Assignment:

Classification of . . . using PCA, Clustering analysis, and ANOVA

Group 7

07/12/2023

ID	Member	Faculty
2370500	Le Duc Khoan	Computer Science
2270522	Tran Trung Thai	Computer Science
1912046	Ngo Minh Hong Thai	Computer Science
2370041	Tran Hoang Nhat Huy	Computer Science

Contents

List of Figures

1 Introduction

Medical analysis encompasses a wide array of methodologies and technologies aimed at understanding, diagnosing, treating, and preventing diseases and medical conditions. It plays a pivotal role in healthcare, utilizing various data sources such as clinical records, medical imaging, genetic information, and laboratory tests to derive insights that aid healthcare professionals in delivering optimal patient care. One significant application of medical analysis is in the field of oncology, particularly in the diagnosis and treatment of breast cancer. The Breast Cancer Wisconsin (Diagnostic) Data Set serves as a fundamental resource in this domain, offering a comprehensive collection of clinical and imaging data derived from fine needle aspirates (FNAs) of breast masses. This dataset, widely utilized in machine learning and data mining research, provides detailed features describing cell nuclei characteristics, allowing for the development and validation of predictive models to distinguish between benign and malignant breast tumors. Through medical analysis of datasets like the Breast Cancer Wisconsin (Diagnostic) Data Set, researchers and healthcare practitioners strive to improve early detection, prognosis, and treatment outcomes for breast cancer patients, ultimately advancing the field of oncology and enhancing patient care.

In the domain of medical analysis, the integration of advanced methodologies such as Principal Component Analysis (PCA), Hierarchical Clustering play pivotal roles in unraveling complex datasets like the Breast Cancer Wisconsin (Diagnostic) Data Set. This report embarks on an exploration of these analytical techniques and their application in unraveling crucial insights into breast cancer diagnosis and prognosis.

PCA serves as a cornerstone in dimensionality reduction, offering a means to distill the multidimensional complexity of medical datasets into a more manageable form. By transforming the original features into a smaller set of uncorrelated variables, PCA facilitates visualization, pattern recognition, and the identification of critical factors contributing to breast cancer diagnosis. In the context of the Breast Cancer Wisconsin (Diagnostic) Data Set, PCA holds promise in elucidating key biomarkers and phenotypic traits indicative of tumor malignancy.

Complementing PCA, Hierarchical Clustering offers a powerful method for uncovering inherent structures within the dataset. By grouping similar data points into clusters based on their similarity or dissimilarity, Hierarchical Clustering provides insights into the natural organization of the data, potentially revealing distinct subgroups of breast tumors or patient profiles. Through this clustering approach, patterns and correlations among variables can be discerned, contributing to a deeper understanding of breast cancer heterogeneity.

In addition to PCA and Hierarchical Clustering, the inclusion of ANOVA adds another layer of statistical rigor to the analysis. ANOVA enables the assessment of differences in means across multiple groups, allowing for the identification of significant variations in clinical or molecular features among different tumor types or patient cohorts. By incorporating ANOVA into the analytical framework, this report aims to elucidate the underlying factors driving breast cancer progression and treatment response, ultimately guiding clinical decision-making and personalized therapeutic interventions.

Through the integration of PCA, Hierarchical Clustering, and ANOVA, this report endeav-

ors to unravel the complex landscape of breast cancer and pave the way for more precise diagnostic and therapeutic strategies. By leveraging these advanced analytical techniques, healthcare practitioners can gain deeper insights into the underlying mechanisms of breast cancer pathogenesis, ultimately leading to improved patient outcomes and enhanced precision in clinical practice.

2 Research Methods

2.1 Data description

The breast cancer dataset contains various features related to breast mass characteristics derived from digitized images of fine needle aspirate (FNA) of breast masses. It contains 32 variables, with 2 information variables: 1. ID: Identifier for each patient. 2. Diagnosis: The diagnosis of the breast mass (M = malignant, B = benign). and 30 feature variables: 1. Radius (mean, se, worst): Mean of distances from the center to points on the perimeter. 2. Texture (mean, se, worst): Standard deviation of gray-scale values. 3. Perimeter (mean, se, worst): Perimeter of the mass. 4. Area (mean, se, worst): Area of the mass. 5. Smoothness (mean, se, worst): Local variation in radius lengths. 6. Compactness (mean, se, worst): 7. Concavity (mean, se, worst): Severity of concave portions of the contour. 8. Concave points (mean, se, worst): Number of concave portions of the contour. 9. Symmetry (mean, se, worst): Symmetry of the mass. 10. Fractal dimension (mean, se, worst): "Coastline approximation" - a measure of the complexity of the contour.

2.2 Data preprocessing using R

```
## # A tibble: 6 x 31
##   diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
##   <chr>           <dbl>        <dbl>          <dbl>     <dbl>           <dbl>
## 1 M                18.0         10.4          123.      1001           0.118
## 2 M                20.6         17.8          133.      1326           0.0847
## 3 M                19.7         21.2          130       1203           0.110
## 4 M                11.4         20.4           77.6      386.          0.142
## 5 M                20.3         14.3          135.      1297           0.100
## 6 M                12.4         15.7           82.6      477.          0.128
## # i 25 more variables: compactness_mean <dbl>, concavity_mean <dbl>,
## #   'concave points_mean' <dbl>, symmetry_mean <dbl>,
## #   fractal_dimension_mean <dbl>, radius_se <dbl>, texture_se <dbl>,
## #   perimeter_se <dbl>, area_se <dbl>, smoothness_se <dbl>,
## #   compactness_se <dbl>, concavity_se <dbl>, 'concave points_se' <dbl>,
## #   symmetry_se <dbl>, fractal_dimension_se <dbl>, radius_worst <dbl>,
## #   texture_worst <dbl>, perimeter_worst <dbl>, area_worst <dbl>, ...
```

2.2.1 Data preparation

- Missing data

```
# Check for missing values
missing_values <- colSums(is.na(data))

# Print columns with missing values
print(missing_values)
```

```
##               diagnosis             radius_mean            texture_mean
##                       0                       0                       0
##          perimeter_mean               area_mean         smoothness_mean
##                       0                       0                       0
##        compactness_mean          concavity_mean     concave_points_mean
##                       0                       0                       0
##           symmetry_mean  fractal_dimension_mean              radius_se
##                       0                       0                       0
##              texture_se            perimeter_se                 area_se
##                       0                       0                       0
##           smoothness_se          compactness_se            concavity_se
##                       0                       0                       0
##       concave_points_se             symmetry_se    fractal_dimension_se
##                       0                       0                       0
##            radius_worst           texture_worst         perimeter_worst
##                       0                       0                       0
##              area_worst        smoothness_worst       compactness_worst
##                       0                       0                       0
##         concavity_worst    concave_points_worst          symmetry_worst
##                       0                       0                       0
## fractal_dimension_worst
##                       0
```

```
#Remove NA rows:
data <- na.omit(data)
```

- Outliers

2.2.2 Data exploration

- Boxplot

```r
for (col in colnames(data)){
  if (col != 'diagnosis'){
    # boxplot(df.clean[, col], main=col)
    boxplot(data[, col] ~ data[, 'diagnosis'], xlab='diagnosis', ylab=col, main = NULL)
  }
}
```

