Exploring Breast Cancer Dataset with Statistical Methods and Identify Benign or Malignant with the Logistic Regression Model

Tianchuan Gao - Li Sun - Jaya Sruthi Koppada

December 2023

1 Introduction

Breast cancer, a complex and multifaceted disease, manifests through the uncontrolled growth of cells in the breast tissue, leading to significant health challenges worldwide. According to Mayo Clinic, "After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States." [1] Statistics play a crucial role in understanding how this disease varies across different populations, including distinctions between benign and malignant tumors. Therefore, we applied statistical methods such as logistic regression to a cancer dataset to uncover the characteristics of breast cancer, as described in the Breast Cancer Wisconsin (Diagnostic) dataset [link].

2 Data

We chose the breast cancer dataset from the UC Irvine Machine Learning Repository. This dataset originates from the 'Nuclear Feature Extraction For Breast Tumor Diagnosis' project, published in Electronic Imaging. It involves analyzing fine needle aspirate (FNA) samples of breast tumors through an interactive process that includes digitizing slide images to precisely define cell nuclei boundaries. The system computes ten nuclear features from these boundaries, such as size, shape, and texture, which are crucial for distinguishing between benign and malignant tumors.[2]

The 'BreastCancer' dataset comprises 569 instances (rows), including 357 benign and 212 malignant cases, and 32 columns. These columns include an ID column, a diagnosis column indicating whether a case is benign or malignant, and 10 real-valued feature columns. Each feature is represented in three ways: mean (_Mean), standard error (_SE), and 'worst' or the largest mean of the three largest values (_Worst), resulting in 30 feature columns in total. The column names are constructed by concatenating the feature name with these three measurement types. The term 'real-valued features' refers to those represented by continuous numerical values, as opposed to categorical or discrete values. Additionally, the dataset has no missing values. These features are listed in the appendix.[2]

3 Methods of Data Exploration

3.1 Univariate Analysis - Sample Mean

To explore the distribution of individual features, density plots were generated for the means of each attribute, identified by the column names ending with '_Mean'. This analysis is instrumental in evaluating if the data distributions align with a normal distribution. The outcomes imply a deviation from normality (refer to Figure 3).

Subsequently, normal probability plots were utilized to visually compare the feature means against a theoretical normal distribution (see Figure 4).

The Kolmogorov-Smirnov (K-S) test was also conducted, contrasting the cumulative distribution function of the feature means with a normal distribution. The K-S test was selected for its non-parametric attribute, which does not presuppose the sample's normal distribution (see Figure 5).

While normality tests can shed light on the data characteristics, their importance may vary, particularly in large samples where the Central Limit Theorem (CLT) is applicable. Given that the Breast Cancer Wisconsin (Diagnostic) dataset contains 569 samples, it is sizeable enough for the CLT to ensure that the sample means will likely be normally distributed. Moreover, the sensitivity of the K-S test to large sample sizes could accentuate

	Radius_Mean	Texture_Mean	Perimeter_Mean	Area_Mean	Smoothness_Mean	Compactness_Mean	Concavity_Mean	ConcavePoints_Mean	Symmetry_Mean	FractalDimension_Mean
Malignant	3.203971	3.779470	21.85465	367.9380	0.01260824	0.05398750	0.07501933	0.03437391	0.02763809	0.007573315
Benign	1.780512	3.995125	11.80744	134.2871	0.01344608	0.03374995	0.04344215	0.01590878	0.02480676	0.006747343

Figure 1: Standard deviation of each feature mean for Malignant group and Benign group

negligible deviations from normality that are not substantively significant. Consequently, statistical tests like the t-Test and ANOVA generally exhibit robustness to violations of normality, more so with large or equal-sized samples. Hence, proceeding with a Two-sample Independent t-Test for comparing the benign and malignant groups within this dataset is justified.

3.2 Bivariate Analysis

To explore the relationship between the continuous variables that contain "-mean" in the name, we generated scatter plots by using "Diagnosis" as a color code for the data points, which can reveal the degree of separation between the benign and malignant tumors across different attributes. We observed that there is no negative correlation in the selected variables and there is most of the positive correlations are between the 'Radius_Mean' vs other variables and 'Area_Mean' and 'Perimeter_Mean'. By observing the distribution and relationship of these continuous variables, we can also highlight outliers and the potential correlation(see Figure 7).

