Predicting Breast Cancer with Simple Logistic Regression

Final Semester Project

Name: Sayan Ghosh **Registration**

Number: 012-1111-0763-20

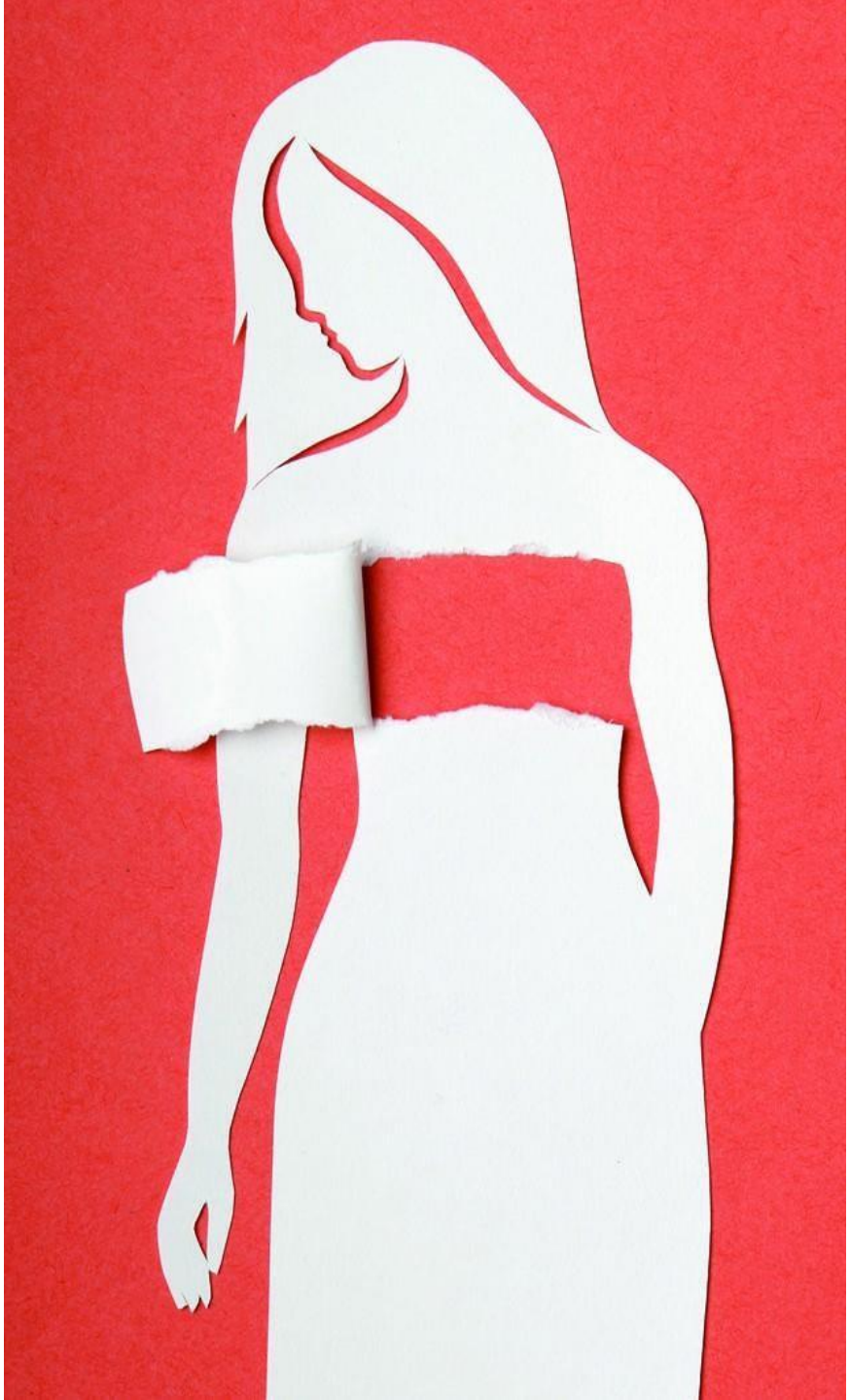
Roll Number: 203012-21-0083

Batch: 2020-2023

Semester: VI

Paper: DSE-B2

INDEX



ABSTRACT :

Breast cancer is worldwide & and a very serious problem. In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer. There are more lost disability-adjusted life years (DALYs) by women to breast cancer globally than any other type of cancer. Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life.

Breast cancer mortality changed little from the 1930s through to the 1970s. Improvements in survival began in the 1980s in countries with early detection programmes combined with different modes of treatment to eradicate invasive disease. Survival of breast cancer for at least 5 years after diagnosis ranges from more than 90% in high-income countries, to 66% in India and 40% in South Africa. Early detection and treatment has proven successful in high-income countries and should be applied in countries with limited resources where some of the standard tools are available.

INTRODUCTION :

Breast cancer arises in the living cells (epithelium) of the ducts (85%) or lobules (15%) in the glandular tissue of the breast. Initially, the cancerous growth is confined to the duct or lobule ("in situ") where it generally causes no symptoms and has minimal potential for spread (metastasis).

Over time, these in situ (stage 0) cancers may progress and invade the surrounding breast tissue (invasive breast cancer) then spread to the nearby lymph nodes (regional metastasis) or to other organs in the body (distant metastasis). If a woman dies from breast cancer, it is because of widespread metastasis.

Breast cancer treatment can be highly effective, especially when the disease is identified early. Treatment of breast cancer often consists of a combination of surgical removal, radiation therapy and medication (hormonal therapy, chemotherapy and/or targeted biological therapy) to treat the microscopic cancer that has spread from the breast tumour through the blood. Such treatment, which can prevent cancer growth and spread, thereby saves lives.

Who is at risk ?

Breast cancer is not a transmissible or infectious disease. Unlike some cancers that have infection-related causes, such as human papillomavirus (HPV) infection and cervical cancer, there are no known viral or bacterial infections linked to the development of breast cancer.

Approximately half of breast cancers develop in women who have no identifiable breast cancer risk factor other than gender (female) and age (over 40 years). Certain factors increase the risk of breast cancer including increasing age, obesity, harmful use of alcohol, family history of breast

cancer, history of radiation exposure, reproductive history (such as age that menstrual periods began and age at first pregnancy), tobacco use and postmenopausal hormone therapy.

Behavioural choices and related interventions that reduce the risk of breast cancer include:

- prolonged breastfeeding;
- regular physical activity;
- weight control;
- avoidance of harmful use of alcohol;
- avoidance of exposure to tobacco smoke;
- avoidance of prolonged use of hormones; and
- avoidance of excessive radiation exposure.

Unfortunately, even if all of the potentially modifiable risk factors could be controlled, this would only reduce the risk of developing breast cancer by at most 30%.

Female gender is the strongest breast cancer risk factor. Approximately 0.5-1% of breast cancers occur in men. The treatment of breast cancer in men follows the same principles of management as for women.

Family history of breast cancer increases the risk of breast cancer, but the majority of women diagnosed with breast cancer do not have a known family history of the disease. Lack of a known family history does not necessarily mean that a woman is at reduced risk.

Certain inherited “high penetrance” gene mutations greatly increase breast cancer risk, the most dominant being mutations in the genes BRCA1, BRCA2 and PALB-2. Women found to have mutations in these major genes could consider risk reduction strategies such as surgical removal of both breasts. Consideration of such a highly invasive approach only concerns a very limited number of women, should be carefully evaluated considering all alternatives and should not be rushed.

Signs and symptoms

Breast cancer most commonly presents as a painless lump or thickening in the breast. It is important that women finding an abnormal lump in the breast consult a health practitioner without a delay of more than 1-2 months even when there is no pain associated with it. Seeking medical attention at the first sign of a potential symptom allows for more successful treatment.

Generally, symptoms of breast cancer include:

- a breast lump or thickening;
- alteration in size, shape or appearance of a breast;
- dimpling, redness, pitting or other alteration in the skin;
- change in nipple appearance or alteration in the skin surrounding the nipple (areola); and/or

- abnormal nipple discharge.

There are many reasons for lumps to develop in the breast, most of which are not cancer. As many as 90% of breast masses are not cancerous. Non-cancerous breast abnormalities include benign masses like fibro adenomas and cysts as well as infections.

Breast cancer can present in a wide variety of ways, which is why a complete medical examination is important. Women with persistent abnormalities (generally lasting more than one month) should undergo tests including imaging of the breast and in some cases tissue sampling (biopsy) to determine if a mass is malignant (cancerous) or benign.

Advanced cancers can erode through the skin to cause open sores (ulceration) but are not necessarily painful. Women with breast wounds that do not heal should have a biopsy performed.

Breast cancers may spread to other areas of the body and trigger other symptoms. Often, the most common first detectable site of spread is to the lymph nodes under the arm although it is possible to have cancer-bearing lymph nodes that cannot be felt.

Over time, cancerous cells may spread to other organs including the lungs, liver, brain and bones. Once they reach these sites, new cancer-related symptoms such as bone pain or headaches may appear.

What is Breast Cancer Prediction ?

Combining multiple risk factors in modelling for breast cancer prediction could help the early diagnosis of the disease with necessary care plans .And we discussed before how important to find out early detection and treatment to become fully cure . If we can find out the problem at the root the chance of surviving of the patients increases . We use various statistical tools to formulate or we can say to create a valid Breast Cancer Prediction model by using logistic regression .

Source of the Data :

The dataset is collected form Kaggle , Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners.

The dataset is D:\chrome downloads\breast cancer.csv

METHODS

1. Data Extraction :

Data extraction is where data is considered and moved through to fetch relevant information from data sources (such as database) in a definite design. Further data processing is completed, which contains inserting metadata and other data integration; another procedure in the data workflow & Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. So here we are Exploring Data set and features and Data cleaning.

2. Exploratory Data Analysis :

I used this technique to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques I am considering for data analysis are appropriate.

3. Analysing the model again to proceed further steps :

"*It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts,"* Sherlock Holmes's proclaims in Sir Arthur Conan Doyle's A Scandal in Bohemia.

Analyse the data is By manipulating the data using various data analysis techniques and tools, we can begin to find trends, correlations, outliers, and variations that tell a story. During this stage, you might use data mining to discover patterns within databases or data visualization software to help transform data into an easy-to-understand graphical format. We analysis the model to jump into the final steps of our process to make a valid breast cancer prediction model.