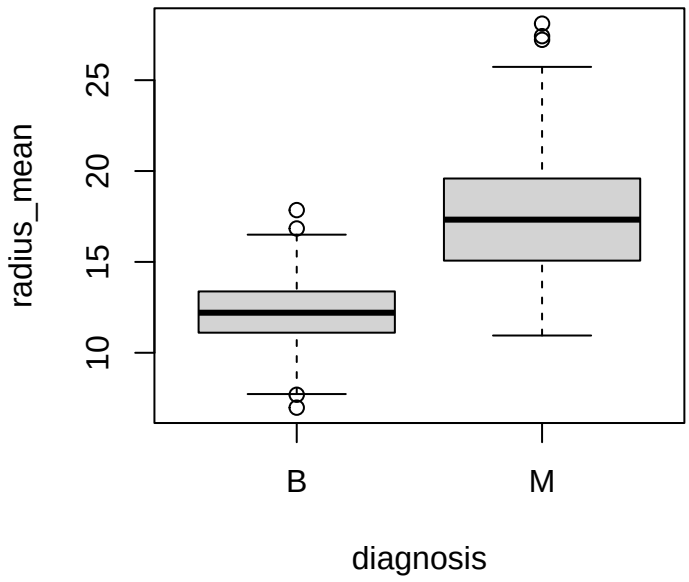


Figure 1: Boxplot of Diagnosis vs. radius_mean

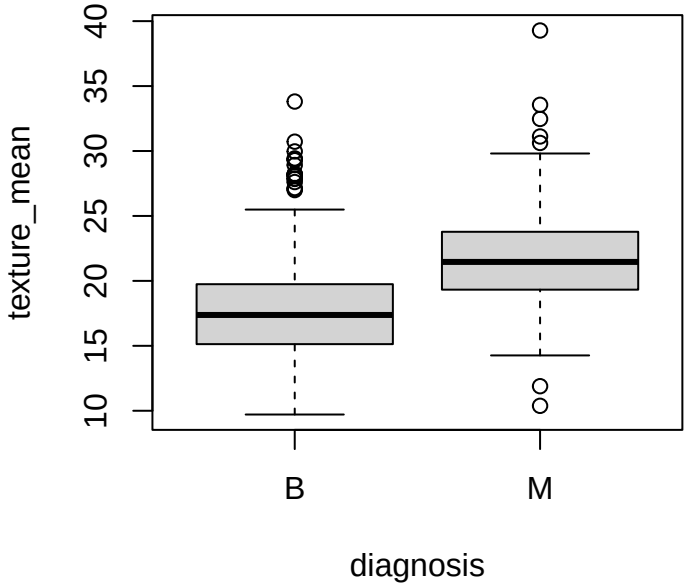


Figure 2: Boxplot of Diagnosis vs. texture_mean

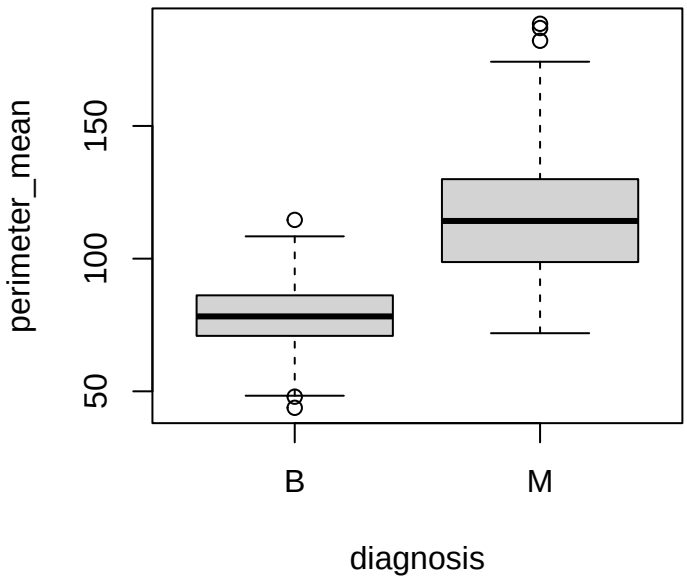


Figure 3: Boxplot of Diagnosis vs. perimeter_mean

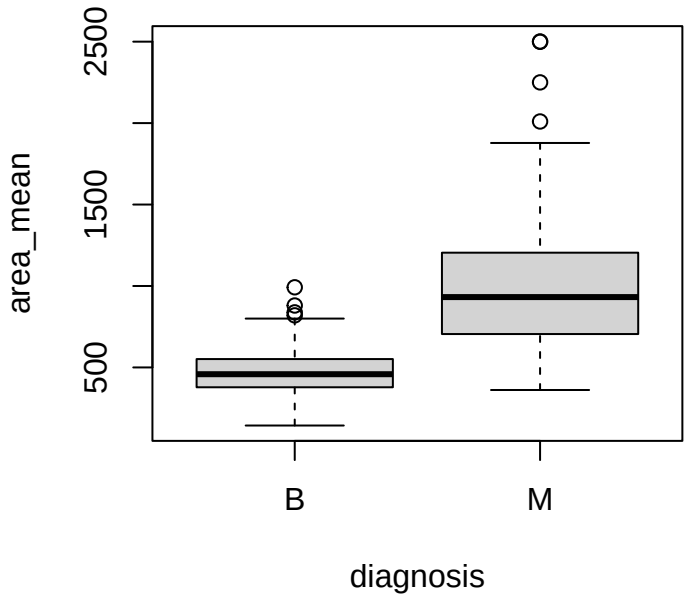


Figure 4: Boxplot of Diagnosis vs. area_mean

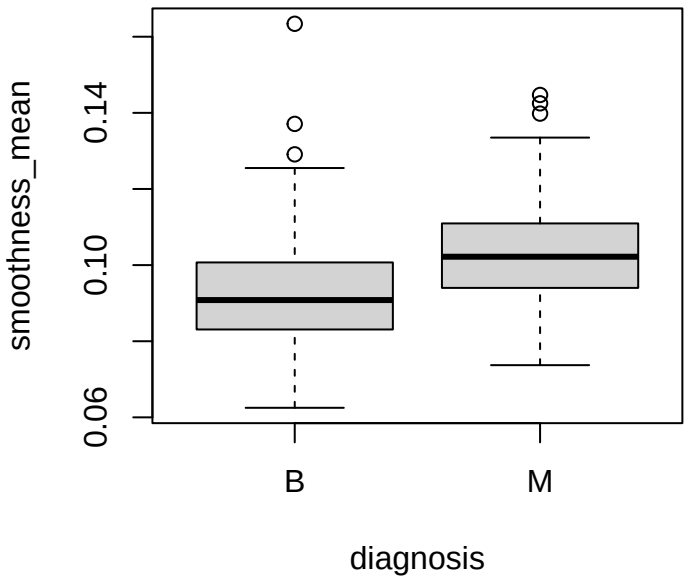


Figure 5: Boxplot of Diagnosis vs. smoothness_mean

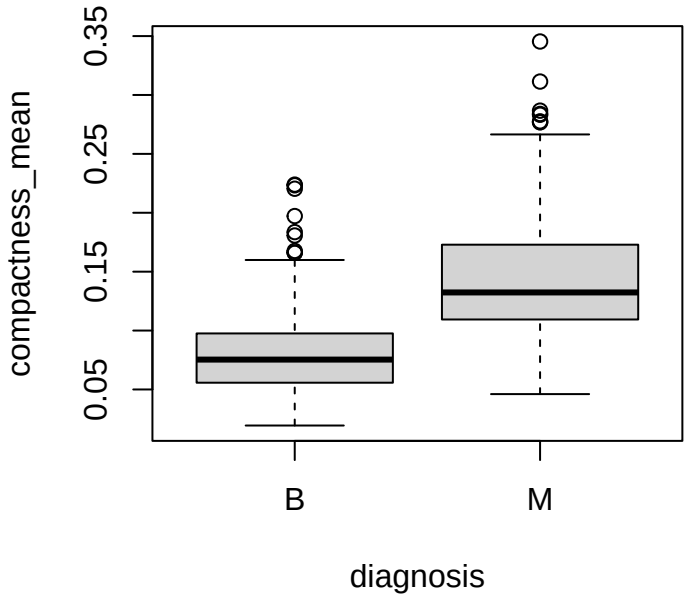


Figure 6: Boxplot of Diagnosis vs. compactness_mean

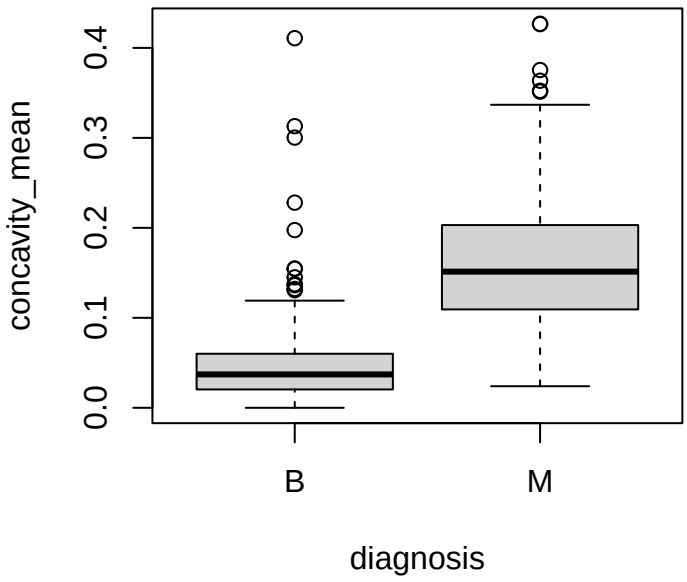


Figure 7: Boxplot of Diagnosis vs. concavity_mean

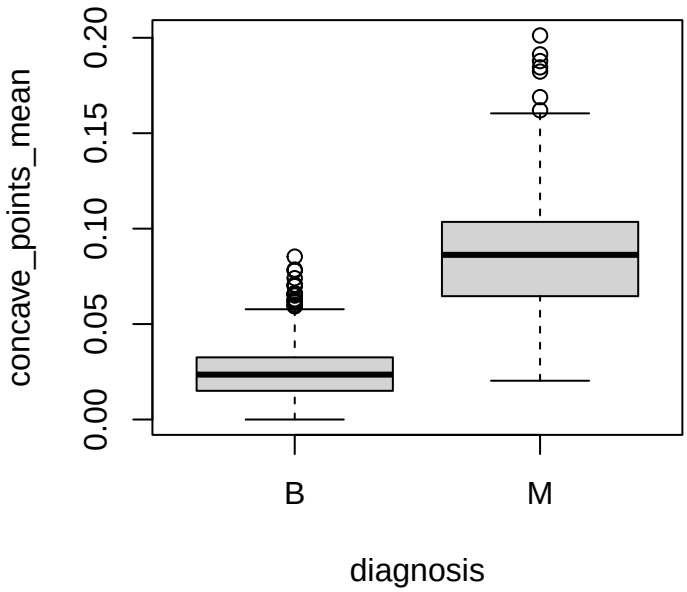


Figure 8: Boxplot of Diagnosis vs. concave_points_mean

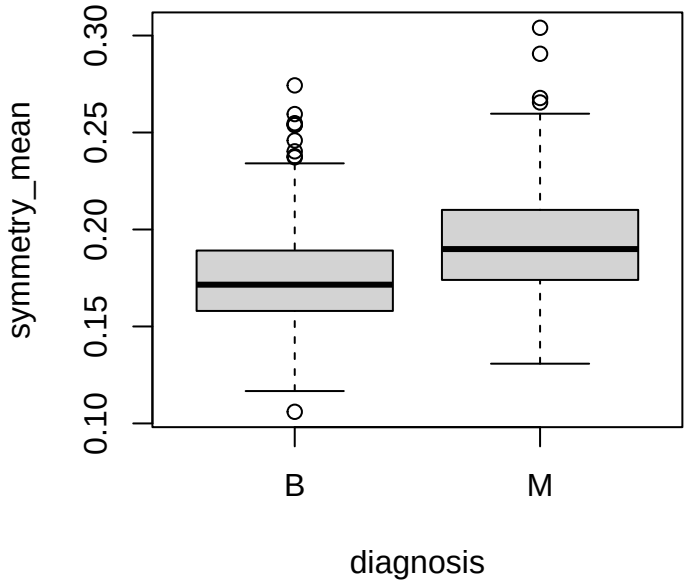


Figure 9: Boxplot of Diagnosis vs. symmetry_mean

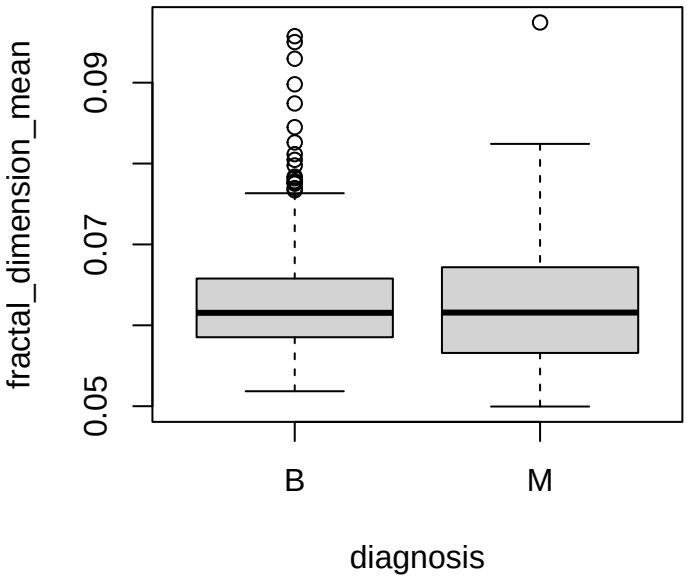


Figure 10: Boxplot of Diagnosis vs. fractal_dimension_mean

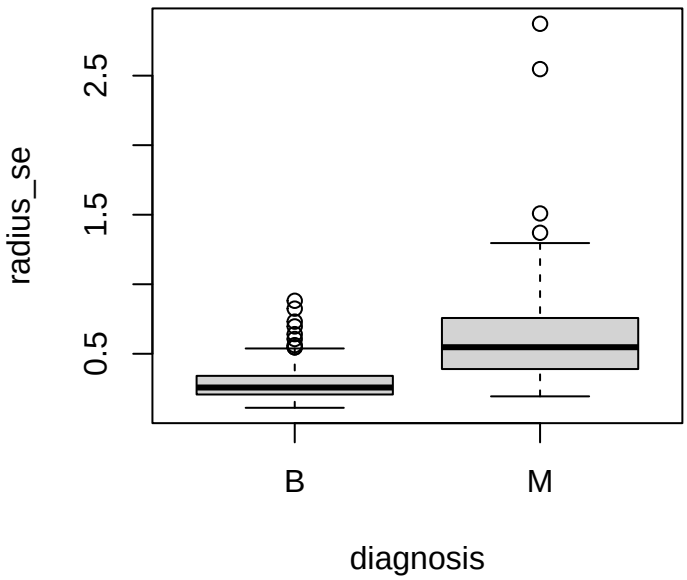


Figure 11: Boxplot of Diagnosis vs. radius_se

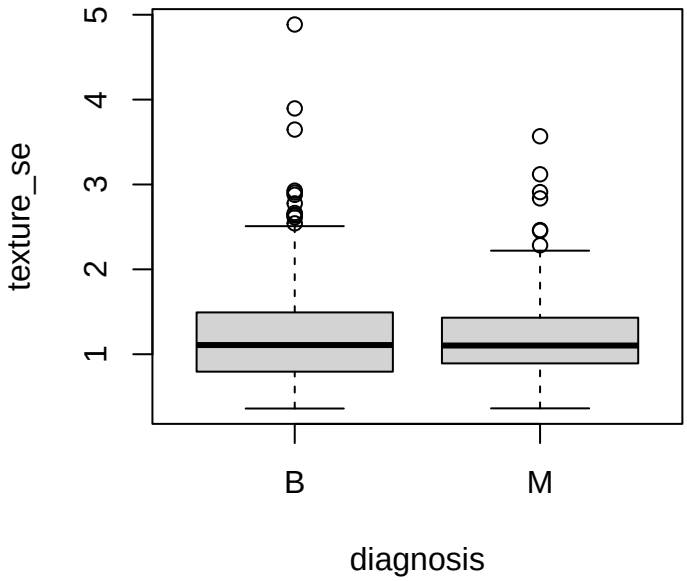


Figure 12: Boxplot of Diagnosis vs. texture_se

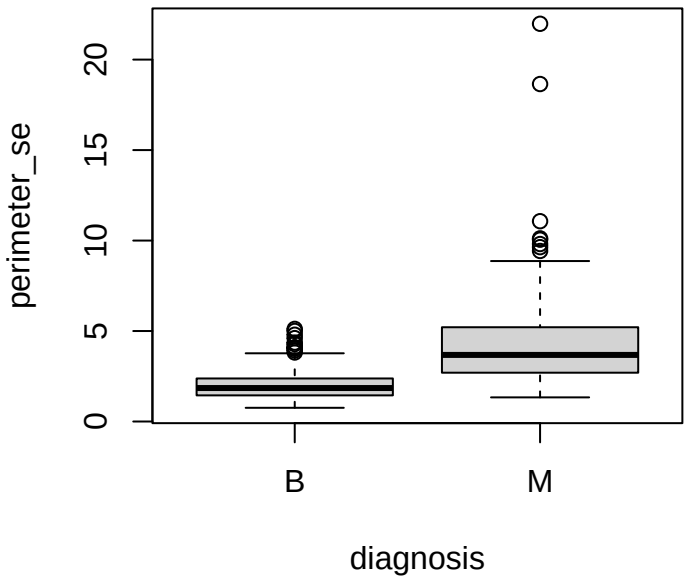


Figure 13: Boxplot of Diagnosis vs. perimeter_se

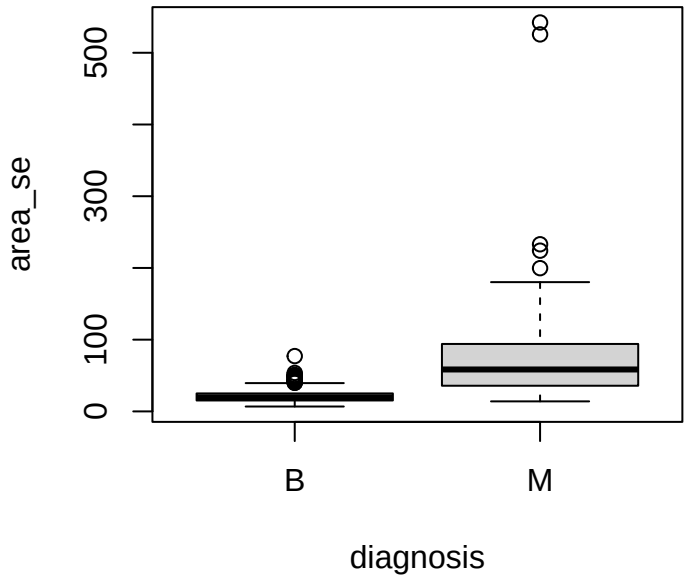


Figure 14: Boxplot of Diagnosis vs. area_se

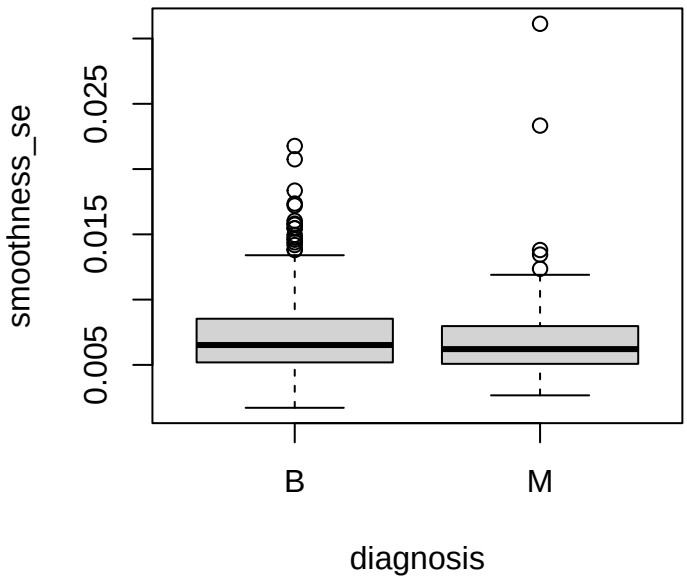


Figure 15: Boxplot of Diagnosis vs. smoothness_se

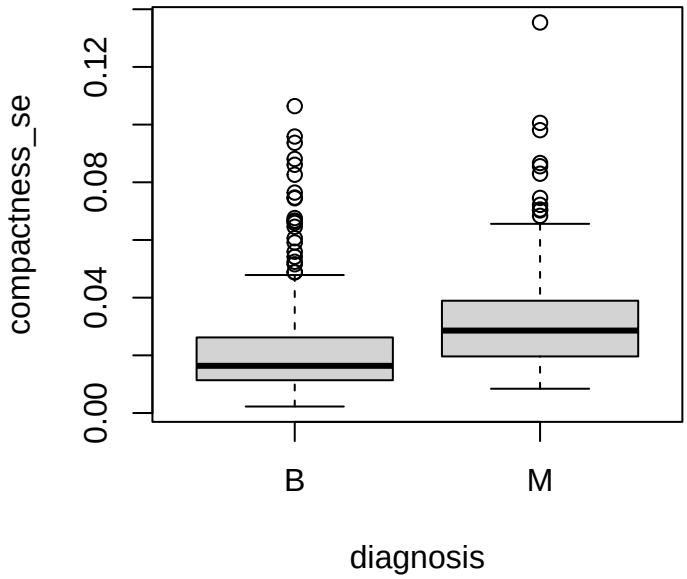


Figure 16: Boxplot of Diagnosis vs. compactness_se

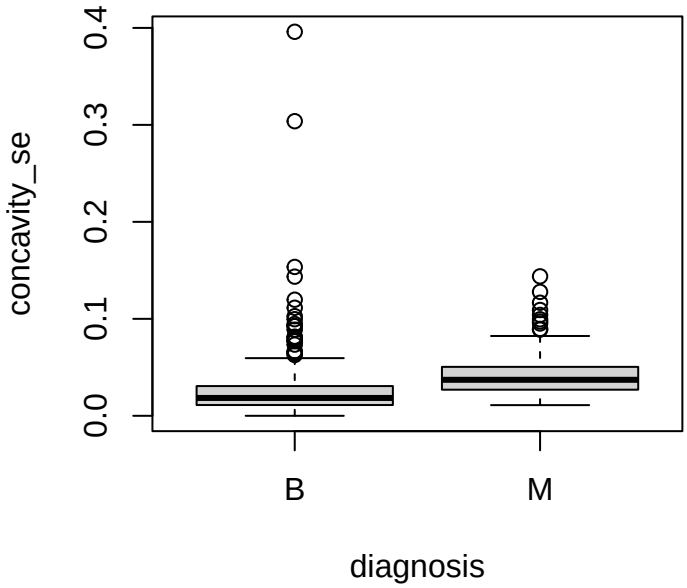


Figure 17: Boxplot of Diagnosis vs. concavity_se

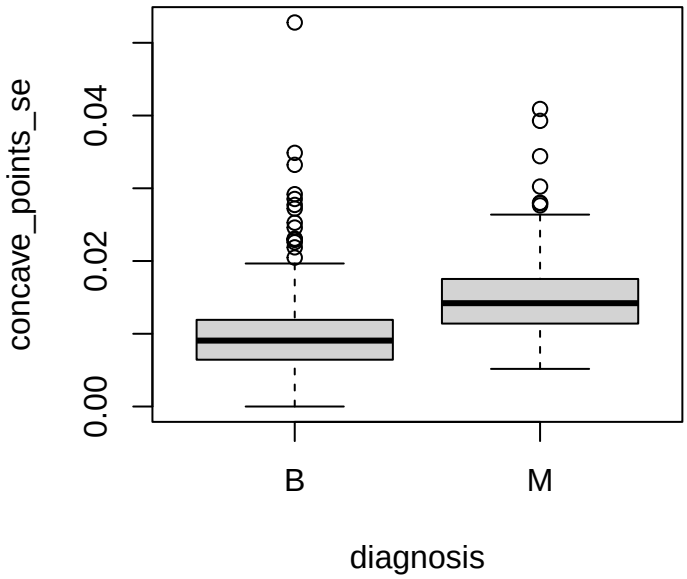


Figure 18: Boxplot of Diagnosis vs. concave_points_se

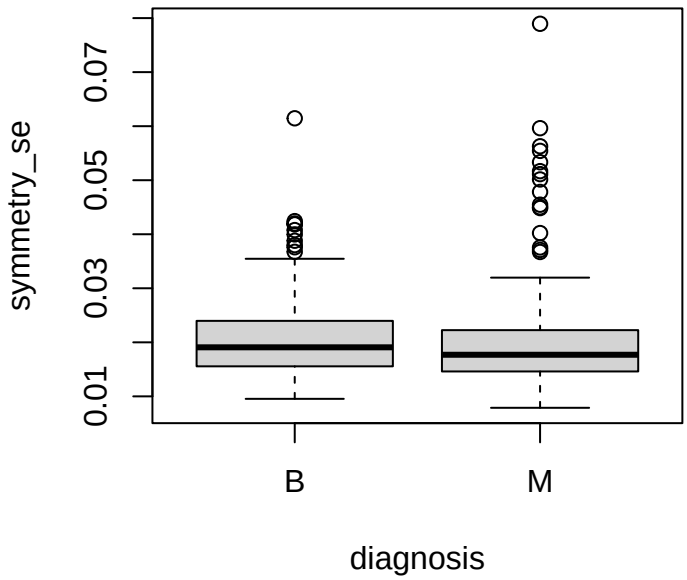


Figure 19: Boxplot of Diagnosis vs. symmetry_se

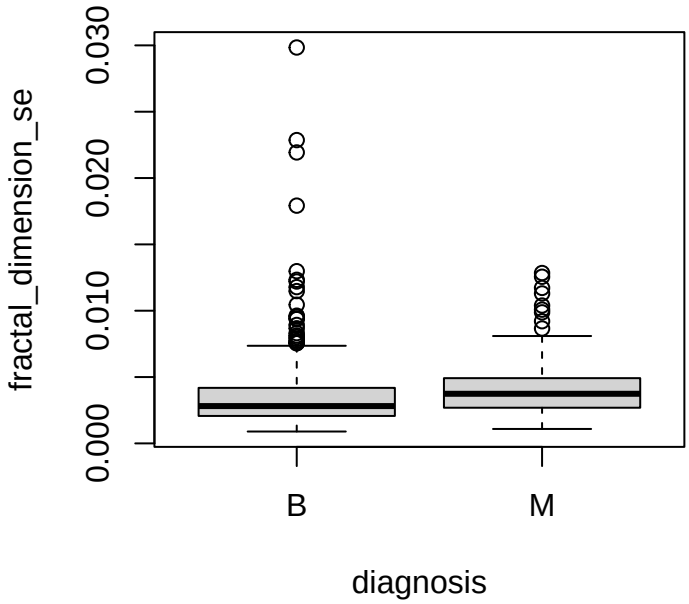


Figure 20: Boxplot of Diagnosis vs. fractal_dimension_se

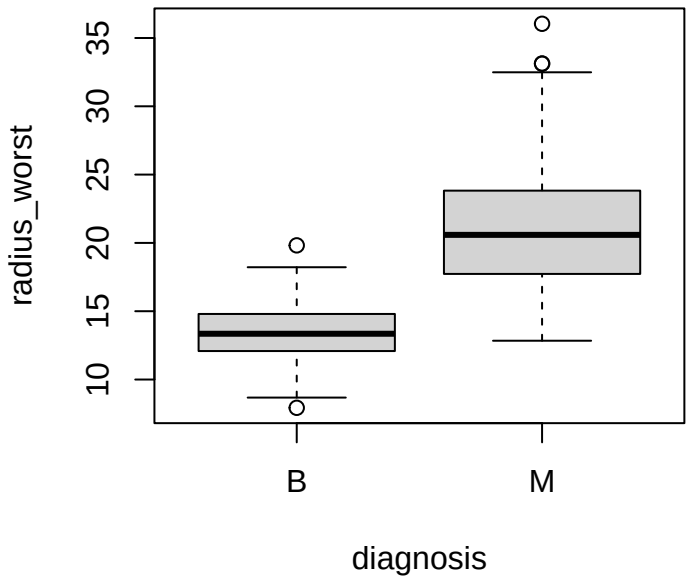


Figure 21: Boxplot of Diagnosis vs. radius_worst

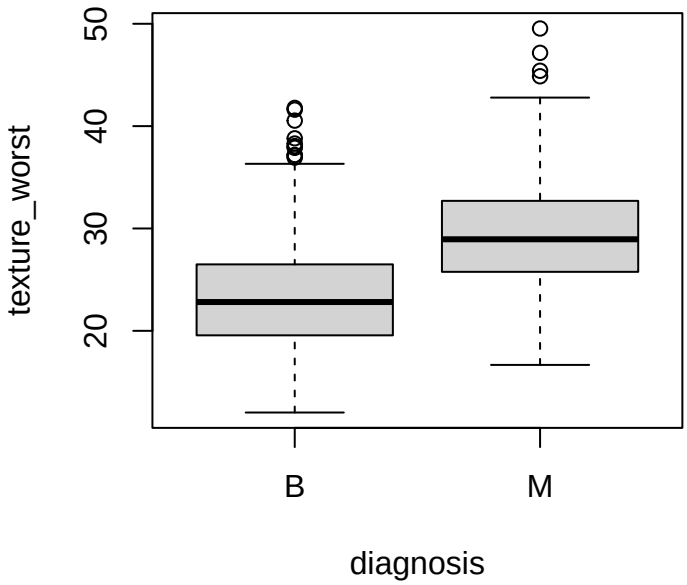


Figure 22: Boxplot of Diagnosis vs. texture_worst

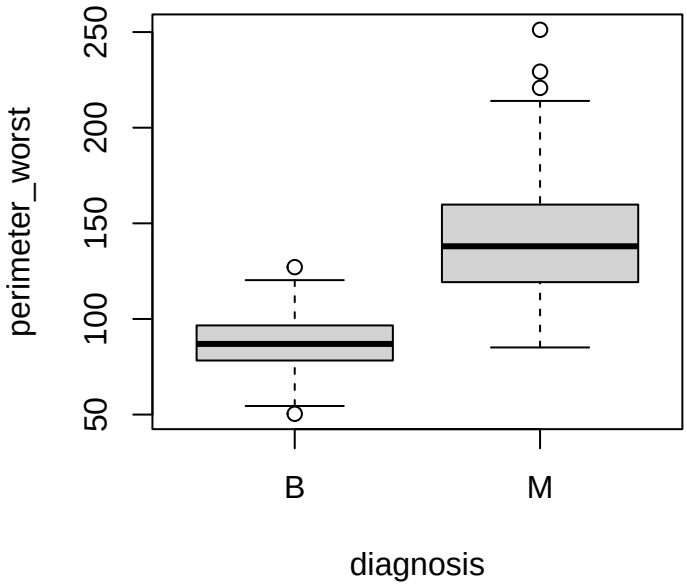


Figure 23: Boxplot of Diagnosis vs. perimeter_worst

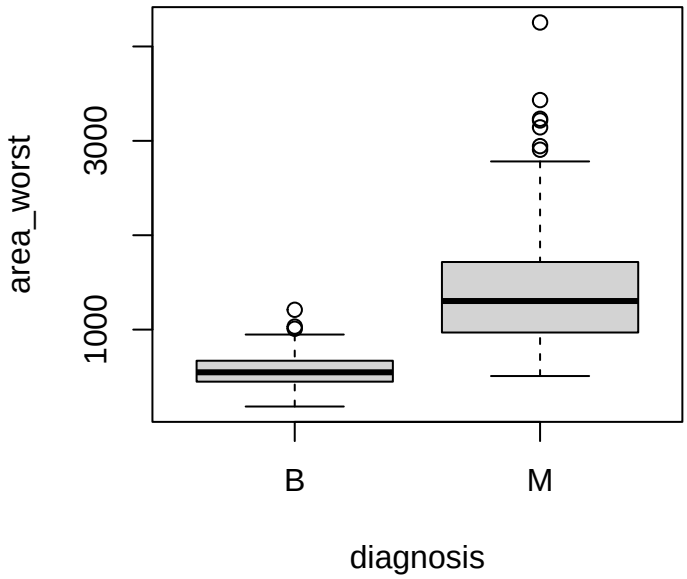


Figure 24: Boxplot of Diagnosis vs. area_worst

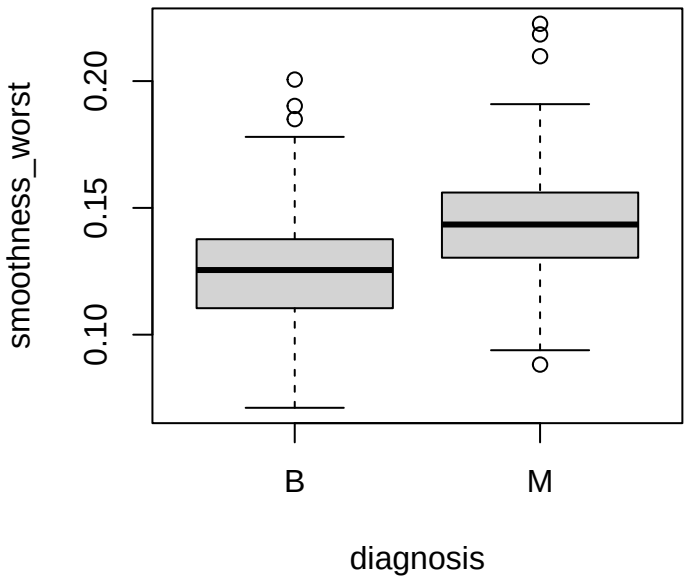


Figure 25: Boxplot of Diagnosis vs. smoothness_worst

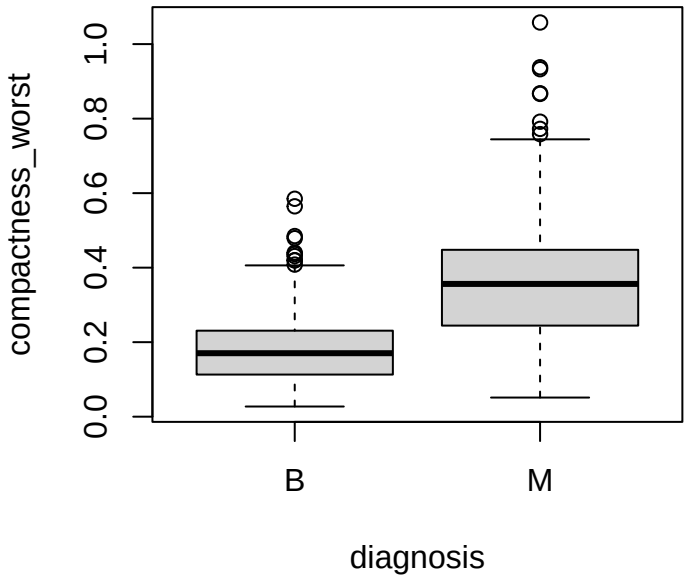


Figure 26: Boxplot of Diagnosis vs. compactness_worst

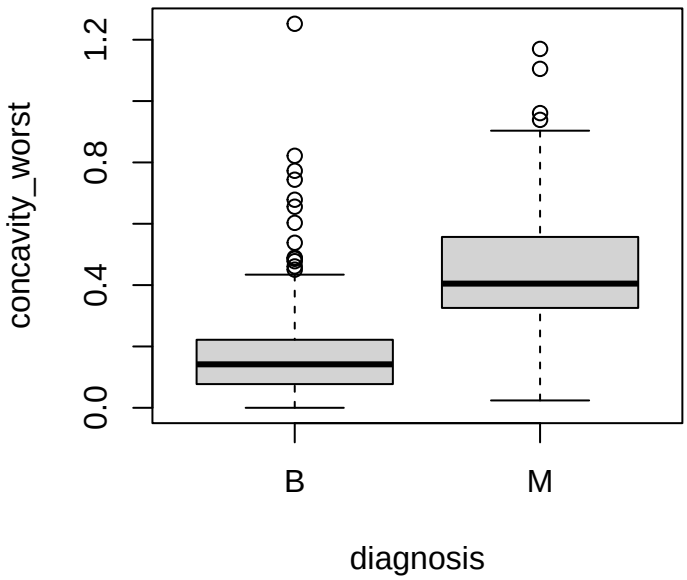


Figure 27: Boxplot of Diagnosis vs. concavity_worst

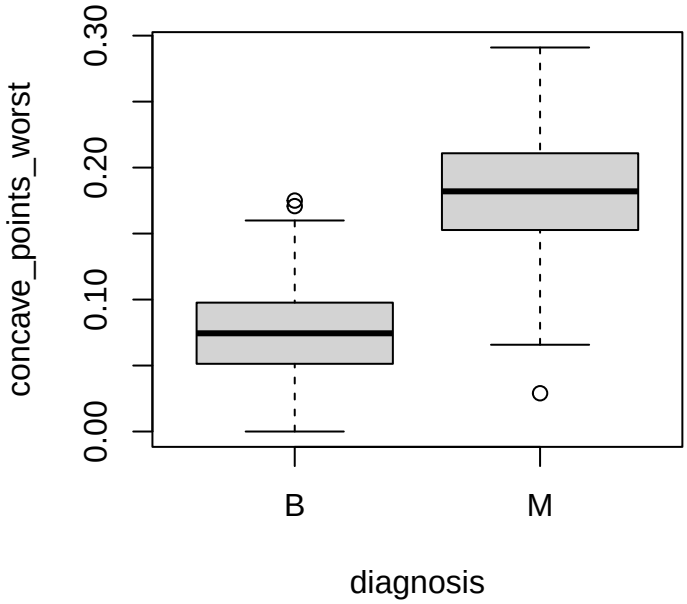


Figure 28: Boxplot of Diagnosis vs. concave_points_worst

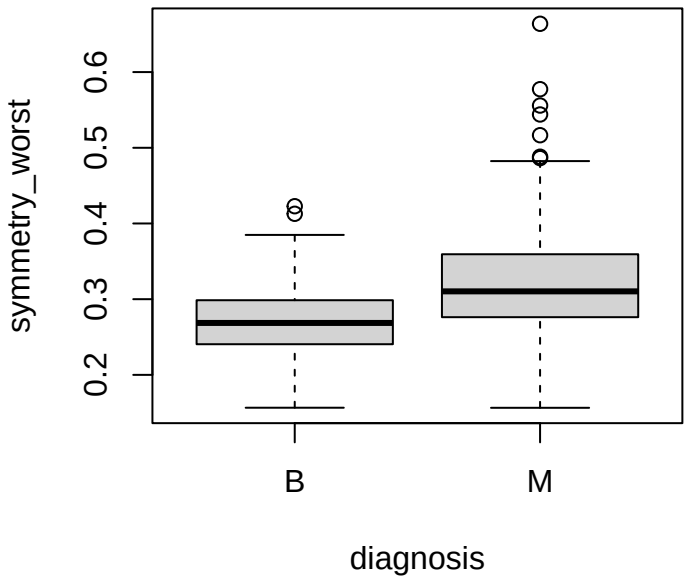


Figure 29: Boxplot of Diagnosis vs. symmetry_worst

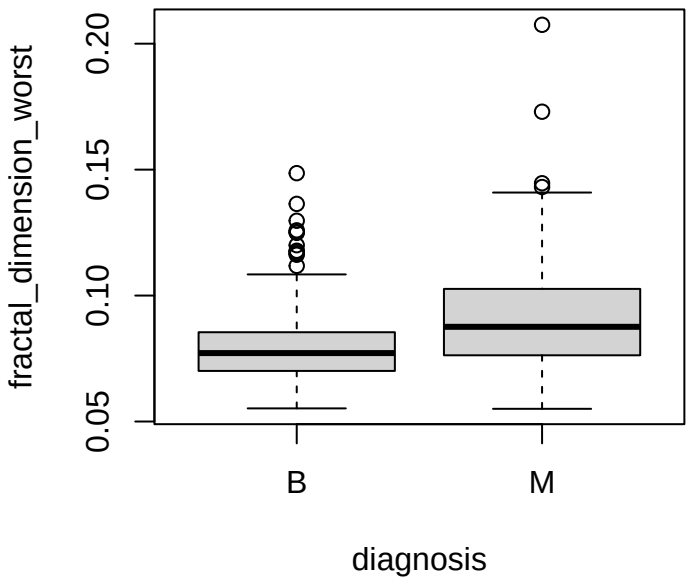


Figure 30: Boxplot of Diagnosis vs. fractal_dimension_worst

- Histogram

```
for (col in colnames(data)){
  if(col != 'diagnosis'){
    hist(data[, col], xlab=col, main = NULL)
  }
  else{
    barplot(table(data[, col]), xlab=col, ylab='frequency', main = NULL)
  }
}
```

