**Page 1: A. Report Cover Page #1:**

**Golden Gate University**

**Summer 2017**

**Adarsh Gupta**

**SID: 0585957**

**MSBA320.SF1**

**(Breast Cancer Wisconsin Data Analysis)**



Differentiating between Malignant and Benign

Breast Cancer

**Contents**

Introduction to Breast cancer and its Diagnosis…………………………………………………………………4

Gathering data information on Breast Cancer Wisconsin………………………………………………….5

Descriptive Statistic………………………………………………………………………………………………………….5

Correlation between the variables……………………………………………………………………………………12

Logistic Regression…………………………………..………………………………………………………………………..13

Hypothesis Testing ……………………………………………………………………………………………………………17

Results………………………………………………………………………………………………………………………………21

Conclusion………………………………………………………………………………………………………………………..21

Recommendation……………………………………………………………………………………………………………..21

Final Notes………………………………………………………………………………………………………………………..22

Appendix(References)……………………………………………………………………………………………………….23

Coding………………………………………………………………………………………………………………………………24

**Introduction to Breast Cancer and its diagnosis:**

Breast Cancer is a cluster of unwanted form of cells in the form of lump or any hard clot in the breast. Breast cancer is growing rapidly in the throughout the world but the cases are at rise at a very high rate in USA. U.S. Breast Cancer statistics report that about 1 in 8 U.S. women (about 12%) will develop invasive breast cancer over the course of her lifetime. In 2017, an estimated 255,180 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 63,410 new cases of non-invasive. Breast cancer is one of the main major leading cause of death across the world and the cases are substantially at a rise. Diagnosing the disease is almost said as a prevention but sometimes it is not diagnosed properly which leads it to become more chronic and deadly. If something is suspected then it is needed to take care of it and do proper testing with the help of clinical procedure. Medical science has number of procedure to diagnose it. Before going to the hospital, self-examination is also necessary and must not be ignored. If something suspected then medical help must be taken to prevent it and cure it in case. Medical science has mainly two testing procedure to do it:

**Blood testing**

**Imaging**

Both are non-invasive. In rare case, t is required to perform invasive procedure when it is uncertain that what kind of clot is it.

Apart from all these, machine learning is also evolving at a very fast rate. It is proving to be very effective in different field such as diagnosing disease, recommending treatment etc. doctors are also taking help of it and using it additionally to give accurate information. Doing so, it is saving millions of lives to reach to the conclusion in time. Machine learning applies to the large sets of data and find relevant information through modeling techniques which become standard and provide result for the upcoming cases. Machine learning result are then tested on another set of data and we call them prediction. If the prediction accuracy is high then it is included in the modeling algorithm and serve as final model indicator.

**Gathering data information on Breast Cancer Wisconsin:**

We are going to make model here and predict whether the tumor is malignant or benign based on the set of information provided in the data set related to the breast cancer. We downloaded the data from [www.Kaggle.com](http://www.Kaggle.com) which is source of dataset information and implementing data modeling. This analysis is not for professional use and it is just only for analyzing the information given in the data set.

The dataset contains information on about 569 patient records and the type of tumor is also defined depending upon the size of tumor and all other attributes. Such as radius, texture, symmetry, smoothness, concavity etc. there are total 32 attributes but they are defined according to the mean and standard error of the radius, texture etc. Also, there are worst case radius, texture and all other attribute information is given. We will be considering 11 attributes to perform statistical operations on malignant and benign data. Based on the result, we would be able to conclude which are malignant or benign and which attributes are most significant and can be used as a result indicator.

**Descriptive statistics:**

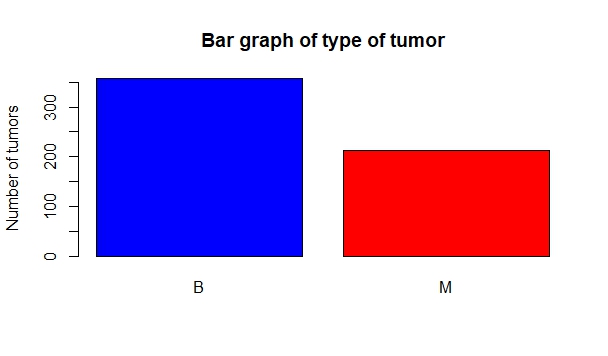
Descriptive statistics comprises of basic operations and hypothesis testing to reach and conclude what c=kind of information is present in to the data set. It also helps to differentiate between the data. Sometimes, most of the conclusion is achieved only with the use of descriptive statistics.

