National University of Science and Technology, Islamabad

School of Interdisciplinary Engineering and Sciences

Course: Data Analysis and Statistics (CSE-883)

Submitted By: Aliya Aftab

Registration no: 53855

Submitted to: Dr. Zamir Hussain

Linear Regression Models

In the **preprocessing stage** of the analysis, missing values were removed, typographical errors were corrected, and outliers were carefully handled to improve data quality. Categorical variables were encoded appropriately and all measurement units were standardized to ensure consistency across the dataset. Correlation checks were then performed to identify the most influential predictors.

For the **simple linear regression**, the single best predictor was selected based on its correlation with the response variable, and the resulting model was developed by reporting its coefficient, p-value, and R^2 value.

In the **multiple linear regression** phase, all meaningful risk factors were included to build a more comprehensive and reliable model.

Descriptive Statistics: Age, BMI, CVD Risk, WeightKg, HeightCm, Height 2

Results do not use current data. Update these results or create new results using the current data.

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Age	1305	0	56.1992	0.233858	8.44806	40	50	57	63	74
BMI	1305	0	29.1813	0.269531	9.73673	13.5959	25.3244	28.3937	31.9663	320.216
CVD Risk	1297	8	7.59214	0.123239	4.43830	1	4	6	10	28
WeightKg	1305	0	76.0751	0.391281	14.1349	31	66	75	85	166
HeightCm	1305	0	162.159	0.277393	10.0208	72	155	162	170	190
Height 2	1305	0	26395.8	88.6399	3202.10	5184	24025	26244	28900	36100

Variable	IQR	Mode	N for Mode	Skewness	Kurtosis
Age	13	57	68	-0.00	-0.88
BMI	6.64193	29.2101	7	20.90	614.17
CVD Risk	6	6	154	1.08	1.01
WeightKg	19	80	48	0.64	1.63
HeightCm	15	160	63	-0.58	5.07
Height 2	4875	25600	63	-0.06	1.28

In this step, all variables were sorted to organize the dataset properly. Missing values were then identified and removed to ensure clean and reliable data for analysis.

SHEET1

Descriptive Statistics: Sorted Age, Diastolic BP, Systolic BP, Sorted BMI, Sorted CVD Risk, Sorted WeightKg, Sorted HeightCm, Sorted Height 2

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Sorted Age	1297	0	56.1904	0.235065	8.46561	40	49.5	57	63
Diastolic BP	1297	0	80.2467	0.269231	9.69604	35	70	80	90
Systolic BP	1297	0	134.337	0.455608	16.4082	90	120	130	140
Sorted BMI	1297	0	28.8827	0.137875	4.96543	13.5959	25.3333	28.3937	31.9576
Sorted CVD Risk	1297	0	7.59214	0.123239	4.43830	1	4	6	10
Sorted WeightKg	1297	0	75.9491	0.385296	13.8760	31	66	75	85
Sorted HeightCm	1297	0	162.256	0.263716	9.49744	130	155	161	170
Sorted Height 2	1297	0	26417.1	86.0173	3097.82	16900	24025	25921	28900

Variable	Maximum	IQR	Mode	N for Mode	Skewness	Kurtosis
Sorted Age	74	13.5	57	68	-0.00	-0.88
Diastolic BP	120	20	80	472	0.08	0.38
Systolic BP	210	20	140	335	0.47	1.37
Sorted BMI	53.2544	6.62430	29.2101	7	0.63	0.78
Sorted CVD Risk	28	6	6	154	1.08	1.01
Sorted WeightKg	135	19	80	48	0.46	0.58
Sorted HeightCm	190	15	160	63	0.15	-0.51
Sorted Height 2	36100	4875	25600	63	0.28	-0.46

For discrete variables, a tally table was constructed to count the frequency of each value. This helped in understanding the distribution and identifying any unusual patterns in the data.

Tally for Discrete Variables:...

SHEET1

Tally for Discrete Variables: Sorted education, Sorted Gender, Sorted Smoking

Results do not use current data.

Tally

Sorted education	Count	Sorted Gender	Count	Sorted Smoking	Count
Graduation	142	F	636	Cigar	3
M.Phil/Phd	12	M	661	Cigarettes	157
Masters	73	N=	1297	No	1132
No	393			Tobacco	5
NOT DONE	59			N=	1297
Primary	368				
Secondary	250				
N=	1297				

We recoded the discrete variables, including education, smoking status, and gender, to ensure they were properly formatted and suitable for statistical analysis.

SHEET1

Recode

Results do not use current data.

Summary

Original Value	Recoded Value	Number of Rows
Cigar	1	3
Cigarettes	1	157
No	0	1132
Tobacco	1	5

Source data column: Sorted Smoking
Recoded data column: Recoded Sorted Smoking

SHEET1

Recode

Results do not use current data.

on, Sorted Gender, Sorted Smoking

Summary

Original Value	Recoded Value	Number of Rows
F	0	636
M	1	661

Source data column: Sorted Gender
Recoded data column: Recoded Sorted Gender

Recode

Results do not use current data.