4. Creating our model :

creating our model by logistic regression method which is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. Analyze the data. By manipulating the data using various data analysis techniques and tools, you can begin to find trends,

correlations, outliers, and variations that tell a story. During this stage, you might use data mining to discover patterns within databases or data visualization software to help transform data into an easy-to-understand graphical format. Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The main advantage is to avoid confounding effects by analysing the association of all variables together. In this article, we explain the logistic regression procedure using examples to make it as simple as possible. After definition of the technique, the basic interpretation of the results is highlighted and then some special issues are discussed.

5. Fitting our model & Visualizing the model :

we fit our model where our input variable is the area_mean and our output variable is the diagnosis. And we visualize the model where we get a curve from our model and we can see that the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability . Generally Analyze the data. By manipulating the data using various data analysis techniques and tools, you can begin to find trends, correlations, outliers, and variations that tell a story. During this stage, you might use data mining to discover patterns within databases or data visualization software to help transform data into an easy-to-understand graphical format.

6. Interpretation :

Logistic regression is a powerful tool, especially in epidemiologic studies, allowing multiple explanatory variables being analysed simultaneously, meanwhile reducing the effect of confounding factors. However, researchers must pay attention to model building, avoiding just feeding software with raw data and going forward to results. Some difficult decisions on model building will depend entirely on the expertise of researcher on the field.

7. Checking accuracy of our model :

Any data science project starts with exploring the data. When we perform an analysis on a sample through exploratory data analysis and inferential statistics we get information about the sample. Now, we want to use this information to predict values for the entire population. Hypothesis testing is done to confirm our observation about the population using sample data, within the desired error level. Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude if the hypothesis about the population is true or not.

Now in this stage we are checking accuracy of the model by performing a simple hypothesis test.

8. **Real life implication of the model :**

Giving an example how this model can be used in real life scenario .

9. **Final conclusion** :

Now all of my project work is done , it's the time to end up with a conclusion **.**

1. Data Extraction

The dataset is downloaded from Kaggle and saved in the data folder. We use **read.csv()** function to read the dataset and put in **bcw_df** data frame.

```
bcw df <- read.csv("D:/chrome downloads/breast cancer.csv")
```

```
> bcw_df <- read.csv("D:/chrome downloads/breast cancer.csv")
> view(bcw_df)
> dim(bcw_df)
[1] 569  32
> view(bcw_df)
> str(bcw_df)
'data.frame':   569 obs. of  32 variables:
 $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844981 84501001 ...
 $ 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 ...
>
```

To see the number of rows and column, we used **dim()** function. The dataset has 569 rows and 33 columns.

```
dim(bcw_df)
```

```
[1] 569  32
```

2. Exploratory Data Analysis

To find out the column names and types, we used **str()** function.

```
str(bcw_df)
```

```
## 'data.frame':    569 obs. of  33 variables:
##  $ id                      : int  842302 842517 84300903 84348301 84358402  843786 844
359 84458202 844981 84501001 ...
##  $ 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 ...
```

From the result above, we know the following:

1. The first column is **id**. It is unique and unnecessary for prediction. So, it should be removed.
2. The second column is **diagnosis**. Currently the type is **char** and it should be converted to **factor**.
3. The last column is **X**. All the values are NA. So, it should be removed.

```
# remove unnecessary columns
> bcw_df$id <-NULL
> bcw_df$X <-NULL
>
> # change to factor for target variable
> bcw_df$diagnosis <- as.factor(bcw_df$diagnosis)
> View(bcw df)
```

2.1. Univariate Data Analysis

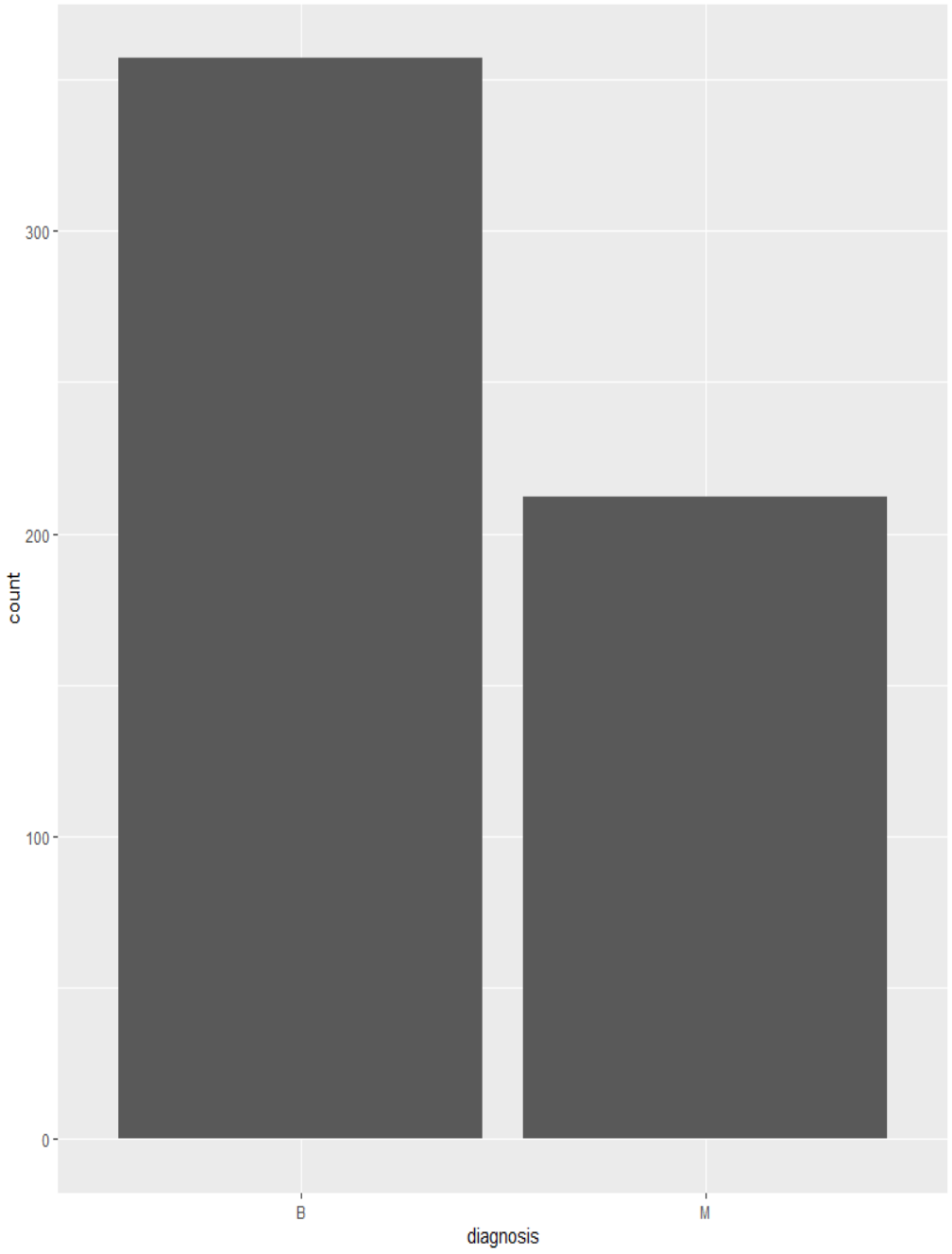
Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own.

Bar charts

Bar charts are one of the many techniques used to present data in a visual form so that the reader may readily recognize patterns or trends. Bar charts usually present categorical variables, discrete variables or continuous variables grouped in class intervals.

Analysis of a single variable. Number of benign (B) and malignant (M) in **dignosis** column.

```
library(ggplot2)
> ggplot(data=bcw_df, aes(x=diagnosis)) + geom_bar()
```



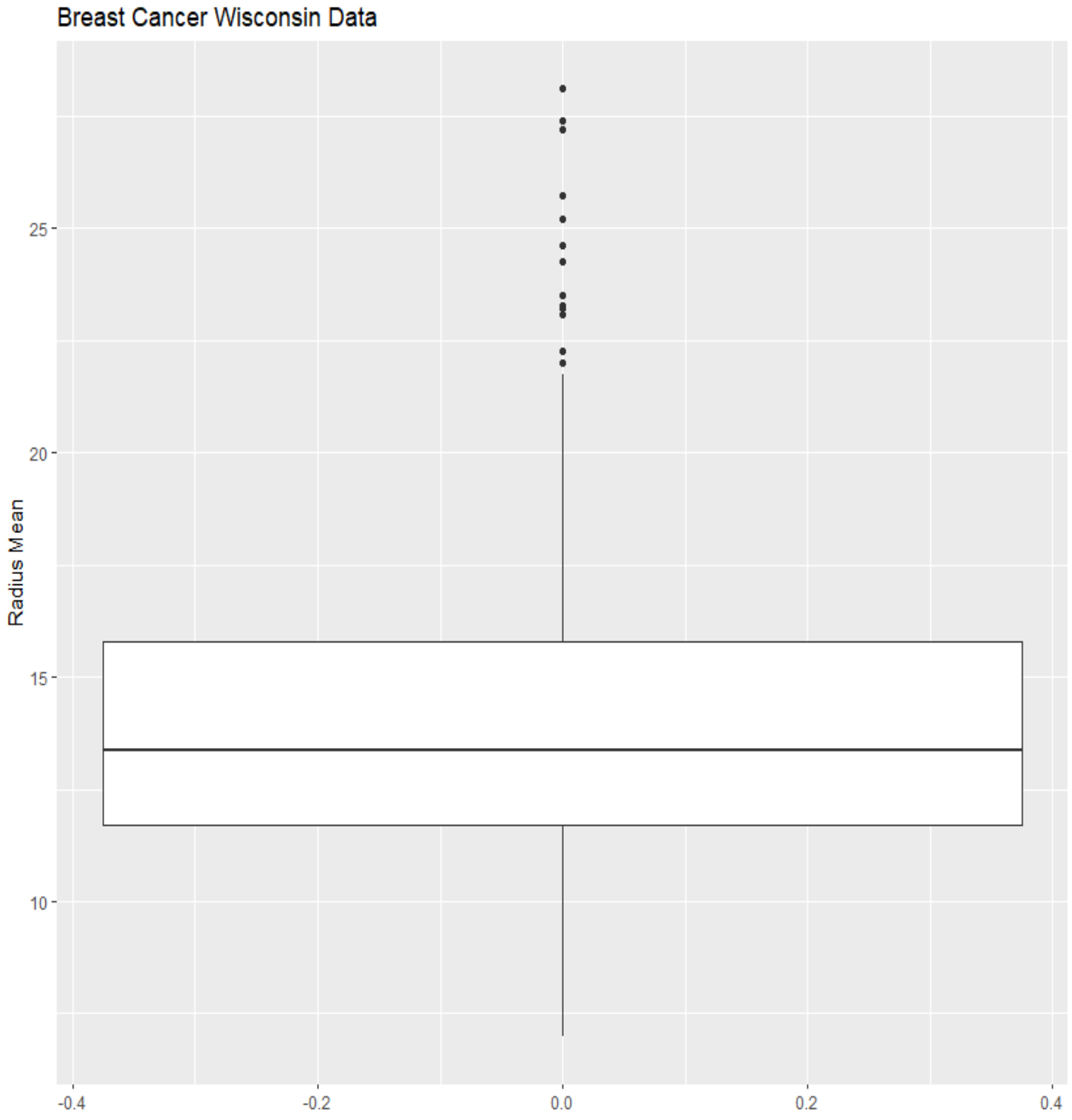
we can see count wise comparison of two diagnosis variable

