PROJECT REPORT

ON

BMI PREDICTION

Submitted as a part of Curriculum of
Bachelor of Technology
In

Computer Science Engineering

Under the guidance of

Mr. D. KRISHNA

Associate professor

Department of Computer Science and Engineering

ACE Engineering College (An Autonomous Institution)

NBA ACCREDITED B. TECH COURSES: EEE, ECE, CSE & MECH

Ankushapur (V), Ghatkesar(M), Medchal.Dist.–501301

(Affiliated to Jawaharlal Nehru Technological University Hyderabad 2021-2025

ACE
Engineering College
An AUTONOMOUS Institution
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project work entitled BMI PREDICTION is being submitted by B.Nagaraju (21AG1A05D8) , B.Sree Kavya Sudha (21AG1A05D9) D.Sai Kiran (21AG1A05E4) , B.Manipal (21AG1A05D7) , R.Sahithi (21AG1A05H4) , R.Meri Prasanna (21AG1A05H5) , R.Sri Mallika (21AG1A05H6) , D.Bhanu Prasad (21AG1A05E3) , D.Sai Varshith (21AG1A05E5) , R.Archana (22AG5A0517) as a part of Curriculum of Degree of Bachelor of Technology in Computer Science and Engineering to the ACE Engineering College during the academic year 2021- 2025 is a record of bonafide work carried out by them under our guidance and supervision.

Internal Guide	**Head of Department**
Mr.D.Krishna	**Dr .M.V.VIJAYA SARADHI**
Associate Professor	**Professor and Head of the Dept CSE**

ACKNOWLEDGEMENT

We would like to express our gratitude to all the people behind the screen who have helped us to transform an idea into a real time application. We would like to express our heart-felt gratitude to our parents without whom we would not have been privileged to achieve and fulfill our dreams.

A special thanks to our Secretary, **Prof. Y. V. GOPALA KRISHNA MURTHY**, for having founded such an esteemed institution. We are also grateful to our beloved principal, **Dr. B. L. RAJU** for permitting us to carry out this project. We profoundly thank **Dr.M.V.VIJAYA SARADHI**, Head of the Department of Computer Science & Engineering.

We are very thankful to our guide **D.KRISHNA**, **Associate Professor** who has been an excellent and also given continuous support for the completion of our project work.
The satisfaction and euphoria the accompany the successful completion of the task would be great, but incomplete without the mention of the people who made it possible, whose guidance and encouragement crown all the efforts with success. In this context, we would like to thank all the other staff members, both teaching and non-teaching, which have extended their timely help and easier our task.

B. Nagaraju (21AG1A05D8)
D. Sai Kiran (21AG1A05E4)
B.Manipal (21AG1A05D7)
D.Bhanu Prasad (21AG1A05E3)
D. Sai Varshith (21AG1A05E5)

B. Sree Kavya Sudha (21AG1A05D9)
R. Sahithi (21AG1A05H4)
R. Meri Prasanna (21AG1A05H5)
R. Sri Mallika (21AG1A05H6)
R. Archana (22AG5A0517)

DECLARATION

We hereby declare that project entitled “**BMI Prediction**” submitted as a part of Curriculum of Bachelor of Technology in Computer Science and Engineering. This dissertation isour original work and the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles and no part of it has been published or sent for the publication at the time of submission.

B. Nagaraju (21AG1A05D8)

B. Sree Kavya Sudha (21AG1A05D9)

D. Sai Kiran (21AG1A05E4)

R. Sahithi(21AG1A05H4)

B. Manipal (21AG1A05D7)

R. Meri Prasanna (21AG1A05H5)

D. Bhanu Prasad (21AG1A05E3)

R. Sri Mallika (21AG1A05H6)

D. Sai Varshith (21AG1A05E5)

R. Archana (22AG5A0517)

Place: Hyderabad

Date: 03/02/2024

ABSTRACT

Body Mass Index(BMI): The Body Mass Index (BMI) is a statistical measurement that uses your weight and height to estimate your body fat. It is a screening tool, not a diagnostic tool, for identifying potential health risks associated with weight. Traditional BMI calculations might not accurately reflect health risks for individuals across different age groups. This study aimed to develop an age-adjusted BMI prediction model using linear regression in **R**.

This project utilizes linear regression in R to develop a model predicting Body Mass Index (BMI) based on relevant parameters. Data is collected, preprocessed, and analyzed to determine significant predictors like height, weight, and potentially lifestyle factors. **R's lm()** function builds the model, which is then evaluated through performance metrics and visualization. The final abstract will report these findings, including the impact of significant predictors on BMI and the model's potential for health assessments and public health initiatives. Further research may refine the model and explore additional predictor.

INDEX

1.INTRODUCTION

What is the body mass index (BMI)?

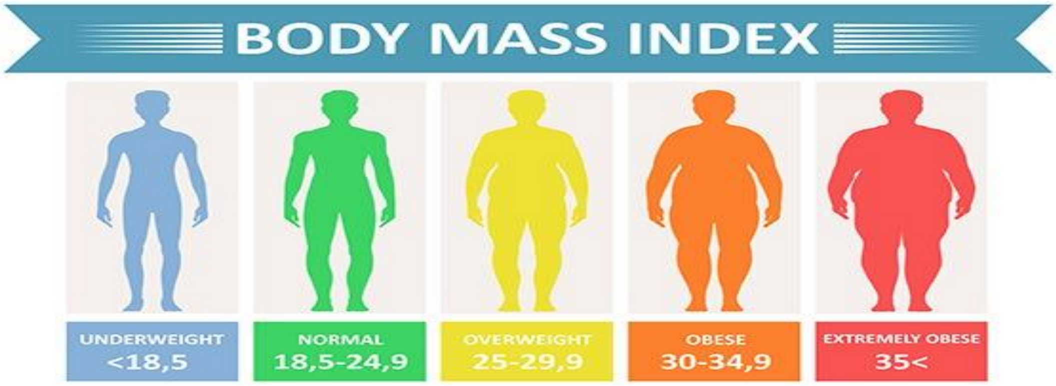
The body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy. The BMI calculation divides an adult's weight in kilograms by their height in metres squared.

Formula: weight (kg) / [height (m)]^2

BMI Ranges

For most adults, an ideal BMI is in the 18.5 to 24.9 range. For children and young people aged 2 to 18, the BMI calculation takes into account age and gender as well as height and weight. If your BMI is:

- below 18.5 — you're in the underweight range
- between 18.5 and 24.9 — you're in the healthy weight range
- between 25 and 29.9 — you're in the overweight range
- between 30 and 39.9 — you're in the obese range

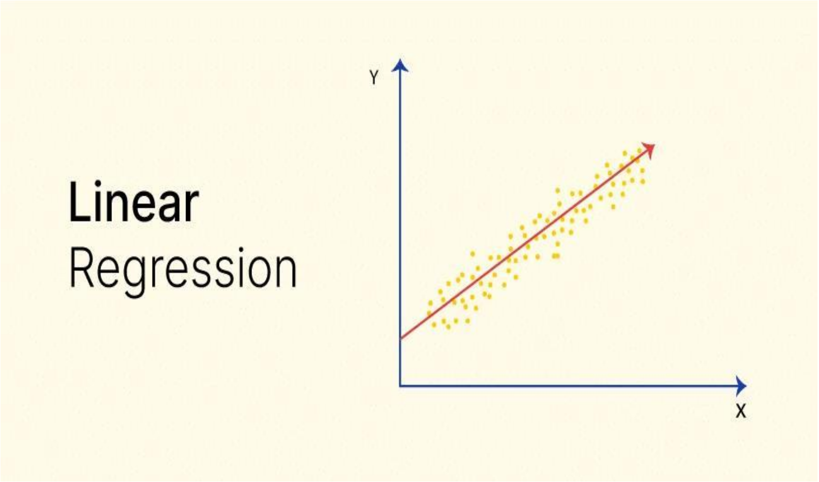


1.1 WHAT IS LINEAR REGRESSION?

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable)