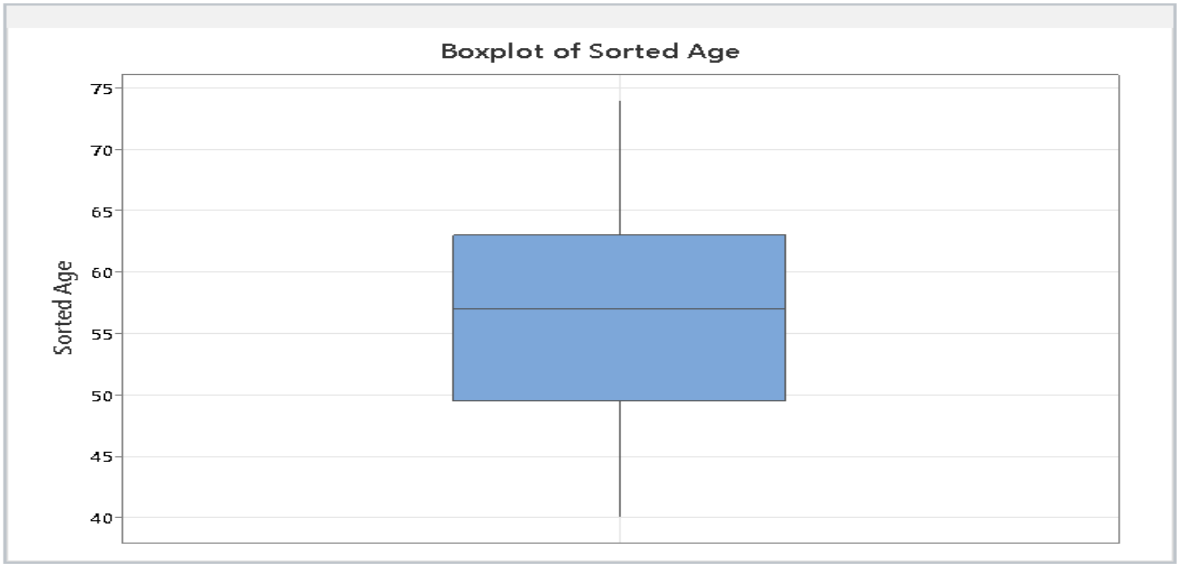
Summary

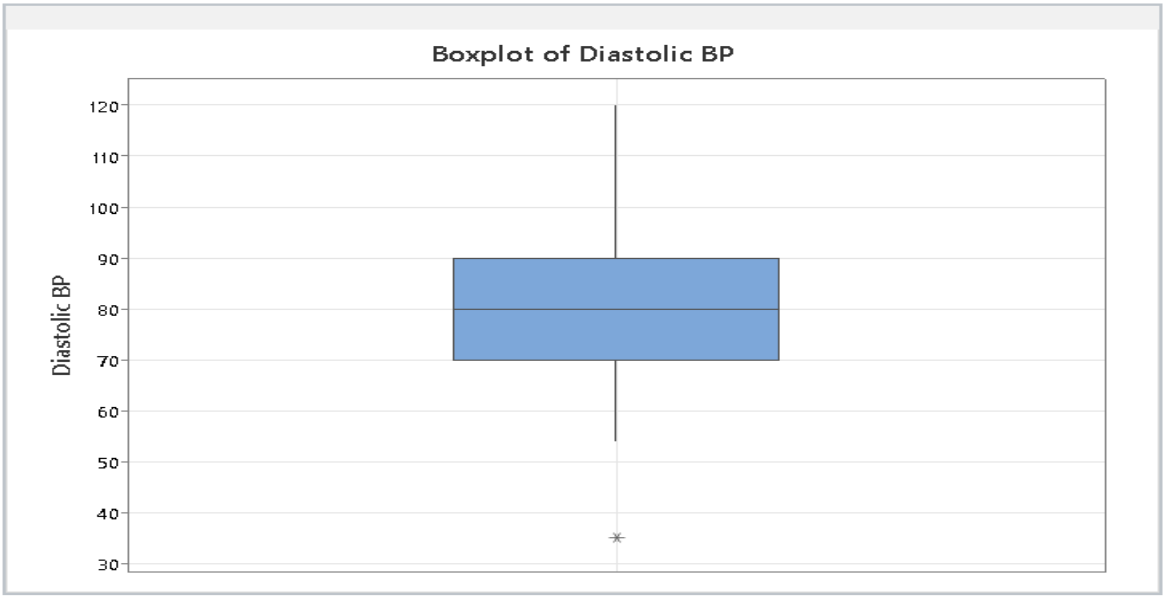
Original Value	Recoded Value	Number of Rows
Graduation	3	142
M.Phil/Phd	4	12
Masters	5	73
No	0	393
NOT DONE	0	59
Primary	1	368
Secondary	7	250

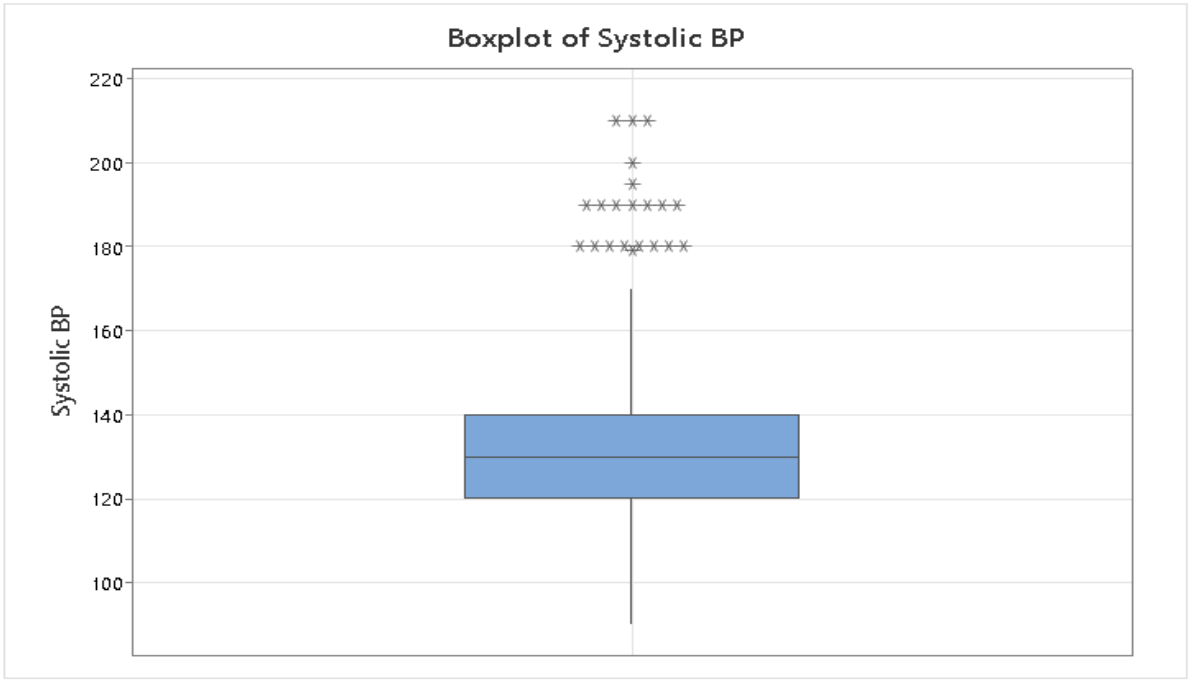
Source data column: Sorted education
Recoded data column: Recoded Sorted education

