Department of Electrical and Electronics Engineering

GE 461: Introduction to Data Science

Class Project I Report

Name and Surname: Ali Aral Takak

Student ID: 22001758

Department: EEE

Introduction

The first project of the course focuses on linear regression, a simple yet effective method of supervised learning. Students are required to solve the problems **3.7.8** and **3.7.9** from the book *"An Introduction to Statistical Learning with Applications in Python and R"*, written by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor.

Before moving further in the implementations and results of the problems, the report will briefly explain what linear regression is, with provided formulations and visualizations. Then, we will report our findings on the specified questions, analyze our results, and conclude the results.

Overview of Linear Regression

Linear regression is a statistical method that aims to model the relationship between two variables by fitting a linear equation to a set of observed data. In simple linear regression model, we have a quantitative response, denoted as Y, on the basis of a single predictor variable, X [1]. Mathematically, one can define this relationship as:

$$Y \approx \beta_0 + \beta_1 X$$

Where β_0 is named as the intercept term, which eliminates origin bias from the linear model, and β_1 is named as the slope. Together, these terms are labeled as the **model coefficients** or **parameters** [1]. One can generate an approximation after using the training data in order to produce estimates, and define a new equation as:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

In the equation provided above, $\hat{y}$ represents our prediction of Y on the basis of $X = x$, with the estimate coefficients $\widehat{\beta_0}$ and $\widehat{\beta_1}$. Now, a new question arises from these definitions: We know that in practice, β_0 and β_1 are unknown. Hence, how can we learn these parameters in order to make reliable estimations?

Let us define some data. Let:

$$(x_1, y_1), \dots (x_n, y_n)$$

Represent n observation pairs, in which each consists of a measurement of X and a measurement of Y. We desire to obtain coefficient estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$, so that the linear model that we will define will fit the data well. The most common method that is used is minimizing the least squares criterion [1]. Let

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

Denote the prediction for Y based on the i^{th} value of X. In this case, one can calculate the difference between the observed value and the true value as:

$$e_i = y_i - \widehat{y_i}$$

In this case, the total error for all observations, or the Residual Sum of Squares (RSS) can be calculated as:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

One can rewrite the Residual Sum of Squares as:

$$RSS = (y_1 - \widehat{\beta_0} - \widehat{\beta_1} x_1)^2 + (y_2 - \widehat{\beta_0} - \widehat{\beta_1} x_2)^2 + \cdots + (y_n - \widehat{\beta_0} - \widehat{\beta_1} x_n)^2$$

Finally, one can determine the model coefficients as:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Where $\bar{y}$ and $\bar{x}$ are sample means, defined as follows:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Although simple linear regression is an effective method, in realistic models, there exists more than one predictor variable. In this scenario, we need to extend our model to be constructed from multiple predictors. Hence, we take a closer look to multiple linear regression model. Suppose that we have k distinct predictor variables. Then, we can write the multiple linear regression model as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

In a similar manner, the regression coefficients are unknown and must be estimated. Given estimates $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}$, we can make predictions by:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_k} x_k$$

The Residual Sum of Squares can be written as:

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \cdots - \widehat{\beta_k} x_{ik})^2$$

Problem I

The first problem involves the use of simple linear regression on the Auto data set, and expects students to answer the questions given below:

1. **Use the sm.OLS() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summarize() function to print the results. Comment on the output. For example:**

a) **Is there a relationship between the predictor and the response?**

After fitting our, model, we have obtained the regression line which can be observed in the figure provided below:

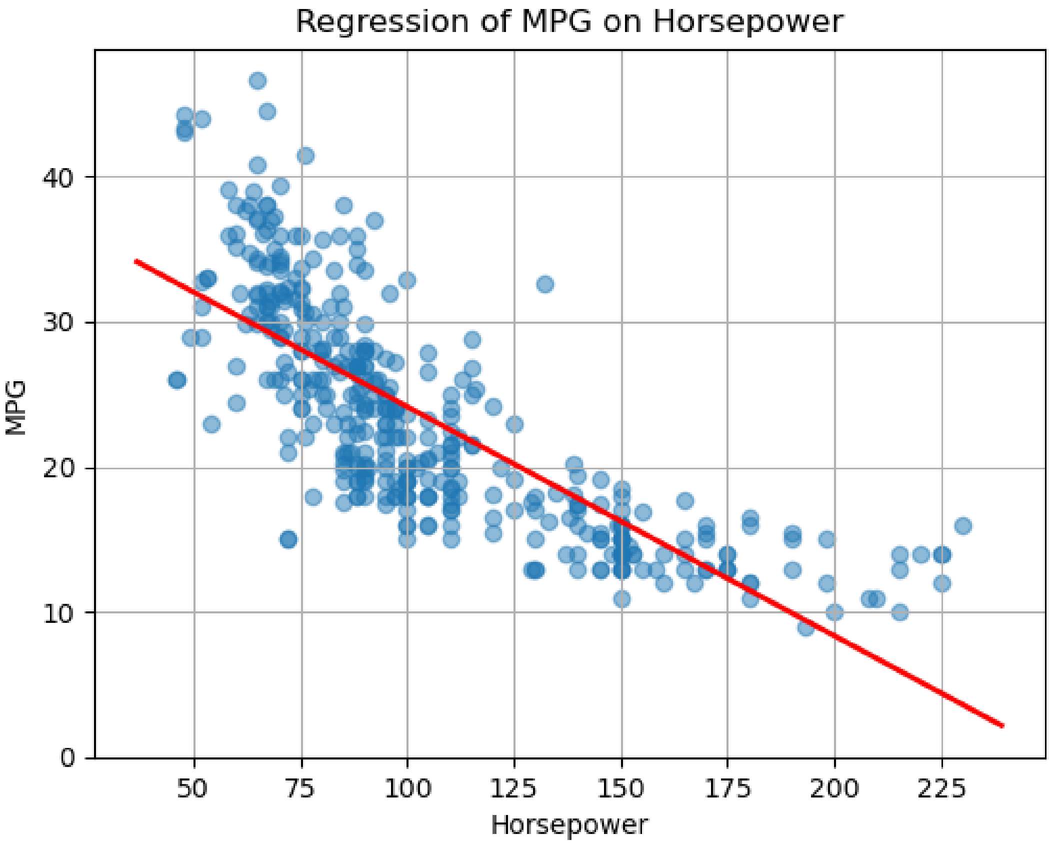
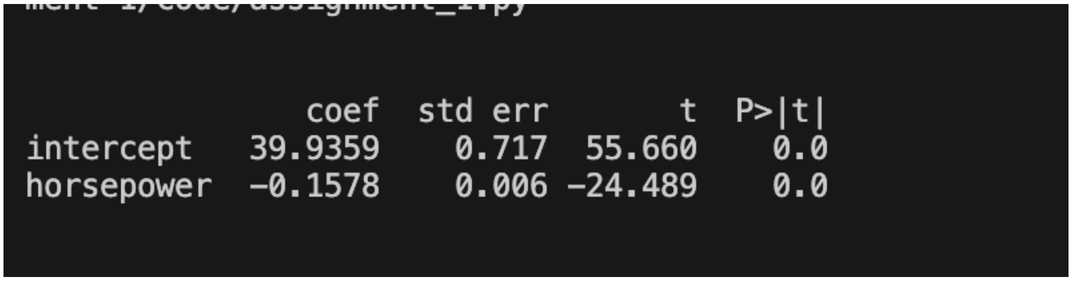


Figure 1: Generated simple regression line.

Then, we have determined our model parameters and various statistical measurements as provided in the figure below:



```
ment 1/code/assignment_1.py


                  coef  std err        t  P>|t|
intercept      39.9359    0.717   55.660    0.0
horsepower     -0.1578    0.006  -24.489    0.0
```

Figure 2: Model parameters along with statistical measurements.

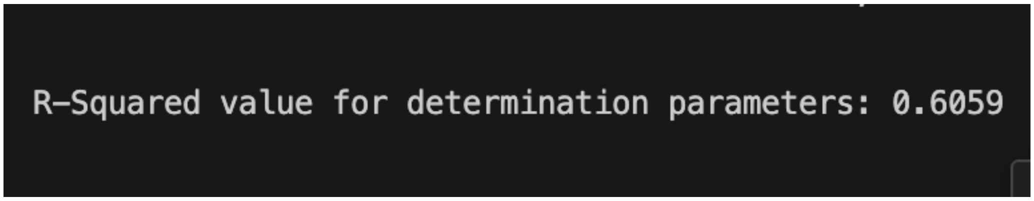
Hence, we can determine our simple linear regression model as:

$$Y = 39.9359 + (-0.1578)X_1$$

b) How strong is the relationship between the predictor and the response?

In order to determine the strength of the relationship between the predictor and the response, we can control two variables, namely the R-squared value and the p-value. The p-value can be observed in *Figure 2,* and it is determined as 0.0 for *horsepower*. This indicates that statistically, there is a strong relationship between mpg and horsepower parameters.

Additionally, we can check the R-squared value too. The figure given below displays the R-squared value:



```
R-Squared value for determination parameters: 0.6059
```

Figure 3: R-squared value between the parameters.

This also indicates that there is a strong relationship between our parameters, mpg and horsepower.

c) Is the relationship between the predictor and the response positive or negative?

In order to answer this question, we have two different observations to make. First of all, the horsepower coefficient, in other words, the slope of the model, is given as a negative coefficient. Hence, an increase in horsepower would gradually decrease the value of the miles per gallon. Second of all, investigating the generated plot confirms our intuition of negative impact on miles per gallon. Hence, the relationship between the predictor and the response is negative.

d) What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

We can determine the predicted miles per gallon value for a horsepower of 98 manually. Given our model:

$$Y = 39.9359 + (-0.1578)X_1$$

When we plug in the value $X_1 = 98$, we can obtain that $Y = 24.4715$. We can also inspect our plot to confirm our results. The figure given below displays the values at $X \approx 98$:

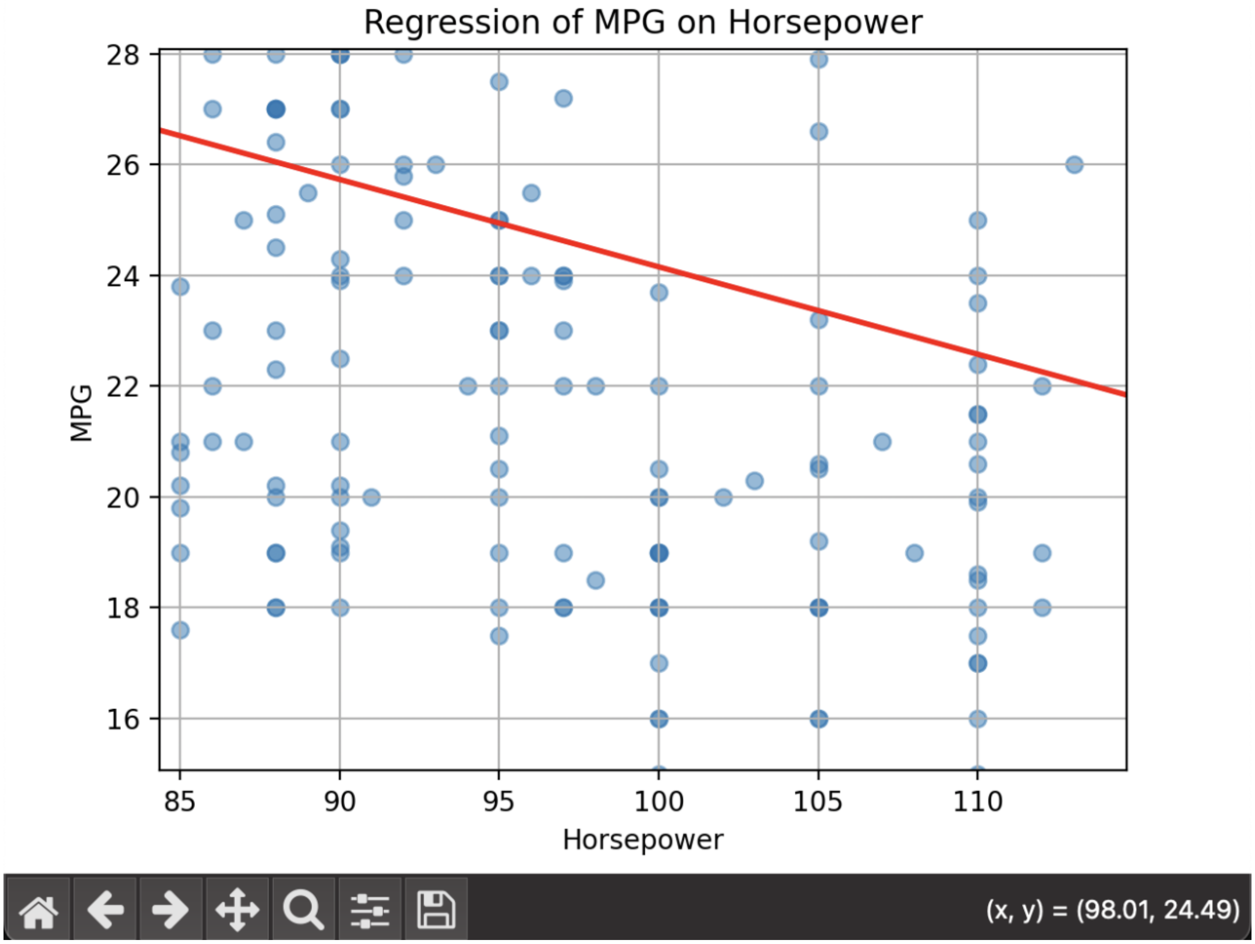
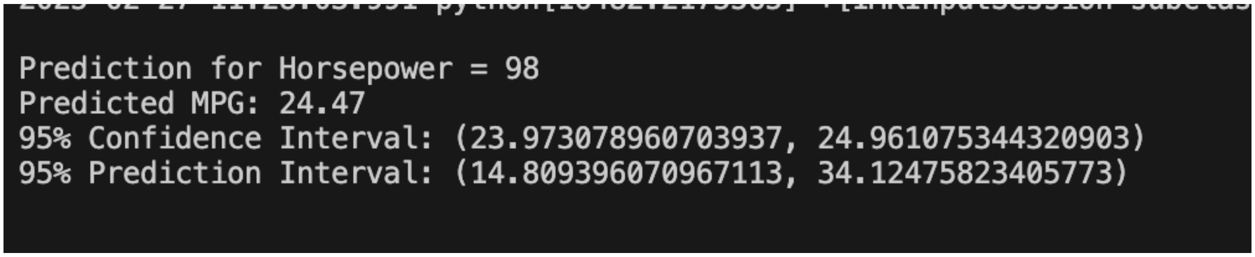


Figure 4: Regression line results at (x,y)=(98.01,24.49) [check right below corner].

The figure given below displays the calculations made via Python in order to determine the prediction at $X = 98$, and the corresponding 95% correspondence and prediction intervals:



```
2025-02-27 11:28:05.591 python[10482:2179363] +[IMKInputSession subclas

Prediction for Horsepower = 98
Predicted MPG: 24.47
95% Confidence Interval: (23.973078960703937, 24.961075344320903)
95% Prediction Interval: (14.809396070967113, 34.12475823405773)
```

Figure 5: Predictions and statistics for 98 horsepower.

Problem II

The second question involves the usage of multiple linear regression on the Auto data set, and expects students to answer the questions provided below:

a) Produce a scatterplot matrix which includes all of the variables in the data set.

The figures given below display the scatterplots pairwise for all variables in the data set:

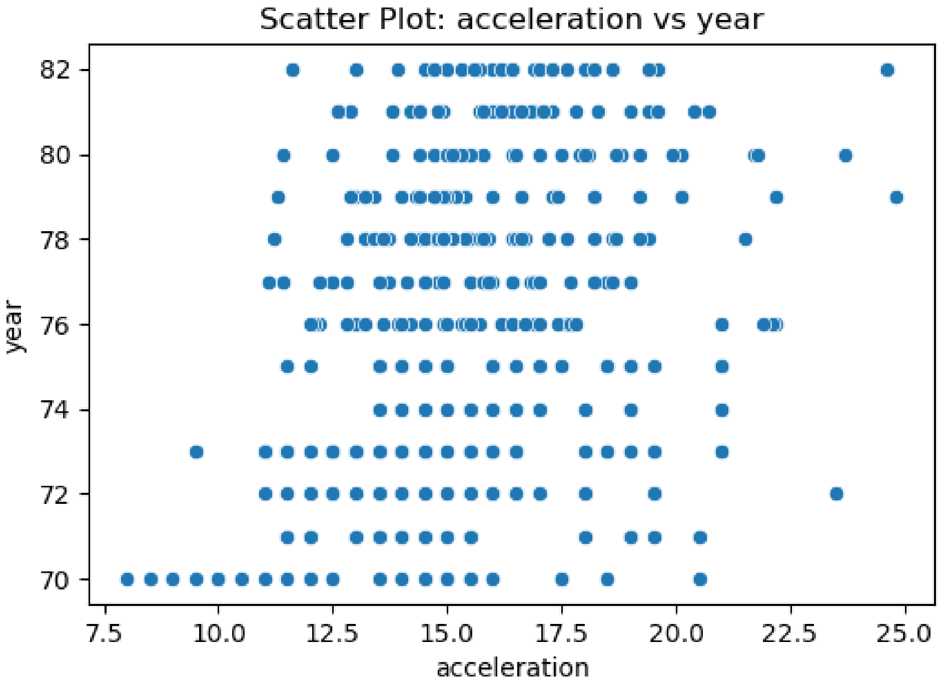


Figure 6: Acceleration versus year scatterplot.

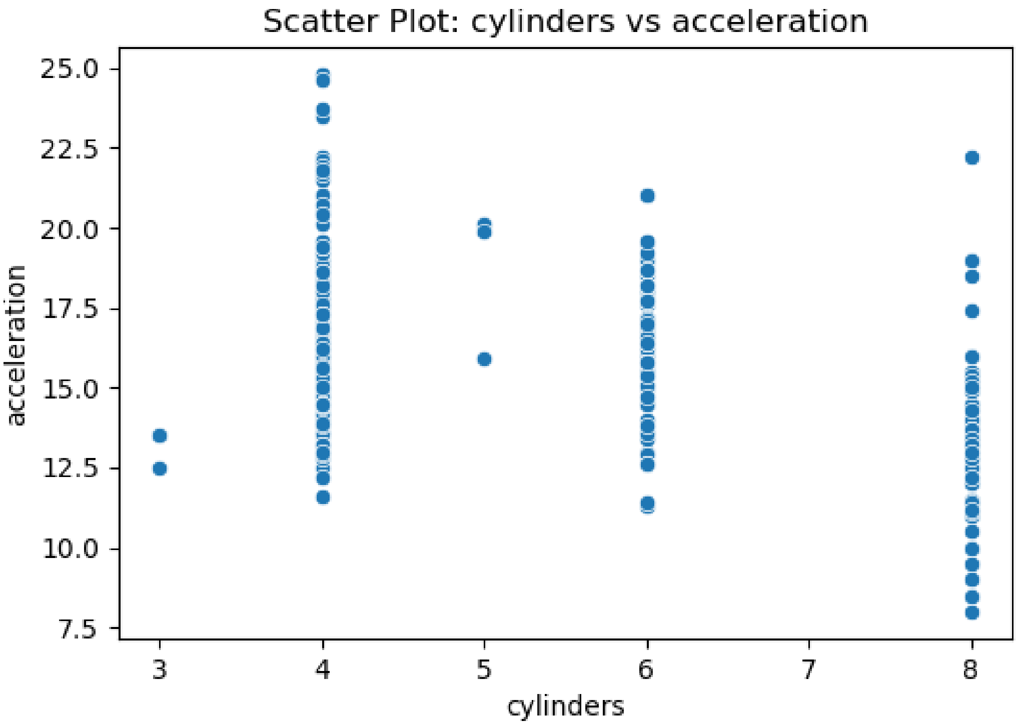


Figure 7: Cylinders versus acceleration scatterplot.

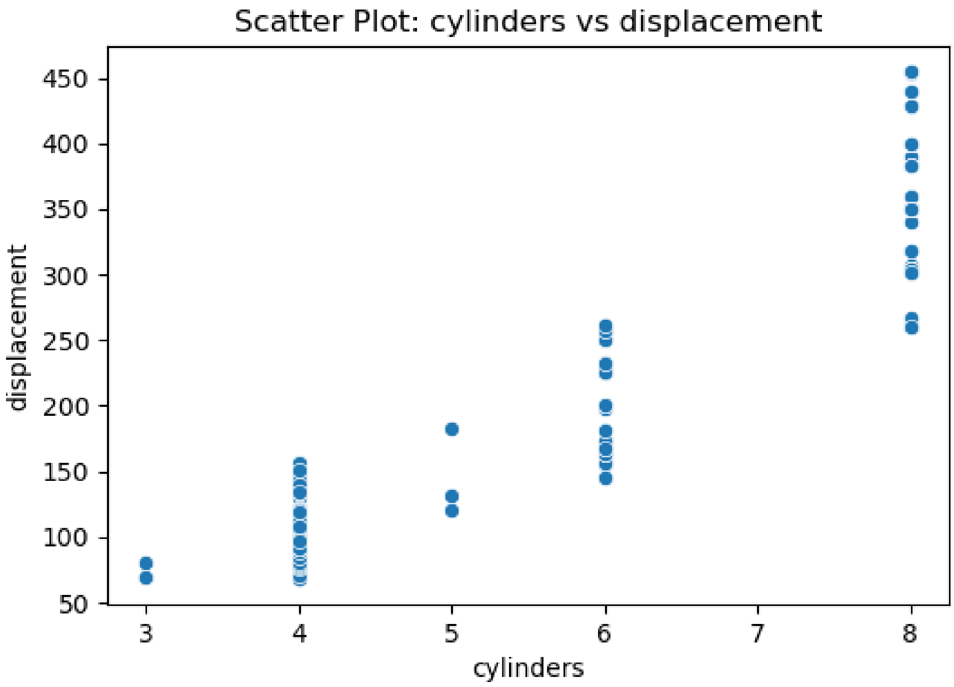


Figure 8: Cylinders versus displacement scatterplot.