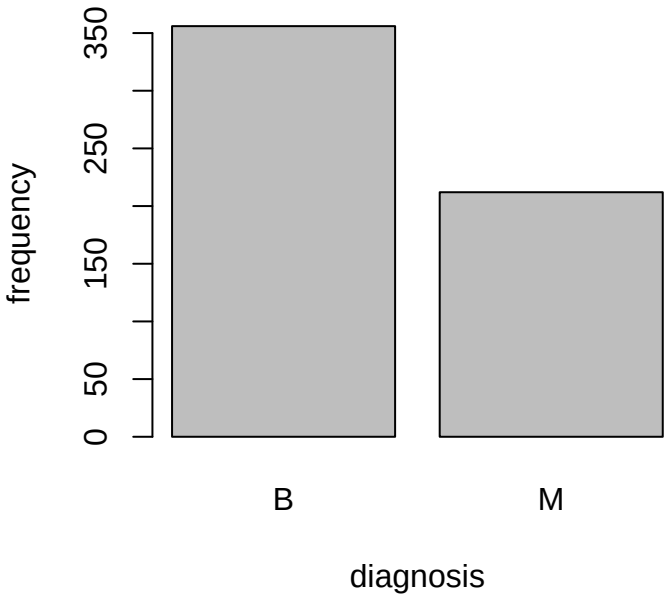


Figure 31: Histogram of diagnosis

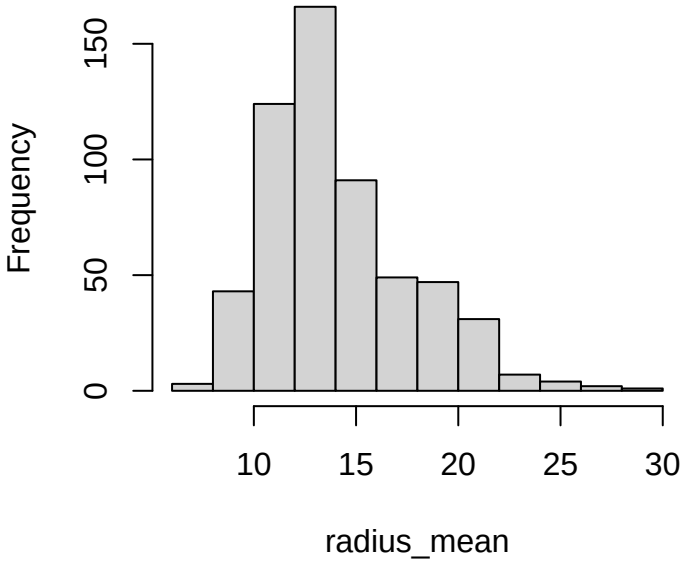


Figure 32: Histogram of radius_mean

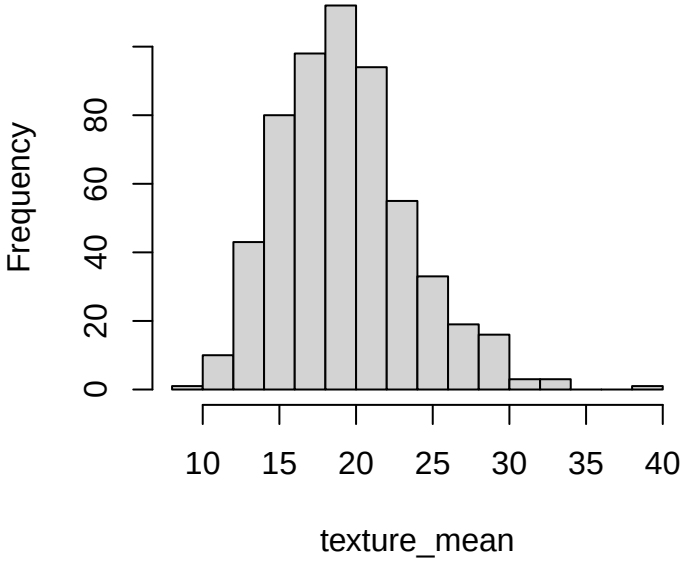


Figure 33: Histogram of texture_mean

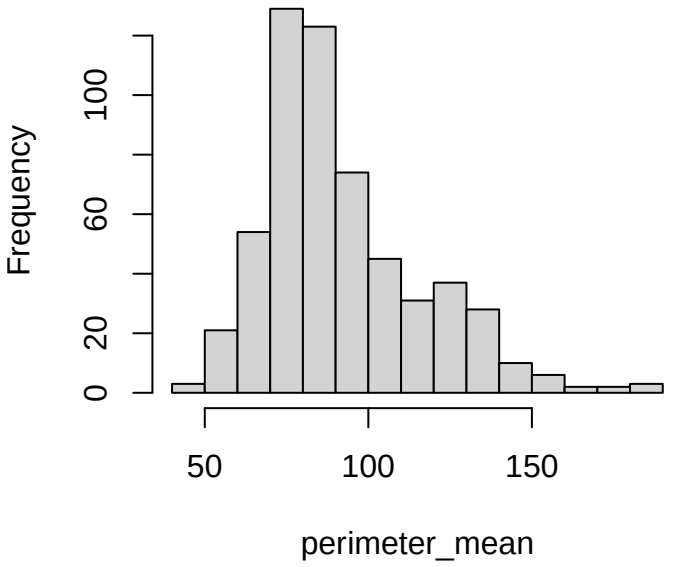


Figure 34: Histogram of perimeter_mean

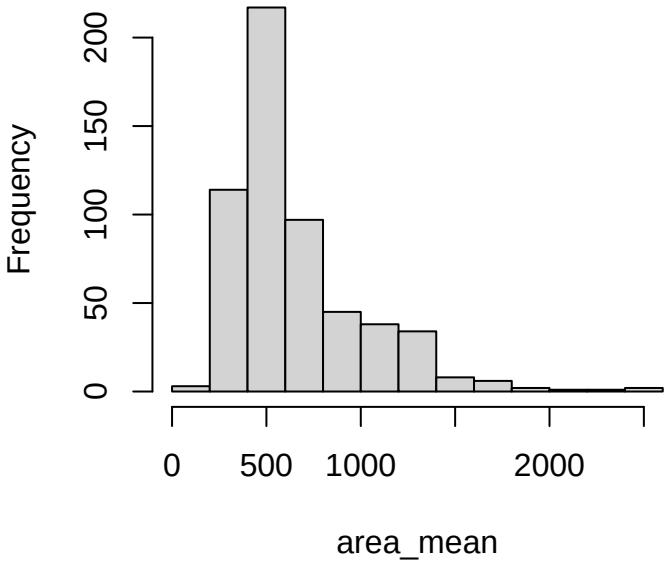


Figure 35: Histogram of area_mean

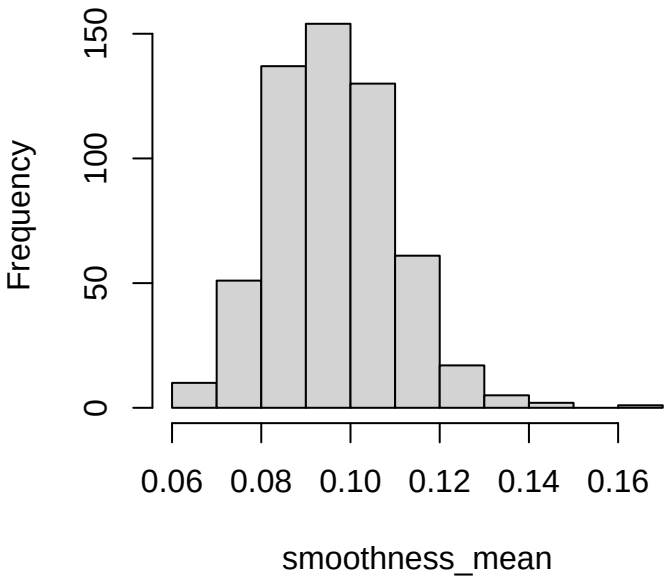


Figure 36: Histogram of smoothness_mean

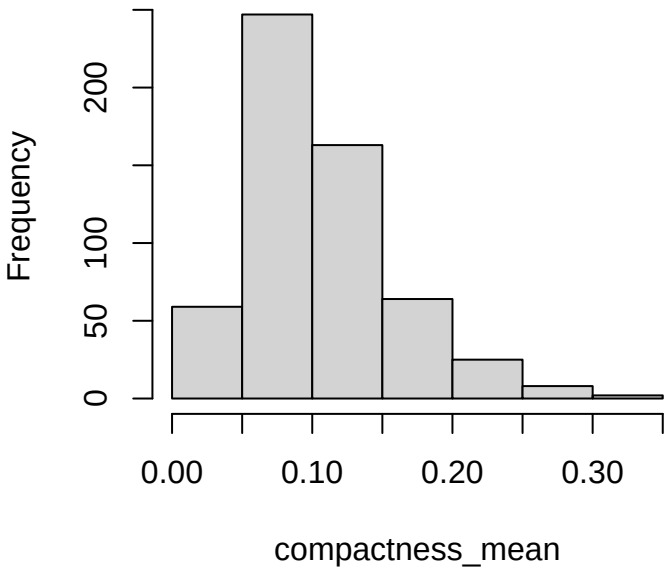


Figure 37: Histogram of compactness_mean

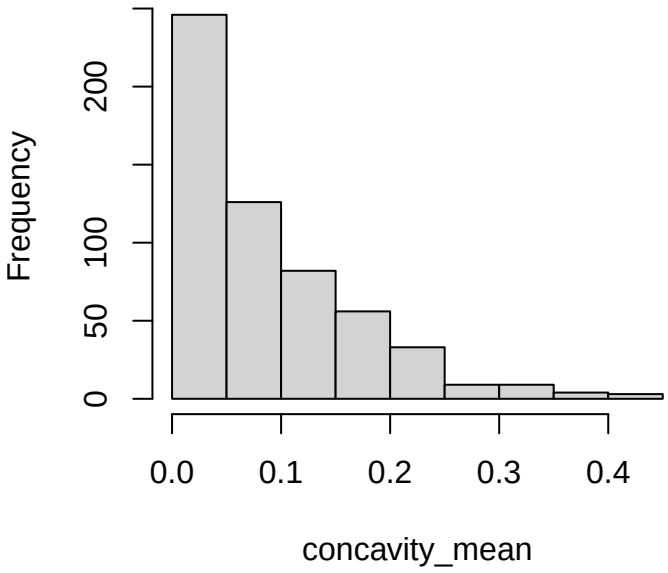


Figure 38: Histogram of concavity_mean

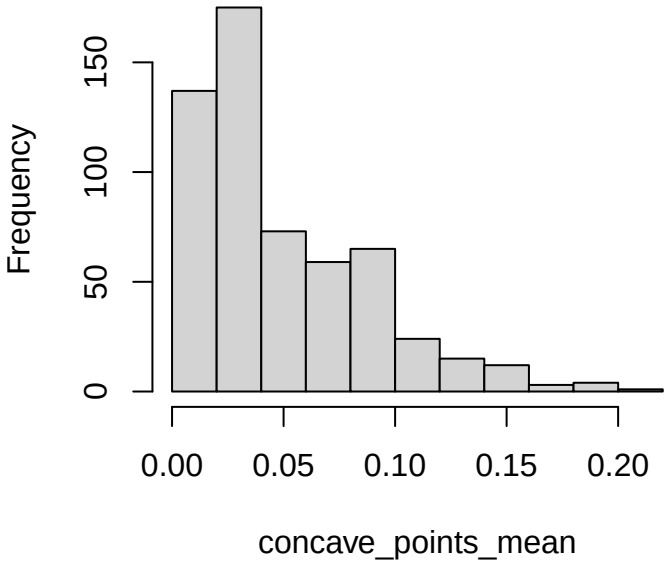


Figure 39: Histogram of concave_points_mean

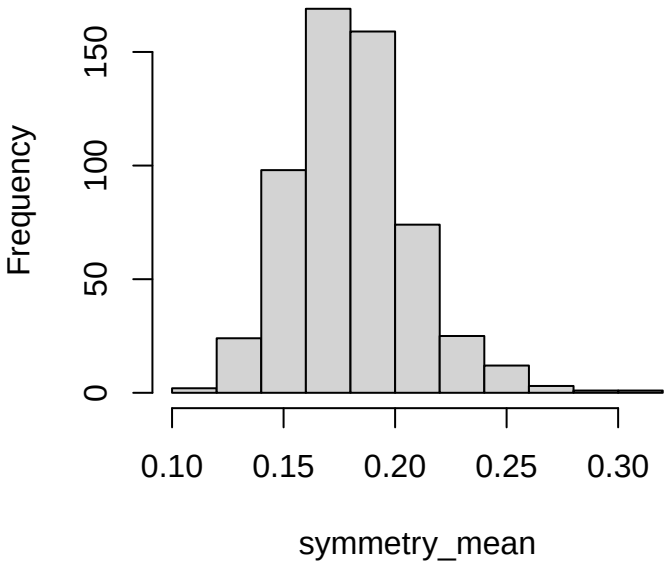


Figure 40: Histogram of symmetry_mean

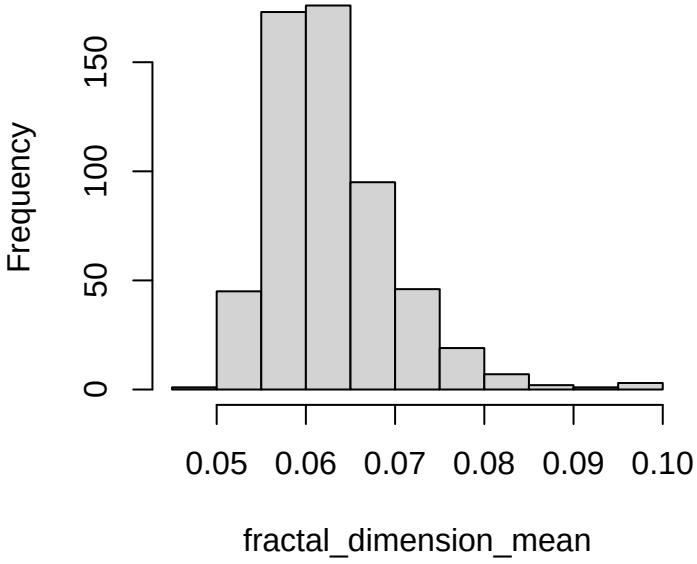


Figure 41: Histogram of fractal_dimension_mean

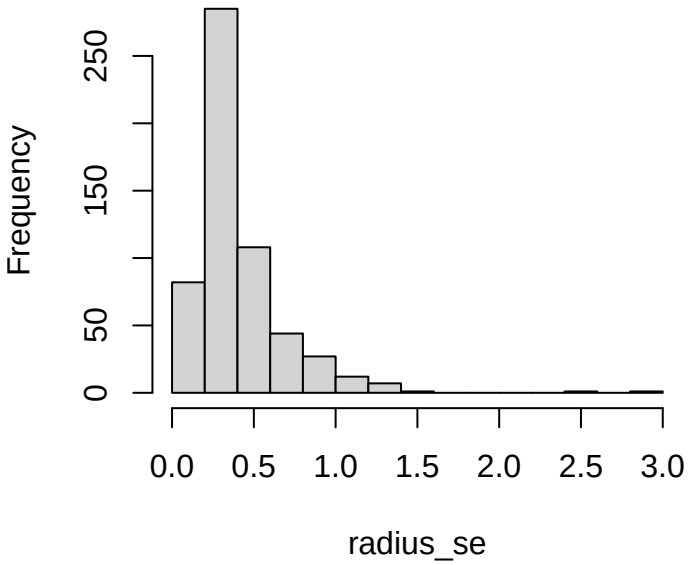


Figure 42: Histogram of radius_se

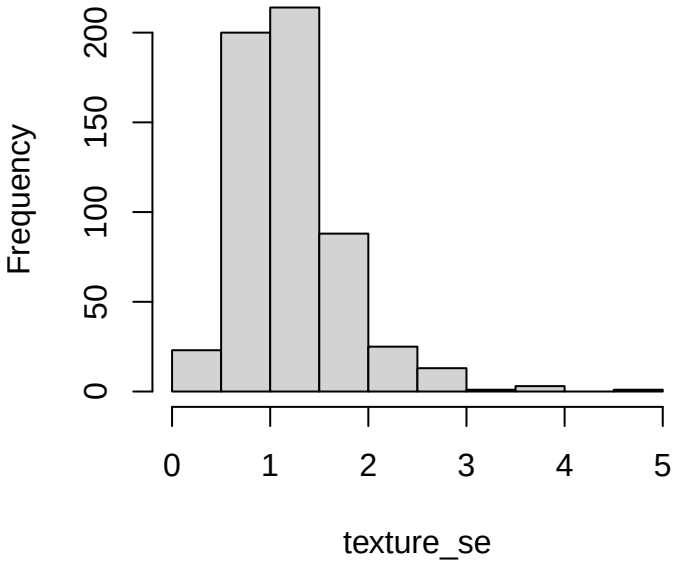


Figure 43: Histogram of texture_se

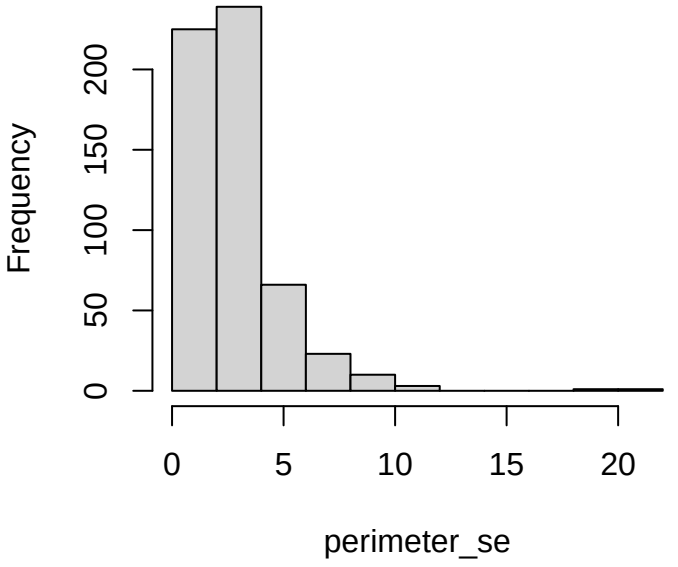


Figure 44: Histogram of perimeter_se

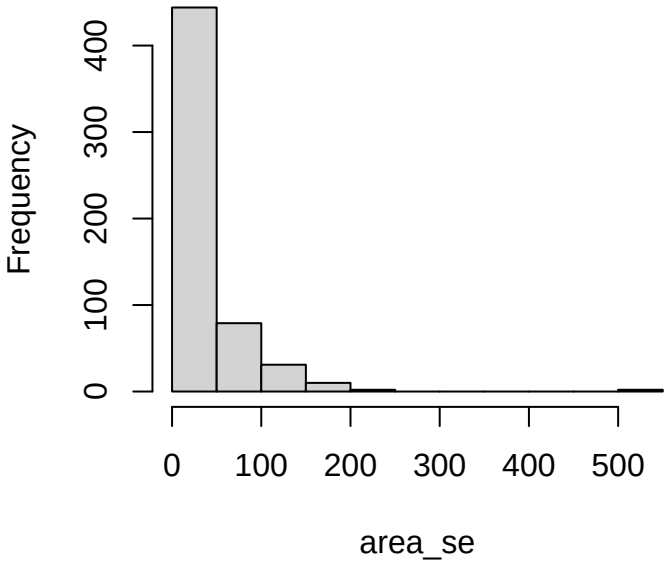


Figure 45: Histogram of area_se

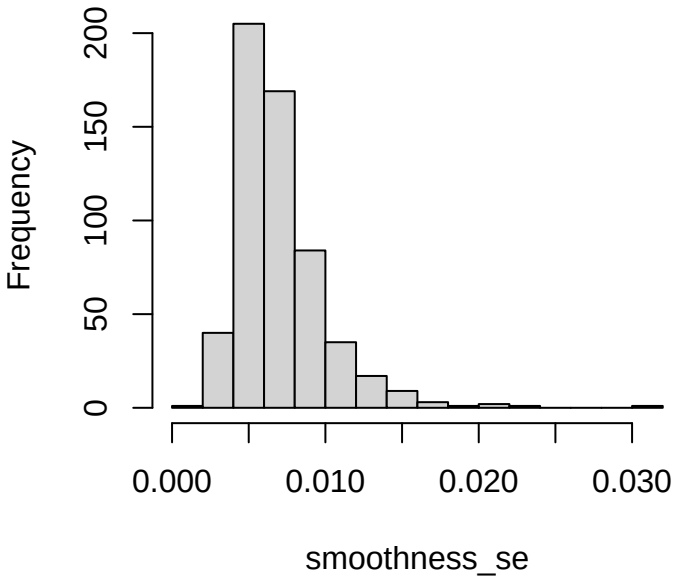


Figure 46: Histogram of smoothness_se

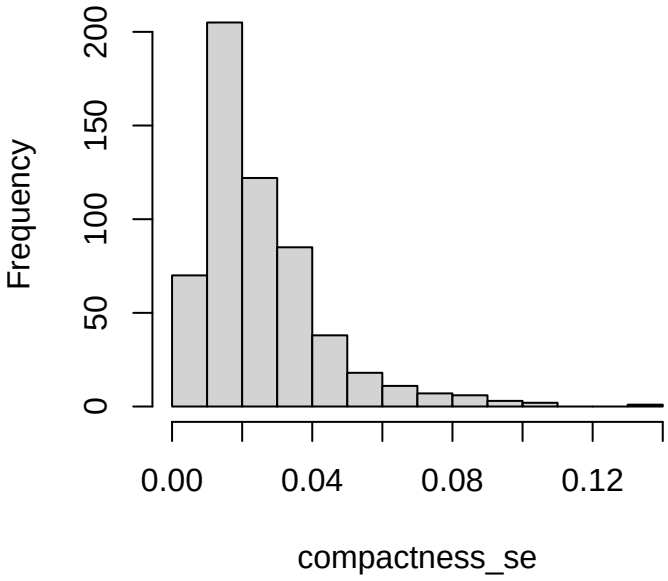


Figure 47: Histogram of compactness_se

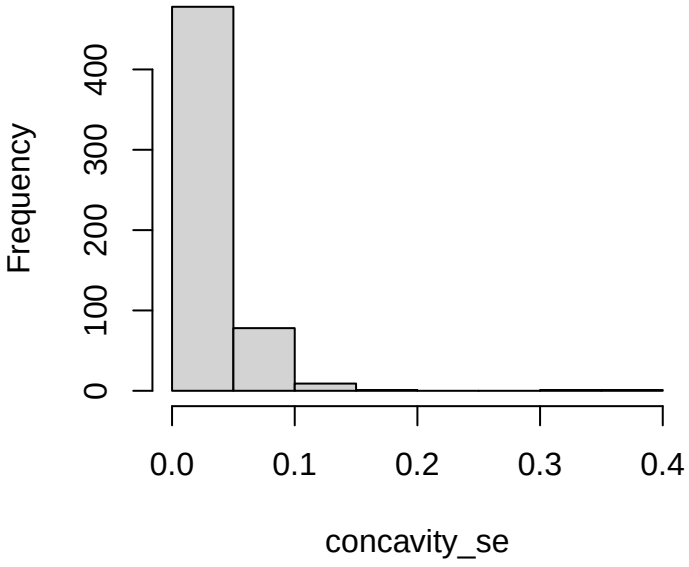


Figure 48: Histogram of concavity_se

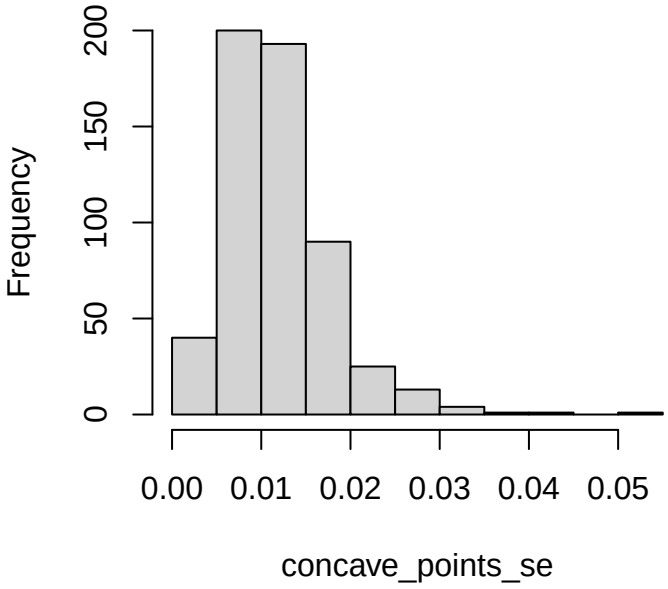


Figure 49: Histogram of concave_points_se

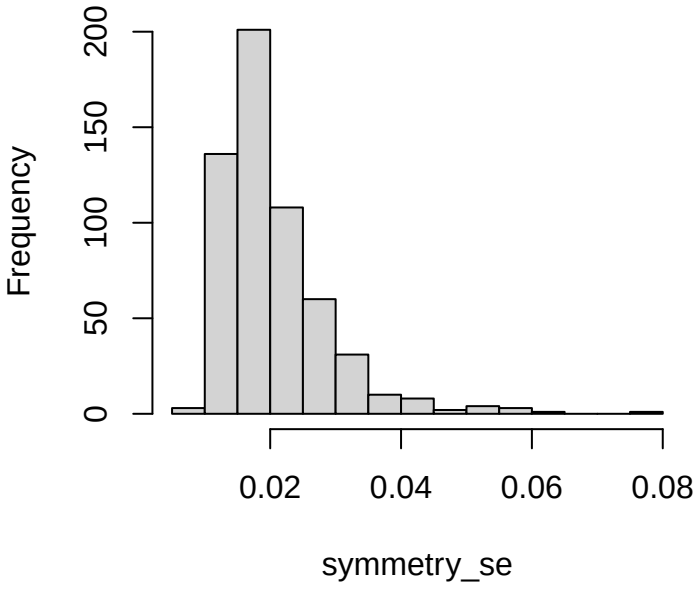


Figure 50: Histogram of symmetry_se

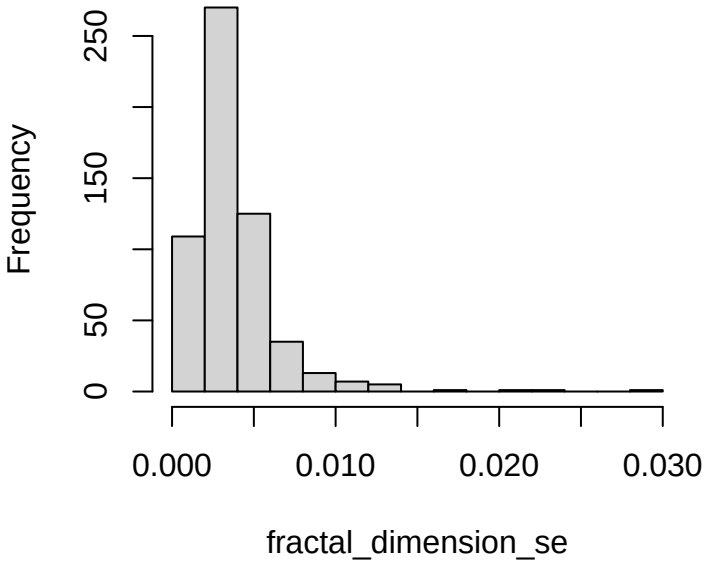


Figure 51: Histogram of fractal_dimension_se

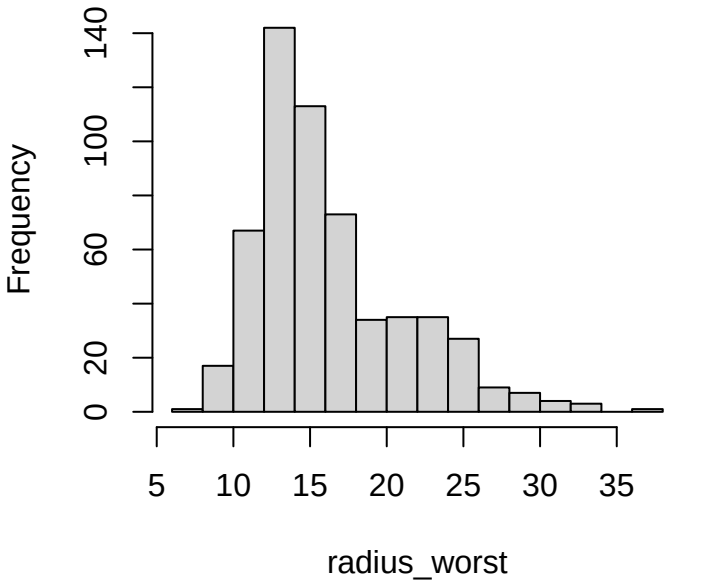


Figure 52: Histogram of radius_worst

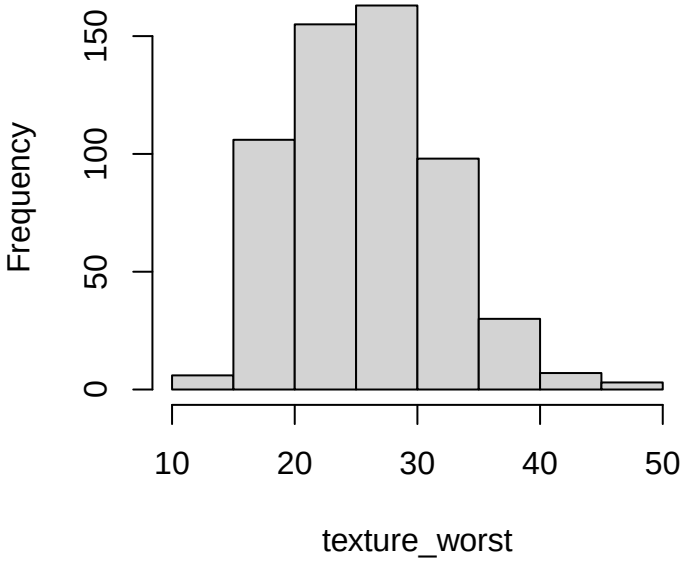


Figure 53: Histogram of texture_worst

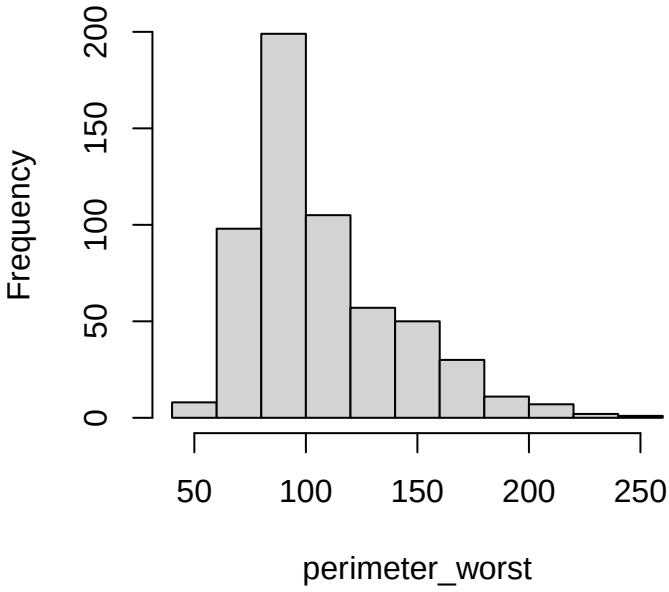


Figure 54: Histogram of perimeter_worst

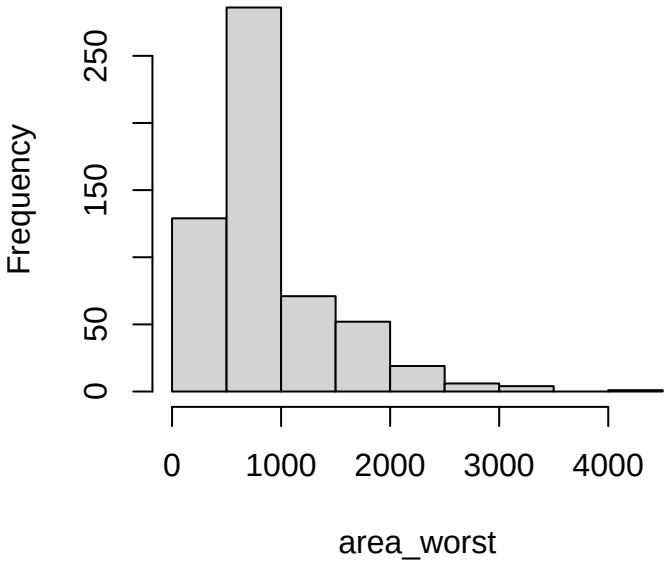


Figure 55: Histogram of area_worst

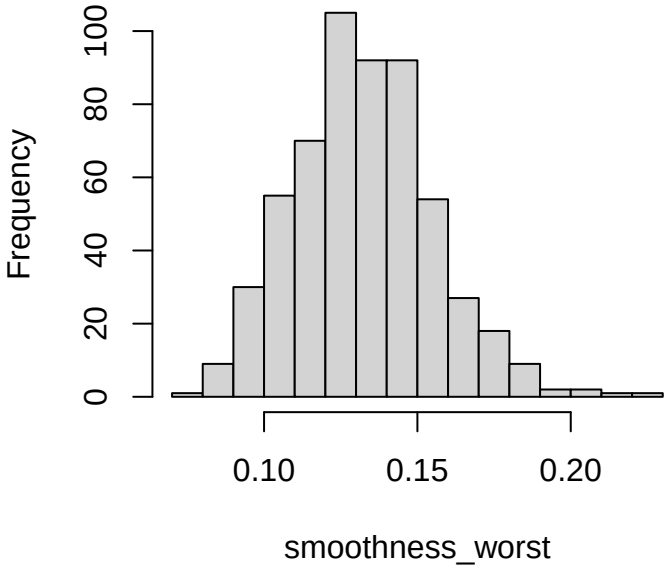


Figure 56: Histogram of smoothness_worst

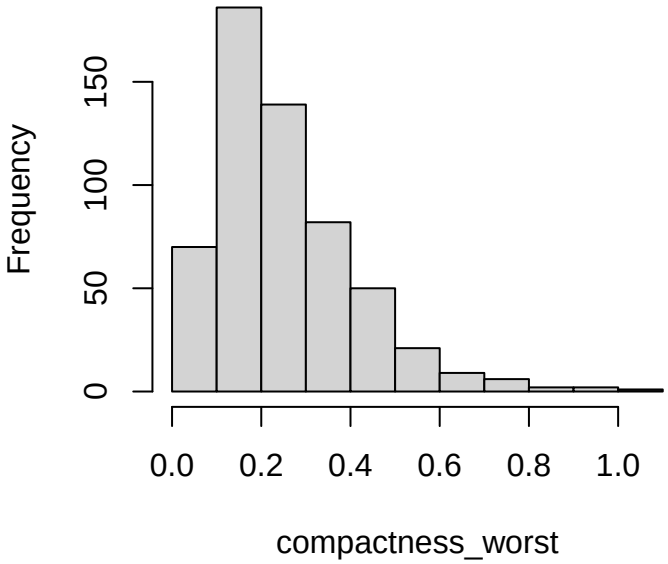


Figure 57: Histogram of compactness_worst

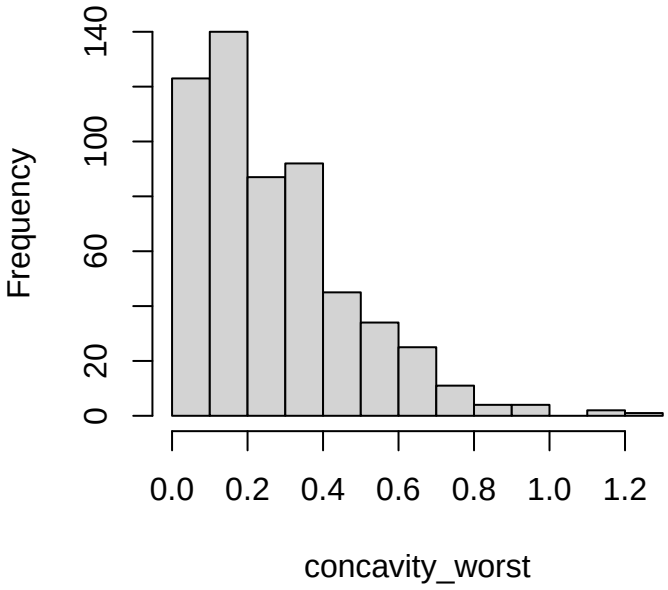


Figure 58: Histogram of concavity_worst

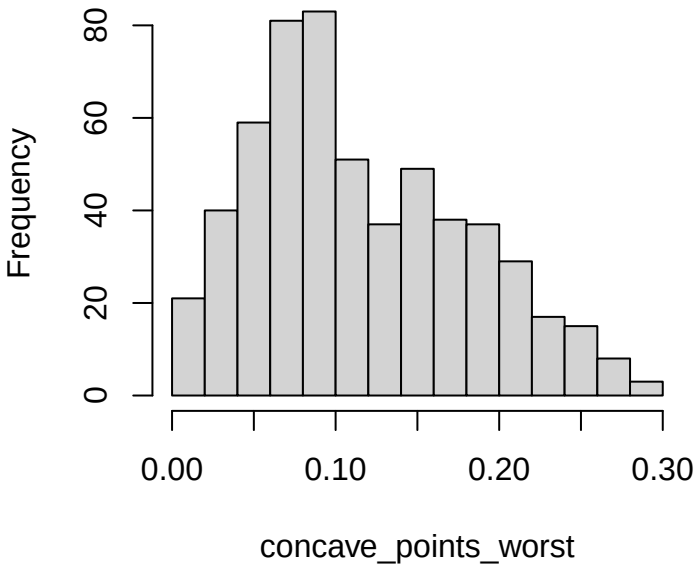


Figure 59: Histogram of concave_points_worst

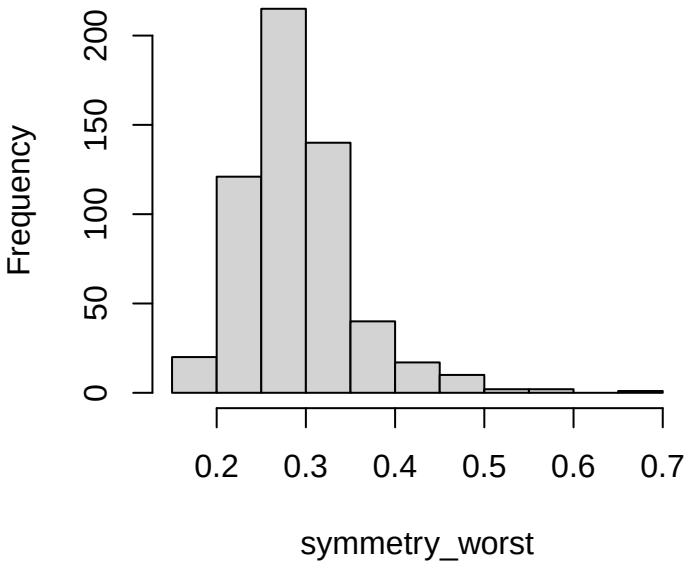


Figure 60: Histogram of symmetry_worst

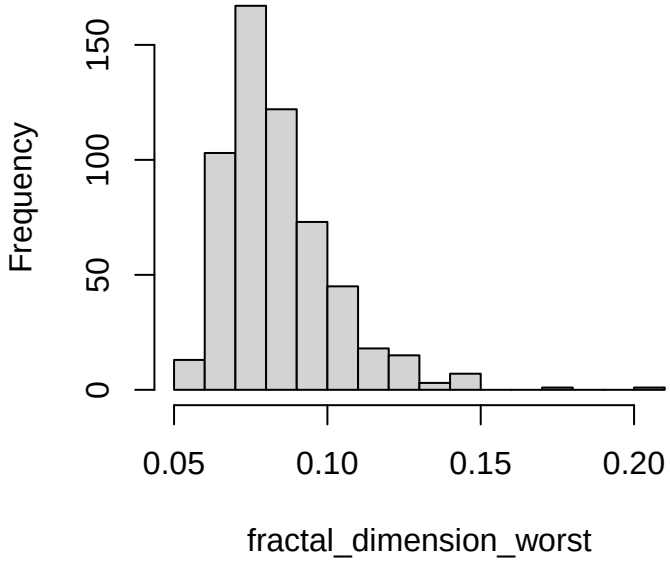


Figure 61: Histogram of fractal_dimension_worst

- Correlation matrix

```
cor_matrix <- cor(data[2:31])
corrplot(cor_matrix, method = "color", tl.cex = 0.5,
         title = "Correlation Heatmap", mar = c(0,0,1,0))
```

2.3 Principal Component Analysis (PCA)

```
res.pca <- PCA(data, scale.unit=T, quali.sup="diagnosis", graph=F)
```

2.3.1 Eigen values

```
res.pca$eig
```

```
##           eigenvalue percentage of variance cumulative percentage of variance
## comp 1  1.327110e+01           4.423701e+01                           44.23701
```

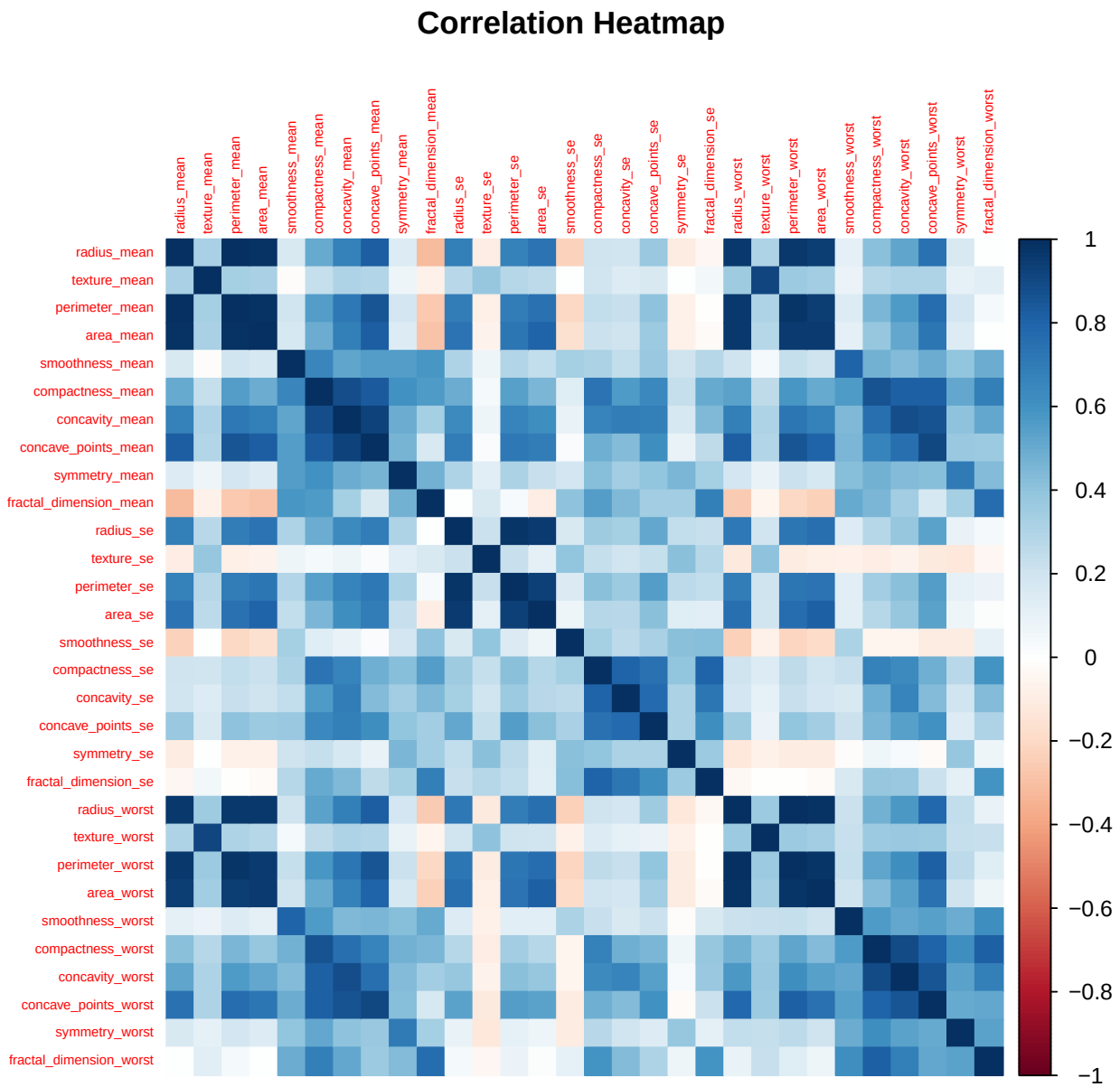
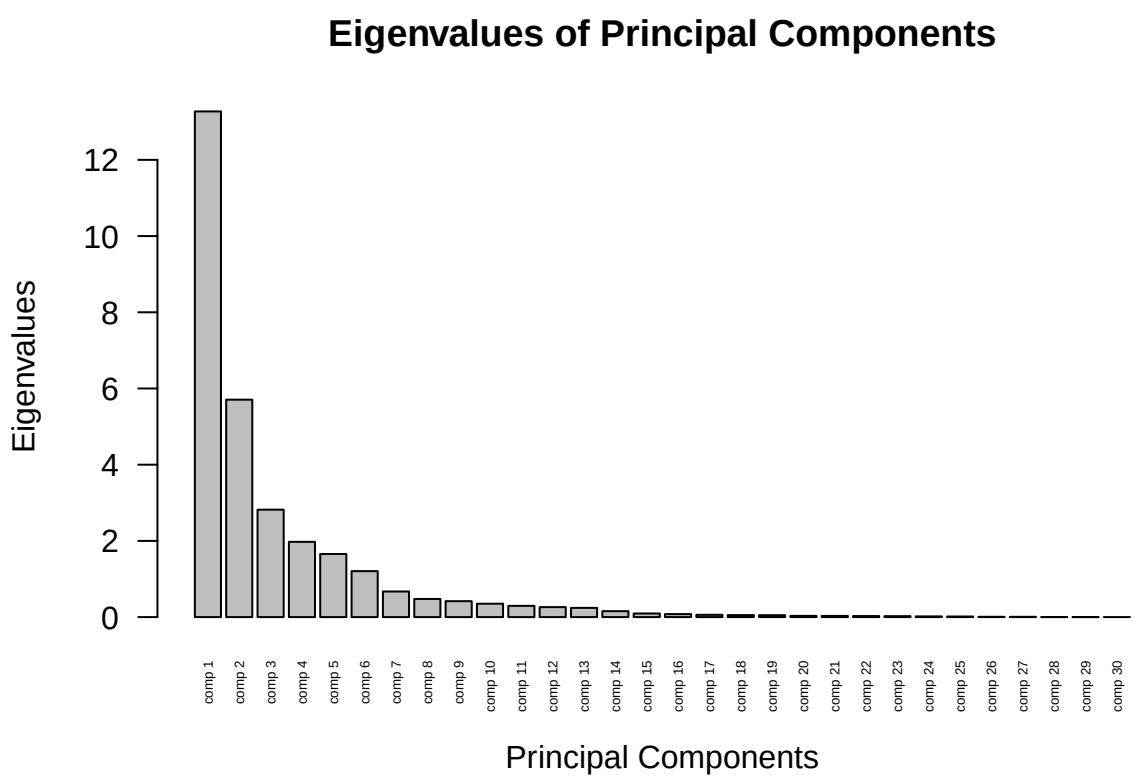


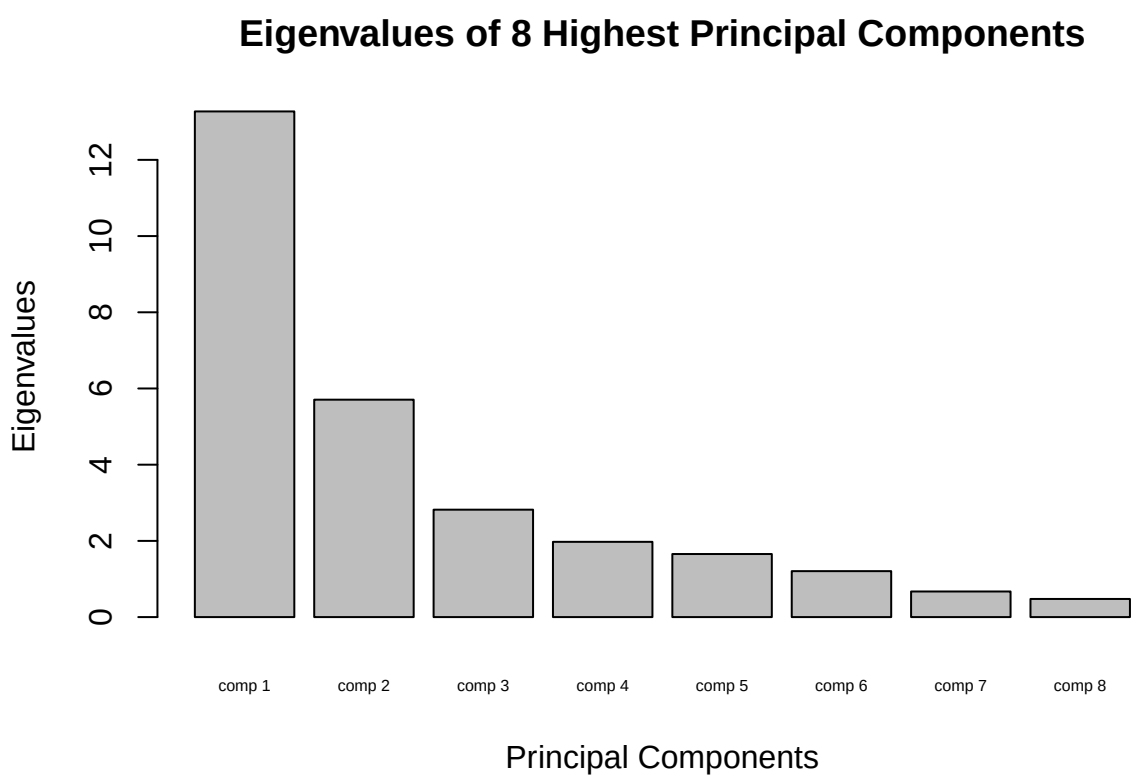
Figure 62: Correlation plot of breast cancer dataset

```
## comp 2  5.705847e+00          1.901949e+01                    63.25650
## comp 3  2.818852e+00          9.396173e+00                    72.65267
## comp 4  1.975260e+00          6.584200e+00                    79.23687
## comp 5  1.655397e+00          5.517989e+00                    84.75486
## comp 6  1.205956e+00          4.019852e+00                    88.77471
## comp 7  6.715595e-01          2.238532e+00                    91.01324
## comp 8  4.757283e-01          1.585761e+00                    92.59900
## comp 9  4.175436e-01          1.391812e+00                    93.99082
## comp 10 3.512424e-01          1.170808e+00                    95.16162
## comp 11 2.946575e-01          9.821917e-01                    96.14382
## comp 12 2.618867e-01          8.729558e-01                    97.01677
## comp 13 2.413406e-01          8.044687e-01                    97.82124
## comp 14 1.553811e-01          5.179370e-01                    98.33918
## comp 15 9.422671e-02          3.140890e-01                    98.65327
## comp 16 7.852423e-02          2.617474e-01                    98.91501
## comp 17 5.937738e-02          1.979246e-01                    99.11294
## comp 18 5.280787e-02          1.760262e-01                    99.28897
## comp 19 4.953305e-02          1.651102e-01                    99.45408
## comp 20 3.117460e-02          1.039153e-01                    99.55799
## comp 21 2.988242e-02          9.960808e-02                    99.65760
## comp 22 2.737953e-02          9.126509e-02                    99.74886
## comp 23 2.442628e-02          8.142094e-02                    99.83028
## comp 24 1.806954e-02          6.023181e-02                    99.89052
## comp 25 1.551898e-02          5.172994e-02                    99.94225
## comp 26 7.973523e-03          2.657841e-02                    99.96882
## comp 27 6.881183e-03          2.293728e-02                    99.99176
## comp 28 1.594249e-03          5.314165e-03                    99.99708
## comp 29 7.440627e-04          2.480209e-03                    99.99956
## comp 30 1.330431e-04          4.434769e-04                   100.00000
```

```r
# Draw eigenvalues dist
barplot(
  res.pca$eig[, 1],
  main = "Eigenvalues of Principal Components",
  xlab = "Principal Components",
  ylab = "Eigenvalues",
  cex.names = 0.4,  # Adjust size of the labels if necessary
  las = 2
)
```




```r
# Draw eigenvalues dist
barplot(
  res.pca$eig[, 1][1:8],
  main = "Eigenvalues of 8 Highest Principal Components",
  xlab = "Principal Components",
  ylab = "Eigenvalues",
  cex.names = 0.5,  # Adjust size of the labels if necessary
)
```



2.3.2 Dimension Description

