DATA SCIENCE AND ANALYSIS (DSA)

Lab

Experiment No. 7

Clustering Model

**AIM:** To study and analyze Clustering Model **SOFTWARE USED**: Python, Jupyter Notebook **THEORY:**

# Simple Linear Regression:

Simple linear regression models the relationship between two variables: an independent variable (x) and a dependent variable (y). The equation of a straight line is used: y = mx + b, where:

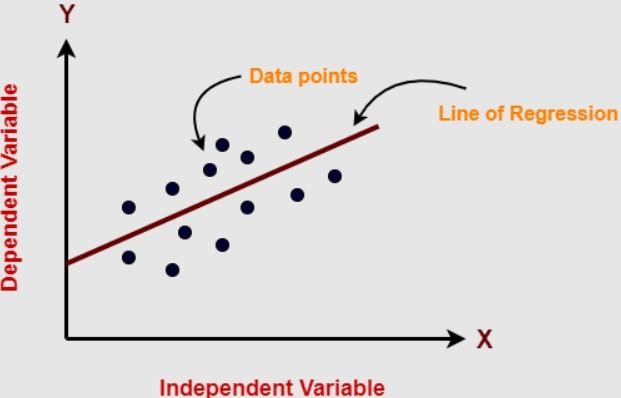
* y is the dependent variable (the predicted value),
* x is the independent variable (used for predictions),
* m is the slope of the line (indicating the direction and steepness of the relationship),
* b is the y-intercept (the value of y when x = 0).

Simple linear regression also evaluates the relationship's strength and direction through the correlation coefficient (r), which ranges from -1 to 1:

r = 1 signifies a perfect positive linear relationship,

r = -1 indicates a perfect negative linear relationship, r = 0 suggests no linear relationship.

In summary, simple linear regression provides a framework for understanding and modeling the association between two variables, aiding in prediction and inference based on observed data.



# Multi Linear Regression:

Multi linear regression models the relationship between multiple independent variables (x1, x2, etc.) and a dependent variable (y) using the equation:

# y = b0 + b1\*x1 + b2\*x2 + ... + bn\*xn

Where:

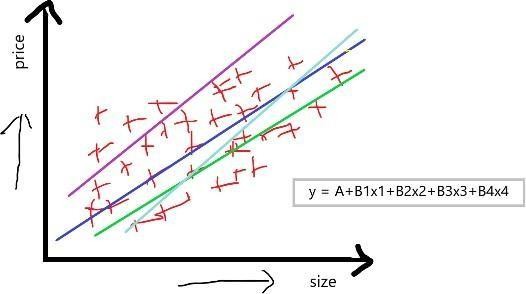
* y is the dependent variable.
* x1, x2, etc., are independent variables.
* b0 is the y-intercept.
* b1, b2, etc., are coefficients representing the impact of each independent variable.

Similar to simple linear regression, multiple linear regression allows for the assessment of the strength and direction of the relationship between the independent variables and the dependent variable through the correlation coefficients.

These coefficients range from -1 to 1, where:

* r = 1: perfect positive linear relationship
* r = -1: perfect negative linear relationship
* r = 0: no linear relationship.

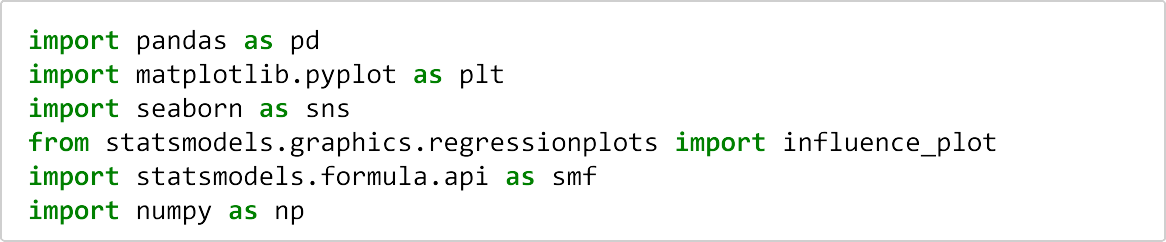
Multi linear regression models the association between multiple independent variables and a dependent variable, aiding in prediction and inference based on observed data.