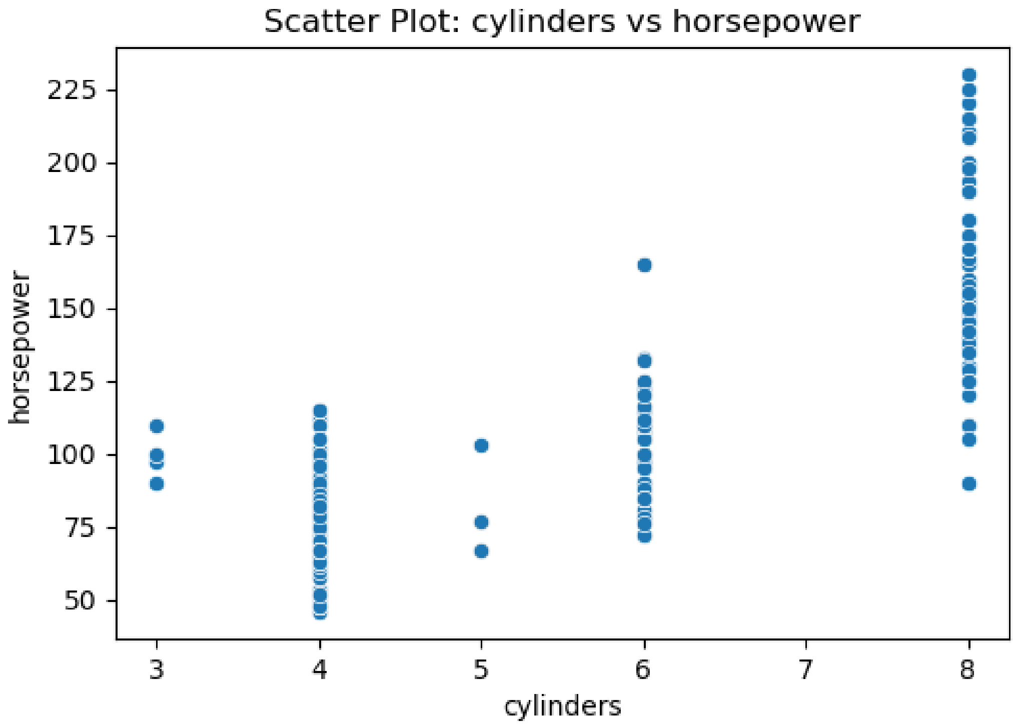


Figure 9: Cylinders versus horsepower scatterplot.

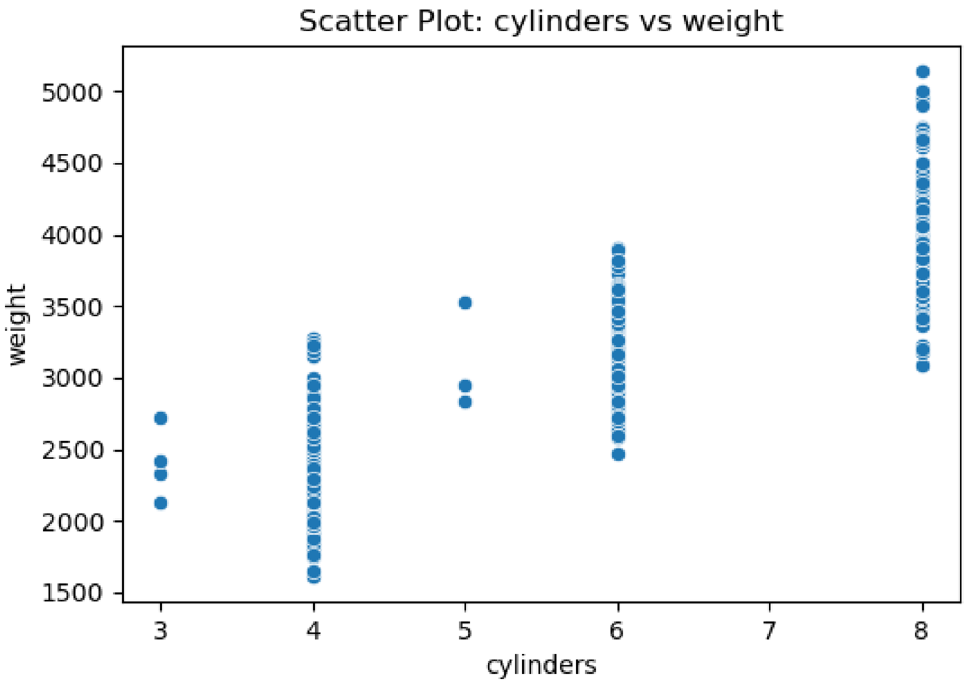


Figure 10: Cylinders versus weight scatterplot.

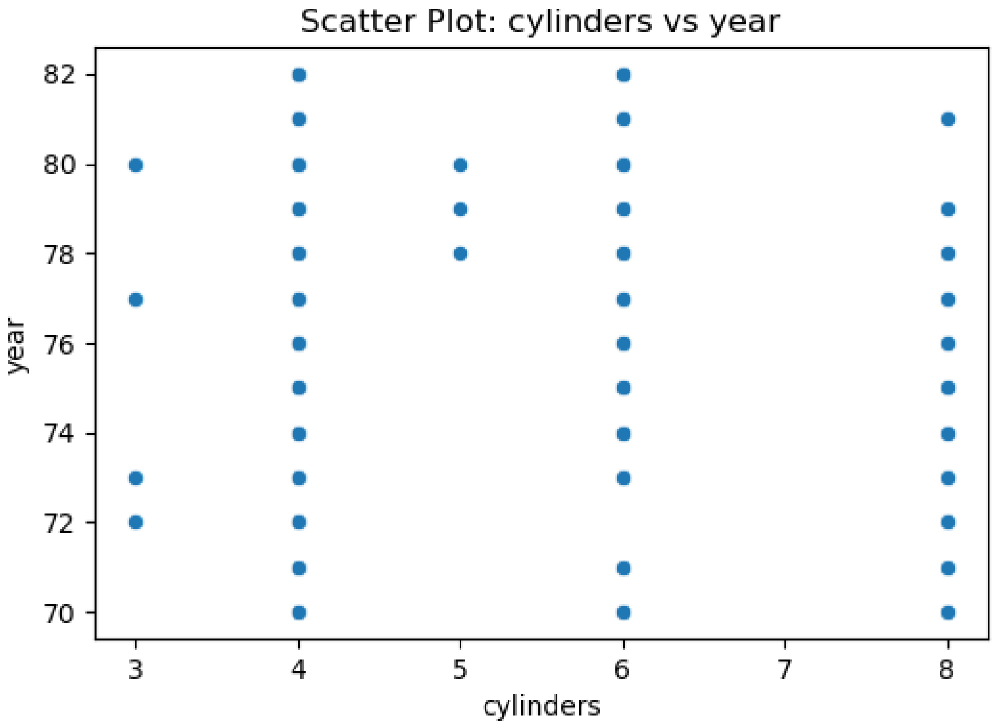


Figure 11: Cylinders versus year scatterplot.

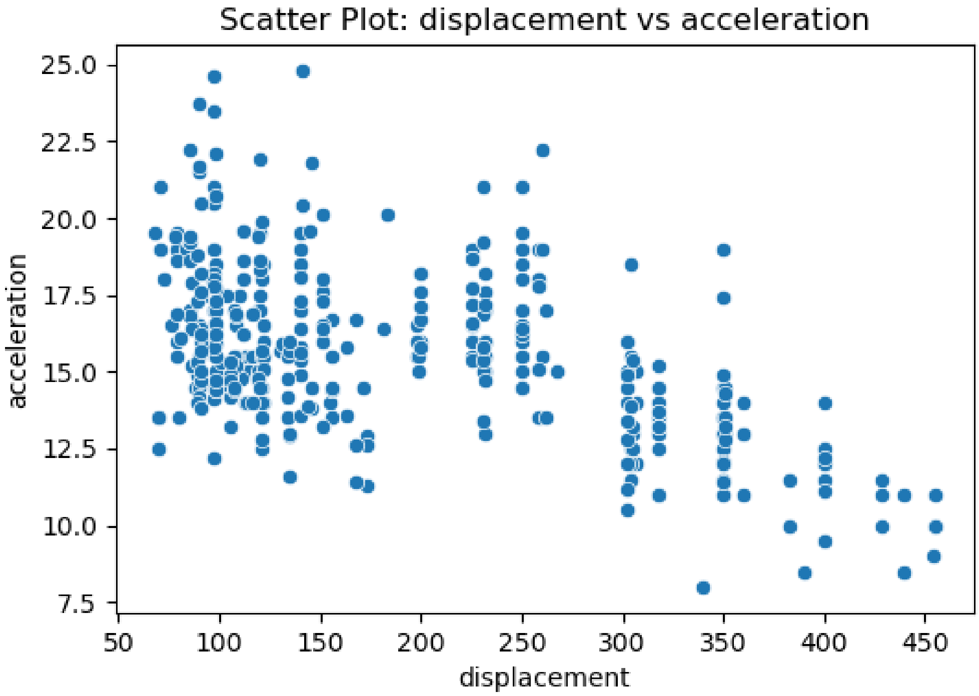


Figure 12: Displacement versus acceleration scatterplot.