3.3 Multivariate Analysis

To understand correlation better we employed the multivariant analysis by plotting heatmap[3].There is a notable positive correlation among the 'Worst' set of features, suggesting that the most extreme values of these features tend to increase together, which could be important for distinguishing between benign and malignant tumors. Features classified as 'Mean' and 'Worst' are also positively correlated, indicating that higher average measurements are often associated with more extreme values. However, the 'Standard Error (SE)' measurements show weaker correlations with 'Mean' and 'Worst,' implying they may represent different aspects of the tumor's characteristics. The strongest correlations are generally seen at the edges of the heatmap, highlighting potential areas of interest for further analysis(See Figure 8).

4 Statistical Testing

Based on the *Diagnosis* column, the data were divided into two subsets. The first subset consists of 212 rows labeled as *Malignant*, and the second comprises 357 rows labeled as *Benign*. Initially, a boxplot was constructed to compare the mean values of each feature in the two subsets (see Fig **??**). Subsequently, a Two-sample Independent t-Test was performed to determine whether the differences in the mean values of features between the subsets were statistically significant.

The null and alternative hypothesis for each _Mean feature are:

- H0: _Mean in Malignant subset = _Mean in Benign subset
- H1: _Mean in Malignant subset $\neq$ _Mean in Benign subset

Check for Variance Homogeneity: The classical two-sample t-test assumes equal variances between the two groups. This assumption was examined by comparing the standard deviations of the two groups. As depicted in Fig 1, significant differences in the standard deviations of certain features (e.g., *Area_Mean*) were observed between the two groups. This indicates a potential violation of the equal variance assumption for some features. Consequently, Welchś t-test, which does not assume equal variances, was employed instead of the traditional Student's t-test.

The t-statistic is

$$t = \frac{(\bar{x}_B - \bar{x}_M) - (\mu_B - \mu_M)}{\sqrt{\frac{{s_B}^2}{n_B} + \frac{{s_M}^2}{n_M}}}$$

In which, μ_B and μ_M are population mean, we assume $\mu_B - \mu_M = 0$ for the null hypothesis.

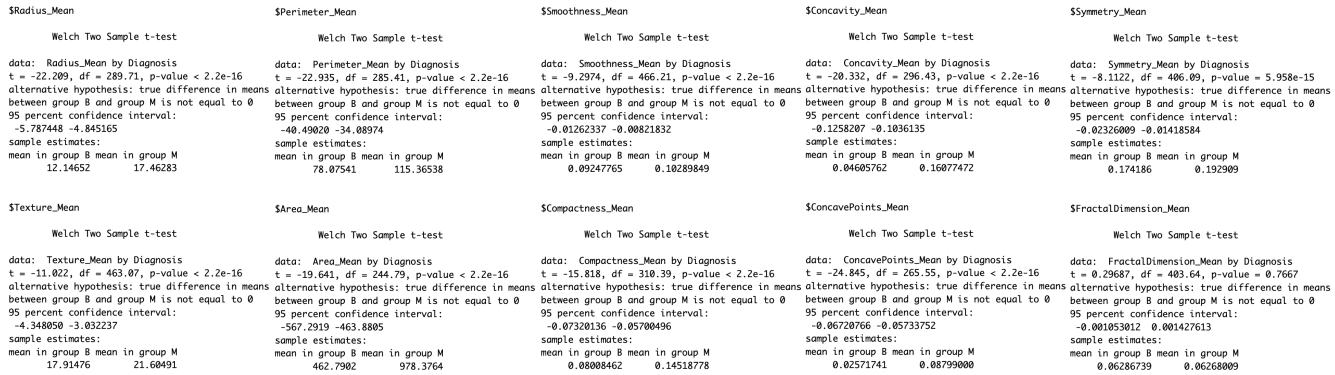

Figure 2: Two-sample Independent Welchś t-Test

Degrees of freedom The df in the Welch's t-test can vary between different features due to different variances between the two groups (malignant and benign). The formula for calculating the degrees of freedom in Welchś t-test is:

$$df = \frac{\left(\frac{s_B^2}{n_B} + \frac{s_M^2}{n_M}\right)^2}{\frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B - 1} + \frac{\left(\frac{s_M^2}{n_M}\right)^2}{n_M - 1}}$$

Where:

- s_B^2 and s_M^2 are the variances of the two samples.
- n_B and n_M are the sample sizes of the two groups.

Testing result (Fig 2) indicates that for most of the _Mean features, there is a statistically significant difference in means between the benign (B) and malignant (M) groups, as the p-values are less than 0.05. However, for FractalDimension_Mean, the p-value is not less than 0.05, suggesting there is not a statistically significant difference in means for this particular feature.