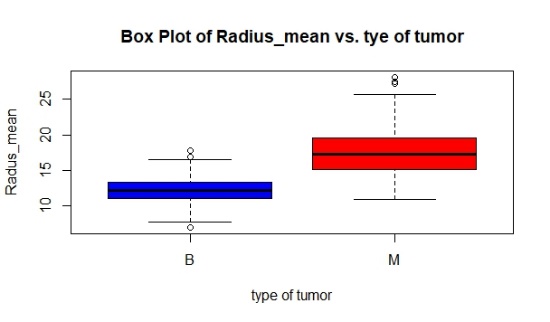
Trend, spread, deviation, repletion of the data can be easily seen through descriptive analysis. In our dataset on breast Cancer, we will be performing following implementation using R:

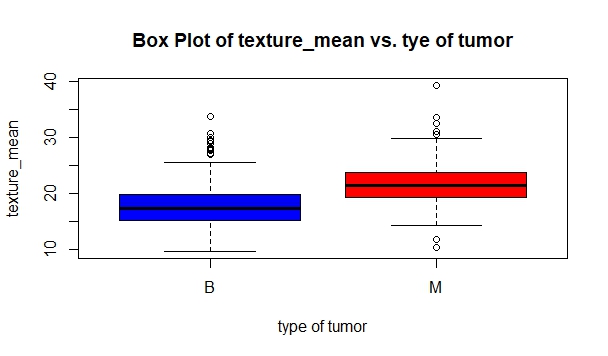
**Visualization of the data using box plot technique;**

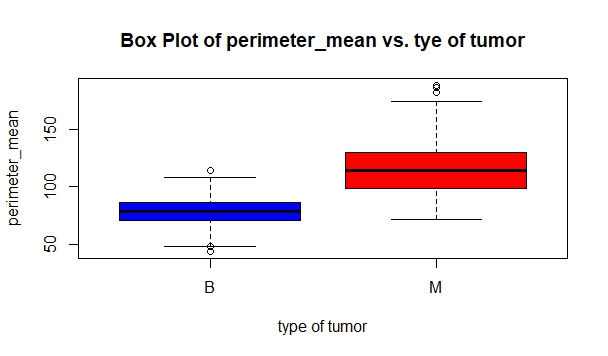
Visualization of technique involves the use of applying functions such as scatter plot, box plot, histogram to visualize the spread of data and check for significant attribute. In our dataset, response variable is of binary type, so using R, we created box plot because it is not possible to create scatter plot and visualize binary data. Box plot gives us an estimate of significant attributes that should be included further in the modeling and remove non-significant attribute. If the data is not overlapping or less overlapping in the box plot that attribute can be include in the significant list otherwise not. We created scatterplot of each attribute such as radius, texture etc. depending upon malignant or benign type.

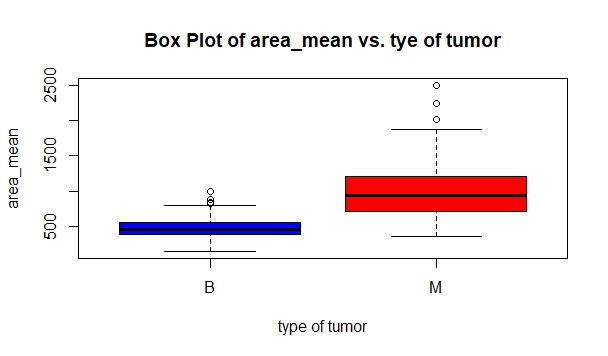
Before analyzing boxplot let us check the total number of malignant and benign type in the data set with the help of bar chart and their count.