Boxplot

Box and whisker plots, sometimes known as box plots, are a great chart to use when showing the distribution of data points across a selected measure. These charts display ranges within variables measured. This includes the outliers, the median, the mode, and where the majority of the data points lie in the “box”.

Distribution of **radius mean** variable in boxplot.

```
ggplot(data=bcw_df, aes(y=radius_mean)) +
     geom_boxplot() +
    labs(title="Breast Cancer Wisconsin Data", y="Radius Mean")
```



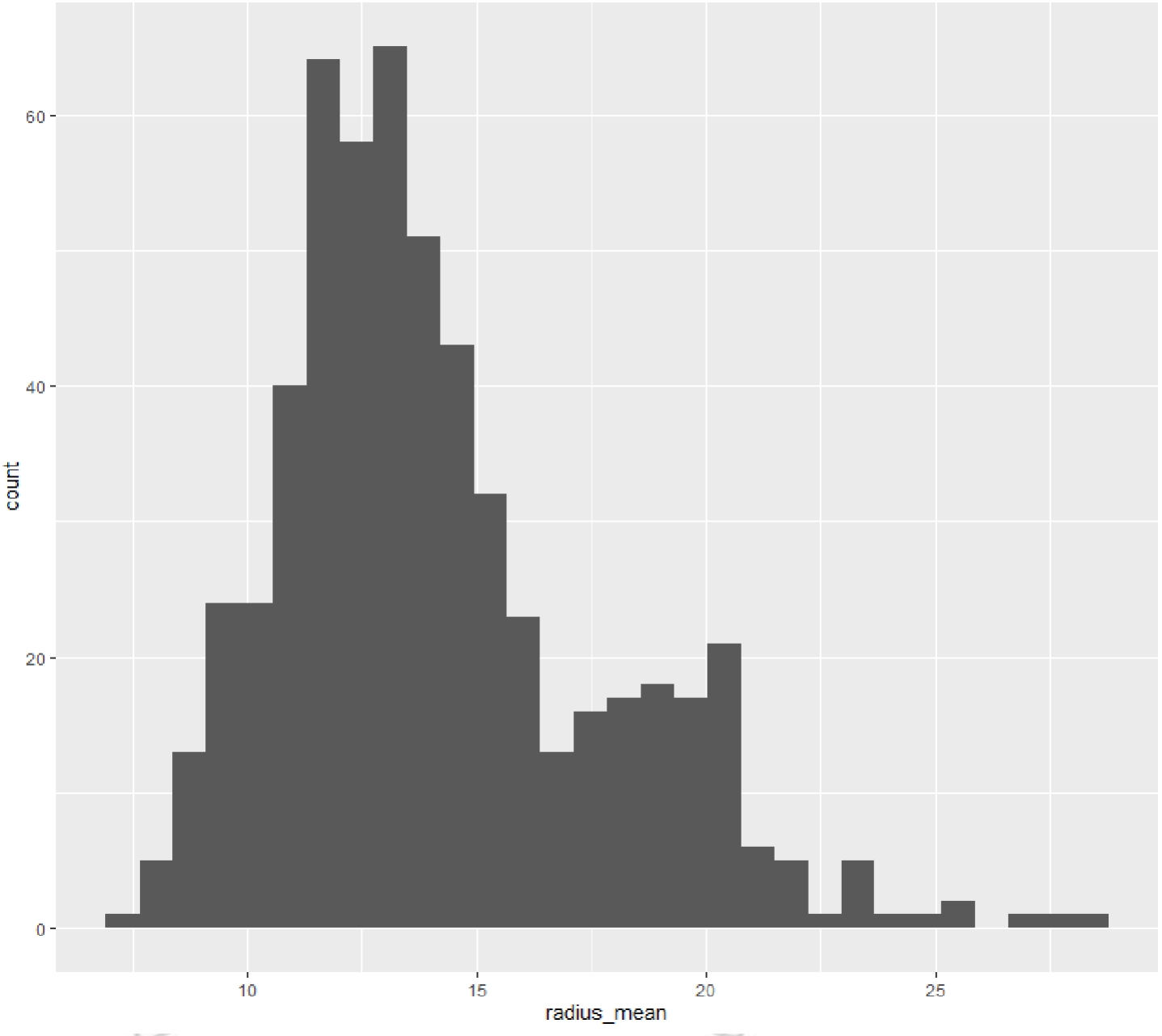
Histogram

n

appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins. The histogram is a popular graphing tool. It is used to summarize discrete or continuous data that are measured on an interval scale. It is often used to illustrate the major features of the distribution of the data in a convenient form.

Distribution of **radius mean** variable in histogram.

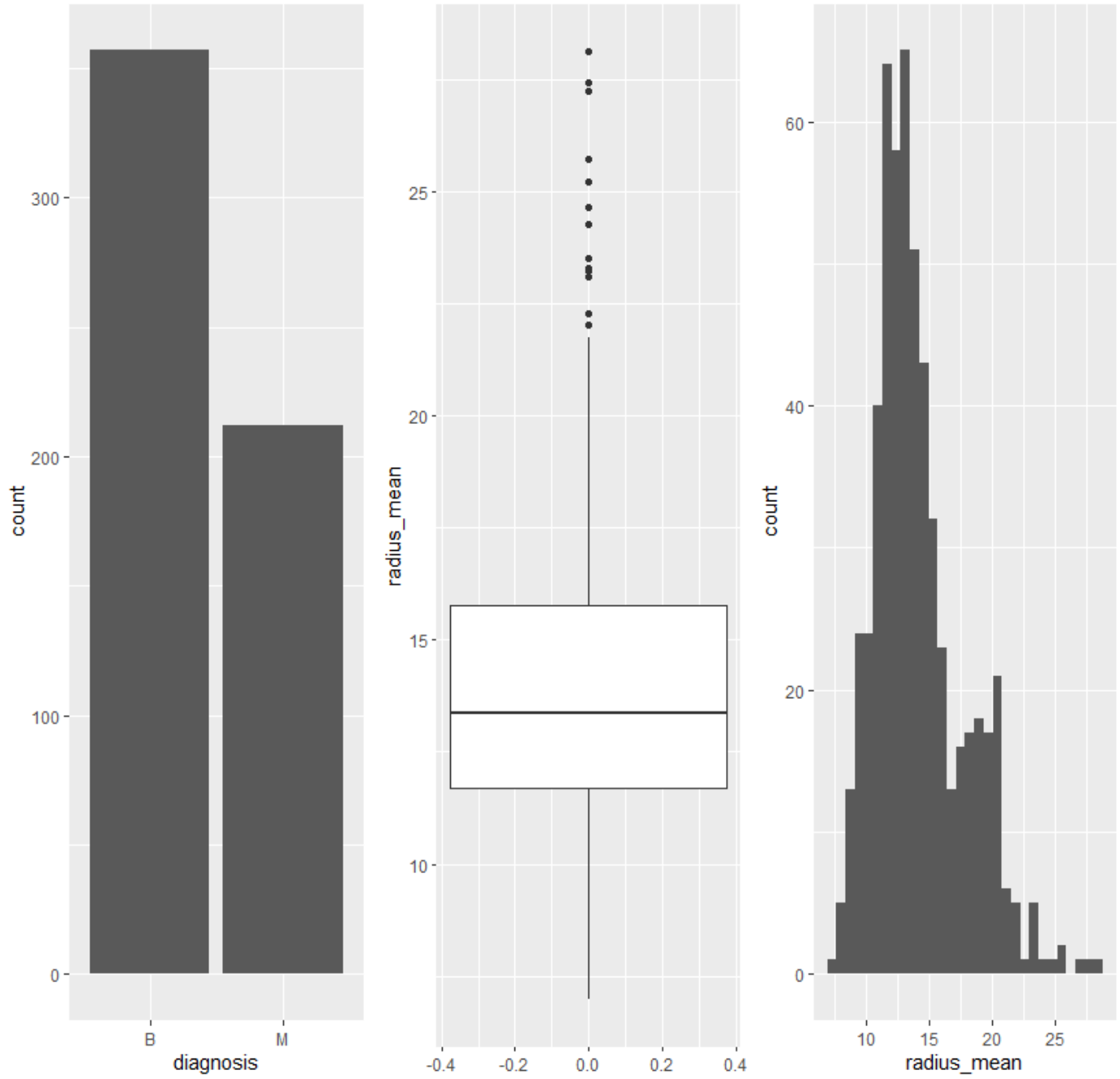
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
P1 <- ggplot(data=bcw_df, aes(x=diagnosis)) + geom_bar()

> p2 <- ggplot(data=bcw_df, aes(y=radius_mean)) + geom_boxplot()
> p3 <- ggplot(data=bcw_df, aes(x=radius_mean)) + geom_histogram()
>
> library(gridExtra)
> grid.arrange(p1, p2, p3, ncol = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



