Mini Project 1: Breast Cancer Prediction

Shilpa Gopal

February 24, 2024

Copyright Notice

Mini Project 1: Breast Cancer Prediction"

1. Dataset Description

The selected dataset is the "Breast Cancer Wisconsin (Diagnostic) Dataset" from the UCI Machine Learning Repository. This dataset contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The task is to predict whether the tumor is malignant (M) or benign (B) based on these features.

2. Operations Plan

1. Loading the dataset and explore its structure.
2. Calculate summary statistics (mean, median, minimum, and maximum) for numerical columns
3. Check for missing values(NA) and handling them
4. Data Visualization
5. Compute and visualize correlation.
6. Barplot comparision for analyzing the tumors of the affected women
7. Using histograms visualize the distribution of numerical features.
8. Perform feature scaling or normalization.
9. Train and evaluate machine learning models for breast cancer diagnosis prediction.
10. Predict using the trained random forest model on the test set
11. Calculate confusion matrix

3. R Scripts starts here

Loading Packages

1. Loading the dataset and explore its structure

```
library("ggplot2")
library("caTools")
library("corrplot")
```

```
## corrplot 0.92 loaded
```

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("corrplot")
library("caret")
```

```
## Loading required package: lattice
```

Read Dataset

```
main_dataset <- read.csv("/Users/sheme/sheme/Codes/R_File/multimedia_mini_project/breast_cancer.csv", h
```

Viewing first few rows of Dataset

```
head(main_dataset)
```

```
##         id diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1   842302         M       17.99        10.38         122.80    1001.0
## 2   842517         M       20.57        17.77         132.90    1326.0
## 3 84300903         M       19.69        21.25         130.00    1203.0
## 4 84348301         M       11.42        20.38          77.58     386.1
## 5 84358402         M       20.29        14.34         135.10    1297.0
## 6   843786         M       12.45        15.70          82.57     477.1
##   smoothness_mean compactness_mean concavity_mean concave.points_mean
## 1         0.11840          0.27760         0.3001             0.14710
## 2         0.08474          0.07864         0.0869             0.07017
## 3         0.10960          0.15990         0.1974             0.12790
## 4         0.14250          0.28390         0.2414             0.10520
## 5         0.10030          0.13280         0.1980             0.10430
## 6         0.12780          0.17000         0.1578             0.08089
##   symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se
## 1        0.2419                0.07871    1.0950     0.9053        8.589
## 2        0.1812                0.05667    0.5435     0.7339        3.398
## 3        0.2069                0.05999    0.7456     0.7869        4.585
## 4        0.2597                0.09744    0.4956     1.1560        3.445
## 5        0.1809                0.05883    0.7572     0.7813        5.438
## 6        0.2087                0.07613    0.3345     0.8902        2.217
##   area_se smoothness_se compactness_se concavity_se concave.points_se
## 1  153.40      0.006399        0.04904      0.05373           0.01587
## 2   74.08      0.005225        0.01308      0.01860           0.01340
## 3   94.03      0.006150        0.04006      0.03832           0.02058
## 4   27.23      0.009110        0.07458      0.05661           0.01867
## 5   94.44      0.011490        0.02461      0.05688           0.01885
## 6   27.19      0.007510        0.03345      0.03672           0.01137
##   symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst
## 1     0.03003             0.006193        25.38         17.33          184.60
## 2     0.01389             0.003532        24.99         23.41          158.80
## 3     0.02250             0.004571        23.57         25.53          152.50
```

```
## 4     0.05963             0.009208        14.91         26.50           98.87
## 5     0.01756             0.005115        22.54         16.67          152.20
## 6     0.02165             0.005082        15.47         23.75          103.40
##   area_worst smoothness_worst compactness_worst concavity_worst
## 1     2019.0           0.1622            0.6656          0.7119
## 2     1956.0           0.1238            0.1866          0.2416
## 3     1709.0           0.1444            0.4245          0.4504
## 4      567.7           0.2098            0.8663          0.6869
## 5     1575.0           0.1374            0.2050          0.4000
## 6      741.6           0.1791            0.5249          0.5355
##   concave.points_worst symmetry_worst fractal_dimension_worst  X
## 1               0.2654         0.4601                 0.11890 NA
## 2               0.1860         0.2750                 0.08902 NA
## 3               0.2430         0.3613                 0.08758 NA
## 4               0.2575         0.6638                 0.17300 NA
## 5               0.1625         0.2364                 0.07678 NA
## 6               0.1741         0.3985                 0.12440 NA
```

Display the structure of the dataset

2. Calculate summary statistics (mean, median, minimum, and maximum) for numerical columns

```
str(main_dataset)
```

```
## 'data.frame':    569 obs. of  33 variables:
##  $ id                      : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844
##  $ diagnosis               : chr  "M" "M" "M" "M" ...
##  $ radius_mean             : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean            : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean          : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean               : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean         : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean        : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean          : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points_mean     : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean           : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean  : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se               : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se              : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se            : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                 : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se           : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se          : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se            : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points_se       : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se             : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se    : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst            : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst           : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst         : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst              : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst        : num  0.162 0.124 0.144 0.21 0.137 ...
```

```
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
##  $ X                      : logi  NA NA NA NA NA NA ...
```

Displaying Dimension of Dataset

```
dim(main_dataset)
```

```
## [1] 569  33
```

Summary of the Dataset

2. Calculate summary statistics (mean, median, minimum, and maximum) for numerical columns

```
summary(main_dataset)
```

```
##        id              diagnosis          radius_mean      texture_mean
##  Min.   :     8670   Length:569         Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   Class :character   1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024   Mode  :character   Median :13.370   Median :18.84
##  Mean   : 30371831                      Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129                      3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502                      Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##  concavity_mean    concave.points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##    radius_se        texture_se      perimeter_se       area_se
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se     concave.points_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
##  Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
```

```
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
##   symmetry_se       fractal_dimension_se  radius_worst   texture_worst
##  Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02
##  1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08
##  Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41
##  Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68
##  3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72
##  Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54
##  perimeter_worst    area_worst     smoothness_worst  compactness_worst
##  Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729
##  1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720
##  Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190
##  Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427
##  3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910
##  Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800
##  concavity_worst  concave.points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.0000   Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2267   Median :0.09993      Median :0.2822   Median :0.08004
##  Mean   :0.2722   Mean   :0.11461      Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :1.2520   Max.   :0.29100      Max.   :0.6638   Max.   :0.20750
##     X
##  Mode:logical
##  NA's:569
##
##
##
##
```

Remove missing value NAs (if applicable)

Check for missing values(NA) and handling them