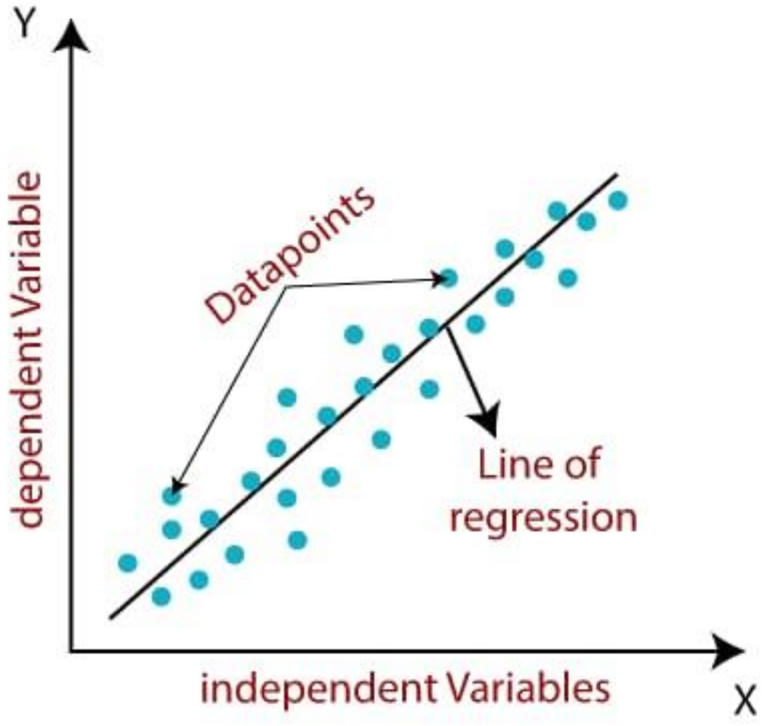


Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Why Linear Regression is Important?

The interpretability of linear regression is a notable strength. The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.



TYPES:-

There are mainly two types in linear regression:

- **Simple Linear Regression:** A linear regression model with one independent and one dependent variable.

Equation :- y = mx + b

- **Multiple Linear Regression:** A linear regression model with more than one independent variable and one dependent variable.

Equation :- y = b + m1x1 + m2x2 + ... + mnxn

KEY ASSUMPTIONS:

Assumptions to be considered for success with linear-regression analysis:

1.The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.

2.Use a scatterplot to find out quickly if there is a linear relationship between those two variables.

3.The observations should be independent of each other (that is, there should be no dependency) i.e No Multicolllinearity.

4.Your data should have no significant outliers.

5.Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.

6.The residuals (errors) of the best-fit regression line follow normal distribution.

1.2 Correlation Matrix

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

A correlation matrix consists of rows and columns that show the variables. Each cell in a table contains the correlation coefficient.

It’s very useful for regression techniques like simple linear regression, multiple linear regression and lasso regression models. In the regression technique, we have several independent variables, and based on that, we are predicting the dependent variable.

A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where 1 is considered a strong relationship between variables, 0 a neutral relationship and -1 a not strong relationship. It’s most commonly used in building regression models.

1.3 Regression Model

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values.

A regression model provides a function that describes the relationship between one or more independent variables and a response, dependent, or target variable.
A linear regression model is a statistical technique that describes the relationship between **a dependent variable (what you're trying to predict) and one or more independent variables (what you think might influence it)**. It assumes this relationship can be represented by a straight line.A regression model determines a relationship between an independent variable and a dependent variable, by providing a function. Formulating a regression analysis helps you predict the effects of the independent variable on the dependent one.

Example: we can say that age and height can be described using a linear regression model. Since a person’s height increases as age increases, they have a linear relationship.

Regression models are commonly used as statistical proof of claims regarding everyday facts.

2.DATASET

This dataset is open-source data collected from ScienceDirect under a Creative Commons license. This dataset had collected some information of residents in Mexico, Peru and Colombia about their lifestyle. The original file type is off so I transformed it into a CSV file then did some clean processes and transformations.

Dataset link: https://www.kaggle.com/datasets/mandysia/obesity-dataset-cleaned-and-data-sinthetic

The objective of the dataset is to predict Body Mass Index (BMI) of the Person.

Target variable or Dependent variable:

- ✓ BMI: Body mass index (weight in kg/(height in m)^2).

Predictors or Independent variable:

- ✓ ID: Indicates the id of the Person.
- ✓ Age: Age of the Person.
- ✓ Gender: Gender of the Person.
- ✓ Height: Height of the Person(meters).
- ✓ Weight: Weight of the Person(kilogram).
- ✓ Result: Indicates the Obesity Level

This dataset contains 2111 rows and 7 columns.

3.Exploratory Analysis

Installing Required Packages into R:

```
install.packages("tidyverse")
install.packages("base")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("broom")
install.packages("ggpubr")
install.packages("tidyverse")
install.packages("catTools")
install.packages("predict3d")
install.packages("corrplot")
install.packages("car")
install.packages("Metrics")
```

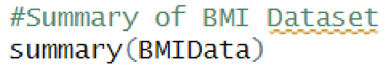
Loading Required Packages into R:

```
library(tidyverse) #to better transform and present Data
library(base) #Contains the basic functions
library(ggplot2) #for Data Visualization
library(dplyr) #to make data manipulation
library(broom) #takes messy output of built in functions
library(ggpubr) #for better creation of beautiful ggplot2 based graphs
library(caTools) #for basic utility function
library(predict3d) #to predicts plots
library(corrplot) #to visualize correlation matrix
library(car) #provides tools for regression analysis
library(Metrics) #Provides Metrics for Model
```

Importing Dataset into R:

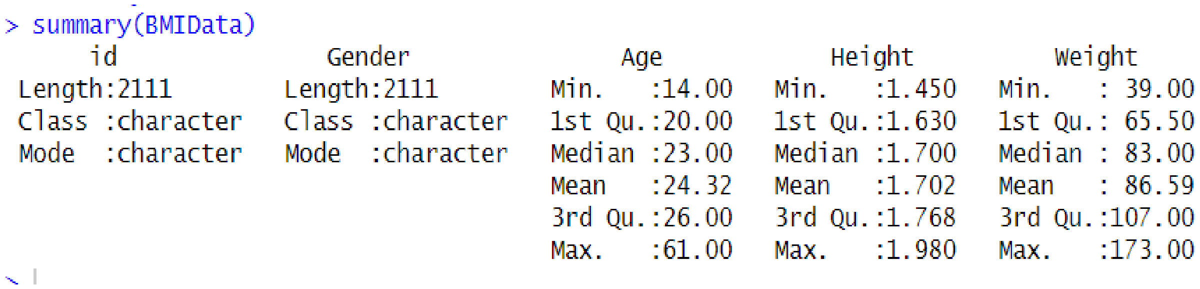
```
#Reading the Dataset and storing it in BMIData Variable
BMIData<-read.csv("BMI.csv")
```

Summary of the Dataset: Getting the Complete Summary of the Dataset



```
#Summary of BMI Dataset
summary(BMIData)
```

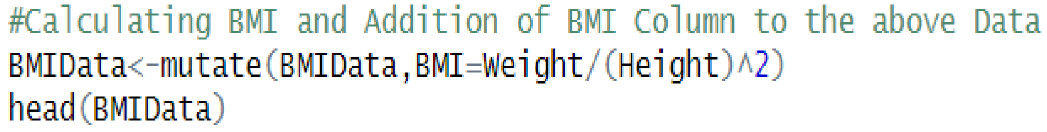
Output:



```
> summary(BMIData)
      id               Gender               Age            Height          Weight      
 Length:2111        Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00  
 Class :character   Class :character   1st Qu.:20.00   1st Qu.:1.630   1st Qu.: 65.50  
 Mode  :character   Mode  :character   Median :23.00   Median :1.700   Median : 83.00  
                                       Mean   :24.32   Mean   :1.702   Mean   : 86.59  
                                       3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.00  
                                       Max.   :61.00   Max.   :1.980   Max.   :173.00  
```

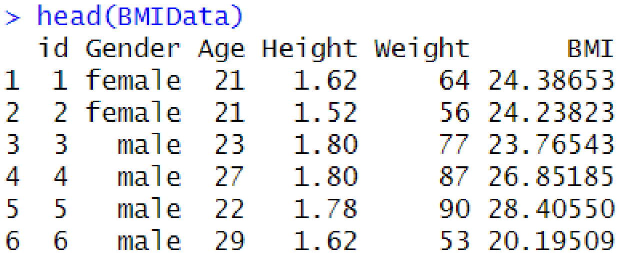
Calculating the BMI And Obesity Level : Calculating BMI and Obesity level and adding these two columns to the Dataset explicitly.