```
dimdesc(res.pca)
```

```
## $Dim.1
##
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##                          correlation       p.value
## concave_points_mean        0.9505468 1.314118e-289
## concavity_mean             0.9415507 1.247666e-269
## concave_points_worst       0.9139173 8.506361e-224
## compactness_mean           0.8715008 2.712857e-177
## perimeter_worst            0.8621240 3.000809e-169
## concavity_worst            0.8329758 1.331367e-147
## radius_worst               0.8305458 5.466752e-146
## perimeter_mean             0.8287401 8.320032e-145
## area_worst                 0.8193953 6.686004e-139
## area_mean                  0.8049936 1.930056e-130
```

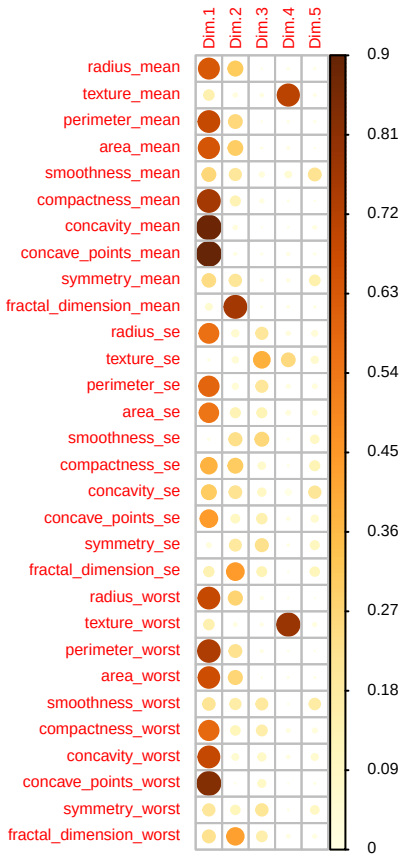
```
## radius_mean               0.7971433 4.025951e-126
## perimeter_se              0.7714610 3.170520e-113
## compactness_worst         0.7647292 4.026779e-110
## radius_se                 0.7521709 1.335627e-104
## area_se                   0.7399265  1.577473e-99
## concave_points_se         0.6666852  2.837226e-74
## compactness_se            0.6195687  1.622660e-61
## concavity_se              0.5583105  7.345259e-48
## smoothness_mean           0.5165531  4.608556e-40
## symmetry_mean             0.5024225  1.163533e-37
## fractal_dimension_worst   0.4790183  6.393386e-34
## smoothness_worst          0.4634294  1.384265e-31
## symmetry_worst            0.4485108  1.846776e-29
## texture_worst             0.3842744  1.989967e-21
## texture_mean              0.3831287  2.671960e-21
## fractal_dimension_se      0.3731811  3.288497e-20
## fractal_dimension_mean    0.2330620  1.911297e-08
## symmetry_se               0.1572424  1.680921e-04
##
## Link between the variable and the categorical variable (1-way anova)
## ============================================
##                  R2       p.value
## diagnosis 0.6171556 4.216856e-120
##
## Link between variable and the categories of the categorical variables
## ============================================================
##              Estimate       p.value
## diagnosis=M  2.958532 4.216856e-120
## diagnosis=B -2.958532 4.216856e-120
##
## $Dim.2
##
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##                          correlation       p.value
## fractal_dimension_mean    0.87492065 2.191906e-180
## fractal_dimension_se      0.66824677  9.768152e-75
## fractal_dimension_worst   0.65737896  1.429786e-71
## compactness_se            0.55574175  2.387444e-47
## smoothness_se             0.48805869  2.484520e-35
## concavity_se              0.47081194  1.122432e-32
## symmetry_mean             0.45479985  2.417092e-30
## smoothness_mean           0.44770058  2.392479e-29
## symmetry_se               0.43939632  3.266079e-28
## smoothness_worst          0.41261927  9.279509e-25
```

```
## compactness_mean         0.36338799  3.580680e-19
## compactness_worst        0.34329935  3.732597e-17
## symmetry_worst           0.33908655  9.486306e-17
## concave_points_se        0.31173145  2.886243e-14
## concavity_worst          0.23436182  1.585122e-08
## texture_se               0.21514945  2.254141e-07
## concavity_mean           0.14431004  5.609971e-04
## concave_points_mean     -0.08237061  4.974748e-02
## texture_worst           -0.10748266  1.036544e-02
## texture_mean            -0.14142132  7.247624e-04
## perimeter_se            -0.21233056  3.262312e-07
## radius_se               -0.25066659  1.379158e-09
## area_se                 -0.36248993  4.439028e-19
## perimeter_worst         -0.47737009  1.143880e-33
## perimeter_mean          -0.51480559  9.263446e-40
## area_worst              -0.52333894  2.946538e-41
## radius_worst            -0.52513529  1.408098e-41
## area_mean               -0.55189199  1.370460e-46
## radius_mean             -0.55952482  4.192226e-48
##
## Link between the variable and the categorical variable (1-way anova)
## ============================================
##                   R2      p.value
## diagnosis 0.03524886 6.644034e-06
##
## Link between variable and the categories of the categorical variables
## ============================================================
##               Estimate      p.value
## diagnosis=B  0.4636155 6.644034e-06
## diagnosis=M -0.4636155 6.644034e-06
##
## $Dim.3
##
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##                          correlation      p.value
## texture_se                 0.6253207 5.881731e-63
## smoothness_se              0.5180705 2.505229e-40
## symmetry_se                0.4824945 1.855034e-34
## radius_se                  0.4491576 1.501158e-29
## perimeter_se               0.4462059 3.850166e-29
## concave_points_se          0.3815011 4.053048e-21
## area_se                    0.3618910 5.121051e-19
## fractal_dimension_se       0.3556920 2.208906e-18
## concavity_se               0.2981390 4.002892e-13
```

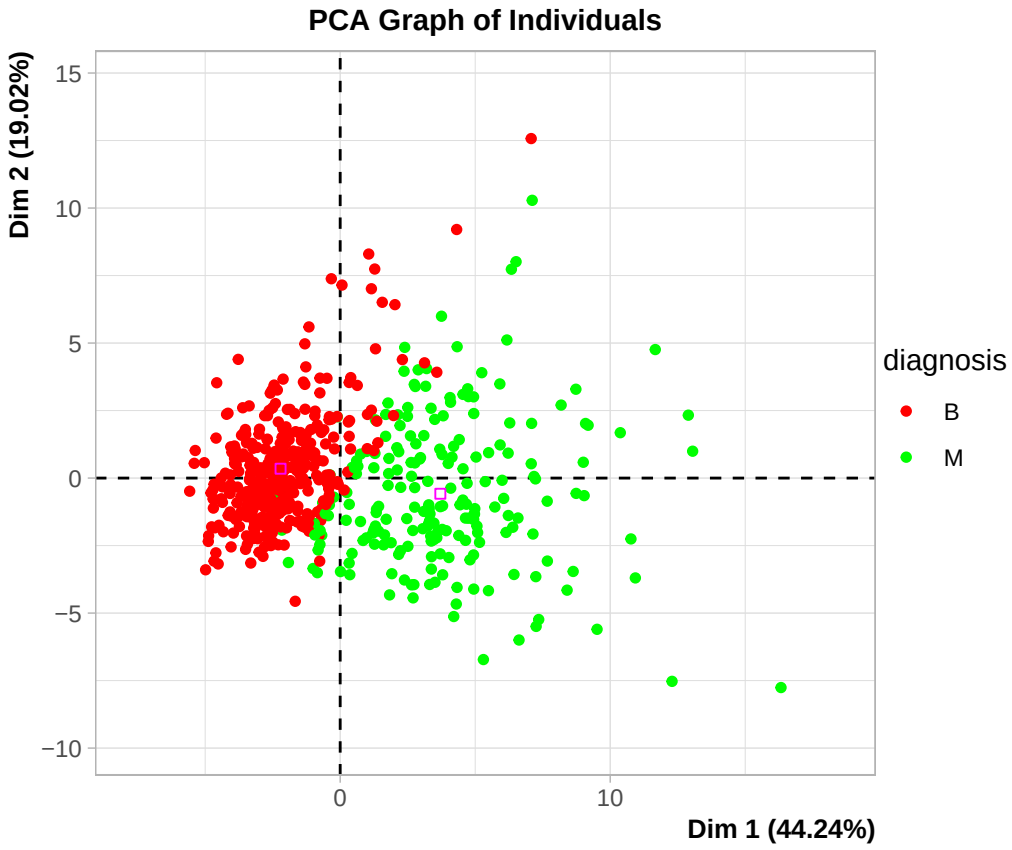
```
## compactness_se             0.2610832 2.640510e-10
## texture_mean               0.1009709 1.607242e-02
## compactness_mean          -0.1248861 2.868368e-03
## smoothness_mean           -0.1731452 3.342047e-05
## concave_points_worst      -0.2861625 3.629808e-12
## concavity_worst           -0.2912819 1.432548e-12
## fractal_dimension_worst   -0.3920670 2.599307e-22
## compactness_worst         -0.3977071 5.756553e-23
## smoothness_worst          -0.4370834 6.678160e-28
## symmetry_worst            -0.4583000 7.652077e-31
##
## Link between the variable and the categorical variable (1-way anova)
## ============================================
##                    R2      p.value
## diagnosis 0.02776489 6.596607e-05
##
## Link between variable and the categories of the categorical variables
## ============================================================
##                Estimate      p.value
## diagnosis=B  0.2892075 6.596607e-05
## diagnosis=M -0.2892075 6.596607e-05
```