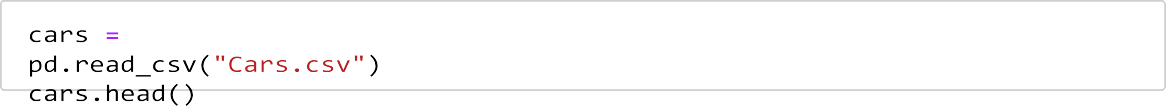


# Application of Multi Linear Regression:

1. **Predictive Modeling**: Building models to predict numerical outcomes based on input features, such as predicting sales revenue based on advertising spending and other factors.
2. **Forecasting**: Using historical data to forecast future trends or values, such as predicting stock prices, demand for products, or energy consumption.
3. **Risk Assessment**: Assessing and quantifying risks, such as predicting the likelihood of loan defaults or identifying factors contributing to patient readmissions in healthcare.
4. **Optimization**: Optimizing processes or systems by identifying key variables that influence outcomes and determining how to manipulate them to achieve desired results, such as optimizing manufacturing processes or supply chain logistics.

In [7]:

In [8]:

Out[8]:

In [9]:

<class 'pandas.core.frame.DataFrame'> RangeIndex: 81 entries, 0 to 80

Data columns (total 5 columns):

# Column Non-Null Count Dtype

* 1. HP 81 non-null int64
  2. MPG 81 non-null float64
  3. VOL 81 non-null int64
  4. SP 81 non-null float64
  5. WT 81 non-null float64 dtypes: float64(3), int64(2)

memory usage: 3.3 KB

In [10]:

Out[10]: HP 0

MPG 0

VOL 0

SP 0

WT 0

dtype: int64

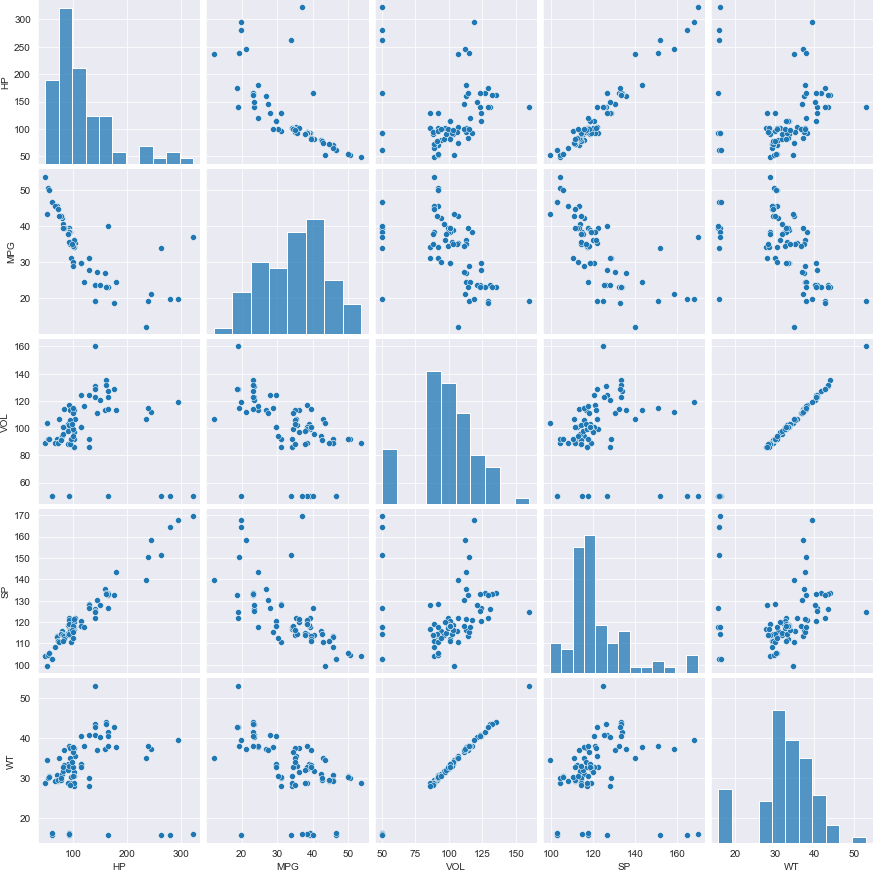
In [11]:

Out[11]:

In [12]:

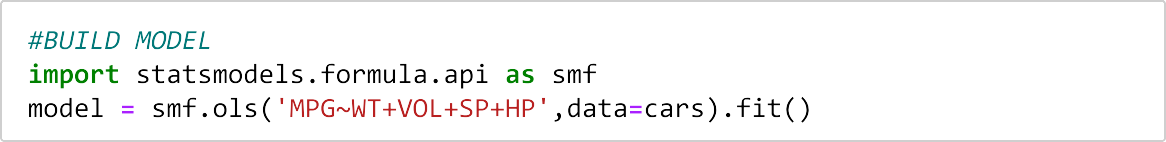
Out[12]: <seaborn.axisgrid.PairGrid at 0x2a0d63bee80>