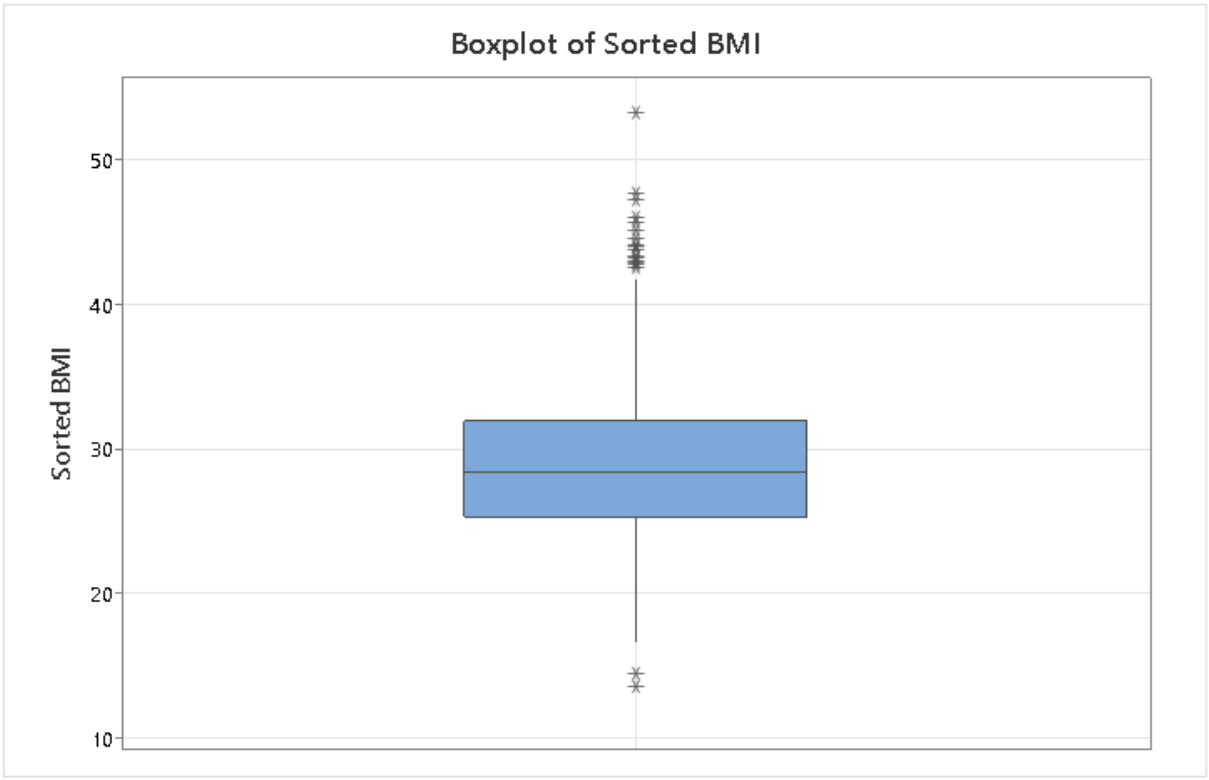
The data was first sorted for better organization. A boxplot was then created to visually detect any outliers.

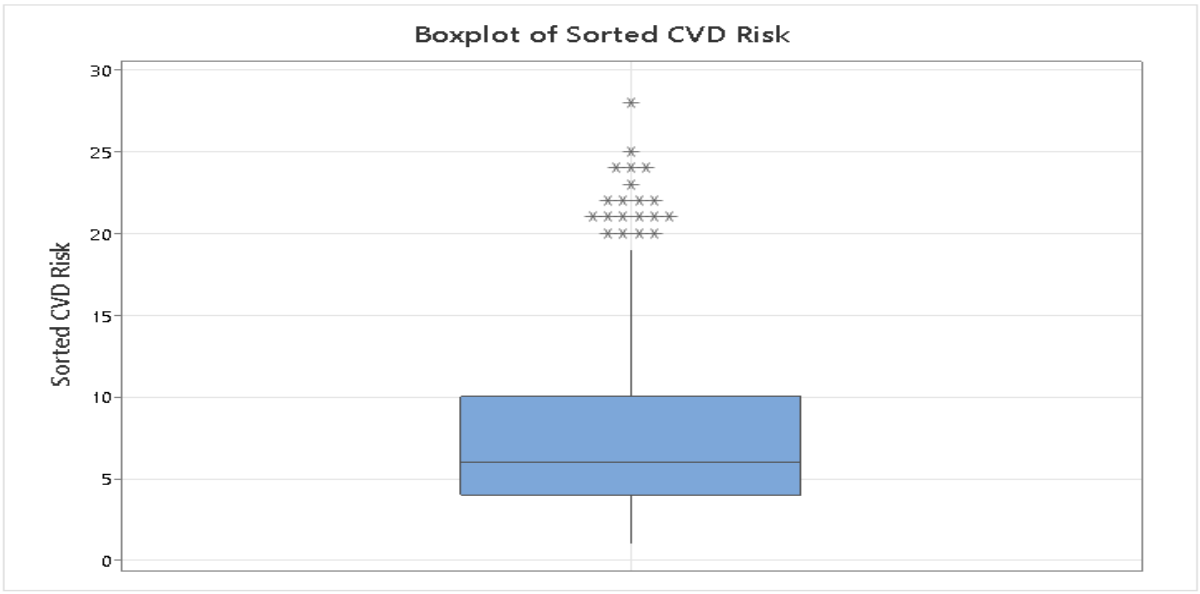
Boxplot of Sorted Age, Diastolic BP, Systolic BP, Sorted BMI, Sorted CVD Risk, Sorted WeightKg, Sorted HeightCm