2.3.3 PCA Graph

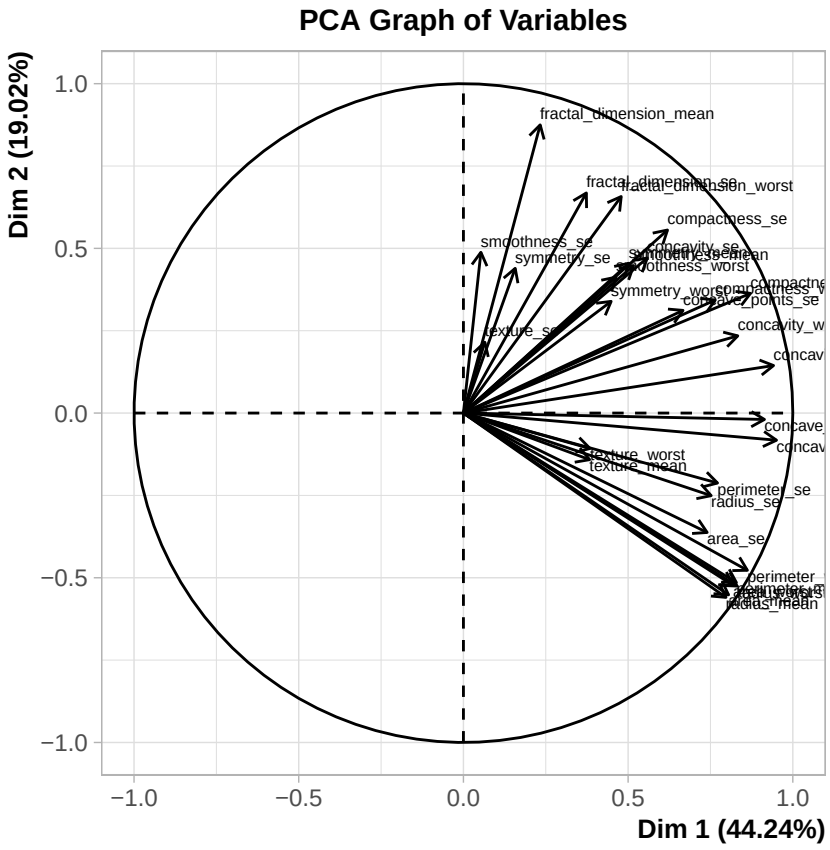
```r
corrplot(
  res.pca$var$cos2,
  is.corr = F,
  tl.cex = 0.5,
  cl.cex = 0.5,
  cl.pos = "r",
  cl.ratio = 0.5,  # Adjust ratio to add space to the legend
  cl.align.text = "l"  # Align the text to the left of the legend
)
```




```r
plot(
  res.pca,
  choix = c("ind"),
  hab = "diagnosis",
  invisible = c("var"),
  cex = 1,
  palette = c("red", "green"),
  autoLab = c("no"),
  title = "PCA Graph of Individuals",
  label = c("none")
)
```




```
plot(
  res.pca,
  choix = c("var"),
  hab = "none",
  invisible = c("ind"),
  cex = 0.5,
  autoLab = c("no"),
  title = "PCA Graph of Variables"
)
```



2.4 Hierarchical Clustering (HC)

One technique for studying clusters in the analysis of data is hierarchical clustering. Data is grouped using hierarchical clustering, which builds a hierarchy of groups. There are two approaches to construct this hierarchy: divisive (top-down), in which all data points begin in a single cluster and divide as we proceed down the hierarchy, or agglomerative (bottom-up), in which each data point forms a cluster and pairs of clusters are merged as we travel up the hierarchy. In this work, the linkage criterion is Ward's linkage, and we employ agglomerative hierarchical clustering. The linking criterion defines appropriate distance between sets of observations to use.

```
res.hcpc <- HCPC(res.pca, nb.clust=2, graph = FALSE)

fviz_dend(res.hcpc,
    cex = 0.7,
    labels_track_height = 0.8
)
```

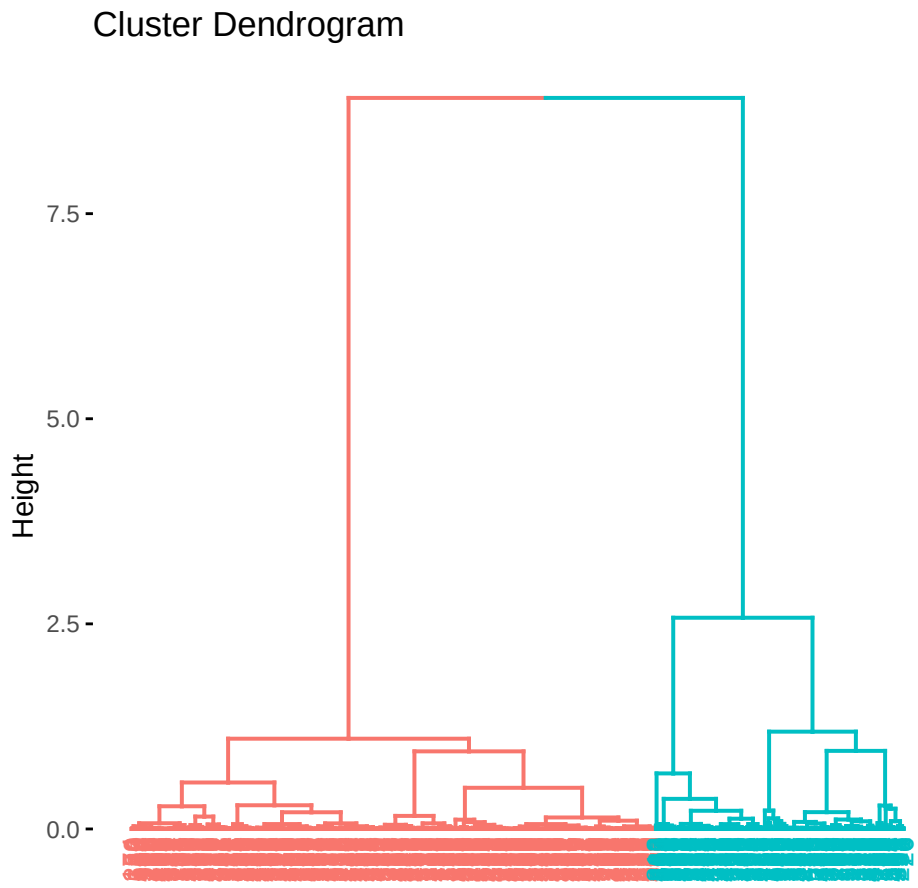


Figure 63: Hierarchical Clustering plot

```
df.hcpc <- res.hcpc$data.clust

plot(df.hcpc$diagnosis, df.hcpc$clust, xlab = 'diagnosis', ylab = '% cluster')
```

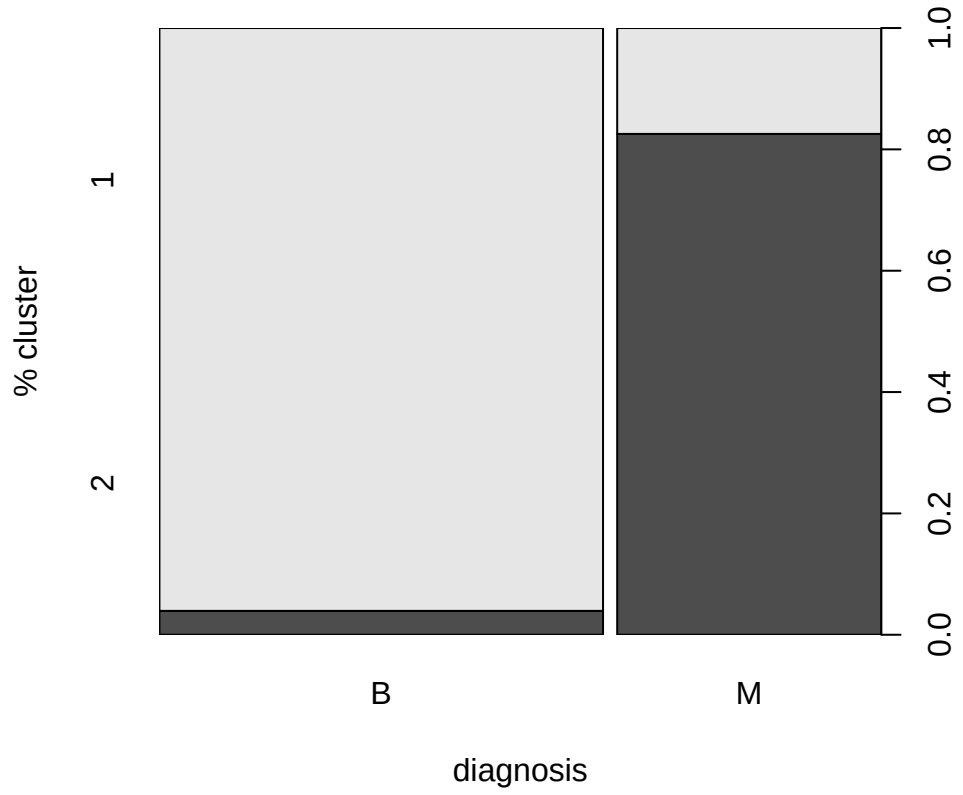


Figure 64: Bar plot of Type vs. cluster percent

```
plot(df.hcpc$clust, df.hcpc$diagnosis, xlab = 'cluster', ylab = '% diagnosis')
```

```
plot.HCPC(res.hcpc,choice='map',draw.tree=FALSE,title='Factor map')
```

```
res.hcpc$desc.var$quanti
```

```
## $`1`
##                            v.test Mean in category Overall mean sd in category
```

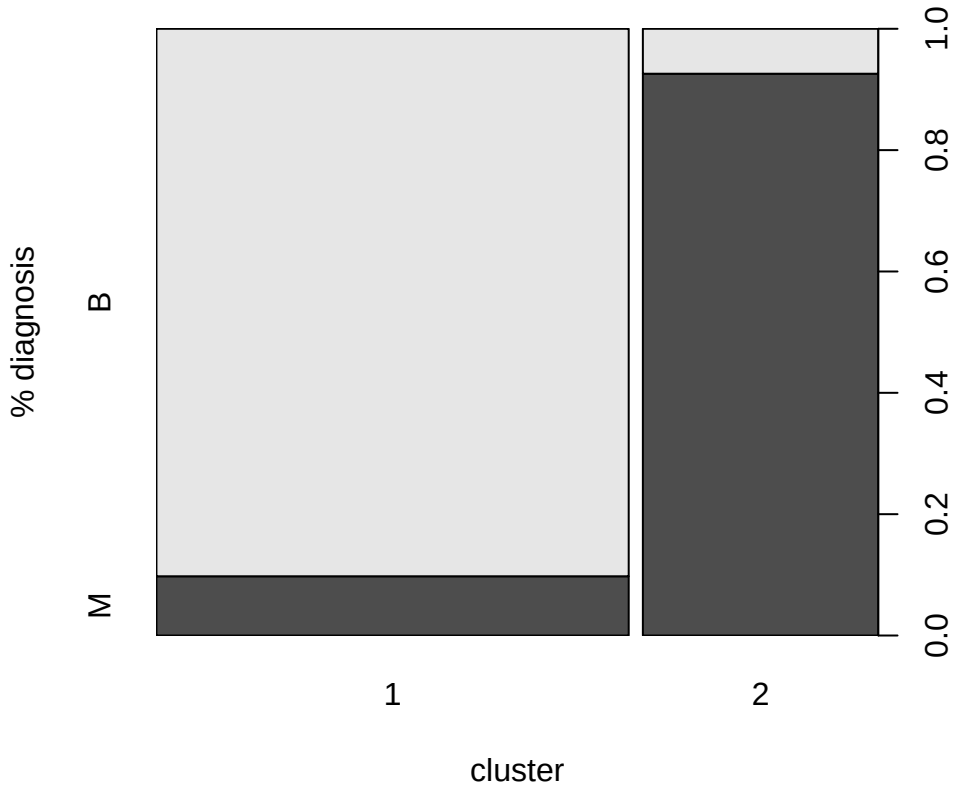


Figure 65: Hierarchical Clustering plot

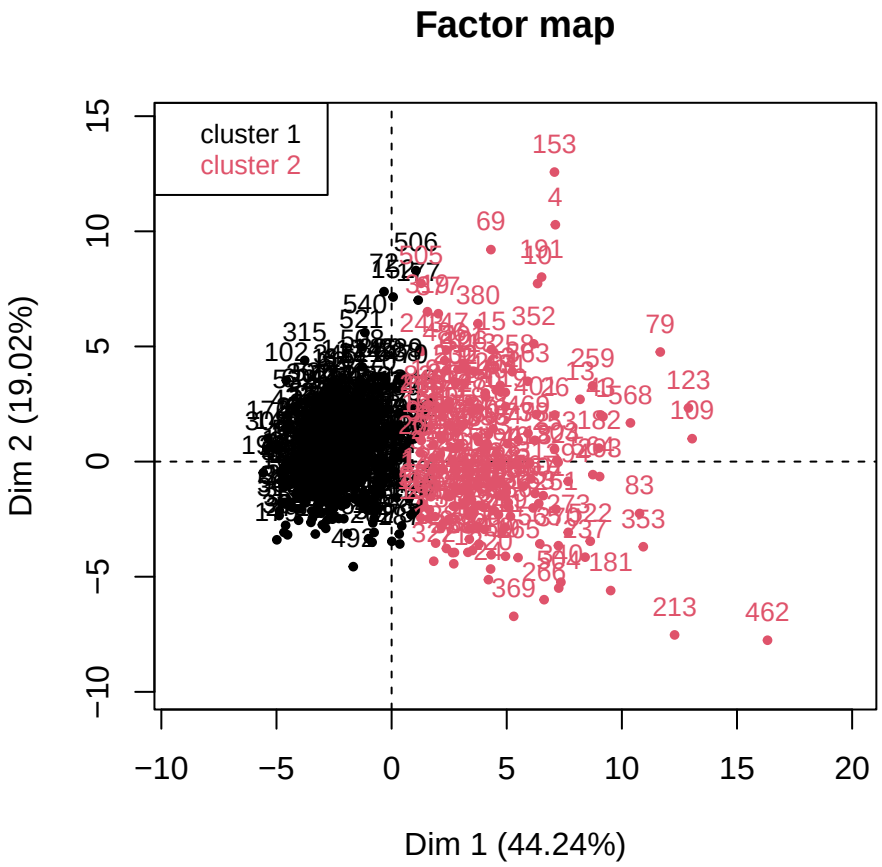


Figure 66: Bar plot of Type vs. cluster percent

```
## symmetry_se              -2.381940     1.994771e-02 2.053135e-02   6.948020e-03
## fractal_dimension_mean   -4.222150     6.192077e-02 6.280458e-02   5.936044e-03
## fractal_dimension_se     -6.962302     3.250402e-03 3.796685e-03   2.110742e-03
## texture_mean             -8.136485     1.824393e+01 1.928040e+01   4.041460e+00
## texture_worst            -8.533234     2.411435e+01 2.566896e+01   5.686349e+00
## symmetry_worst          -10.035456     2.716681e-01 2.900808e-01   4.412183e-02
## smoothness_worst        -10.195669     1.255605e-01 1.324433e-01   1.980582e-02
## smoothness_mean         -10.231294     9.220641e-02 9.643727e-02   1.245037e-02
## symmetry_mean           -10.249214     1.728736e-01 1.812014e-01   2.324094e-02
## fractal_dimension_worst -10.440732     7.838005e-02 8.396968e-02   1.258000e-02
## concavity_se            -10.679362     2.239898e-02 3.194987e-02   1.733584e-02
## compactness_se          -11.656645     1.933140e-02 2.551479e-02   1.278953e-02
## concave_points_se       -13.026611     9.440873e-03 1.181690e-02   4.528762e-03
## area_se                 -13.548575     2.209945e+01 4.037438e+01   1.013949e+01
## radius_se               -14.422787     2.865921e-01 4.052063e-01   1.158245e-01
## perimeter_se            -14.448729     2.000276e+00 2.866619e+00   7.834008e-01
## compactness_worst       -15.959273     1.802282e-01 2.545992e-01   8.520753e-02
## area_mean               -16.173706     4.871960e+02 6.557234e+02   1.544061e+02
## radius_mean             -16.356902     1.243396e+01 1.413850e+01   1.942663e+00
## area_worst              -16.838783     5.976253e+02 8.816606e+02   2.041599e+02
## perimeter_mean          -16.907312     7.989815e+01 9.204658e+01   1.281463e+01
## compactness_mean        -17.134098     7.764306e-02 1.044479e-01   2.858810e-02
## radius_worst            -17.462766     1.378251e+01 1.628118e+01   2.300775e+00
## concavity_worst         -17.532394     1.643566e-01 2.726677e-01   1.134018e-01
## perimeter_worst         -17.898779     8.954142e+01 1.073459e+02   1.543926e+01
## concavity_mean          -19.130891     4.377573e-02 8.895565e-02   3.009071e-02
## concave_points_worst    -19.256934     7.736947e-02 1.148080e-01   3.739687e-02
## concave_points_mean     -19.555607     2.653310e-02 4.900527e-02   1.555627e-02
##                           Overall sd      p.value
## symmetry_se             8.262246e-03 1.722169e-02
## fractal_dimension_mean  7.058406e-03 2.419834e-05
## fractal_dimension_se    2.645730e-03 3.347563e-12
## texture_mean            4.295380e+00 4.069188e-16
## texture_worst           6.143097e+00 1.423123e-17
## symmetry_worst          6.186734e-02 1.064665e-23
## smoothness_worst        2.276286e-02 2.073129e-24
## smoothness_mean         1.394371e-02 1.435876e-24
## symmetry_mean           2.739805e-02 1.193092e-24
## fractal_dimension_worst 1.805229e-02 1.615603e-25
## concavity_se            3.015633e-02 1.271418e-26
## compactness_se          1.788683e-02 2.122420e-31
## concave_points_se       6.150366e-03 8.635882e-39
## area_se                 4.548230e+01 8.078346e-42
## radius_se               2.773115e-01 3.720118e-47
## perimeter_se            2.021810e+00 2.553554e-47
```

```
## compactness_worst       1.571344e-01 2.455748e-57
## area_mean               3.513509e+02 7.730431e-59
## radius_mean             3.513889e+00 3.883885e-60
## area_worst              5.687766e+02 1.268174e-63
## perimeter_mean          2.422846e+01 3.974239e-64
## compactness_mean        5.275116e-02 8.262504e-66
## radius_worst            4.824765e+00 2.752654e-68
## concavity_worst         2.083109e-01 8.108237e-69
## perimeter_worst         3.354177e+01 1.205358e-71
## concavity_mean          7.963254e-02 1.396673e-81
## concave_points_worst    6.555590e-02 1.234776e-82
## concave_points_mean     3.874842e-02 3.695826e-85
##
## $`2`
##                            v.test Mean in category Overall mean sd in category
## concave_points_mean     19.555607     9.406852e-02 4.900527e-02   3.136278e-02
## concave_points_worst    19.256934     1.898831e-01 1.148080e-01   4.079263e-02
## concavity_mean          19.130891     1.795546e-01 8.895565e-02   7.028848e-02
## perimeter_worst         17.898779     1.430490e+02 1.073459e+02   3.150730e+01
## concavity_worst         17.532394     4.898630e-01 2.726677e-01   1.841832e-01
## radius_worst            17.462766     2.129175e+01 1.628118e+01   4.660217e+00
## compactness_mean        17.134098     1.581994e-01 1.044479e-01   4.892723e-02
## perimeter_mean          16.907312     1.164077e+02 9.204658e+01   2.335483e+01
## area_worst              16.838783     1.451234e+03 8.816606e+02   6.343947e+02
## radius_mean             16.356902     1.755661e+01 1.413850e+01   3.468416e+00
## area_mean               16.173706     9.936698e+02 6.557234e+02   3.899051e+02
## compactness_worst       15.959273     4.037349e-01 2.545992e-01   1.622110e-01
## perimeter_se            14.448729     4.603889e+00 2.866619e+00   2.555567e+00
## radius_se               14.422787     6.430624e-01 4.052063e-01   3.455744e-01
## area_se                 13.548575     7.702094e+01 4.037438e+01   6.322988e+01
## concave_points_se       13.026611     1.658154e-02 1.181690e-02   6.207303e-03
## compactness_se          11.656645     3.791429e-02 2.551479e-02   2.007686e-02
## concavity_se            10.679362     5.110217e-02 3.194987e-02   3.975725e-02
## fractal_dimension_worst 10.440732     9.517852e-02 8.396968e-02   2.176554e-02
## symmetry_mean           10.249214     1.979011e-01 1.812014e-01   2.747436e-02
## smoothness_mean         10.231294     1.049214e-01 9.643727e-02   1.286821e-02
## smoothness_worst        10.195669     1.462452e-01 1.324433e-01   2.202453e-02
## symmetry_worst          10.035456     3.270037e-01 2.900808e-01   7.453879e-02
## texture_worst            8.533234     2.878640e+01 2.566896e+01   5.831600e+00
## texture_mean             8.136485     2.135884e+01 1.928040e+01   4.027551e+00
## fractal_dimension_se     6.962302     4.892143e-03 3.796685e-03   3.210016e-03
## fractal_dimension_mean   4.222150     6.457688e-02 6.280458e-02   8.623212e-03
## symmetry_se              2.381940     2.170173e-02 2.053135e-02   1.031005e-02
##                          Overall sd      p.value
## concave_points_mean     3.874842e-02 3.695826e-85
```

```
## concave_points_worst    6.555590e-02 1.234776e-82
## concavity_mean          7.963254e-02 1.396673e-81
## perimeter_worst         3.354177e+01 1.205358e-71
## concavity_worst         2.083109e-01 8.108237e-69
## radius_worst            4.824765e+00 2.752654e-68
## compactness_mean        5.275116e-02 8.262504e-66
## perimeter_mean          2.422846e+01 3.974239e-64
## area_worst              5.687766e+02 1.268174e-63
## radius_mean             3.513889e+00 3.883885e-60
## area_mean               3.513509e+02 7.730431e-59
## compactness_worst       1.571344e-01 2.455748e-57
## perimeter_se            2.021810e+00 2.553554e-47
## radius_se               2.773115e-01 3.720118e-47
## area_se                 4.548230e+01 8.078346e-42
## concave_points_se       6.150366e-03 8.635882e-39
## compactness_se          1.788683e-02 2.122420e-31
## concavity_se            3.015633e-02 1.271418e-26
## fractal_dimension_worst 1.805229e-02 1.615603e-25
## symmetry_mean           2.739805e-02 1.193092e-24
## smoothness_mean         1.394371e-02 1.435876e-24
## smoothness_worst        2.276286e-02 2.073129e-24
## symmetry_worst          6.186734e-02 1.064665e-23
## texture_worst           6.143097e+00 1.423123e-17
## texture_mean            4.295380e+00 4.069188e-16
## fractal_dimension_se    2.645730e-03 3.347563e-12
## fractal_dimension_mean  7.058406e-03 2.419834e-05
## symmetry_se             8.262246e-03 1.722169e-02
```

```
res.hcpc$desc.axes$quanti
```

```
## $`1`
##         v.test Mean in category Overall mean sd in category Overall sd
## Dim.1 -20.2478        -2.187516 1.787498e-15       1.440264   3.642952
##           p.value
## Dim.1 3.71438e-91
##
## $`2`
##        v.test Mean in category Overall mean sd in category Overall sd
## Dim.1 20.2478         4.386606 1.787498e-15       2.624071   3.642952
##           p.value
## Dim.1 3.71438e-91
```

2.5 Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a statistical technique that allows us to test the null hypothesis that the means of any three or more groups are the same against the alternative hypothesis that they are not equal using information from their samples. The fundamental idea behind the ANOVA technique is to compare the degree of variation within each sample to the degree of variance between samples in order to determine whether the population means differ from one another. It is assumed that the samples used in the ANOVA model come from normal populations with similar variances. An ANOVA with a single factor between subjects is used when there is only one factor and the analysis has more than two levels and multiple subjects in both experimental conditions.

3 Result

3.1 PCA and HC classification

3.1.1 PCA interpretation

3.1.1.1 Eigenvalues Eigenvalues, which are measurements of the variance explained by each main component, are important indicators in principal component analysis. Three principal components have been recognized in this investigation, and the variation they account for is shown by their associated eigenvalues. Notably, the first major component accounts for 38.85% of the variation in total, with the second and third accounting for 20.49% and 9.55% of the variance, respectively. Increased eigenvalues indicate a higher degree of information retention from the original dataset, highlighting the fact that PC1 contains the largest amount of variance derived from the wine data.

3.1.1.2 Contributions Analysis of contributions reveals insights into links between continuous and categorical variable (Type) through the data table's display of the correlations between original variables and principal components. Their relevance is clarified by examining the contributions of the first three main components below.

The first principle component (PC1), representing 38.85% of the variance, demonstrates strong correlations with variables like Flavanoids, Phenols, Dilution, Proline, and Hue. Notably, Flavanoids, Phenols, Dilution, and Proline exhibit high positive correlations, signifying their significant contributions to the observed separation or variability in the data along this component. Moreover, the categorical variable 'Type' shows a substantial relationship with PC1, emphasizing its role in differentiating between wine types.

For the second principle component, it accounted for 20.49% of the variance, displays high correlations with variables such as Color, Alcohol, Proline, Ash, and Magnesium. Particularly, Color and Alcohol stand out with strong positive correlations, indicating their substantial contributions to the variation captured by PC2. Similarly, 'Type' showcases a

robust relationship with PC2, underlining its importance in distinguishing between wine types along this component.

Finally, the third principle component represented 9.55% of the variance, exhibits notable correlations with variables like Ash, Alcalinity, Nonflavanoids, Dilution, and Flavanoids. Notably, Ash and Alcalinity display high positive correlations, indicating their significant contributions to the variability represented by PC3. Although the categorical variable 'Type' also demonstrates differentiation along PC3, its impact is not as pronounced as observed in PC1 and PC2.

3.1.2 HC interpretation

As a result of the cluster analysis, specific quantitative factors that identified the distinctive qualities and contributions of every cluster were found to exist. There are 2 cluster:

- Cluster 1 contains about 90% of B tumors and 10% of M tumors.
- Cluster 2 contains about 10% of B tumors and 90% of M tumors.

First of all, Cluster 1, which is composed of more than 90% of benign tumors, has characteristics by having almost all values quite lower when compare to overall. Cluster 1 type are distinguished by characteristics that stand out: symmetry, fractal dimension, texture, smoothness, concavity, compactness, concave points, area and so on.

