Breast Cancer Analysis

- Apeksha Shetty
- Yash Shah
- Roshni Suhanda

Hello, Our data set about breast cancer. Our Problem Statement is as follows :

Breast cancer starts when cells in the breast begin to grow out of control. These cells usually form a tumor that can often be seen on an x-ray or felt as a lump. The tumor is malignant (cancer) if the cells can grow into surrounding tissues or spread to distant areas of the body.

The main aim of our project is to analyse the breast cancer data so as to provide preventive measures that can be taken by patients before the condition becomes too serious.
We also wish to analyse at which stage is the cancer by analysing the causing factors.

For the analysis we will use KNN (K nearest neighbour) method.
We use this method, as it is most preferred for classification for analysis for breast cancer and gives maximum accuracy.

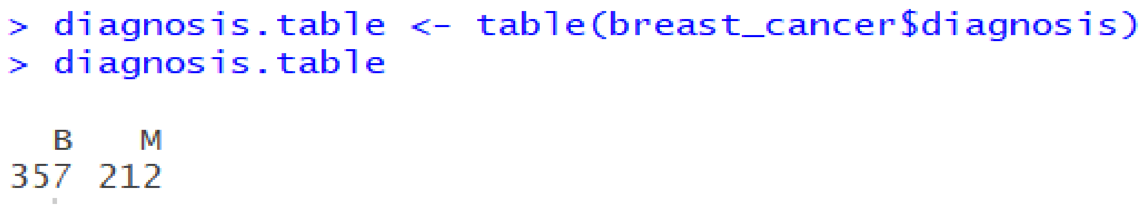
Questions that can be asked :

what can be preventive measures for breast cancer ?
what are the common factors that cause breast cancer ?
Which is more common ? benign or malignant ?
Classify stages of breast cancer.

Analysis:

Following is the Analysis that we have tried to make with respect to Week 3 lecture and general plots. We have tried to explore our data and understand basic relationships.