```
main_dataset <- main_dataset[-33]
summary(main_dataset)
```

```
##        id             diagnosis          radius_mean      texture_mean
##  Min.   :     8670   Length:569         Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   Class :character   1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024   Mode  :character   Median :13.370   Median :18.84
##  Mean   : 30371831                      Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129                      3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502                      Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
```

```
##  concavity_mean    concave.points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##    radius_se        texture_se      perimeter_se       area_se
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se     concave.points_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
##  Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
##   symmetry_se       fractal_dimension_se  radius_worst   texture_worst
##  Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02
##  1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08
##  Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41
##  Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68
##  3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72
##  Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54
##  perimeter_worst    area_worst     smoothness_worst  compactness_worst
##  Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729
##  1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720
##  Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190
##  Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427
##  3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910
##  Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800
##  concavity_worst  concave.points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.0000   Min.   :0.00000     Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.1145   1st Qu.:0.06493     1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2267   Median :0.09993     Median :0.2822   Median :0.08004
##  Mean   :0.2722   Mean   :0.11461     Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3829   3rd Qu.:0.16140     3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :1.2520   Max.   :0.29100     Max.   :0.6638   Max.   :0.20750
```

Frequency of Cancer Diagnosis

```
# Number of women affected with benign and malignant tumor
main_dataset %>% count(diagnosis)
```

```
##   diagnosis   n
## 1         B 357
## 2         M 212
```

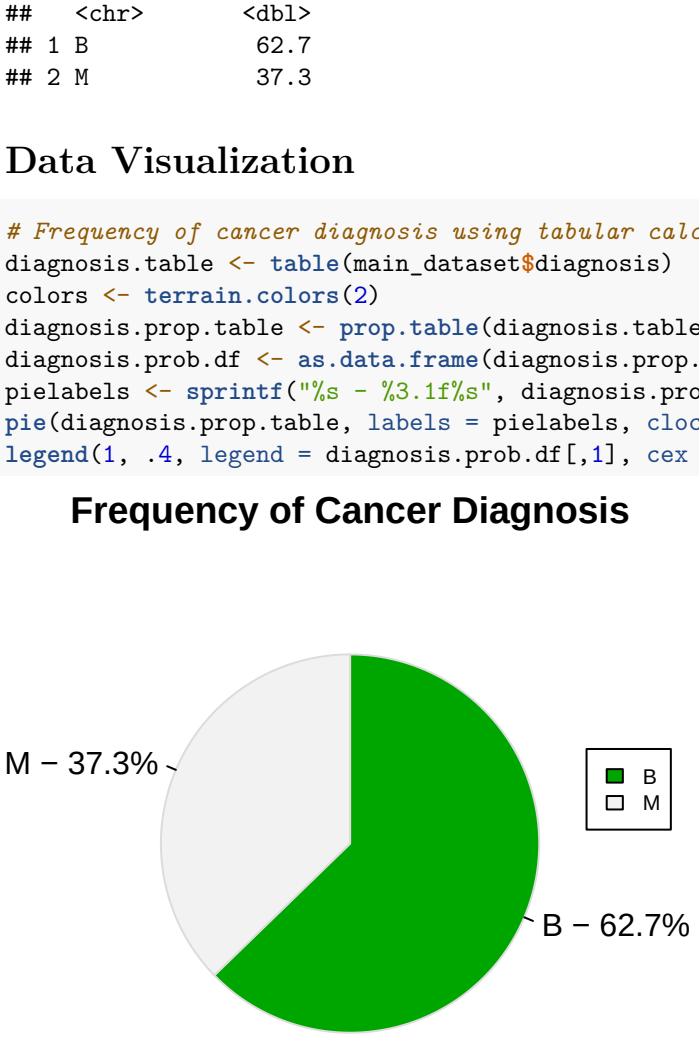
```
# Percentage of women affected with benign and malignant tumor
main_dataset %>% count(diagnosis) %>%
```

```
  group_by(diagnosis) %>%
  summarize(perc_dx = round((n / 569) * 100, 2))
```