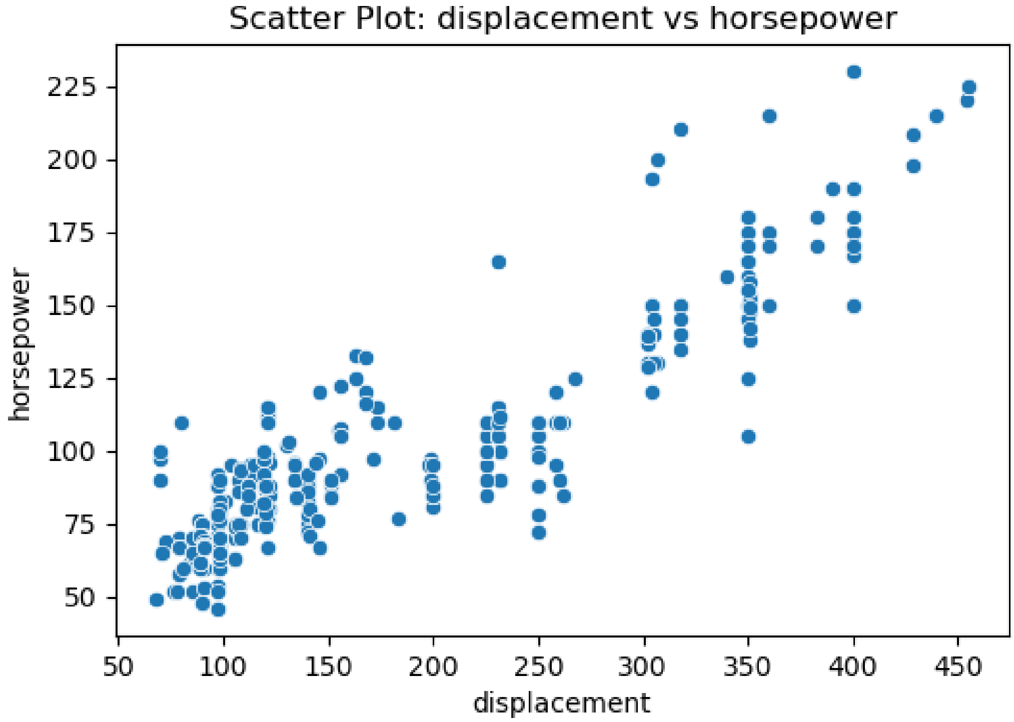


Figure 13: Displacement versus horsepower scatterplot.

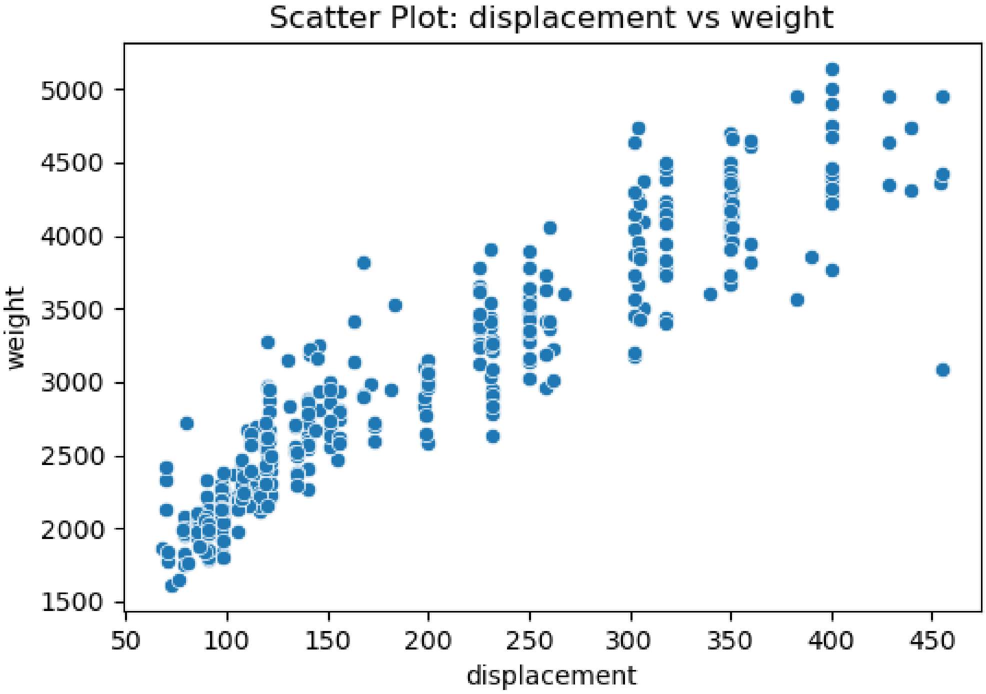


Figure 14: Displacement versus weight scatterplot.

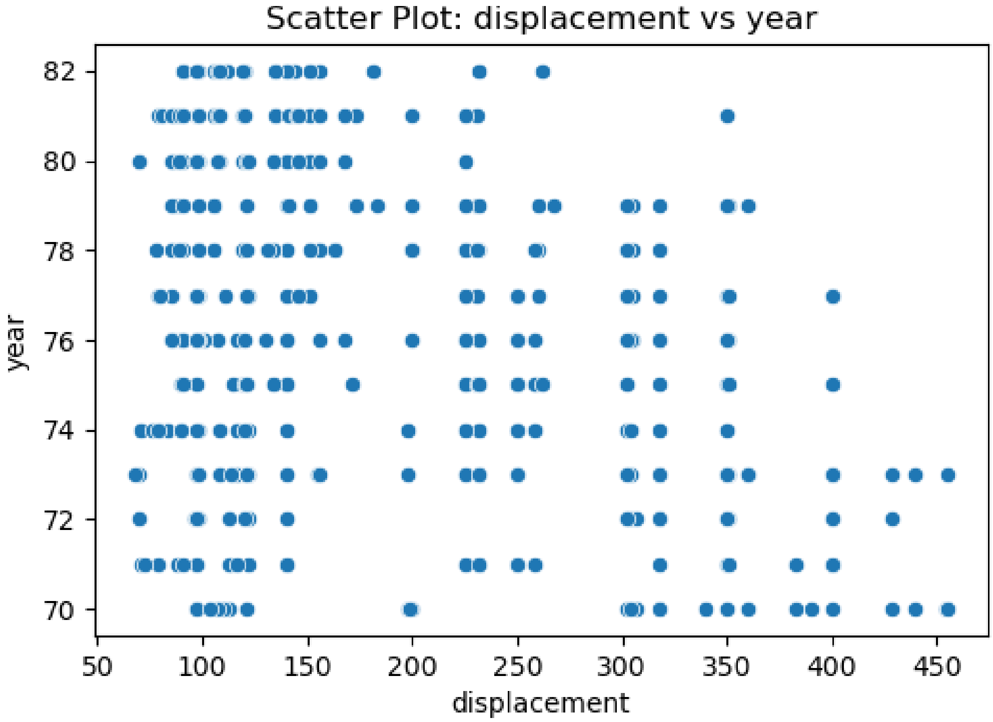


Figure 15: Displacement versus year scatterplot.

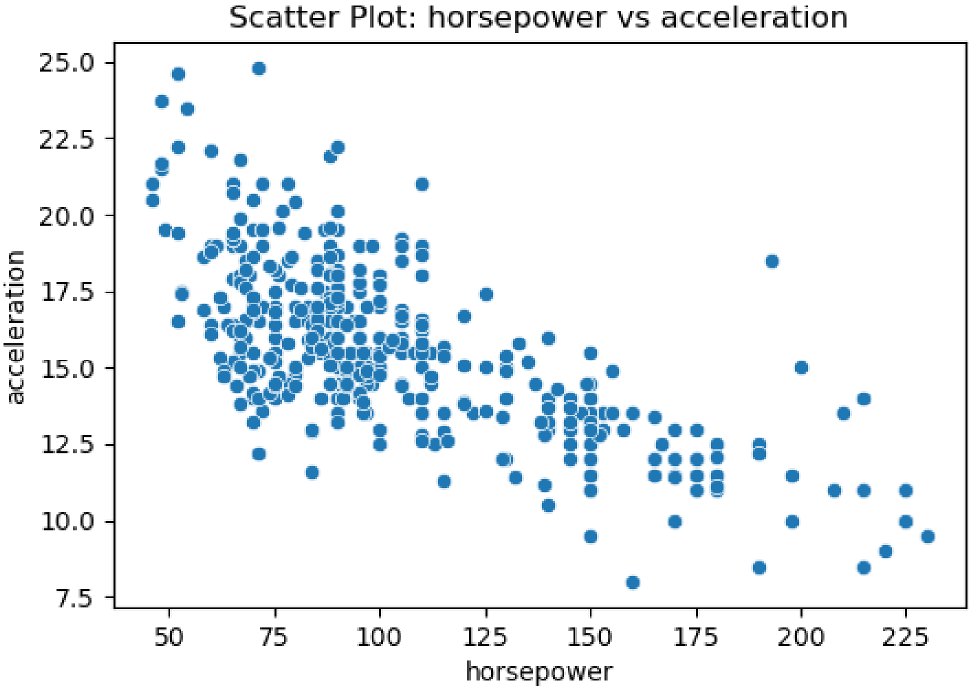


Figure 16: Horsepower versus acceleration scatterplot.

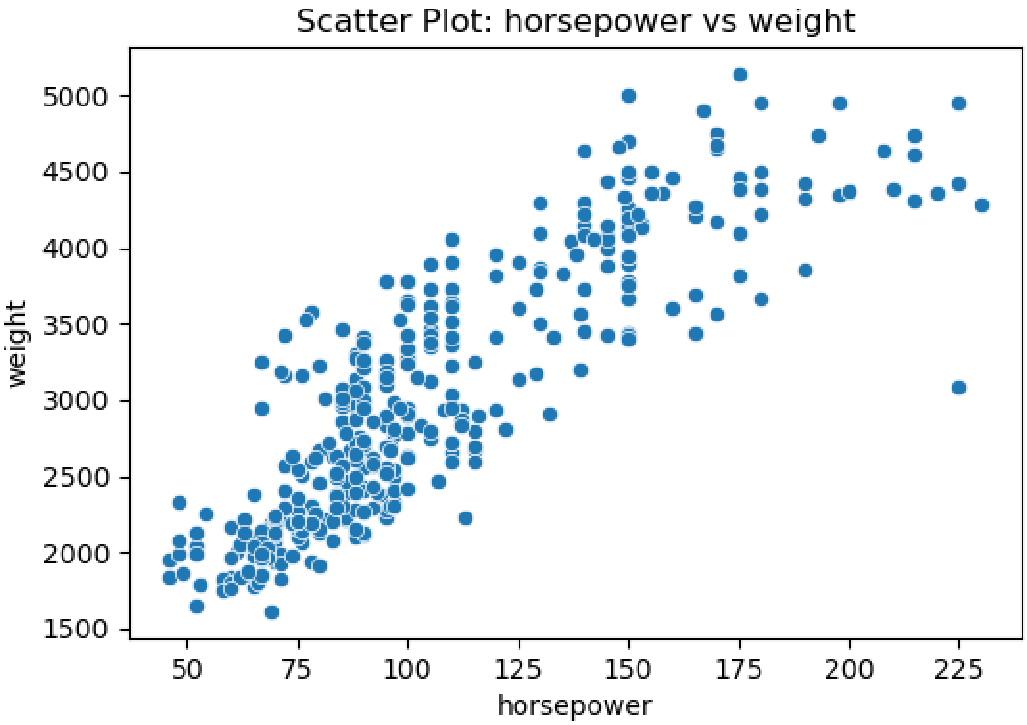


Figure 17: Horsepower versus weight scatterplot.

