Name : Dharshani A

Register No: 23MSP3068

Section : Section - 2

Course Code : CS6103

Course Name : PROBABILITY AND STATISTICS

Project Title : Regression Analysis on Age, Weight, Height, BMI Analysis Dataset using Minitab

Table of Contents

1.Introduction

Determining and measuring the link between one or more independent variables and a dependent variable is the aim of this endeavor.Understanding the ANOVA, model validation, and Chi-square test is another goal of this project.

Age, Weight, Height, BMI Analysis from https://www.kaggle.com/datasets/rukenmissonnier/age-weight-height-bmi-analysis/was selected as the data set.

This project is important because it provides insight into the correlations between variables, enabling data-driven decision-making. Identifying important variables, forming predictions, and testing hypotheses are helpful.

2. Dataset Description

Source of dataset ::

https://www.kaggle.com/datasets/rukenmissonnier/age-weight-height-bmi-analysis

There are 5 columns and 741 rows in the dataset.

Description of the attributes ::

1.Age :: Numerical column to store age(discrete).

2.Height :: Numerical column to store height(continuous).

3.Weight :: Numerical column to store weight(continuous)

4.Bmi:: Numerical column to store Body Mass Index(continuous).

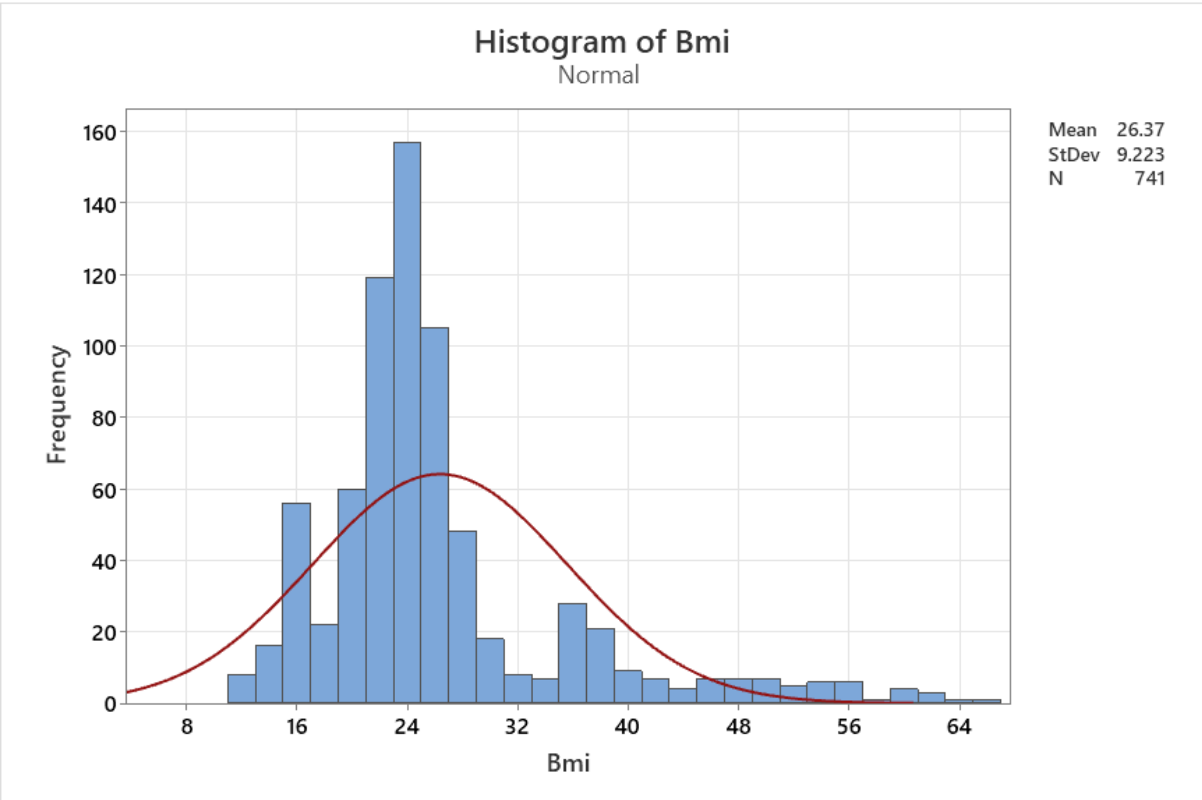
5.BmiClass :: Categorical column to store BmiClass(Ordinal).

3.Exploratory Data Analysis and Visualization

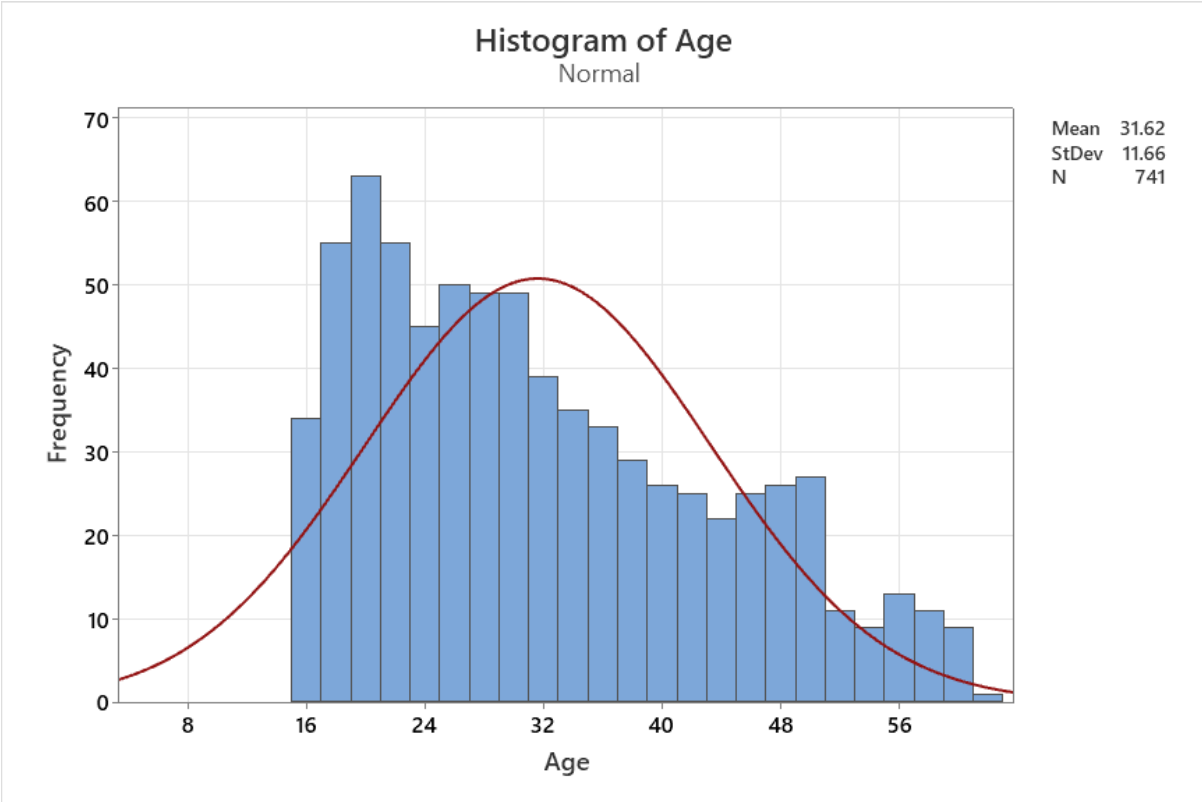
a) Data Analysis:

There are no missing values in the dataset.
I have not detected any anomalies in the dataset.

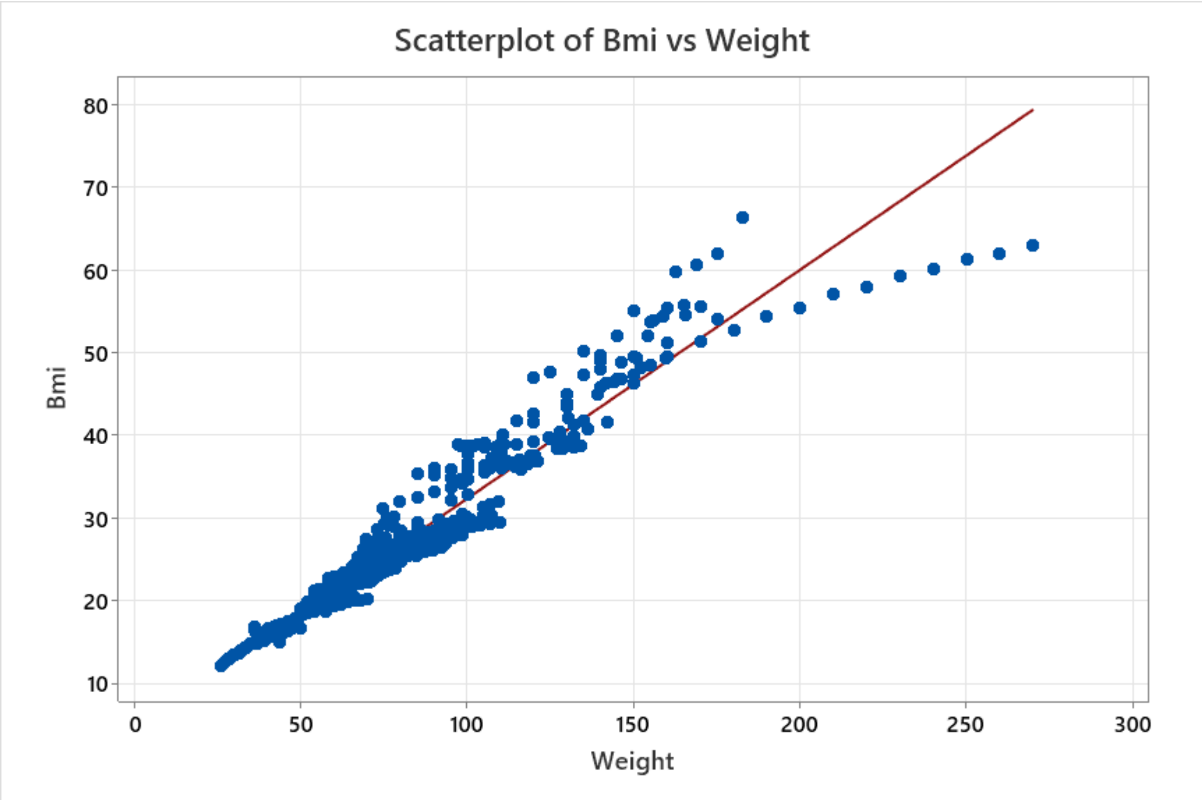
b) Visualisations:



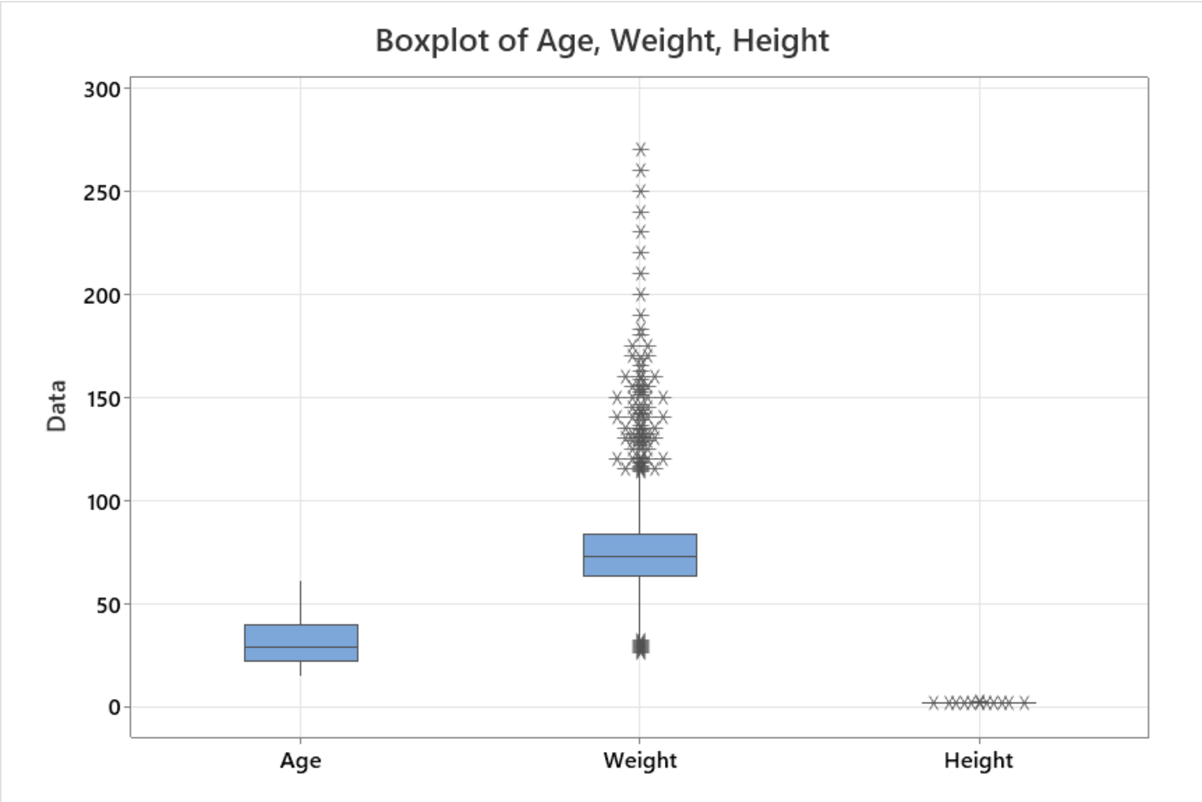
The histogram of Bmi column shows that it is slightly right skewed.The mean of Bmi column is 26.37 and standard deviation is 9.223.



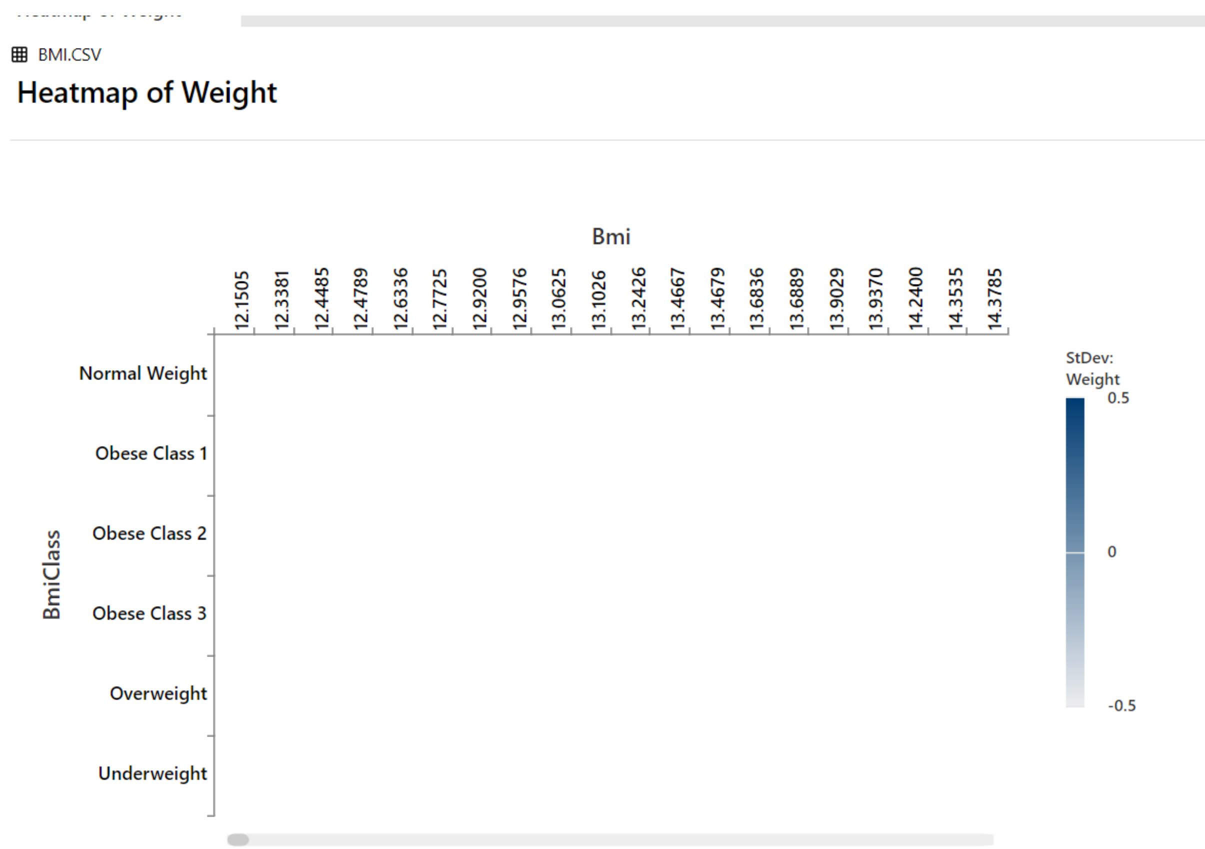
The histogram of Age column shows that it is slightly right skewed.The mean of Bmi column is 31.62 and standard deviation is 11.66.



The scatter plot shows relation between Bmi and Weight column.And the type of relationship is linear positive relationship.There exists strong relationship between Bmi and Weight.



It can be understood that there are very few outliers in the columns Age,Weight and Height because there are very few points that are far away.



There is no deviation between the columns Bmi and BmiClass with respect to the predictor used weight.So we cannot determine the correlation between Bmi and BmiClass using the standard deviation function for predictor weight.

c) Probability Distribution Analysis:

Probability Density Function

BMI.CSV

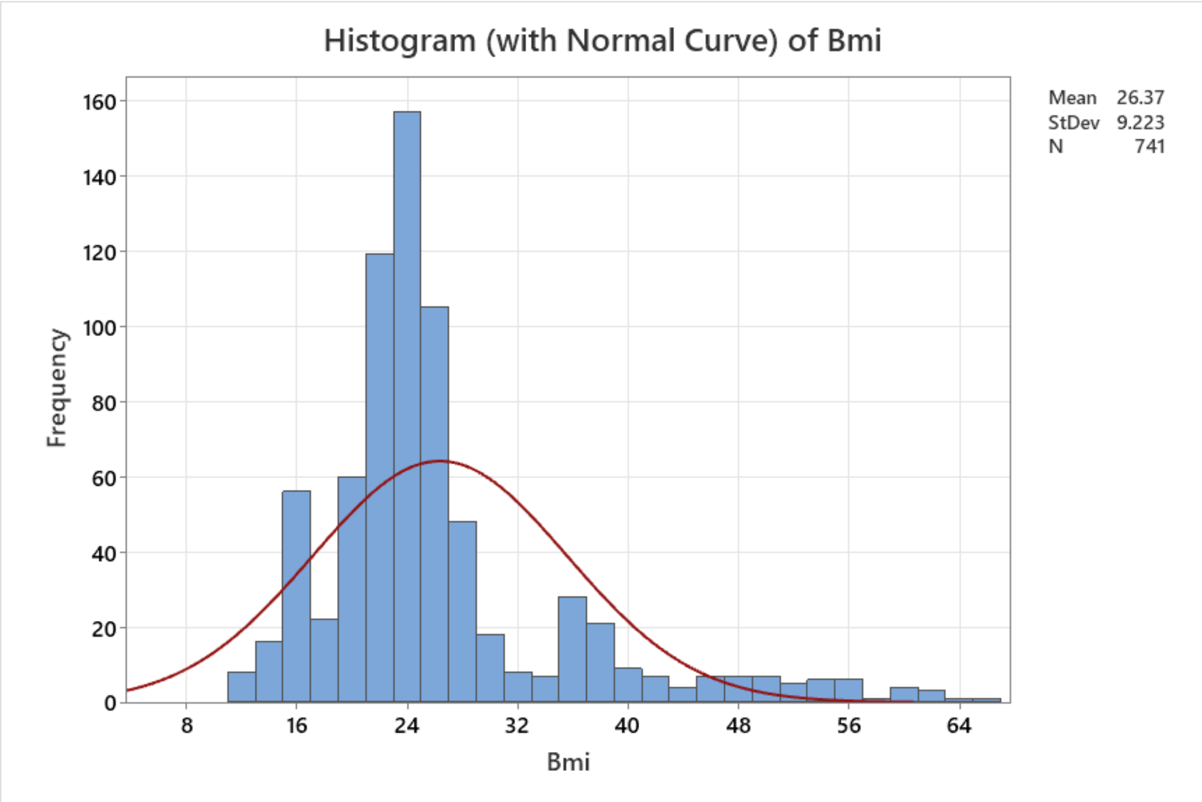
Probability Density Function

Continuous uniform on 12.15 to 66.301

x	f(x)
31.9357	0.0184669
27.0237	0.0184669
31.0926	0.0184669
16.8418	0.0184669
38.8960	0.0184669
27.1263	0.0184669
25.3702	0.0184669
28.8399	0.0184669
16.8887	0.0184669
31.2640	0.0184669
27.1229	0.0184669
25.4014	0.0184669
28.8027	0.0184669
16.6597	0.0184669
27.1195	0.0184669