```
## # A tibble: 2 x 2
##   diagnosis perc_dx
##   <chr>       <dbl>
## 1 B            62.7
## 2 M            37.3
```

Data Visualization

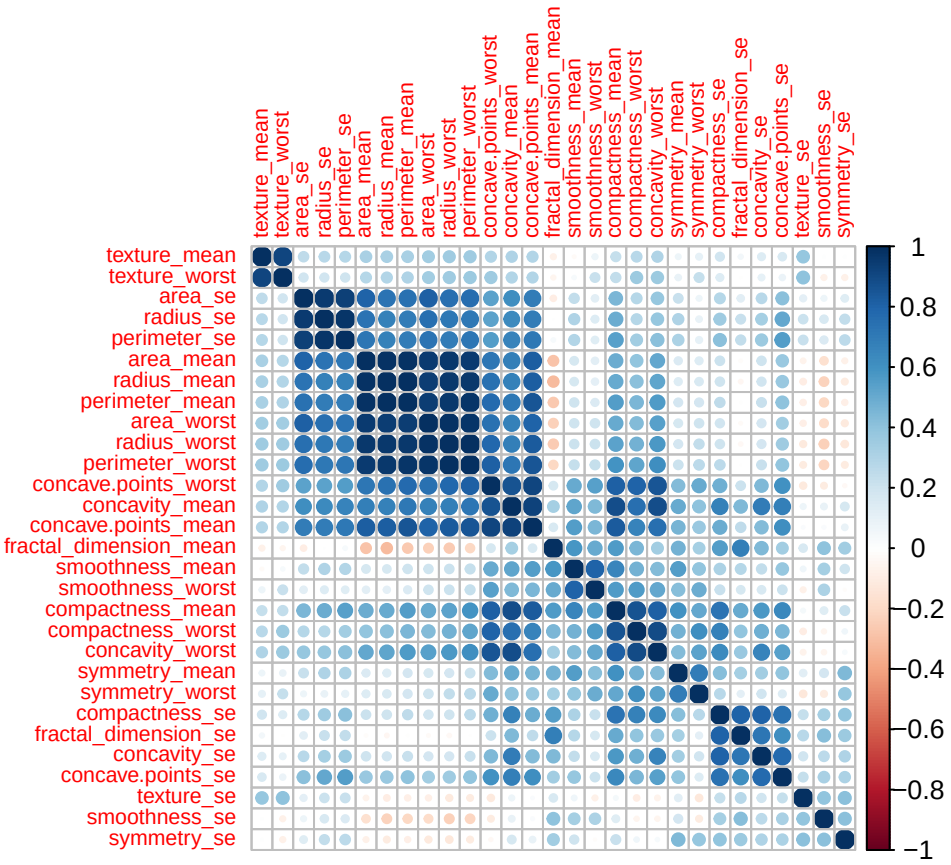
```
# Frequency of cancer diagnosis using tabular calculation
diagnosis.table <- table(main_dataset$diagnosis)
colors <- terrain.colors(2)
diagnosis.prop.table <- prop.table(diagnosis.table) * 100
diagnosis.prob.df <- as.data.frame(diagnosis.prop.table)
pielabels <- sprintf("%s - %3.1f%s", diagnosis.prob.df[,1], diagnosis.prop.table, "%")
pie(diagnosis.prop.table, labels = pielabels, clockwise = TRUE, col = colors, border = "gainsboro", radi
legend(1, .4, legend = diagnosis.prob.df[,1], cex = 0.7, fill = colors)
```



Correlation plot

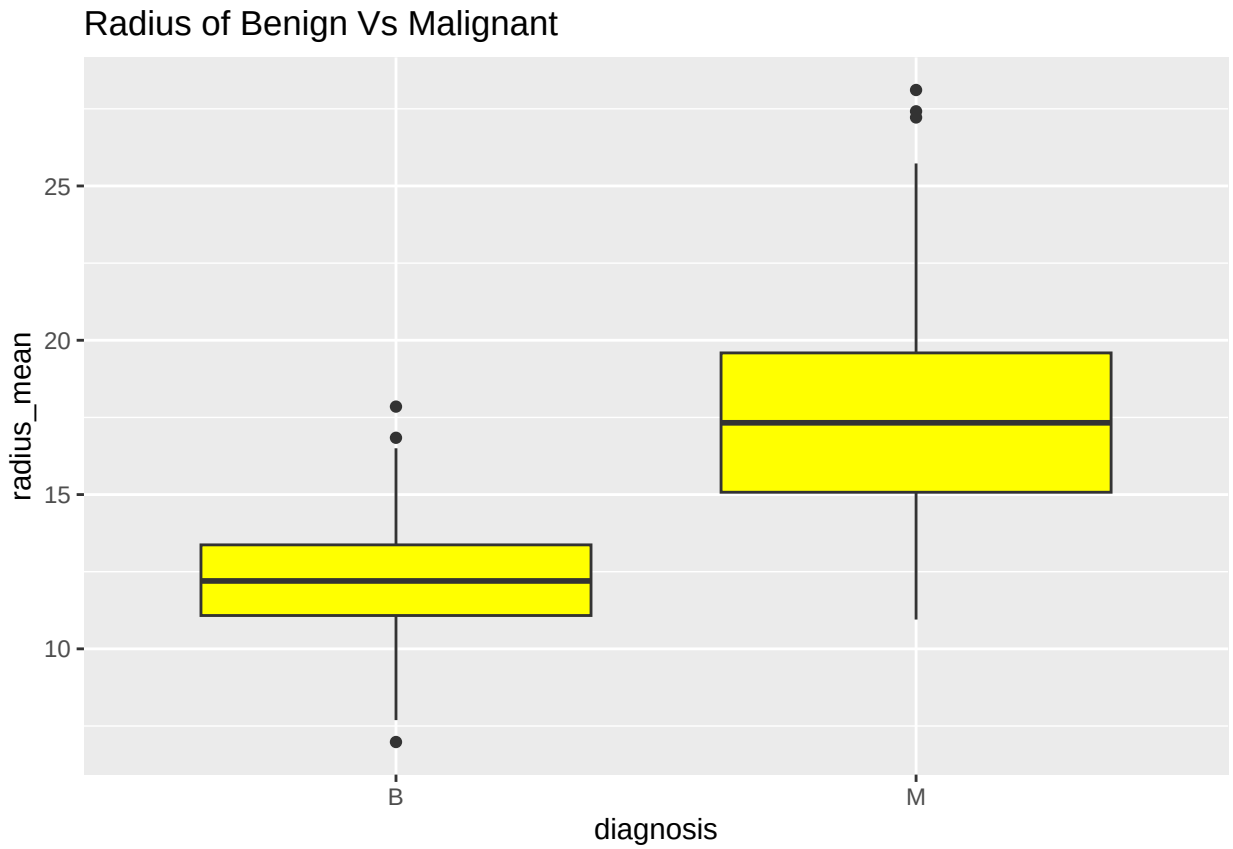
5. Compute and visualize correlation.