5 Logistic Regression

In order to find strong predictors for the cancer being malignant, we can perform logistic regression. At first, we look at the summary statistics for these 10 variables

```
> summary(mean_data[3:12])
  Radius_Mean      Texture_Mean    Perimeter_Mean     Area_Mean      Smoothness_Mean
 Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5   Min.   :0.05263
 1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637
 Median :13.370   Median :18.84   Median : 86.24   Median : 551.1   Median :0.09587
 Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9   Mean   :0.09636
 3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530
 Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0   Max.   :0.16340
 Compactness_Mean  Concavity_Mean    ConcavePoints_Mean Symmetry_Mean    FractalDimension_Mean
 Min.   :0.01938   Min.   :0.00000   Min.   :0.00000    Min.   :0.1060   Min.   :0.04996
 1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031    1st Qu.:0.1619   1st Qu.:0.05770
 Median :0.09263   Median :0.06154   Median :0.03350    Median :0.1792   Median :0.06154
 Mean   :0.10434   Mean   :0.08880   Mean   :0.04892    Mean   :0.1812   Mean   :0.06280
 3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400    3rd Qu.:0.1957   3rd Qu.:0.06612
 Max.   :0.34540   Max.   :0.42680   Max.   :0.20120    Max.   :0.3040   Max.   :0.09744
```

We can see that these 10 variables' scale are significantly different, thus we normalize these columns and perform logistic regression. We first try full model with all 10 variables:

```
Call:
glm(formula = Diagnosis ~ Radius_Mean + Texture_Mean + Perimeter_Mean +
    Area_Mean + Smoothness_Mean + Compactness_Mean + Concavity_Mean +
    ConcavePoints_Mean + Symmetry_Mean + FractalDimension_Mean,
    family = binomial, data = scaled_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95590  -0.14839  -0.03943   0.00429   2.91690

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.48702    0.56432   0.863   0.3881
Radius_Mean             -7.22185   13.09494  -0.551   0.5813
Texture_Mean             1.65476    0.27758   5.961  2.5e-09 ***
Perimeter_Mean          -1.73763   12.27499  -0.142   0.8874
Area_Mean               14.00485    5.89090   2.377   0.0174 *
Smoothness_Mean          1.07495    0.44942   2.392   0.0168 *
Compactness_Mean        -0.07723    1.07434  -0.072   0.9427
Concavity_Mean           0.67512    0.64733   1.043   0.2970
ConcavePoints_Mean       2.59287    1.10701   2.342   0.0192 *
Symmetry_Mean            0.44626    0.29143   1.531   0.1257
FractalDimension_Mean -0.48248    0.60406  -0.799   0.4244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 146.13  on 558  degrees of freedom
AIC: 168.13

Number of Fisher Scoring iterations: 9
```

Logistic regression assumes little or no multicollinearity among the independent variables. We check this by calculating the Variance Inflation Factor (VIF).

```
      Radius_Mean       Texture_Mean     Perimeter_Mean          Area_Mean    Smoothness_Mean
        899.521292           1.806441         698.983071         129.559492           4.372939
Compactness_Mean     Concavity_Mean  ConcavePoints_Mean       Symmetry_Mean FractalDimension_Mean
        15.280847           5.259524           5.856371           1.839511           9.787700
```

Therefore, We get a reduced model and perform logistic regression again:

```
Call:
glm(formula = Diagnosis ~ Texture_Mean + Smoothness_Mean + ConcavePoints_Mean,
    family = binomial, data = scaled_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.80842  -0.21536  -0.08281   0.07419   2.53893

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.6961     0.1933  -3.602 0.000316 ***
Texture_Mean        1.2994     0.2194   5.922 3.19e-09 ***
Smoothness_Mean    -0.5473     0.2552  -2.145 0.031981 *
ConcavePoints_Mean  5.0888     0.5185   9.814  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 198.46  on 565  degrees of freedom
AIC: 206.46

Number of Fisher Scoring iterations: 7
```

The null deviance is 751.44 on 568 degrees of freedom. After adding predictors, the residual deviance is 198.46 on 565 degrees of freedom. A large drop in deviance indicates a good fit, and in this case, it suggests that the model with predictors fits the data significantly better than the null model. After taking exponential of these coefficients, we can interpret this model:

```
exp(coef(fit))
     (Intercept)       Texture_Mean    Smoothness_Mean ConcavePoints_Mean
       0.4985224          3.6670104          0.5784884        162.1898737
```