Code:

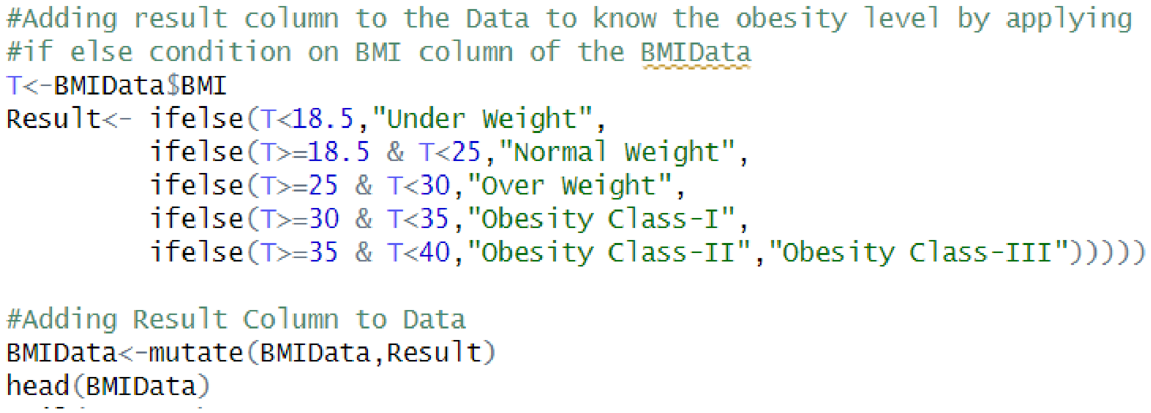


```
#Calculating BMI and Addition of BMI Column to the above Data
BMIData<-mutate(BMIData,BMI=Weight/(Height)^2)
head(BMIData)
```

Output:



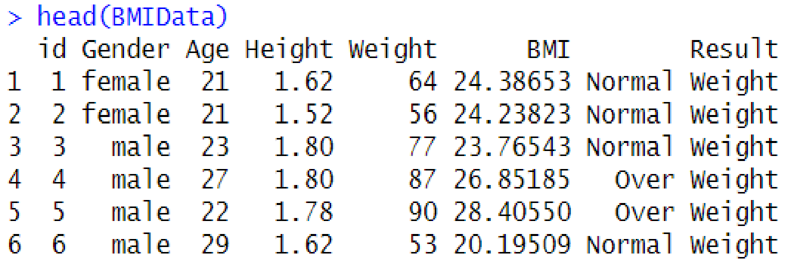
```
> head(BMIData)
  id Gender Age Height Weight      BMI
1  1 female  21   1.62     64 24.38653
2  2 female  21   1.52     56 24.23823
3  3   male  23   1.80     77 23.76543
4  4   male  27   1.80     87 26.85185
5  5   male  22   1.78     90 28.40550
6  6   male  29   1.62     53 20.19509
```



```
#Adding result column to the Data to know the obesity level by applying
#if else condition on BMI column of the BMIData
T<-BMIData$BMI
Result<- ifelse(T<18.5,"Under Weight",
         ifelse(T>=18.5 & T<25,"Normal Weight",
         ifelse(T>=25 & T<30,"Over Weight",
         ifelse(T>=30 & T<35,"Obesity Class-I",
         ifelse(T>=35 & T<40,"Obesity Class-II","Obesity Class-III")))))

#Adding Result Column to Data
BMIData<-mutate(BMIData,Result)
head(BMIData)
```

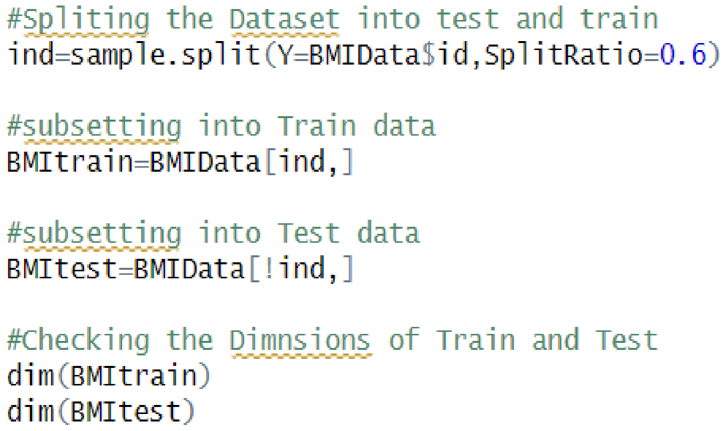
Output:



```
> head(BMIData)
  id Gender Age Height Weight      BMI        Result
1  1 female  21   1.62     64 24.38653 Normal Weight
2  2 female  21   1.52     56 24.23823 Normal Weight
3  3   male  23   1.80     77 23.76543 Normal Weight
4  4   male  27   1.80     87 26.85185   Over Weight
5  5   male  22   1.78     90 28.40550   Over Weight
6  6   male  29   1.62     53 20.19509 Normal Weight
```

Splitting the Dataset into Test and Train Data : Here we are splitting the Dataset into Test and Train Data. Train Data is used to train the Model and Test Data is used to test the model. We are splitting the Dataset in the ratio of 0.6.

Code:



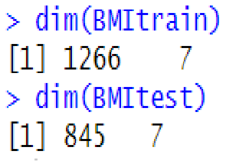
```
#Spliting the Dataset into test and train
ind=sample.split(Y=BMIData$id,SplitRatio=0.6)

#subsetting into Train data
BMItrain=BMIData[ind,]

#subsetting into Test data
BMItest=BMIData[!ind,]

#Checking the Dimnsions of Train and Test
dim(BMItrain)
dim(BMItest)
```

Output:



```
> dim(BMItrain)
[1] 1266    7
> dim(BMItest)
[1] 845   7
```

4.Checking the Assumptions

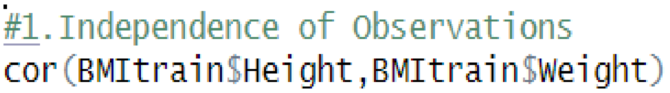
Make sure our data meets Assumptions:

We can use R to check that our data meet the four main assumptions
for linear regression.

4.1 Independence of Observations

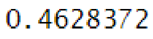
We have one Independent Variable. But because of the more than one dependent Variables we need to make sure that the correlation between the dependent variables is not high.

Code:-



```
#1.Independence of Observations
cor(BMItrain$Height,BMItrain$Weight)
```

Output:-



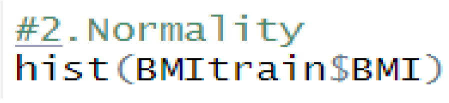
```
0.4628372
```

When we run this code we have the correlation value as 0.4628372. The Correlation between Height and Age is 46.2%.Which is not high. So we can include both the values in our model.

4.2 Normality

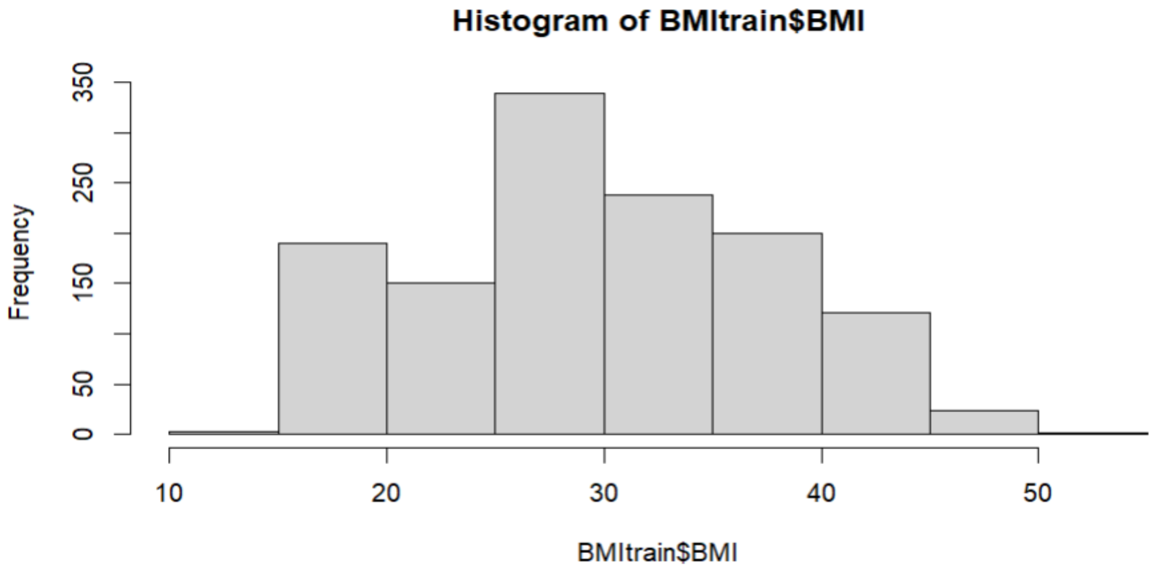
To Check the whether the Dependent is Normal Distribution We use hist() Function in R.

