1. **Description of the Data Set:**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Non-Null Count | Data Type | Description |
| age | 303 | Integer | Age of the individual |
| sex | 303 | Integer | Gender (likely coded as 0 or 1) |
| cp | 303 | Integer | Chest pain type |
| trbps | 303 | Integer | Resting blood pressure |
| chol | 303 | Integer | Cholesterol level |
| fbs | 303 | Integer | Fasting blood sugar (likely coded as 0 or 1) |
| restecg | 303 | Integer | Resting electrocardiographic results |
| thalachh | 303 | Integer | Maximum heart rate achieved |
| exng | 303 | Integer | Exercise-induced angina (likely coded as 0 or 1) |
| oldpeak | 303 | Float | ST depression induced by exercise |
| slp | 303 | Integer | The slope of the peak exercise ST segment |
| caa | 303 | Integer | Number of major vessels colored by fluoroscopy |
| thall | 303 | Integer | Thalassemia (categorical variable) |
| output | 303 | Integer | Target variable indicating  the presence of heart disease. |

The data I have taken is valuable for understanding how age, cholesterol, and exercise influence heart disease. It is complete, with no missing values, and includes numerical, making it suitable for regression analysis, feature exploration, and building predictive models to estimate the presence of heart disease using the given features.

The dataset contains information about 303 individuals. It includes fourteen variables: age, sex, cp, trbps, chol, fbs, restecg, thalachh, exng, old peak, slp, caa, thall, and output. And the description about the variables is written in column 4 which is mentioned above, respectively.

1. **Research Problem:**

**Effects of Lifestyle and Biological Factors:**

•Exploring the association between age, cholesterol, and exercise-induced angina on heart disease.

•Investigating the gender influence on the risk of heart disease.

**Exploring Chest Pain Types:**

•Understanding the relationship between chest pain types (cp) and heart disease.

•Exploring if there are specific chest pain types more significantly associated with higher cholesterol or blood pressure.

**Risk Stratification:**

•Identifying high-risk groups based on combinations of variables such as age, cholesterol, blood pressure, etc.

•Determining thresholds of "old peak" or "thalachh" that significantly elevate risk.

**Feature Interactions:**

•Analyzing how combinations of features (e.g., cholesterol and thalassemia) interact to influence heart disease risk.

1. **Hypothesis:**

Null Hypothesis (H₀): The dependent variable (output) cannot be significantly explained by the independent factors (age, sex, cp, trbps, chol, fbs, restecg, thalachh, exng, old peak, slp, caa, thall) taken together.

H₀: Every regression coefficient is equal to zero.

Alternative Hypothesis(H₁): At least one regression coefficient is not equal to zero.

**Purpose:** The hypotheses will be tested using regression analysis to identify the predictors that significantly influence insurance charges and the hierarchy of their impact.