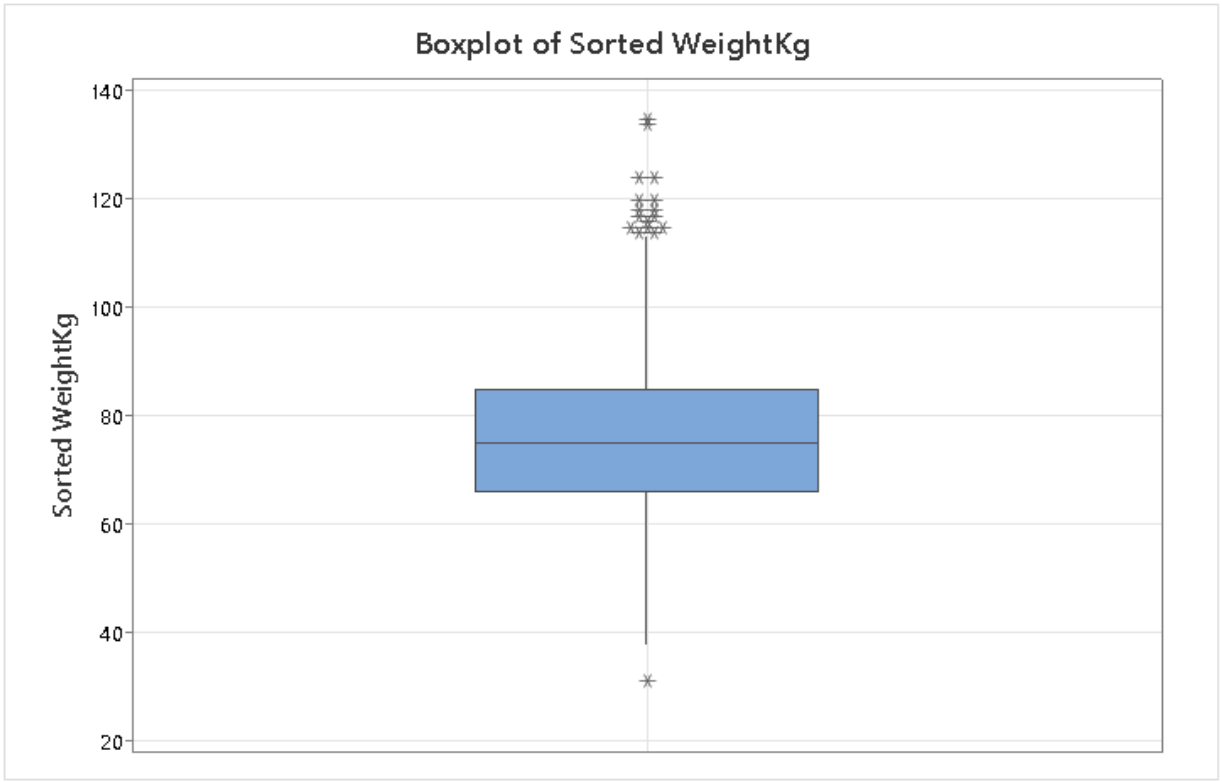


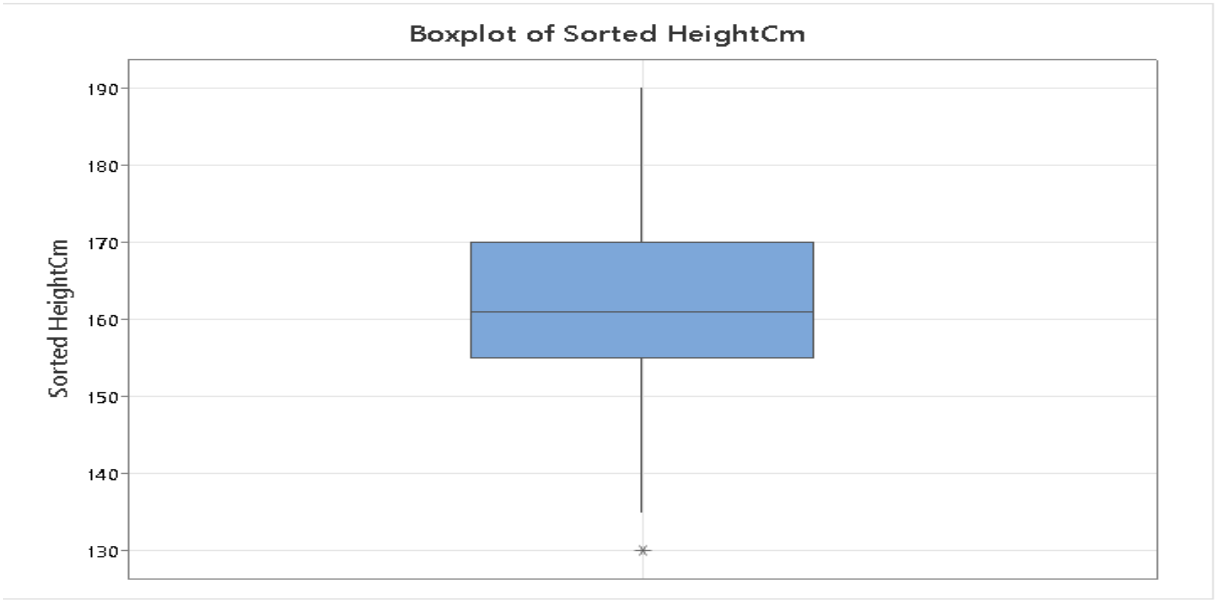



Boxplot of Systolic BP
Systolic BP
220
200
180
160
140
120
100

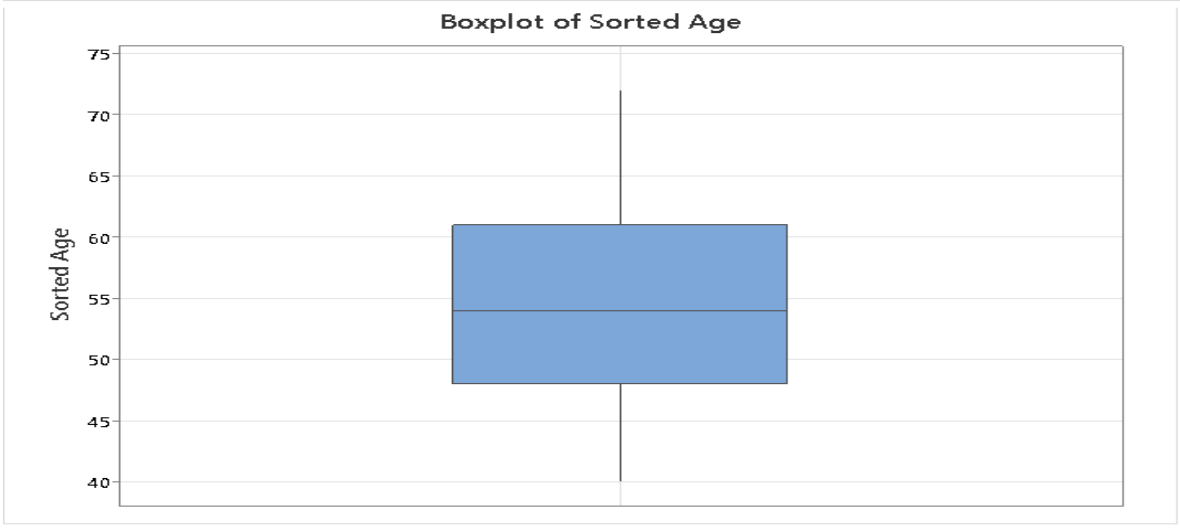

Boxplot of Sorted BMI
Sorted BMI
50
40
30
20
10

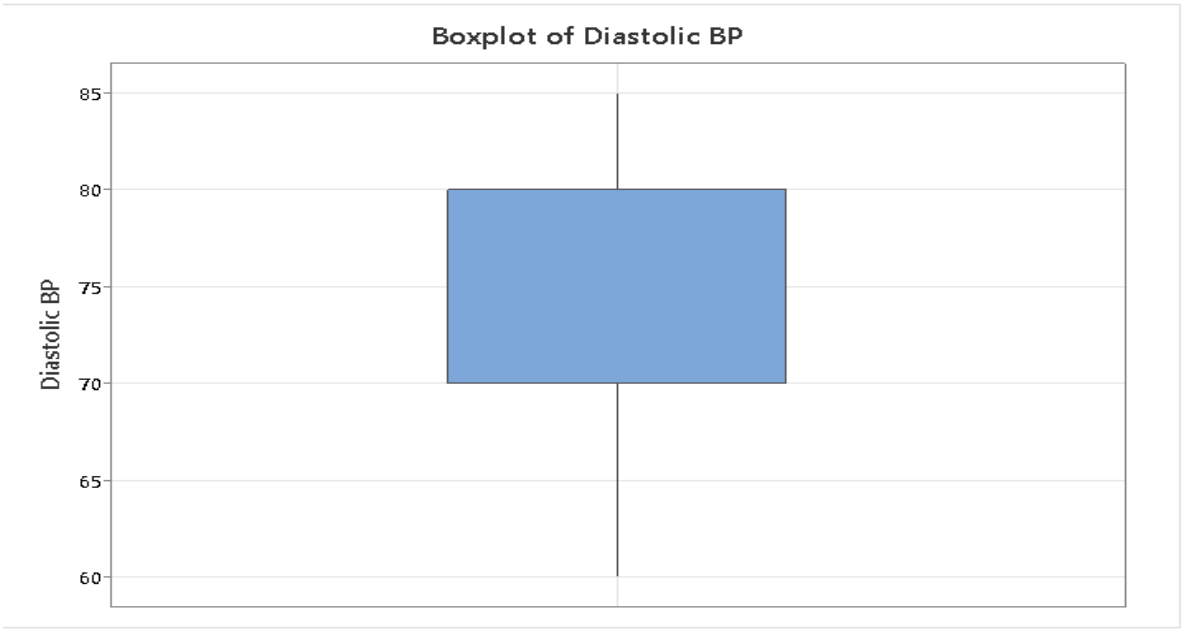

