Breast Cancer Analysis

- Apeksha Shetty
- Yash Shah
- Roshni Suhanda

Reading and Displaying the dataset:

```
> #Reading the dataset
> breast_cancer <- read.csv("C:\\Users\\APEKSHA\\Downloads\\wisc_bc_data.csv")
> #Displaying the dataset using head function
> head(breast_cancer)
        id diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
1 87139402         B       12.32        12.39          78.85     464.1         0.10280
2  8910251         B       10.60        18.95          69.28     346.4         0.09688
3   905520         B       11.04        16.83          70.92     373.2         0.10770
4   868871         B       11.28        13.39          73.00     384.8         0.11640
5  9012568         B       15.19        13.21          97.65     711.8         0.07963
6   906539         B       11.57        19.04          74.20     409.7         0.08546
  compactness_mean concavity_mean points_mean symmetry_mean dimension_mean radius_se texture_se
1          0.06981        0.03987     0.03700        0.1959        0.05955    0.2360     0.6656
2          0.11470        0.06387     0.02642        0.1922        0.06491    0.4505     1.1970
3          0.07804        0.03046     0.02480        0.1714        0.06340    0.1967     1.3870
4          0.11360        0.04635     0.04796        0.1771        0.06072    0.3384     1.3430
5          0.06934        0.03393     0.02657        0.1721        0.05544    0.1783     0.4125
6          0.07722        0.05485     0.01428        0.2031        0.06267    0.2864     1.4400
  perimeter_se area_se smoothness_se compactness_se concavity_se points_se symmetry_se
1        1.670   17.43      0.008045       0.011800      0.01683  0.012410     0.01924
2        3.430   27.10      0.007470       0.035810      0.03354  0.013650     0.03504
3        1.342   13.54      0.005158       0.009355      0.01056  0.007483     0.01718
4        1.851   26.33      0.011270       0.034980      0.02187  0.019650     0.01580
5        1.338   17.72      0.005012       0.014850      0.01551  0.009155     0.01647
6        2.206   20.30      0.007278       0.020470      0.04447  0.008799     0.01868
  dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst
1     0.002248        13.50         15.64           86.97      549.1           0.1385
2     0.003318        11.88         22.94           78.28      424.8           0.1213
3     0.002198        12.41         26.44           79.93      471.4           0.1369
4     0.003442        11.92         15.77           76.53      434.0           0.1367
5     0.001767        16.20         15.73          104.50      819.1           0.1126
6     0.003339        13.07         26.98           86.43      520.5           0.1249
  compactness_worst concavity_worst points_worst symmetry_worst dimension_worst
1            0.1266         0.12420      0.09391         0.2827         0.06771
2            0.2515         0.19160      0.07926         0.2940         0.07587
3            0.1482         0.10670      0.07431         0.2998         0.07881
4            0.1822         0.08669      0.08611         0.2102         0.06784
5            0.1737         0.13620      0.08178         0.2487         0.06766
6            0.1937         0.25600      0.06664         0.3035         0.08284
>
```
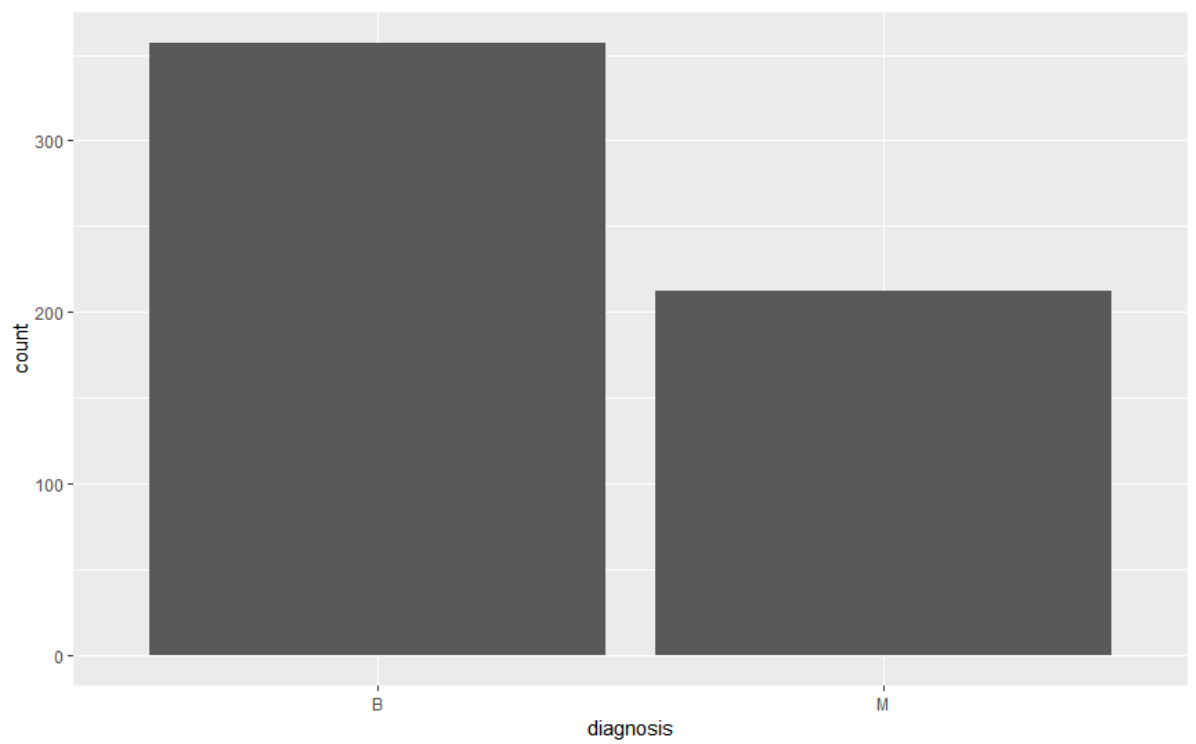
Generate Frequency Table:

> This shows the frequency of two type of diagnosis categories:
  1. Benign(B) - 357
  2. Malignant(M) – 212

```
> diagnosis.table <- table(breast_cancer$diagnosis)
> diagnosis.table

  B   M
357 212
```
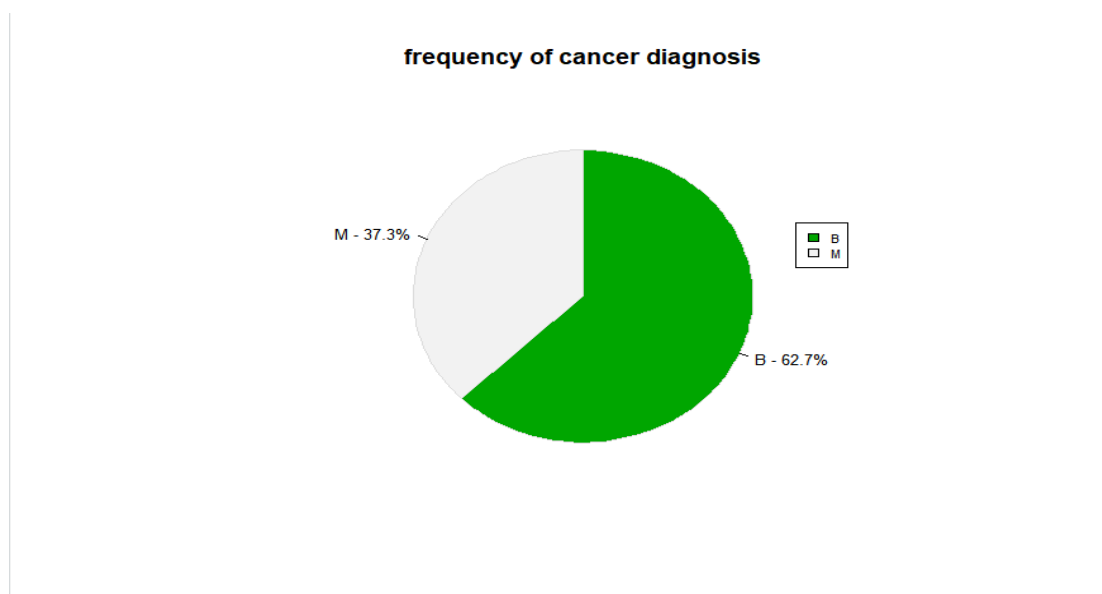
Bar plot:

➤ The above frequency of benign and malignant is shown using bar plot.