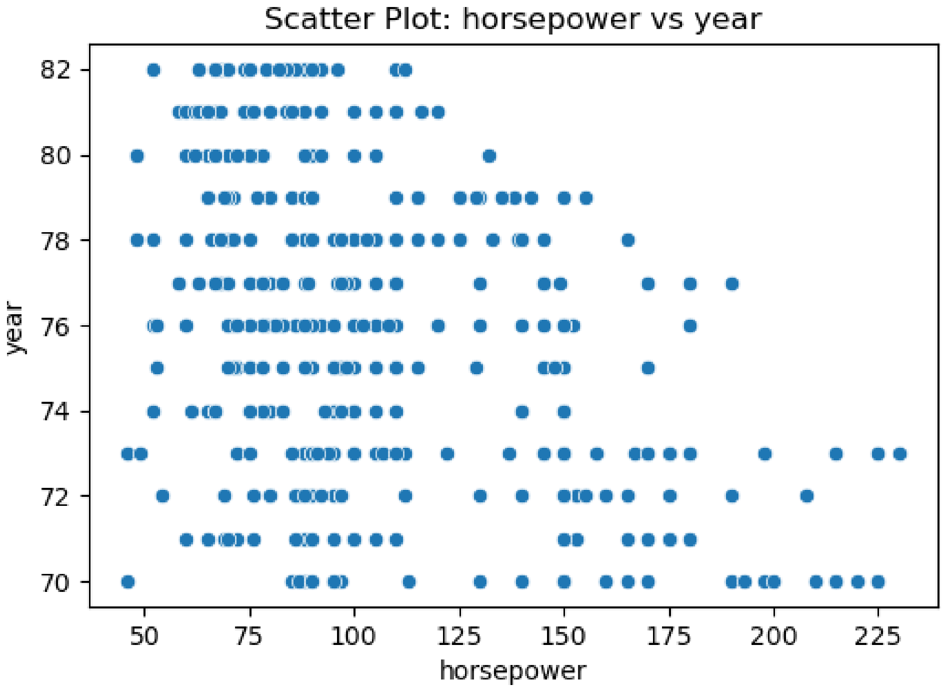


Figure 18: Horsepower versus year scatterplot.

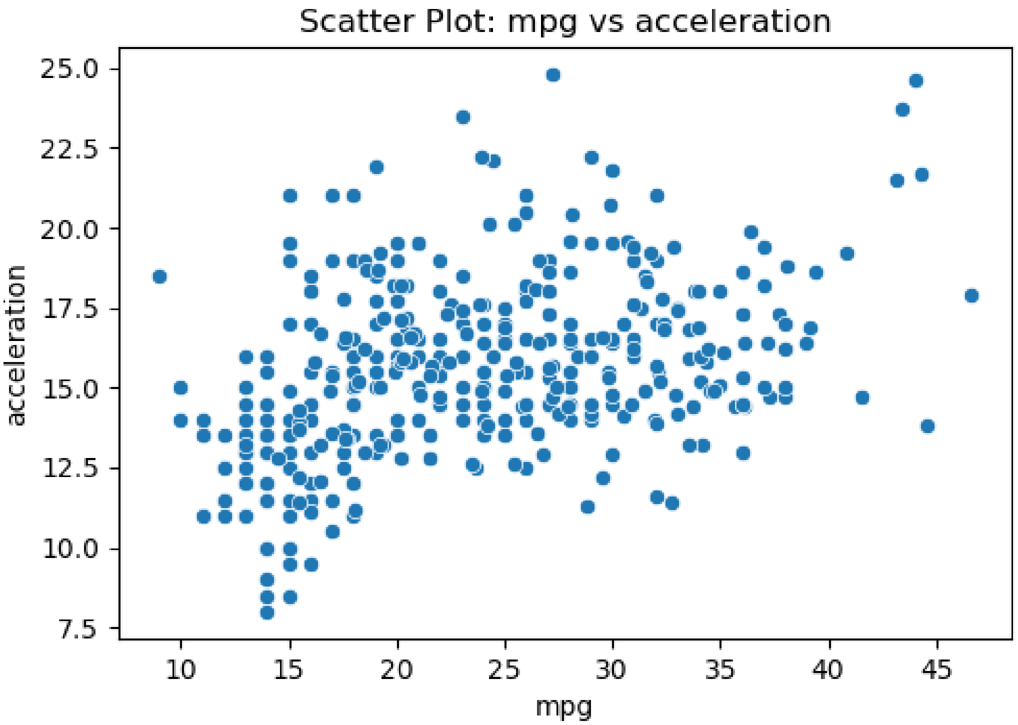


Figure 19: MPG versus acceleration scatterplot.

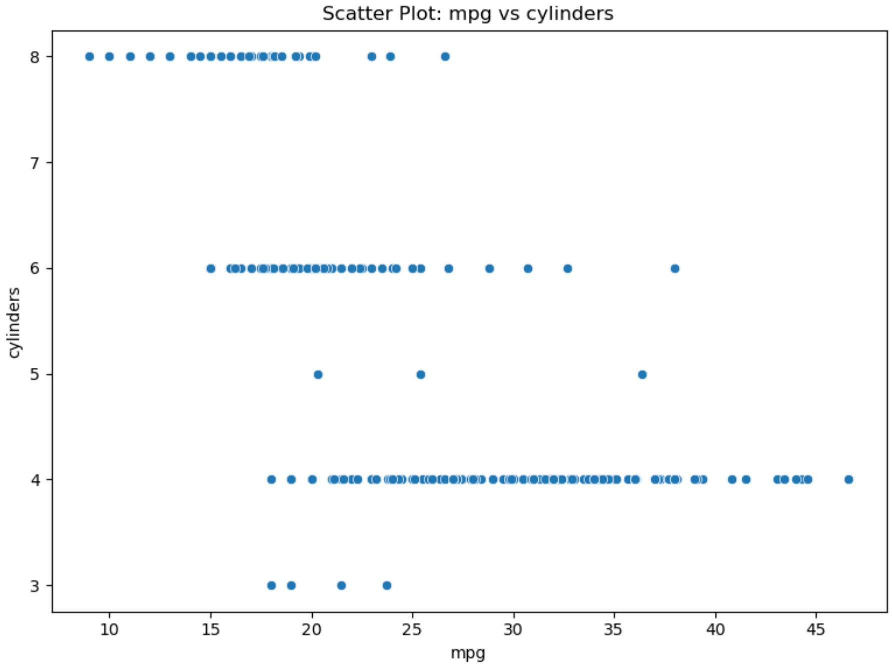


Figure 20: MPG versus cylinders scatterplot.

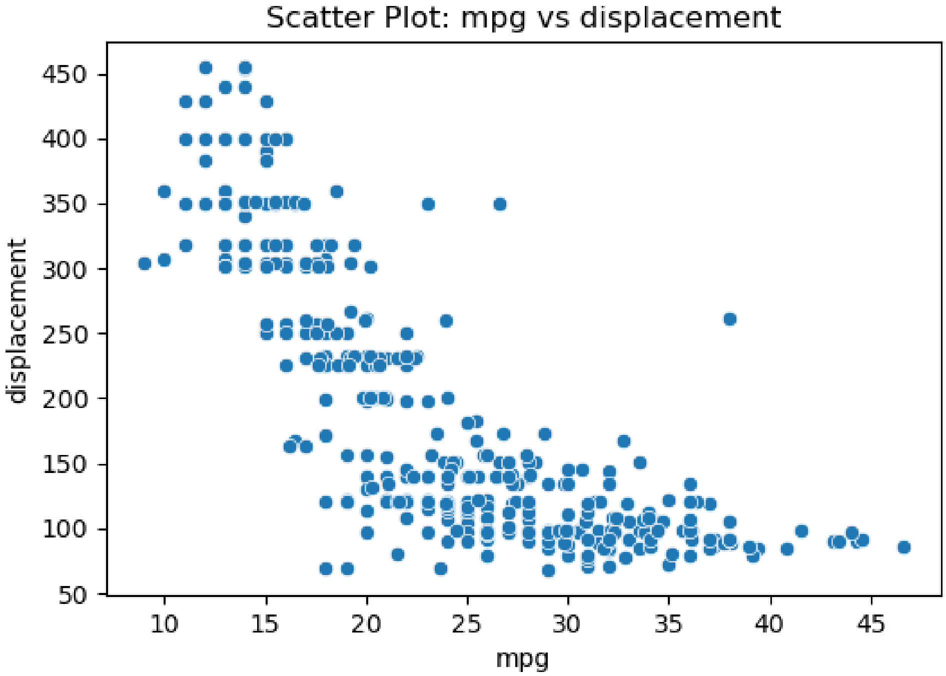


Figure 21: MPG versus displacement scatterplot.

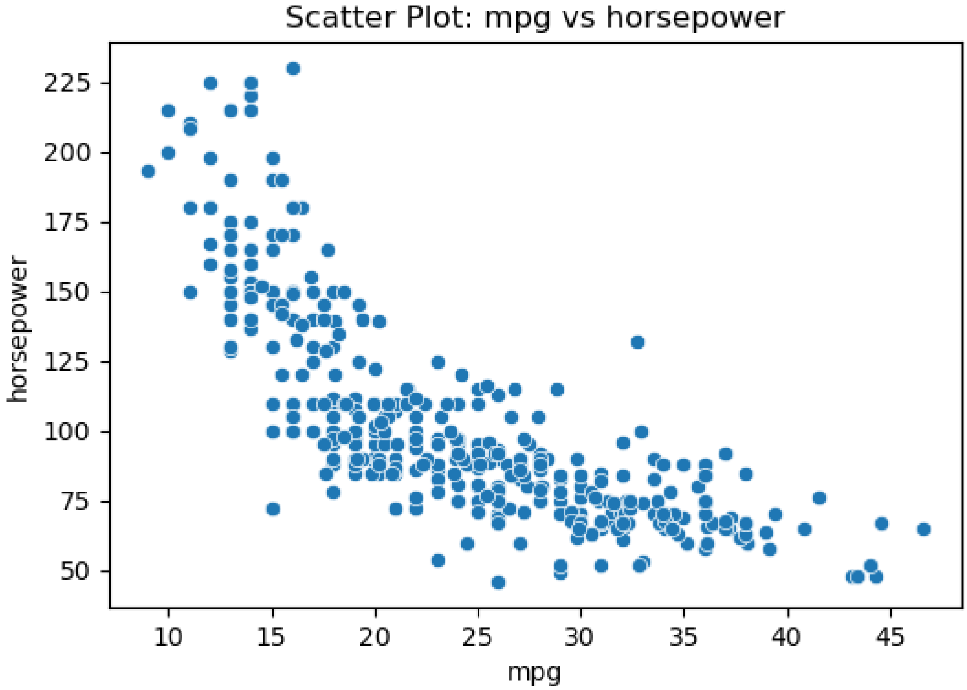


Figure 22: MPG versus horsepower scatterplot.