Boxplot of Sorted CVD Risk
Sorted CVD Risk
30
25
20
15
10
5
0

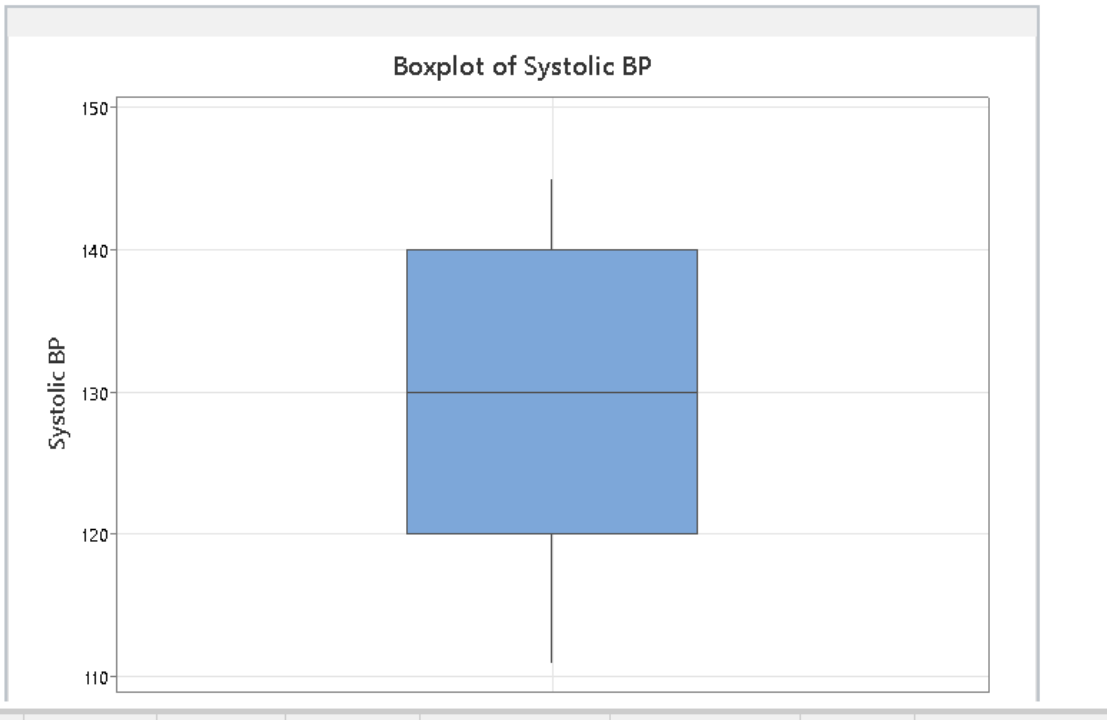

Boxplot of Sorted WeightKg
Sorted WeightKg
140
120
100
80
60
40
20

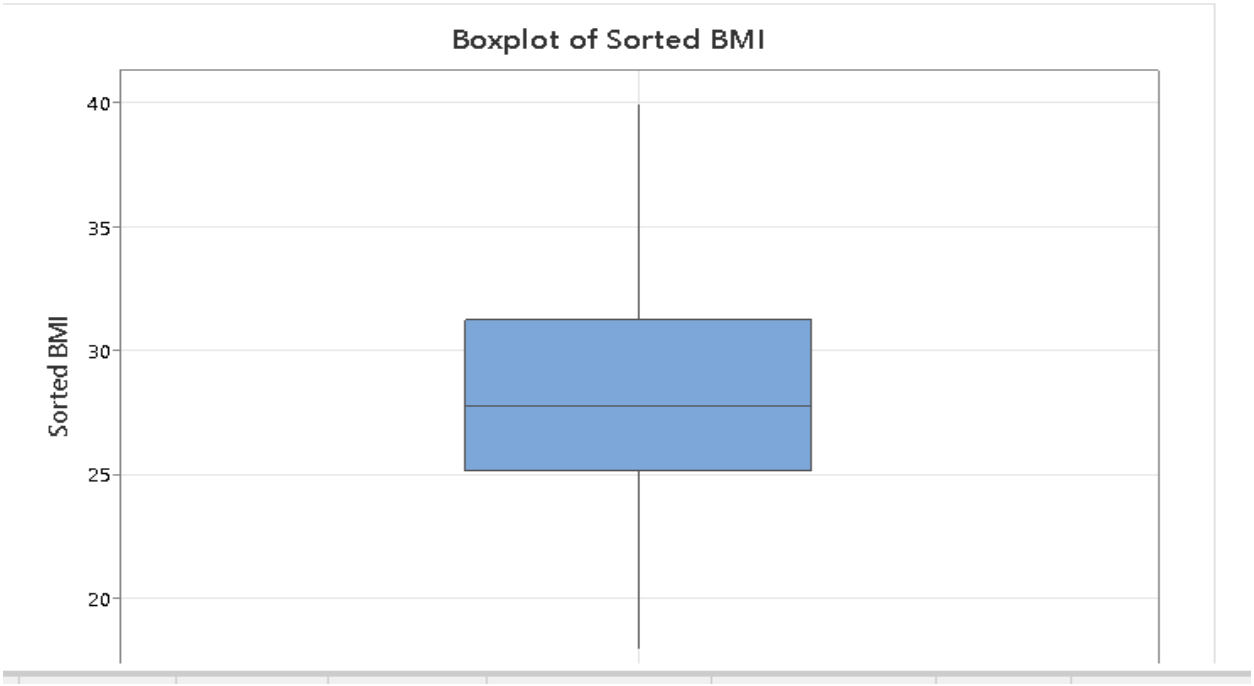


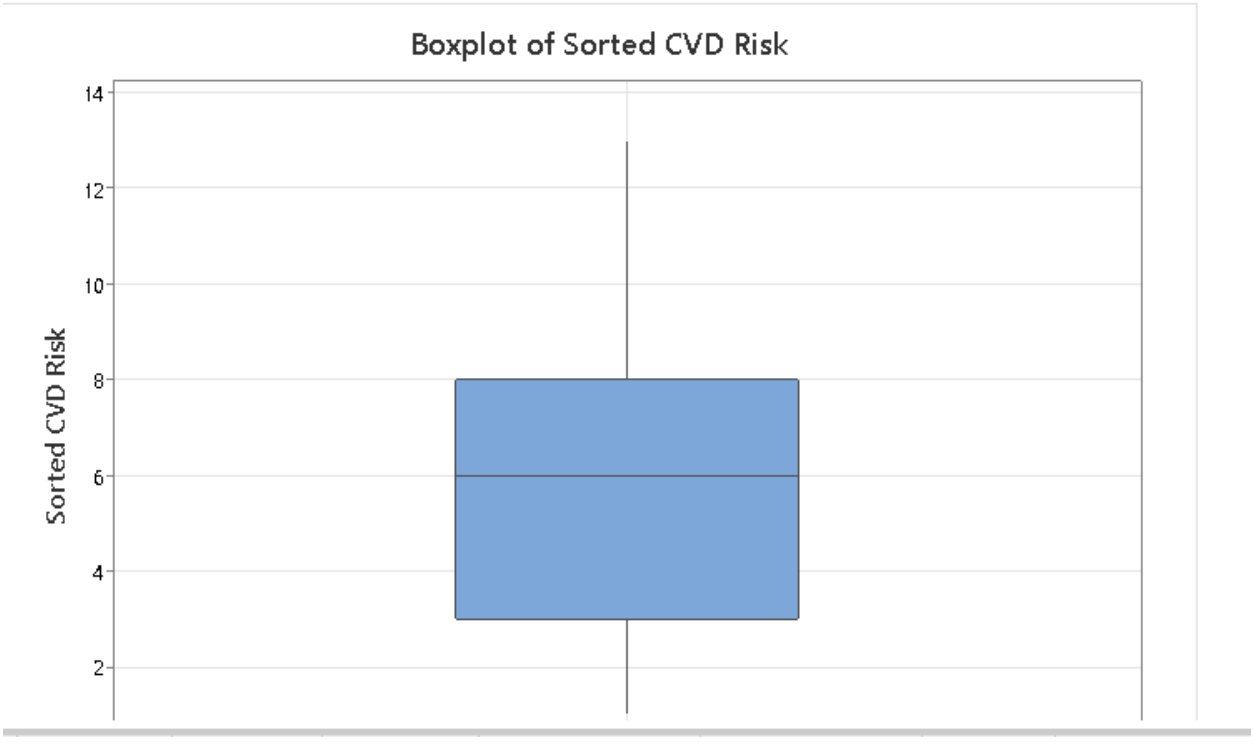
Outlier values were deleted to correct the dataset. This was done using the Subset Worksheet option in Minitab. After this step, all outliers were successfully removed. Boxplots were then recreated to confirm that the dataset was free of outliers.