28.3605	0.0184669
32.1120	0.0184669
55.3633	0.0184669
16.4810	0.0184669
26.8118	0.0184669
25.2454	0.0184669
30.1300	0.0184669
16.5267	0.0184669
38.6945	0.0184669
26.8084	0.0184669
29.0733	0.0184669
16.3056	0.0184669
38.7517	0.0184669
26.8050	0.0184669
25.3069	0.0184669
28.5156	0.0184669
16.3506	0.0184669
39.0764	0.0184669
26.8016	0.0184669
25.3377	0.0184669

The column Bmi has Probability Density Function because it is a continuous random variable.The upper bound is 66.301 and the lower bound is 12.15.So all the x value that is the Bmi between the upper and lower bound has the probability(p(x)) as 0.0184.



The Skewness of Bmi is 1.72 that is greater than 0.So it is positively skewed.

4. Descriptive Statistics

Descriptive Statistics: Weig...

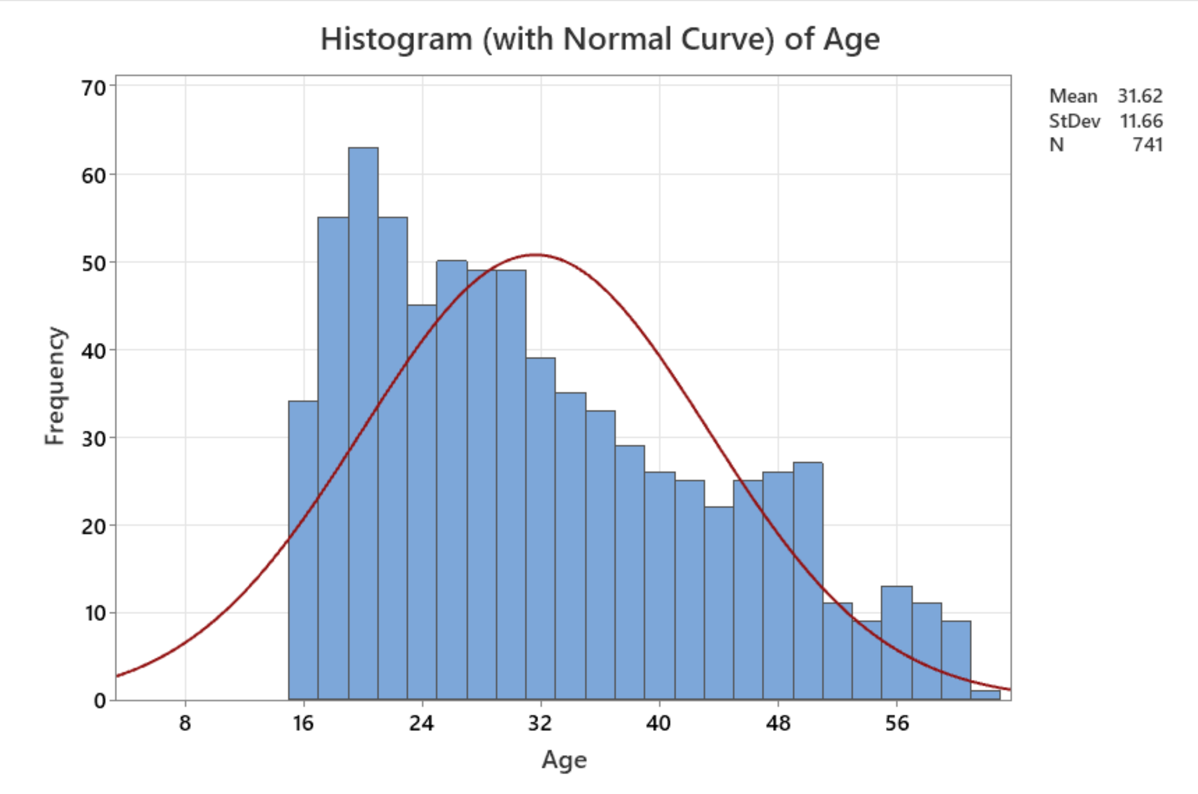
BMI.CSV

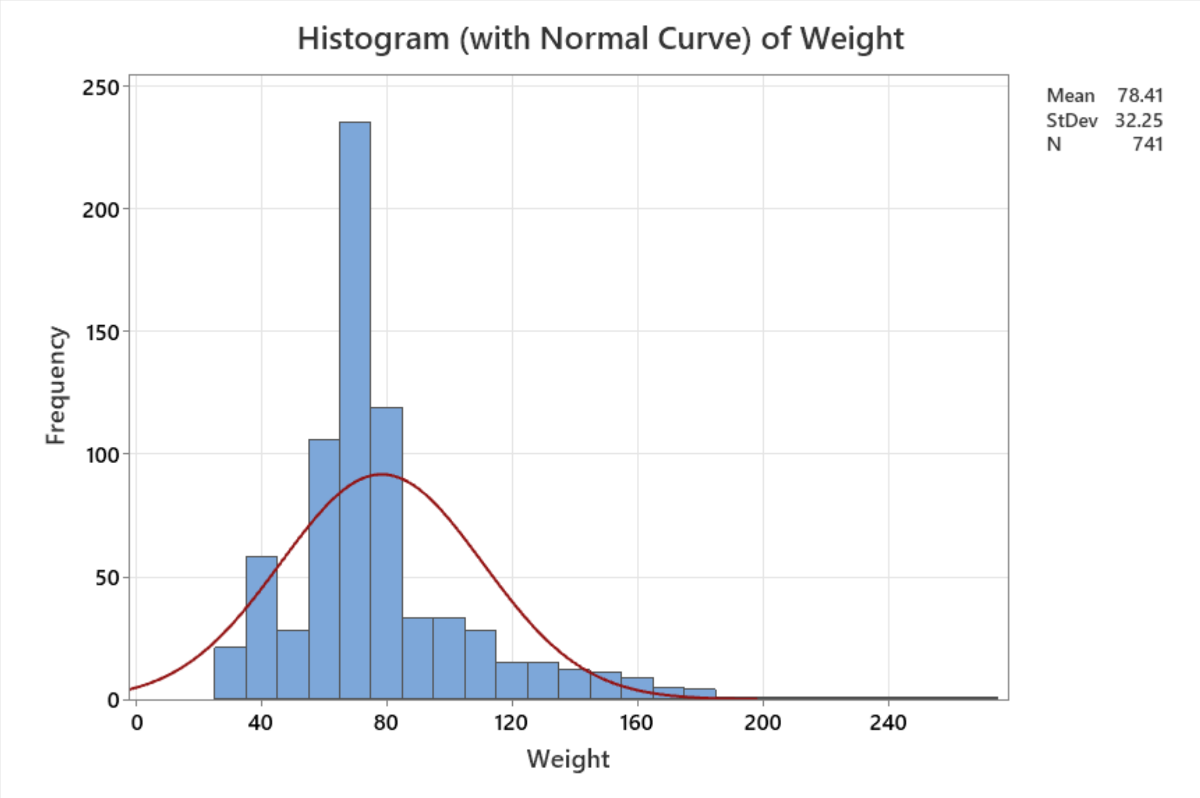
Descriptive Statistics: Weight, Age, Height, Bmi

