method pointed out that 2 is the optimal number of clusters for this data and should be selected as **k** for k means algorithm. Hence, k –means algorithm was performed with k=2 According to the results of the the elbow method described above, 2 clusters should be selected for k means algorithm. We can see the visualizations of the k-means algorithm run with k=2 and nstart=100 below. With nstart=100, the algorithm goes over the steps described above for 100 times. This is to make the results of the algorithm more stable


Cluster plot
Dim2 (19%)
5
0
-5
-10
-15
-10
-5
0
5
Dim1 (44.3%)
cluster
1
2



.

From both the plots we see that the data was clustered into 2 different groups with a very little overlap.

Hierarchical Clustering There are two types of hierarchical clustering: agglomerative (bottom-up) and divisive (topdown). In this analysis, agglomerative approach was performed. Hierarchical clustering is another method for clustering data. Main advantage of this method is that the number of clusters do not have to be given the algorithm apriori. Outcome of this algorithm is a tree , called dendrogram, consisting of leaves and branches. Leaves correspond to every single data points. Similar data points , leaves, fuse into branches and similar branches fuse until there no leaves or branches left to fuse. Height of the tree represents how similar/dissimilar the data points are to each other. Therefore, data points that fuse at the lower part of the tree are very similar to each other.

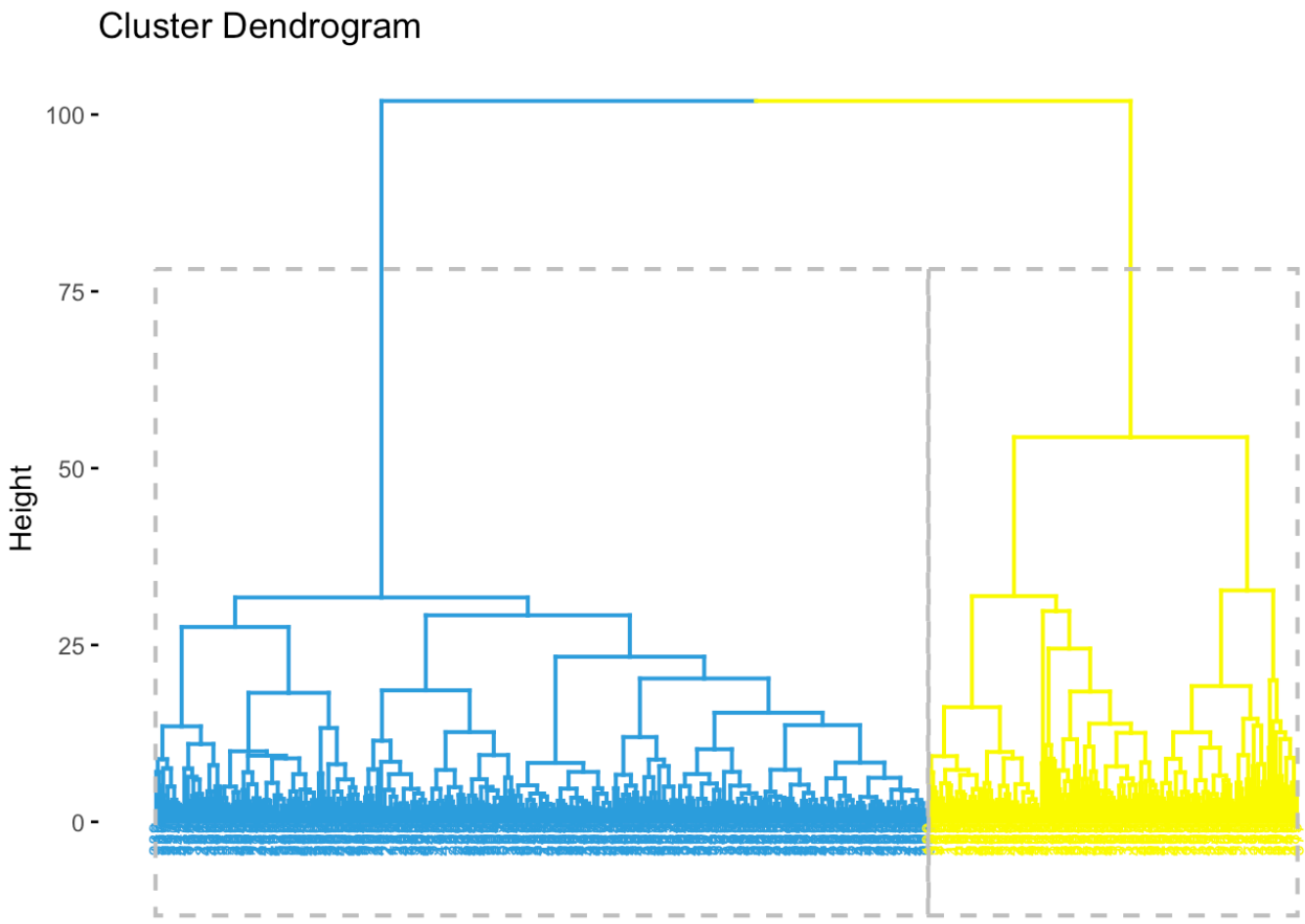
Hierarchical Clustering Algorithm

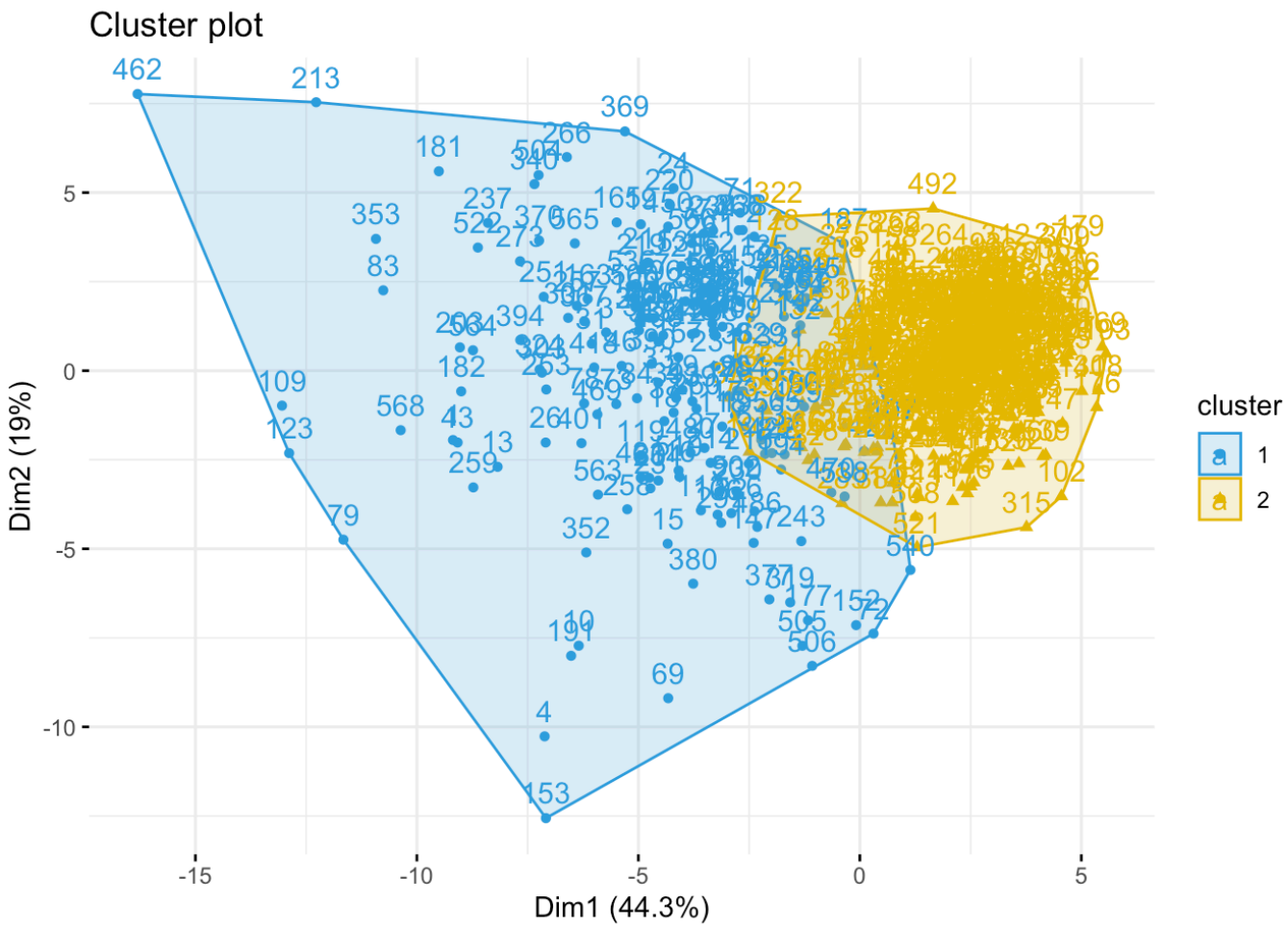
Step 1: Treat each data point as if they are themselves clusters, and then compute pairwise dissimilarities.

Step 2: Fuse the pairs(clusters) that are most similar

Step 3: Compute the new pairwise dissimilarities between clusters and fuse the most similar ones until the algorithm reaches the root node.