Opposite to cluster 2, which consists primarily of malignant tumors, has unique characteristics that are indicated by increased almost all values relative to the dataset's average values.

3.1.3 Conclusion from PCA and HC

The PCA analysis identified key variables (chemical components) that significantly contribute to the differentiation of wine types. PC1, PC2, and PC3 collectively explain a substantial portion of the variance in the dataset. Variables like Flavanoids, Phenols, Dilution, Proline, Color, Alcohol, Ash, Alcalinity, etc., exhibit strong correlations with these principal components and are crucial in distinguishing between different wine types.

On the other hand, the hierarchical clustering analysis, along with associated quantitative variables and principal dimensions, provides a clear demarcation of wines into three distinct clusters (representing Types 1, 2, and 3). Each cluster exhibits unique characteristics based on the 13 attributes. These findings offer valuable insights into how different wine types/classes can be differentiated based on these chemical attributes, aiding in wine classification and potentially understanding the underlying characteristics that differentiate these types.

3.2 ANOVA test

The aim is to test the significance of various chemical components across different wine types. These tests in R will provide insights into which variables significantly differ between wine types, aiding in the classification of wines based on their chemical compositions.

3.2.1 Hypothesis 1: Mean of Concave Points between benign and malignant tumors.

- Null Hypothesis (H0): There is no significant difference in Mean of Concave Points between 2 tumor types.
- Alternative Hypothesis (H1): There is a significant difference in Mean of Concave Points between 2 tumor types.

3.2.1.1 Test: ANOVA for Mean of Concave Points among tumor types.

```
# ANOVA for Mean of Concave Points among tumor types.
h1 <- lm(
  concave_points_mean ~ as.factor(diagnosis),
  data = data
)
anova(h1)
```

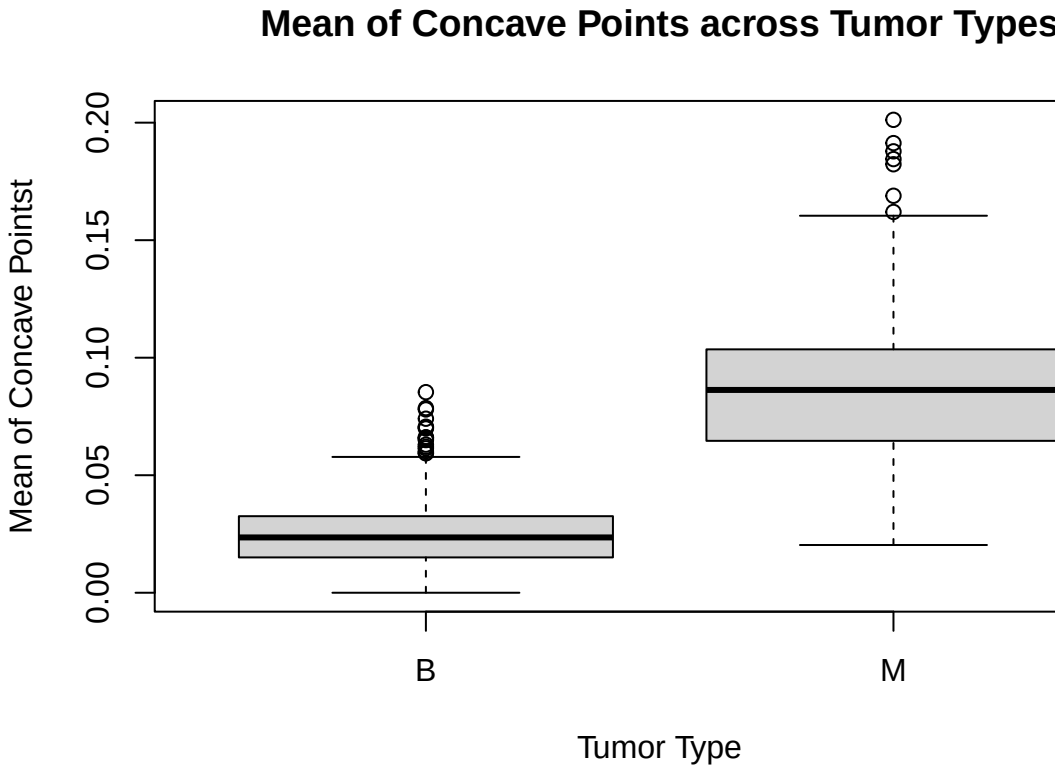
```
## Analysis of Variance Table
##
## Response: concave_points_mean
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)     1 0.51407 0.51407  858.94 < 2.2e-16 ***
## Residuals              566 0.33875 0.00060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.2.1.2 Interpretation:

The ANOVA test indicates a highly significant difference in Mean of Concave Points among different tumor types (p = 2.2e-16 < 0.001). This suggests that Mean of Concave Points vary strongly between benign and malignant tumor. Then **reject Hypothesis**.

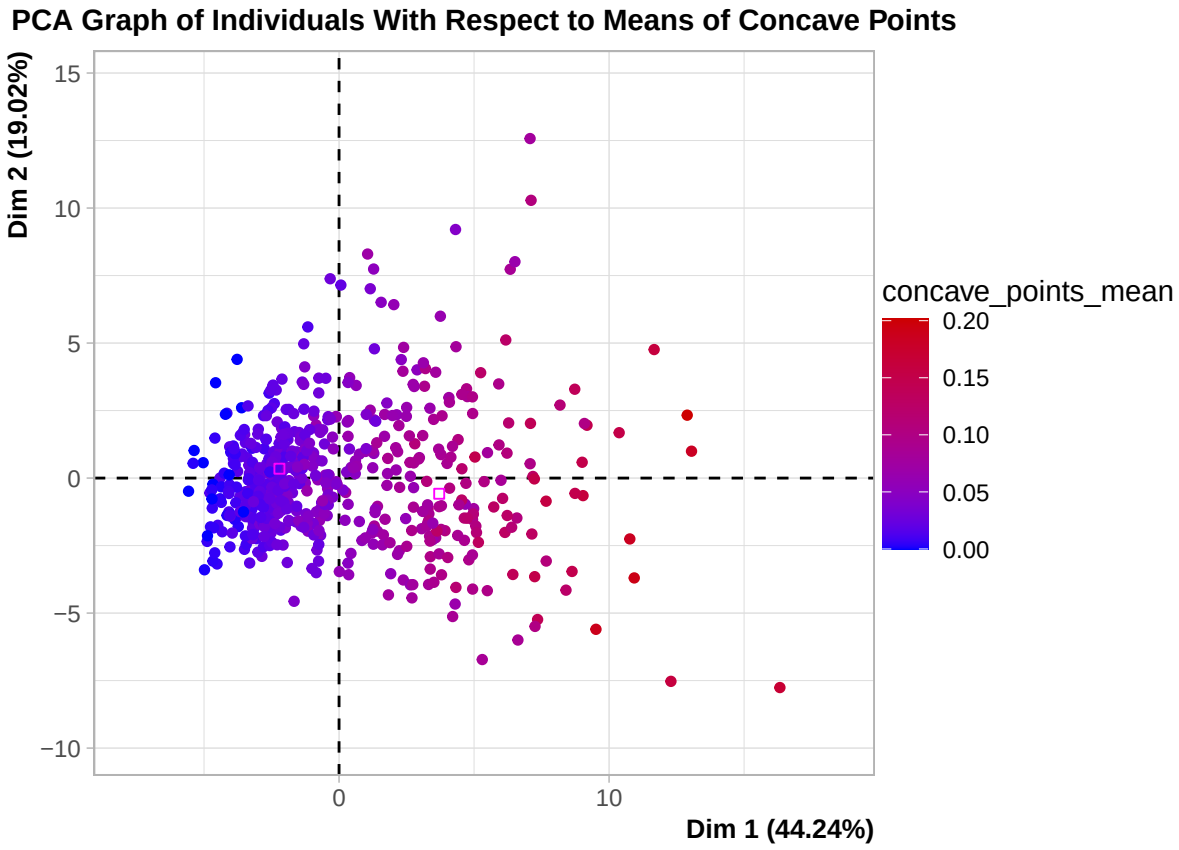
```
# Boxplot for Mean of Concave Points
boxplot(
  concave_points_mean ~ as.factor(diagnosis),
```

```
  data = data,
  xlab = "Tumor Type",
  ylab = "Mean of Concave Pointst",
  main = "Mean of Concave Points across Tumor Types"
)
```



3.2.1.3 Visualization:

```
plot.PCA(res.pca,
   choix = "ind",
   hab = "concave_points_mean",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Means of Concave Points",
   label = c("none")
)
```



We can see that there is significant difference of mean of concave points between malignant and benign breast tumors. The mean of concave points of malignant breast tumors is much higher than that of benign ones as in the Box Plot. Recall the variable PCA plot, the mean of concave points is strongly positive correlated with Dim 1, which is also true in the PCA graph of individuals as we can see almost all malignant breast tumors are aligned to the right of Dim 1. Hence, they have higher mean of concave points. All mentions above prove that a high number of concave points usually the sign of breast cancer.

3.2.2 Hypothesis 2: Mean of Concavity between benign and malignant tumors.

- Null Hypothesis (H0): There is no significant difference in Mean of Concavity between 2 tumor types.
- Alternative Hypothesis (H1): There is a significant difference in Mean of Concavity between 2 tumor types.

3.2.2.1 Test: ANOVA for Mean of Concavity among tumor types.

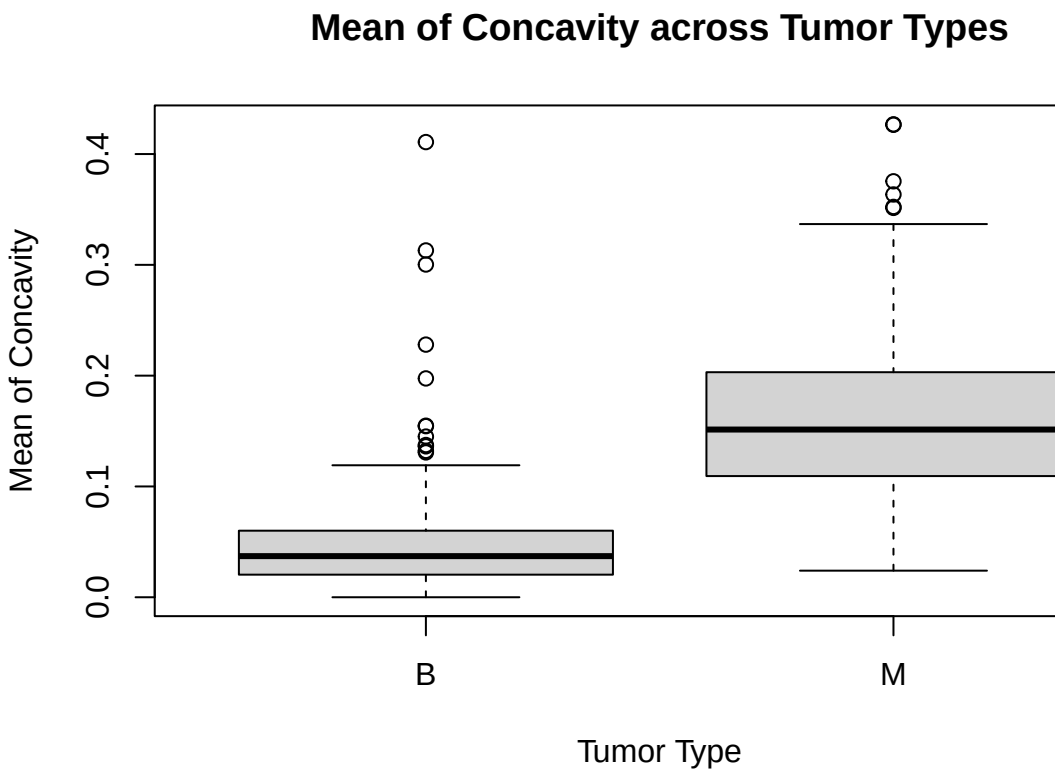
```
# ANOVA for Mean of Concavity among tumor types.
h2 <- lm(
  concavity_mean ~ as.factor(diagnosis),
  data = data
```

```
)
anova(h2)
```

```
## Analysis of Variance Table
##
## Response: concavity_mean
##                        Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)    1 1.7447 1.74467   531.7 < 2.2e-16 ***
## Residuals             566 1.8572 0.00328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

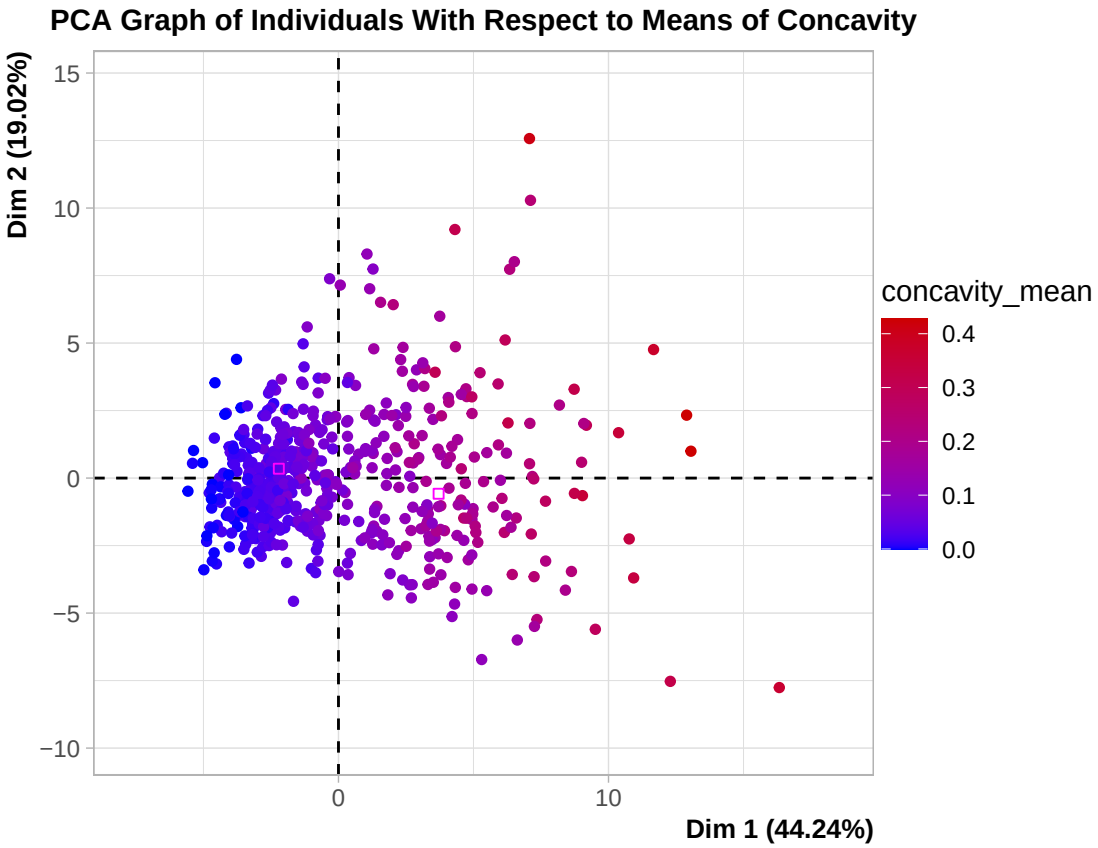
3.2.2.2 Interpretation: The ANOVA test indicates a highly significant difference in Mean of Concavity among different tumor types (p = 2.2e-16 < 0.001). This suggests that Mean of Concavity vary strongly between benign and malignant tumor. Then **reject Hypothesis**.

```
# Boxplot for Mean of Concavity
boxplot(
  concavity_mean ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Mean of Concavity",
  main = "Mean of Concavity across Tumor Types"
)
```



3.2.2.3 Visualization:

```
plot.PCA(res.pca,
   choix = "ind",
   hab = "concavity_mean",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Means of Concavity",
   label = c("none")
)
```



Just like Mean of Concave Points, the mean of concavity is quite different between malignant and benign breast tumors. The mean of concavity of malignant breast tumors is much higher than that of benign ones as in the Box Plot. Recall the variable PCA plot, the mean of concavity is also strongly positive correlated with Dim 1 similar to the mean of concave points. We can see in the PCA graph of individuals that almost all malignant breast tumors are aligned to the right of Dim 1. Therefore, the bigger size of tumor, the higher chance of a person to have breast cancer.

3.2.3 Hypothesis 3: Worst Concave Points between benign and malignant tumors.

- Null Hypothesis (H0): There is no significant difference in Worst Concave Points between 2 tumor types.
- Alternative Hypothesis (H1): There is a significant difference in Worst Concave Points between 2 tumor types.

3.2.3.1 Test: ANOVA for Worst Concave Points among tumor types.

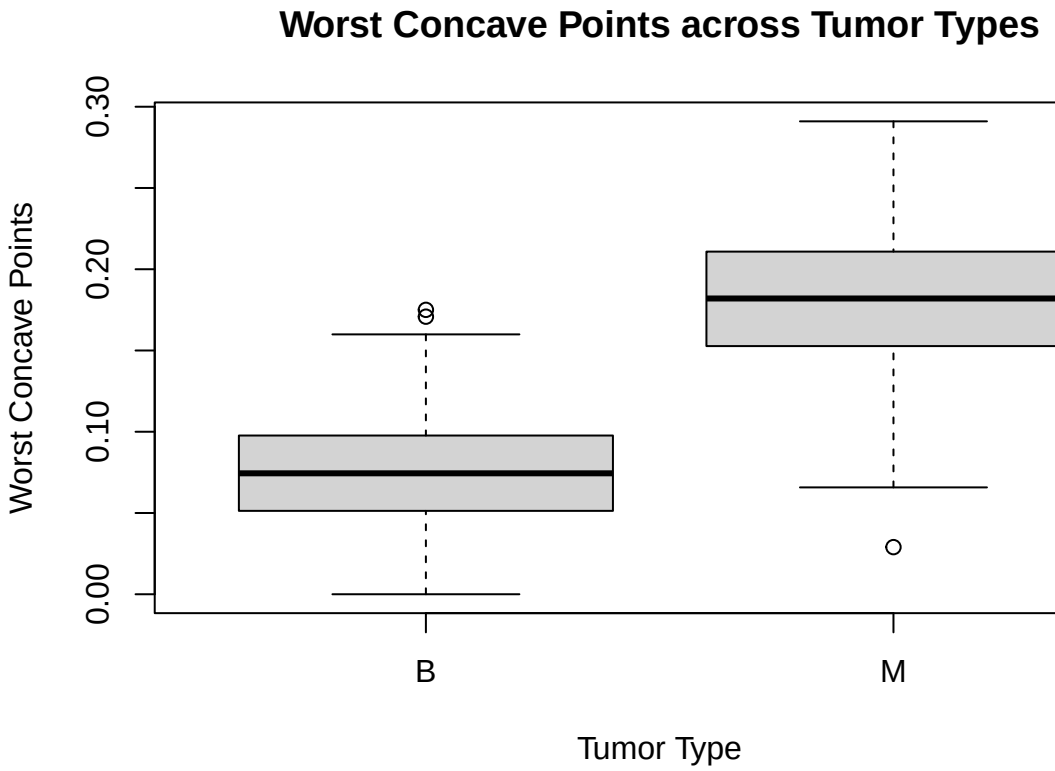
```
# ANOVA for Worst Concave Points among tumor types.
h3 <- lm(
  concave_points_worst ~ as.factor(diagnosis),
```

```
  data = data
)
anova(h3)
```

```
## Analysis of Variance Table
##
## Response: concave_points_worst
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)   1 1.53791  1.5379  963.85 < 2.2e-16 ***
## Residuals            566 0.90311  0.0016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

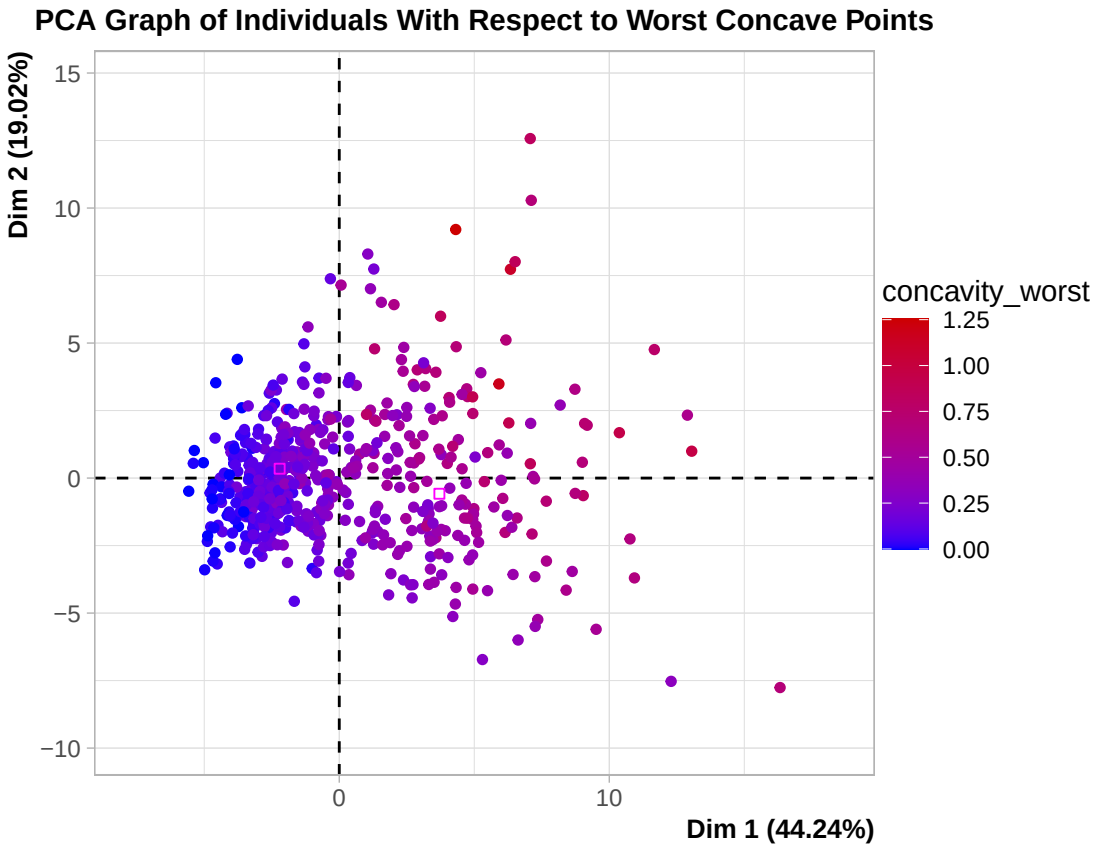
3.2.3.2 Interpretation: The ANOVA test indicates a highly significant difference in Worst Concave Points among different tumor types (p = 2.2e-16 < 0.001). This suggests that Mean of Concavity vary strongly between benign and malignant tumor. Then **reject Hypothesis**.

```
# Boxplot for Worst Concave Points
boxplot(
  concave_points_worst ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Worst Concave Points",
  main = "Worst Concave Points across Tumor Types"
)
```



3.2.3.3 Visualization:

```
plot.PCA(res.pca,
   choix = "ind",
   hab = "concavity_worst",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Worst Concave Points",
   label = c("none")
)
```



Similar to 2 stats of concavity mentioned above, the Worst Concave Points is different between 2 type of breast tumors. The Worst Concave Points of malignant breast tumors is much higher than that of benign ones as in the Box Plot. Also, the Worst Concave Points is positively correlated with Dim 1 and just like the concavity stats above, almost all malignant breast tumors are aligned to the right of Dim 1. Therefore, the worse concavity of tumor, the higher chance of a person to have breast cancer.

3.2.4 Hypothesis 4: Mean of Compactness between benign and malignant tumors.

- Null Hypothesis (H0): There is no significant difference in Mean of Compactness between 2 tumor types.
- Alternative Hypothesis (H1): There is a significant difference in Mean of Compactness between 2 tumor types.

3.2.4.1 Test: ANOVA for Mean of Compactness among tumor types.

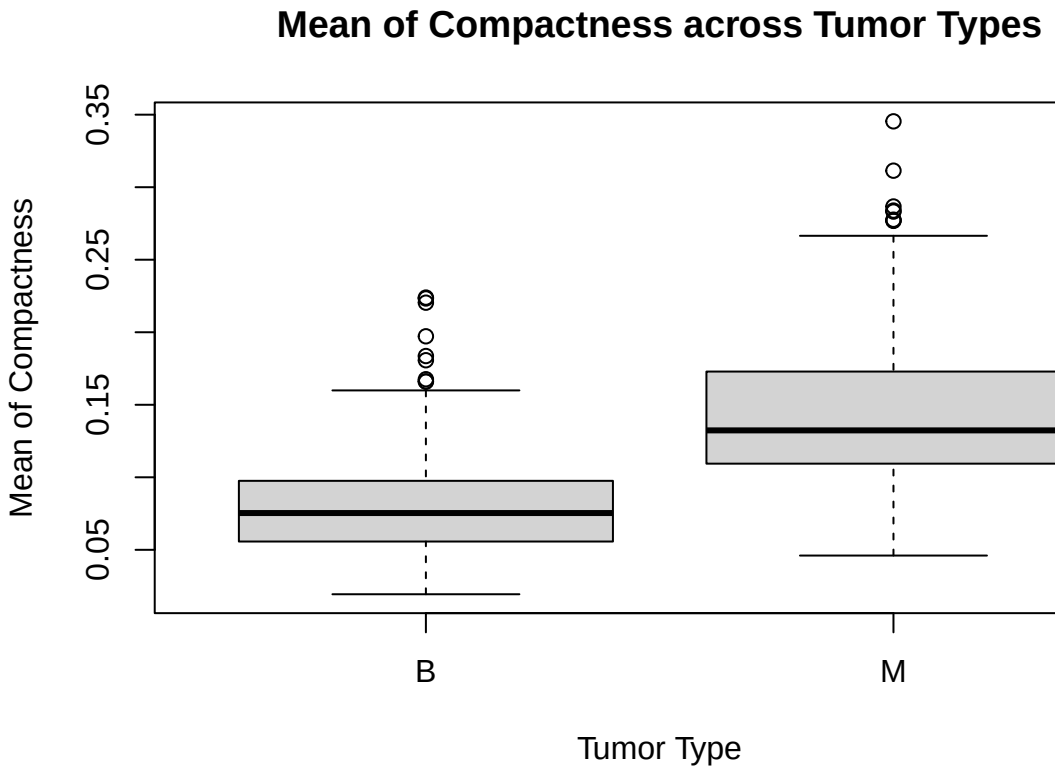
```
# ANOVA for Mean of Compactness among tumor types.
h4 <- lm(
  compactness_mean ~ as.factor(diagnosis),
  data = data
```

```
)
anova(h4)
```

```
## Analysis of Variance Table
##
## Response: compactness_mean
##                        Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)    1 0.5614  0.5614  311.78 < 2.2e-16 ***
## Residuals             566 1.0192  0.0018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

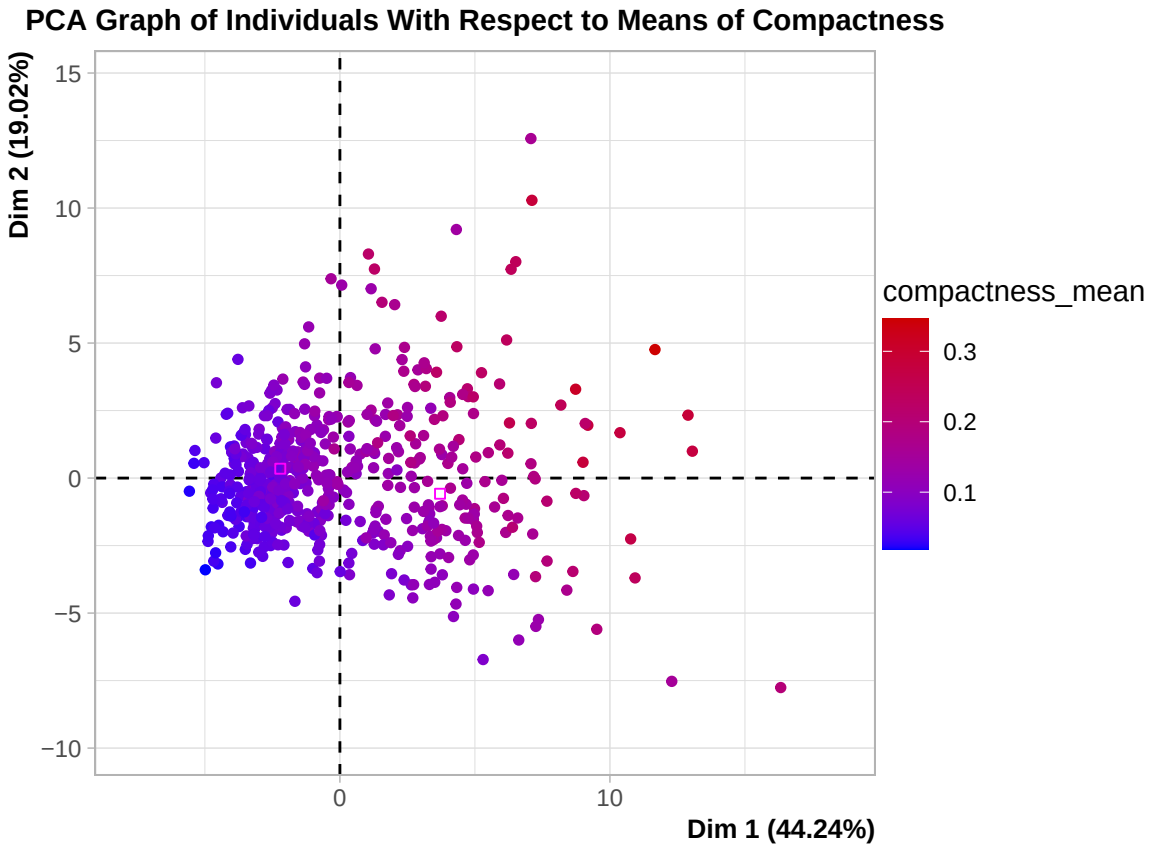
3.2.4.2 Interpretation: The ANOVA test indicates a highly significant difference in Mean of Compactness among different tumor types (p = 2.2e-16 < 0.001). This suggests that Mean of Compactness vary strongly between benign and malignant tumor. Then **reject Hypothesis**.

```
# Boxplot for Mean of Compactness
boxplot(
  compactness_mean ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Mean of Compactness",
  main = "Mean of Compactness across Tumor Types"
)
```



3.2.4.3 Visualization:

```
plot.PCA(res.pca,
   choix = "ind",
   hab = "compactness_mean",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Means of Compactness",
   label = c("none")
)
```



Here, Mean of Compactness is different between 2 type of breast tumors. The Mean of Compactness of malignant breast tumors is also higher than that of benign ones as in the Box Plot. And, the Mean of Compactness is positively correlated with Dim 1 and just like all the concavity stats above, almost all malignant breast tumors are aligned to the right of Dim 1.