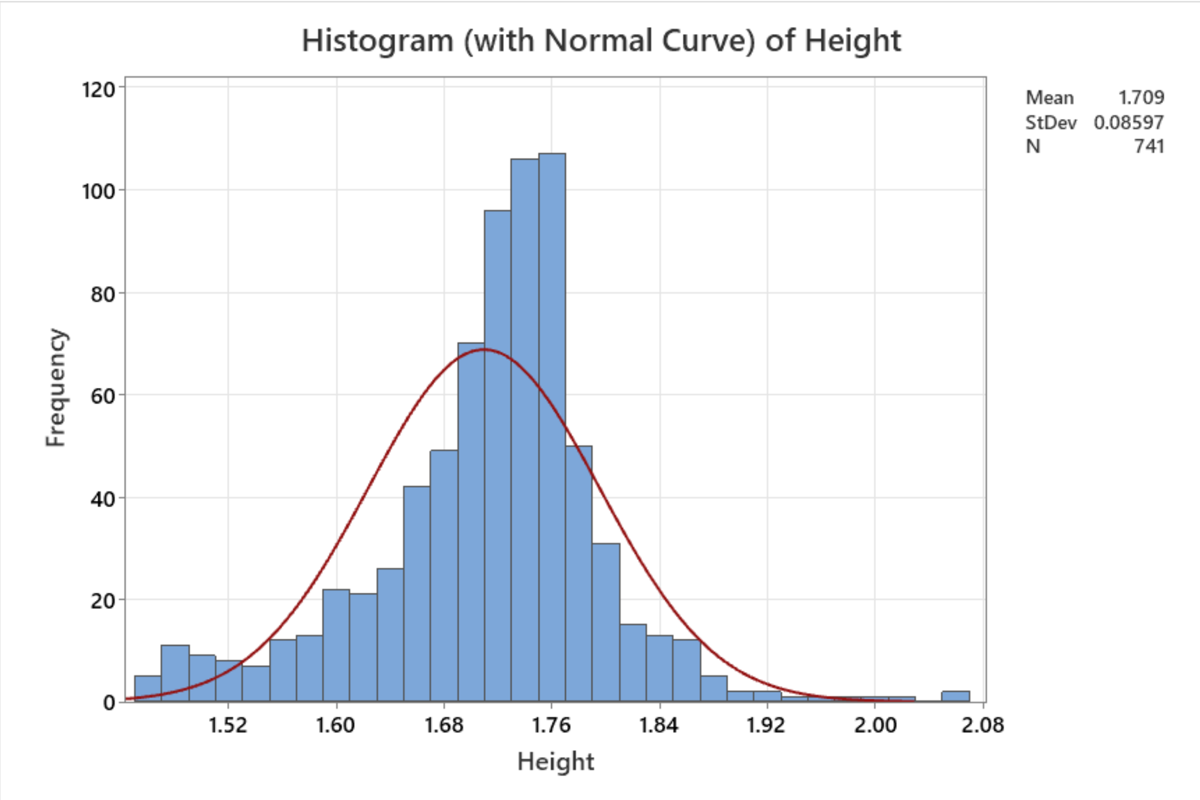
Statistics

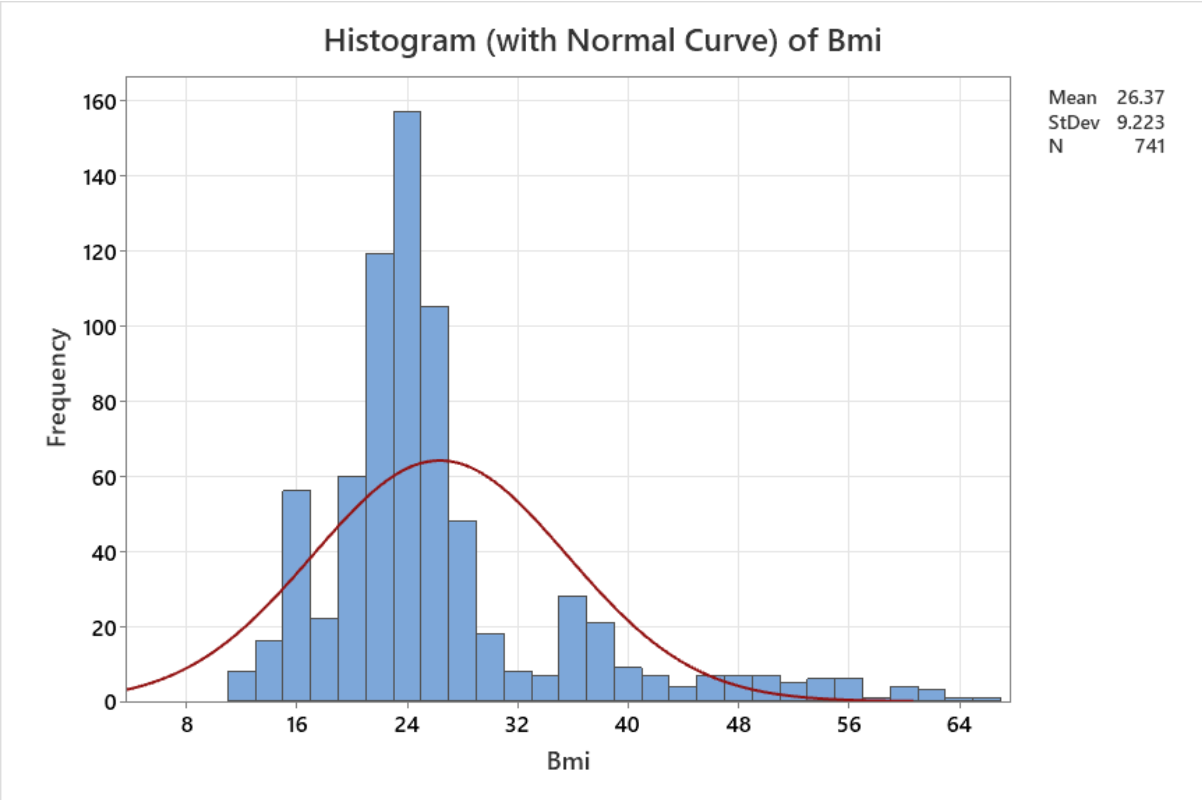
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Weight	741	0	78.41	1.18	32.25	25.90	63.00	72.90	83.30	270.00
Age	741	0	31.618	0.428	11.655	15.000	22.000	29.000	40.000	61.000
Height	741	0	1.7094	0.00316	0.0860	1.4600	1.6700	1.7210	1.7510	2.0700
Bmi	741	0	26.365	0.339	9.223	12.150	22.094	24.132	27.268	66.301

Variable	Skewness	Kurtosis
Weight	2.01	6.54
Age	0.58	-0.66
Height	-0.39	1.71
Bmi	1.72	3.24




Histogram (with Normal Curve) of Weight
Mean 78.41
StDev 32.25
N 741
Frequency
Weight
0
50
100
150
200
250
0
40
80
120
160
200
240


Histogram (with Normal Curve) of Height
Mean 1.709
StDev 0.08597
N 741
Frequency
Height
0
20
40
60
80
100
120
1.52
1.60
1.68
1.76
1.84
1.92
2.00
2.08



The Skewness of Weight is 2.01 that is greater than 0.So it is positively skewed.
The Skewness of Height is -0.39 that is less than 0.So it is negatively skewed.
The Skewness of Age is 0.58 that is greater than 0.So it is positively skewed.
The Skewness of Bmi is 1.72 that is greater than 0.So it is positively skewed.

The Kurtosis of Weight is 6.54 that is greater than 3.So it is Lepto.

The Kurtosis of Height is 1.71 that is less than 3.So it is Platy.

The Kurtosis of Age is -0.66 that is less than 3.So it is Platy.

The Kurtosis of Bmi is 3.24 that is greater than 3.So it is Lepto.

Mean::

Weight - 78.41
Height - 1.7094
Age - 31.618
Bmi - 26.365

Standard Deviation::

Weight - 32.25
Height - 0.0860
Age - 11.655
Bmi - 0.339

Median ::

Weight - 72.90
Height - 1.7210
Age - 29.000
Bmi - 24.132

Minimum Value::

Weight - 25.90
Height - 1.4600
Age - 15.000
Bmi - 12.150

Maximum Value::

Weight - 270.00
Height - 2.0700
Age - 61.000
Bmi - 66.301

5. Regression Analysis

a) Simple Linear Regression:

Regression Analysis: Bmi v...

BMI.CSV

Regression Analysis: Bmi versus Weight

Regression Equation

Bmi = 4.685 + 0.27649 Weight

Coefficients

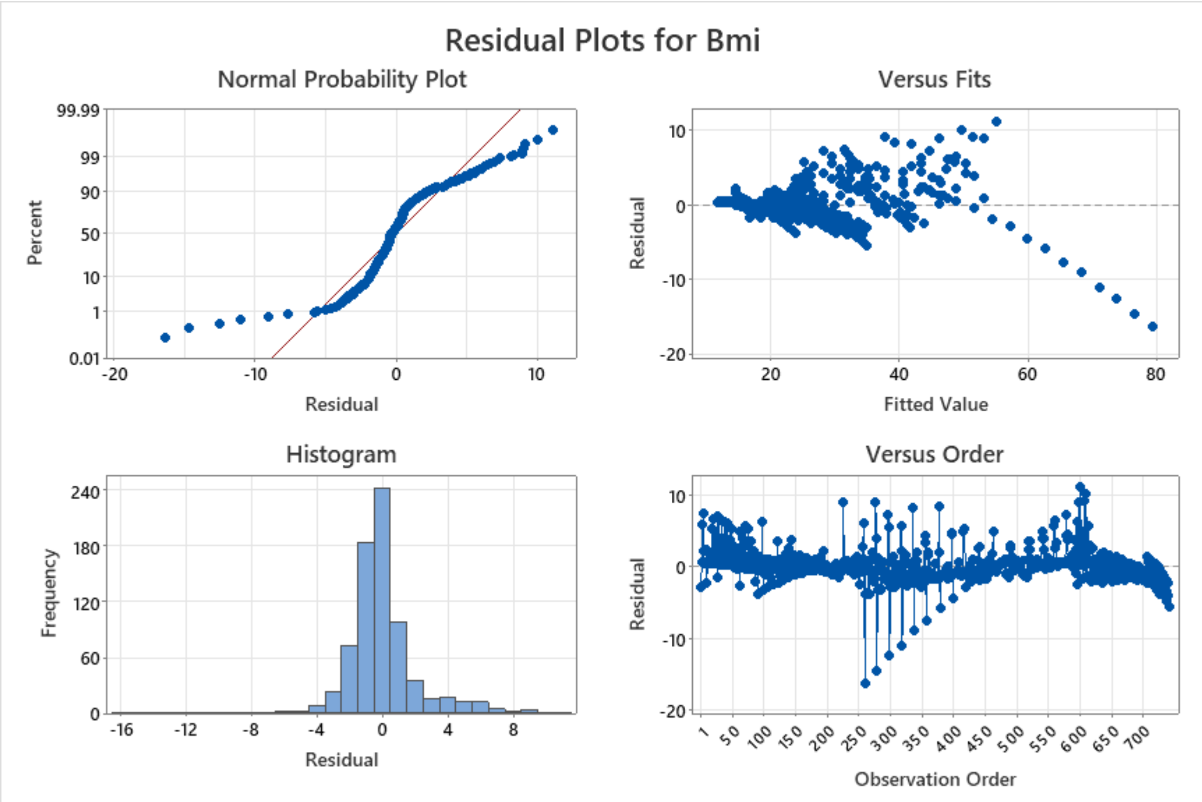
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.685	0.227	20.60	0.000	
Weight	0.27649	0.00268	103.05	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.35422	93.49%	93.48%	93.32%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	58854.0	58854.0	10618.95	0.000
Weight	1	58854.0	58854.0	10618.95	0.000
Error	739	4095.8	5.5		
Lack-of-Fit	449	3451.1	7.7	3.46	0.000
Pure Error	290	644.7	2.2		
Total	740	62949.8			



The predictor chosen is Weight.