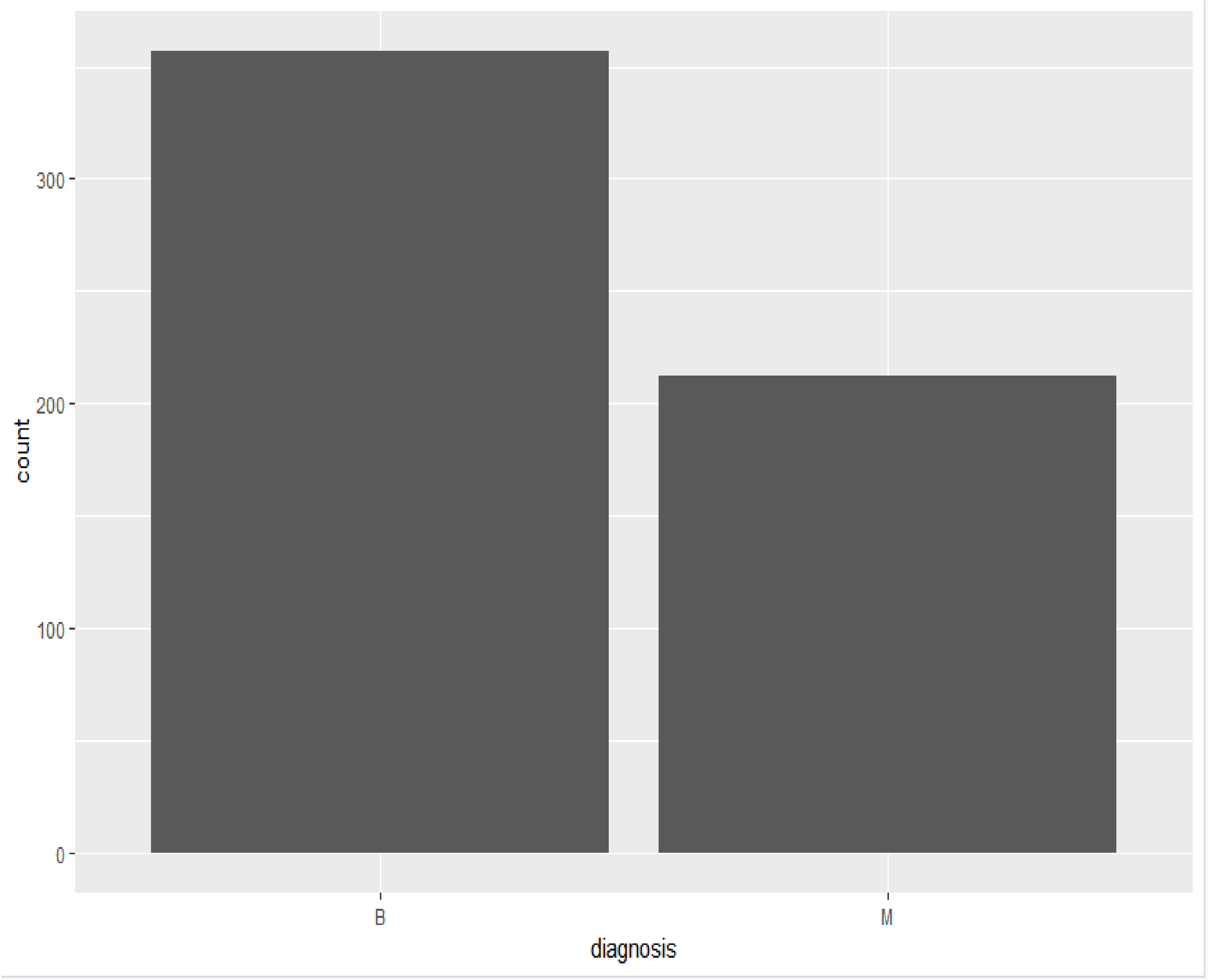
1. Bar Plot



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

  B   M 
357 212 
```

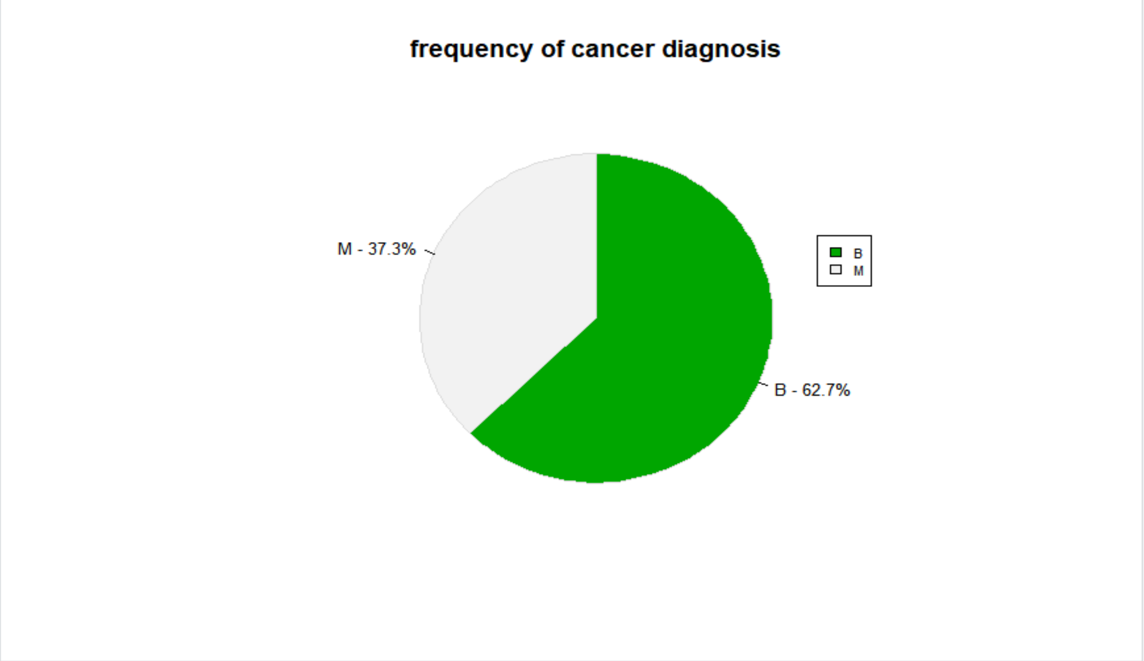
The following plot is a BAR PLOT. We are displaying the count of Benign & Malignant Cancer Affected Patients. As we can see that the count for Benign is greater