- **Intercept**: the odds of being malignant is 0.4985224 when all predictors are at their mean values (which is 0 after scaling). Since it's less than 1, it indicates that the baseline odds (with all predictors at their mean) favor the outcome being benign.
- **Texture Mean**: The odds ratio is approximately 3.6670. This suggests that for each standard deviation increase in standard deviation of gray-scale value, the odds of being malignant are about 3.667 times higher, assuming other variables are held constant. It indicates a strong positive association with the outcome.
- **Smoothness Mean**:The odds ratio is approximately 0.5785. For each standard deviation increase in variation in radius lengths, the odds of being malignant are multiplied by 0.5785. An odds ratio below 1 indicates a negative association with the outcome; in this case, higher values of variation in radius lengths are associated with a lower likelihood of being malignant.
- **ConcavePoints Mean**: The odds ratio is approximately 162.1899. This indicates a very strong positive association; for each standard deviation increase in the number of concave portions of the contour of the nucleus, the odds of being malignant are multiplied by 162.1899. This is a substantial increase and highlights the strong predictive power of the number of concave portions of the contour of the nucleus with respect to the Diagnosis.

6 Discussion

6.1 Logistic Regression

It's always good practice to further assess the model's predictive performance using a validation dataset or cross-validation and to examine other diagnostic measures and plots to check for potential issues such as influential observations or lack of fit.

6.2 Data Quality

In our analysis of breast tumor diagnosis using features derived from FNA samples, a key consideration is the independence of data points. Each observation in our dataset represents a unique measurement; however, the study's methodology does not explicitly confirm if these measurements correspond to distinct patients or multiple samples from the same patient. This ambiguity raises questions about the independence of observations, a fundamental assumption in our statistical tests (the two-sample independent t-test). While larger sample sizes and diverse data might mitigate potential issues, the possibility of correlated data cannot be dismissed without further clarification. This highlights the need for clear data documentation in medical research, particularly in studies involving complex sample collection and feature extraction processes.

[3]

References

[1] M. Clinic. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[2] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Electronic imaging*, 1993. [Online]. Available: https://api.semanticscholar.org/CorpusID:14922543

[3] J. F. n. Hair. [Online]. Available: https://digitalcommons.kennesaw.edu/facpubs/2925/

[4] V. Julie and H. David, *Introductory Statistics for the Life and Biomedical Sciences*, 1st ed. [Online]. Available: https://www.openintro.org/book/biostat/

Appendix

Illustration of the 10 measured features

- **Radius** The mean of distances from center to points on the perimeter.
- **Texture** The standard deviation of gray-scale values.
- **Perimeter** The perimeter of the cell nucleus.
- **Area of the cell nucleus** The size of the nucleus within a cell, measured in square units. It's the space enclosed by the nuclear membrane as captured in the two-dimensional image of a FNA of a breast mass.
- **Smoothness** The variation in radial length from the center of the nucleus to its perimeter, which is also called local variation in radius lengths.
- **Compactness** A higher compactness value indicates a more irregular shape. And a circle (which is regular and smooth) has the lowest possible compactness.

$$\text{compactness} = \frac{(\text{perimeter})^2}{\text{area}} - 1$$

- **Concavity** The extent or severity of its inward curves or indentations. Higher concavity values, indicating irregular shapes, can suggest malignancy in cells.
- **Concave points** The count of concave portions of the contour of the nucleus. As with concavity, more concave points may correlate with malignancy.
- **Symmetry** symmetry of the cell nucleus is measured by comparing the two halves of the nucleus along its major axis with a perfectly symmetrical nucleus showing one half as a mirror image of the other.
- **Fractal dimension** Based on the coastline paradox, fractal dimension is a measure that quantifies the complexity of the cell nucleus shape by looking at the pattern at different scales. In pathology, cancerous cells often have more irregular and complex shapes compared to healthy cells, which might be reflected in a higher fractal dimension.

$$\text{fractal dimension} = \frac{\log(N)}{\log(1/\epsilon)}$$

Where:

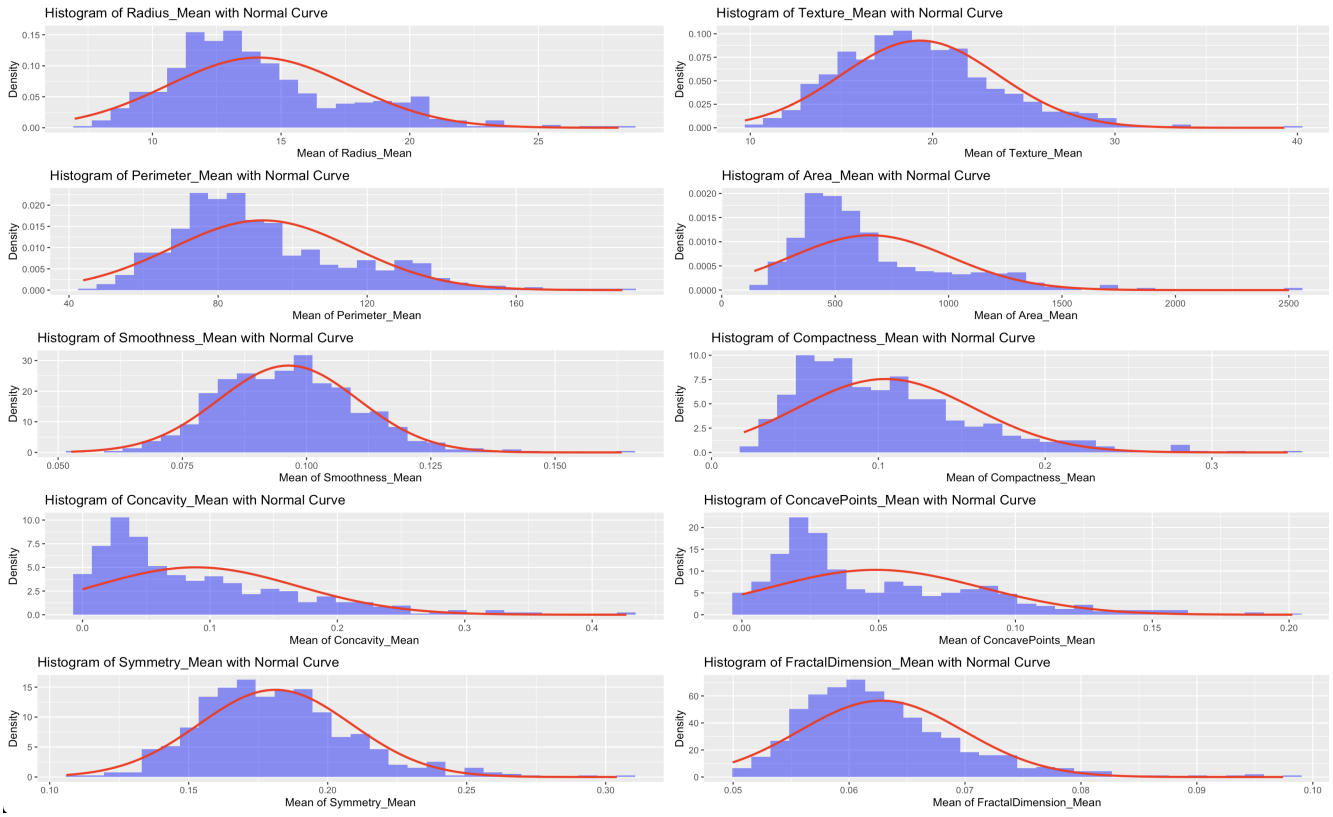

Figure 3: Histogram overlaid with a normal curve fitted using the sample mean $\bar{x}$ and standard deviation s. A closer fit indicates a more reasonable assumption of the normal model. [4]

- N is the number of units needed to cover the boundary.
- ϵ is the scale or size of the measurement unit.