Hierarchical clustering was applied to reduced data which PCA produced. Distances between data points were calculated using Euclidean distance. Several types of linkage determine which clusters should be fused. In this analysis, complete, average, and ward linkage were employed. Cutting the dendrogram into different groups Dendrogram produce by hierarchical clustering with ward linkage was cut into first 2 and then, 4 groups. Numbers of observations in each cluster were examined and it is clear that clusters represent the different groups of patients. Cutting the tree into 2 groups reveals that patients who have been diagnosed with M belong to 1st cluster and patients who have been diagnosed as B belong to 2nd cluster. The tree was also cut into 4 groups, however, 2 clusters represent the data in a much better and useful way, than 4 clusters. When compared to the clusters produced by 2-means clustering algorithm , it can be seen that the clusters obtained by cutting the tree into 2 are consistent with 2-means results.





To understand how well these clustering methods partitioned the data, cluster numbers assigned to each data points were checked according to diagnosis label. Number of patients belong to each cluster appeared to be meaningful when examining those numbers based on diagnosis class; hence it has been clear that the data was clustered well enough to reflect these classes. Additionally, mean of all observations whose true labels and cluster numbers are equal is 0.85 for 2-means and 0.88 for hierarchical clustering with Ward link

Exploratory Factor Analysis

The aim of performing factor analysis on the WBCD dataset from UCI is to identify underlying factors that explain the correlations among the variables in the dataset. The WBCD dataset contains information about the characteristics of different cell nuclei from fine needle aspirates of breast mass. The dataset includes a large number of variables, such as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, fractal dimension, and diagnosis.

Factor analysis can help identify the underlying factors driving the relationships among these variables. This can provide insights into the factors that contribute to breast cancer diagnosis and help develop more accurate diagnostic models. Additionally, factor analysis can help simplify the dataset by identifying the variables most strongly associated with the underlying factors, making it easier to analyze the data. Finally, factor analysis can help identify outliers or unusual patterns in the data that may indicate errors or other issues with the dataset. Overall, factor analysis aims to better understand the complex relationships among the variables in the WBCD dataset.

Steps taken :

1. Perform data cleaning and pre-processing by converting the diagnosis column of the wdbc data frame into a factor variable, with levels "B" (for benign) and "M" (for malignant)
2. Perform a principal component analysis (PCA) on the Wisconsin (WDBC) dataset using the principal() function from the psych package in R. The purpose of this PCA is to reduce the dimensionality of the dataset and to identify the underlying patterns or structure.
3. Note the values of h2, u2, ,com,loadings, communality and scores
4. Perform parallel analysis and note inference from the plot

Inferences:

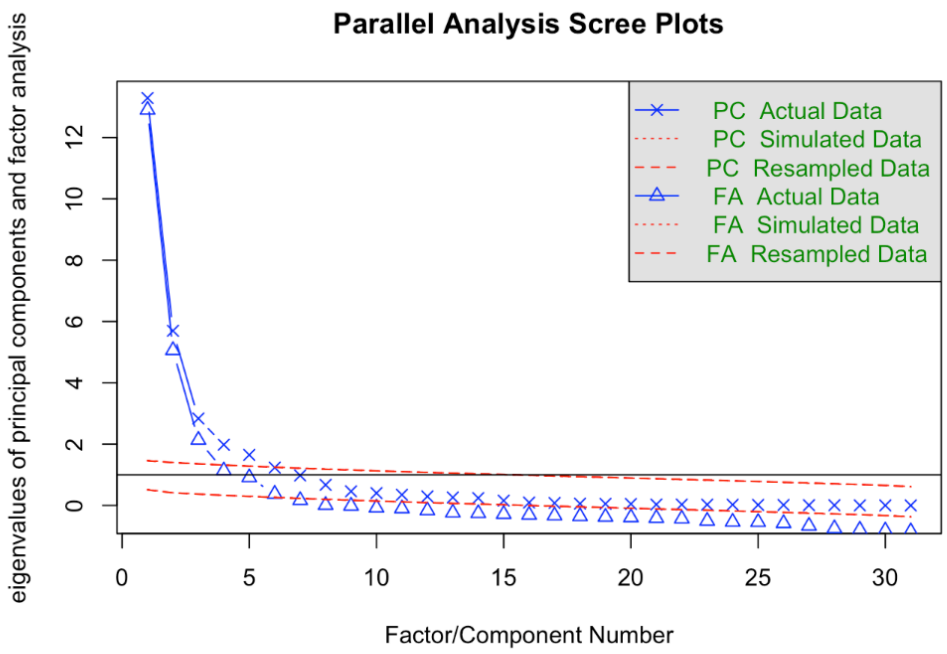
```
##       radius_mean      texture_mean    perimeter_mean         area_mean
##         0.9540666         0.9005072         0.9575588         0.9607121
##   smoothness_mean  compactness_mean    concavity_mean  concave_pts_mean
##         0.8669154         0.9097154         0.9205771         0.9273616
##     symmetry_mean  fractal_dim_mean         radius_se        texture_se
##         0.7809788         0.8463982         0.8891593         0.7636540
##      perimeter_se           area_se     smoothness_se    compactness_se
##         0.8788648         0.8623428         0.7441554         0.8978579
##      concavity_se    concave_pts_se       symmetry_se    fractal_dim_se
##         0.8328287         0.7607929         0.8504862         0.8304588
##      radius_worst     texture_worst   perimeter_worst        area_worst
##         0.9724005         0.9721432         0.9783229         0.9491775
##  smoothness_worst compactness_worst   concavity_worst concave_pts_worst
##         0.9153933         0.9043820         0.9045003         0.9224693
##    symmetry_worst fractal_dim_worst
##         0.9243849         0.8490730
```

##Higher communalities mean the factor solution represents the variable well.