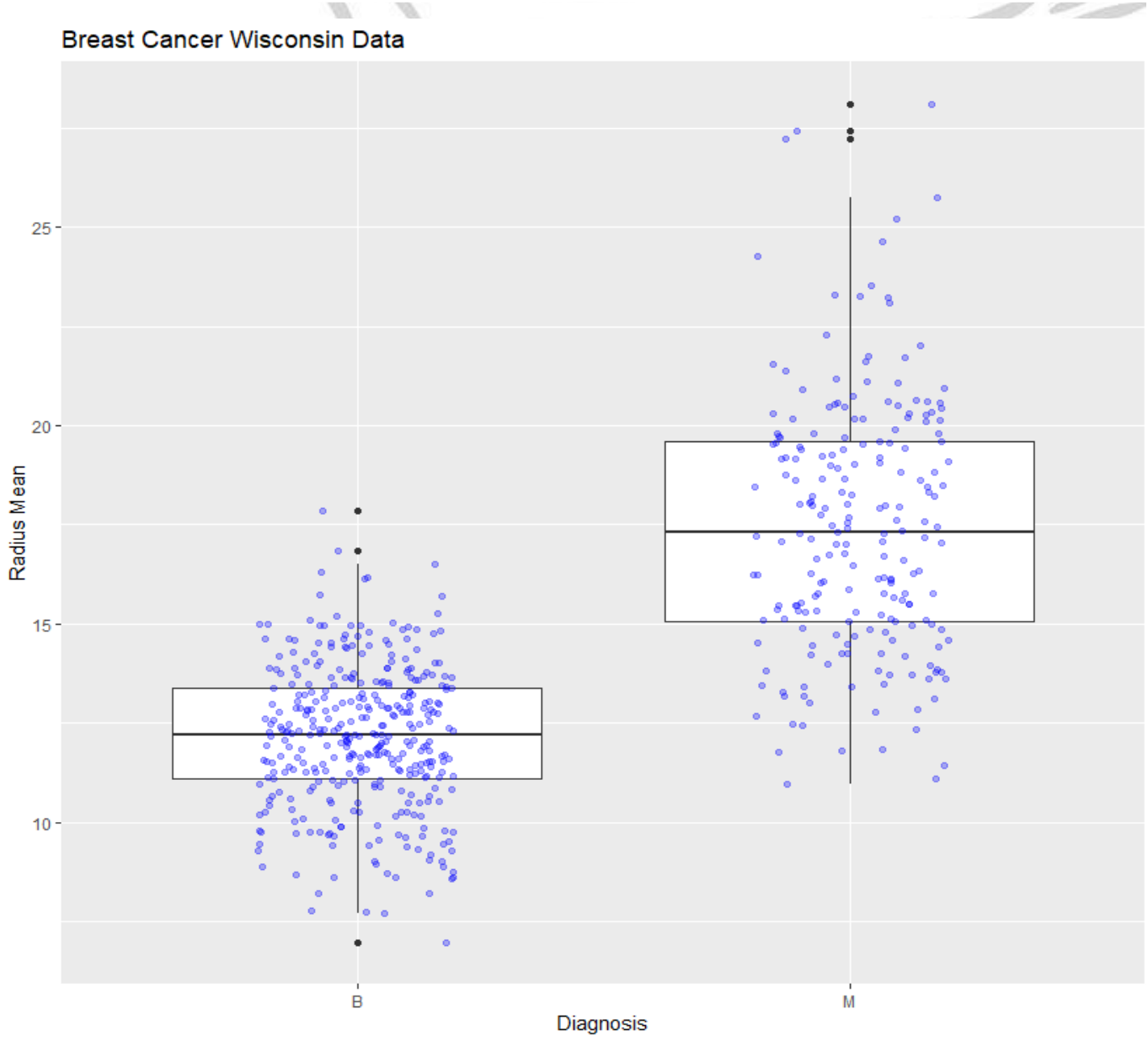
All comparison at a glance

2.2. Bivariate Data Analysis

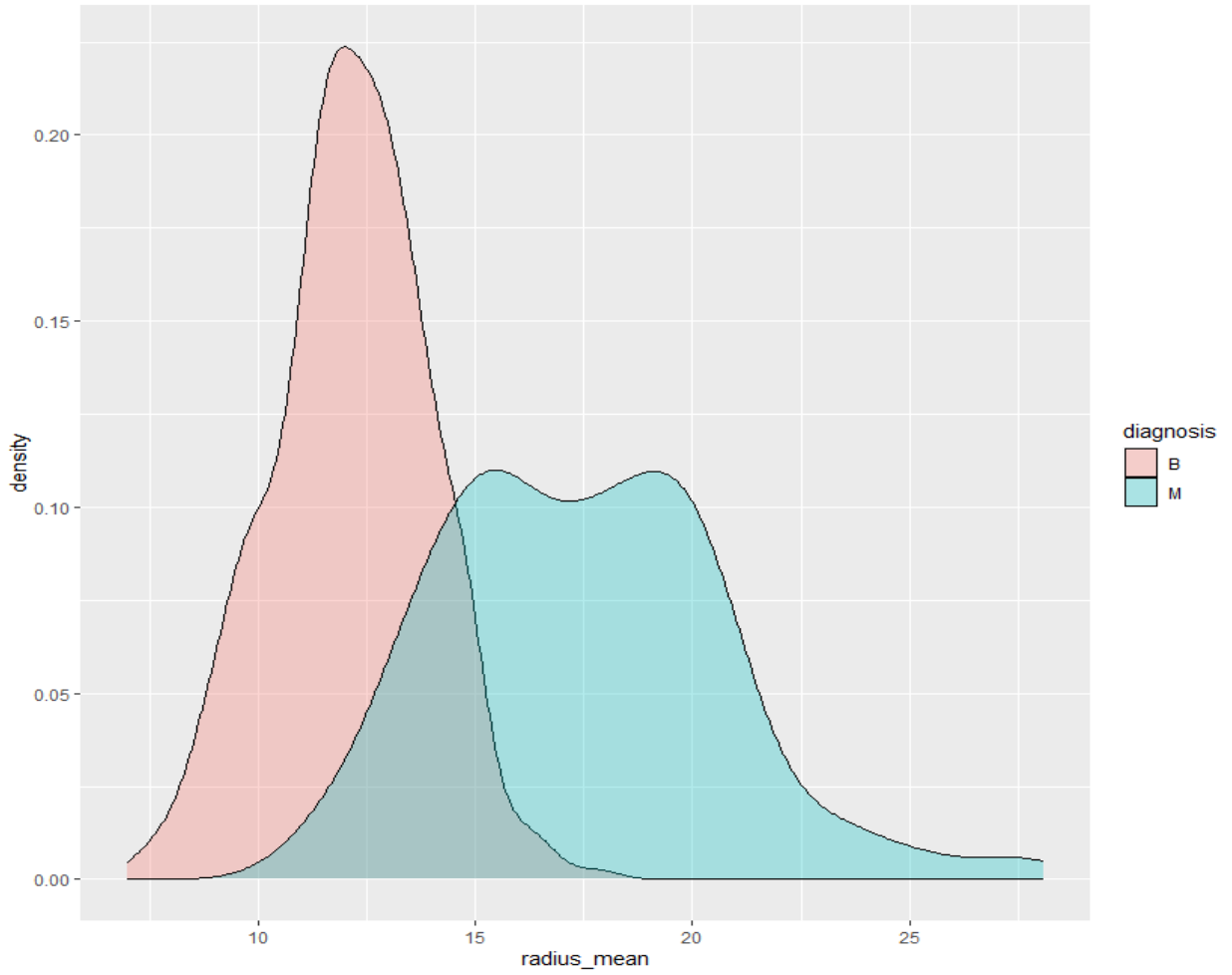
Bivariate analysis is a statistical method examining how two different things are related. The bivariate analysis aims to determine if there is a statistical link between the two variables and, if so, how strong and in which direction that link is.

Analysis of two variables. Distribution of radius mean variable based on diagnosis.

```
ggplot(data=bcw_df, aes(x=diagnosis, y=radius_mean)) +
     geom_boxplot() +
     geom_jitter(alpha = 0.3,
                 color = "blue",
                width = 0.2) +
    labs(title="Breast Cancer Wisconsin Data", x="Diagnosis", y="Radius Mean")
```

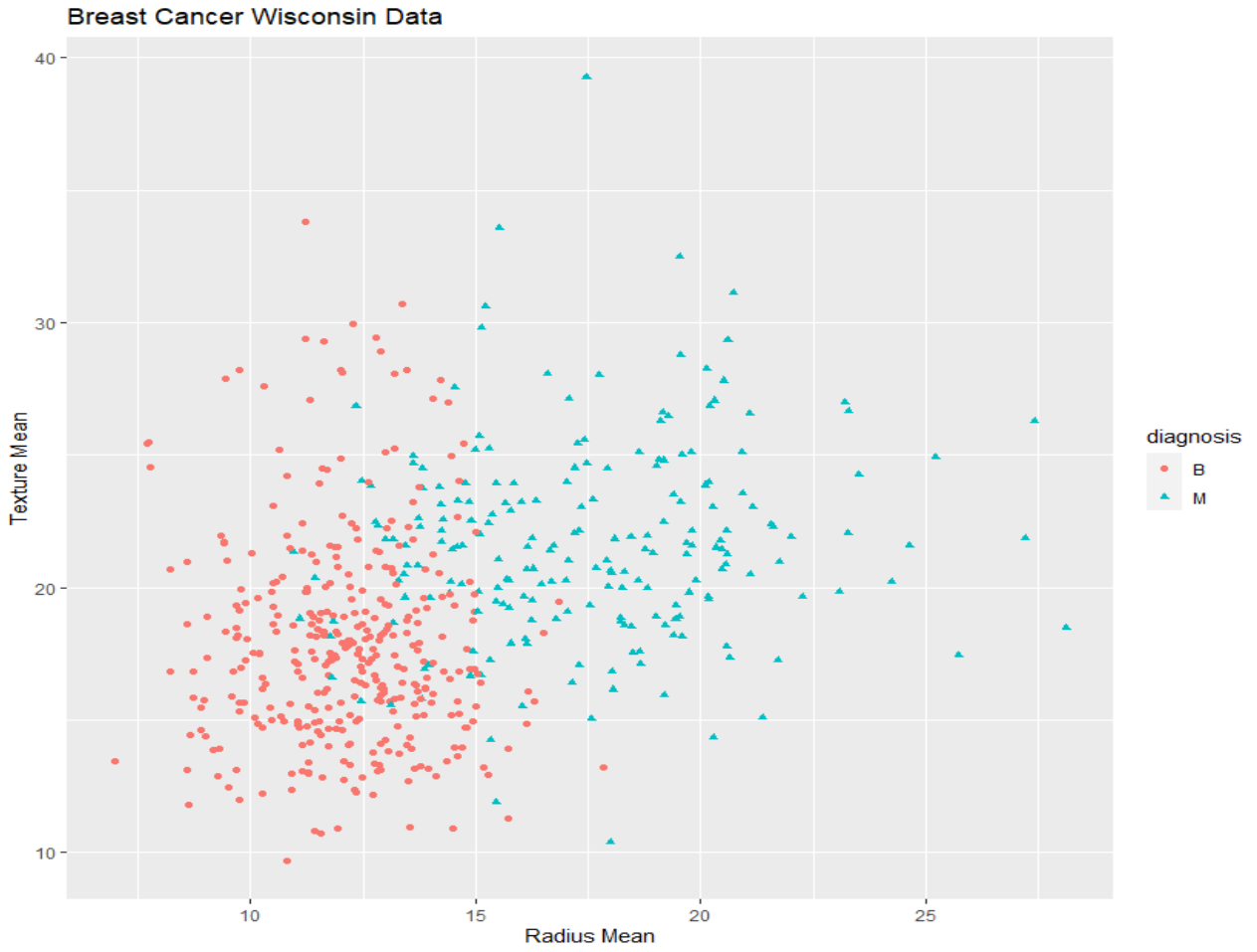


```
ggplot(data=bcw_df, aes(x=radius_mean, fill=diagnosis)) +
    geom_density(alpha=.3)
```