than that of Malignant. With this we can understand that Benign is more common.



2. Pie Chart

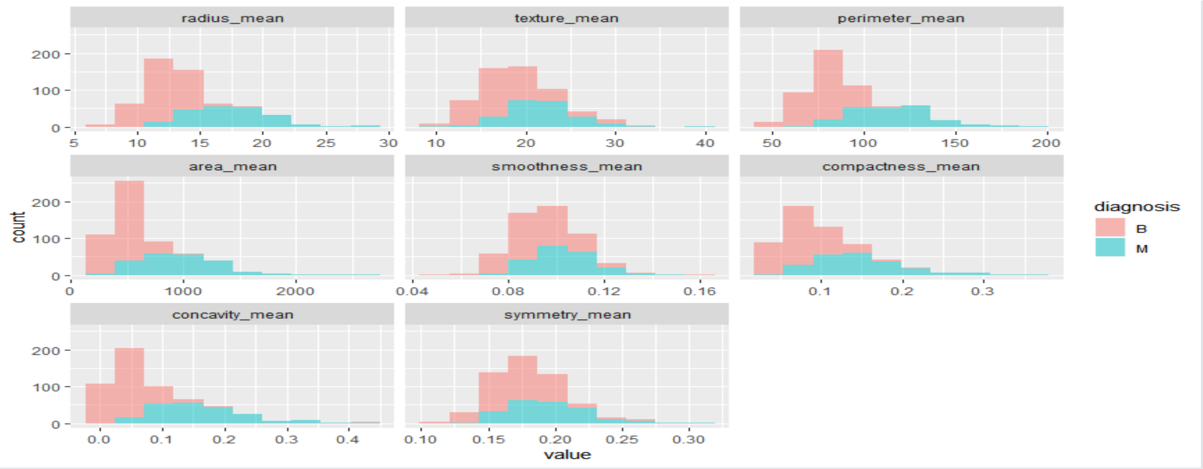
In the following pie chart we have displayed the same frequency in terms on % of both the categories. Using the following code we assigned lables to the pie chart to make it more descriptive.