#From the output, we can see that most of the original variables have high communality values (close to 1), indicating that they are well-represented by the principal components. This suggests that the principal components capture the majority of the variation in the original data.

#Communalities represent the proportion of variance in each variable that can be explained by all the other variables included in the analysis.

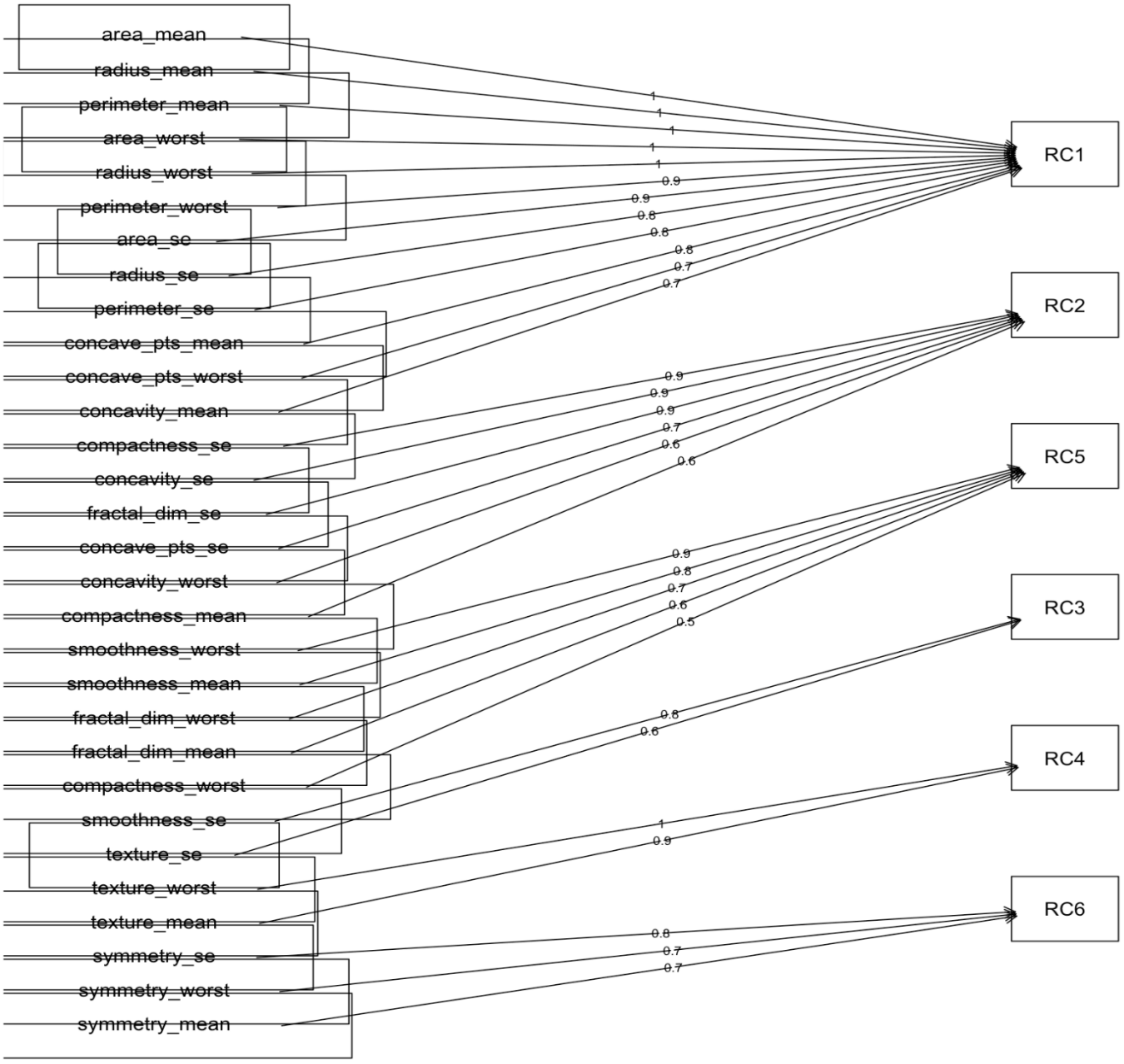
#For example, the communalities for "radius_mean" is 0.954, which means that 95.4% of the variance in "radius_mean" can be explained by all the other variables included in the principal component analysis (PCA).




```
## Parallel analysis suggests that the number of factors =  6  and the number of components =  5
```

Parallel analysis is a method used to determine the optimal number of factors or components in a factor analysis. It compares the eigenvalues of the actual data to the eigenvalues of randomly generated data with the same sample size and number of variables. Based on the output you provided, it seems that the parallel analysis suggests that the number of factors in the data is 6 and the number of components is 5. This means that the data is likely best represented by 6 underlying factors or dimensions, and that a factor analysis with 5 components is appropriate for summarizing the data.