<u>Regression Equation</u>

Bmi=4.685 + 0.27649 Weight

4.685 is the Y intercept,all equations will start with 4.685.
0.27649 is the Weight Coefficient,multiply it by Weight value.

Weight's VIF value is 1.00 that is less than 5.So the model is in good shape and because of that there is no multicollinearity.

P-Value of Weight is 0.000 less than 0.05 so that the variable is significant.

P-Value of F-test = 0.000.Therefore the model is statistically significant.
The R squared value =93.49% so it is nearer to 100 % and greater than 83%.Therefore it is a good model.

b) Multiple Linear Regression:

BMI.CSV

Regression Analysis: Bmi versus Weight, Height, Age

Regression Equation

Bmi = 45.42 + 0.31800 Weight - 26.068 Height + 0.01811 Age

Coefficients

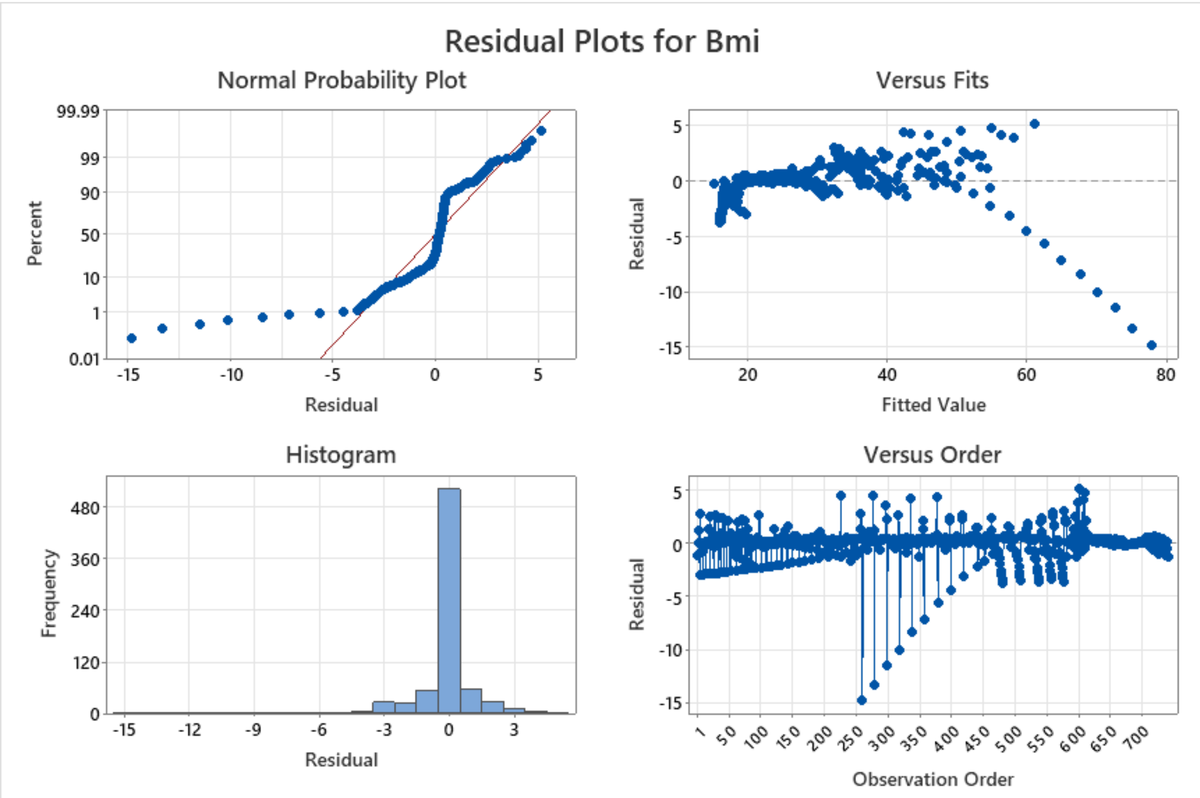
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	45.42	1.31	34.56	0.000	
Weight	0.31800	0.00219	145.34	0.000	1.61
Height	-26.068	0.816	-31.94	0.000	1.59
Age	0.01811	0.00481	3.76	0.000	1.02

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.51356	97.32%	97.31%	97.17%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	61261.4	20420.5	8913.90	0.000
Weight	1	48391.8	48391.8	21123.90	0.000
Height	1	2337.3	2337.3	1020.28	0.000
Age	1	32.5	32.5	14.17	0.000
Error	737	1688.4	2.3		
Total	740	62949.8			



The predictors chosen are Weight,Height,Age.

<u>Regression Equation</u>

Bmi = 45.42 + 0.31800 Weight - 26.068 Height + 0.01811 Age

45.42 is the Y intercept,all equations will start with 45.42.

0.31800 is the Weight Coefficient,multiply it by Weight value.

-26.068 is the Height Coefficient,multiply it by Height value.

0.01811 is the Age Coefficient,multiply it by Age value.

VIF value of Weight,Height and Age are less than 5.So the model is in good shape and because of that there is no multicollinearity.

P-Value of Weight,Height and Age are 0.000 less than 0.05 so that the variables are significant.

P-Value of F-test = 0.000.Therefore the model is statistically significant.

The R squared value =97.32% so it is nearer to 100 % and greater than 83%.Therefore it is a good model.

c) Regression using Bayesian Regression Analysis.Also compare the models between two groups using Bayes Factor.

```
#bayessian regression
install.packages("brms")
install.packages("rstan")
library(brms)
library(rstan)
data<-read.csv("C:/Users/ajesh/OneDrive/Desktop/bmi.csv")
head(data)
fit<-brm(Bmi~Weight+Height,data = data,family = gaussian())
summary(fit)
pred<-data.frame(Weight=100,Height=1.80)
predict(fit,pred)
```

```
> pred<-data.frame(Weight=100,Height=1.80)
> predict(fit,pred)
     Estimate Est.Error     Q2.5    Q97.5
[1,] 30.84407  1.511182 27.98064 33.88548
```

```
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    46.19      1.30    43.63    48.71 1.00     2964     2621
Weight        0.32      0.00     0.31     0.32 1.00     3628     3276
Height      -26.23      0.82   -27.81   -24.64 1.00     2910     2404

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     1.53      0.04     1.45     1.61 1.00     3263     2651
```

The value of Rhat for all the predictors and response is 1.00 therefore it shows that the links are highly converged.

0.32 is the weight coefficient,so one unit of weight increases the bmi by 0.32.

-26.23 is the height coefficient,so one unit of height decreases the bmi by 26.23.

The Bulk_ESS and Tail_ESS are big numbers so that it shows that the model is efficient and reliable.

```
predict(fit,pred)
#bayesfactor
install.packages("BayesFactor")
library(BayesFactor)
data=read.csv("C:/Users/ajesh/OneDrive/Desktop/bmi.csv")
g1=data$Bmi[data$Age==21]
g2=data$Bmi[data$Age==40]
res<-ttestBF(x=g1,y=g2)
summary(res)
```

```
> summary(res)
Bayes factor analysis
--------------
[1] Alt., r=0.707 : 0.4553137 ±0%

Against denominator:
  Null, mu1-mu2 = 0
---
Bayes factor type: BFindepSample, JZS
```

The Bayes Factor < 1 indicates evidence in favour of the null hypothesis.

d) Perform simple MonteCarlo Simulation for ttest

```r
install.packages("MonteCarlo")
library(MonteCarlo)
set.seed(9)
ttest<-function(n,loc,scale){
  sample<-rnorm(n,loc,scale)
  stat<-sqrt(n)*mean(sample)/sd(sample)
  decision<- abs(stat) >1.96
  return(list("decision"=decision))
}
n_grid<-c(50,100,250,500)
loc_grid<-seq(0,1,0.2)
scale_grid<-c(1,2)
param_list<-list("n"=n_grid,"loc"=loc_grid,"scale"=scale_grid)
res<-MonteCarlo(func=ttest,nrep=1000,param_list =param_grid)
summary(res)
rows<-c("n")
cols<-c("loc","scale")
MakeTable(output = res,rows=rows,cols=cols,digit=2)
```

```
 Required time: 1.31 secs for nrep = 1000  repetitions on 1 CPUs

 Parameter grid:

      n : 50 100 250 500
    loc : 0 0.2 0.4 0.6 0.8 1
  scale : 1 2


 1 output arrays of dimensions: 4 6 2 1000
```

```
\hline\hline\\\\
 scale && \multicolumn{ 6 }{c}{ 1 } &  & \multicolumn{ 6 }{c}{ 2 } \\
n/loc &  & 0 & 0.2 & 0.4 & 0.6 & 0.8 & 1 &  & 0 & 0.2 & 0.4 & 0.6 & 0.8 & 1 \\
 &  &  &  &  &  &  &  &  &  &  &  &  &  &  \\
50 &  & 0.05 & 0.28 & 0.80 & 0.99 & 1.00 & 1.00 &  & 0.05 & 0.12 & 0.27 & 0.58 & 0.80 & 0.94 \\
100 &  & 0.05 & 0.51 & 0.97 & 1.00 & 1.00 & 1.00 &  & 0.06 & 0.17 & 0.53 & 0.82 & 0.98 & 1.00 \\
250 &  & 0.06 & 0.89 & 1.00 & 1.00 & 1.00 & 1.00 &  & 0.06 & 0.36 & 0.89 & 1.00 & 1.00 & 1.00 \\
500 &  & 0.04 & 0.99 & 1.00 & 1.00 & 1.00 & 1.00 &  & 0.05 & 0.63 & 1.00 & 1.00 & 1.00 & 1.00 \\
\\
\\
```

The output array will provide results for every combination of the parameters in the grid,repeated 1000 times.

This output array allows for a comprehensive examination of how the t-test performs under a wide range of scenarios.The 1000 repetitions for each combination help assess the variability and reliability of the results under each set of conditions.