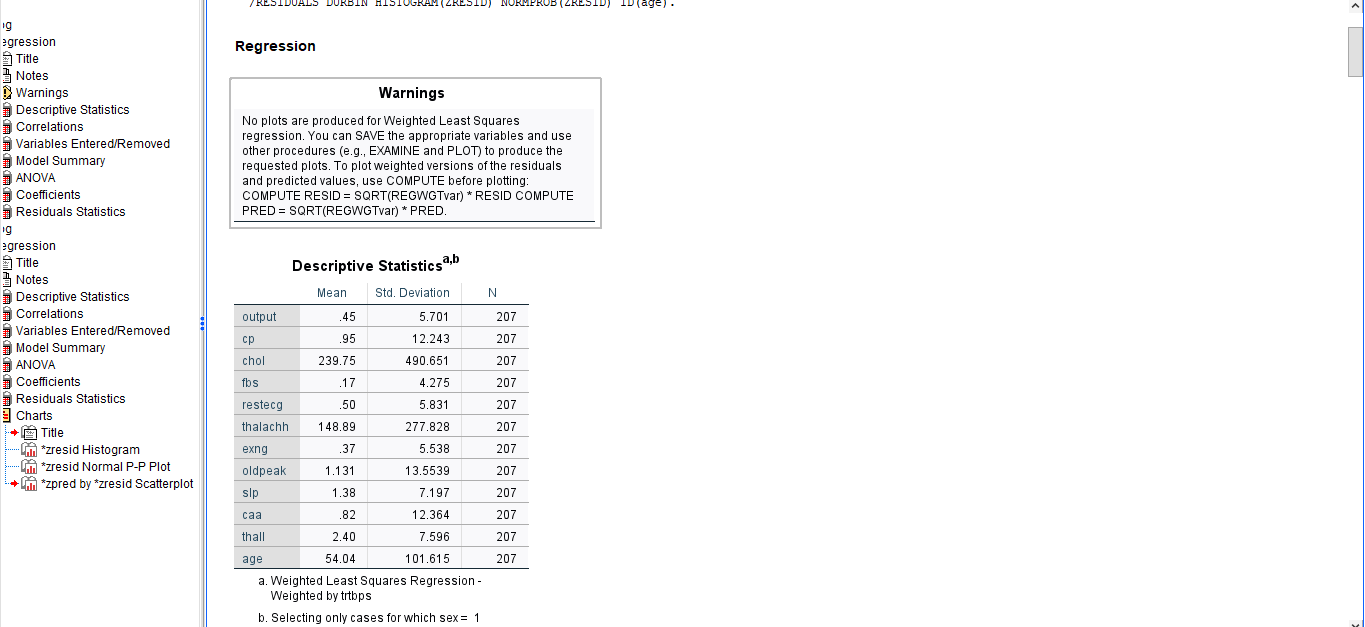
1. **Data-Analysis-Plan:**

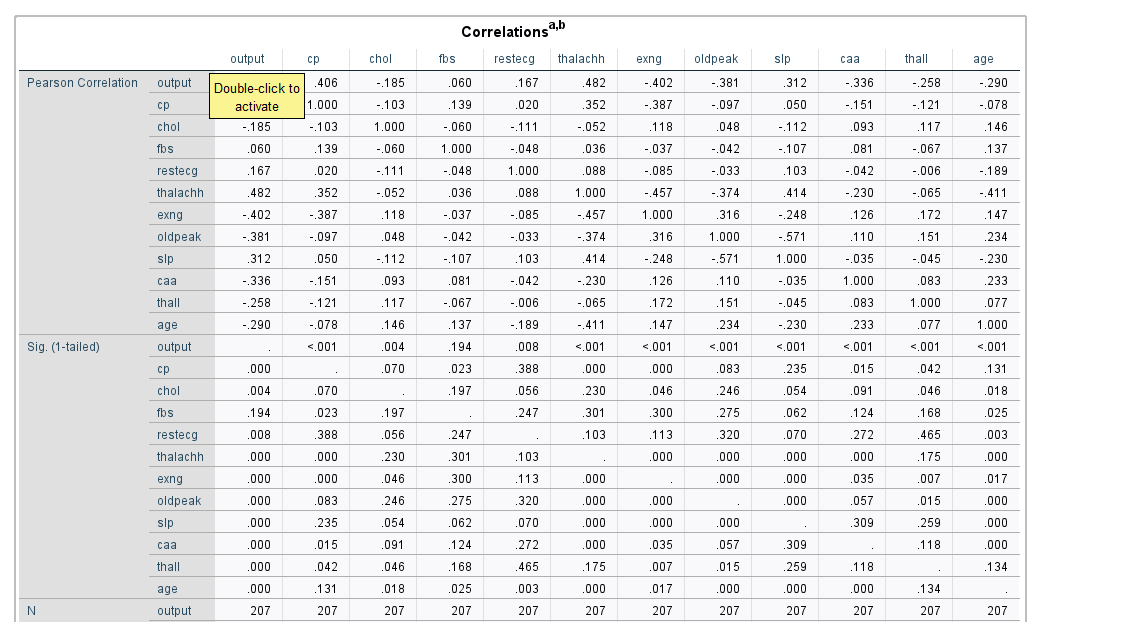
***Exploratory Data Analysis***: The structure, summary statistics, and data types of the dataset will be investigated to know its nature. The distribution of the variables will be checked with histogram, boxplot, and count plots. In this process, missing values will also be checked, and they will be addressed appropriately either through imputation or exclusion. Boxplots will be generated for numerical variables such as charges and BMI to identify outlying values. The effects of removing or transforming the outliers will then be evaluated against the dataset. Additionally, the correlation matrix will be computed for numerical variables to check for multicollinearity, and these relationships will be further depicted with a heatmap.