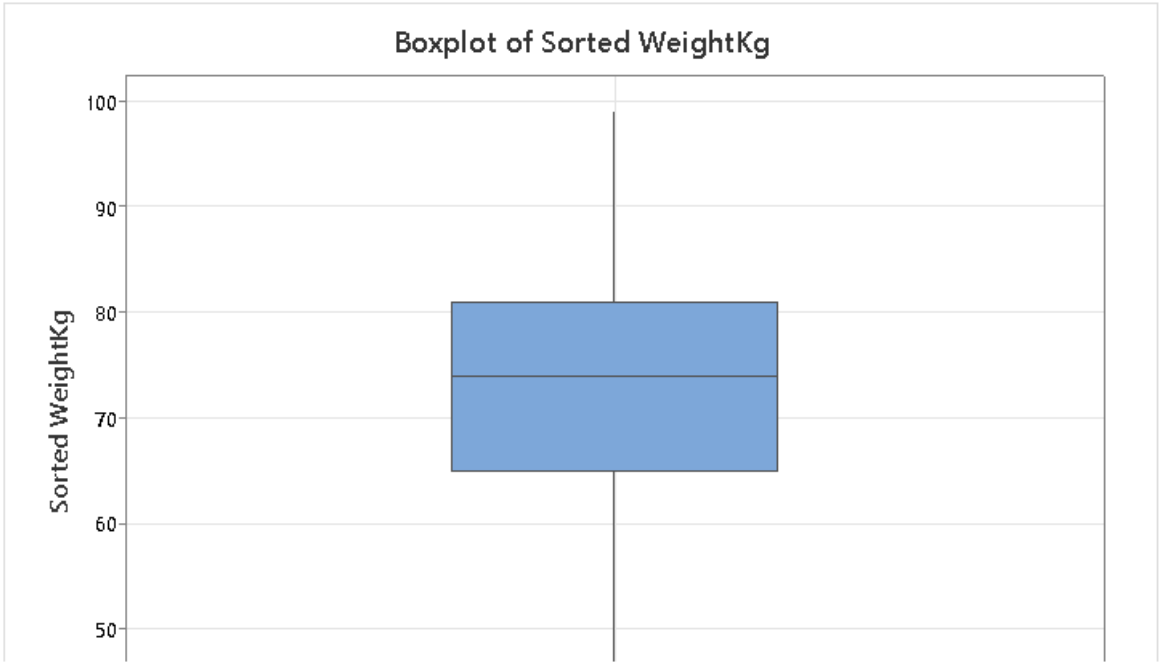


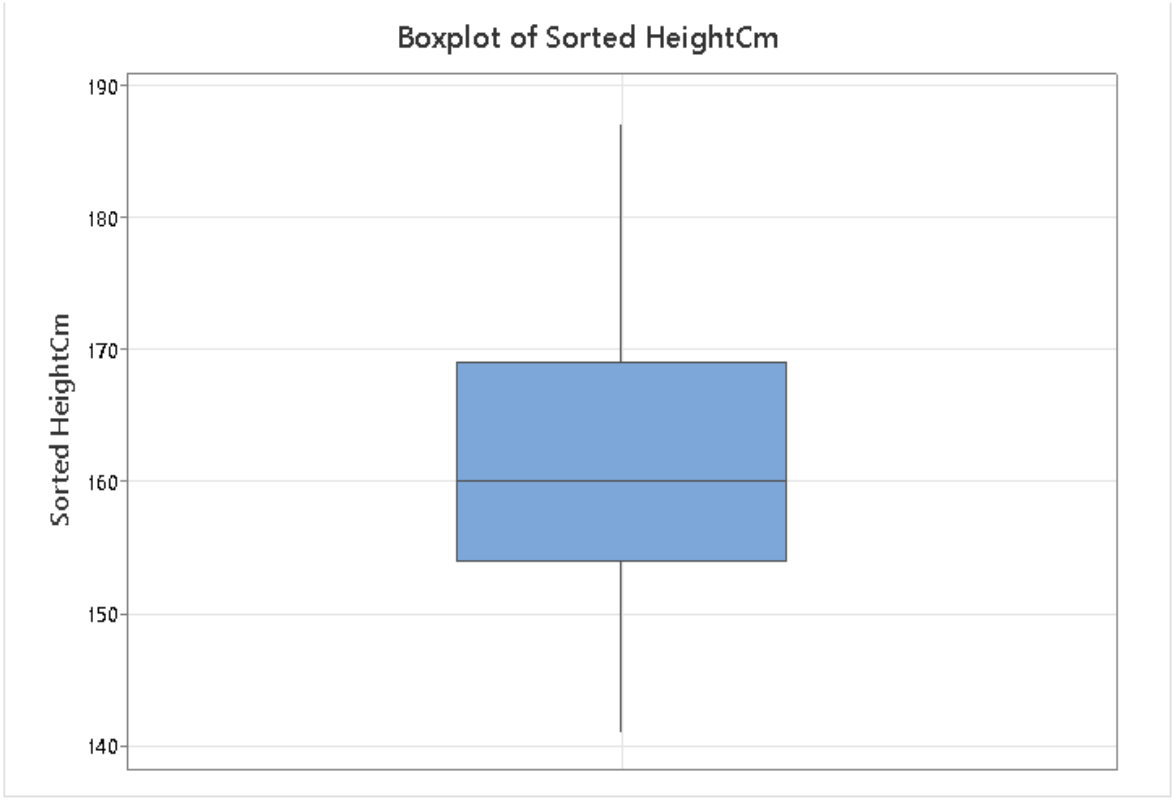

Boxplot of Diastolic BP
Diastolic BP
85
80
75
70
65
60


Boxplot of Systolic BP
Systolic BP
150
140
130
120
110


Boxplot of Sorted BMI
Sorted BMI
40
35
30
25
20


Boxplot of Sorted CVD Risk
Sorted CVD Risk
14
12
10
8
6
4
2


Boxplot of Sorted WeightKg
Sorted WeightKg
100
90
80
70
60
50


Boxplot of Sorted HeightCm
Sorted HeightCm
190
180
170
160
150
140

Building of Multiple Linear Regression Model:

C3 GREATER THAN OR EQUAL TO 111

Regression Analysis: Sorted CVD Risk versus Sorted Age, Diastolic BP, Systolic BP, Sorted BMI, Sorted WeightKg, Sorted HeightCm

Results do not use current data.