6. Chi-square Test

a) Goodness-of-fit:

↓	C1-T	C2	C3	C4
	TEAM	TROPHIES COUNT		
1	EEE	13		
2	CS	8		
3	EC	8		
4	MECH	11		
5	CIVIL	3		
6	EI	9		
7				
8				

GOODNESS

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: TROPHIES COUNT

Using category names in TEAM

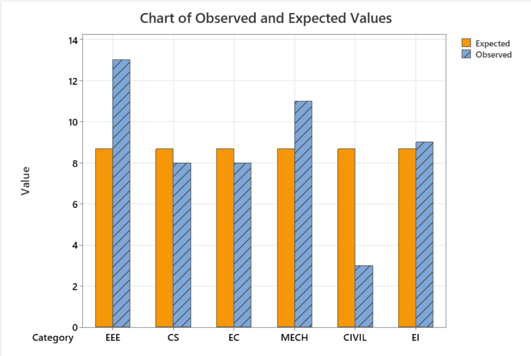
Observed and Expected Counts

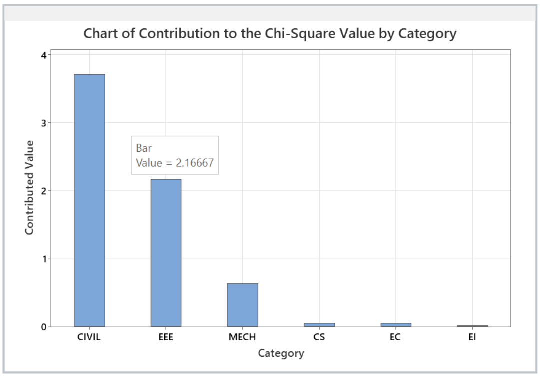
Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
EEE	13	0.166667	8.66667	2.16667
CS	8	0.166667	8.66667	0.05128
EC	8	0.166667	8.66667	0.05128
MECH	11	0.166667	8.66667	0.62821
CIVIL	3	0.166667	8.66667	3.70513
EI	9	0.166667	8.66667	0.01282

Chi-Square Test

N	DF	Chi-Sq	P-Value
52	5	6.61538	0.251

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: TROPHIES COUNT





The p-value from the output is 0.251

P > 0.05 therefore we fail to reject the null hypothesis.

There is no evidence that the trophies were not randomly selected from a population with equal proportions of Teams.

b) Test of Association:

↓	C1-T	C2	C3	C4	C5	C
	Age Group	Excercise-YES	Excercise-NO			
1	Teens	43	63			
2	Young Adults	95	113			
3	Adults	32	56			
4	Older Adults	12	60			
5						
6						
7						
8						

AASSOCIATION

Tabulated Statistics: Age Group, Worksheet columns

Rows: Age Group Columns: Worksheet columns

	Excercise-YES	Excercise-NO	All
Teens	43	63	106
	40.70	65.30	
	0.1299	0.0810	
Young Adults	95	113	208
	79.86	128.14	
	2.8682	1.7877	
Adults	32	56	88
	33.79	54.21	
	0.0947	0.0590	
Older Adults	12	60	72
	27.65	44.35	
	8.8544	5.5188	
All	182	292	474

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	19.394	3	0.000
Likelihood Ratio	21.156	3	0.000

Ho :Taking exercises and age groups are not related in the population(independent).
Ha :Taking exercises and age groups are related in the population(dependent).

The p-value from the output is 0.000.

P < 0.05 therefore we can reject the null hypothesis.

There is enough evidence of a relationship between taking exercises and age groups in the population(dependent).

7. ANOVA

↓	C1	C2	C3	C4	C5	C6	C7
	Feed_1	Feed_2	Feed_3	Feed_4			
1	60.8	68.3	102.6	87.9			
2	57.1	67.7	102.2	84.7			
3	65.0	74.0	100.5	83.2			
4	58.7	66.3	97.5	85.8			
5	61.8	69.9	98.9	90.3			
6							
7							
8							

WORKSHEET 4

One-way ANOVA: Feed_1, Feed_2, Feed_3, Feed_4

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Factor	4	Feed_1, Feed_2, Feed_3, Feed_4

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	3	4703.2	1567.73	206.72	0.000
Error	16	121.3	7.58		
Total	19	4824.5			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.75386	97.48%	97.01%	96.07%

Means

Factor	N	Mean	StDev	95% CI
Feed_1	5	60.68	3.03	(58.07, 63.29)
Feed_2	5	69.24	2.96	(66.63, 71.85)
Feed_3	5	100.340	2.164	(97.729, 102.951)
Feed_4	5	86.38	2.78	(83.77, 88.99)

Pooled StDev = 2.75386

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

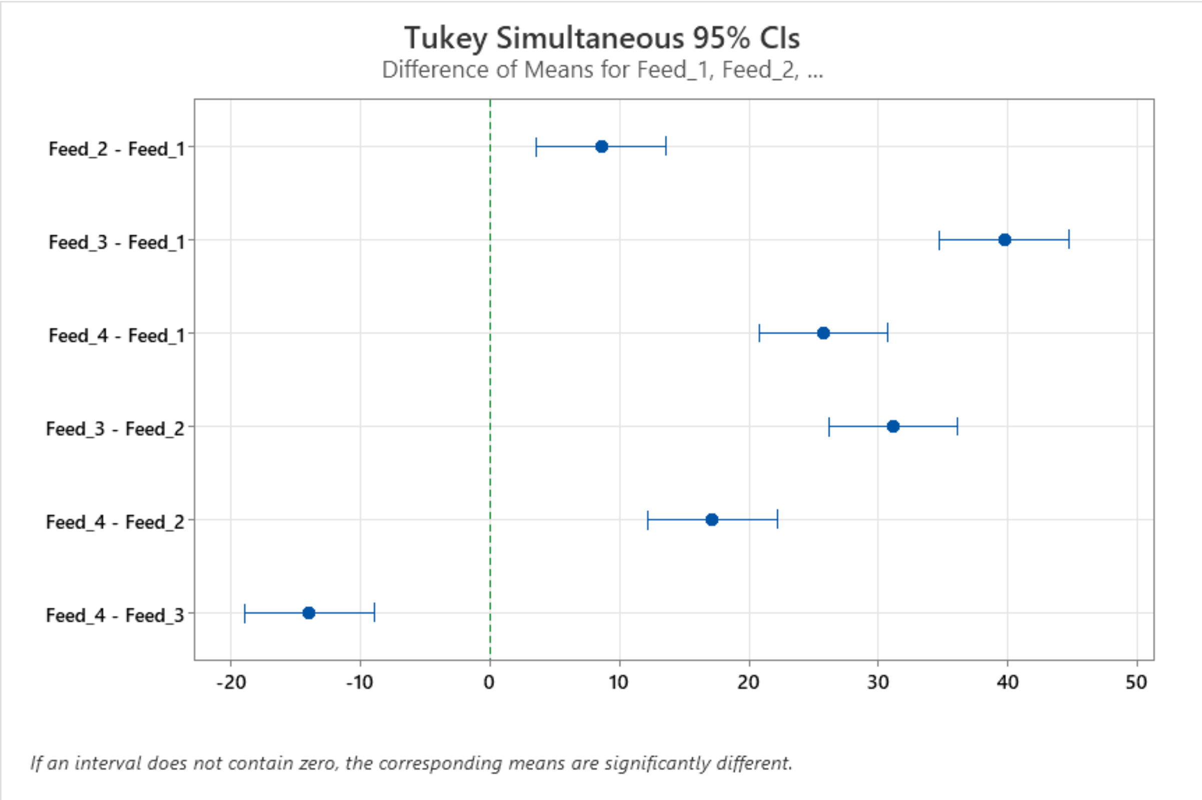
Factor	N	Mean	Grouping			
Feed_3	5	100.340	A			
Feed_4	5	86.38		B		
Feed_2	5	69.24			C	
Feed_1	5	60.68				D

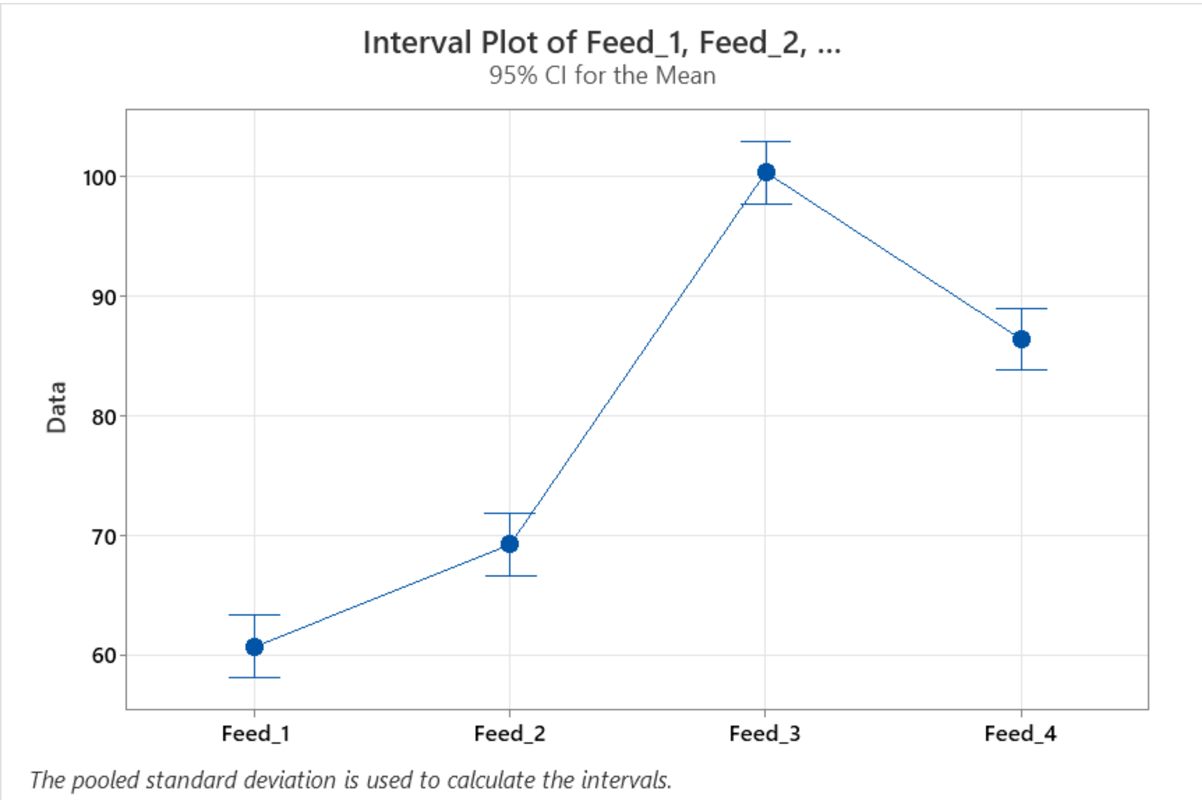
Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
Feed_2 - Feed_1	8.56	1.74	(3.57, 13.55)	4.91	0.001
Feed_3 - Feed_1	39.66	1.74	(34.67, 44.65)	22.77	0.000
Feed_4 - Feed_1	25.70	1.74	(20.71, 30.69)	14.76	0.000
Feed_3 - Feed_2	31.10	1.74	(26.11, 36.09)	17.86	0.000
Feed_4 - Feed_2	17.14	1.74	(12.15, 22.13)	9.84	0.000
Feed_4 - Feed_3	-13.96	1.74	(-18.95, -8.97)	-8.02	0.000