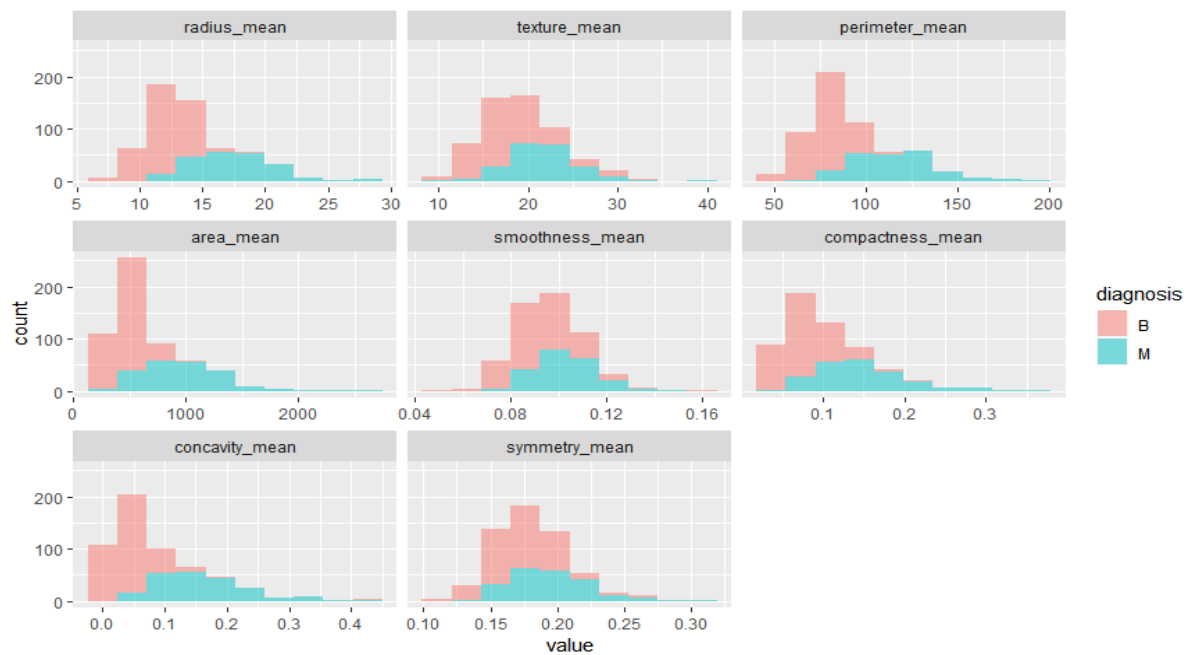➤ It is seen that the patients are more diagnosed as benign compared to malignant.

Pie Chart:

➤ Pie Chart gives us the frequency of the two categories on the basis of percentage.
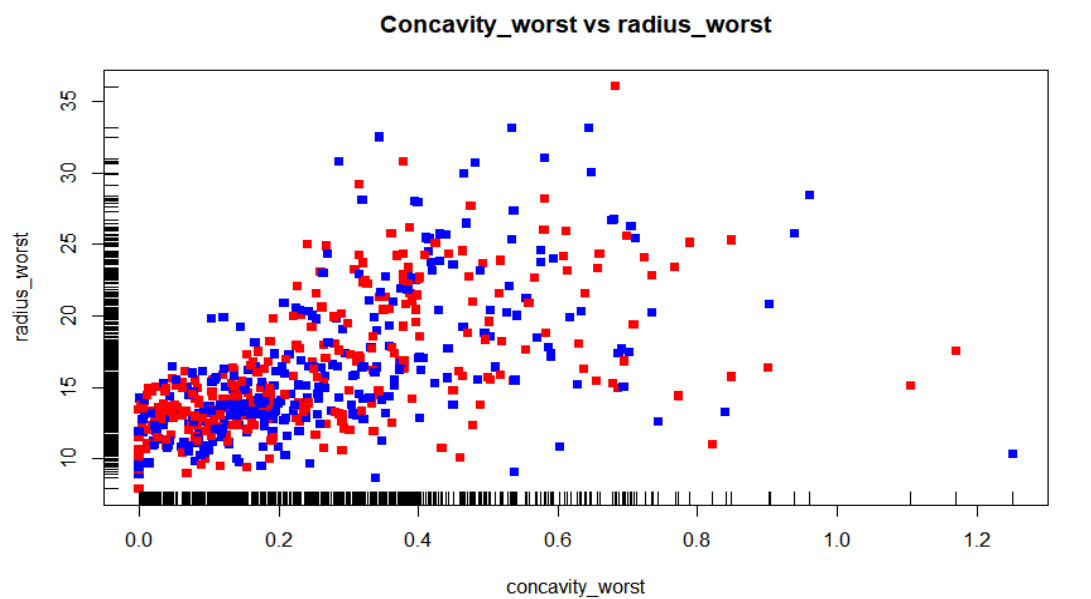
Benign - 62.7%