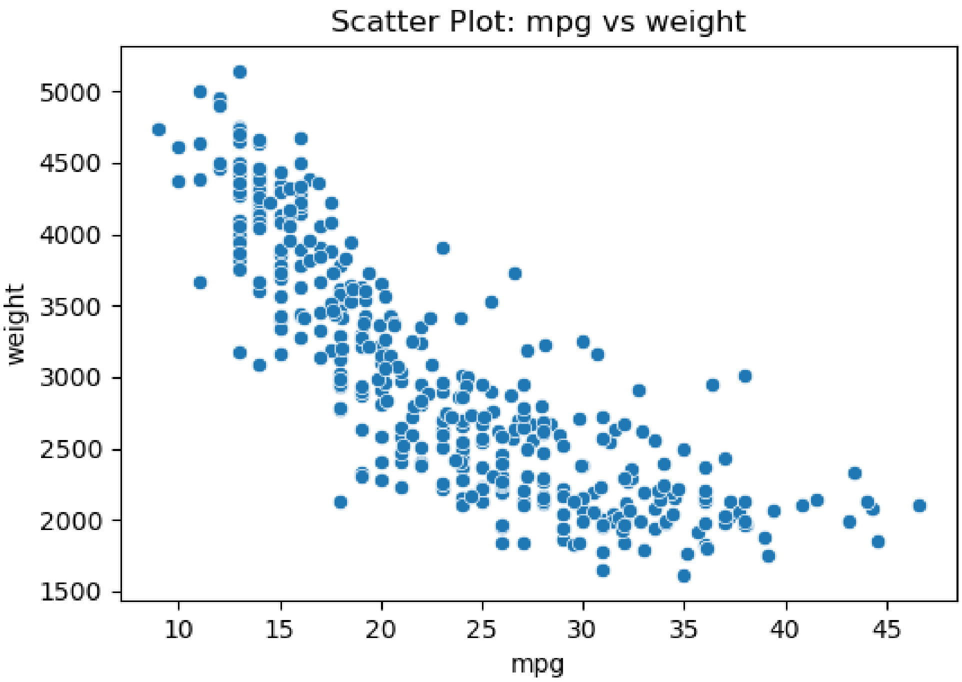


Figure 23: MPG versus weight scatterplot.

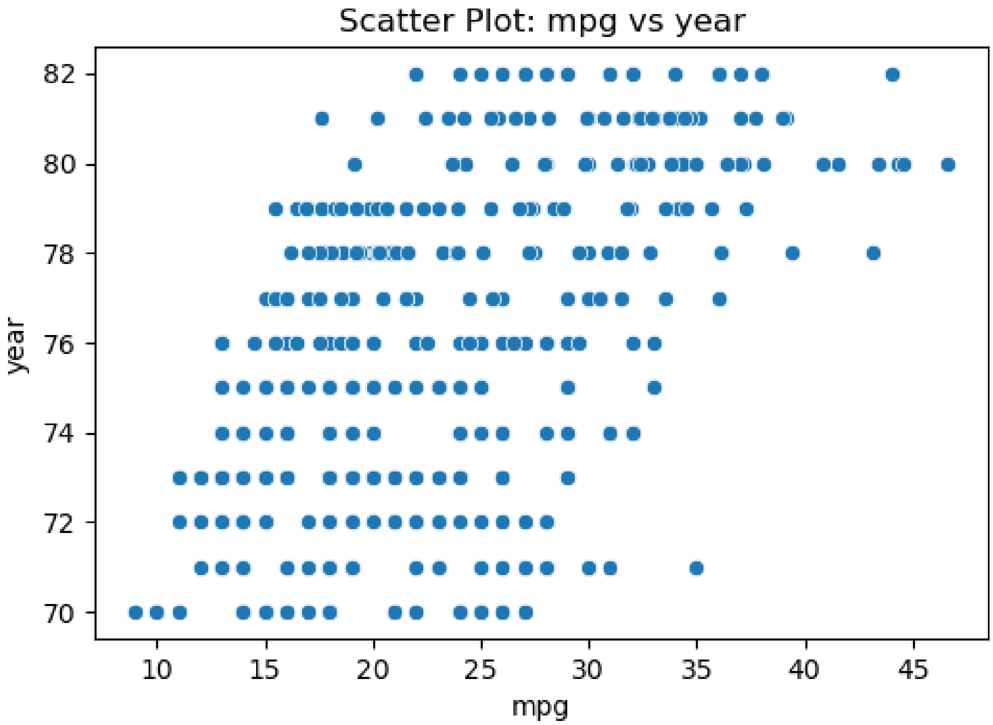


Figure 24: MPG versus year scatterplot.

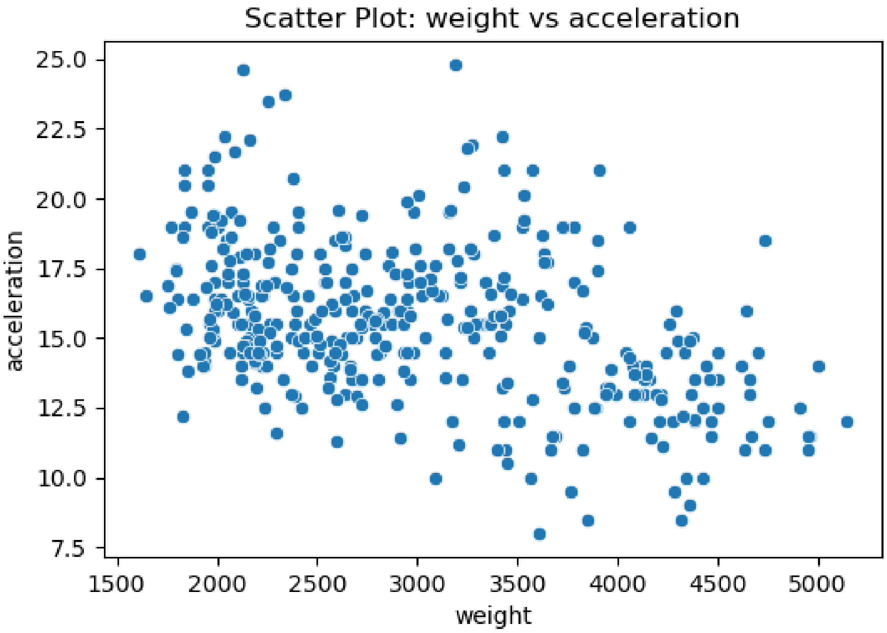


Figure 25: Weight versus acceleration scatterplot.

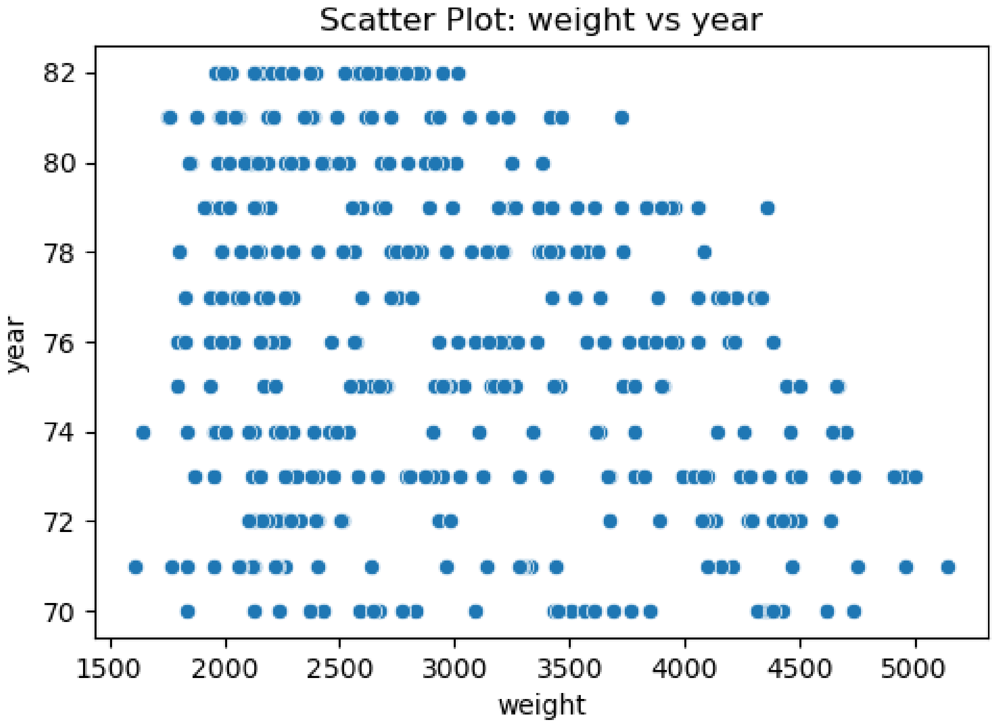


Figure 26: Weight versus year scatterplot.

b) Compute the matrix of correlations between the variables using the DataFrame.corr() method.

The figure provided below displays the heatmap of correlations between the variables:

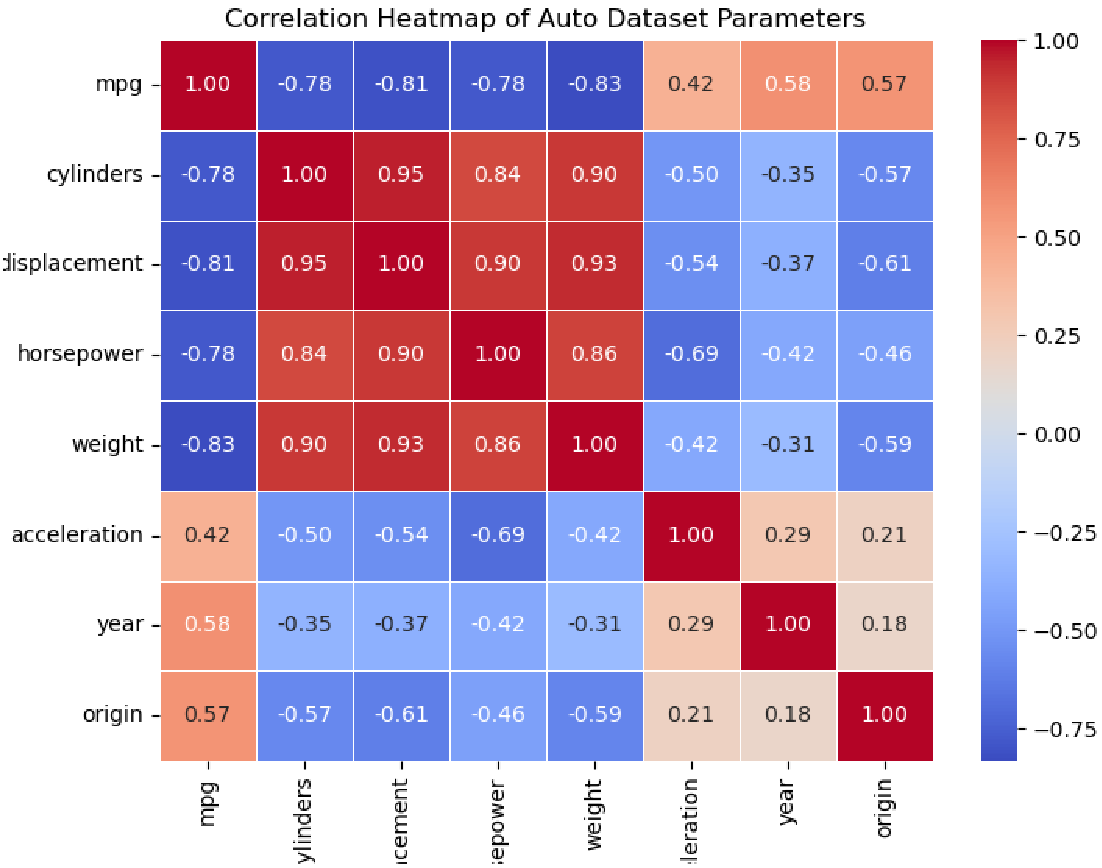
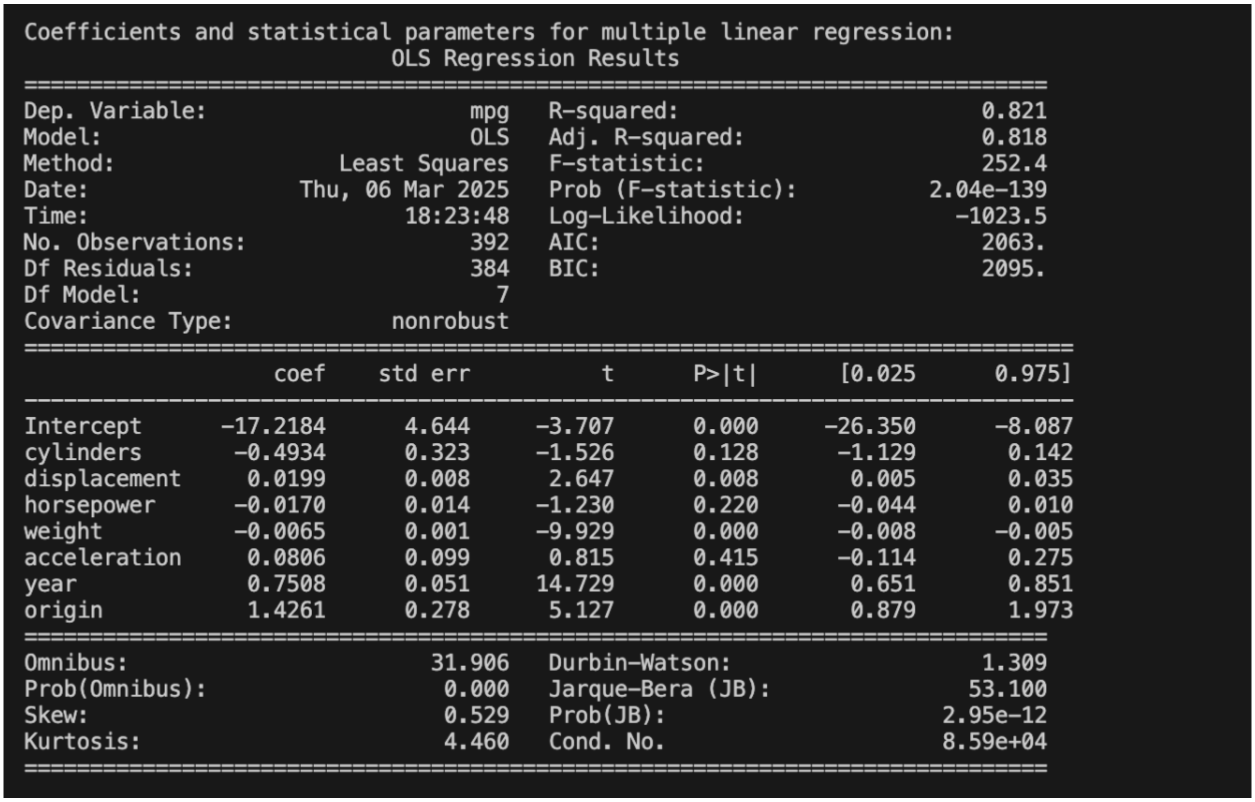


Figure 27: Correlation heatmap of Auto dataset parameters.

c) Use the sm.OLS() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summarize function to print the results. Comment on the output. For example:

   a. Is there a relationship between the predictors and the response? Use the anova_lm() function from statsmodels to answer this question.

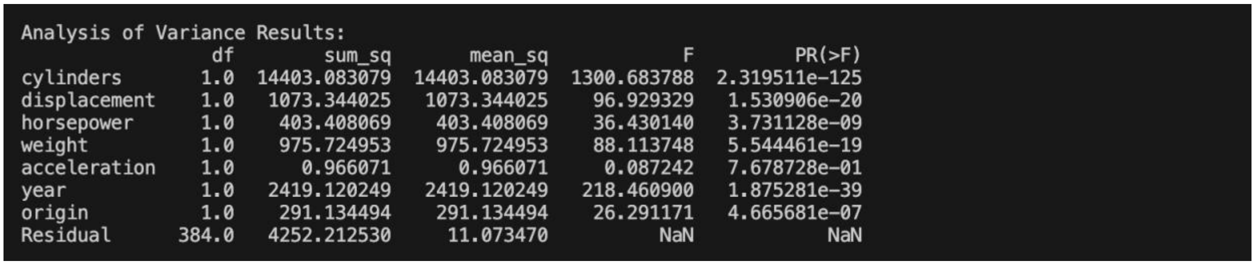
The figure given below displays the results of the OLS on the Auto dataset with multiple variables:



```
Coefficients and statistical parameters for multiple linear regression:
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.821
Model:                            OLS   Adj. R-squared:                  0.818
Method:                 Least Squares   F-statistic:                     252.4
Date:                Thu, 06 Mar 2025   Prob (F-statistic):          2.04e-139
Time:                        18:23:48   Log-Likelihood:                -1023.5
No. Observations:                 392   AIC:                             2063.
Df Residuals:                     384   BIC:                             2095.
Df Model:                           7
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      -17.2184      4.644     -3.707      0.000     -26.350      -8.087
cylinders       -0.4934      0.323     -1.526      0.128      -1.129       0.142
displacement     0.0199      0.008      2.647      0.008       0.005       0.035
horsepower      -0.0170      0.014     -1.230      0.220      -0.044       0.010
weight          -0.0065      0.001     -9.929      0.000      -0.008      -0.005
acceleration     0.0806      0.099      0.815      0.415      -0.114       0.275
year             0.7508      0.051     14.729      0.000       0.651       0.851
origin           1.4261      0.278      5.127      0.000       0.879       1.973
==============================================================================
Omnibus:                       31.906   Durbin-Watson:                   1.309
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               53.100
Skew:                           0.529   Prob(JB):                     2.95e-12
Kurtosis:                       4.460   Cond. No.                     8.59e+04
==============================================================================
```

Figure 28: Results for the multiple linear regression.

The figure given below displays the results for Analysis of Variance (ANOVA):



```
Analysis of Variance Results:
                 df        sum_sq       mean_sq            F         PR(>F)
cylinders       1.0  14403.083079  14403.083079  1300.683788  2.319511e-125
displacement    1.0   1073.344025   1073.344025    96.929329   1.530906e-20
horsepower      1.0    403.408069    403.408069    36.430140   3.731128e-09
weight          1.0    975.724953    975.724953    88.113748   5.544461e-19
acceleration    1.0      0.966071      0.966071     0.087242   7.678728e-01
year            1.0   2419.120249   2419.120249   218.460900   1.875281e-39
origin          1.0    291.134494    291.134494    26.291171   4.665681e-07
Residual      384.0   4252.212530     11.073470          NaN            NaN
```

Figure 29: ANOVA results.

b. Which predictors appear to have a statistically significant relationship to the response?

There are a few takeaways from the results provided above, which can be listed as:

- An R-squared result of 0.821 shows that the variability in mpg is explained by the predictors.
- There exist few significant predictors with $p - value < 0.05$, which are:
    - Displacement: Positive coefficient of displacement suggests that higher displacement leads to higher mpg.
    - Weight: Negative coefficient suggests that heavier cars have lower mpg.
    - Year: Positive coefficient suggests that newer cars tend to have higher mpg.

c. What does the coefficient for the year variable suggest?

The coefficient for the year variable suggests that newer cars tend to have higher mpg, whereas older cars tend to have lower mpg.

d) Produce some of diagnostic plots of the linear regression fit as described in the lab. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

The figure given below displays the residual plot for the multiple linear regression model:

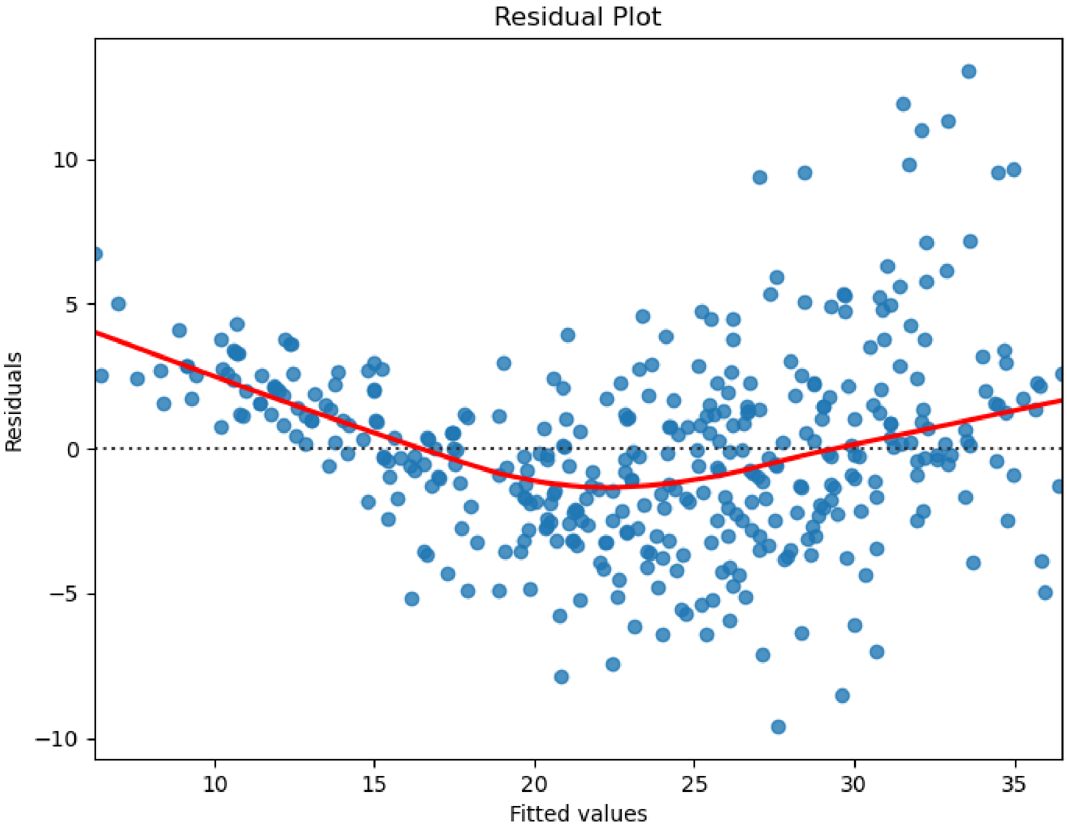
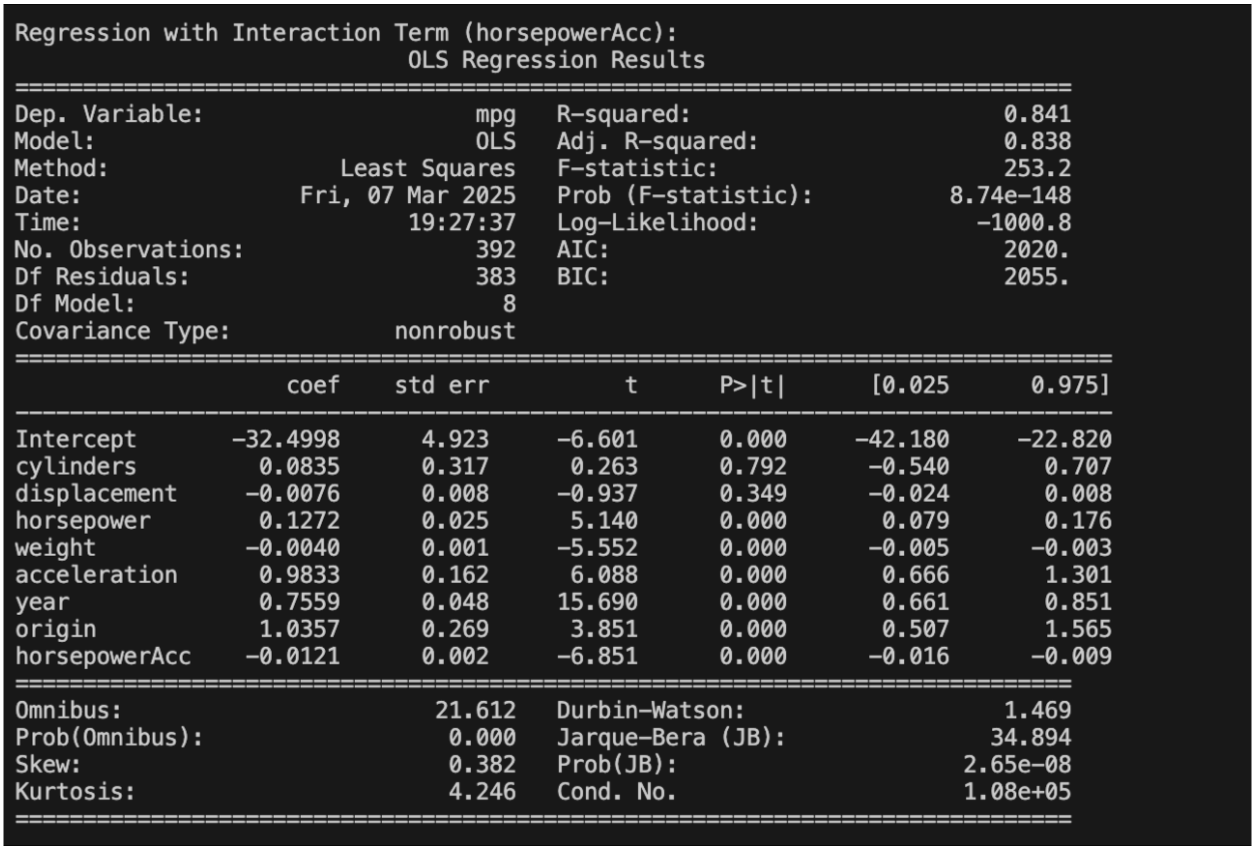


Figure 30: Residual plot for the multiple regression model.

Inspecting the plot provided above yields some points deviate significantly from the horizontal zero line, particularly at higher fitted values.

e) Fit some models with interactions as described in the lab. Do any interactions appear to be statistically significant?

The figure given below displays the results of the multiple linear regression with the interaction term, which is defined as $horsepower \times acceleration$:



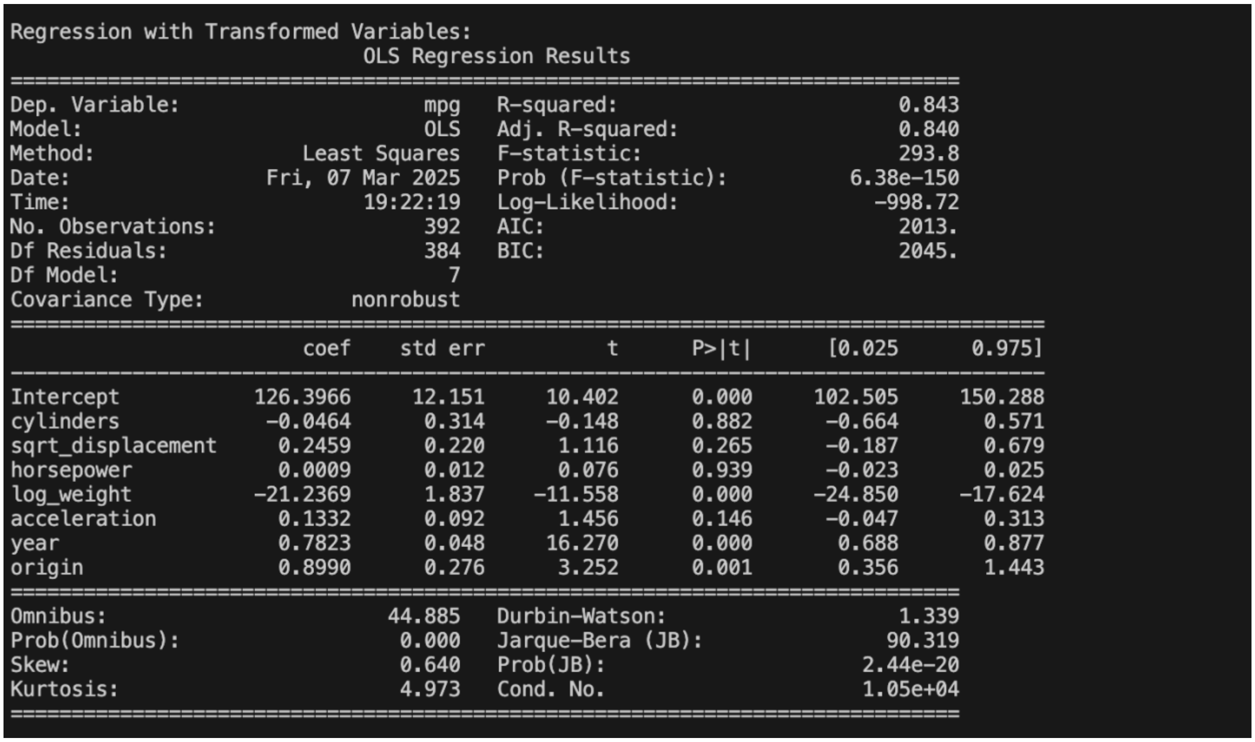
```
Regression with Interaction Term (horsepowerAcc):
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.841
Model:                            OLS   Adj. R-squared:                  0.838
Method:                 Least Squares   F-statistic:                     253.2
Date:                Fri, 07 Mar 2025   Prob (F-statistic):          8.74e-148
Time:                        19:27:37   Log-Likelihood:                -1000.8
No. Observations:                 392   AIC:                             2020.
Df Residuals:                     383   BIC:                             2055.
Df Model:                           8                                         
Covariance Type:            nonrobust                                         
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept       -32.4998      4.923     -6.601      0.000     -42.180     -22.820
cylinders         0.0835      0.317      0.263      0.792      -0.540       0.707
displacement     -0.0076      0.008     -0.937      0.349      -0.024       0.008
horsepower        0.1272      0.025      5.140      0.000       0.079       0.176
weight           -0.0040      0.001     -5.552      0.000      -0.005      -0.003
acceleration      0.9833      0.162      6.088      0.000       0.666       1.301
year              0.7559      0.048     15.690      0.000       0.661       0.851
origin            1.0357      0.269      3.851      0.000       0.507       1.565
horsepowerAcc    -0.0121      0.002     -6.851      0.000      -0.016      -0.009
==============================================================================
Omnibus:                       21.612   Durbin-Watson:                   1.469
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.894
Skew:                           0.382   Prob(JB):                     2.65e-08
Kurtosis:                       4.246   Cond. No.                     1.08e+05
==============================================================================
```

*Figure 31: Regression results with interaction term defined as horsepower*acceleration.*

It is evident that our interaction term is statistically significant, which can be proved by the statistical measurements given in the figure.

f) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2.

The figure given below displays the results of the multiple linear regression with the provided transformations applied:



```
Regression with Transformed Variables:
                             OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.843
Model:                            OLS   Adj. R-squared:                  0.840
Method:                 Least Squares   F-statistic:                     293.8
Date:                Fri, 07 Mar 2025   Prob (F-statistic):          6.38e-150
Time:                        19:22:19   Log-Likelihood:                -998.72
No. Observations:                 392   AIC:                             2013.
Df Residuals:                     384   BIC:                             2045.
Df Model:                           7
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept           126.3966     12.151     10.402      0.000     102.505     150.288
cylinders            -0.0464      0.314     -0.148      0.882      -0.664       0.571
sqrt_displacement     0.2459      0.220      1.116      0.265      -0.187       0.679
horsepower            0.0009      0.012      0.076      0.939      -0.023       0.025
log_weight          -21.2369      1.837    -11.558      0.000     -24.850     -17.624
acceleration          0.1332      0.092      1.456      0.146      -0.047       0.313
year                  0.7823      0.048     16.270      0.000       0.688       0.877
origin                0.8990      0.276      3.252      0.001       0.356       1.443
==============================================================================
Omnibus:                       44.885   Durbin-Watson:                   1.339
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               90.319
Skew:                           0.640   Prob(JB):                     2.44e-20
Kurtosis:                       4.973   Cond. No.                     1.05e+04
==============================================================================
```

Figure 32: Regression results with transformed variables.

Conclusion

In this project, we explored linear regression using the Auto dataset and analyzed its results. We first applied simple linear regression to examine the relationship between miles per gallon (mpg) and horsepower, showing a strong negative correlation. Then, we extended our analysis to multiple linear regression, adding additional predictors to improve our model. The results showed that weight, displacement, and year had significant effects on mpg. We also investigated diagnostic plots, interactions, and transformations to refine our understanding of the model's performance.

References

[1] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: With Applications in Python*. Cham, Cham: Springer International Publishing Springer, 2023.