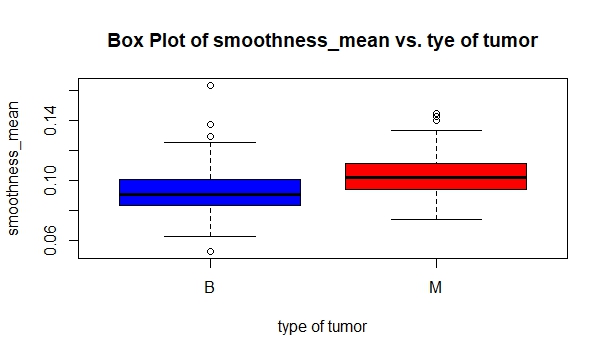
**Box plots of different attributes:**

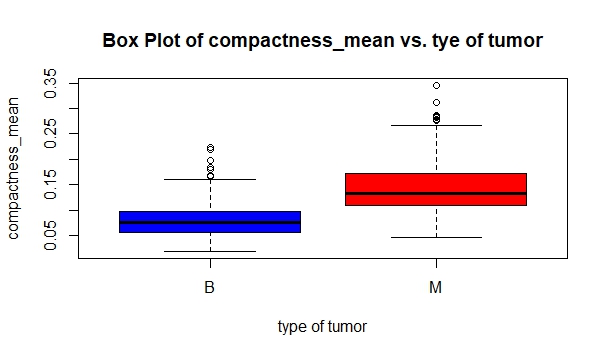


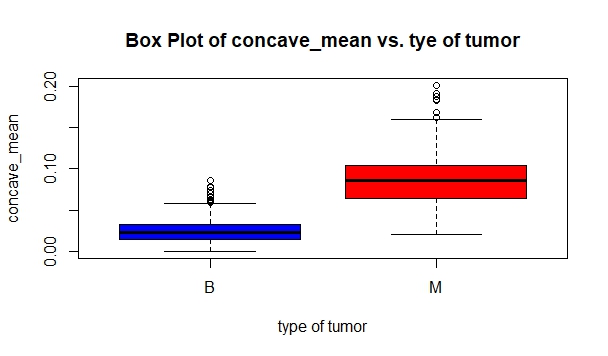


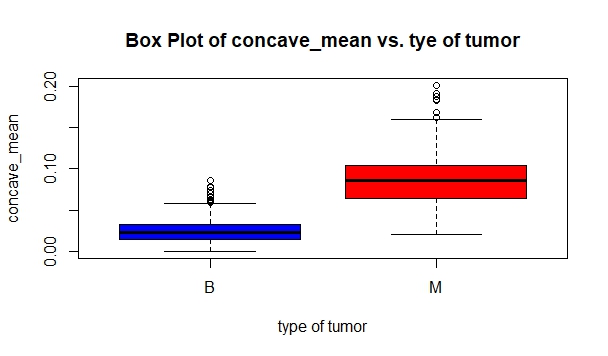


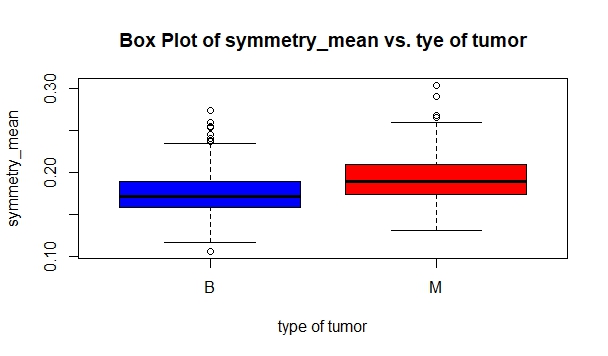


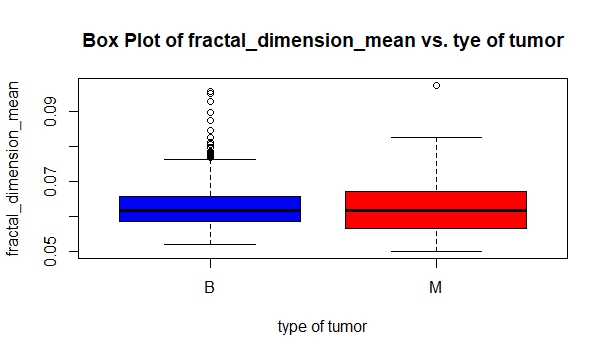








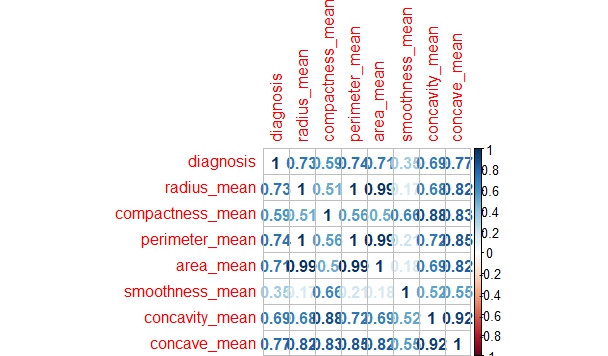




From the above box plot, non-overlapping box plots will be much more significant than the over-lapping one. It can be seen from the above box plots that the radius, area, smoothness, compactness, concavity and concave mean are more significant than the other one so we will take them to make our regression model and check for which will contribute more as a whole through regression technique.