Normality test

Hypothesis testing

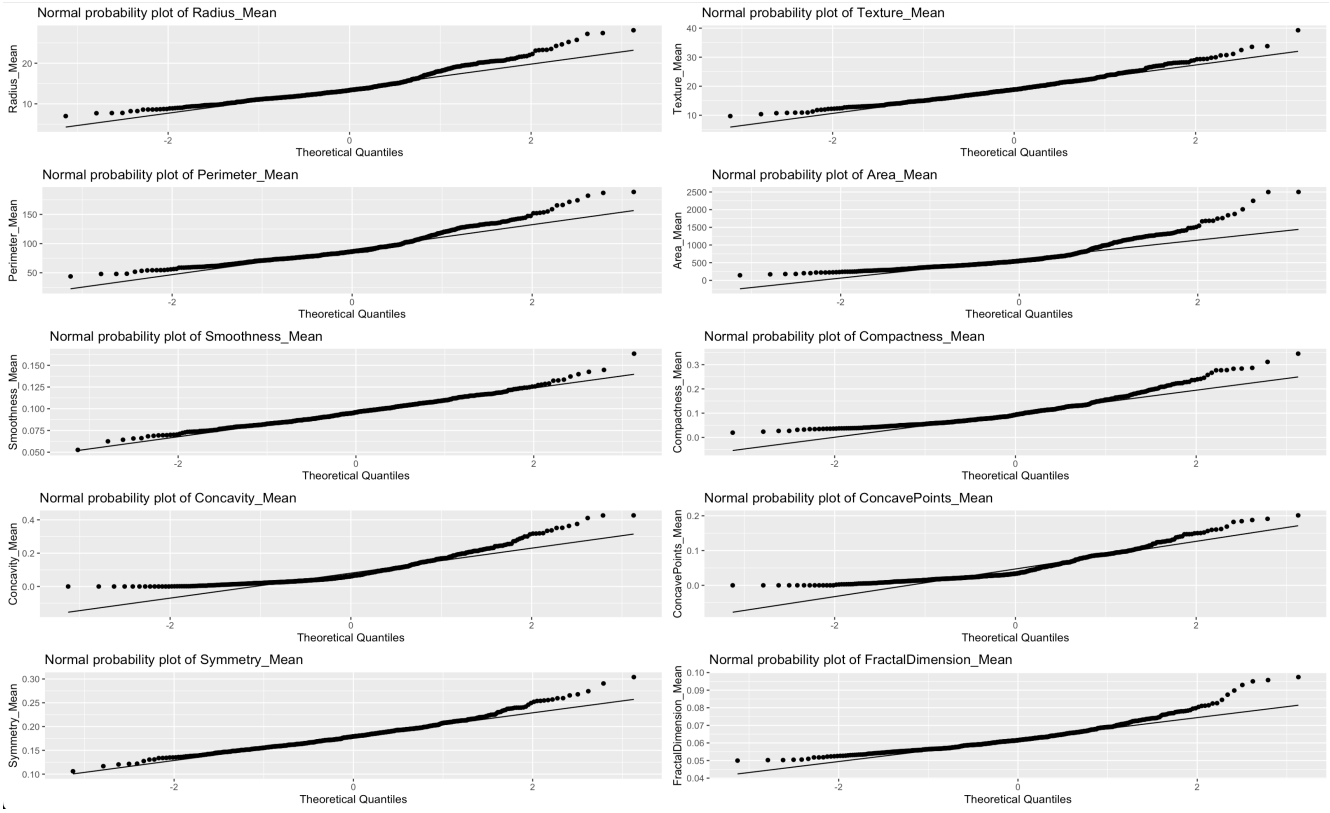

Figure 4: Normal probability plots. "Typically, the normal approximation is reasonable even if there are some small observed departures from normality in the tails"[4]

```
Column: Radius_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.11271, p-value = 1.054e-06
alternative hypothesis: two-sided


Column: Texture_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.049075, p-value = 0.129
alternative hypothesis: two-sided


Column: Perimeter_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.12068, p-value = 1.267e-07
alternative hypothesis: two-sided


Column: Area_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.156, p-value = 1.88e-12
alternative hypothesis: two-sided


Column: Smoothness_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.035967, p-value = 0.4533
alternative hypothesis: two-sided


Column: Compactness_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.10111, p-value = 1.771e-05
alternative hypothesis: two-sided


Column: Concavity_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.13962, p-value = 4.648e-10
alternative hypothesis: two-sided


Column: ConcavePoints_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.1576, p-value = 1.061e-12
alternative hypothesis: two-sided


Column: Symmetry_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.051685, p-value = 0.09566
alternative hypothesis: two-sided


Column: FractalDimension_Mean

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data[[column_name]]
D = 0.094753, p-value = 7.308e-05
alternative hypothesis: two-sided
```

Figure 5: Kolmogorov-Smirnov test. D-statistic is the maximum difference between the cumulative distribution function of the sample data and the cumulative distribution function of the comparison normal distribution. The smaller the D-statistic, the closer the fit to a normal distribution.