Observations based on **radius mean** and **texture** mean variables. Each point is a single observation. The color and shape of the observations are based on diagnosis (benign or malignant).

```
ggplot(data=bcw_df, aes(x=radius_mean, y=texture_mean,
                        shape=diagnosis, color=diagnosis)) +
    geom_point() +
    labs(title="Breast Cancer Wisconsin Data", x="Radius Mean", y="Texture Mean")
```



In general, **benign** has lower radius mean and texture mean measurement than **malignant**. However, these two variables are not enough two separate the classes.

2.3. Multivariate Data Analysis

There are three type of measurements: mean, standard error (se), and worst (mean of the three largest values). Each measurement has 10 variables so the total is 30 variables. We want to compute and visualize correlation coefficient of each measurement.

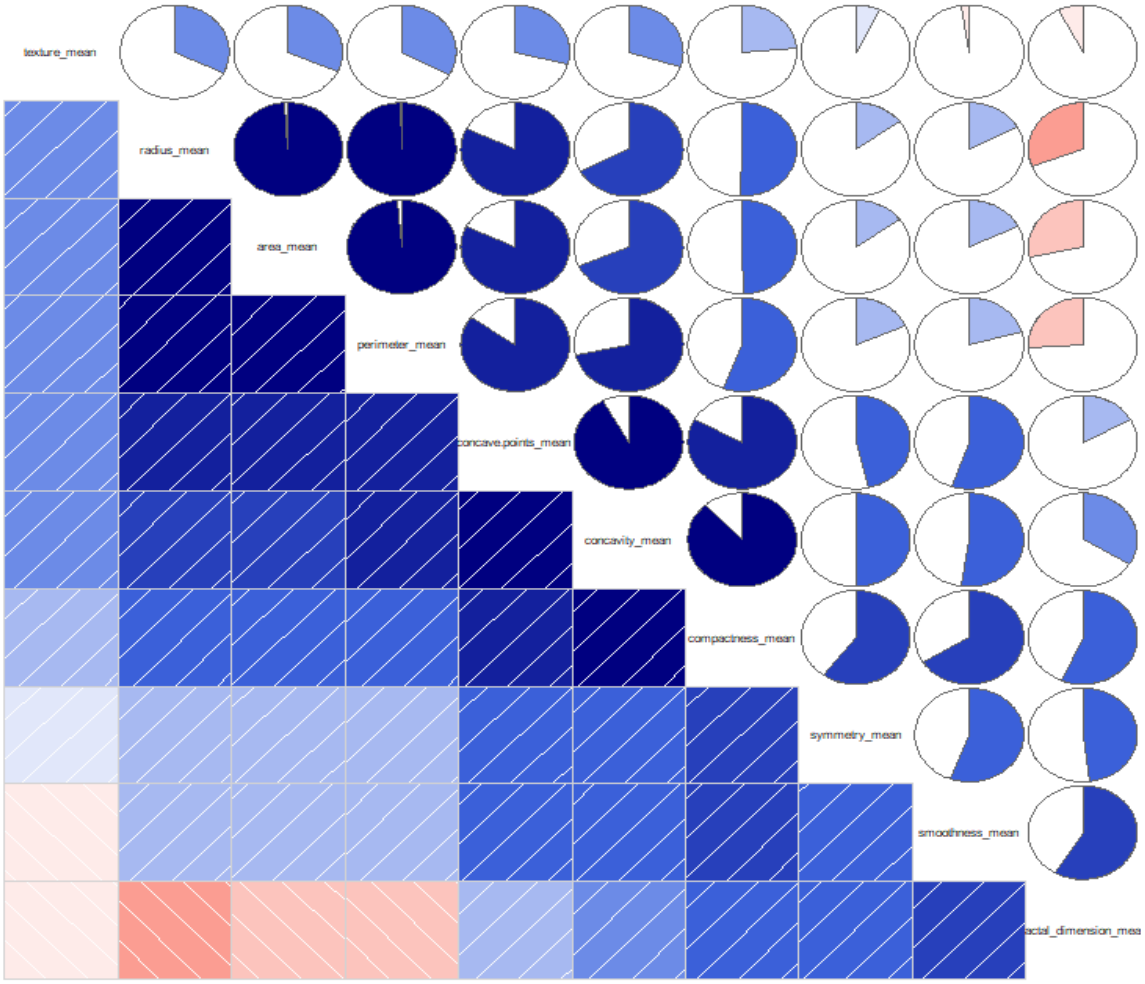
Visualize Pearson's Correlation Coefficient for *_mean variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where r = Pearson's correlation coefficient

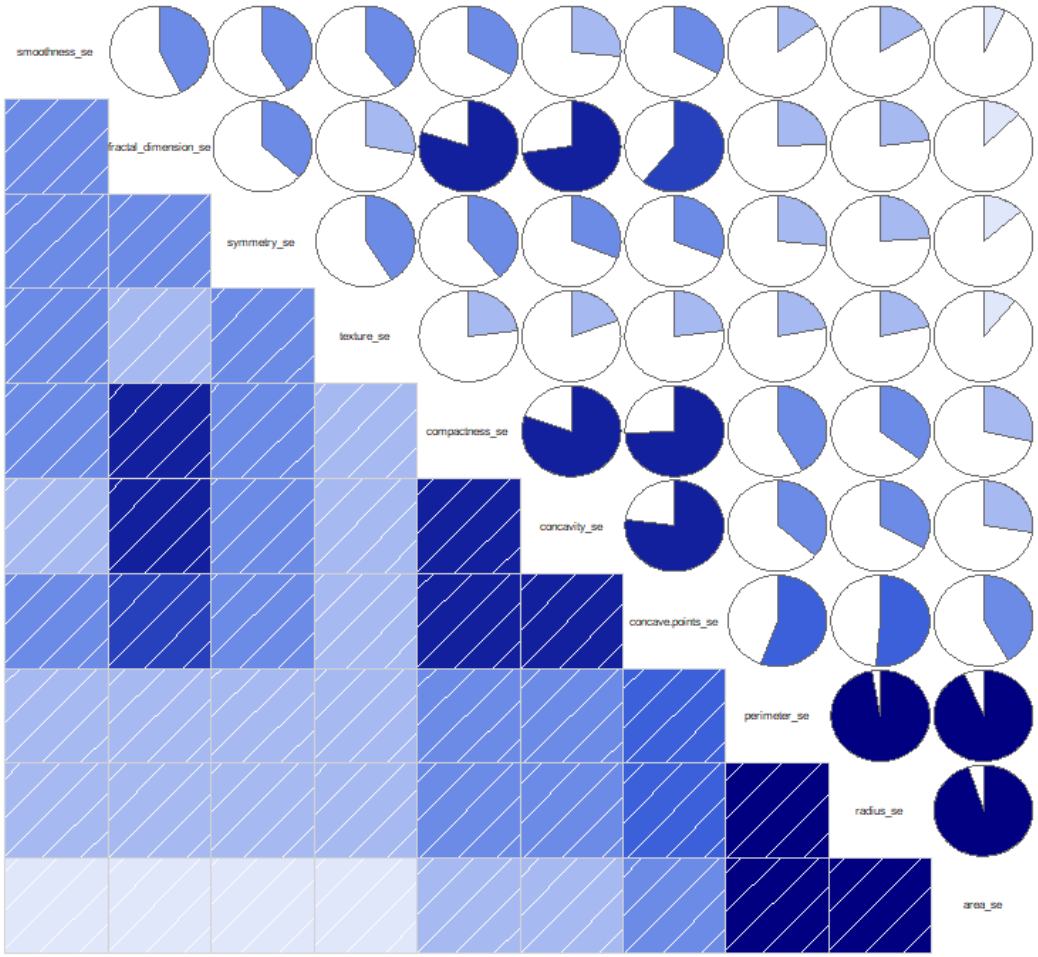
- Pearson correlation coefficient, also known as Pearson r, is a statistical test that estimates the strength between the different variables and their relationships. Hence, whenever any statistical test is performed between the two variables, it is always a good idea for the person to estimate the correlation coefficient value to know the strong relationship between them.
- The correlation coefficient of -1 means a robust negative relationship. Therefore, it imposes a perfect negative relationship between the variables. If the correlation coefficient is 0, it displays no relationship. Moreover, if the correlation coefficient is 1, it means a strong positive relationship. Therefore, it implies a perfect positive relationship between the variables.
- The Pearson correlation coefficient shows the relationship between the two variables calculated on the same interval or ratio scale. In addition, It estimates the relationship strength between the two continuous variables.

```
library(corrgram)
> corrgram(bcw_df[2:11], order = TRUE,
          upper.panel = panel.pie)
```