**Correlation matrix:**

The aim of correlation matrix is to check statistical significance between two or more variables. Here, we can see some variables are highly correlated with each other such as radius vs. area, perimeter vs. area. If we do linear model of each independent variable then it is fine to but in case of multiple regression it is required to check it because two highly correlated variables might cause collinearity. But in our case, logistic regression in R will provide only significant variables result that we can use to make final model.



**Regression analysis:**

In statistical modeling, regression analysis is a statistical process for estimating the relationship among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

In our data, we will be performing logical regression suing R because our response variable is of binary type and independent variable will be radius, area, concavity, perimeter, concavity, concave mean.

**Steps Followed to Perform Logistic Regression**

**Creating a model:** First, we have created a generalized linear model using logit function to perform logistic regression and saved it into model variable.

**Multi co-linearity Check:** Next step is to check for multicollinearity. Multicollinearity can be checked by two ways: VIF or Co relation matrix. We followed VIF function and performed VIF function on model variable which contains regression result and applying VIF function on model gives us the result and we need to remove the variables which have VIF value greater than 5. So, we checked and removed radius and perimeter variables one by one.

**P value Significance:** Next step is to check for p value significance for each variable and checking for P value, we found compactness non-significant so, we removed that one too from our regression model.

**Final Model:** Finally, we ran regression once again and found area, smoothness and concavity mean significant. This will be our final regression model with area, smoothness and concavity\_mean.

**R square analysis in GLM model:** As we have glm model, we need to found R square through another method using manual computation formula which is

**1-Residual Deviance/Null Deviance=> 0.70**

**Regression Result:**

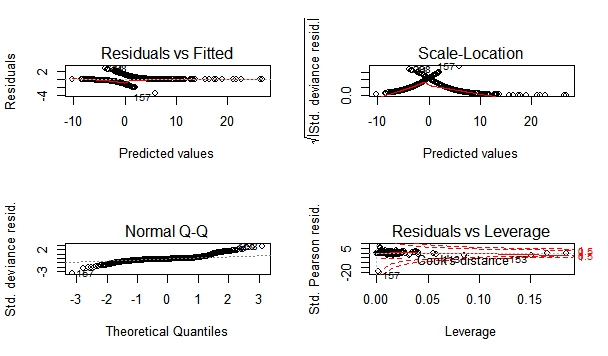
From the regression result, we found Area, Smoothness and Concavity means significant. These three variables can be used to tell the difference between malignancy and benign. Form the regression result, it is found that the 70 percent of variation in Malignancy and Benign can be explained by Area, Smoothness and Concavity Mean variables.

**Normality Checking using Shapiro Test:**

We also checked the Area, Smoothness and Concavity using **SHAPIRO.TEST** function to check for the normality of the data. The p-value is significant for each of the variables so we can assume that the data for three variables come from the normal population.

**Residual Plot and normal Q-Q Plot Analysis:**

In logistic regression, the residual plots have not that much importance but it is better to have a look at them. Residual and normal probability plots are plotted to check the normal of the data and constant variance of the data (homoscedasticity). In our residual plot, it shows a homoscedastic pattern. Normal probability plot shows little bit deviation in the fitted line but we assume it ok because we already have checked the normality of the data through SHAPIRO.TEST.



> model<-glm(formula=diagnosis~.,family=binomial("logit"),data=bc)

> summary(model)

Call:

glm(formula = diagnosis ~ ., family = binomial("logit"), data = bc)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.4417 -0.2761 -0.1127 0.0475 2.7821

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -15.598075 1.941956 -8.032 0.000000000000000958 \*\*\*

area\_mean 0.011675 0.001338 8.727 < 0.0000000000000002 \*\*\*

smoothness\_mean 62.606008 15.004352 4.173 0.000030124464988908 \*\*\*

concavity\_mean 17.716603 3.425409 5.172 0.000000231461712562 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 749.33 on 568 degrees of freedom

Residual deviance: 223.05 on 565 degrees of freedom

AIC: 231.05

Number of Fisher Scoring iterations: 7

**ANOVA and T Testing on Area, Smoothness and Concavity:**

**Hypothesis testing:**

Hypothesis testing is done here to check, is there any mean difference in between the statistical significant variables that we have selected for further data modeling such as radius, texture, smoothness, area, perimeter, concavity, concave etc. the aim of hypothesis is to check the mean difference between the malignant and benign radius, texture and so on. It provides us the clue about the population mean difference where the sample is taken from.