Components Analysis



Based on the loadings output, the variables that contribute the most to each of the 6 components (RC1, RC2, RC3, RC4, RC5, and RC6) are as follows:

#

RC1: radius_mean, perimeter_mean, area_mean, radius_worst, perimeter_worst, area_area_se, radius_se,perimeter_se,concave_pts_mean,concavity_mean,concave_pts_first

RC2: smoothness_mean, smoothness_worst, fractal_dim_worst,fractal_dim_mean, compactness_worst

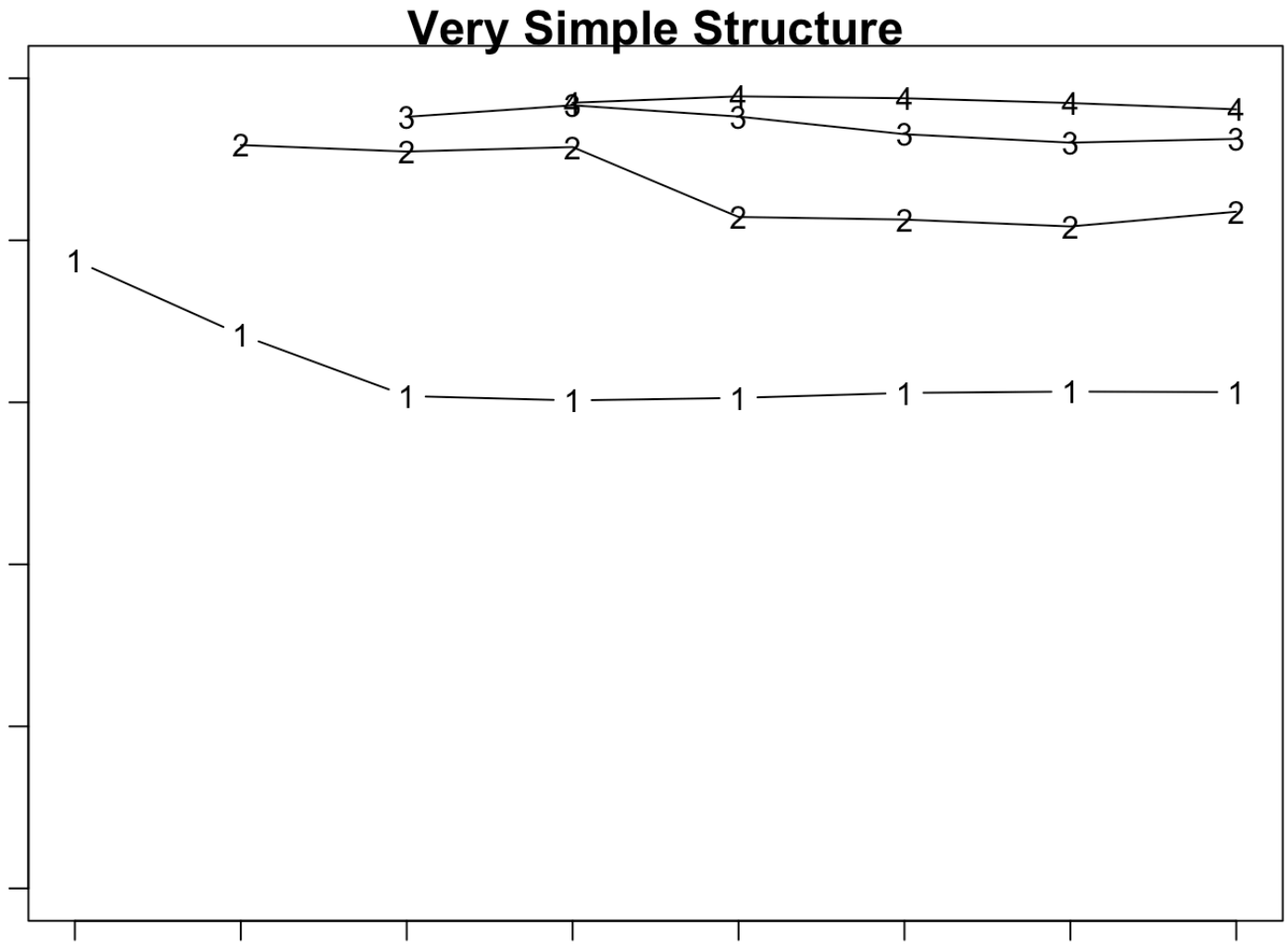
RC3: smoothness_se, texture_se

RC5: texture_worst, texture_mean

RC5: smoothness_mean, compactness_mean, concavity_mean, concave_pts_mean, symmetry_mean

RC6: symmetry_se,symmetry_worst_symmetry_mean

Therefore, the variables contributing the most to the first five components are quite varied, while symmetry-related variables and measures of irregularity in cell shape mostly characterize the sixth component.