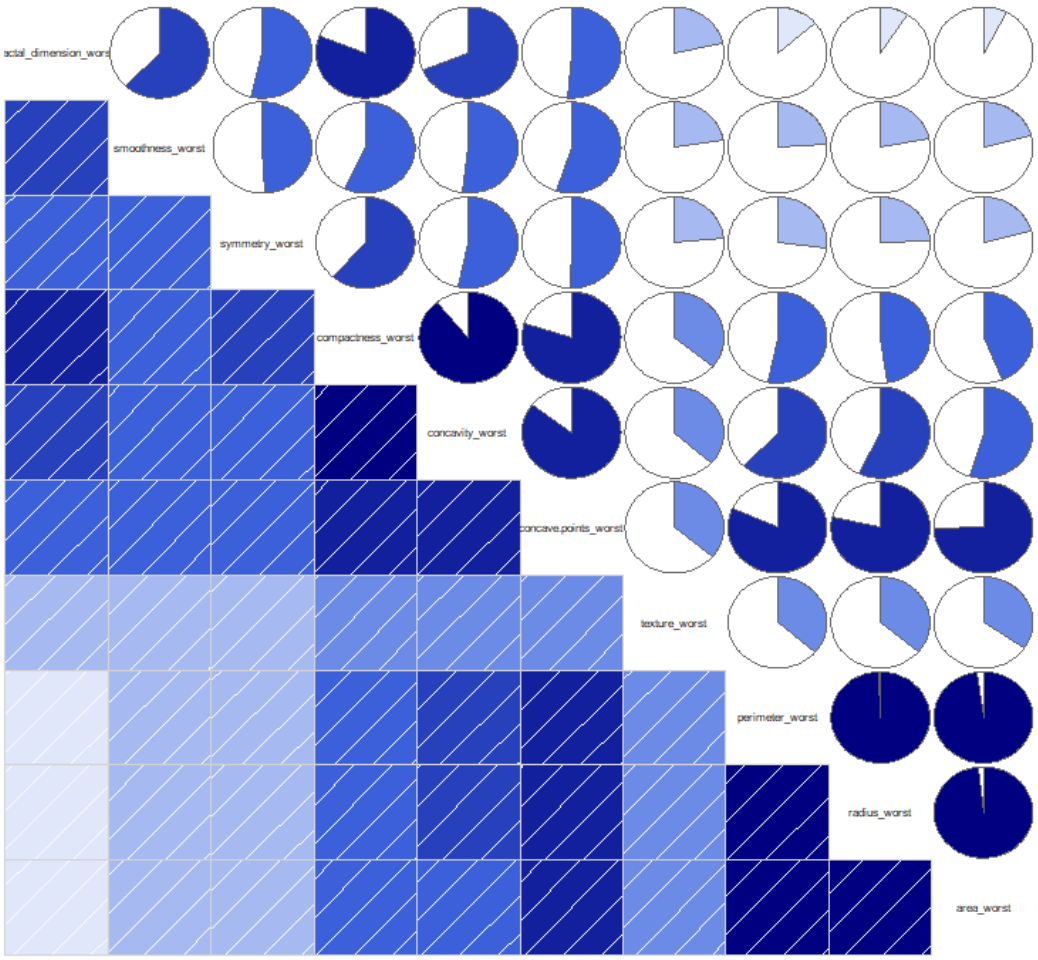
Visualize Pearson's Correlation Coefficient for *_se variables.

```
corrgram(bcw_df[12:21], order = TRUE,
         upper.panel = panel.pie)
```



Visualize Pearson's Correlation Coefficient for *_worst variables.

```
corrgram(bcw_df[22:31], order = TRUE,
         upper.panel = panel.pie)
```



Form the correlation coefficient, we can see that area, radius, and perimeter are co-linear. So, we need to remove two of them: area and perimeter.

We can also see that compactness, concavity, and concave points are co-linear. So, we need to remove two of them: compactness and concave points.

3. Analysing the model again to proceed further steps:

Now we start by taking a peak at the data. (We'll only be using the `area_mean` for this analysis for simplicity's sake).

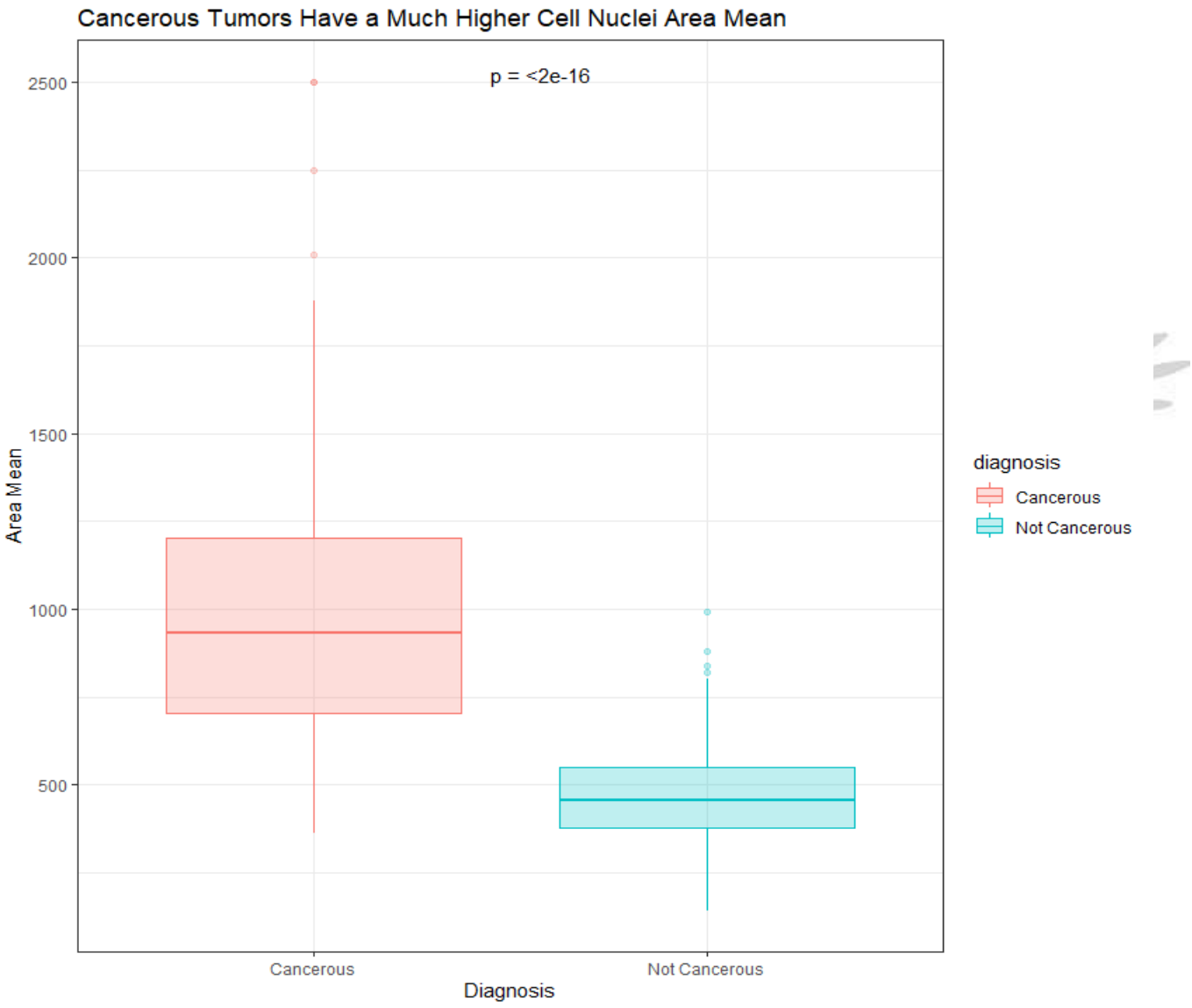
```
#Bringing in our data and cleaning it up a bit
> data <- read_csv("D:/chrome downloads/breast cancer.csv") %>%
      select(diagnosis, area_mean) %>%
      mutate(diagnosis = ifelse(diagnosis == "M", "Cancerous", "Not Cancerous")) %>%
      as.data.frame()
```

```
## # A tibble: 6 x 2
##   diagnosis area_mean
##   <chr>         <dbl>
## 1 Cancerous     1001.0
## 2 Cancerous     1326.0
## 3 Cancerous     1203.0
## 4 Cancerous      386.1
## 5 Cancerous     1297.0
## 6 Cancerous      477.1
```

Let's start by visualizing the distribution of area mean for both cancerous and not cancerous tumour cells using boxplots. This will help us see if the means are consistently different between cancerous and non-cancerous tumours.

We will also run a quick t-test to compare the means and see what our p-value is. The p-value will show up on the plot above the two boxplots.