**ANOVA RESULT:**

**ANOVA TESTING:**

aov(area\_mean~diagnosis)

Call:

aov(formula = area\_mean ~ diagnosis)

Terms:

diagnosis Residuals

Sum of Squares 35358547 34984592

Deg. of Freedom 1 567

Residual standard error: 248.3973

Estimated effects may be unbalanced

> summary(aov(area\_mean~diagnosis))

Df Sum Sq Mean Sq F value Pr(>F)

diagnosis 1 35358547 35358547 573.1 <2e-16 \*\*\*

Residuals 567 34984592 61701

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(aov(smoothness\_mean~diagnosis))

Df Sum Sq Mean Sq F value Pr(>F)

diagnosis 1 0.01444 0.014444 83.65 <2e-16 \*\*\*

Residuals 567 0.09791 0.000173

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(aov(concavity\_mean~diagnosis))

Df Sum Sq Mean Sq F value Pr(>F)

diagnosis 1 1.750 1.7504 533.8 <2e-16 \*\*\*

Residuals 567 1.859 0.0033

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

We have performed ANOVA testing to check the mean difference in AREA, SMOOTHNESS and CONCAVITY mean. Through ANOVA testing, we would be able to tell only the mean difference in the area, smoothness and concavity among Malignant and Benign sample but to tell the difference whether it is large or see which one is large, we need to apply T Test. So, we have performed T Testing on all these significant variables.

ANOVA works best on categorical data where we need to analyze samples. As we have categorical data, we performed ANOVA to check whether there is truly a mean difference exists or not. We checked each variable one by one and compare it to p value. the p values are significant for each variable so we rejected the null hypothesis and conclude that there is mean difference in two sample means among Area, Smoothness and Concavity. But the limitation of ANOVA here is that we can only tell the difference exists or not among different samples. To check for actual mean difference or whether it is greater than particular value, we need to look for T Test.

**T TEST RESULT:**

**T TESTING:**

> t.test(area\_mean~diagnosis,alt="less",data=bc)

Welch Two Sample t-test

data: area\_mean by diagnosis

t = -19.641, df = 244.79, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -472.2439

sample estimates:

mean in group B mean in group M

462.7902 978.3764

t.test(smoothness\_mean~diagnosis,alternative="less",var.equal=F,data=bc)

Welch Two Sample t-test

data: smoothness\_mean by diagnosis

t = -9.096, df = 460.49, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.008384321

sample estimates:

mean in group 0 mean in group 1

0.09258111 0.10282086

> t.test(concavity\_mean~diagnosis,alternative="less",var.equal=F,data=bc)

Welch Two Sample t-test

data: concavity\_mean by diagnosis

t = -20.003, df = 294.81, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.1046091

sample estimates:

mean in group 0 mean in group 1

0.04672025 0.16073448

T Test in R provides a better way to analyze the sample means. It also has tailed testing option to define our criteria. In our sample means for three different variables, we looked for mean difference, whether it is significantly greater than 0 or not.

Level of significance selected: 5%

Null Hypothesis: True difference in mean is greater than 0

Alternative Hypothesis: True difference in mean is less than 0

For Area, smoothness and concavity, we reject the null hypothesis at p value <2.2e.06 and conclude that is true difference in sample mean is less than 0.

If we see the actual sample mean for two samples the sample mean for Malignant is larger than Benign but here we are concluding that the mean difference is less than 0. The reason behind this is the Benign data values are coming first whose total mean is less than the malignant one. So, on that basis the difference coming into negative otherwise the true mean difference is larger than zero. The mean value can also be changed using mu= function and we can set up particular mean difference test.

**Results:**

After performing different operations on the data set we got different result. These results are correlated and step by step procedure to reach to the final result. Visualization of data provided us general view of spread of data then correlation provided us the correlated variables. Hypothesis testing is used to check for the true difference in population data. Finally, regression analysis provided us the final model of the data.

**Conclusion:**

From the data set given, it is concluded that there is significant relationship between the all the variables. Some variables have high effect on malignancy while some of them has less effect. Two major statistically significant and major malignancy and benign defining variables are area and concave points mean.

**Recommendation:**

The operation performed above very basic in nature but the true modeling techniques that are implemented and also gives model as well as accuracy can also be implemented such as Random Forest technique, LDR, PCA algorithms. R provides in built functionality to perform that in R.

Later, we can compare the models in terms of which one has better accuracy.