In this case, where the line is almost horizontal for 4 factors, a 4-factor model is likely to be a good fit for the data, and further increasing the number of factors may not be meaningful or valuable.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a statistical method that is used to classify data into different categories. In the case of the WBCD dataset, the goal is to classify tumors as either malignant or benign based on their characteristics. LDA can be used to create a linear combination of features that maximizes the separation between the two classes. This linear combination is then used to classify new tumors as either malignant or benign. By performing LDA on the WBCD dataset, one can determine which features are most important in distinguishing between malignant and benign tumors, and develop a model that can accurately classify new tumors. The proportion of variance explained by each LD tells us how much of the total variability in the data is accounted for by that LD.

Steps taken :

1. Perform data cleaning and pre-processing by converting the selected columns to a matrix using as.matrix(),setting the row names to the id column values, converting the diagnosis column of the wdbc data frame into a factor variable, with levels "B" (for benign) and "M" (for malignant) and later to numeric so that "B" is represented by 0 and "M" is represented by 1.
2. Calculate the sample size for the training set in a machine learning model. In this case, we use 75% of the data for training and 25% for testing. We can reduce the risk of overfitting by randomly selecting a subset of observations to be used for training.
3. Train a Linear Discriminant Analysis (LDA) model on the training set of the WBCD. The trained LDA model can be used to predict new data. The model will have learned a linear decision boundary that separates the two classes of diagnosis based on the values of the predictor variables. By using LDA, we aim to find a lower-dimensional representation of the data that maximizes the separation between the two classes, which can help us better classify new observations in the future.
4. Plot the output of LDA to visualize the separation between the two classes (malignant and benign tumors) in the reduced-dimensional space.
5. Perform prediction on the test data using the previously fitted LDA mode
6. Generate a ROC (Receiver Operating Characteristic) curve using the predictions generated by the LDA model on the test set. The ROC curve shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) at different probability thresholds for classification.
7. Note the inferences

Inferences from the linear discriminant coefficients (LD1) :

The output below shows the linear discriminant coefficients (LD1) for each feature of the breast cancer dataset. The coefficients indicate each feature's importance in distinguishing between malignant and benign tumors. Features with larger absolute values of LD1 are more critical in discriminating between tumor types.

	LD1
radius_mean	-0.802397587
texture_mean	0.038661119
perimeter_mean	0.075455004
area_mean	0.002475268
smoothness_mean	4.872158718
compactness_mean	-19.916911526
concavity_mean	12.522041295
concave_pts_mean	-3.912173036
symmetry_mean	1.764002079
fractal_dim_mean	-17.632908602
radius_se	2.895924023
texture_se	0.044777925
perimeter_se	-0.042849165
area_se	-0.011868576
smoothness_se	-2.734982178
compactness_se	5.756083232
concavity_se	-21.986709973
concave_pts_se	44.511925528
symmetry_se	8.071379854
fractal_dim_se	-6.968610253
radius_worst	0.764197359
texture_worst	0.027196108
perimeter_worst	-0.005271662
area_worst	-0.004157875
smoothness_worst	8.625163030
compactness_worst	0.530275050
concavity_worst	0.849721791
concave_pts_worst	6.333886066
symmetry_worst	1.826353636
fractal_dim_worst	19.860808775

These coefficients determine the direction of the linear combination of the input variables (predictors) that best separates the two groups (benign and malignant) in the training data

The LD1 column represents the coefficients for the first linear discriminant. Positive coefficients suggest that

higher values of that variable are associated with the malignant group, while negative coefficients suggest the opposite. The magnitude of the coefficient indicates the strength of the association. For example, we can see those larger values of concavity_se and compactness_mean are strongly associated with the malignant group, while larger values of area_mean and perimeter_mean are associated with the benign group.

A coefficient of -21.986709973 for concavity_se in the LD1 direction means this variable is strongly negatively associated with the first linear discriminant. In other words, lower values of concavity_se are associated with the malignant group, while higher values are associated with the benign group. This may seem counterintuitive given the coefficient's sign. Still, it's important to remember that the sign and magnitude of the coefficients depend on the direction of the linear discriminant and the scaling of the variables.