```
# Correlation plot - relationship with variator
c <- cor(main_dataset[,3:31])
corrplot(c, order = "hclust", tl.cex = 0.7)
```

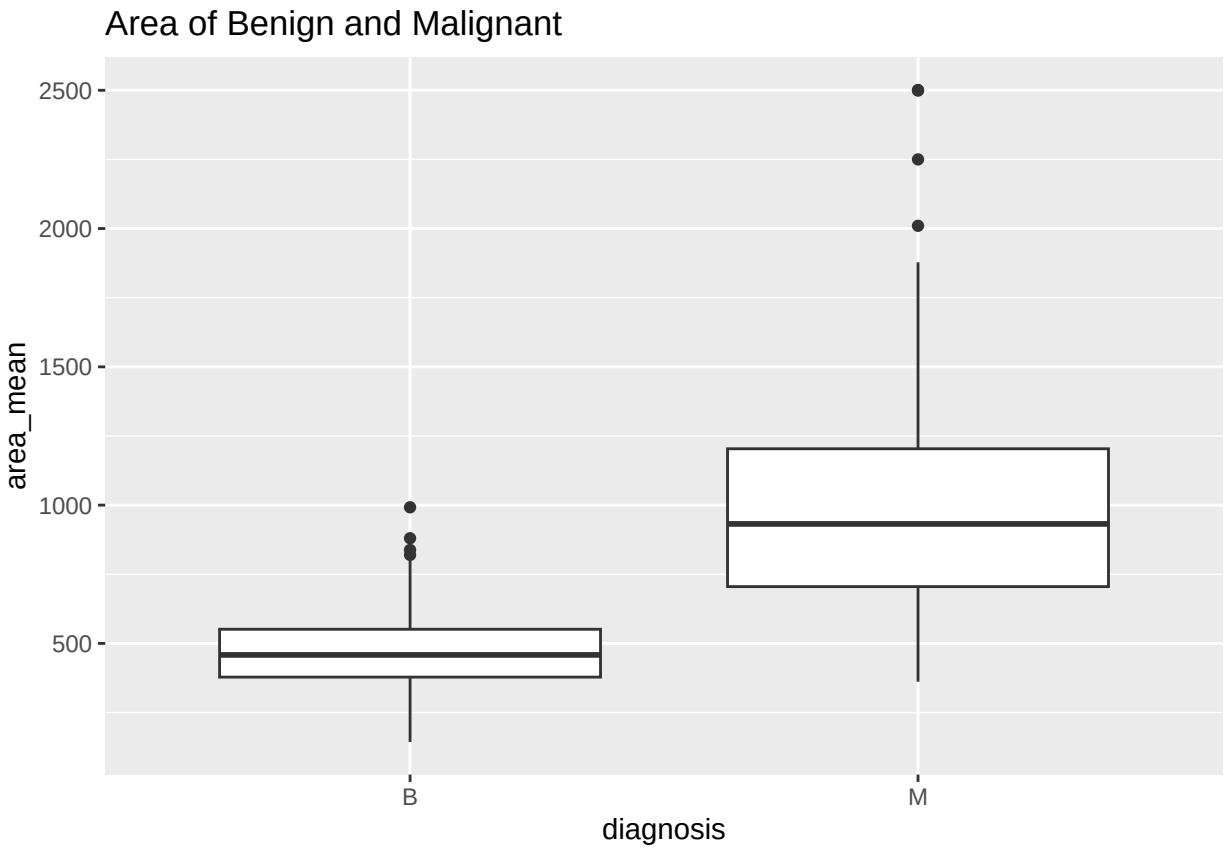


Columns comparions

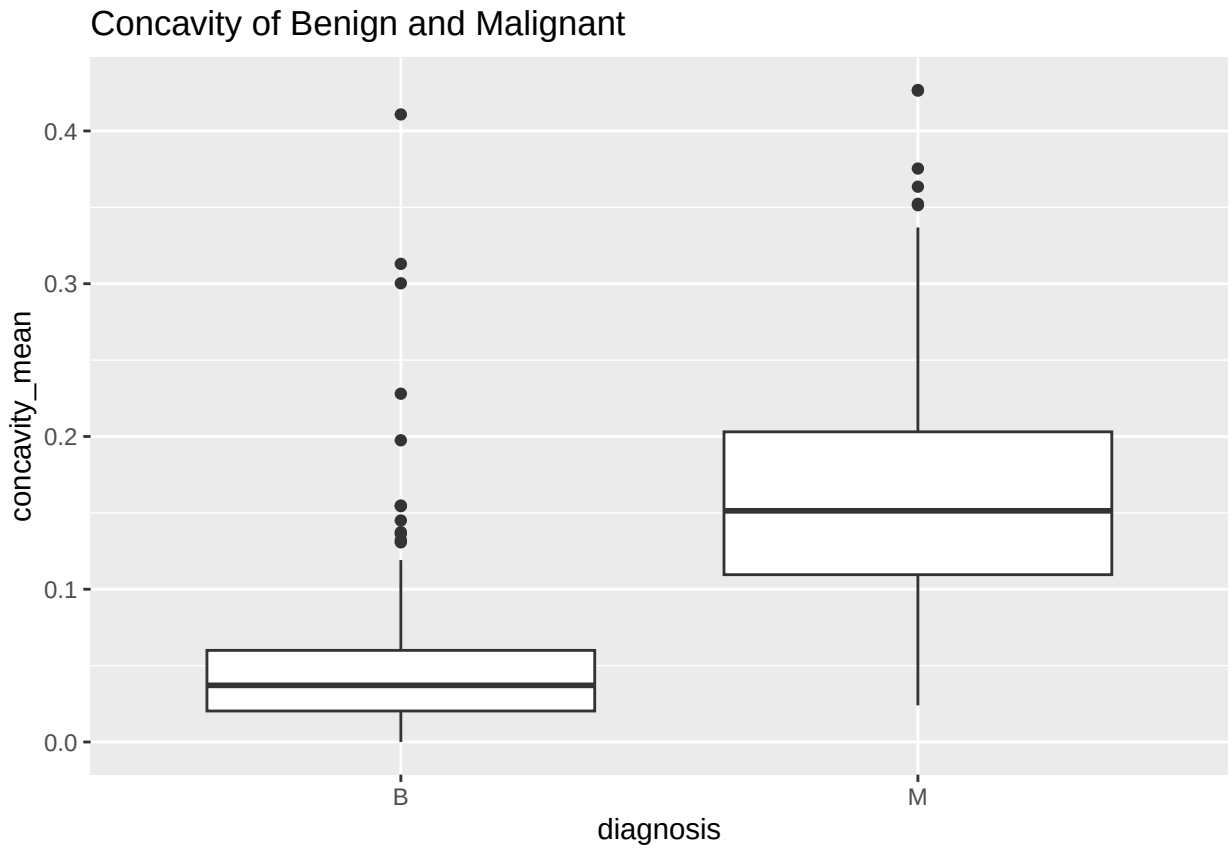
```
# Comparing the radius column, area column, concavity column of benign and malignant stage
ggplot(main_dataset, aes(x=diagnosis, y=radius_mean, fill="pink")) + geom_boxplot(fill = "yellow") + gg
```




```
ggplot(main_dataset, aes(x=diagnosis, y=area_mean)) + geom_boxplot() + ggtitle("Area of Benign and Mali
```