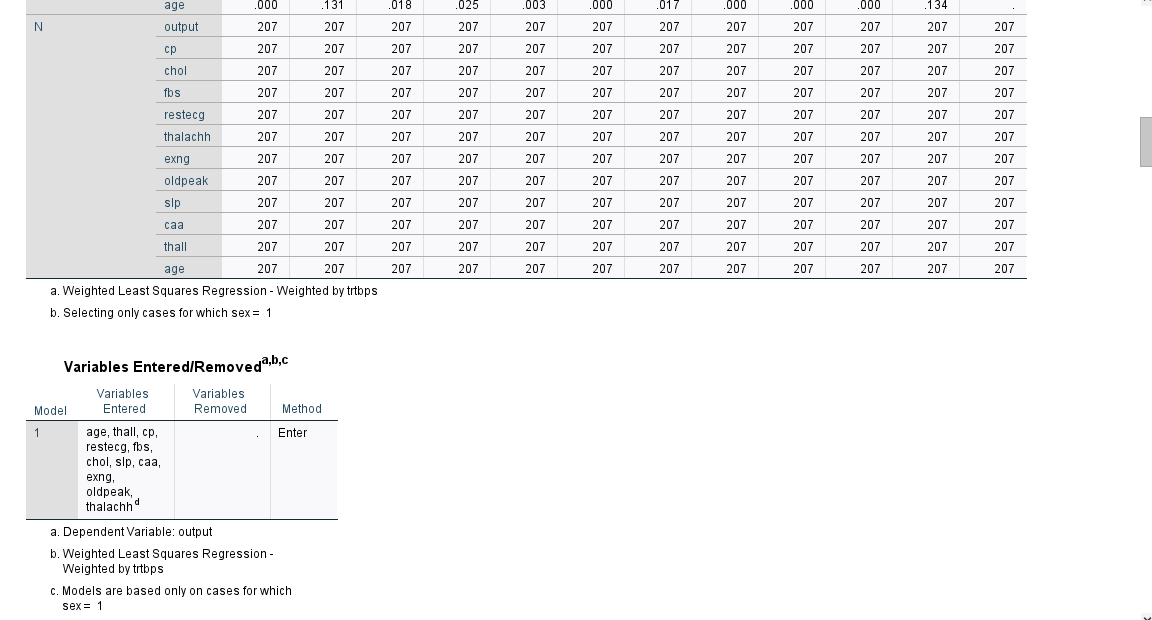
***Regression Analysis:*** This would require creating a linear regression model to predict charges. Here, output is the dependent variable, and the other variables as independent predictors: age, sex, cp, trbps, chol, fbs, restecg, thalachh, exng, old peak, slp, caa, thall. The significant predictors are picked up based on the p-value, where p < 0.05 will be used as a cutoff point. The relative importance can then be ranked by the standardized coefficients or equivalent. Diagnostic checks will be carried out to verify the validity of the regression model. Heteroscedasticity will be tested by plotting residuals against fitted values. Autocorrelation in residuals will be checked by using the Durbin-Watson test.