Code:



```
#2.Normality
hist(BMItrain$BMI)
```

Output:



The Observations are Roughly bell-shaped (more observations in the middle of the Distribution, fewer observations in the tail). So we can we move with Regression.

4.3 Multi Collinearity

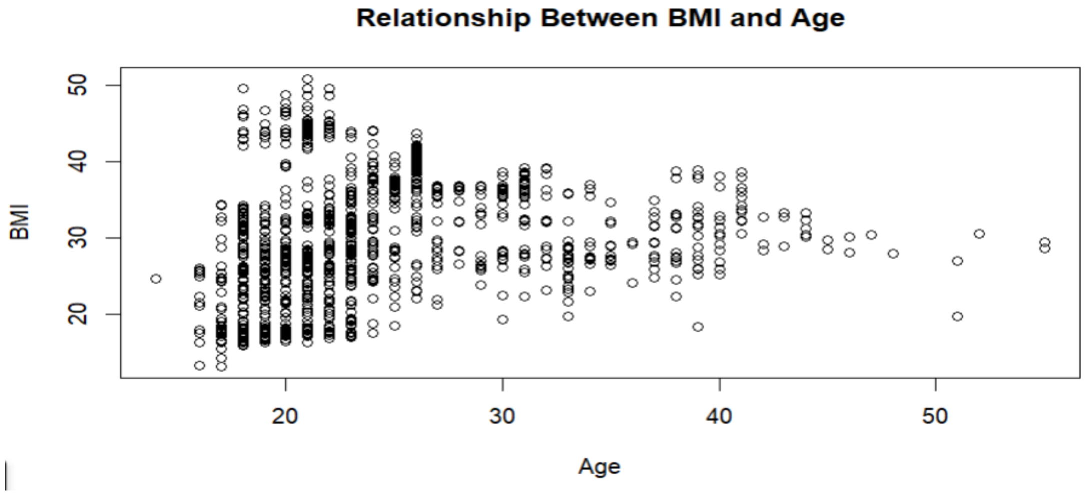
4.3.1 Linearity:-

We can check this using three scatter plots, One for Age, One for Height and another for Weight.

Code:- Relation Between Age and BMI

```
plot(BMI~Age,data=BMItrain,main="Relationship Between BMI and Age")
```

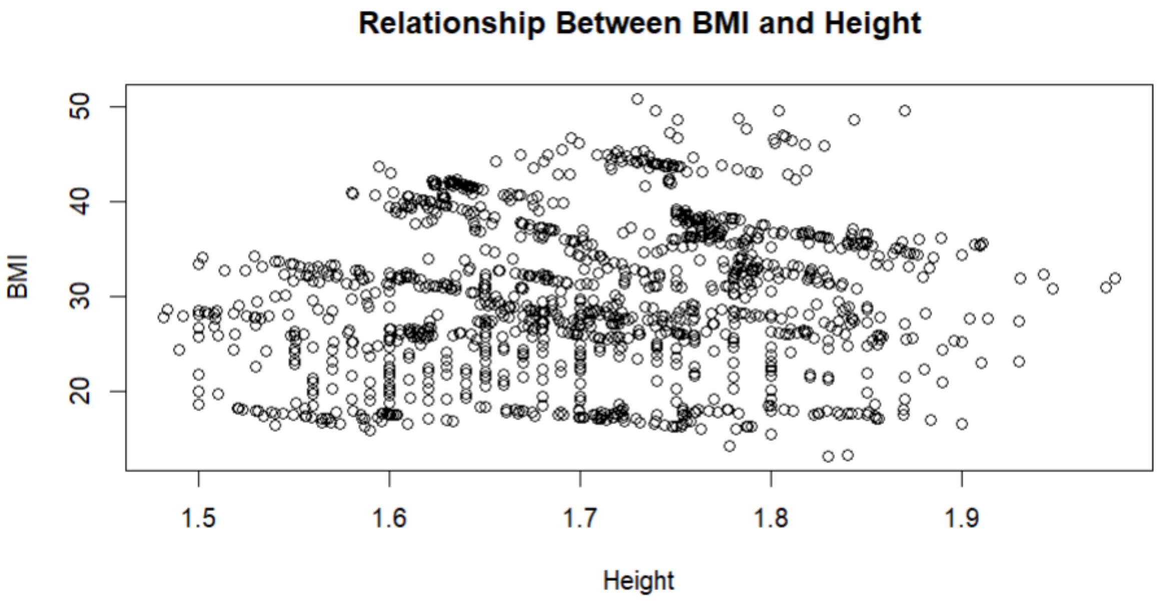
Output:-



Code: Relation Between BMI and Height

```
plot(BMI~Height,data=BMItrain,main="Relationship Between BMI and Height")
```

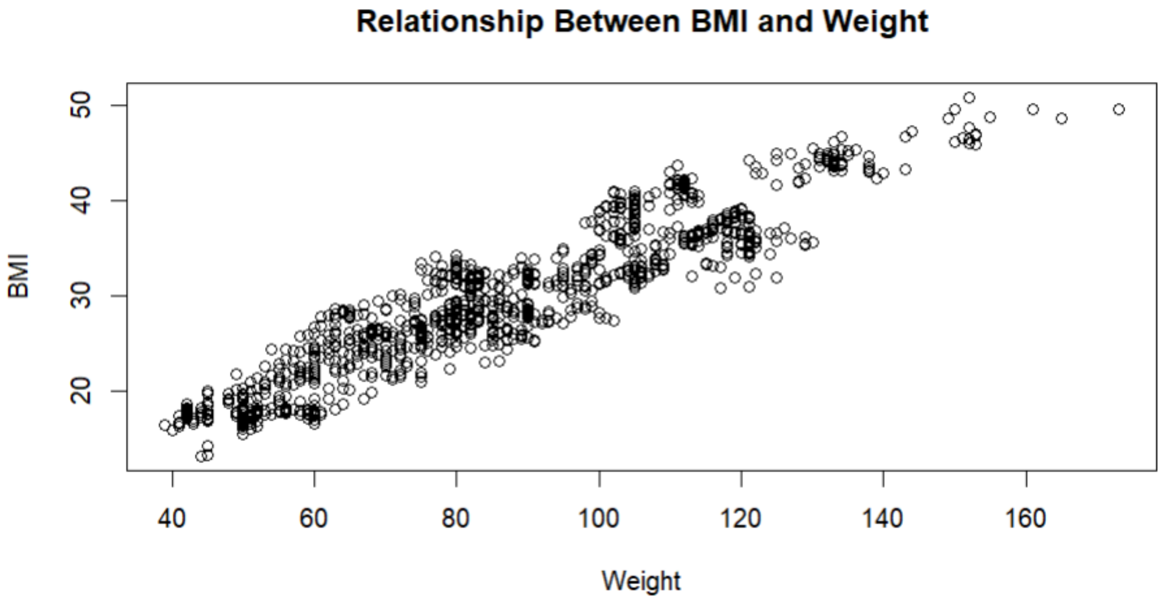
Output:-



Code:- Relation Between BMI And Weight

```
plot(BMI~Weight,data=BMItrain,main="Relationship Between BMI and Weight")
```