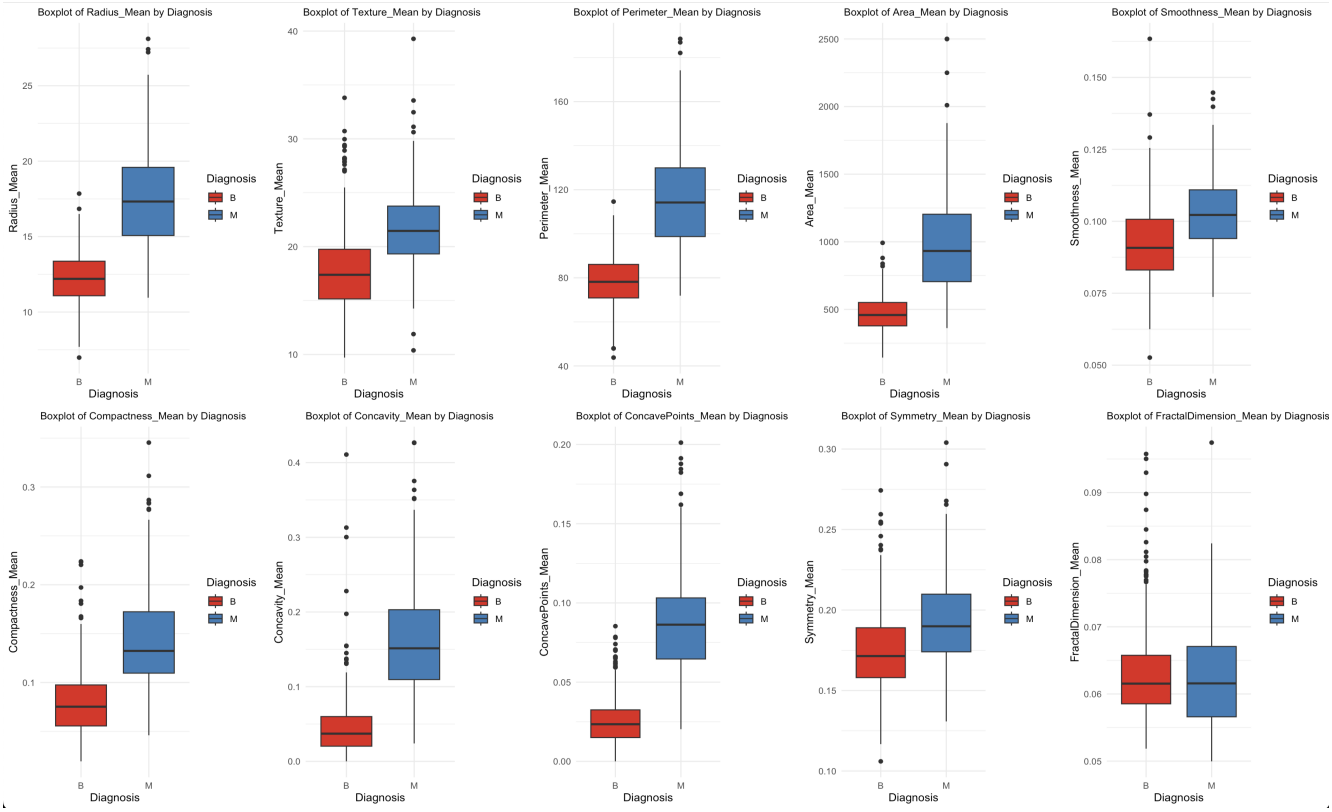

Figure 6: Boxplot to compare the feature in Malignant diagnosis subset and Benign diagnosis subset.