```
pielabels <- sprintf("%s - %3.1f%s", diagnosis.prop.df[,1], diagnosis.prop.table, "%")
```



3. Histogram

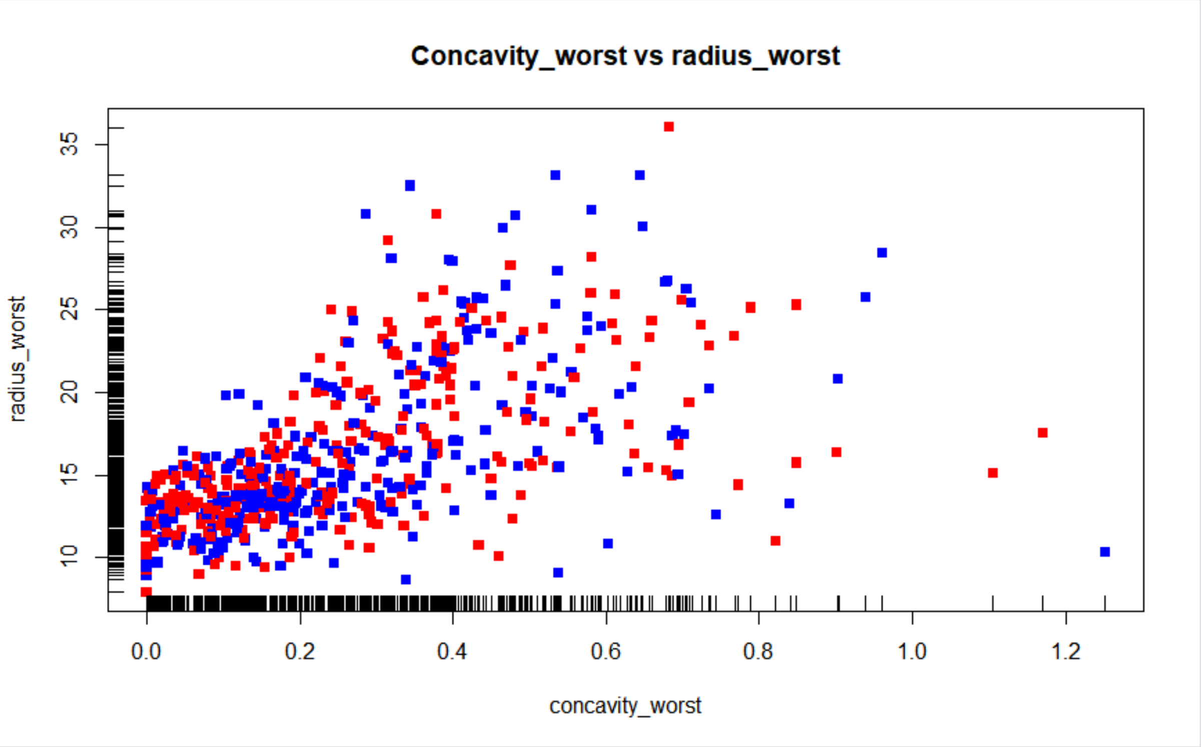
In this plot we are comparing the mean values of the 9 columns.



4. Scatter Plot

In this plot the blue dots represent the concavity_worst and red represents the radius_worst. It represents the concentration of most occurring values and can be used to plot outliers as well as develop a relationship between them.

From this we can conclude that more the radius and more of concavity may interpret a worst case scenario of breast cancer.

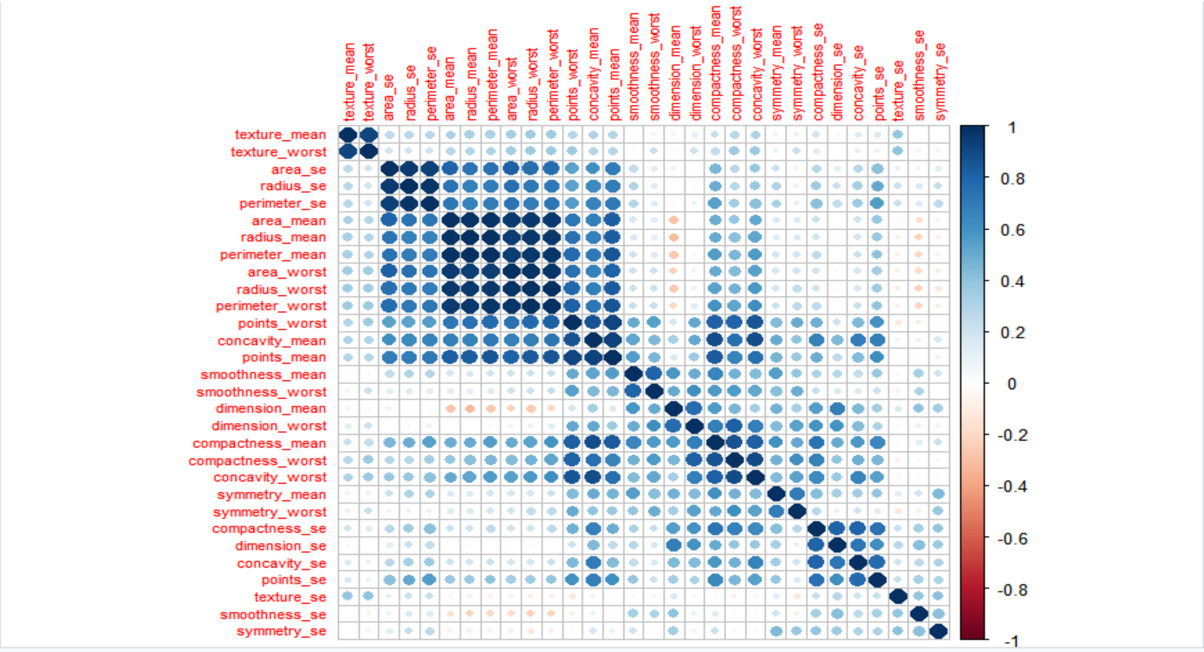


5. Correlation Matrix