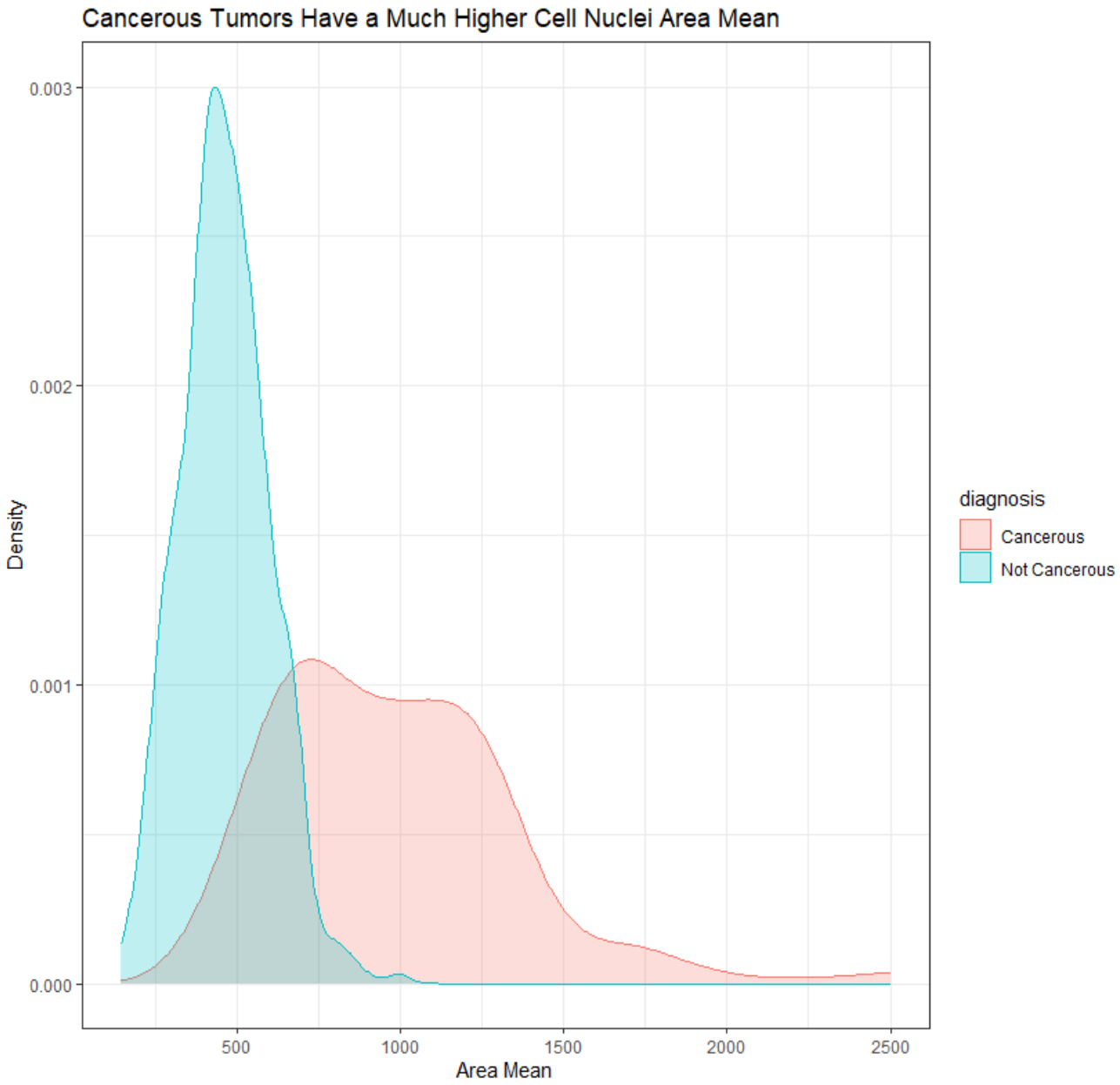
```
library(ggpubr)
>
> data %>%
      ggplot(aes(x = diagnosis,
                 y = area_mean,
                 color = diagnosis,
                fill = diagnosis)) +
      geom_boxplot(alpha=.25) +
      stat_compare_means(inherit.aes = TRUE,
                         label = "p.format",
                         method = "t.test",
                         label.x = 1.5,
                         show.legend=FALSE) +
      theme_bw()
      labs(x = "Diagnosis",
           y = "Area Mean",
          title = "Cancerous Tumors Have a Much Higher Cell Nuclei Area Mean")
```



There's a pretty large gap between the area_mean for cancerous tumours and the area_mean for not cancerous tumours! We also have an extremely small p-value, which implies that the means of are very consistently different.

We can also visualize our cancerous and not cancerous measurements with density plots:

```
data %>%
     ggplot(aes(x = area_mean,
                color = diagnosis,
                fill = diagnosis)) +
     geom_density(alpha=.25) +
     theme_bw() +
     labs(x = "Area Mean",
          y = "Density",
          title = "Cancerous Tumors Have a Much Higher Cell Nuclei Area Mean")
```



Insights:

- Because the cancerous and not cancerous area_mean's are consistely different, the area_mean will make a good predictor of cancer.
- Cancerous cells have higher variance (more spread out) than not cancerous cells

4. Creating Our Model

Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no.

Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model, like the Adeline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification.

To best predict a binary outcome (1 or 0, TRUE or FALSE), we'll run a logistic regression model.

Here's all the math parts of what we're going to do. The equation for a logistic regression model is as follows:

$$P(Y_i=1 \mid x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \pi_i$$

Y_i**= 1** denotes that the tumour is cancerous,

Y_i**= 0** denotes that the tumour is not cancerous, and

x_i**=**denotes the area_mean of the cell nuclei in the tumour.

Note that if β_1 is zero in the above model, then x_i (area_mean) provides no insight about the probability of a tumour being cancerous.
Thus, we could test the hypothesis that

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

In simple English : if the area_mean truly does help us predict if a breast tumour is cancerous, then it will show up in our formula as a number greater than 0!

5. Fitting our model

Here's where we fit our model! Remember, our input variable is the area_mean and our output variable is the diagnosis.

```
#Creating a column with 1's and 0's instead of "Cancerous" and "Not Cancerous"
> data <- data %>% mutate(diagnosis_0_1 = ifelse(diagnosis == "Cancerous", 1, 0))
>
> #Splitting our data into 80% training and 20% testing data sets
> library(caret)
> training <- createDataPartition(data$diagnosis_0_1, p=0.8, list=FALSE)
> train <- data[ training, ]
> test <- data[ -training, ]
>
> #fitting our model
> breast_cancer_glm <- glm(diagnosis_0_1 ~ area_mean, data = train, family = "binomial")
> summary(breast_cancer_glm)
```

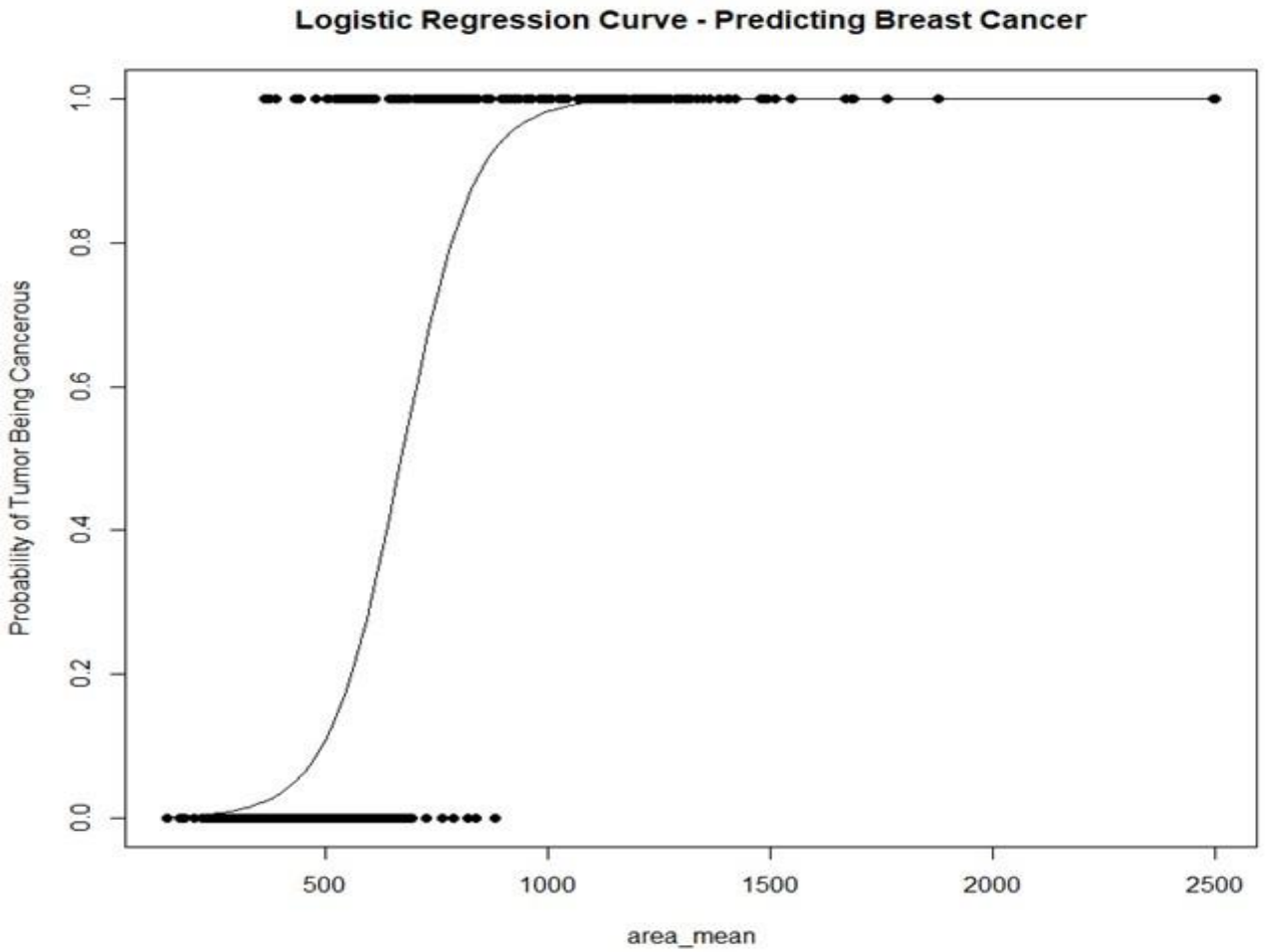
```
##
## Call:
## glm(formula = diagnosis_0_1 ~ area_mean, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7563  -0.4715  -0.2117   0.1226   2.7544
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.098550   0.769236 -10.528   <2e-16 ***
## area_mean    0.011969   0.001222   9.793   <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: 598.00  on 455  degrees of freedom
## Residual deviance: 266.99  on 454  degrees of freedom
## AIC: 270.99
##
## Number of Fisher Scoring iterations: 6
```

Just as we found before, our p-value is very low, meaning that we can use this feature in our model with confidence.

6. Visualizing the model

To visualize this simple logistic regression we could make the following plot.

```
 plot(diagnosis_0_1 ~ area_mean,
data=train,
main="Logistic Regression Curve - Predicting Breast Cancer",
ylab='Probability of Tumor Being Cancerous', +      pch=16)
 >
> curve(
exp(breast_cancer_glm$coef[1]+breast_cancer_glm$coef[2]*x)/
(1+exp(breast_cancer_glm$coef[1]+breast_cancer_glm$coef[2]*x)),
add=TRUE
)
```



Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group

7. Interpretation

Here's what the mathematical model for πi would now look like:

$$P(Y_i = 1 \mid x_i) \sim \frac{e^{-8.1-.01\,x_i}}{1 + e^{-8.1-.01x_i}} = \hat{\pi}_i$$

Where b_0= -8.1 is the value of the `Intercept` which estimates $\beta_0$and b_1= 0.01 is the value of `area_mean` which estimates β_1. Because the p-value for the logistic regression is <0.05, we can reject our null hypothesis that `area_mean` doesn't predict the `diagnosis`. It actually does!
But , by how much?