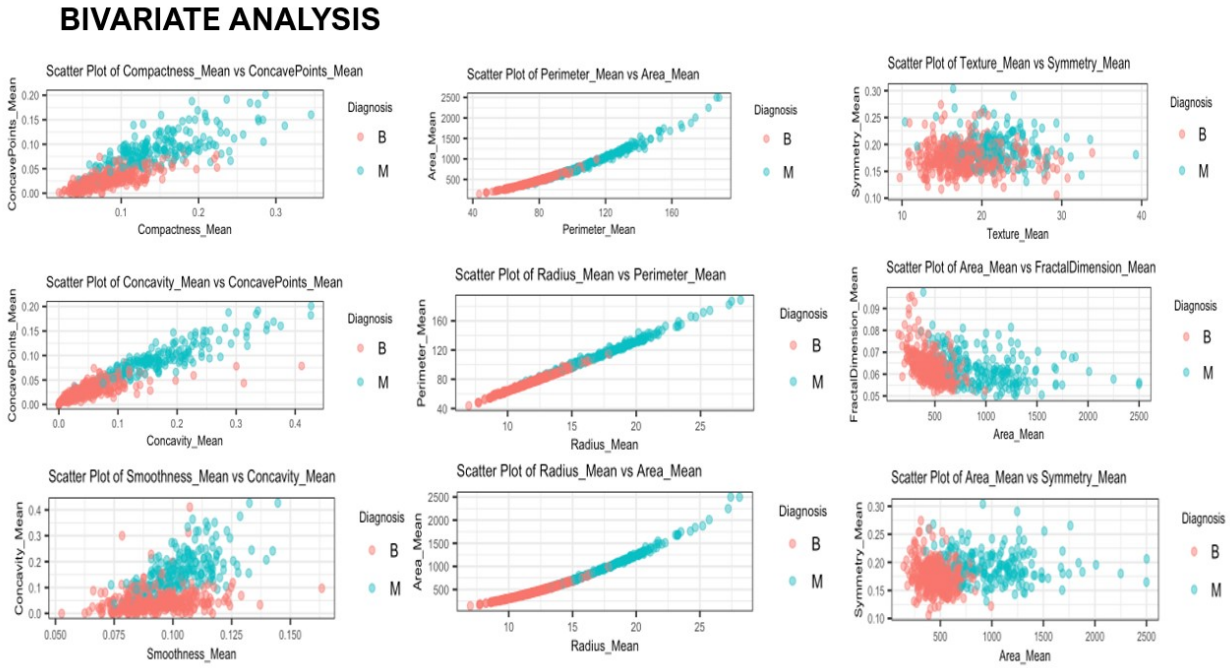

Figure 7: Bivariate Analysis-Scatter plots

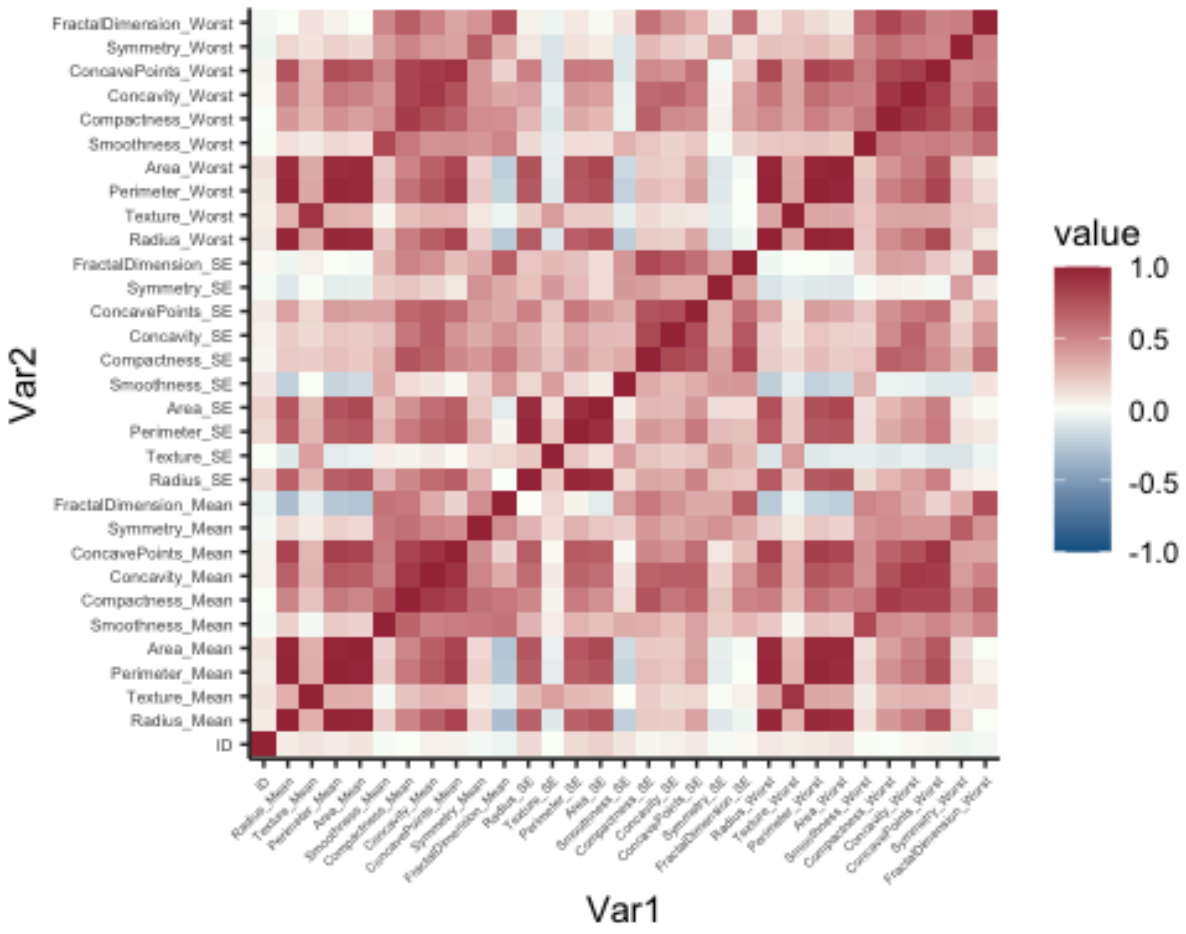

Figure 8: Multivariate Analysis-Heatmap