Based on the output of the linear discriminant analysis (LDA) on the breast cancer dataset, the 10 most important features in distinguishing between malignant and benign tumors are:

concave_pts_se (44.511925528)

fractal_dim_worst (19.860808775)

compactness_mean (-19.916911526)

concavity_se (-21.986709973)

symmetry_se (8.071379854)

smoothness_worst (8.62516303)

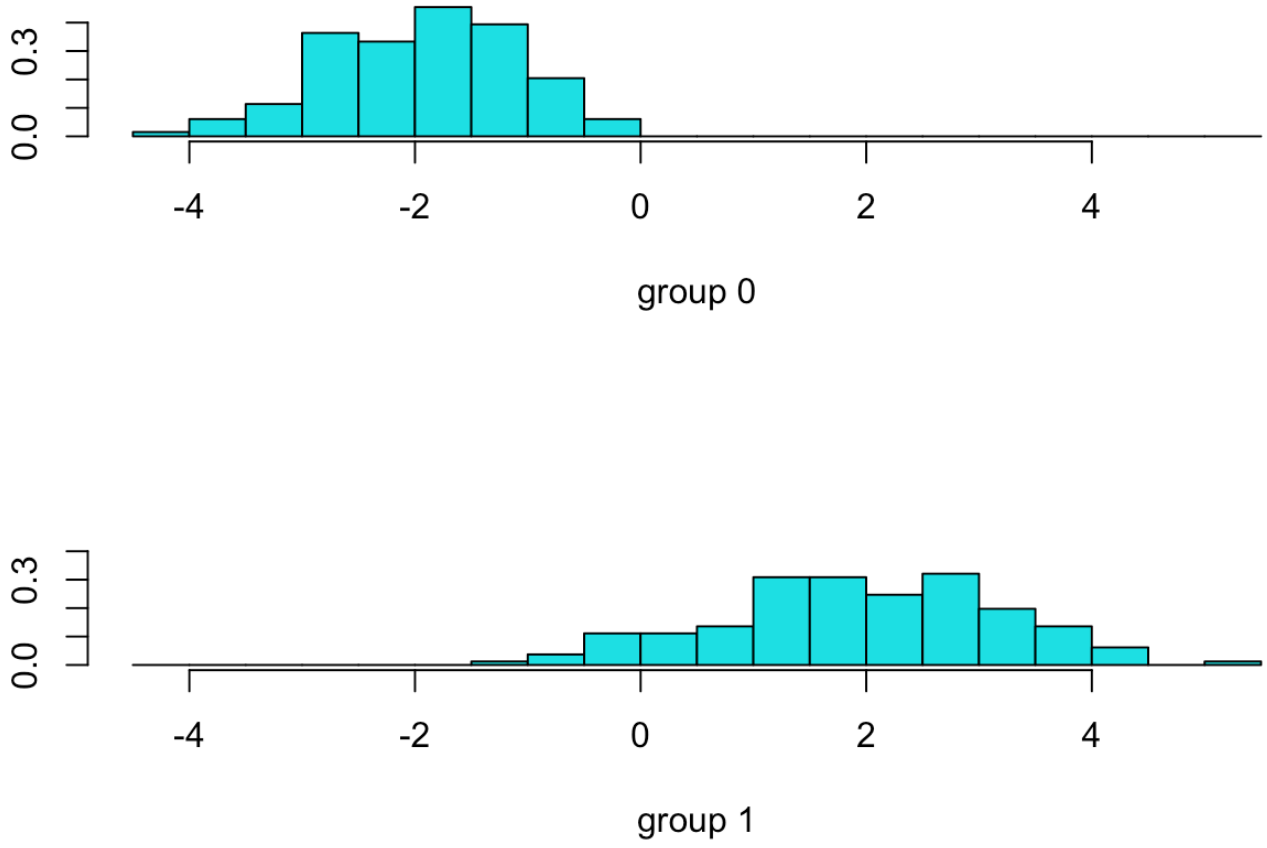
concavity_mean (12.522041295)

radius_se (2.895924023)

symmetry_mean (1.764002079)