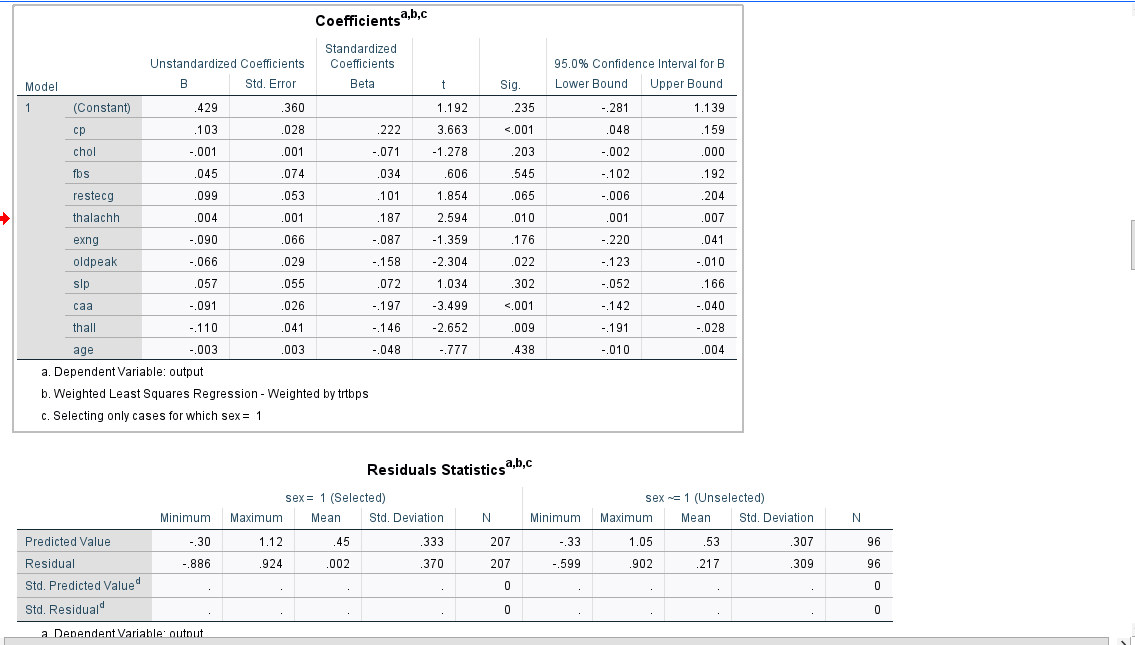
***Final Insights and Recommendations***: Based on the Analysis, we’ll interpret and summarize findings from EDA, ANOVA, and regression analysis. We will also highlight which are the most significant variables affecting charges and their hierarchy of influence.

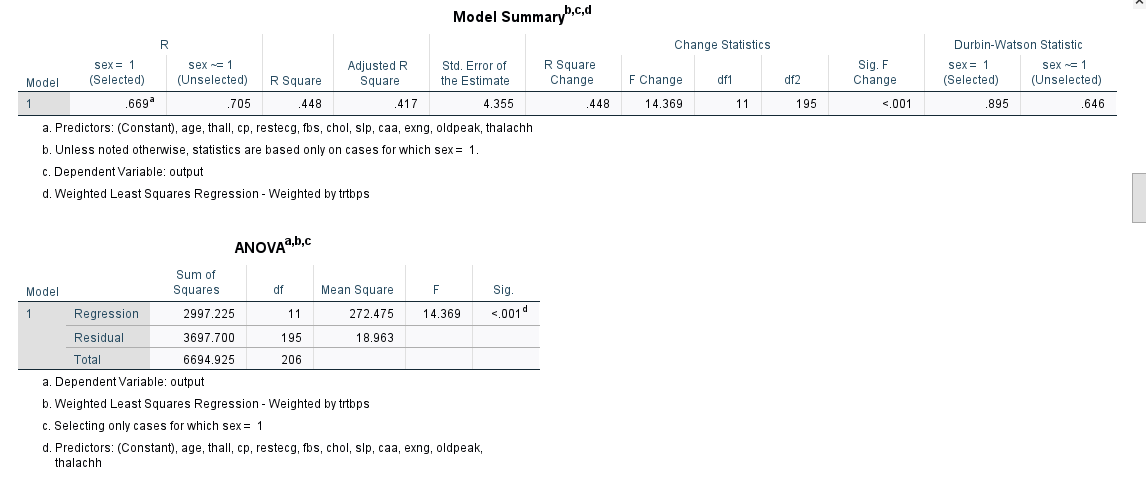
1. Exploratory data analysis for data cleaning and Comprehension Data Analysis:

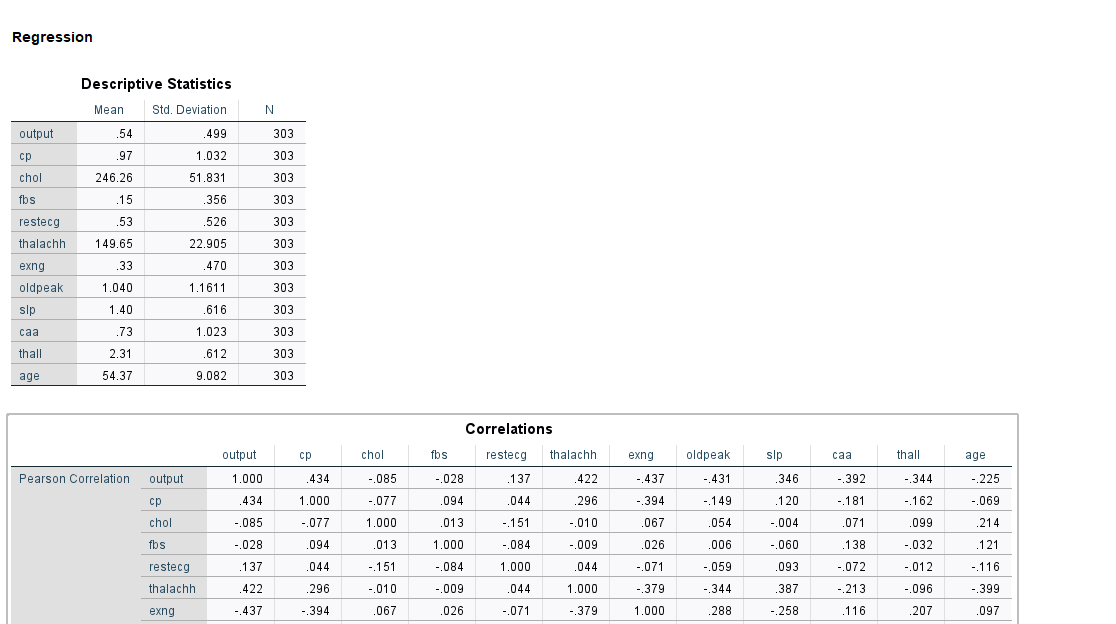


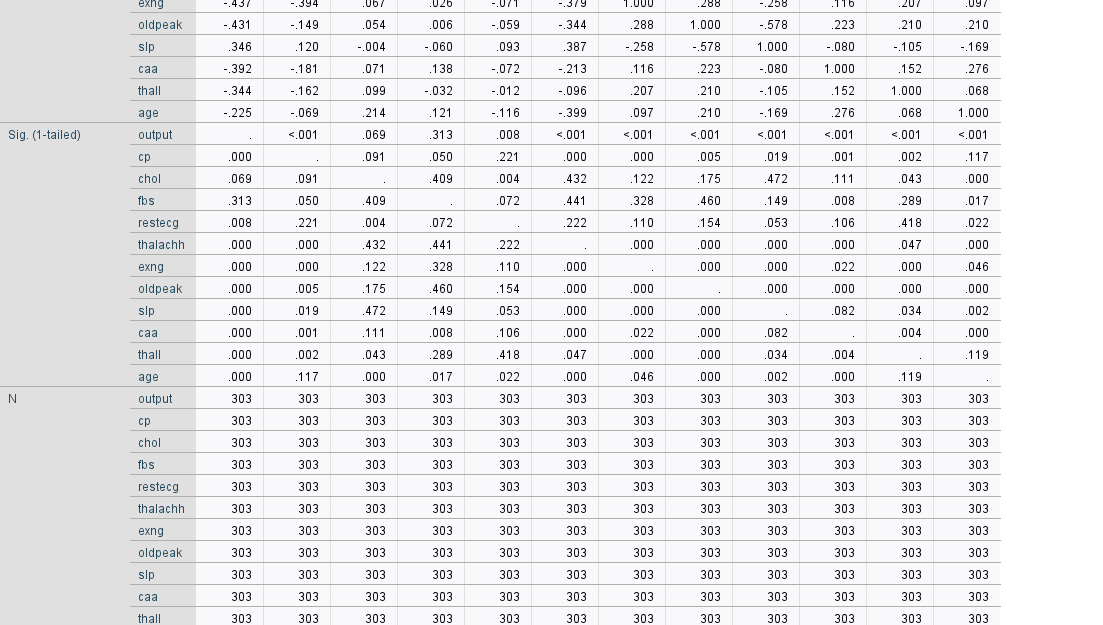


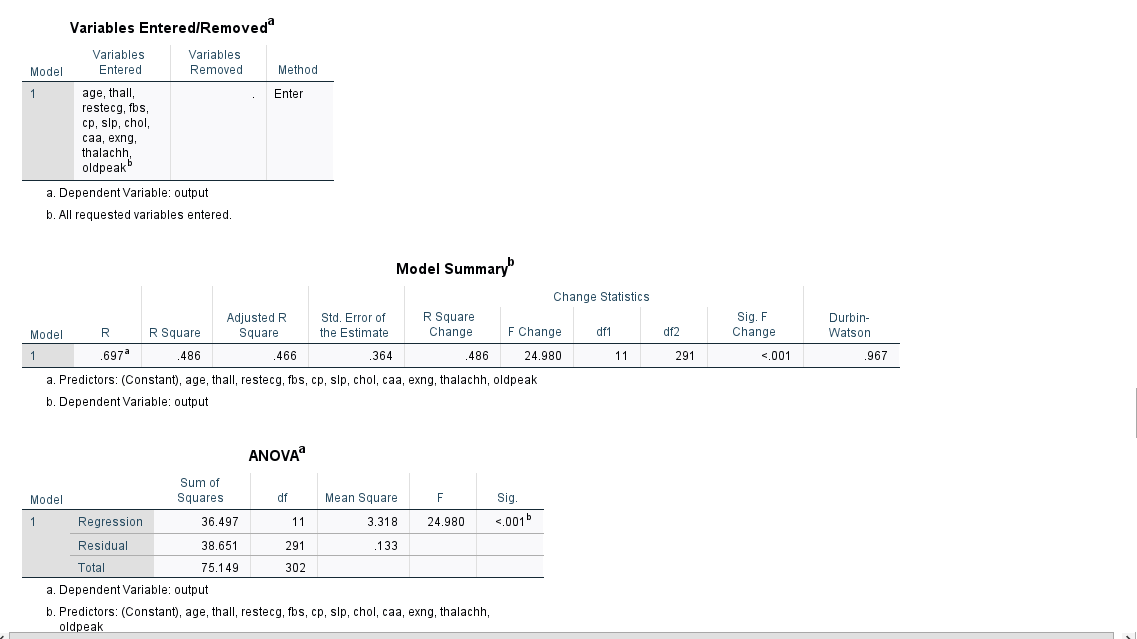


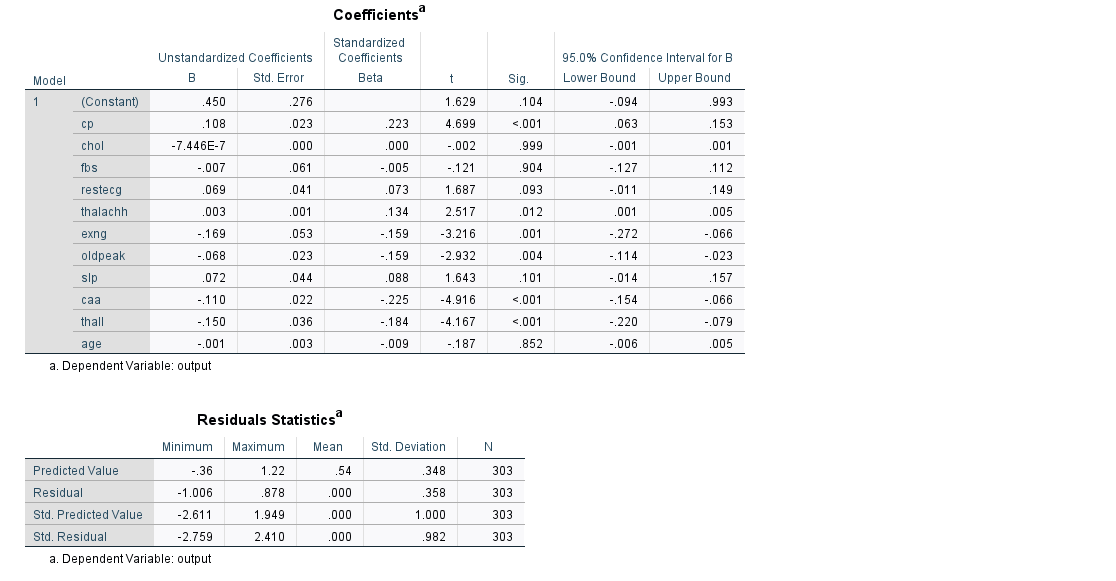


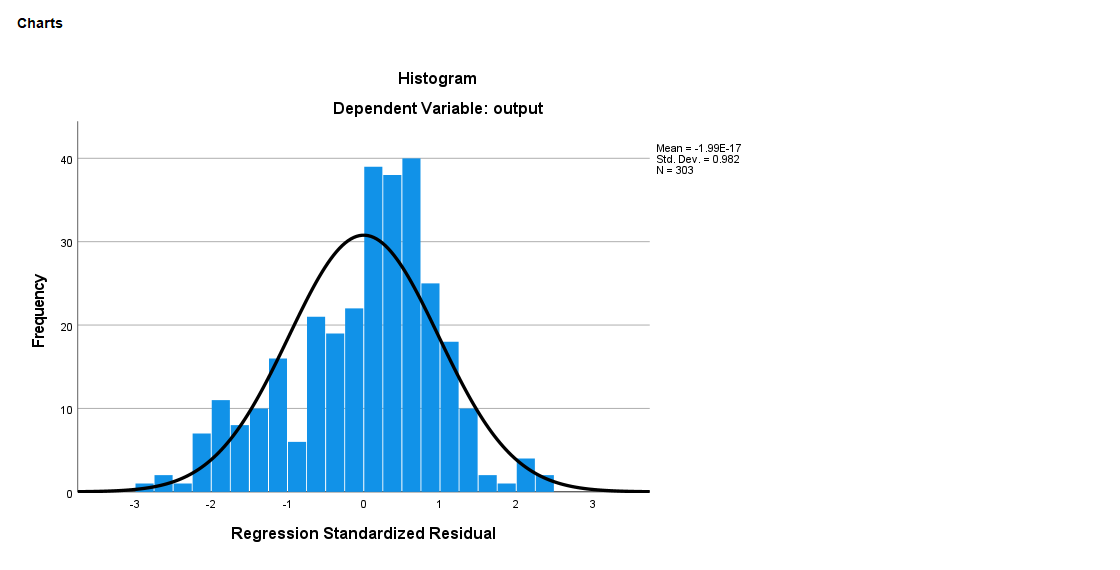


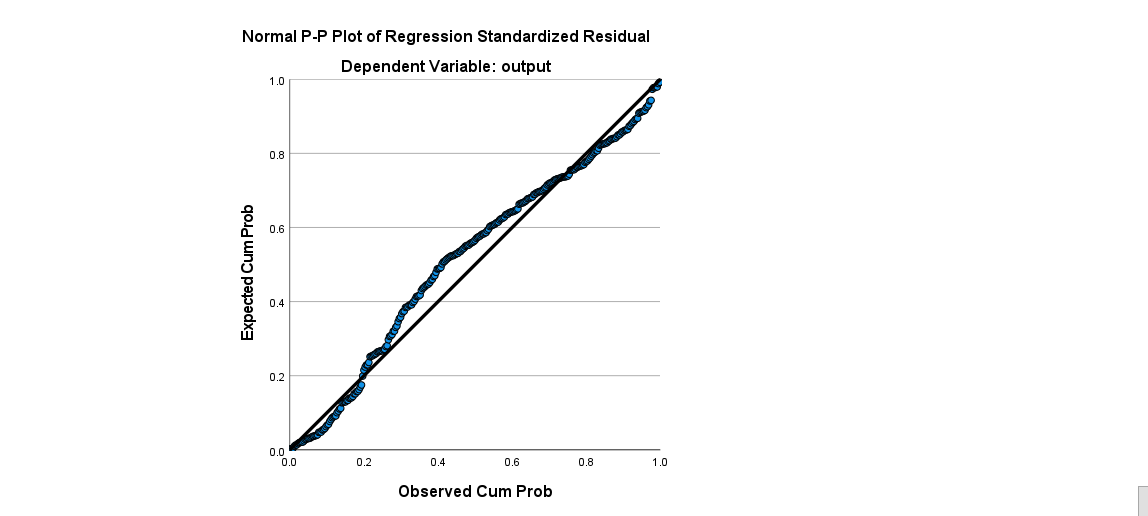


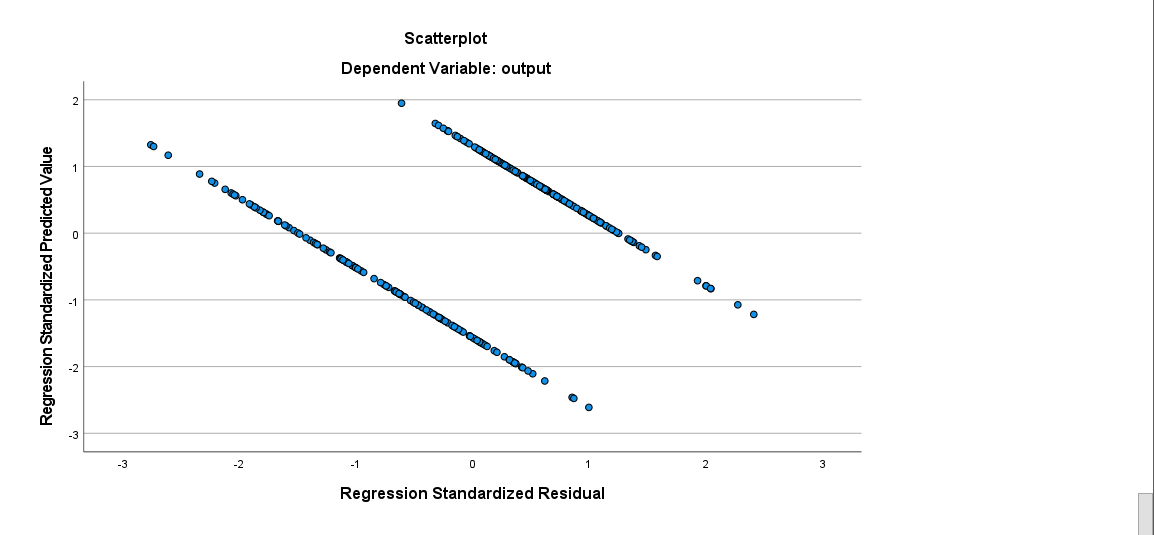












1. Interpretations of results:
2. To check whether the model is significant or not, we focus on the ANOVA table. In the table we check the significance value, and for this model

Let, H0 = Model is Insignificant.

and H1 = Model is Significant.

Now, sig = 0.001 < p = 0.05 ………. Insignificant.

The null hypothesis is rejected, the model is significant.

1. Adjusted R2 = 0.417 = 41.7%