Corrplot is a package that we use to plot correlation matrix. We use this to show how 2 or more variables are related to each other. Correlation establishes a relationship or connection between two or more measures statistically and tells us that there is interdependence of variable quantities.

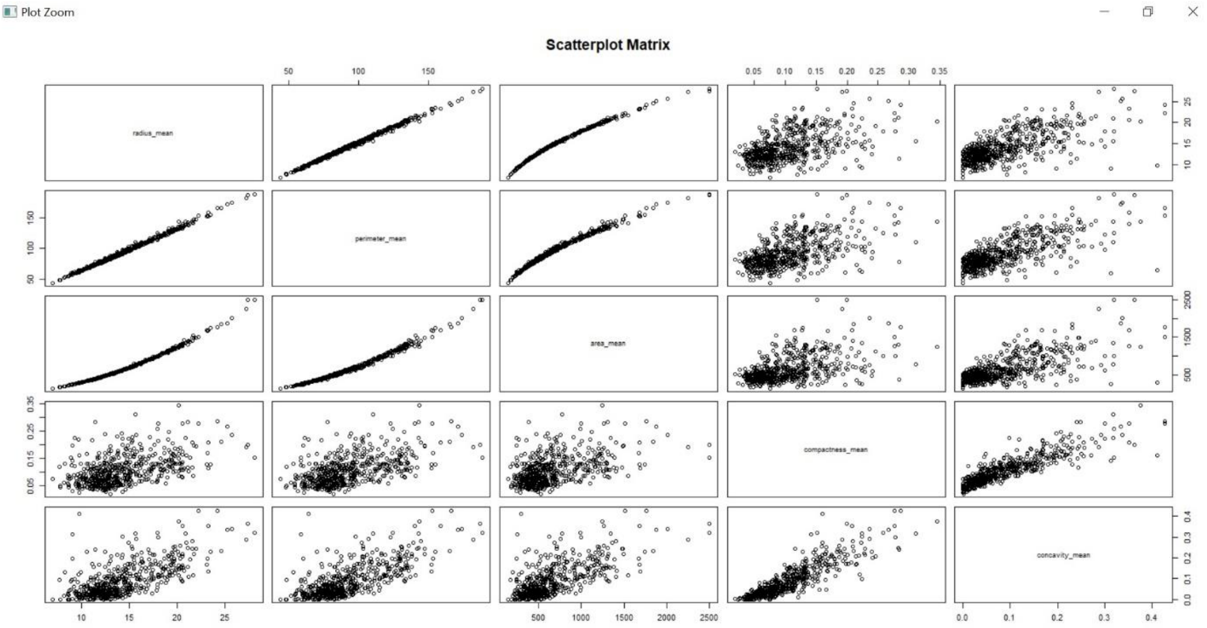
The color key on right side gives you the values of the correlation.

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



6. Scatterplot Matrix

A scatter plot matrix is table of scatter plots. Each plot is small so that many plots can be fit on a page. When you need to look at several plots.


Plot Zoom
Scatterplot Matrix
radius_mean
perimeter_mean
area_mean
compactness_mean
concavity_mean