Regression Equation

Sorted CVD Risk = -18.46 + 0.31621 Sorted Age + 0.0057 Diastolic BP + 0.05687 Systolic BP - 0.219 Sorted BMI + 0.1043 Sorted WeightKg - 0.0126 Sorted HeightCm

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-18.46	8.56	-2.16	0.032	
Sorted Age	0.31621	0.00804	39.35	0.000	1.07
Diastolic BP	0.0057	0.0116	0.49	0.623	1.10
Systolic BP	0.05687	0.00823	6.91	0.000	1.13
Sorted BMI	-0.219	0.150	-1.46	0.145	100.45
Sorted WeightKg	0.1043	0.0588	1.77	0.077	116.47
Sorted HeightCm	-0.0126	0.0526	-0.24	0.810	60.57

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.49174	75.86%	75.61%	75.24%

Variables with a p-value greater than 0.05 were removed from the analysis. This included diastolic blood pressure, BMI, height, and other non-significant predictors. Only statistically significant variables were retained to build a more reliable regression model.

Regression Analysis: Sorted CVD Risk versus Sorted Age, Systolic BP, Sorted BMI

Regression Equation

Sorted CVD Risk = -17.48 + 0.3166 Sorted Age + 0.0509 Systolic BP - 0.0132 Sorted BMI

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-17.48	2.24	-7.79	0.000	
Sorted Age	0.3166	0.0104	30.51	0.000	1.01
Systolic BP	0.0509	0.0159	3.19	0.002	1.01
Sorted BMI	-0.0132	0.0203	-0.65	0.516	1.00

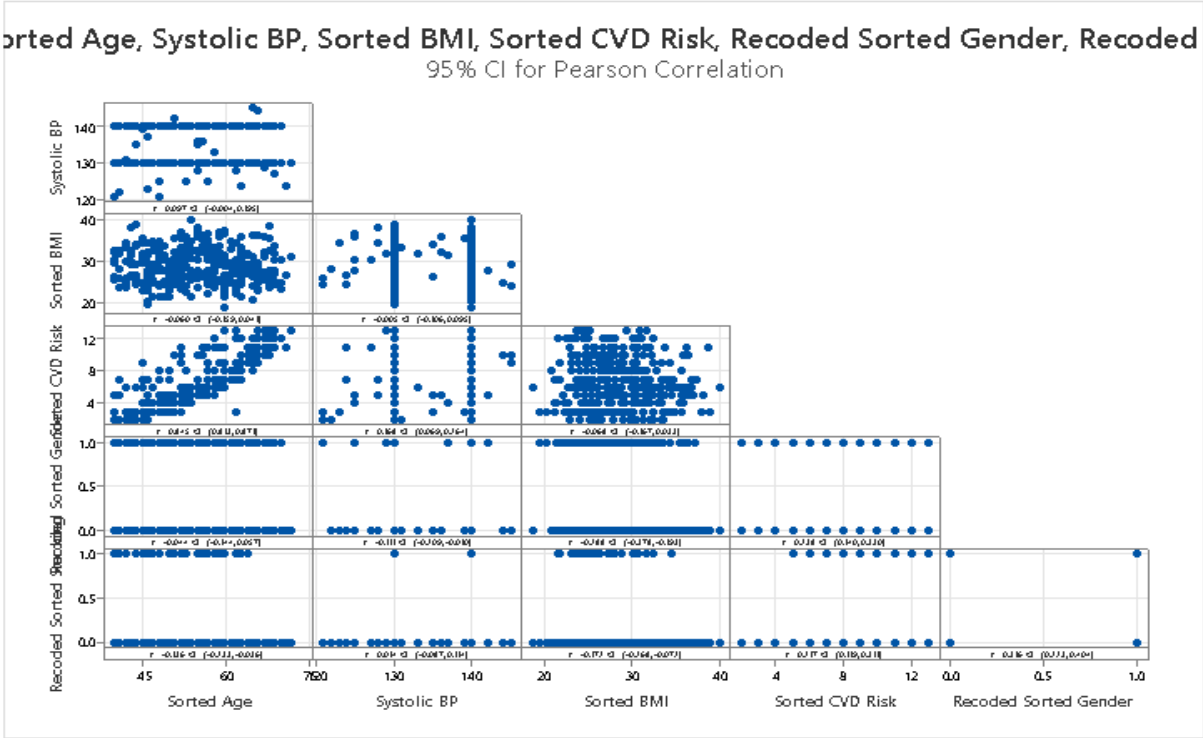
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.61691	72.18%	71.95%	71.61%

Building of Simple Linear Regression Model:

We checked the correlation between each independent variable and the dependent variable to identify the strongest relationship. The variable with the highest and most significant correlation was selected to build the simple linear regression model.

Correlation: Sorted Age, Systolic BP, Sorted BMI, Sorted CVD Risk, Recoded Sorted Gender, Recoded Sorted Smoking



Method

Correlation type	Pearson
Number of rows used	380

Correlations

	Sorted Age	Systolic BP	Sorted BMI	Sorted CVD Risk	Recoded Sorted Gender
Systolic BP	0.097				
Sorted BMI	-0.060	-0.005			
Sorted CVD Risk	0.845	0.168	-0.068		
Recoded Sorted Gender	-0.044	-0.111	-0.288	0.238	
Recoded Sorted Smoking	-0.136	0.014	-0.172	0.217	0.316

After examining the correlation matrix, we found that age has the highest correlation with CVD risk. Therefore, a simple linear regression model was constructed using age as the predictor and CVD risk as the response variable.

Regression Analysis: Sorted CVD Risk versus Sorted Age

Regression Equation

Sorted CVD Risk = -11.243 + 0.3202 Sorted Age

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-11.243	0.584	-19.27	0.000	
Sorted Age	0.3202	0.0104	30.71	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.63524	71.39%	71.31%	71.11%

Interpretation of Results

The simple linear regression model showed that **age** is strongly associated with CVD risk. The multiple regression model retained only statistically significant variables for better accuracy. Non-significant factors such as BMI and height were removed based on p-values greater than 0.05. This helped improve the reliability and interpretability of the final models.

Conclusion

Data preprocessing steps improved the overall quality of the dataset. Regression analysis helped identify important predictors of CVD risk. The models provided a clear understanding of how risk factors influence disease outcomes. Overall, the assignment demonstrated successful application of linear regression techniques.