Individual confidence level = 98.87%


Tukey Simultaneous 95% CIs
Difference of Means for Feed_1, Feed_2, ...
Feed_2 - Feed_1
Feed_3 - Feed_1
Feed_4 - Feed_1
Feed_3 - Feed_2
Feed_4 - Feed_2
Feed_4 - Feed_3
-20
-10
0
10
20
30
40
50
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of Feed_1, Feed_2, ...
95% CI for the Mean
Data
100
90
80
70
60
Feed_1
Feed_2
Feed_3
Feed_4
The pooled standard deviation is used to calculate the intervals.

Ho :The groups have equal means.

Ha :At Least one group mean is different from the other group means.

The F-test statistic is F-value = 206.72

The P-Value = 0.000

$P < 0.05$ therefore we can reject the null hypothesis.

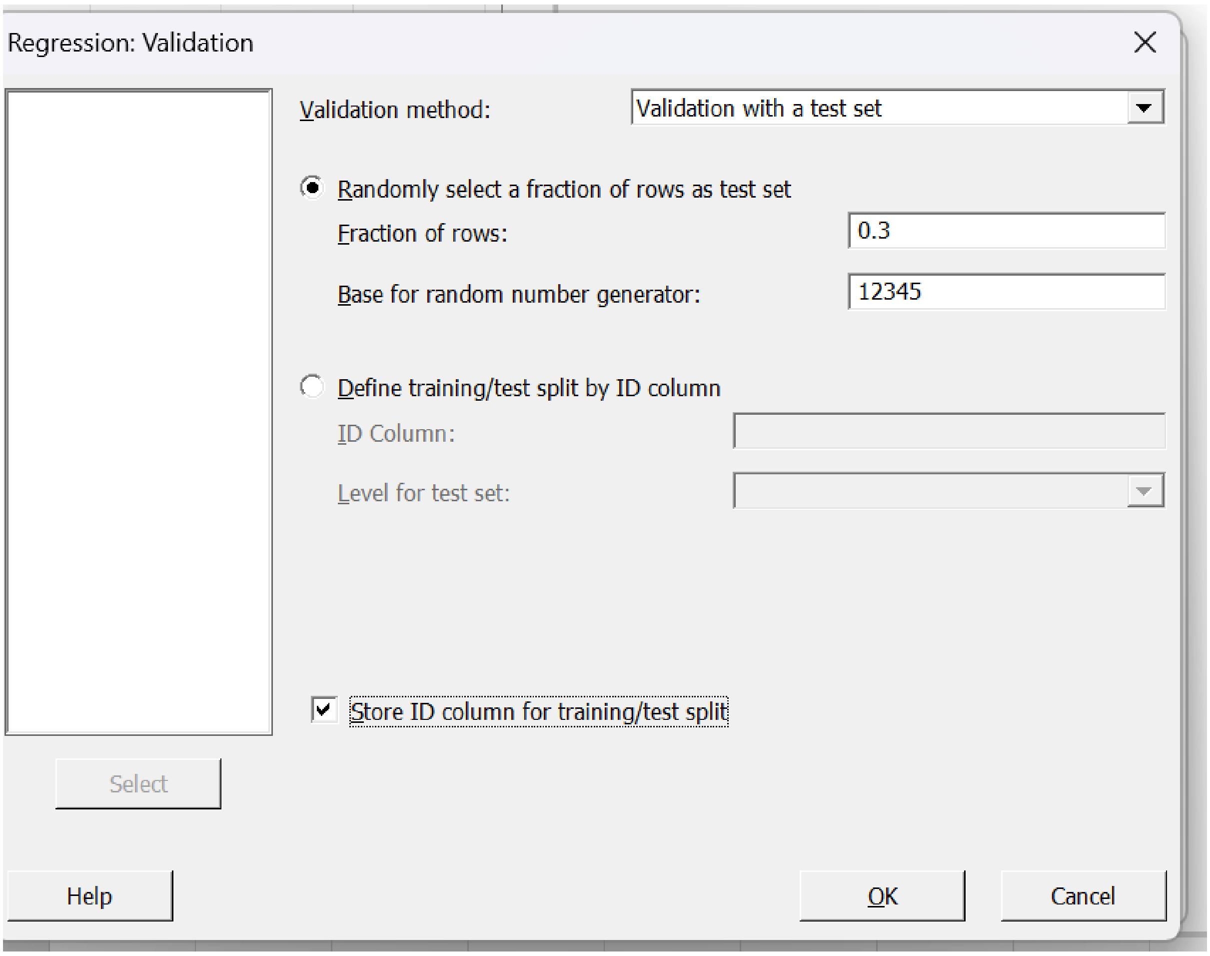
Therefore atleast one group mean is different from the other group means.

In the Tukey simultaneous tests, all the 6 groups have adjusted p value less than 0.05 so all the 6 groups are statistically different.

From the interval plot for Feed_1,Feed_2,Feed_3,Feed_4 we can see that all the six groups mean does not overlap each other so we can conclude that all the six groups are statistically different.

8. Model Validation, Diagnostic, and Prediction

a) Mention the size of the training and testing sets.



The 70 percent of the dataset is used for training the model whereas 30 percent is used for testing the model.

b) Provide performance metrics on the test set: RMSE, MAE, etc

Regression Analysis: Bmi v... ⌄ ✕

BMI.CSV

Regression Analysis: Bmi versus Weight, Height, Age

Method

Test set fraction 30.0%

Regression Equation

Bmi = 43.87 + 0.32129 Weight - 25.300 Height + 0.01817 Age

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	43.87	1.46	29.96	0.000	
Weight	0.32129	0.00252	127.34	0.000	1.61
Height	-25.300	0.909	-27.82	0.000	1.58
Age	0.01817	0.00539	3.37	0.001	1.03

Age	0.01817	0.00539	3.37	0.001	1.03

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	Test S	Test R-sq
1.39502	97.57%	97.56%	97.36%	1.78135	96.74%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	40265.6	13421.9	6896.86	0.000
Weight	1	31558.9	31558.9	16216.62	0.000
Height	1	1506.4	1506.4	774.07	0.000
Age	1	22.1	22.1	11.36	0.001
Error	515	1002.2	1.9		
Total	518	41267.8			

Fits and Diagnostics for Unusual Observations

Training Set

Obs	Bmi	Fit	Resid	Std Resid		
9	16.889	19.572	-2.683	-1.95		X
14	16.660	19.301	-2.641	-1.92		X
38	16.527	19.040	-2.513	-1.82		X
226	46.875	42.528	4.347	3.15	R	
259	54.012	55.102	-1.089	-0.79		X
277	55.096	50.847	4.250	3.09	R	X
279	61.868	76.068	-14.200	-10.57	R	X
297	51.992	48.716	3.276	2.37	R	
298	55.773	53.877	1.897	1.38		X
299	61.269	73.596	-12.327	-9.14	R	X
357	57.857	65.674	-7.817	-5.74	R	X
379	56.966	63.202	-6.236	-4.56	R	X

Fits and Diagnostics for Unusual Observations

Test Set

Obs	Bmi	Fit	Resid	Std Resid		
4	16.842	19.558	-2.716	-1.92		X
5	38.896	36.185	2.711	1.92		X
20	16.660	19.283	-2.623	-1.85		X
26	16.706	19.297	-2.591	-1.83		X
33	55.363	53.267	2.097	1.48		X
34	16.481	19.026	-2.545	-1.80		X
260	63.012	78.793	-15.781	-10.89	R	X
278	55.510	54.742	0.768	0.54		X
319	60.000	70.871	-10.871	-7.58	R	X
336	50.193	46.226	3.968	2.81	R	
338	59.265	68.399	-9.134	-6.38	R	X
377	47.630	43.483	4.147	2.94	R	
399	55.402	60.477	-5.075	-3.57	R	X

The standard error of regression in the test set is 1.78135.

The R squared value of the test set is 96.74%.

Regression Equation

Bmi =43.87 + 0.01817 Age - 25.300 Height + 0.32129 Weight

43.87 is the Y intercept,all equations will start with 45.42.

0.32129 is the Weight Coefficient,multiply it by Weight value.

-25.300 is the Height Coefficient,multiply it by Height value.

0.01817 is the Age Coefficient,multiply it by Age value.