```
ggplot(main_dataset, aes(x=diagnosis, y=concavity_mean)) + geom_boxplot() + ggtitle("Concavity of Benig
```

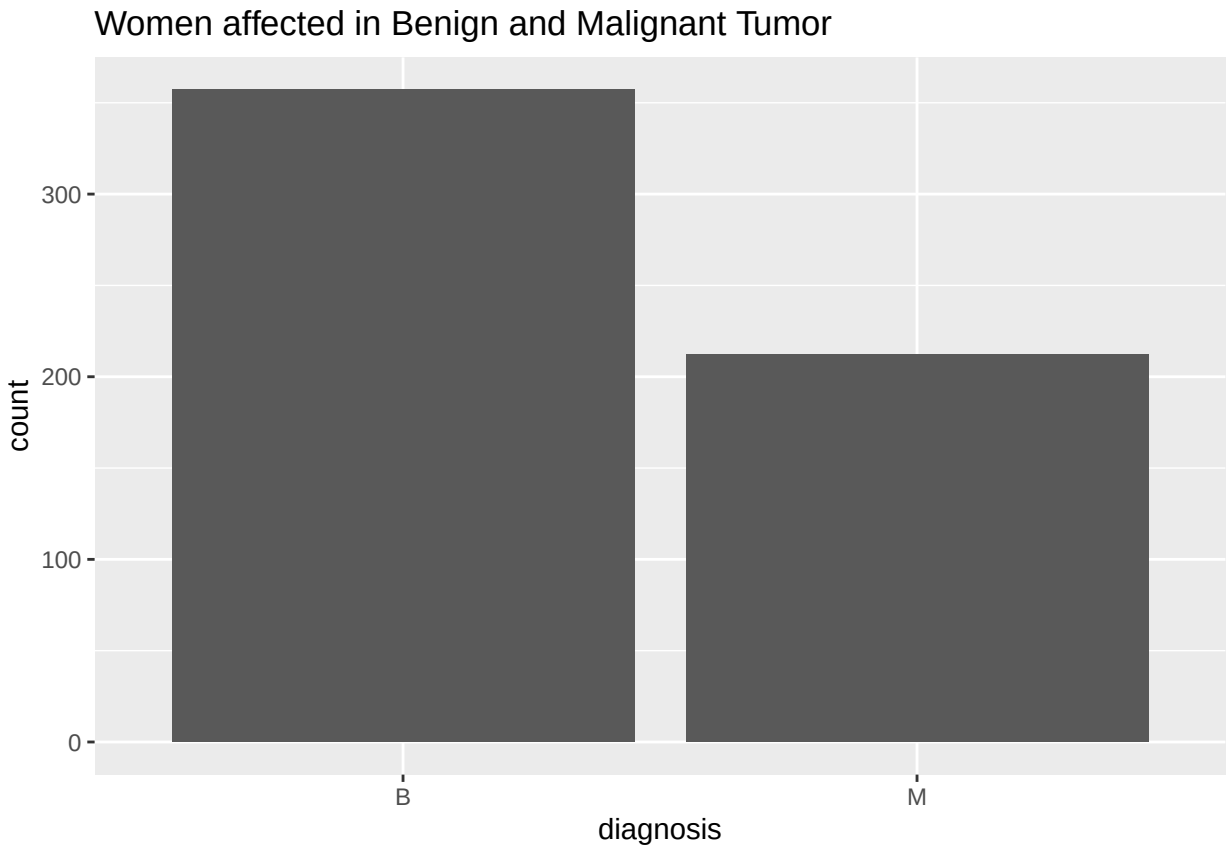



```
# Observation from the box plot - malignant cells have higher radius, area, concavity mean than benign
```

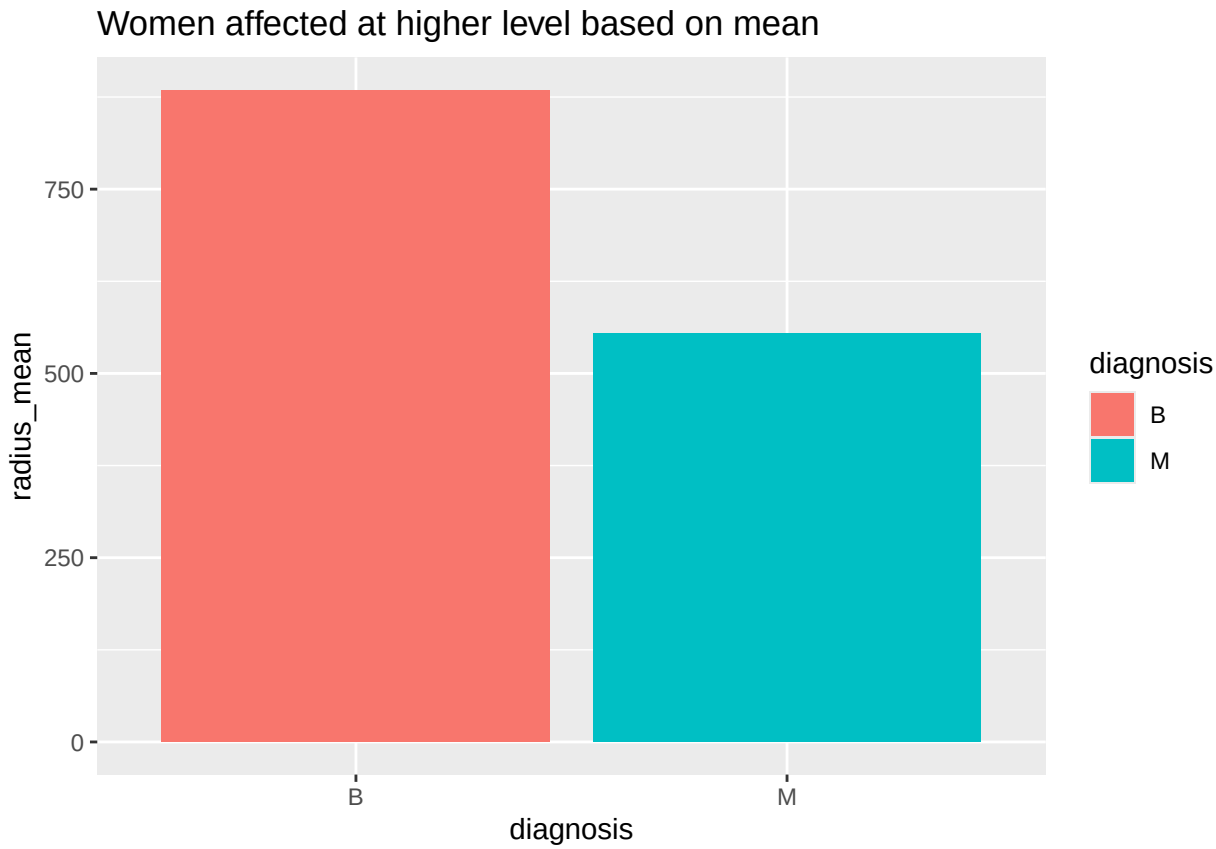
6. Barplot comparision