That means 41.7% of the variance in the DV or Outcome Variable is explained by the IVs or predictor variable.

1. Now, we’ll check which are the variables significant, for the hierarchy of the variables.

|  |  |
| --- | --- |
| Variables | Significant Values |
| cp | 0.001 |
| thalachh | 0.010 |
| Oldpeak | 0.022 |
| caa | 0.001 |
| thall | 0.009 |

These variables are significant. Now we’ll order them for check which variable is the most impactful.

So,

**For Hierarchy:** we check the Beta Value of these variables.

The bigger the Beta, the Higher the Impact.

Always check magnitude value.

|  |  |  |
| --- | --- | --- |
| Variables | Beta value | Order |
| cp | 0.222 | I |
| thalachh | 0.187 | III |
| oldpeak | -0.158 | IV |
| caa | -0.197 | II |
| thall | -0.146 | V |

The biggest beta value is of “cp” i.e. 0.222, followed by caa, thalachh, oldpeak, and thall. Hence, cp is the most impactful variable.

This means that, if a person feels chest pain then it is a maximum possibility that the person is a heart patient or higher chance of getting a heart attack.

1. Conclusions and Recommendations:

From our analysis, we can conclude that certain factors i.e. “cp” play a critical role in determining the heart disease patient, while others are less impactful. Among these factors, CP stands out as the most significant.

To reduce heart disease and chest pain, patients should start morning and evening exercise.

And start eating a healthy diet, less workload, etc.

1. Limitations:

* The dataset should be large.
* The categorical variables, such as sex, were treated as fixed categories without exploring potential interactions or more granular subcategories, which might have revealed subtle effects.
* The analysis relies on a linear regression model, assuming linear relationships between predictors and the target variable. However, real-world relationships might be non-linear, leading to potential inaccuracies.
* Interaction effects between variables, such as how Chest pain and Cholesterol level might jointly impact charges, were not explored. This could provide deeper insights.