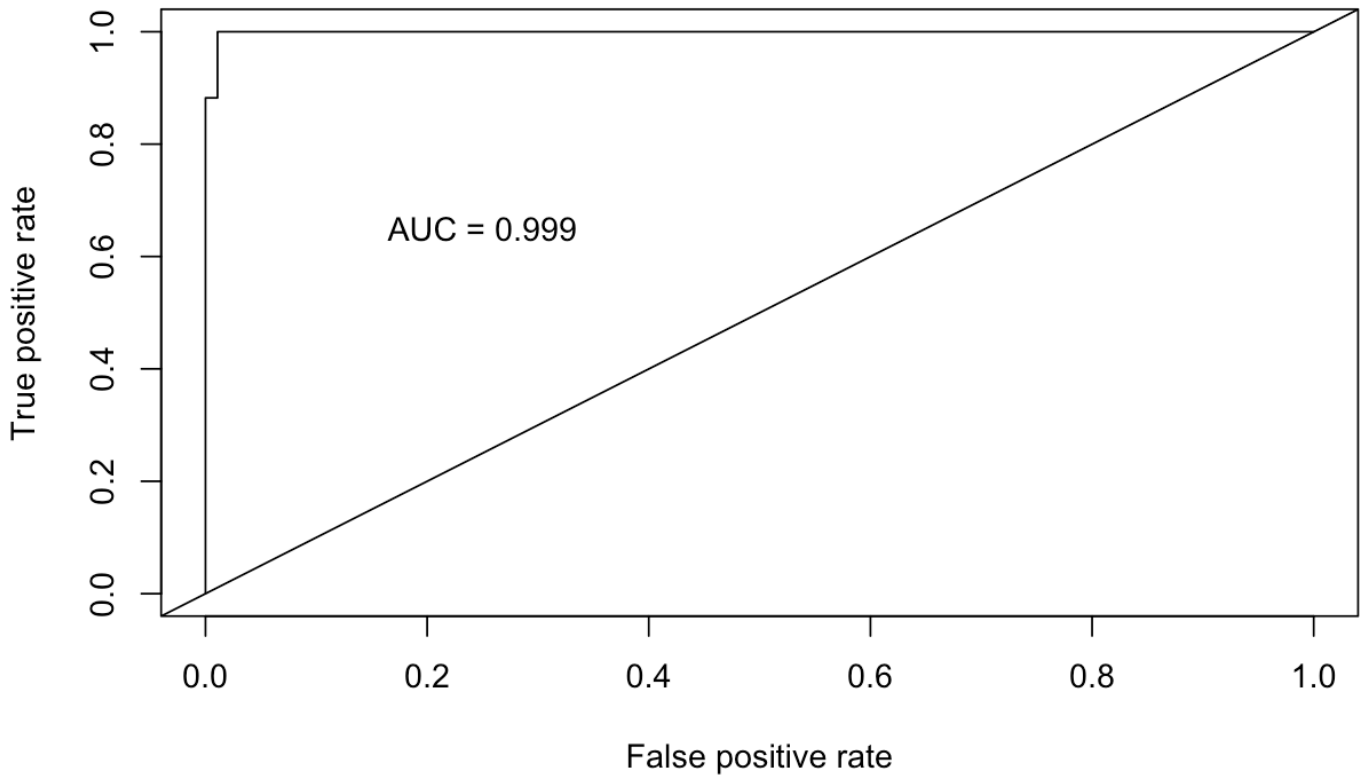
concave_pts_worst (6.333886066)

Note that the ranking is based on the absolute values of the LD1 coefficients obtained from the LDA, with the highest absolute value indicating the highest importance.



These plots show the distribution of the linear discriminant scores for each group. The linear discriminant score is a measure of the distance between an observation and the group centroids in the transformed space. they show the distribution of the linear discriminant scores for the group, with the x-axis representing the linear discriminant score and the y-axis representing the densit

Inference from these plots can be used to understand the separation between the two groups in the transformed space. If the distributions of the linear discriminant scores for the two groups overlap substantially, this indicates that the LDA model cannot separate the groups based on the predictors fully and may need to be more effective at discriminating between the two groups. On the other hand, if the distributions are well separated, the LDA model can effectively differentiate between the two groups based on the predictors. Here, we observe a slight overlap between the two groups.



The area under the curve (AUC) measures the classifier's performance, with a higher AUC indicating better performance. In this case, the plot shows that the LDA classifier has a very high performance, with an AUC of 0.999, close to the maximum possible value of 1. The fact that the ROC curve is very close to the top-left corner of the plot indicates that the classifier can achieve high TPRs with low FPRs, which is a desirable property for a binary classifier. Overall, the plot indicates that the LDA classifier effectively distinguishes malignant and benign tumors in the test dataset.

Conclusion

- Are there any missing values or outliers in the data?

  No there are no missing values or outliers.
- Is the response variable distributed in a balanced manner?

  No it is not.
- Can we reduce the dimensionality of the data while preserving the important information?

  Yes we can and that was proven by the PCA
- Can we identify any relationships or correlations between the factors?

  Yes we have found many variables exhibiting positive correations
- Can we identify any natural groupings or clusters in the data?

  Yes, we could find two naturally occurring cluster by applying cluster analysis
- Can we find a linear combination of variables that can effectively discriminate between different groups or classes in the data?

  Yes, we could apply LDA concept on the data
- How accurate is the classification based on this linear combination of variables?

  The area under the curve (AUC) measures the classifier's performance, with a higher AUC indicating better performance. In this case, the plot shows that the LDA classifier has a very high performance, with an AUC of 0.999, close to the maximum possible value of 1