**Final Notes:**

Machine learning modeling techniques are playing vital role in every sector such as automation, IT, healthcare. Research is going on in this field. Data modeling is not making model through couple of data sets. It needs decades of data and then the testing is done.

Prevention is better than Cure

Appendix:

References

Breast-cancer-Wisconsin, Retrieved from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

**Coding:**

**Box Plots Coding:**

attach(BreastCancer)

> head(BreastCancer)

# A tibble: 6 x 32

id diagnosis radius\_mean texture\_mean perimeter\_mean area\_mean

<dbl> <chr> <dbl> <dbl> <dbl> <dbl>

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

# ... with 26 more variables: smoothness\_mean <dbl>, compactness\_mean <dbl>,

# concavity\_mean <dbl>, concave\_mean <dbl>, symmetry\_mean <dbl>,

# fractal\_dimension\_mean <dbl>, radius\_se <dbl>, texture\_se <dbl>,

# perimeter\_se <dbl>, area\_se <dbl>, smoothness\_se <dbl>, compactness\_se <dbl>,

# concavity\_se <dbl>, `concave points\_se` <dbl>, symmetry\_se <dbl>,

# fractal\_dimension\_se <dbl>, radius\_worst <dbl>, texture\_worst <dbl>,

# perimeter\_worst <dbl>, area\_worst <dbl>, smoothness\_worst <dbl>,

# compactness\_worst <dbl>, concavity\_worst <dbl>, `concave points\_worst` <dbl>,

# symmetry\_worst <dbl>, fractal\_dimension\_worst <dbl>

summary(BreastCancer)

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\_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

as.factor(BreastCancer$diagnosis)

dia<-as.factor(BreastCancer$diagnosis)

plot(dia,main="Bar graph of type of tumor",col=c("blue","red"),ylab="Number of tumors")

> table(dia)

dia

B M

357 212

> plot(radius\_mean~dia)

> plot(radius\_mean~dia,main="Box Plot of Radius\_mean vs. tye of tumor",xlab="type of tumor",ylab="Radus\_mean",col=c("blue","red"))

> plot(texture\_mean~dia,main="Box Plot of texture\_mean vs. tye of tumor",xlab="type of tumor",ylab="texture\_mean",col=c("blue","red"))

> plot(perimeter\_mean~dia,main="Box Plot of perimeter\_mean vs. tye of tumor",xlab="type of tumor",ylab="perimeter\_mean",col=c("blue","red"))

> plot(area\_mean~dia,main="Box Plot of area\_mean vs. tye of tumor",xlab="type of tumor",ylab="area\_mean",col=c("blue","red"))

> plot(smoothness\_mean~dia,main="Box Plot of smoothness\_mean vs. tye of tumor",xlab="type of tumor",ylab="smoothness\_mean",col=c("blue","red"))

> plot(compactness\_mean~dia,main="Box Plot of compactness\_mean vs. tye of tumor",xlab="type of tumor",ylab=compactness\_mean",col=c("blue","red"))

Error: unexpected string constant in "plot(compactness\_mean~dia,main="Box Plot of compactness\_mean vs. tye of tumor",xlab="type of tumor",ylab=compactness\_mean",col=c(""

> plot(compactness\_mean~dia,main="Box Plot of compactness\_mean vs. tye of tumor",xlab="type of tumor",ylab="compactness\_mean",col=c("blue","red"))

> plot(concave\_mean~dia,main="Box Plot of concave\_mean vs. tye of tumor",xlab="type of tumor",ylab="concave\_mean",col=c("blue","red"))

> plot(symmetry\_mean~dia,main="Box Plot of symmetry\_mean vs. tye of tumor",xlab="type of tumor",ylab="symmetry\_mean",col=c("blue","red"))

> plot(fractal\_dimension\_mean~dia,main="Box Plot of fractal\_dimension\_mean vs. tye of tumor",xlab="type of tumor",ylab="fractal\_dimension\_mean",col=c("blue","red"))

**ANOVA Coding:**

aov(area\_mean~diagnosis)

Call:

aov(formula = area\_mean ~ diagnosis)

Terms:

diagnosis Residuals

Sum of Squares 35358547 34984592

Deg. of Freedom 1 567

Residual standard error: 248.3973

Estimated effects may be unbalanced

> summary(aov(area\_mean~diagnosis))

Df Sum Sq Mean Sq F value Pr(>F)

diagnosis 1 35358547 35358547 573.1 <2e-16 \*\*\*

Residuals 567 34984592 61701

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(aov(smoothness\_mean~diagnosis))