**HP MPG VOL SP WT**

4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **0** | 53.700681 | 89 | 104.18535 | 28.762059 | |
| 49 |  |  | 3 |  | |
| **1** | 50.013401 | 92 | 105.46126 | 30.466833 | |
| 55 |  |  | 4 |  | |
| **2** | 50.013401 | 92 | 105.46126 | 30.193597 | |
| 55 |  |  | 4 |  | |
| **3** | 45.696322 | 92 | 113.46126 | 30.632114 | |
| 70 |  |  | 4 |  | |
| **4**  53 | 50.504232 | 92 | 104.46126 | 29.889149 | |
|  | **HP** | **MPG** | **VOL** | **SP** | **WT** |
| **HP** | 1.000000 | -0.725038 | 0.077459 | 0.973848 | 0.076513 |
| **MPG** | -0.725038 | 1.000000 | -0.529057 | -0.687125 | -0.526759 |
| **VOL** | 0.077459 | -0.529057 | 1.000000 | 0.102170 | 0.999203 |
| **SP** | 0.973848 | -0.687125 | 0.102170 | 1.000000 | 0.102439 |
| **WT** | 0.076513 | -0.526759 | 0.999203 | 0.102439 | 1.000000 |

In [13]:

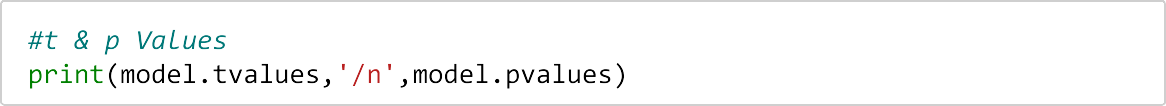
 

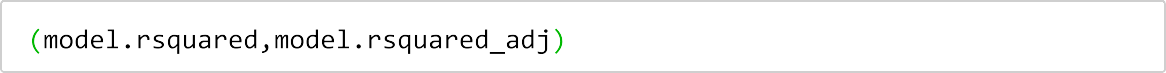
|  |  |
| --- | --- |
| In [14]: |  |
| Out[14]: Intercept | 30.677336 |
| WT | 0.400574 |
| VOL | -0.336051 |
| SP | 0.395627 |
| HP | -0.205444 |

dtype: float64

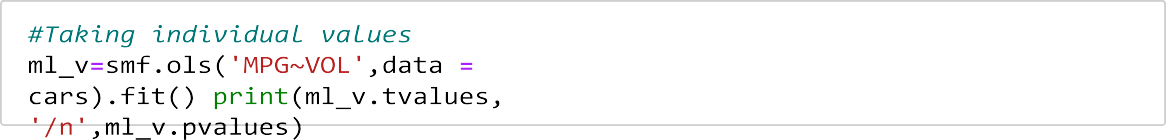
 

In [15]:

0.042936

In [16]:

Out[16]: (0.7705372737359844, 0.7584602881431415)

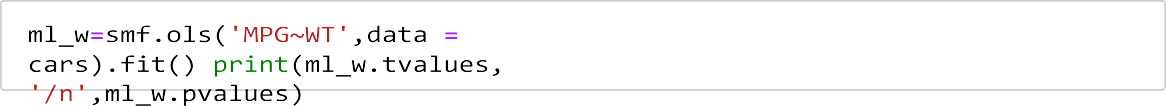
In [17]:

Intercept 14.106056

VOL -5.541400

dtype: float64 /n Intercept 2.753815e-23 VOL 3.822819e-07

dtype: float64

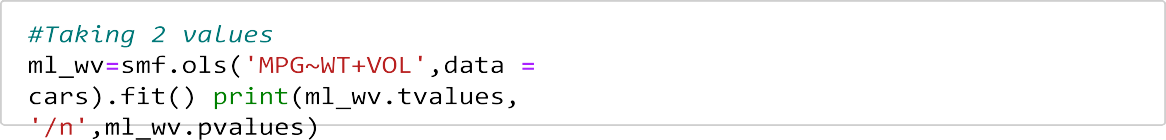
In [18]:

Intercept 14.248923

WT -5.508067

dtype: float64 /n Intercept 1.550788e-23 WT 4.383467e-07

dtype: float64

In [19]:

Intercept 12.545736

WT 0.489876

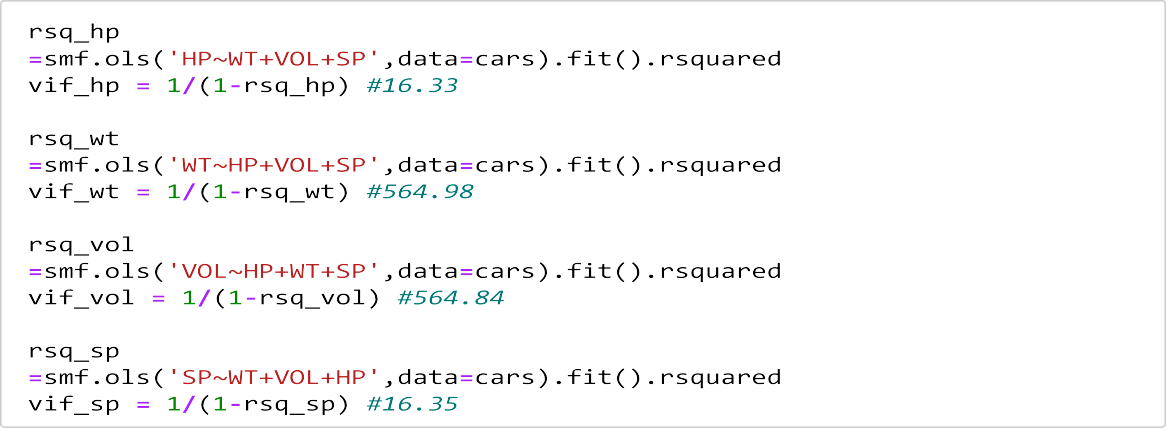
VOL -0.709604

dtype: float64 /n Intercept 2.141975e-20 WT 6.255966e-01

VOL 4.800657e-01

dtype: float64

In [20]:





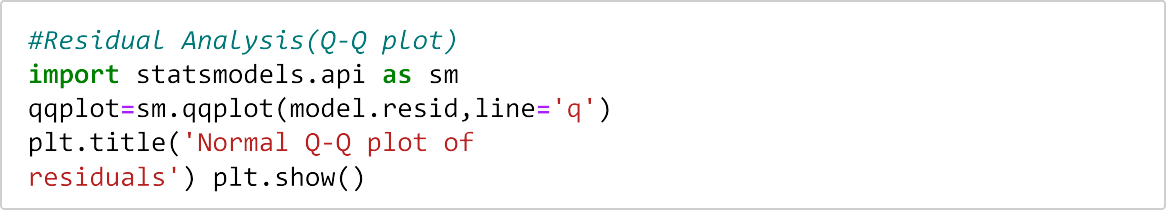
Out[20]: **Variables VIF**

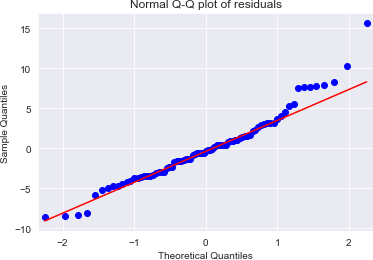
**0** HP 19.926589

**1** WT 639.533818

**2** VOL 638.806084

**3** SP 20.007639

In [21]:





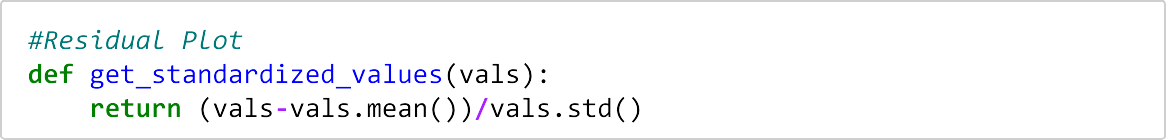
In [22]:

|  |  |  |
| --- | --- | --- |
|  | WT | 0.813649 |
| VOL | 0.556294 |
| Intercept 2.058841 | SP | 0.014579 |
| WT 0.236541 | HP | 0.000001 |
| VOL -0.590970 | dtype: | float64 |
| SP 2.499880 |  |  |
| HP -5.238735 |  |  |
| dtype: float64 /n Intercept |  |  |

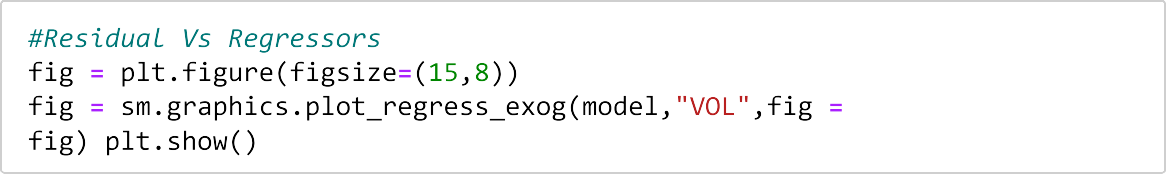
Out[22]: [array([ 0, 76], dtype=int64)]