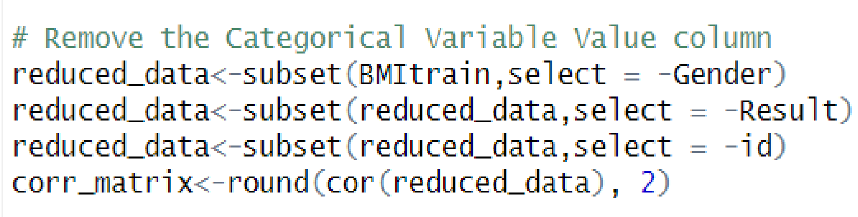
Output:-



4.3.2 Multi Collinearity:

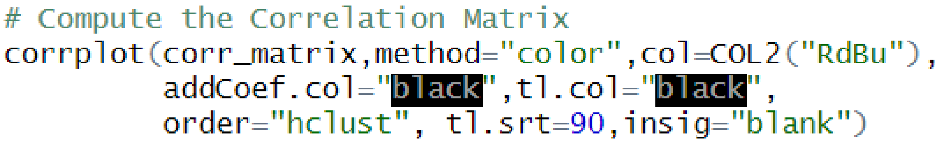
We have to remove the Gender, Result and id column in order to get the correlation matrix. Because they are Categorical variables.

Code:-



```
# Remove the Categorical Variable Value column
reduced_data<-subset(BMItrain,select = -Gender)
reduced_data<-subset(reduced_data,select = -Result)
reduced_data<-subset(reduced_data,select = -id)
corr_matrix<-round(cor(reduced_data), 2)
```

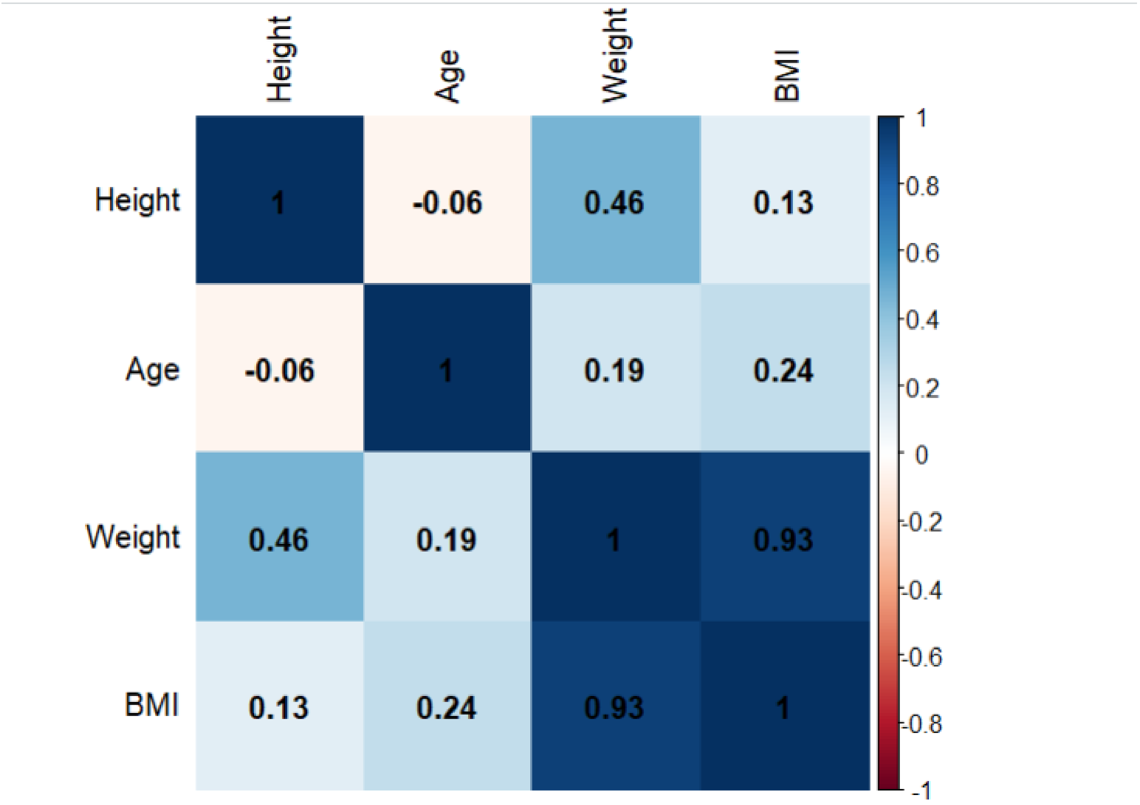
Code:- To compute the Correlation Matrix



```
# Compute the Correlation Matrix
corrplot(corr_matrix,method="color",col=COL2("RdBu"),
           addCoef.col="black",tl.col="black",
           order="hclust", tl.srt=90,insig="blank")
```

Output:-

Correlation Matrix



	Height	Age	Weight	BMI
Height	1	-0.06	0.46	0.13
Age	-0.06	1	0.19	0.24
Weight	0.46	0.19	1	0.93
BMI	0.13	0.24	0.93	1

With this Correlation Matrix we can Build our Regression Model.

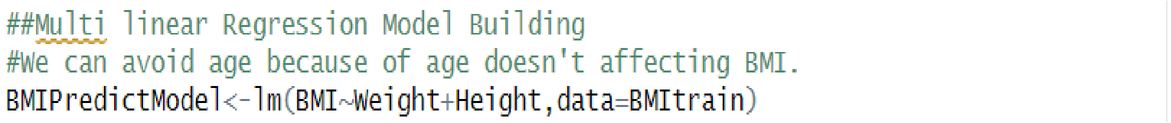
5.Model Building

5.1 Regression Model

Now we have determined that our Data meets Assumptions. Now we can perform the Linear Regression Analysis to determine the relationship between the Independent Variable and Dependent Variables.

We first fit the BMI as a Dependent variable and Independent variables are Weight and Height.

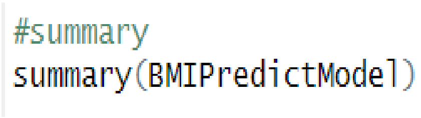
Code:-



```
##Multi linear Regression Model Building
#We can avoid age because of age doesn't affecting BMI.
BMIPredictModel<-lm(BMI~Weight+Height,data=BMItrain)
```

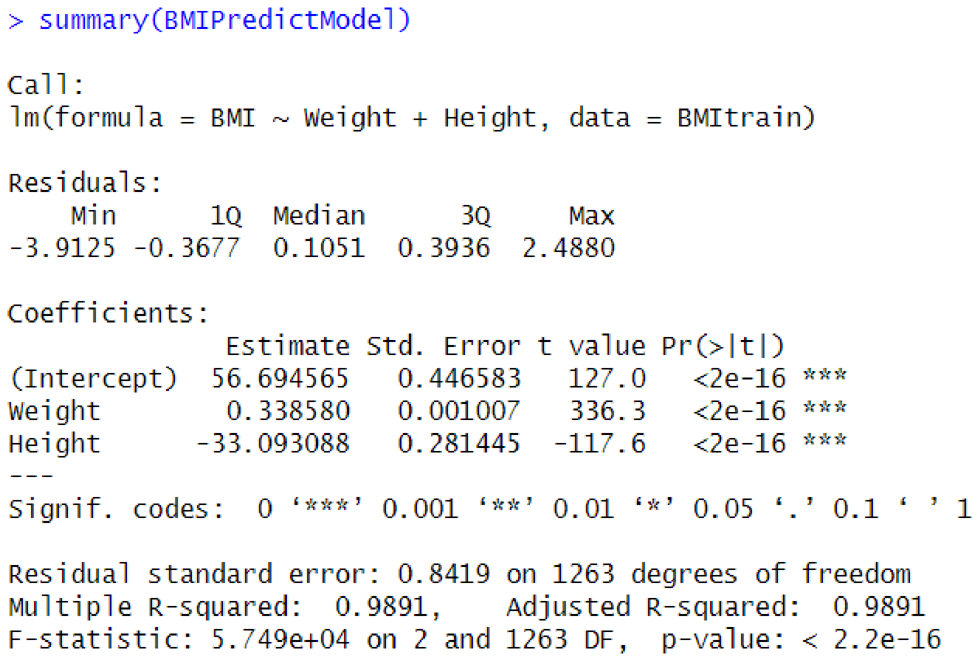
This line of code will create and Regression Model which predicts the BMI

Code:-To get the Summary of the Model



```
#summary
summary(BMIPredictModel)
```

Output:-



```
> summary(BMIPredictModel)

Call:
lm(formula = BMI ~ Weight + Height, data = BMItrain)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9125 -0.3677  0.1051  0.3936  2.4880 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  56.694565   0.446583   127.0   <2e-16 ***
Weight        0.338580   0.001007   336.3   <2e-16 ***
Height      -33.093088   0.281445  -117.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8419 on 1263 degrees of freedom
Multiple R-squared:  0.9891,	Adjusted R-squared:  0.9891 
F-statistic: 5.749e+04 on 2 and 1263 DF,  p-value: < 2.2e-16
```

In the above summary, the significant relationships between the Weight and BMI and the Height and BMI were found to be p<0.001.