```
# Barplot for analyzing the tumors of the affected women
ggplot(main_dataset, aes(x=diagnosis, fill = texture_mean)) + geom_bar() + ggtitle("Women affected in B
```

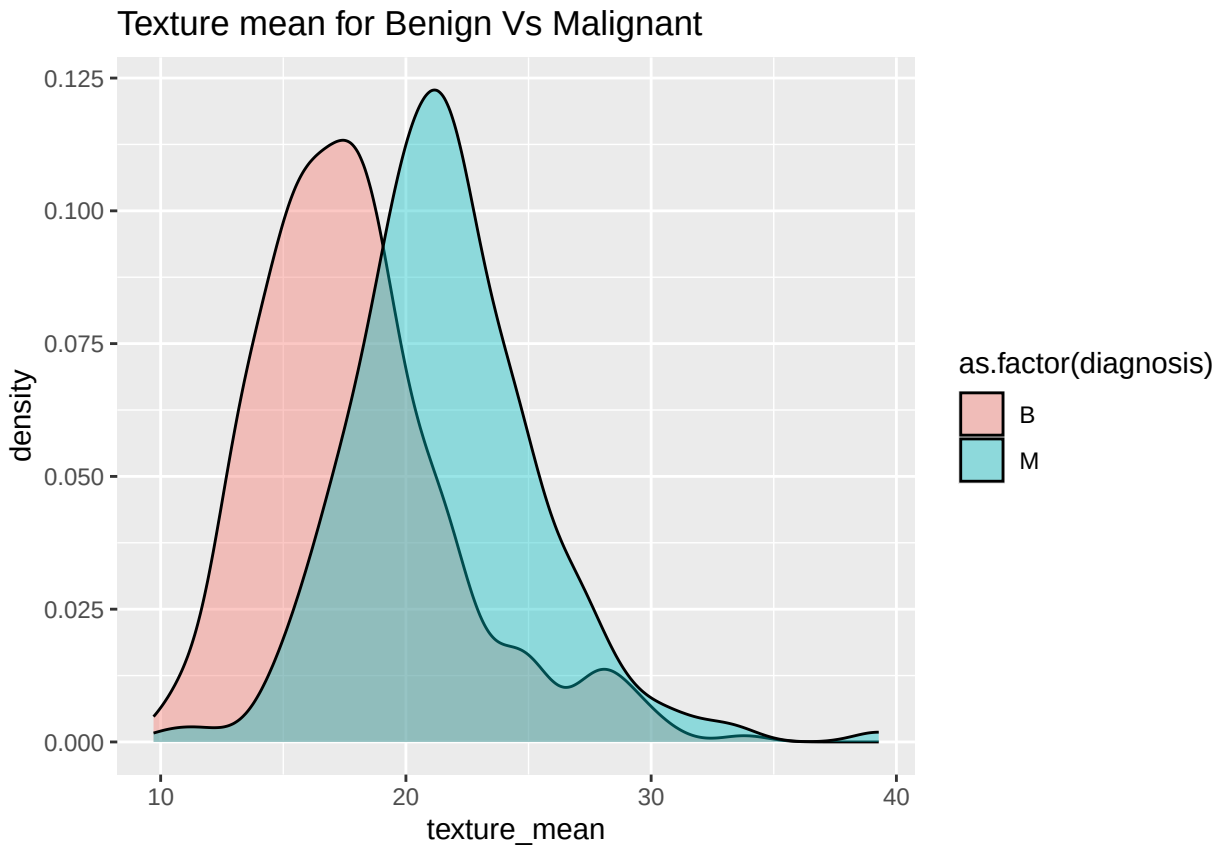
```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



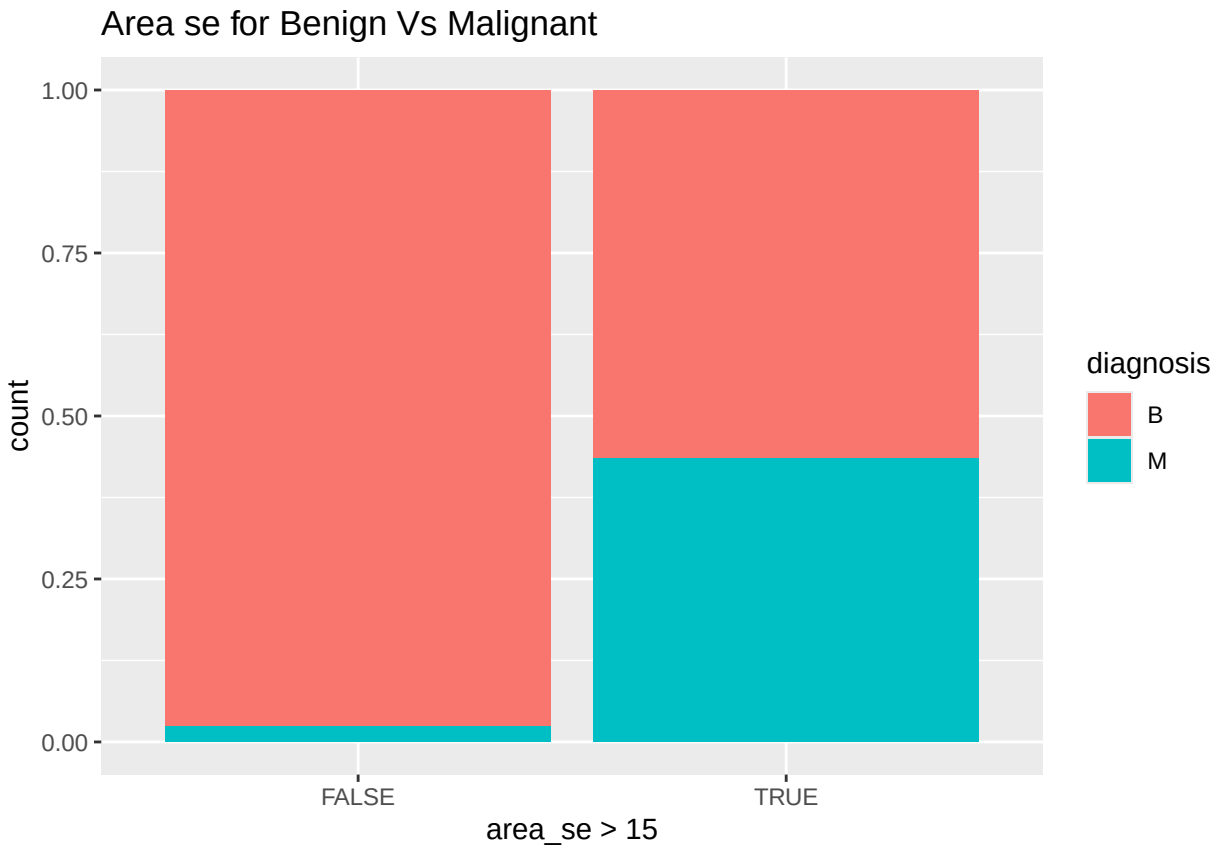

```
# Women affected at higher levels based on mean from the analysis of boxplot
sel_data <- main_dataset[main_dataset$radius_mean > 10 & main_dataset$radius_mean < 15 & main_dataset$c

ggplot(sel_data, aes(x=diagnosis, y=radius_mean, fill = diagnosis)) + geom_col() + ggtitle("Women affec
```




```
# Density plot based on texture mean
ggplot(main_dataset, aes(x=texture_mean, fill = as.factor(diagnosis))) + geom_density(alpha = 0.4) + gg
```