3.2.5 Other hypotheses on an variable between benign and malignant tumors.

Similar to the analysis above and the Dimension Description of the PCA on the dataset, we can find some other variables that are: - Significantly different between benign and malignant tumors. - Positively correlated with Dim 1 of PCA variable graph. Means that higher value correspond to higher chance of breast cancer.

These variables are: - perimeter_worst - concavity_worst - radius_worst - perimeter_mean - area_worst - area_mean

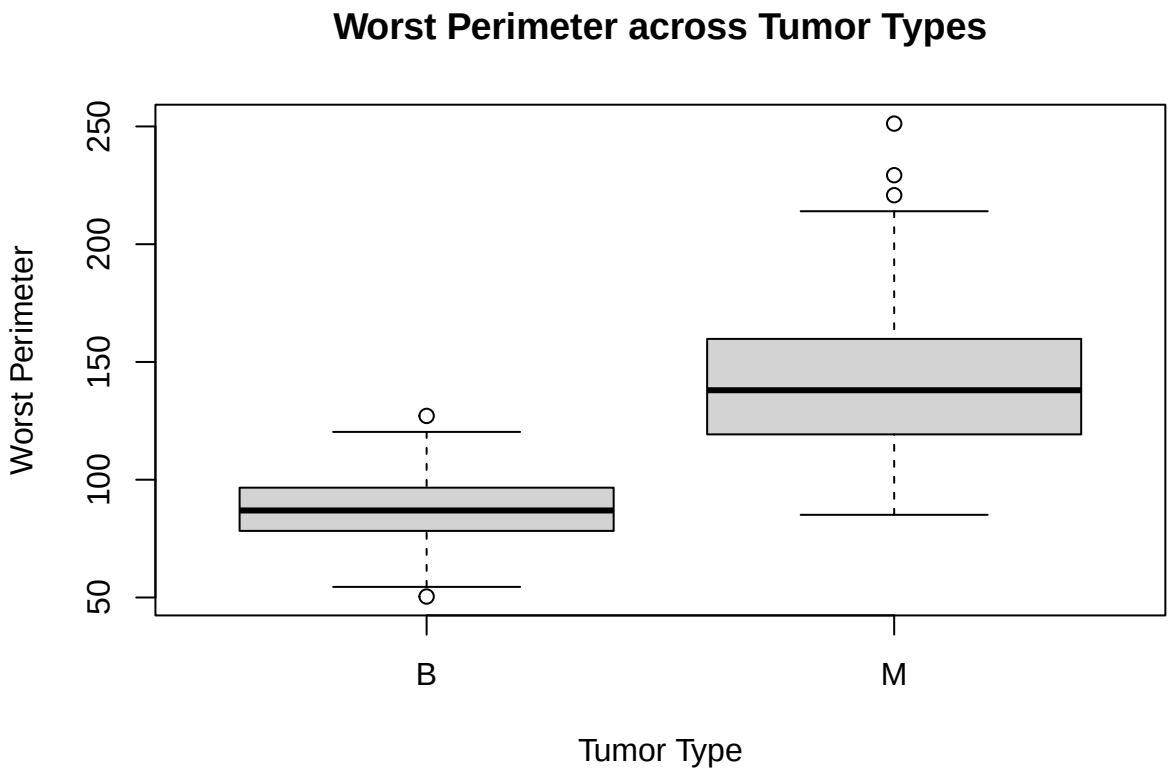
We will perform ANOVA analysis and visualize the relations of these variable below with corresponding order: #### Worst Perimeter

```
h4.1 <- lm(
  perimeter_worst ~ as.factor(diagnosis),
  data = data
```

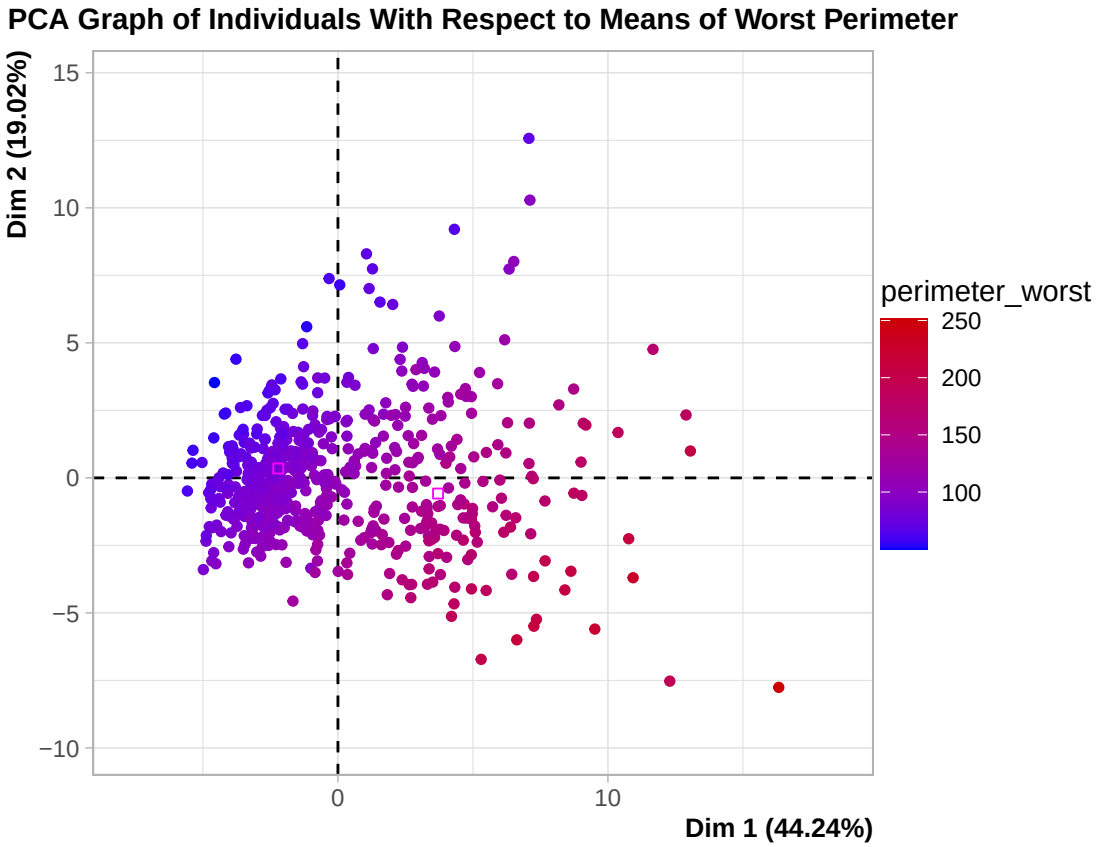
```
)
anova(h4.1)
```

```
## Analysis of Variance Table
##
## Response: perimeter_worst
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)     1 391576  391576  895.65 < 2.2e-16 ***
## Residuals              566 247453     437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Boxplot for Worst Perimeter
boxplot(
  perimeter_worst ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Worst Perimeter",
  main = "Worst Perimeter across Tumor Types"
)
```



```r
plot.PCA(res.pca,
    choix = "ind",
    hab = "perimeter_worst",
    invisible = c("var"),
    cex = 1,
    autoLab = c("no"),
    title = "PCA Graph of Individuals With Respect to Means of Worst Perimeter",
    label = c("none")
)
```

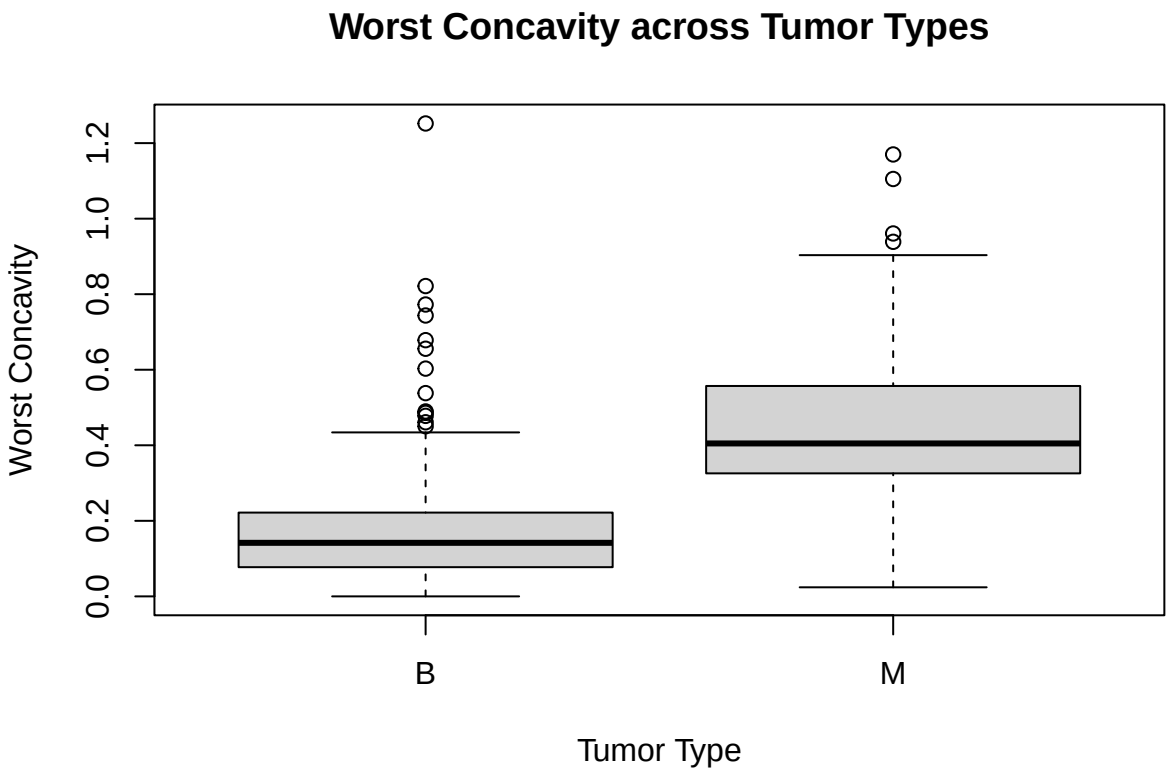



```r
h4.2 <- lm(
  concavity_worst ~ as.factor(diagnosis),
  data = data
)
anova(h4.2)
```

3.2.5.1 Worst Concavity

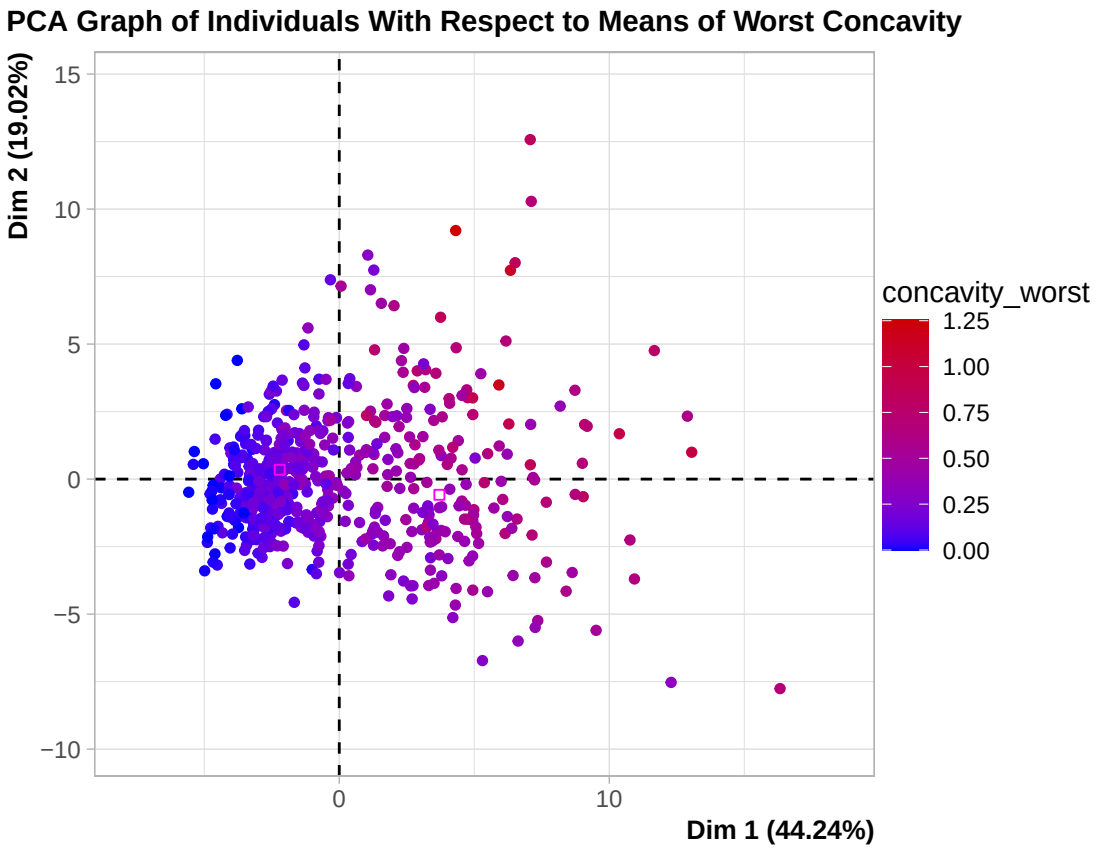
```
## Analysis of Variance Table
##
## Response: concavity_worst
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)     1 10.710 10.7095   434.9 < 2.2e-16 ***
## Residuals              566 13.938  0.0246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Boxplot for Worst Perimeter
boxplot(
  concavity_worst ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Worst Concavity",
  main = "Worst Concavity across Tumor Types"
)
```




```r
plot.PCA(res.pca,
   choix = "ind",
   hab = "concavity_worst",
```

```
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Means of Worst Concavity",
   label = c("none")
)
```



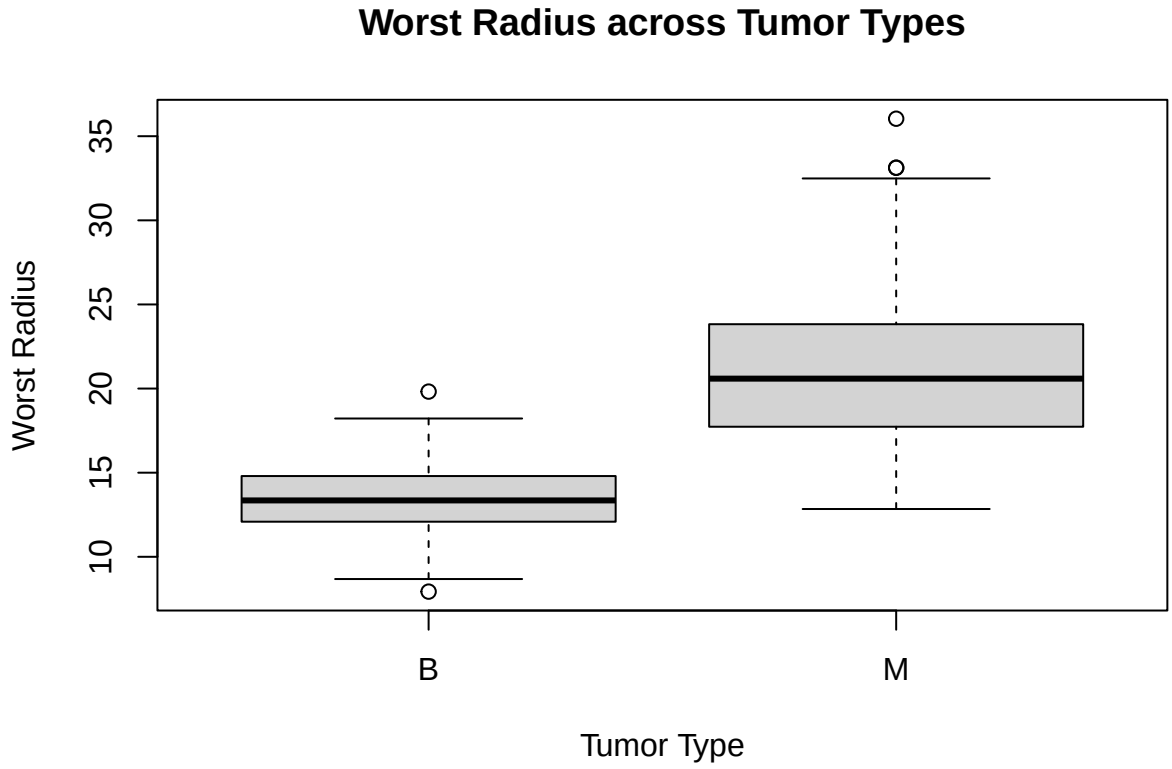

```
h4.3 <- lm(
  radius_worst ~ as.factor(diagnosis),
  data = data
)
anova(h4.3)
```

3.2.5.2 Worst Radius

```
## Analysis of Variance Table
##
## Response: radius_worst
```

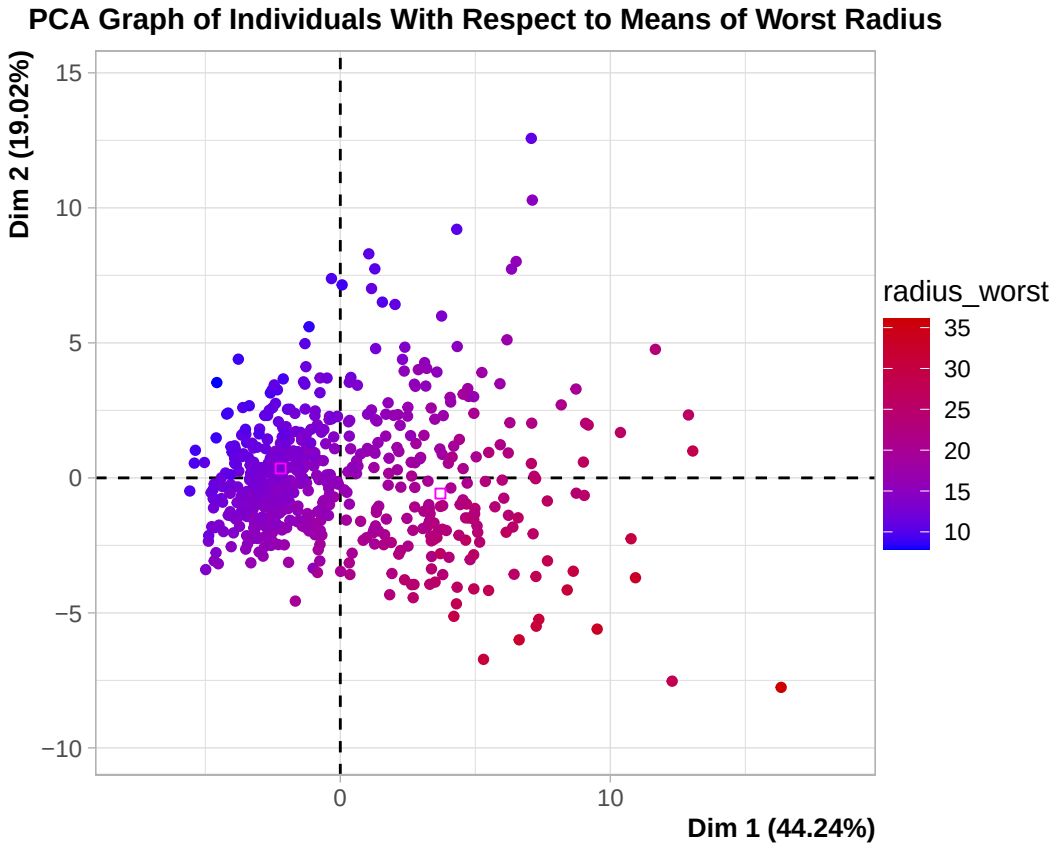
```
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)     1 7968.3  7968.3  858.44 < 2.2e-16 ***
## Residuals              566 5253.8     9.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Boxplot for Worst Radius
boxplot(
  radius_worst ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Worst Radius",
  main = "Worst Radius across Tumor Types"
)
```




```r
plot.PCA(res.pca,
   choix = "ind",
   hab = "radius_worst",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
```

```
   title = "PCA Graph of Individuals With Respect to Means of Worst Radius",
   label = c("none")
)
```



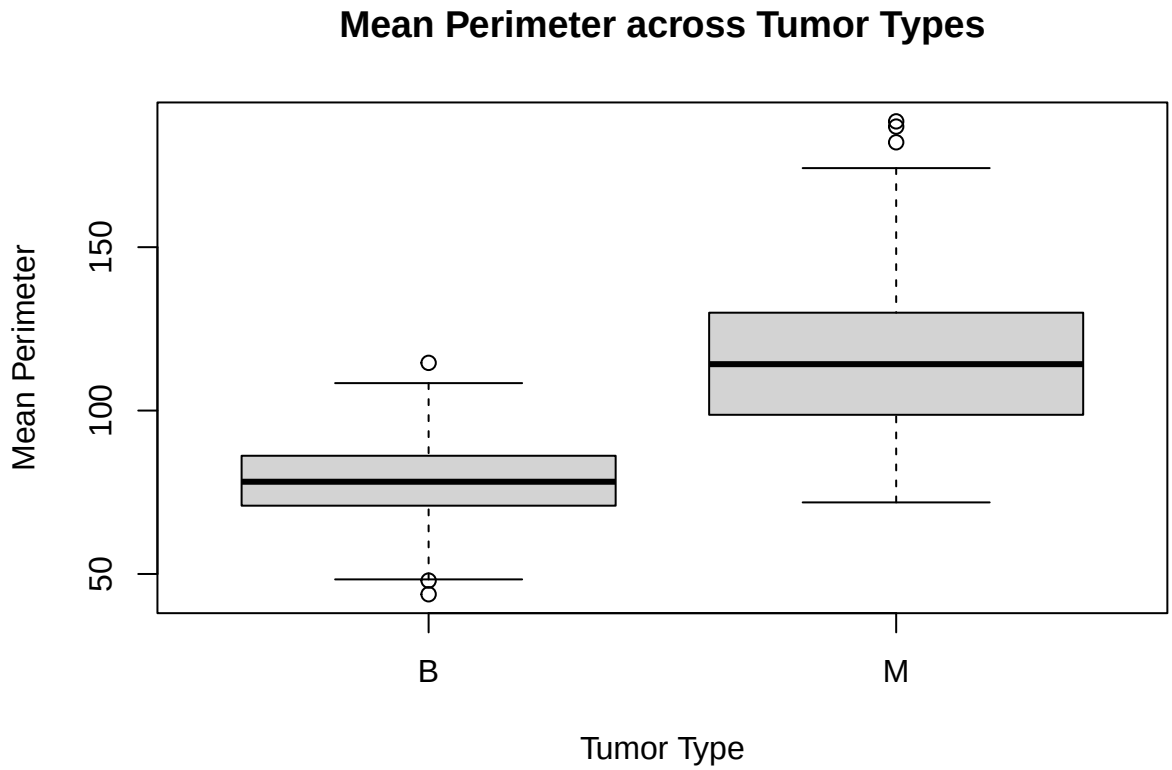

```
h4.4 <- lm(
  perimeter_mean ~ as.factor(diagnosis),
  data = data
)
anova(h4.4)
```