Malignant – 37.3%

Histogram:

➢ Using mean values of the dataset we plot histogram for the given data.

➢ In this plot, the histogram of two categories(B & M) is generated in one graph for better comparison and for visualization.
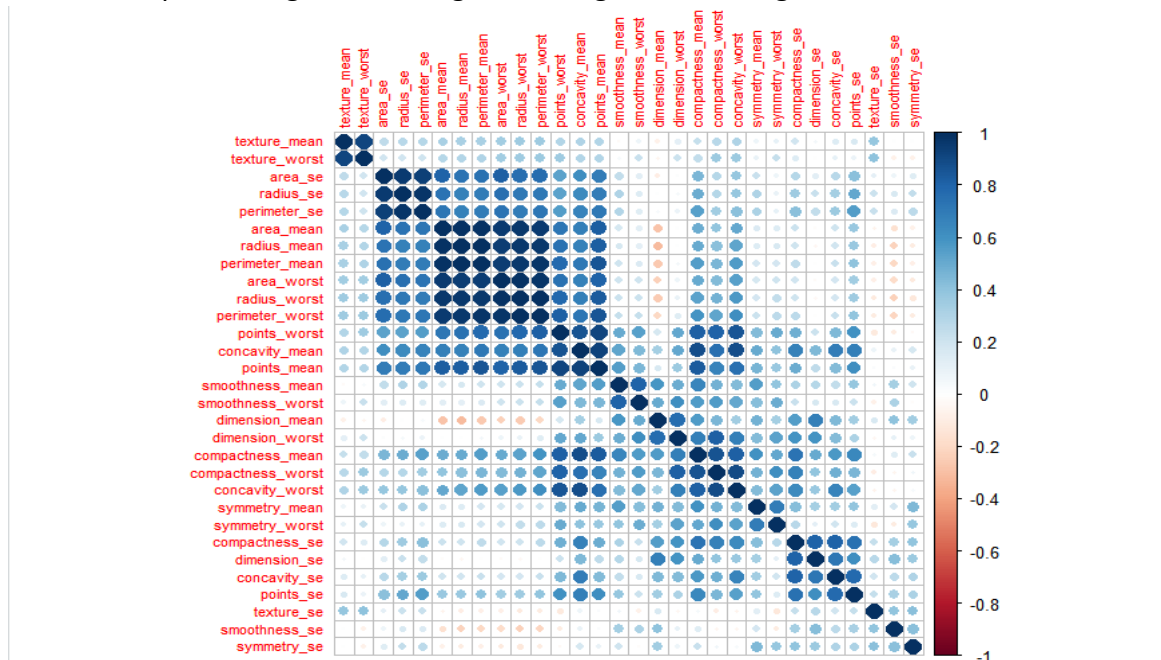

Scatterplot:

➢ The fig. shows the relation between two variables ie. concavity_worst vs radius_worst.

➢ This represents that as concavity and radius increases, the patient can be more prone towards breast cancer.

Correlation Matrix:

➢ Correlation Matrix establishes a relationship or connection between two or more measures statistically speaking tells you interdependence of variable quantities.

➢ The color key on right side gives you the values of the correlation.

➢ The dark blue color represents number one strong positive correlation whereas dark red representing number negative one gives weak negative correlation.

Scatterplot Matrix:

➢ Scatterplot Matrix is a table of scatter plots. It is used when we have to visualize several plots together.

➢ The names in the row represent Y-axis and the names in the columns represent X-axis.