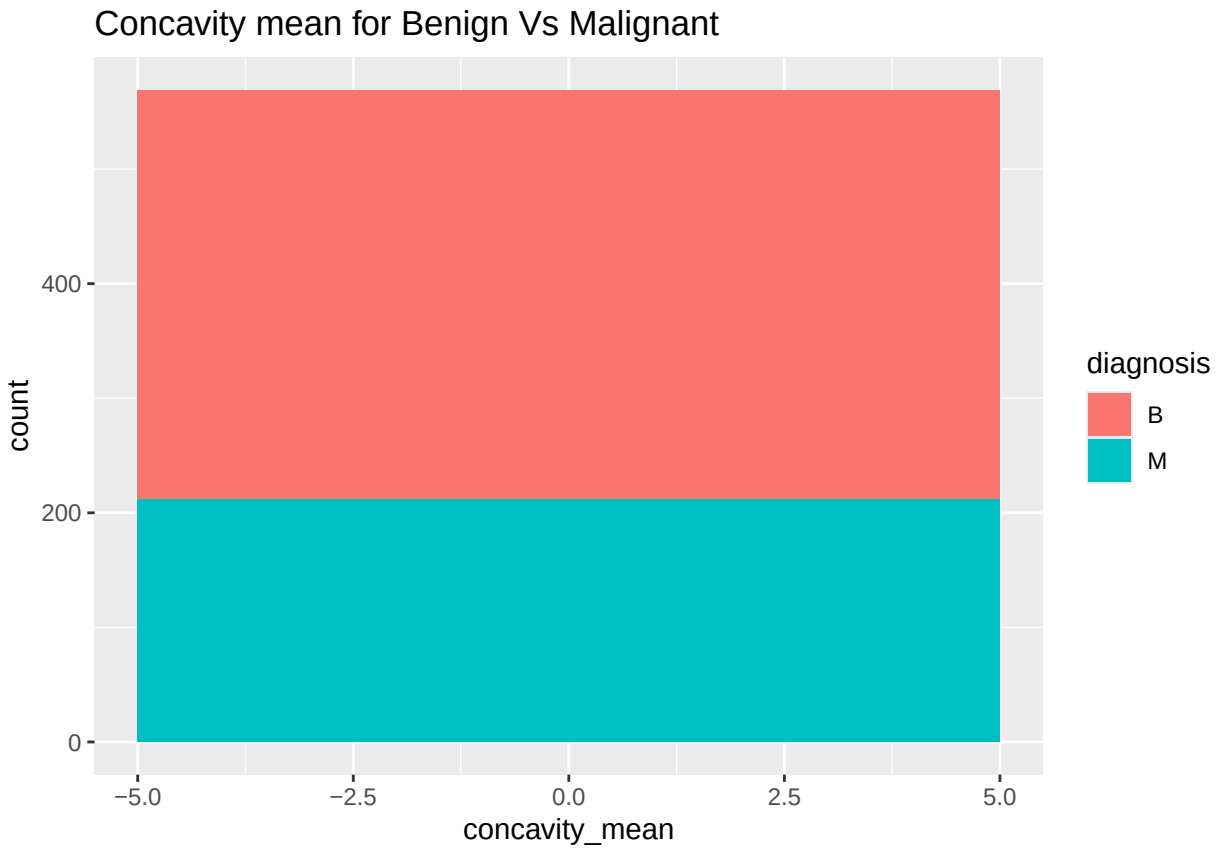

```
# Barplot for area_se
ggplot(main_dataset, aes(x=area_se > 15, fill = diagnosis)) + geom_bar(position = "fill") + ggtitle("Ar
```



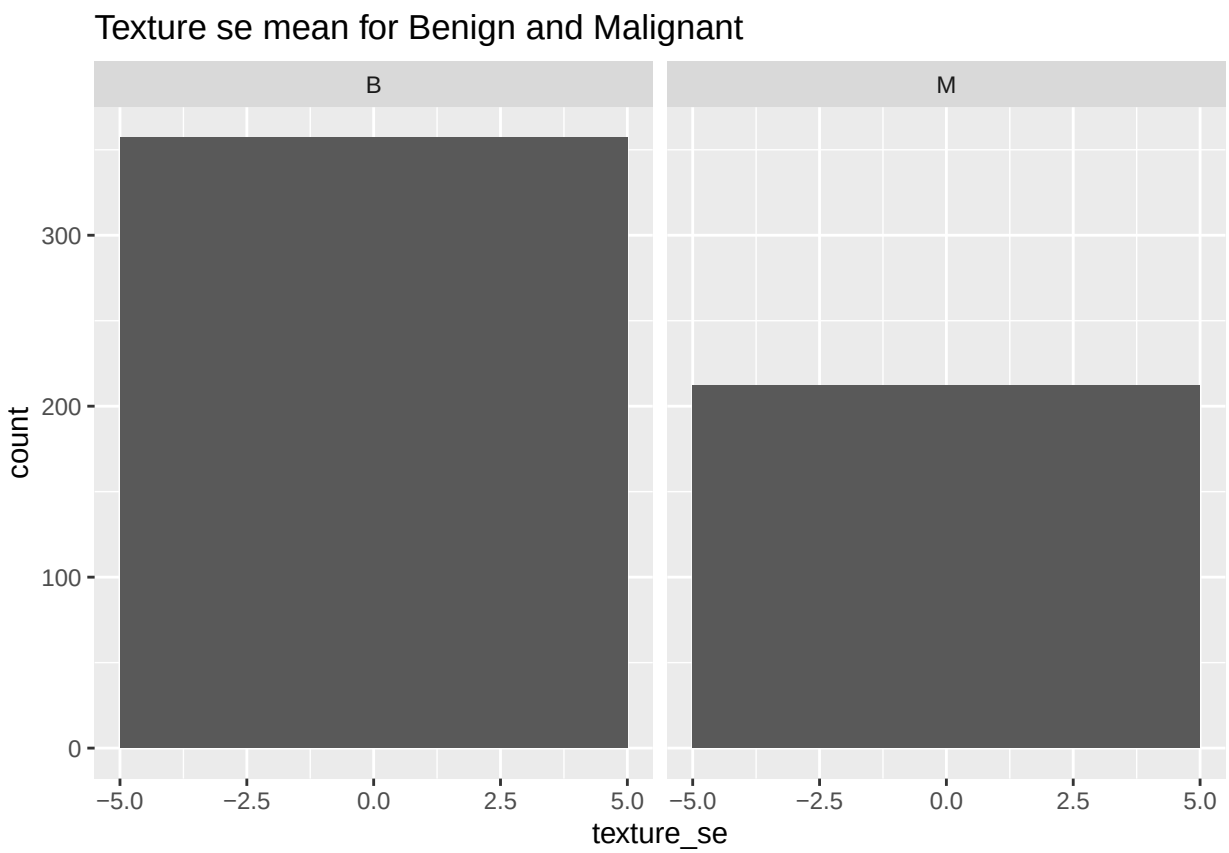
Checking distribution of Data via histograms