Multiple R-Squared is 0.9891 which means 98% indicates the strength of the Linear Relationship.

Residual Standard Error Calculates the average distance that the observed values fall from the Regression Line.

Finally we conclude that, The BMI value is increased by 0.33 for every increase in one kilogram of weight. The BMI values is decreased by 33 for every increase in 1 meter of Height.

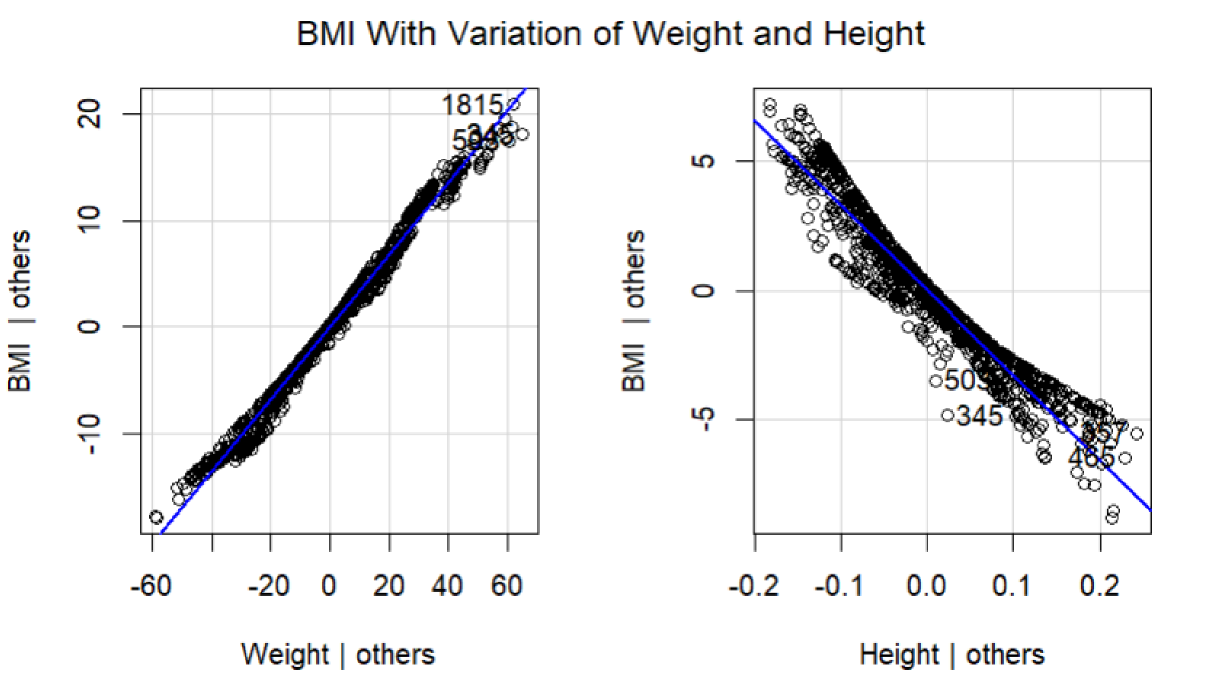
Added Variable Plots For the Regression Model:-

AV plots are used in Multiple Linear Regression. It is used to plot the Dependent Variable with Independent Variable Separately.

Code:-

```
#Added Variable plots for the BMIPrediction Regression Model
avPlots(BMIPredictModel,main="BMI With Variation of Weight and Height")
```

Output:-



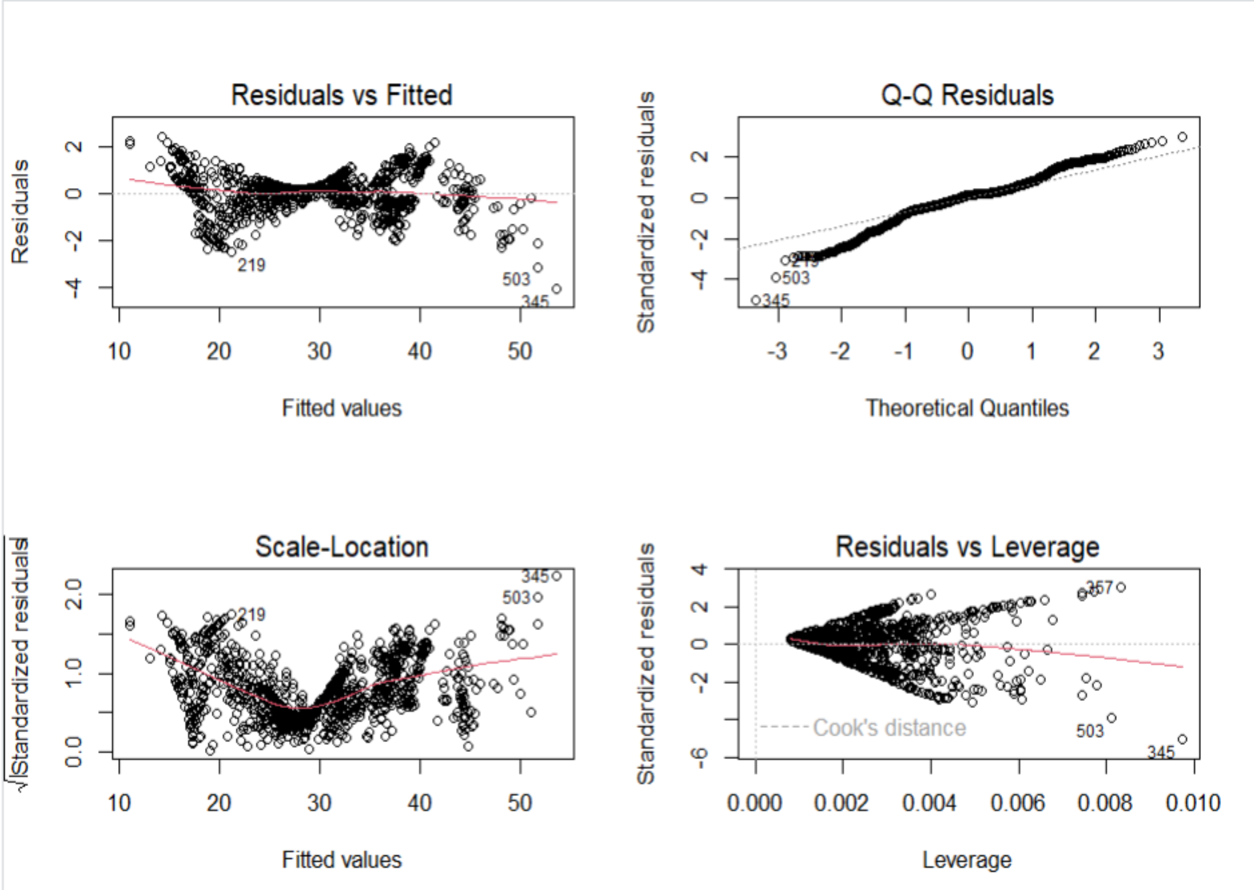
Check for the Homoscedasticity of The Model:-

Before proceeding with data visualization, we should make sure that our models fit the homoscedasticity assumption of the linear model.