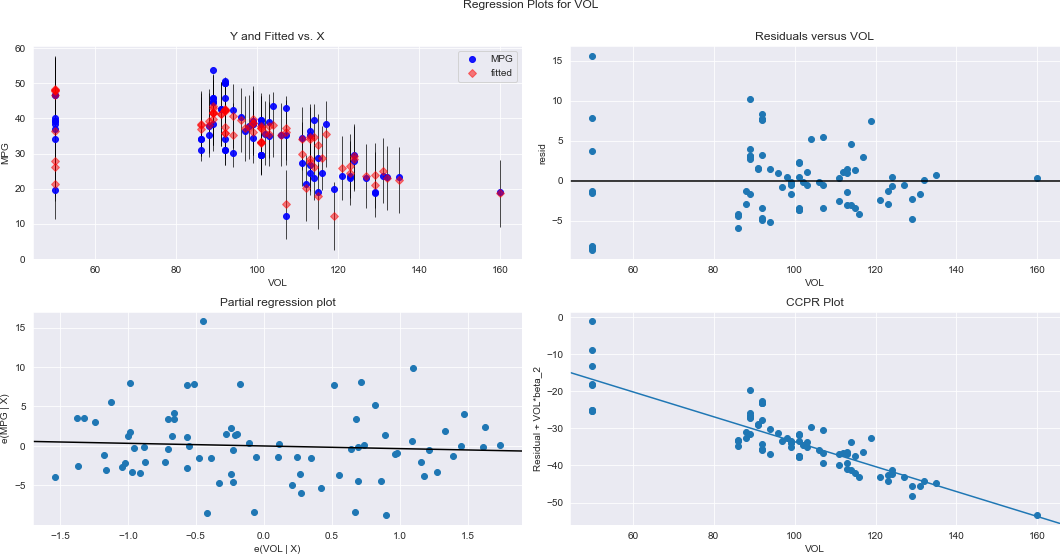
 

In [23]:

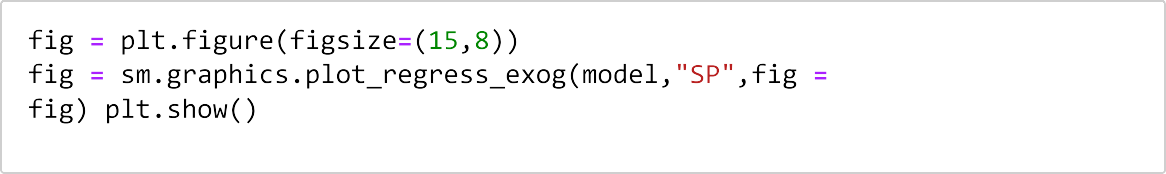
In [24]:

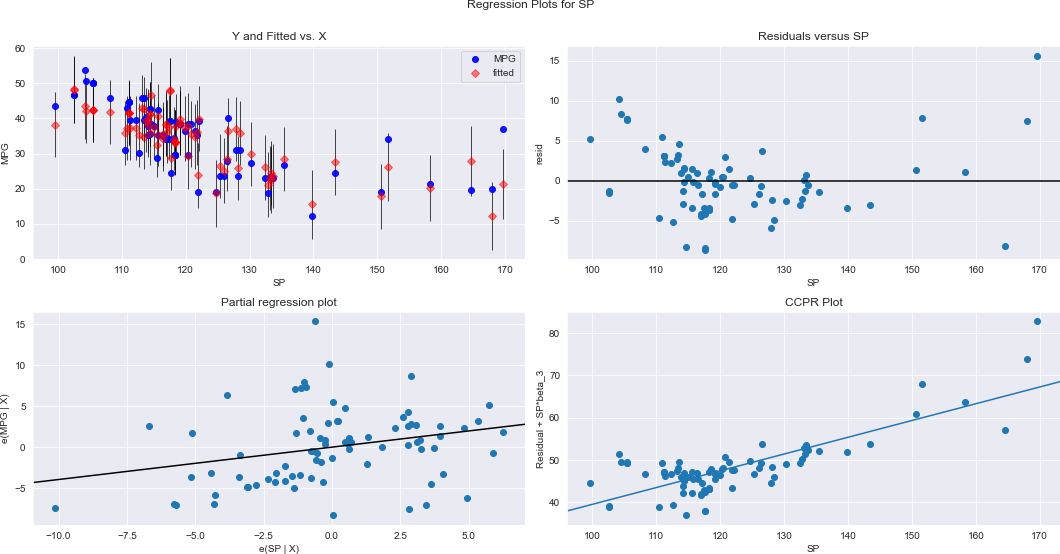


In [25]:

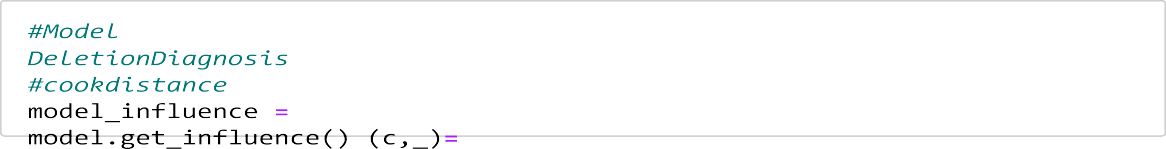


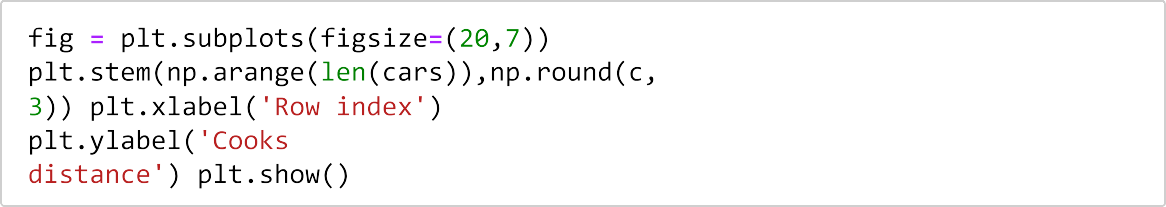
 

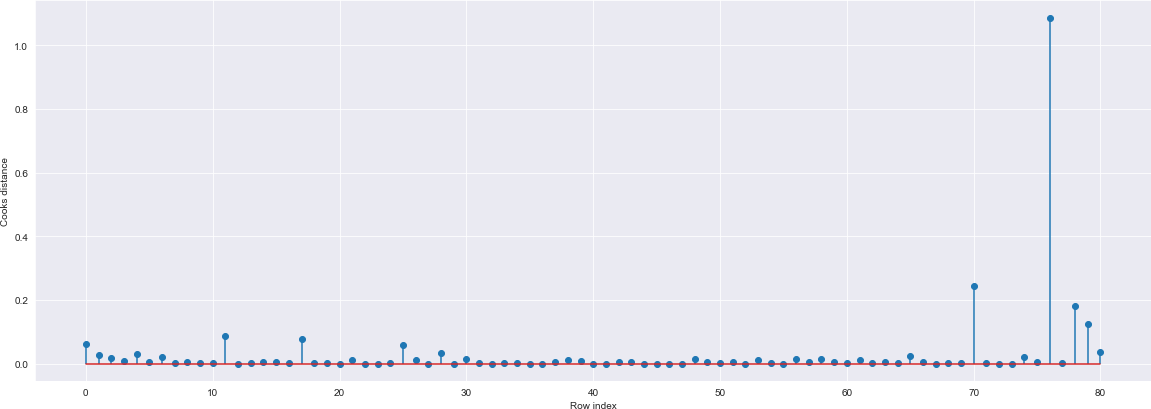
In [26]:



In [28]:



In [30]:



In [31]:

Out[31]: (76, 1.0865193998180098)

In [32]:

In [34]:

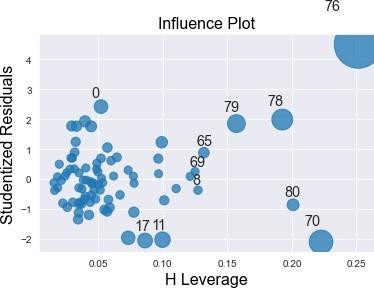
In [35]:

Out[35]:

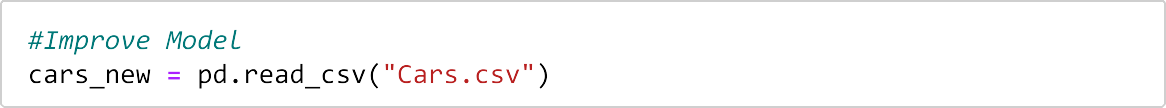
In [40]:

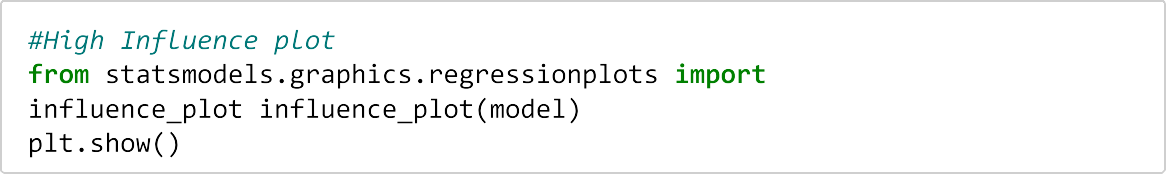
Out[40]:

In [42]:

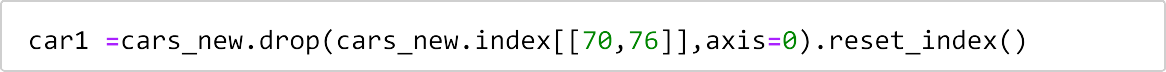


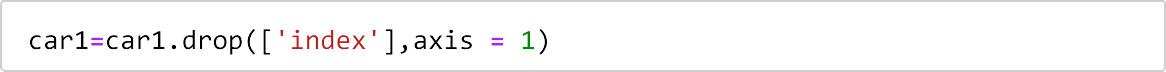




Type *Markdown* and LaTeX: 𝛼 2

In [43]: 

In [44]: 

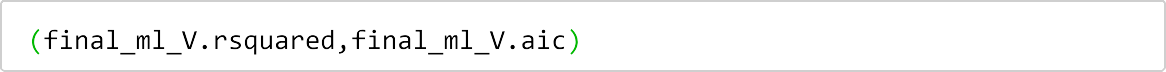
In [46]:

Out[46]:

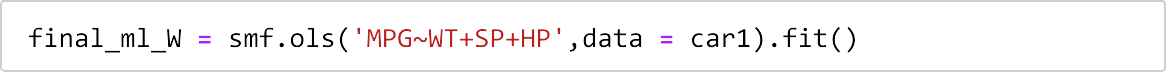
79 rows × 5 columns

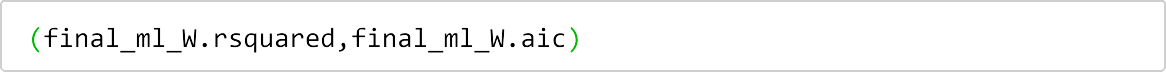
In [47]:

In [50]: 

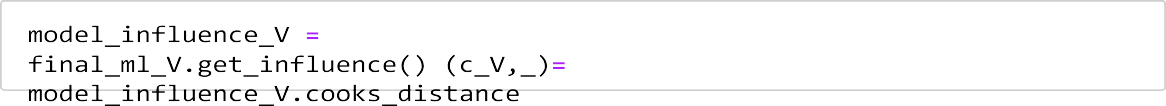
In [52]: 

Out[52]: (0.8161692010376007, 446.11722639447726)

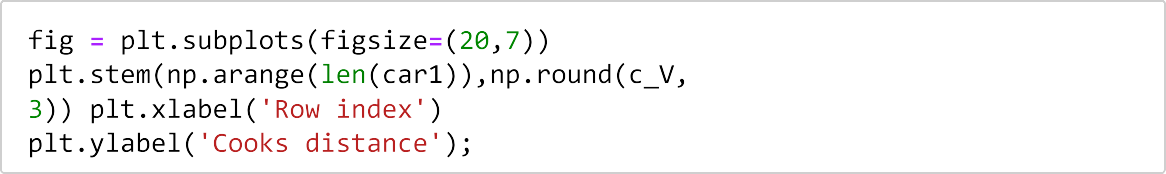
In [53]:

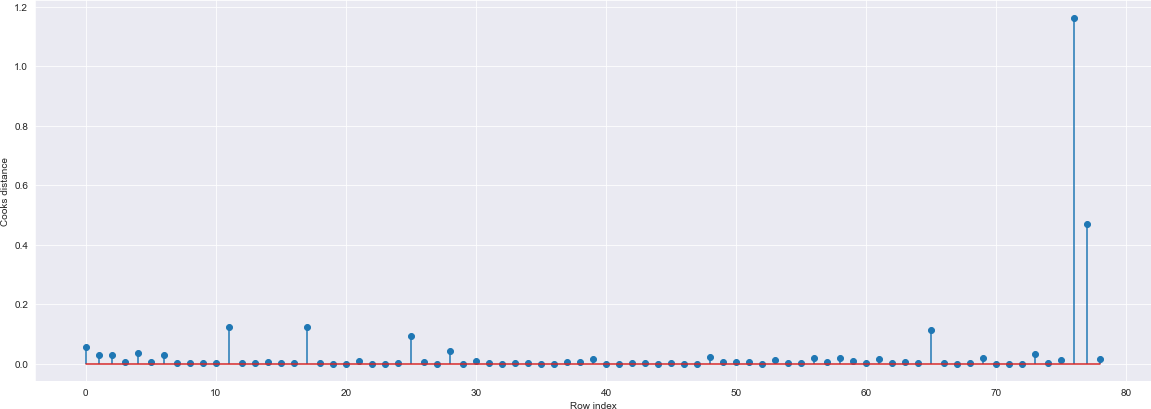
In [54]:

Out[54]: (0.8160034320495304, 446.1884323575032)

In [56]:

In [57]:



**HP MP VOL SP WT**

**G**

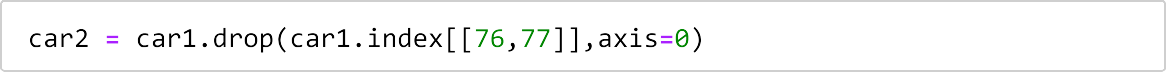
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **70** | 19. | 50 | 164.5 | 15.823060 | |
| 280 | 678 |  | 9851 |  | |
|  | 507 |  | 3 |  | |
| **76** | 36. | 50 | 169.5 | 16.132947 | |
| 322 | 900  000 |  | 9851 |  | |
| **HP** | **MP** | **VOL** | **SP** | **WT** | |
| **G** | | | | | |
| **0** | 53. | 89 | 104.1 | 28.762059 | |
| 49 | 700 |  | 8535 |  | |
|  | 681 |  | 3 |  | |
| **1** | 50. | 92 | 105.4 | 30.466833 | |
| 55 | 013 |  | 6126 |  | |
|  | 401 |  | 4 |  | |
| **2** | 50. | 92 | 105.4 | 30.193597 | |
| 55 | 013 |  | 6126 |  | |
|  | 401 |  | 4 |  | |
| **3** | 45. | 92 | 113.4 | 30.632114 | |
| 70 | 696 |  | 6126 |  | |
|  | 322 |  | 4 |  | |
| **4** | 50. | 92 | 104.4 | 29.889149 | |
| 53 | 504 |  | 6126 |  |  |
|  | 232 |  | 4 |  |  |
| car1 |  |  |  |  |  |
|  | **HP** | **MPG** | **VOL** | **SP** | **WT** |
| **0** | 49 | 53.700681 | 89 | 104.185353 | 28.762059 |
| **1** | 55 | 50.013401 | 92 | 105.461264 | 30.466833 |
| **2** | 55 | 50.013401 | 92 | 105.461264 | 30.193597 |
| **3** | 70 | 45.696322 | 92 | 113.461264 | 30.632114 |
| **4** | 53 | 50.504232 | 92 | 104.461264 | 29.889149 |
| **...** | ... | ... | ... | ... | ... |
| **74** | 175 | 18.762837 | 129 | 132.864163 | 42.778219 |
| **75** | 238 | 19.197888 | 115 | 150.576579 | 37.923113 |
| **76** | 263 | 34.000000 | 50 | 151.598513 | 15.769625 |
| **77** | 295 | 19.833733 | 119 | 167.944460 | 39.423099 |
| **78** | 236 | 12.101263 | 107 | 139.840817 | 34.948615 |

3

In [58]:

Out[58]: (76, 1.1629387469135095)

In [59]:

In [60]:

Out[60]:

77 rows × 5 columns

In [61]: 

In [62]:

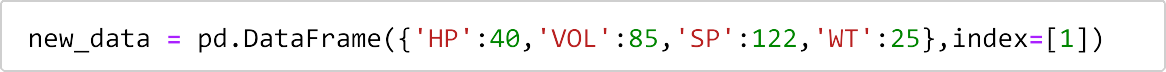
 

In [63]:

Out[63]:

77 rows × 5 columns

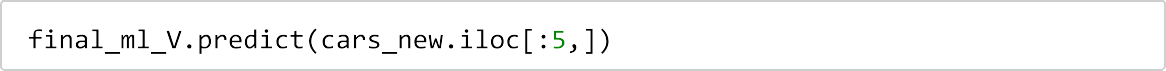
In [64]:

In [70]:

In [71]:

Out[71]: 1 54.709949

dtype: float64

In [72]: 

Out[72]: 0 44.053224

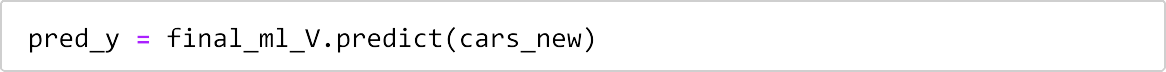
1 42.694424

2 42.694424

3 42.790029

4 42.711109

dtype: float64

In [73]:

|  |  |  |
| --- | --- | --- |
| In [74]: | pred\_ y |  |
| Out[74]: | 0 | 44.053224 |
|  | 1 | 42.694424 |
|  | 2 | 42.694424 |
|  | 3 | 42.790029 |
|  | 4 | 42.711109  ... |
| 76 | | 17.510844 |
| 77 | | 16.491552 |
| 78 | | 23.080092 |
| 79 | | 10.371281 |
| 80 | | 13.670311 |

Length: 81, dtype: float64

In [ ]:

# Conclusion:

In this study, we employed multiple linear regression to analyze the dataset and understand the relationships between multiple independent variables and a dependent variable. Our objective was to uncover patterns and associations within the data by fitting a regression model that best captures the relationships between the variables. Multiple linear regression provides a valuable tool for exploring complex relationships and making predictions based on the observed data, facilitating insightful insights into the underlying structure of the dataset.