7. Using histograms visualize the distribution of numerical features.

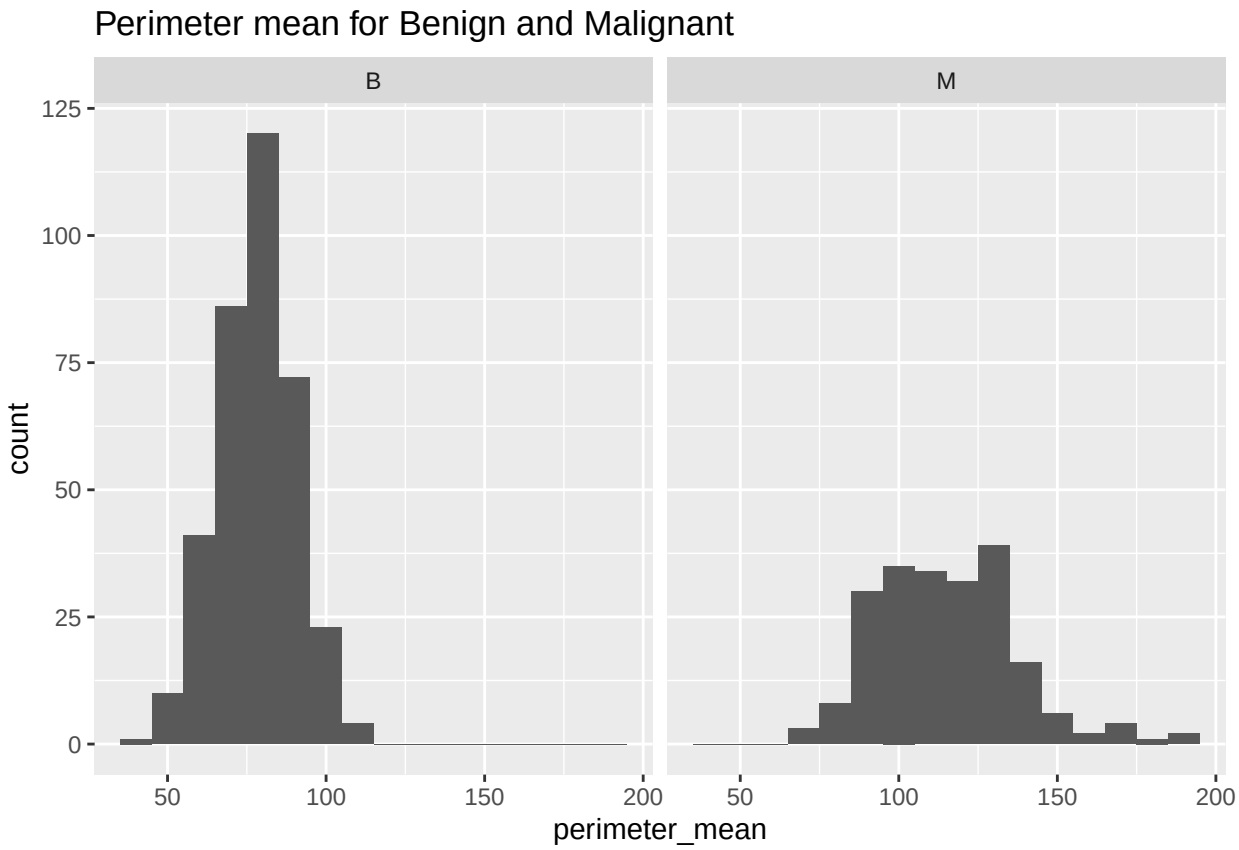
```
ggplot(main_dataset, aes(x=concavity_mean, fill = diagnosis)) + geom_histogram(binwidth = 10) + ggtitle
```




```
ggplot(main_dataset, aes(x = texture_se)) + geom_histogram(binwidth = 10) + facet_wrap(~ diagnosis) + g
```




```
ggplot(main_dataset, aes(x = perimeter_mean)) + geom_histogram(binwidth = 10) + facet_wrap(~ diagnosis)
```



Train the Algorithm

Split the data into training and testing sets using Logistic Regression

```
main_dataset$diagnosis <- factor(main_dataset$diagnosis, levels = c("B", "M"))

split <- sample.split(main_dataset$diagnosis, SplitRatio = 0.65)
main_dataset <- main_dataset[-33]
training_set <- subset(main_dataset, split == TRUE)
test_set <- subset(main_dataset, split == FALSE)
```

Normalization process

8. Perform feature scaling or normalization.

```
training_set[,3:32] <- scale(training_set[,3:32])
test_set[,3:32] <- scale(test_set[,3:32])
```

9. Train and evaluate machine learning models for breast cancer diagnosis prediction.

Create training and testing sets

```
set.seed(1234)
data_index <- createDataPartition(main_dataset$diagnosis, p = 0.7, list = FALSE)
train_data <- main_dataset[data_index, -1]
test_data <- main_dataset[-data_index, -1]
```

Building Model

10. Predict using the trained random forest model on the test set

```
fitControl <- trainControl(
  method="cv",
  number = 5,
  preProcOptions = list(thresh = 0.99), # threshold for PCA preprocess
  classProbs = TRUE,
  summaryFunction = twoClassSummary
)

# Random Forest model
model_rf <- train(
  diagnosis ~ .,
  train_data,
  method = "ranger",
  metric = "ROC",
  preProcess = c('center', 'scale'),
  trControl = fitControl
)
```

Predict using the trained random forest model on the test set

```
pred_rf <- predict(model_rf, test_data)

# Convert the actual column to a factor with levels "B" and "M"
results_df <- tibble(
  predicted = pred_rf,
  actual = test_data$diagnosis
)
results_df$actual <- factor(results_df$actual, levels = c("B", "M"))
```

Calculate confusion matrix

11. Calculate confusion matrix

```
cm_rf <- confusionMatrix(data = results_df$predicted, reference = results_df$actual, positive = "M")
```

```r
# printing
print(cm_rf)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 106   6
##          M   1  57
##
##                Accuracy : 0.9588
##                  95% CI : (0.917, 0.9833)
##     No Information Rate : 0.6294
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9103
##
##  Mcnemar's Test P-Value : 0.1306
##
##             Sensitivity : 0.9048
##             Specificity : 0.9907
##          Pos Pred Value : 0.9828
##          Neg Pred Value : 0.9464
##              Prevalence : 0.3706
##          Detection Rate : 0.3353
##    Detection Prevalence : 0.3412
##       Balanced Accuracy : 0.9477
##
##        'Positive' Class : M
##
```