Code:-

```
#Checking for the Homoscedasicity
par(mfrow=c(2,2))
plot(BMIPredictModel)
par(mfrow=c(1,1))
```

Output:-



A Horizontal line in Residuals vs Fitted indicates Good Model.

Q-Q indicates that residuals follows the dotted straight line it is good for model.

Scale-Location indicates the Homogeneity of the Variance.

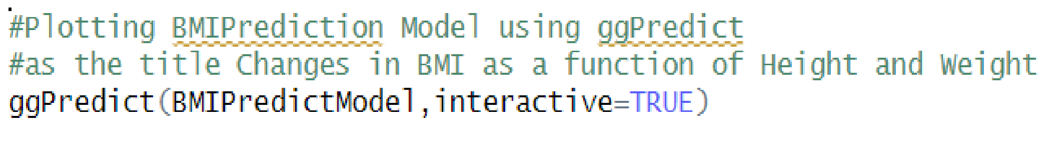
Residuals vs Leverage indicates outliers and hight leverage points.

5.2 Data Visualization:-Visualize the Results with Graph

Plotting the Regression Model:-

In this plotting the Regression Model using the ggPredict(), which is used only for multiple linear regression Model. Which plots the Dependent Variable as a function of independent variables.

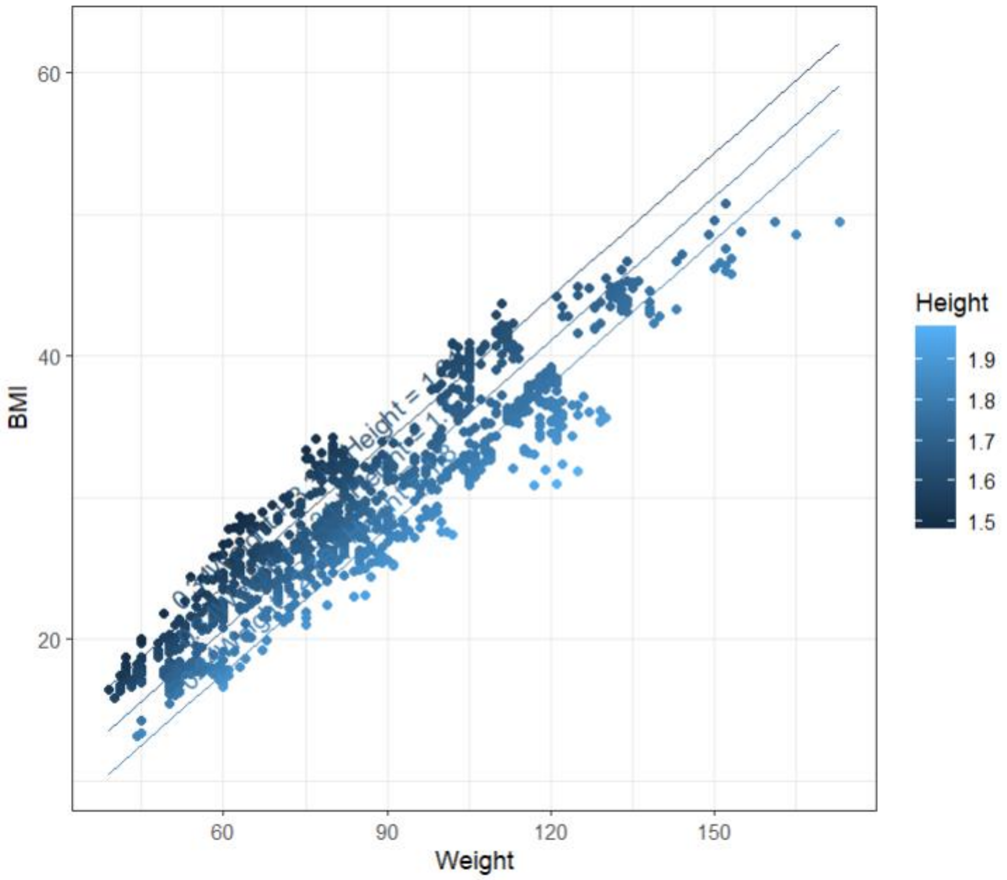
Code:



```
#Plotting BMIPrediction Model using ggPredict
#as the title Changes in BMI as a function of Height and Weight
ggPredict(BMIPredictModel,interactive=TRUE)
```

Output:-

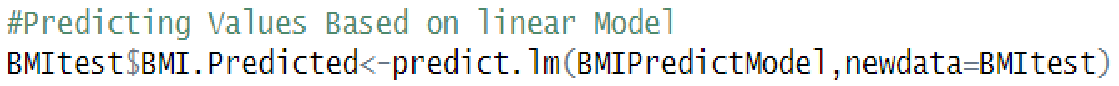
Changes in BMI As a Function of Weight and Height



Predicting the Model:-

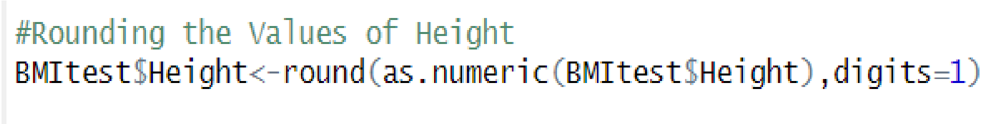
Now we have to Test the Regression Model on Test Dataset. And Plotting the Predicted Values and Actual Values.

Code:-To Predict the BMI values on Test Dataset



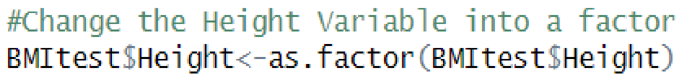
```
#Predicting Values Based on linear Model
BMItest$BMI.Predicted<-predict.lm(BMIPredictModel,newdata=BMItest)
```

Code:-Rounding the values of Height up to 2 decimals so that we can group according to Height



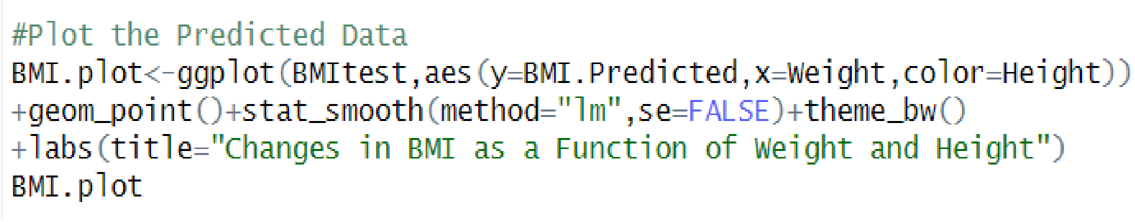
```
#Rounding the Values of Height
BMItest$Height<-round(as.numeric(BMItest$Height),digits=1)
```