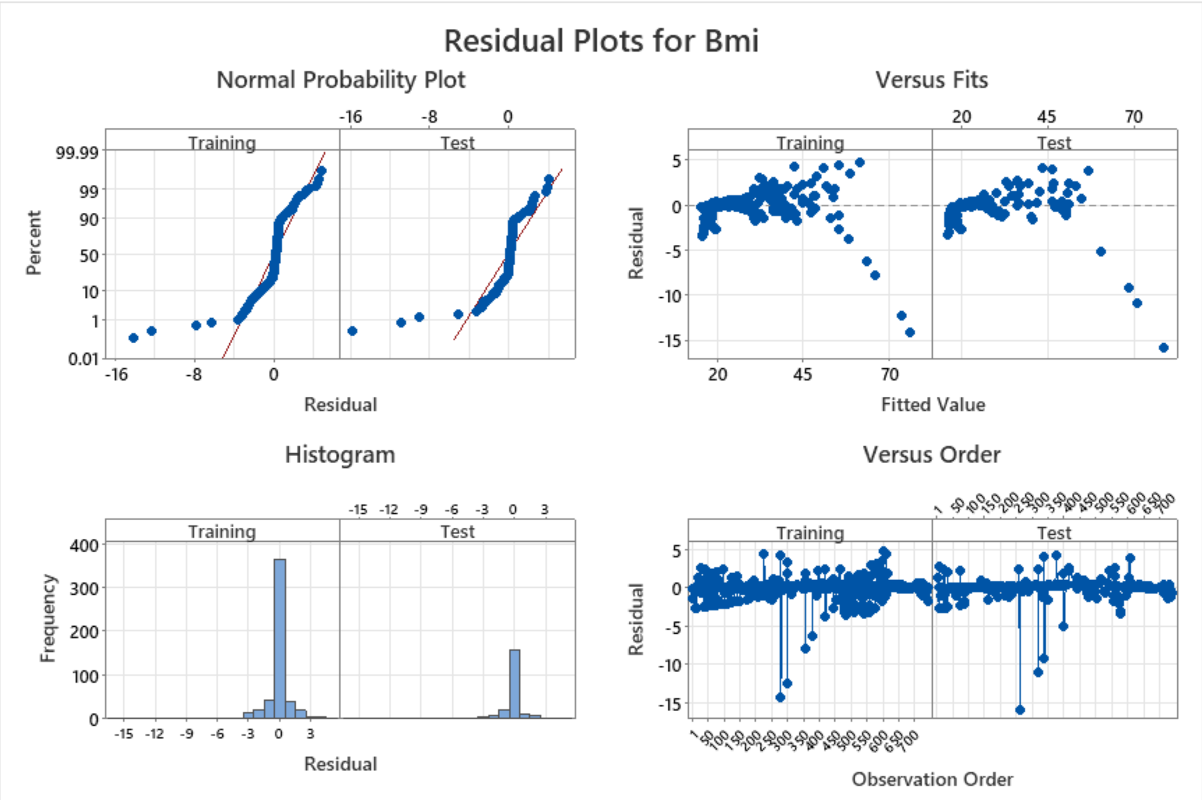
The value of R squared is 97.57 % and the value of R squared predicted 97.36%.

R-sq > R-sq(pred)

But there is no large range of difference between R-sq, R-sq(pred)so the model is not overfitted.

VIF value of Weight,Height and Age are less than 5.So the model is in good shape and because of that there is no multicollinearity.

c) Attach residual plots and interpret them.



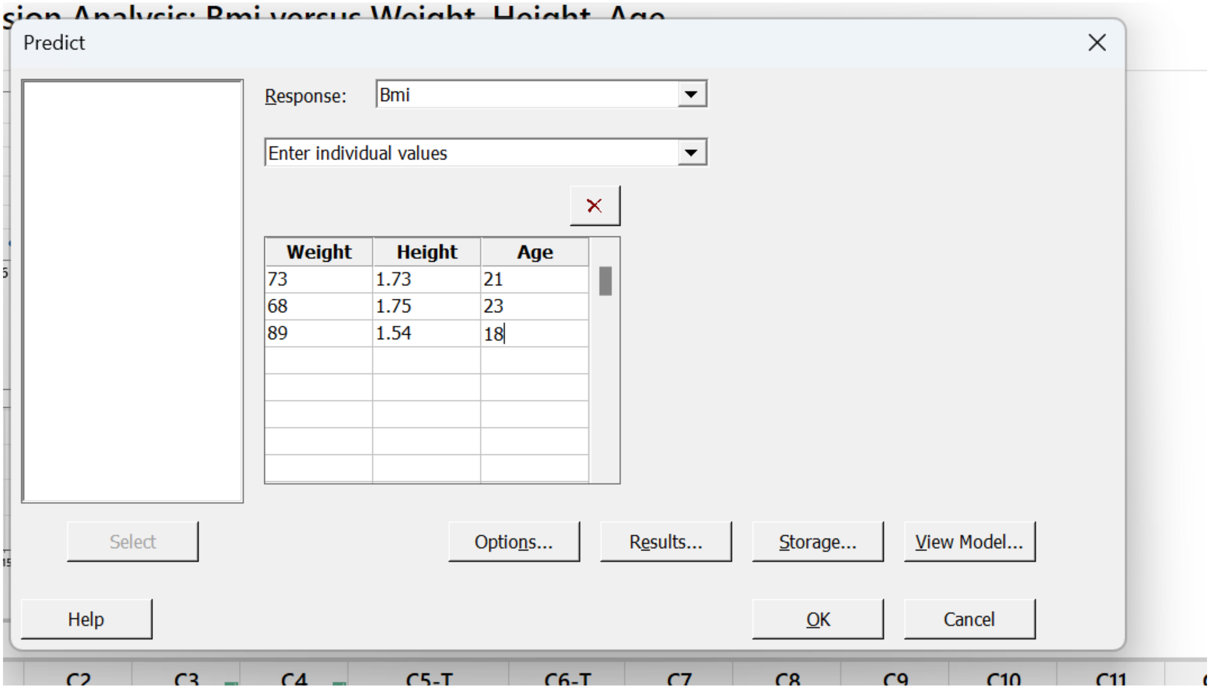
We can say that the residuals of both train and test sets are normally distributed from the normal probability plot because most of the points roughly follow the line.

From the Observation Order plot we can infer that there is no wave-like pattern created for the both train and test set.So the model is an efficient model.

From the Versus fit plot of both train and test set we can infer that there is no funnel-like pattern created and the points are mostly scattered .So the model is an efficient model.

By interpreting the histogram of both train and test set we can conclude that there is a slight positive skewness.

d) Provide a sample prediction on new, unseen data and interpret the results.



Prediction for Bmi

BMI.CSV

Prediction for Bmi

Regression Equation

Bmi = 43.87 + 0.01817 Age - 25.300 Height + 0.32129 Weight

Settings

Variable	Setting
Age	12
Height	1.34
Weight	60

Prediction

Fit	SE Fit	95% CI	95% PI	
29.4651	0.338158	(28.8008, 30.1294)	(26.6451, 32.2851)	XX

XX denotes an extremely unusual point relative to predictor levels used to fit the model.

Settings

Variable	Setting
Age	23
Height	1.8
Weight	78

Prediction

Fit	SE Fit	95% CI	95% PI
23.8100	0.109188	(23.5955, 24.0245)	(21.0610, 26.5590)

Settings

Variable	Setting
Age	28
Height	1.23
Weight	43

Settings

Variable	Setting
Age	28
Height	1.23
Weight	43

Prediction

Fit	SE Fit	95% CI	95% PI	
27.0770	0.394521	(26.3019, 27.8520)	(24.2288, 29.9251)	XX

XX denotes an extremely unusual point relative to predictor levels used to fit the model.

Regression Equation

Bmi =43.87 + 0.01817 Age - 25.300 Height + 0.32129 Weight

43.87 is the Y intercept,all equations will start with 45.42.

0.32129 is the Weight Coefficient,multiply it by Weight value.

-25.300 is the Height Coefficient,multiply it by Height value.

0.01817 is the Age Coefficient,multiply it by Age value.

The Standard Error Fit(SE Fit) for all the predicted values are very small numbers so we can say that the prediction model has a more precise predicted mean response.

The Fit values of the predicted unseen data are 29.46,23.81,27.07.The Fit estimates of the mean response for given values of the predictors.

All the Confidence Intervals(CI) and the Prediction Intervals(PI) are not wide so it shows that we need not want to increase the samples in the dataset.

9. Conclusion

The project's main discovery was that the Body Mass Index depends on a number of variables, including weight, height, and age.

The BMI is most influenced by the weight factor.

Additionally, BMI falls as height increases.

Strengths ::

Interpretability : The model can be easily understood and is comparatively simple. Explaining that BMI is determined using specific coefficients based on age, weight, and height is a simple task.

Predictive Power : Based on the provided predictor variables and assuming the model's assumptions are met, the model can yield valuable BMI predictions.

Variable Significance : We can ascertain which factors have the greatest influence on BMI by looking at the coefficients associated with each predictor variable (weight, height, and age).

Limitations ::

Linearity Assumption : The model makes the assumption that the predictor variables and BMI have a linear relationship. In real-world situations, this might not hold true because there could be a nonlinear relationship between these variables and BMI.

Independence Assumption : The predictor variables are assumed by the model to be independent of one another.

This assumption might not hold true if there is a strong correlation between, say, height and weight. This would cause problems with multicollinearity.

Data Quality : The quality of the data affects the results' dependability and accuracy. Results from the model may be skewed by mistakes or anomalies in the data.

Prospective enhancements and additional research fields

Data augmentation : Add more pertinent predictor variables, such as gender, degree of physical activity, food preferences, genetics, and medical background, that may affect BMI.

Model Comparison : To determine whether a more complex model increases predictive accuracy, compare this linear regression model's performance with that of other regression models, such as random forests or logistic regression.

Outside Verification : To determine whether the model is generalizable beyond the current dataset, validate its performance on other datasets.

Biological Considerations : Examine the biological processes that underlie the associations between BMI and the predictor variables. Investigating hormone levels, metabolic variables, and other physiological metrics may be part of this.

References

Data Sources:

https://www.kaggle.com/datasets/rukenmissonnier/age-weight-height-bmi-analysis/

Literature Sources:

PDF'S uploaded in the collpol.

MinitabGettingStarted_EN
Tutorial::(https://www.minitab.com/content/dam/www/en/uploadedfiles/documents/getting-started/MinitabGettingStarted_EN.pdf)

https://www.quora.com/

https://support.minitab.com/en-us/minitab/20

Tools used:

1.MINITAB

2.R STUDIO