Well, in our model's case, the y-intercept isn't very helpful because none of the `area_mean`'s were 0. That would imply that the tumour simply didn't exist. However, the value of $e^{b_1} = e^{0.0117} \approx 1.012$ shows that the odds of a tumour being cancerous increase by approximately 1.012 for every increase of 1 in `area_mean`.

8. Checking accuracy of our model

We can test our model's accuracy by predicting our test-data values with our model and seeing how many we got right!

```
train$pred_prob <- predict(breast_cancer_glm, data=train, type='response')
> train$pred <- ifelse(train$pred_prob >= .5, 1, 0)
>
> #Testing our predictions against our training set
> results <- train %>%
+     mutate(correct = (diagnosis_0_1 == pred)) %>%
+     group_by(correct) %>%
+     tally()
>
> accuracy <- round(results$n[2]/(nrow(train)),3) #Calculating accuracy
>
> print(paste0("We made ", results$n[2], " correct predictions,"))

[1] "We made 406 correct predictions,"


print(paste0(results$n[1], " incorrect predictions, "))


[1] "50 incorrect predictions, "


print(paste0("thus giving us an accuracy rating of: ", accuracy*100, "%"))


[1] "thus giving us an accuracy rating of: 89%"
```

9. Real life implication of the model :

One of my good friend Reetama recently went to the doctor to get a check up. After the check up, she said me

Something that nobody wants to hear from their friend after a doctor's visit. "They found something."

The doctor has found a tumour in her breast that could be cancerous. The doctors, however, weren't sure if it was cancerous or not. If it's cancerous, your friend has to go through rounds of chemo therapy and have to pay for mountains of medical bills. There's also a chance that the tumour is harmless.

This diagnosis means everything to your friend!

So by this logistic regression model we can predict whether or not her tumour is cancerous! Perhaps this can save her from a lot of unnecessary suffering.

Her doctor give me some data that they gleaned from some 3D images of the tumour cells. You find out that her tumour cell nuclei have an `area_mean` of 511 units.

Using the Model to Predict Reetama's Diagnosis

So let's put this model to the test! Recall that the `area_mean` of Reetama's tumour cell nuclei is 511 unit.

```
prediction <- predict(breast_cancer_glm, data_frame(area_mean = 511), type="response")

(paste0("According to our model, the probability of Reetama's tumor being cancerous is
", round(prediction[1], 4)*100, "%."))
```

```
## [1] "According to our model, the probability of Reetama's tumour being cancerous is
12.11%."
```

With such a low probability, we can diagnose Reetama as "not cancerous" !.We can show the doctors this results, and they can decide against performing chemo-therapy.

10. Final conclusion

After developing and evaluating a breast cancer prediction model, several key conclusions can be drawn:

1. Accuracy: Assess the accuracy of the model by evaluating its performance on a test dataset. Measure metrics such as sensitivity, specificity, precision, and F1 score to determine how well the model predicts breast cancer. High accuracy indicates that the model is reliable and effective in identifying potential cases.

2. Feature Importance: Determine the most influential features or variables in the model. Identify the factors that contribute significantly to the prediction of breast cancer. This information can help researchers and medical professionals understand the key indicators and risk factors associated with the disease.

3. Model Validation: Validate the prediction model using different datasets or employing cross-validation techniques. This process helps ensure that the model's performance is consistent and not biased towards a specific dataset. A well-validated model will provide more reliable predictions in real-world scenarios.

4. Clinical Applicability: Assess the practicality and clinical applicability of the model. Consider factors such as ease of use, interpretability, and integration with existing medical systems or workflows. A user-friendly model that aligns with clinical requirements will have a higher chance of adoption and implementation in healthcare settings.

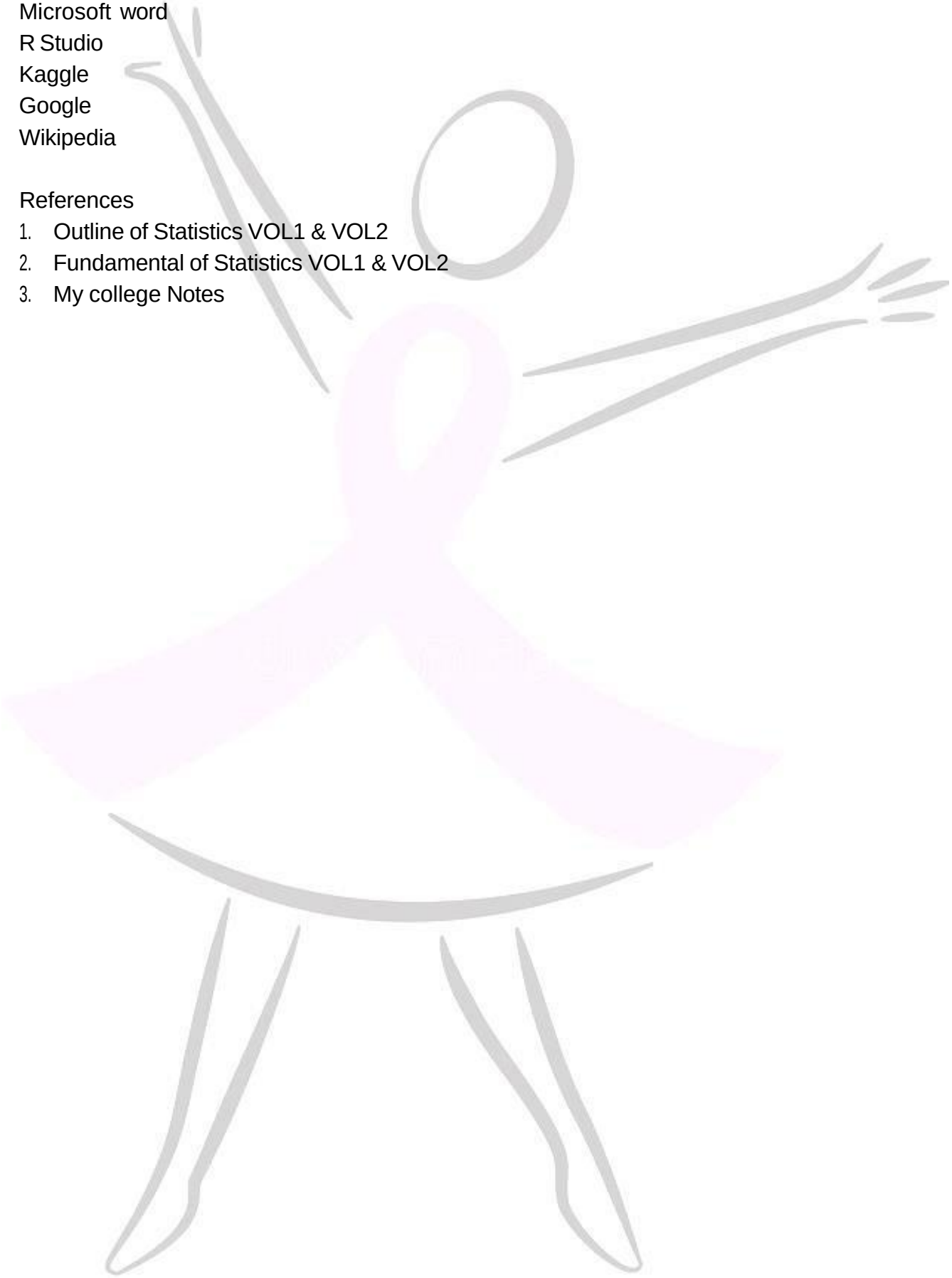
5. Future Improvements: Identify areas for further improvement and research. Machine learning models are continuously evolving, and advancements in technology and data availability can lead to enhanced prediction capabilities. Feedback from medical experts, additional data sources, and novel algorithms can contribute to refining the model and increasing its accuracy.

11. The Tools I used

1. Microsoft Power point
2. Microsoft word
3. R Studio
4. Kaggle
5. Google
6. Wikipedia

References

1. Outline of Statistics VOL1 & VOL2
2. Fundamental of Statistics VOL1 & VOL2
3. My college Notes



Acknowledgement:

I am indebted to a number of people for helping me in the preparation of this project. Firstly, Prof. (Dr.) Manas Kabi Principal, Asutosh College, University of Calcutta Without whose help I couldn't have been a part of this prestigious college. I owe a deep debt of gratitude to my supervisor Dr. Dhiman Dutta sir for necessary guidance, for this presentation of this dissertation, valuable comments and suggestions. I am extremely grateful to him for giving me the necessary stimulus, support and valuable time. Again Special thanks to Dr. Dhiman Dutta, Head of the Department of Statistics, Asutosh college. I am greatly indebted to Dr. Parthasarathi Bera. Dr. Shirsendu Mukhopadhay and Oindrila Bose (Faculty members) often took pains and stood by me in adverse circumstances. Without their encouragement and inspiration, it was not possible for me to complete this project. Finally, my earnest thanks go to my friends who were always beside me when I needed them without any excuses and made these three years worthwhile. This project is not only a mere project. It is the memories spent with the whole department which has created a mutual understanding among us. There are many emotions related to this piece of work, especially respect and duty towards teachers and vice versa; educational attachment with my friends; social attachment with my college.

SAYAN GHOSH

Student, Department of Statistics

DECLARATION

I am Sayan Ghosh, a student of B.Sc.Sem-6, Statistics Honours, of university of calcutta, Registration no- 012-1111-0763_20, Roll no- 203012-21-0083 hereby declare that I have done this piece of project work entitled as "Breast cancer Prediction" under the supervision of Dr. Dhiman Dutta (HOD , Department of Statistics, Asutosh college) as a part of B.Sc.Sem-6 examination according to the syllabus paper DSE B2. I further declare that the piece of project work has not been published elsewhere for any degree or diploma or taken from any published project.