Df Sum Sq Mean Sq F value Pr(>F)

diagnosis 1 0.01444 0.014444 83.65 <2e-16 \*\*\*

Residuals 567 0.09791 0.000173

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(aov(concavity\_mean~diagnosis))

Df Sum Sq Mean Sq F value Pr(>F)

diagnosis 1 1.750 1.7504 533.8 <2e-16 \*\*\*

Residuals 567 1.859 0.0033

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**T Tests:**

t.test(area\_mean~diagnosis,var.equal=F,data=bc)

Welch Two Sample t-test

data: area\_mean by diagnosis

t = -19.521, df = 243.02, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-568.0747 -463.9398

sample estimates:

mean in group 0 mean in group 1

464.4471 980.4543

t.test(smoothness\_mean~diagnosis,alternative="less",var.equal=F,data=bc)

Welch Two Sample t-test

data: smoothness\_mean by diagnosis

t = -9.096, df = 460.49, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.008384321

sample estimates:

mean in group 0 mean in group 1

0.09258111 0.10282086

> t.test(concavity\_mean~diagnosis,alternative="less",var.equal=F,data=bc)

Welch Two Sample t-test

data: concavity\_mean by diagnosis

t = -20.003, df = 294.81, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.1046091

sample estimates:

mean in group 0 mean in group 1

0.04672025 0.16073448

**Regression Coding:**

> library(readxl)

> bc <- read\_excel("C:/Users/adarsh/Desktop/bc.xlsx")

> View(bc)

> attach(bc)

The following objects are masked from BreastCancer:

area\_mean, compactness\_mean, concave\_mean, concavity\_mean, diagnosis,

perimeter\_mean, radius\_mean, smoothness\_mean

model<-glm(formula=diagnosis~.,family=binomial("logit"),data=bc)

Warning message:

glm.fit: fitted probabilities numerically 0 or 1 occurred

> vif(model)

radius\_mean compactness\_mean perimeter\_mean area\_mean smoothness\_mean concavity\_mean concave\_mean

699.579958 8.724303 556.755319 98.771389 3.471871 4.642638 5.867357

> sqrt(vif(model))

radius\_mean compactness\_mean perimeter\_mean area\_mean smoothness\_mean concavity\_mean concave\_mean

26.449574 2.953693 23.595663 9.938380 1.863296 2.154678 2.422263

> bc$radius\_mean<-NULL

> model<-glm(formula=diagnosis~.,family=binomial("logit"),data=bc)

Warning message:

glm.fit: fitted probabilities numerically 0 or 1 occurred

> sqrt(vif(model))

compactness\_mean perimeter\_mean area\_mean smoothness\_mean concavity\_mean concave\_mean

2.316867 8.930729 8.830367 1.831089 2.101856 2.426046

> bc$perimeter\_mean<-NULL

> model<-glm(formula=diagnosis~.,family=binomial("logit"),data=bc)

> sqrt(vif(model))

compactness\_mean area\_mean smoothness\_mean concavity\_mean concave\_mean

1.897305 1.470271 1.792477 1.979709 2.335866

> bc$concave\_mean<-NULL

> model<-glm(formula=diagnosis~.,family=binomial("logit"),data=bc)

> sqrt(vif(model))

compactness\_mean area\_mean smoothness\_mean concavity\_mean

1.890527 1.131664 1.417999 1.602894

> model<-glm(formula=diagnosis~.,family=binomial("logit"),data=bc)

> summary(model)

Call:

glm(formula = diagnosis ~ ., family = binomial("logit"), data = bc)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.4500 -0.2767 -0.1137 0.0462 2.7564

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -15.782485 2.001965 -7.883 0.00000000000000318 \*\*\*

compactness\_mean -3.310324 7.756634 -0.427 0.669545

area\_mean 0.011697 0.001346 8.691 < 0.0000000000000002 \*\*\*

smoothness\_mean 66.622466 17.789401 3.745 0.000180 \*\*\*

concavity\_mean 19.306031 5.132614 3.761 0.000169 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 749.33 on 568 degrees of freedom

Residual deviance: 222.87 on 564 degrees of freedom

AIC: 232.87

Number of Fisher Scoring iterations: 7

> bc$compactness\_mean<-NULL

> model<-glm(formula=diagnosis~.,family=binomial("logit"),data=bc)

> summary(model)

Call:

glm(formula = diagnosis ~ ., family = binomial("logit"), data = bc)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.4417 -0.2761 -0.1127 0.0475 2.7821