3.2.5.3 Mean Perimeter

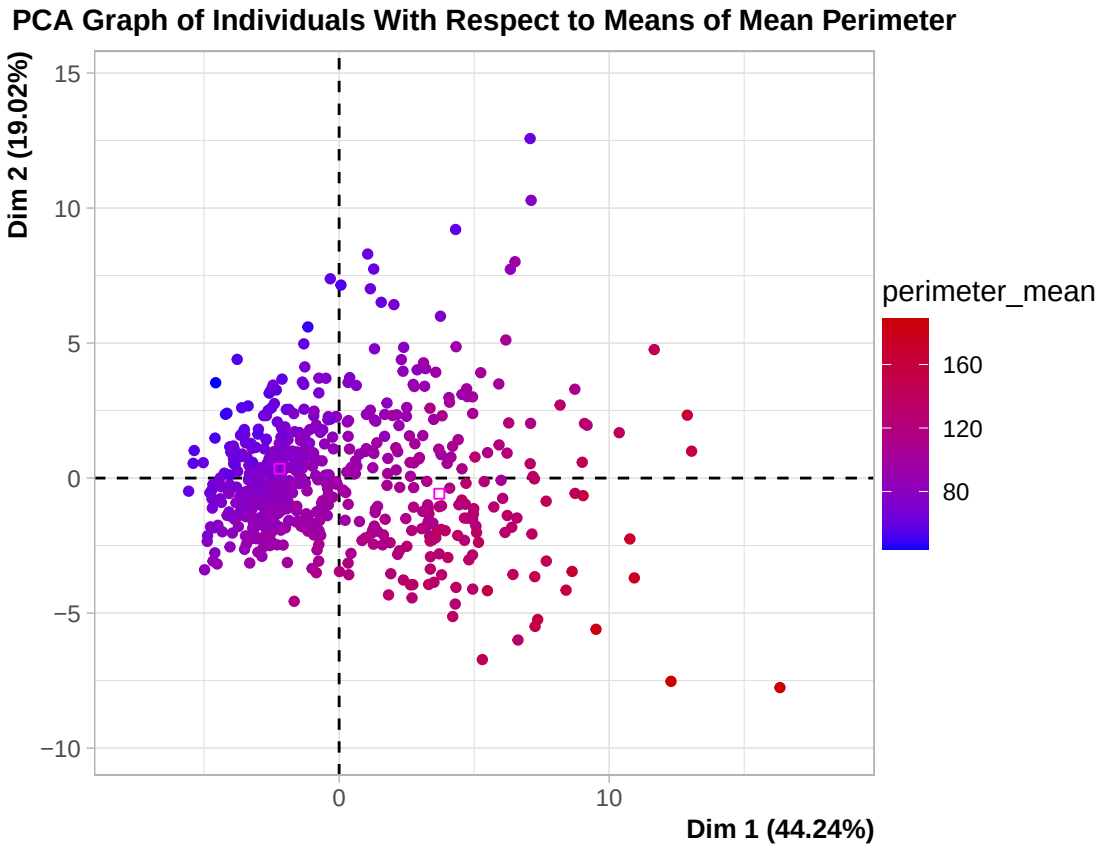
```
## Analysis of Variance Table
##
## Response: perimeter_mean
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(diagnosis)   1 183927  183927  696.34 < 2.2e-16 ***
## Residuals            566 149499     264
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Boxplot for Mean Perimeter
boxplot(
  perimeter_mean ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Mean Perimeter",
  main = "Mean Perimeter across Tumor Types"
)
```




```r
plot.PCA(res.pca,
   choix = "ind",
   hab = "perimeter_mean",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Means of Mean Perimeter",
   label = c("none")
)
```

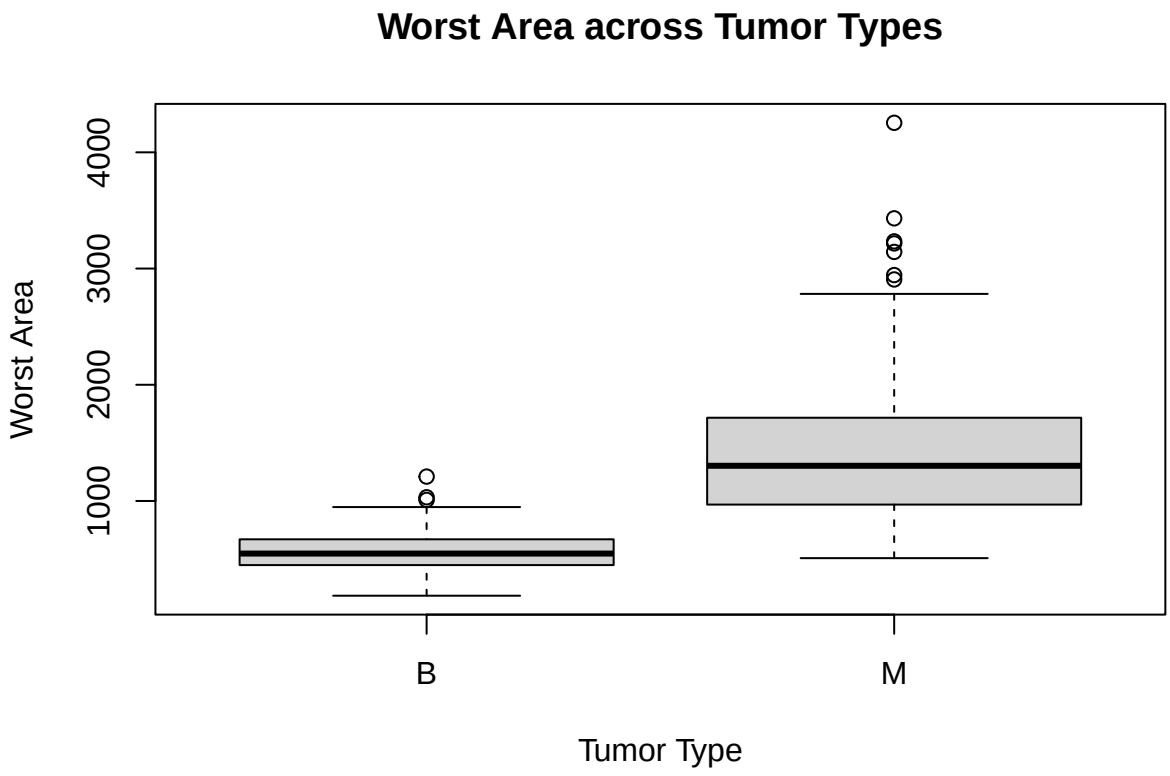



```
h4.5 <- lm(
  area_worst ~ as.factor(diagnosis),
  data = data
)
anova(h4.5)
```

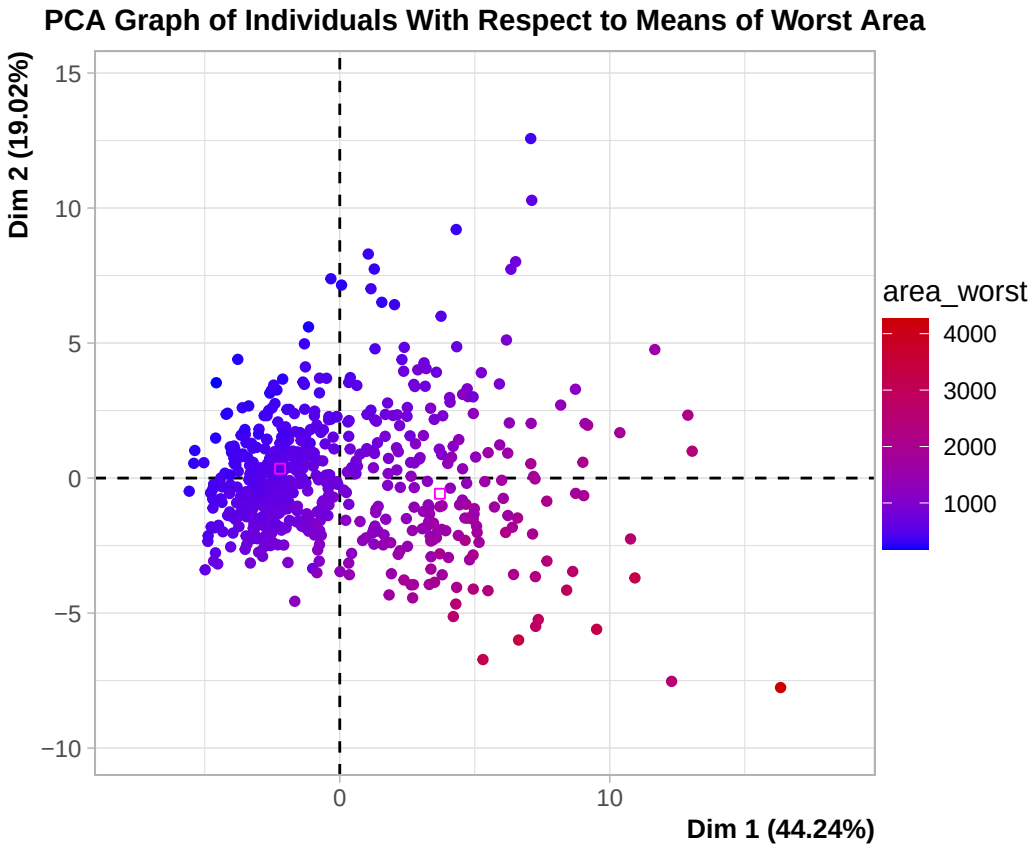
3.2.5.4 Worst Area

```
## Analysis of Variance Table
##
## Response: area_worst
##                        Df   Sum Sq  Mean Sq F value    Pr(>F)
## as.factor(diagnosis)    1 98861607 98861607  659.15 < 2.2e-16 ***
## Residuals             566 84890285   149983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Boxplot for Worst Area
boxplot(
  area_worst ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Worst Area",
  main = "Worst Area across Tumor Types"
)
```




```r
plot.PCA(res.pca,
   choix = "ind",
   hab = "area_worst",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Means of Worst Area",
   label = c("none")
)
```

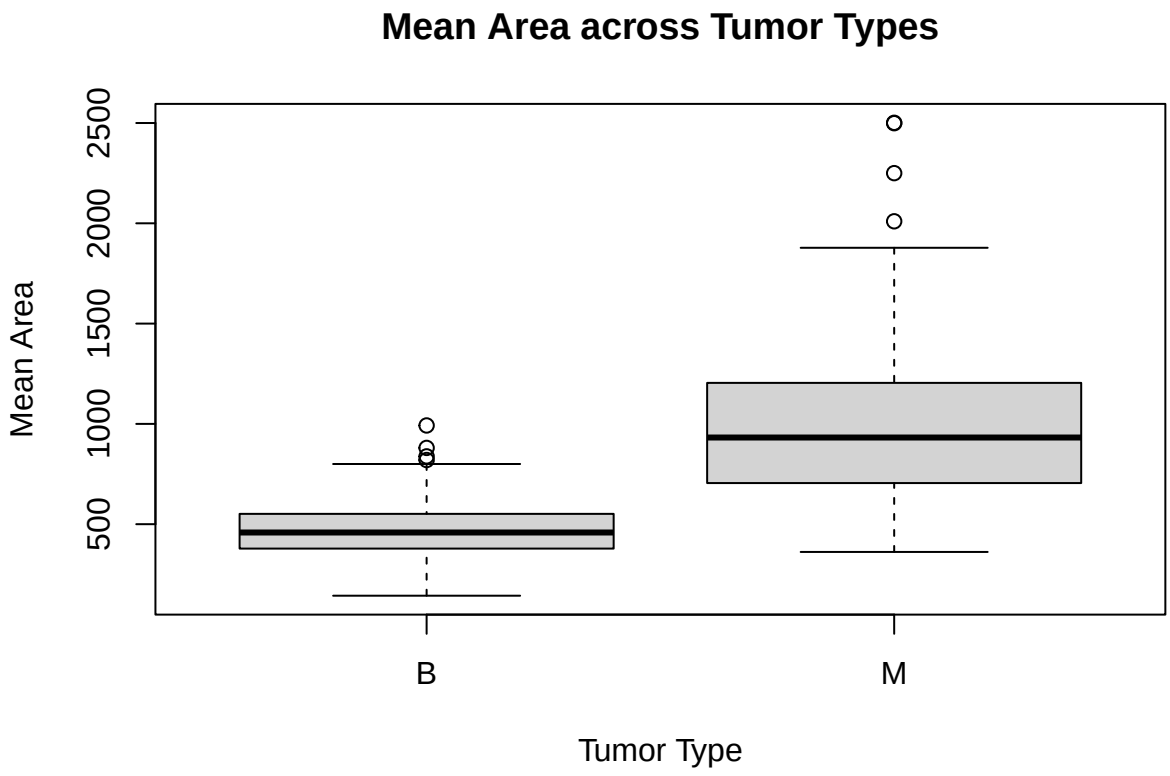



```
h4.6 <- lm(
  area_mean ~ as.factor(diagnosis),
  data = data
)
anova(h4.6)
```

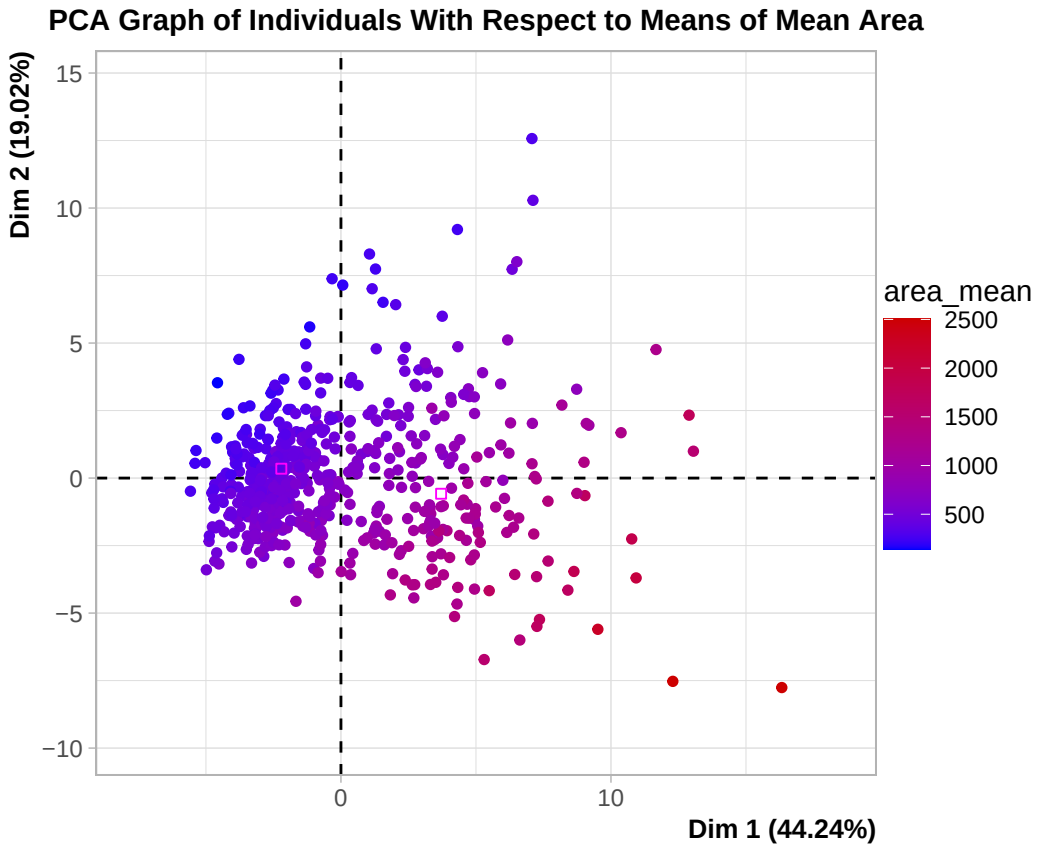
3.2.5.5 Mean Area

```
## Analysis of Variance Table
##
## Response: area_mean
##                        Df   Sum Sq  Mean Sq F value    Pr(>F)
## as.factor(diagnosis)    1 35213210 35213210     571 < 2.2e-16 ***
## Residuals             566 34904963    61670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Boxplot for Mean Area
boxplot(
  area_mean ~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Mean Area",
  main = "Mean Area across Tumor Types"
)
```




```r
plot.PCA(res.pca,
   choix = "ind",
   hab = "area_mean",
   invisible = c("var"),
   cex = 1,
   autoLab = c("no"),
   title = "PCA Graph of Individuals With Respect to Means of Mean Area",
   label = c("none")
)
```



3.2.6 Hypothesis 5: Mean Fractal dimension between benign and malignant tumors.

- Null Hypothesis (H0): There is no significant difference in Mean Fractal dimension between 2 tumor types.
- Alternative Hypothesis (H1): There is a significant difference in Mean Fractal dimension between 2 tumor types.

3.2.6.1 Test: ANOVA for Mean Fractal dimension among tumor types.

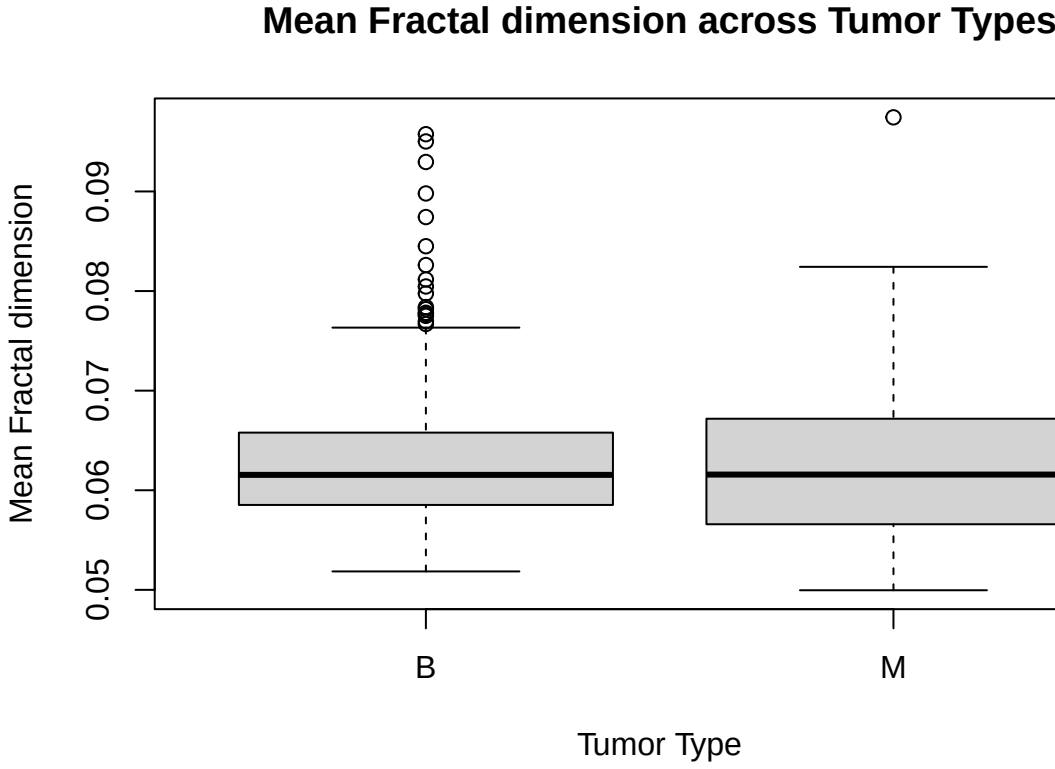
```
# ANOVA for Mean Fractal dimension among tumor types.
h5 <- lm(
  fractal_dimension_mean ~ as.factor(diagnosis),
  data = data
)
anova(h5)
```

```
## Analysis of Variance Table
##
## Response: fractal_dimension_mean
```

```
##                          Df    Sum Sq    Mean Sq F value Pr(>F)
## as.factor(diagnosis)      1 0.0000052 5.2410e-06  0.1049 0.7462
## Residuals               566 0.0282931 4.9988e-05
```

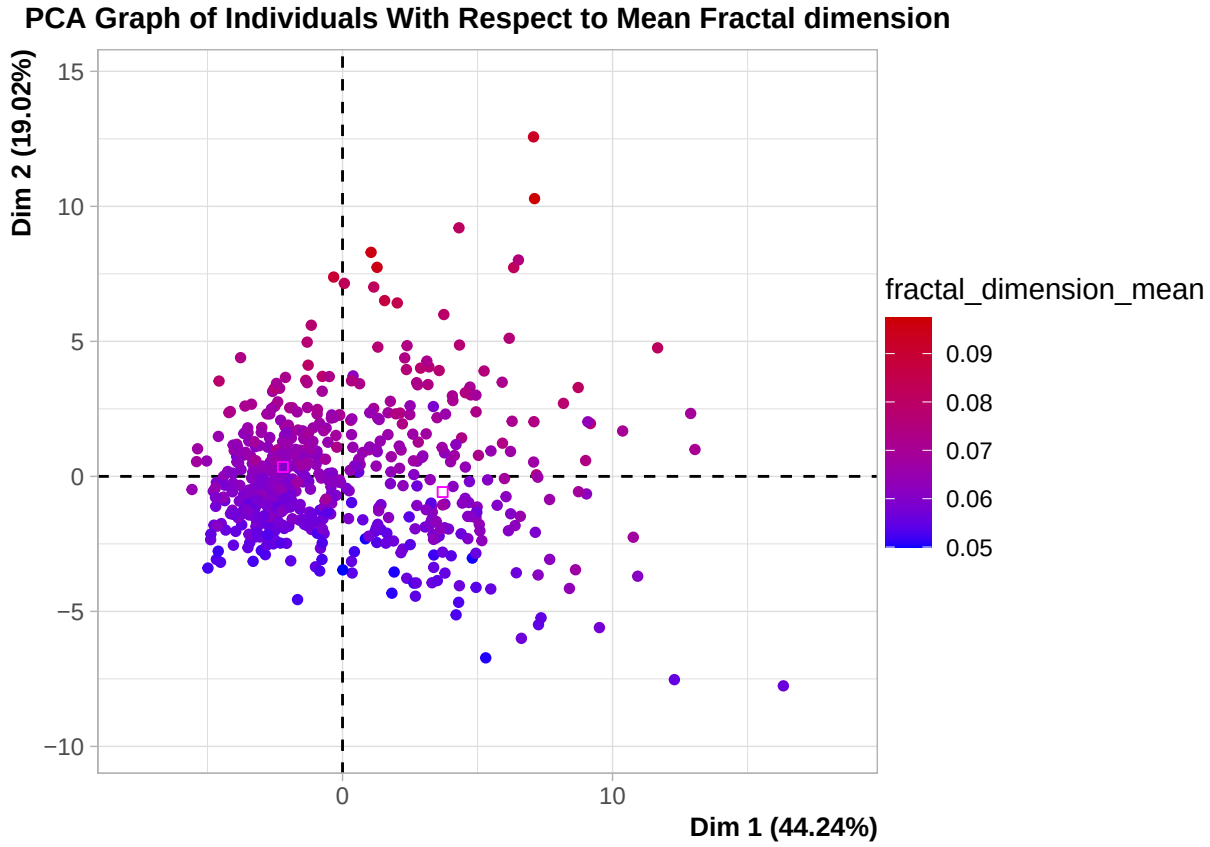
3.2.6.2 Interpretation: Eventhough the high correlations with the second principal component (Dim.2) suggest that Mean Fractal dimension significantly contribute to the variability captured by Dim 2. However, the high p-value in the ANOVA test indicates that the Mean Fractal dimension does not significantly differ between the diagnosed groups (malignant and benign) when considered alone. ($p = 0.7599$). Then **accept Hypothesis**.

```r
# Boxplot for Mean Fractal dimension
boxplot(
  fractal_dimension_mean~ as.factor(diagnosis),
  data = data,
  xlab = "Tumor Type",
  ylab = "Mean Fractal dimension",
  main = "Mean Fractal dimension across Tumor Types"
)
```



3.2.6.3 Visualization:

```
plot.PCA(res.pca,
    choix = "ind",
    hab = "fractal_dimension_mean",
    invisible = c("var"),
    cex = 1,
    autoLab = c("no"),
    title = "PCA Graph of Individuals With Respect to Mean Fractal dimension",
    label = c("none")
)
```



For this variables, the high correlations with the second principal component (Dim.2) does not indicate that this variable varies across different tumor types. It suggests that Mean Fractal dimension significantly contribute to the variability captured by Dim 2. As we can see in the box plot that the mean values of 2 types are no different at all. Therefore we cannot reject the null hypothesis.

3.2.7 Hypothesis 6: All 3 Fractal dimension variables effect on benign and malignant tumors.

- Null Hypothesis (H0): There is no association between the diagnosis of the patients and their combined fractal dimension values.

- Alternative Hypothesis (H1): There is an association between the diagnosis of the patients and their combined fractal dimension values.

3.2.7.1 Test: ANOVA for 3 Mean Fractal dimension variables among tumor types.

```
# ANOVA for 3 Mean Fractal dimension variables among tumor types.
h6 <- lm(
  fractal_dimension_mean + fractal_dimension_se + fractal_dimension_worst ~ as.factor(di
  data = data
)
anova(h6)
```

```
## Analysis of Variance Table
##
## Response: fractal_dimension_mean + fractal_dimension_se + fractal_dimension_worst
##                         Df  Sum Sq   Mean Sq F value    Pr(>F)
## as.factor(diagnosis)     1 0.02006 0.0200625  31.984 2.469e-08 ***
## Residuals              566 0.35503 0.0006273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.2.7.2 Interpretation: When we used the combination of the 3 variables of Fractal dimension, this leads to a small p-value (2.155e-08 < 0.001), then there must be an association between the diagnosis of the patients and their combined fractal dimension values => **reject Hypothesis**.

4 Conclusion and Discussion

4.1 Summary of results

4.1.1 Keypoints from analysis and testing

The findings, particularly from PCA, delineate the pivotal role of specific features related to breast mass characteristics derived from digitized images of FNA of breast masses in diagnosing the existing of breast cancer.

With hierarchical clustering analysis, the study further emphasized the heterogeneity in the benign and malignant breast cancer patient groups, as evident by the clear division into two clusters (M or B tumors).

The conducted ANOVA tests further validate and emphasize the significance of these chemical components in distinguishing between the wine types.

4.1.2 Statistical significance to the real-world problem

4.2 Comments and Limitations

4.2.1 Recommendations on data quality

In the pursuit of comprehensive data analysis, refining or improving the data table is crucial: firstly, ensuring consistency across columns, especially in categorical variables like 'diagnose,' to eliminate discrepancies and ensure accurate labeling. Next, leveraging feature engineering techniques to enhance analysis depth by creating new features or interaction terms based on domain knowledge. Employing cross-validation techniques becomes crucial for predictive modeling, assessing model performance comprehensively for robust findings. Thorough documentation of data preprocessing steps, including handling missing values, ensures transparency and reproducibility. Additionally, establishing a systematic process for ongoing dataset maintenance and updates guarantees sustained data quality and accuracy, especially for periodically updated datasets. These practices collectively fortify the integrity and reliability of the dataset and analyses conducted upon it.

4.2.2 Data and classification method limitations

PCA is a powerful and flexible technique but it may lose some information and details when reducing data size. This can lead to simplification or distortion of the data and make it more difficult to identify outliers or anomalies. Another disadvantage of PCA is that it can be sensitive to scaling and outliers, which can affect the quality and stability of the results.

For HC, this is a solid, flexible technique that is easy to understand and apply. However, it does not guarantee optimal results or the best possible clustering. The results can depend on the order in which data points are processed, making it difficult to reproduce or generalize them to other data sets with similar characteristics. It cannot handle categorical variables effectively because they cannot be converted to numeric values and thus will not produce meaningful clusters if used in hierarchical clustering algorithms.

Although applying PCA combined with hierarchical clustering and ANOVA makes it easier to visualize taxonomies and select the aspects of the data that contribute the most, this method cannot yield satisfactory classification results and high accuracy. Also, each classification method has its limitations.

Thus, this study has limited capacity for precise interpretation, and require integratation of additional analytical techniques or refining existing ones might enhance the method's precision, allowing for more precise of breast cancer patients.

** Propose some method like K-means, DNN, ... **

5 References