Code:-Changing the Height Variable into Factor



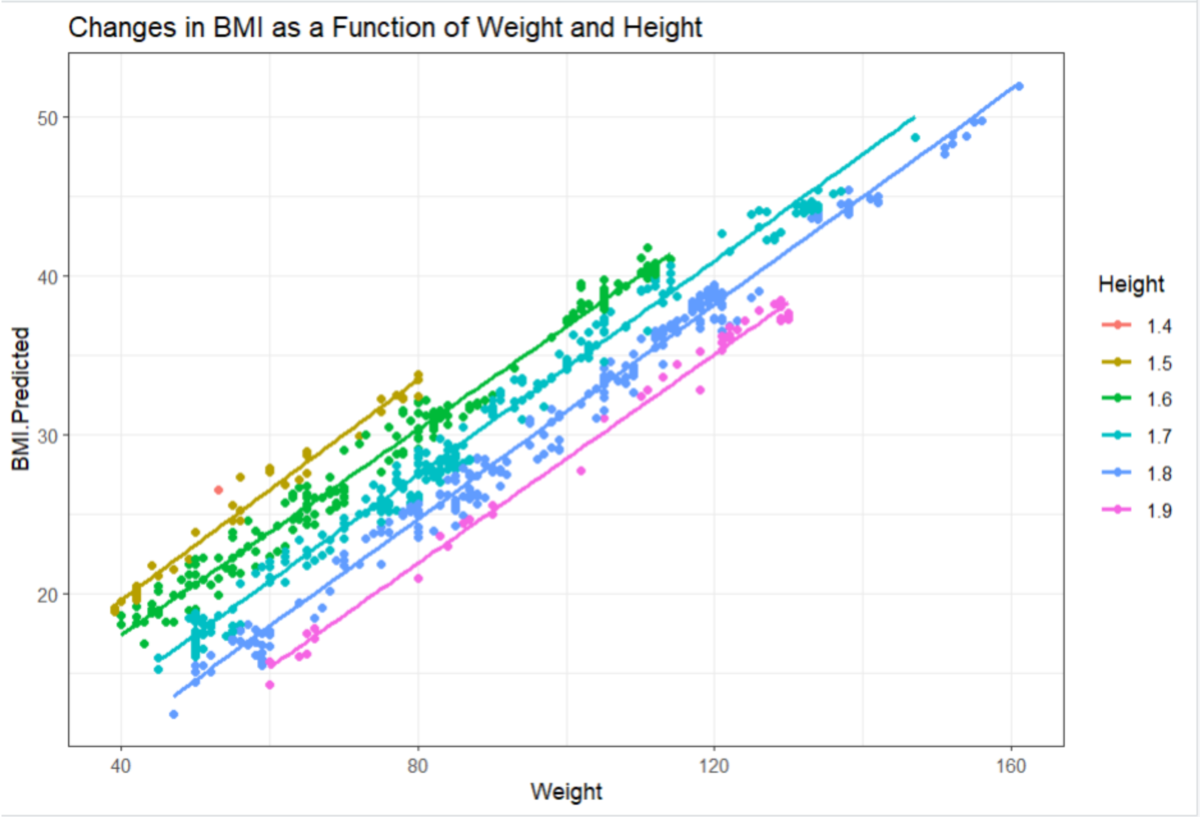
```
#Change the Height Variable into a factor
BMItest$Height<-as.factor(BMItest$Height)
```

Code:-Plotting the Predicted using ggplot()



```
#Plot the Predicted Data
BMI.plot<-ggplot(BMItest,aes(y=BMI.Predicted,x=Weight,color=Height))
+geom_point()+stat_smooth(method="lm",se=FALSE)+theme_bw()
+labs(title="Changes in BMI as a Function of Weight and Height")
BMI.plot
```

Output:-

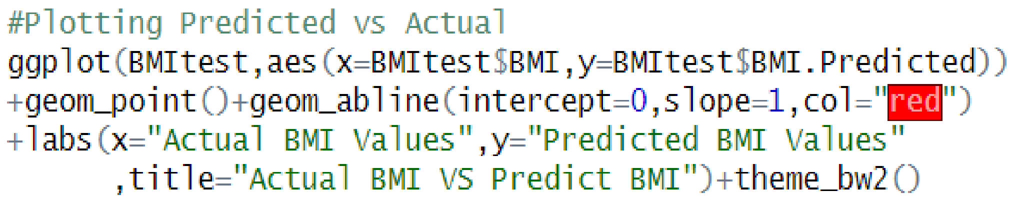


Plotting Predicted vs Actual On Test Data:-

Now we are plotting the Actual and Predicted BMI values on Test Dataset so that we can visualize the model accuracy.

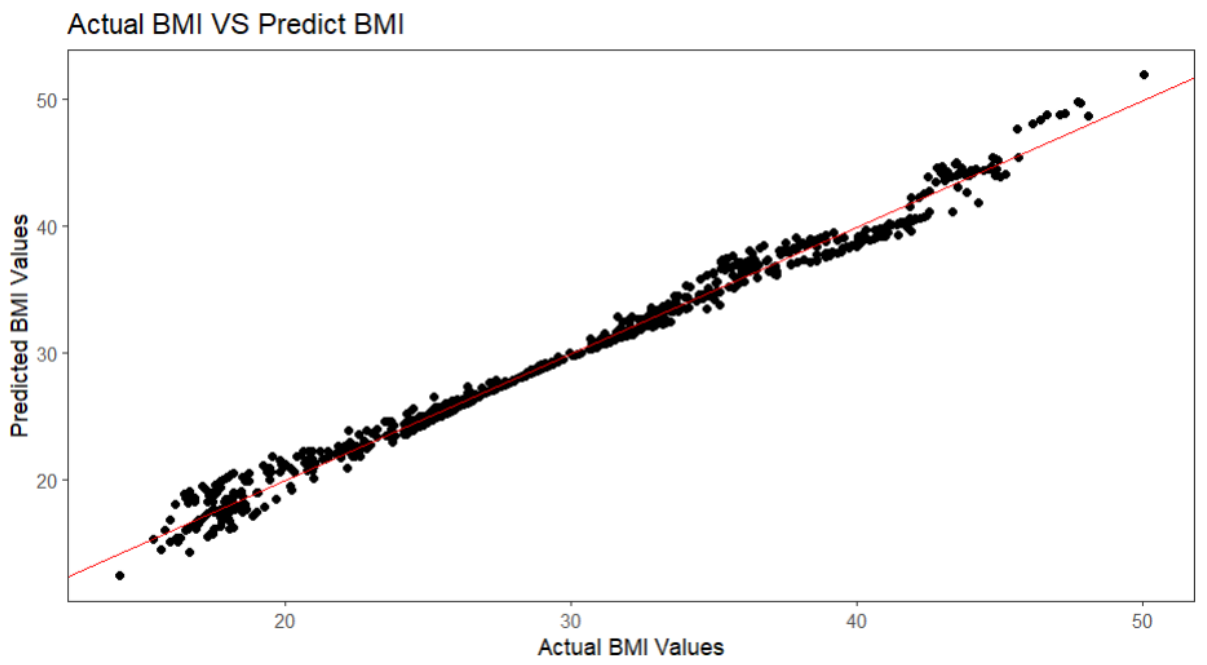
Plotting the graph by taking Predicted BMI on Y axis and Actual BMI on X-axis

Code:-



```
#Plotting Predicted vs Actual
ggplot(BMItest,aes(x=BMItest$BMI,y=BMItest$BMI.Predicted))
+geom_point()+geom_abline(intercept=0,slope=1,col="red")
+labs(x="Actual BMI Values",y="Predicted BMI Values"
      ,title="Actual BMI VS Predict BMI")+theme_bw2()
```

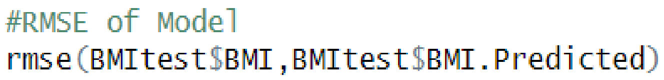
Output:-



The above graph is linear that indicates out model is good.

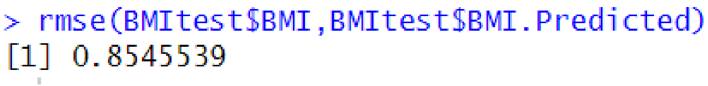
RMSE Value of The Model:-

Code:- Indicates Root Mean Square Error i.e Adjusted R^2 Value For Test Dataset



```
#RMSE of Model
rmse(BMItest$BMI,BMItest$BMI.Predicted)
```

Output:-



```
> rmse(BMItest$BMI,BMItest$BMI.Predicted)
[1] 0.8545539
```

Reporting the Results in Multi Linear Regression:-

In our Dataset we significantly found that Relationship between BMI and Height, BMI and Weight.

Significantly we conclude that, The BMI value is increased by 0.33 for every increase in one kilogram of weight. The BMI values is decreased by 33 for every increase in 1 meter of Height.

6. Conclusion:-

The multiple linear regression model developed in this project significantly improved BMI prediction compared to simpler models. Using independent variables i.e using Weight and Height, the model explained 98.91% of the variance in BMI with an adjusted R-squared of 0.9891.

Weight showed the strongest positive association with BMI, suggesting that 0.33. Conversely, Height had a negative association, indicating that approximately thirty. These findings highlight the complex interplay of factors influencing BMI and provide valuable insights for healthcare and fitness.

However, it's important to acknowledge that the model has limitations. Future research could address these limitations and further refine the model for even more accurate BMI prediction.

7. References:-

Dataset Link:- https://www.kaggle.com/datasets/mandysia/obesity-dataset-cleaned-and-data-sinthetic

YouTube Reference Video:-
https://youtu.be/UomnHPBfvBc?si=P0TJk0GDP80xSqlU

Regression Analysis using R:-

https://www.scribbr.com/statistics/linear-regression-in-r/