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -15.598075 1.941956 -8.032 0.000000000000000958 \*\*\*

area\_mean 0.011675 0.001338 8.727 < 0.0000000000000002 \*\*\*

smoothness\_mean 62.606008 15.004352 4.173 0.000030124464988908 \*\*\*

concavity\_mean 17.716603 3.425409 5.172 0.000000231461712562 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 749.33 on 568 degrees of freedom

Residual deviance: 223.05 on 565 degrees of freedom

AIC: 231.05

Number of Fisher Scoring iterations: 7

**Checking For R square:**

**1-residula deviance/Null deviance**

|  |
| --- |
| 223.05/749.33  [1] 0.2976659  > 1-0.2976  [1] 0.7024 |
|  |
| |  | | --- | | > | |

> layout(matric(c(1,2,3,4)2,2))

Error: unexpected numeric constant in "layout(matric(c(1,2,3,4)2"

> layout(matrix(c(1,2,3,4)2,2))

Error: unexpected numeric constant in "layout(matrix(c(1,2,3,4)2"

> layout(matrix(c(1,2,3,4),2,2))

> plot(fit)

> plot(model)

**Co-Relation Coding:**

> M<-corr(bc)

Error in corr(bc) : could not find function "corr"

> M<-cor(bc)

> col1 <- colorRampPalette(c("#7F0000","red","#FF7F00","yellow","white",

+ "cyan", "#007FFF", "blue","#00007F"))

>

> col2 <- colorRampPalette(c("#67001F", "#B2182B", "#D6604D", "#F4A582", "#FDDBC7",

+ "#FFFFFF", "#D1E5F0", "#92C5DE", "#4393C3", "#2166AC", "#053061"))

>

> col3 <- colorRampPalette(c("red", "white", "blue"))

>

> col4 <- colorRampPalette(c("#7F0000","red","#FF7F00","yellow","#7FFF7F",

+ "cyan", "#007FFF", "blue","#00007F"))

>

> wb <- c("white","black")

>

>

> par(ask = TRUE)

>

> corrplot(M, method="number", col="black", addcolorlabel="no")

Hit <Return> to see next plot:

Warning messages:

1: In text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt = tl.srt, :

"addcolorlabel" is not a graphical parameter

2: In text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col = tl.col, :

"addcolorlabel" is not a graphical parameter

3: In title(title, ...) : "addcolorlabel" is not a graphical parameter

> corrplot(M, method="number")

Hit <Return> to see next plot:

Warning messages:

1: In doTryCatch(return(expr), name, parentenv, handler) :

"addcolorlabel" is not a graphical parameter

2: In doTryCatch(return(expr), name, parentenv, handler) :

"addcolorlabel" is not a graphical parameter

3: In doTryCatch(return(expr), name, parentenv, handler) :

"addcolorlabel" is not a graphical parameter

4: In doTryCatch(return(expr), name, parentenv, handler) :

"addcolorlabel" is not a graphical parameter

5: In doTryCatch(return(expr), name, parentenv, handler) :

"addcolorlabel" is not a graphical parameter

6: In doTryCatch(return(expr), name, parentenv, handler) :

"addcolorlabel" is not a graphical parameter

> corrplot(M)

Hit <Return> to see next plot:

> corrplot(M, order ="AOE")

Hit <Return> to see next plot:

> corrplot(M, order ="AOE", addCoef.col="grey")

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col1(20), cl.length=21,addCoef.col="grey")

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col1(10),addCoef.col="grey")

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col2(200))

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col2(200),addCoef.col="grey")

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col2(20), cl.length=21,addCoef.col="grey")

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col2(10),addCoef.col="grey")

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col3(100))

Hit <Return> to see next plot:

> corrplot(M, order="AOE", col=col3(10))

Hit <Return> to see next plot:

>

>

>

> corrplot(M, method="color", col=col1(20), cl.length=21,order = "AOE", addCoef.col="grey")

Hit <Return> to see next plot:

>

> if(TRUE){

+ corrplot(M, method="square", col=col2(200),order = "AOE")

+

+ corrplot(M, method="ellipse", col=col1(200),order = "AOE")

+

+

+ corrplot(M, method="shade", col=col3(20),order = "AOE")

+

+ corrplot(M, method="pie", order = "AOE")

+

+

+ corrplot(M, col = wb, order="AOE", outline=TRUE, addcolorlabel="no")

+ corrplot(M, col = wb, bg="gold2", order="AOE", addcolorlabel="no")

+ }

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot: