THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

TIME SERIES MODELING & ANALYSIS

Instructor Name: Reza Jafari

Lab#: 5

Submitted by Dinesh Kumar Padmanabhan

Date: 16-Oct-2020

ABSTRACT

In LAB #5, we learned concepts of Regression. Using dataset auto.csv we implemented the multi regression model. For the very model using Least squares estimation, the sum square of errors was minimized. using datasets, we plotted various time series plots and made a comparison with respect to training, testing and different regression methods and using Coefficient of determination or R-squared measured we assessed how well a model explains and predicts future outcome.

INTRODUCTION

The auto.csv dataset will be used for this LAB. The dependent variable is 'price' and the independent variables are 'normalized-losses', 'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-size', 'bore', 'stroke', 'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg' and 'highway-mpg'.

In this LAB you want to eliminate features and find the best multiple linear regression model.

The multiple linear regression model is:

$yt = \beta 0 + \beta 1 x1,t + \beta 2 x2,t + \cdots + \beta k xk,t + \varepsilon t$

$\beta 0, \beta 1, \ldots \beta k$ are unknown values which needs to be estimated using LSE using the following equation:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where X and Y are given as :

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \cdots & x_{k,T} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}$$

Matrix X has T rows and (k+1) columns where T is the number of samples and k is number of independent variable.

METHOD, THEORY & PROCEDURES

Method:

1. Programming Language: Python

Libraries used: Some basic libraries used for analysis & model building are mentioned below

library(Numpy) - large collection of high-level mathematical functions to operate on these arrays.

library (Pandas) – For Data manipulation and analysis

library(Matplotlib) – is a system for declaratively creating graphics

library(Math) –To Compute mathematical calculations

library (statsmodels) – Import statistical models

Theory:

We want to eliminate features and find the best multiple linear regression model. To Plot the forecast accuracy of above-mentioned methods for the given data set and determine which method performs better.

Procedure:

I shall be looking at the results of various regression accuracy methods and time series plots and infer about it in my analysis. And through my exploration I shall try to identify which methods perform better and draw inferences.

The Dataset will be explored in following stages:

1. **Data Exploration (EDA)** – looking at the models and making inferences about the data.
2. **Data Visualization** – Plotting different time series plots for the regression method and forecast accuracy.
3. **Testing** – Running Autocorrelation, Pearson correlation test to identify the correlation between errors.

ANSWERS TO QUESTIONS

File - unknown

```
C:\ProgramData\Anaconda3\python.exe "C:\Program Files\JetBrains\PyCharm 2019.3.1\plugins\python\helpers\pydev\pydevconsole.py" --mode=client --port=53461

import sys; print('Python %s on %s' % (sys.version, sys.platform))
sys.path.extend(['C:\\Users\\nsree_000\\Desktop\\Python-Quiz', 'C:/Users/nsree_000/Desktop/Python-Quiz'])

Python 3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)]
Type 'copyright', 'credits' or 'license' for more information
IPython 7.8.0 -- An enhanced Interactive Python. Type '?' for help.
PyDev console: using IPython 7.8.0

Python 3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)] on win32
In[2]: runfile('C:/Users/nsree_000/Desktop/Python-Quiz/TIME SERIES/LAB5.py', wdir='C:/Users/nsree_000/Desktop/Python-Quiz/TIME SERIES')


Unknown Coefficients: [-58079.71740468      1.65121682    164.25166205    -59.57672822
    393.49354813    161.54990942      1.00056352    117.72013471
   -152.30810365  -3019.31506977    319.07356781     48.53290032
      3.07146384   -281.5442421     224.5818695 ]


                            OLS Regression Results

==============================================================================
Dep. Variable:                      y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.856
Method:                 Least Squares   F-statistic:                     68.63
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           1.25e-56
Time:                        13:49:07   Log-Likelihood
```

File - unknown

```
:                -1514.4
No. Observations:                 160   AIC
:                               3059.
Df Residuals:                     145   BIC
:                               3105.
Df Model:                          14

Covariance Type:            nonrobust

==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.888e+04   1.85e+04     -3.138      0.002   -9.47e+04   -2.15e+04
x1             1.6512     10.580      0.156      0.876      -19.259      22.561
x2           164.2517    115.175      1.426      0.156      -63.388     391.891
x3           -59.5767     62.746     -0.949      0.344    -183.592      64.438
x4           393.4935    285.660      1.377      0.170    -171.101     958.088
x5           161.5499    162.238      0.996      0.321    -159.107     482.207
x6             1.0006      1.997      0.501      0.617      -2.947       4.948
x7           117.7201     17.209      6.841      0.000      83.707     151.733
x8          -152.3081   1521.939     -0.100      0.920   -3160.359    2855.742
x9         -3019.3151    849.671     -3.554      0.001   -4698.655   -1339.975
x10          319.0736    103.795      3.074      0.003    113.927     524.220
x11           48.5329     22.888      2.120      0.036      3.296      93.770
x12            3.0715      0.936      3.283      0.001      1.222       4.921
x13         -281.5442    214.788     -1.311      0.192    -706.064     142.976
x14          224.5819    196.248      1.144      0.254    -163.293     612.457
==============================================================================
```

File - unknown

```
==================
Omnibus:                       18.174   Durbin-Watson
:                   0.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB
):               73.988
Skew:                          -0.070   Prob(JB
):                    8.58e-17
Kurtosis:                       6.328   Cond. No
.                    4.10e+05
==========================================================
==================

Warnings:
[1] Standard Errors assume that the covariance matrix of
the errors is correctly specified.
[2] The condition number is large, 4.1e+05. This might
indicate that there are
strong multicollinearity or other numerical problems.


variance of prediction error:  10752385.280026225
variance of forecast error:  16120350.653554574


0.82


                            OLS Regression Results

==========================================================
==================
Dep. Variable:                        y   R-squared
:                       0.869
Model:                              OLS   Adj. R-squared
:                   0.857
Method:                 Least Squares   F-statistic
:                     74.41
Date:                Fri, 16 Oct 2020   Prob (F-statistic
):           1.43e-57
Time:                        13:49:15   Log-Likelihood
:                -1514.4
No. Observations:                 160   AIC
:                                3057.
Df Residuals:                     146   BIC
:                                3100.
Df Model:                          13
```

Page 3 of 15

File - unknown

```
79
80 Covariance Type:            nonrobust

81 ==============================================================================
82                  coef    std err          t      P>|t|      [0.025      0.975]
83 ------------------------------------------------------------------------------
84 const       -5.88e+04    1.7e+04     -3.457      0.001   -9.24e+04   -2.52e+04
85 x1             1.7897     10.453      0.171      0.864      -18.869      22.449
86 x2           163.3571    114.438      1.427      0.156      -62.812     389.526
87 x3           -60.2266     62.197     -0.968      0.334     -183.150      62.697
88 x4           394.0880    284.628      1.385      0.168     -168.435     956.611
89 x5           163.2132    160.836      1.015      0.312     -154.655     481.082
90 x6             1.0186      1.982      0.514      0.608       -2.899       4.937
91 x7           117.8979     17.059      6.911      0.000       84.183     151.612
92 x8         -3005.5437    835.606     -3.597      0.000   -4656.990   -1354.098
93 x9           317.1114    101.580      3.122      0.002      116.354     517.869
94 x10           47.9373     22.026      2.176      0.031        4.407      91.467
95 x11            3.1064      0.865      3.590      0.000        1.396       4.817
96 x12         -279.5013    213.090     -1.312      0.192     -700.640     141.638
97 x13          224.7105    195.577      1.149      0.252     -161.817     611.238
98 ==============================================================================
99 Omnibus:                       17.965   Durbin-Watson:                   0.990
100 Prob(Omnibus):                 0.000   Jarque-Bera (JB):               72.315
101 Skew:                         -0.064   Prob(JB):                     1.98e-16
102 Kurtosis:                      6.291   Cond. No
```

File - unknown

```
.                     3.78e+05
==============================================================
====================

Warnings:
[1] Standard Errors assume that the covariance matrix of
the errors is correctly specified.
[2] The condition number is large, 3.78e+05. This might
indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results

==============================================================
====================
Dep. Variable:                          y   R-squared
:                       0.869
Model:                                OLS   Adj. R-squared
:                   0.858
Method:                  Least Squares   F-statistic
:                     81.15
Date:                 Fri, 16 Oct 2020   Prob (F-statistic
):           1.57e-58
Time:                         13:49:15   Log-Likelihood
:                -1514.4
No. Observations:                   160   AIC
:                                3055.
Df Residuals:                       147   BIC
:                                3095.
Df Model:                           12

Covariance Type:            nonrobust

==============================================================
====================
                  coef    std err          t      P>|t
|      [0.025      0.975]
--------------------------------------------------------------
--------------------
const      -5.871e+04   1.69e+04     -3.465      0.001   -
9.22e+04   -2.52e+04
x1           166.0377    112.987      1.470      0.144
       -57.251     389.327
x2           -60.6161     61.950     -0.978      0.329
      -183.044      61.812
x3           398.8909    282.306      1.413      0.160
      -159.011     956.792
```

Page 5 of 15

File - unknown

```
:                      88.93
Date:                Fri, 16 Oct 2020   Prob (F-statistic
):           1.87e-59
Time:                        13:49:15   Log-Likelihood
:                -1514.6
No. Observations:                 160   AIC
:                              3053.
Df Residuals:                     148   BIC
:                              3090.
Df Model:                          11

Covariance Type:            nonrobust

==============================================================
====================
                 coef    std err          t      P>|t
|      [0.025      0.975]
--------------------------------------------------------------
--------------------
const      -6.006e+04   1.67e+04     -3.594      0.000   -
9.31e+04    -2.7e+04
x1           178.8338    110.135      1.624      0.107
       -38.807     396.475
x2           -53.8332     60.480     -0.890      0.375
      -173.348      65.682
x3           425.3626    277.222      1.534      0.127
      -122.461     973.186
x4           158.2753    150.296      1.053      0.294
      -138.727     455.278
x5           119.7368     16.301      7.345      0.000
        87.523     151.950
x6         -2973.4431    828.249     -3.590      0.000   -
4610.164   -1336.722
x7           337.4087     93.440      3.611      0.000
       152.760     522.057
x8            51.3240     21.016      2.442      0.016
         9.795      92.853
x9             2.9923      0.827      3.620      0.000
         1.359       4.626
x10         -285.6199    207.760     -1.375      0.171
       -696.179     124.939
x11          208.6321    184.739      1.129      0.261
       -156.436     573.700
==============================================================
====================
Omnibus:                       17.907   Durbin-Watson
```

Page 7 of 15

File - unknown

```
:                   1.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB
):               72.325
Skew:                          -0.043   Prob(JB
):                1.97e-16
Kurtosis:                       6.293   Cond. No
.                    3.35e+05
==============================================================
====================

Warnings:
[1] Standard Errors assume that the covariance matrix of
the errors is correctly specified.
[2] The condition number is large, 3.35e+05. This might
indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results

==============================================================
====================
Dep. Variable:                      y   R-squared
:                      0.868
Model:                            OLS   Adj. R-squared
:                   0.859
Method:                 Least Squares   F-statistic
:                    97.88
Date:                Fri, 16 Oct 2020   Prob (F-statistic
):           2.72e-60
Time:                        13:49:16   Log-Likelihood
:                  -1515.0
No. Observations:                 160   AIC
:                              3052.
Df Residuals:                     149   BIC
:                              3086.
Df Model:                          10

Covariance Type:            nonrobust

==============================================================
====================
                 coef    std err          t      P>|t
|      [0.025      0.975]
--------------------------------------------------------------
--------------------
const      -5.875e+04   1.66e+04     -3.531      0.001   -
9.16e+04   -2.59e+04
```

Page 8 of 15

File - unknown

```
199 x1            133.0737     97.332      1.367      0.174
        -59.256     325.404
200 x2            348.9836    263.423      1.325      0.187
       -171.544     869.511
201 x3            120.7793    144.171      0.838      0.404
       -164.104     405.662
202 x4            118.9571     16.266      7.313      0.000
         86.814     151.100
203 x5          -2943.7248    826.999     -3.560      0.000   -
    4577.885   -1309.564
204 x6            328.8318     92.877      3.541      0.001
        145.306     512.358
205 x7             50.6306     20.986      2.413      0.017
          9.161      92.100
206 x8              3.0961      0.818      3.786      0.000
          1.480       4.712
207 x9           -218.9137    193.637     -1.131      0.260
       -601.544     163.716
208 x10           171.0471    179.724      0.952      0.343
       -184.089     526.183
209 ==============================================================
    ====================
210 Omnibus:                       17.374   Durbin-Watson
    :                   1.006
211 Prob(Omnibus):                  0.000   Jarque-Bera (JB
    ):                66.791
212 Skew:                          -0.078   Prob(JB
    ):                     3.14e-15
213 Kurtosis:                       6.161   Cond. No
    .                       3.34e+05
214 ==============================================================
    ====================
215
216 Warnings:
217 [1] Standard Errors assume that the covariance matrix of
    the errors is correctly specified.
218 [2] The condition number is large, 3.34e+05. This might
    indicate that there are
219 strong multicollinearity or other numerical problems.
220                             OLS Regression Results

221 ==============================================================
    ====================
222 Dep. Variable:                         y   R-squared
    :                       0.867
223 Model:                             OLS   Adj. R-squared
```

Page 9 of 15

File - unknown

```
:                  0.859
Method:                 Least Squares   F-statistic
:                      108.9
Date:                Fri, 16 Oct 2020   Prob (F-statistic
):           3.58e-61
Time:                        13:49:16   Log-Likelihood
:                  -1515.4
No. Observations:                 160   AIC
:                                3051.
Df Residuals:                     150   BIC
:                                3082.
Df Model:                           9

Covariance Type:            nonrobust

==============================================================
====================
                 coef    std err          t      P>|t
|      [0.025      0.975]
--------------------------------------------------------------
--------------------
const      -5.297e+04   1.51e+04     -3.502      0.001   -
8.29e+04   -2.31e+04
x1           178.8177     80.493      2.222      0.028
       19.771     337.865
x2           305.6295    258.033      1.184      0.238
     -204.219     815.478
x3           117.4667     16.153      7.272      0.000
       85.550     149.383
x4         -3048.7616    816.626     -3.733      0.000   -
4662.338   -1435.185
x5           331.0819     92.746      3.570      0.000
      147.825     514.339
x6            50.5417     20.965      2.411      0.017
        9.116      91.967
x7             3.0132      0.811      3.715      0.000
        1.411       4.616
x8          -215.8108    193.409     -1.116      0.266
     -597.970     166.348
x9           168.0934    179.510      0.936      0.351
     -186.602     522.789
==============================================================
====================
Omnibus:                       17.709   Durbin-Watson
:                  1.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB
```

File - unknown

```
246 ):                70.213
247 Skew:                           -0.059   Prob(JB
    ):                    5.67e-16
248 Kurtosis:                        6.243   Cond. No
    .                     3.04e+05
249 ============================================================
    ====================
250
251 Warnings:
252 [1] Standard Errors assume that the covariance matrix of
    the errors is correctly specified.
253 [2] The condition number is large, 3.04e+05. This might
    indicate that there are
254 strong multicollinearity or other numerical problems.
255                             OLS Regression Results

256 ============================================================
    ====================
257 Dep. Variable:                       y   R-squared
    :                       0.866
258 Model:                            OLS   Adj. R-squared
    :                   0.859
259 Method:                 Least Squares   F-statistic
    :                    122.5
260 Date:                Fri, 16 Oct 2020   Prob (F-statistic
    ):            4.83e-62
261 Time:                        13:49:16   Log-Likelihood
    :                -1515.9
262 No. Observations:                 160   AIC
    :                              3050.
263 Df Residuals:                     151   BIC
    :                              3077.
264 Df Model:                           8

265 Covariance Type:            nonrobust

266 ============================================================
    ====================
267                  coef    std err          t      P>|t
    |      [0.025      0.975]
268 ------------------------------------------------------------
    --------------------
269 const      -4.952e+04   1.47e+04     -3.377      0.001   -
    7.85e+04   -2.06e+04
270 x1           168.3419     79.679      2.113      0.036
         10.911     325.772
```

Page 11 of 15

File - unknown

```
x2           281.3419    256.621      1.096      0.275      -225.690     788.374
x3           115.4832     16.007      7.215      0.000        83.857     147.109
x4         -2979.3949    812.928     -3.665      0.000   -4585.577   -1373.213
x5           328.8373     92.677      3.548      0.001       145.726     511.948
x6            52.5411     20.848      2.520      0.013        11.350      93.732
x7             3.0272      0.811      3.735      0.000         1.426       4.629
x8           -52.2924     83.114     -0.629      0.530      -216.508     111.923
==============================================================================
Omnibus:                       17.933   Durbin-Watson:                   0.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               72.250
Skew:                          -0.056   Prob(JB):                     2.05e-16
Kurtosis:                       6.290   Cond. No.                     2.94e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.94e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results

==============================================================================
Dep. Variable:                      y   R-squared:                       0.866
Model:                            OLS   Adj. R-squared:                  0.860
Method:                 Least Squares   F-statistic:                     140.5
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           4.77e-63
Time:                        13:49:16   Log-Likelihood
```

Page 12 of 15

File - unknown

```
295 :                -1516.1
296 No. Observations:                 160   AIC
    :                             3048.
297 Df Residuals:                     152   BIC
    :                             3073.
298 Df Model:                           7

299 Covariance Type:            nonrobust

300 ==============================================================
    ====================
301                   coef    std err          t      P>|t
    |      [0.025      0.975]
302 --------------------------------------------------------------
    --------------------
303 const      -5.448e+04   1.23e+04     -4.414      0.000   -
    7.89e+04   -3.01e+04
304 x1           183.9395     75.574      2.434      0.016
         34.628     333.251
305 x2           314.8122    250.547      1.256      0.211
       -180.193     809.817
306 x3           112.2363     15.122      7.422      0.000
         82.360     142.113
307 x4         -2984.4499    811.271     -3.679      0.000   -
    4587.273   -1381.627
308 x5           305.5303     84.783      3.604      0.000
        138.026     473.035
309 x6            60.6620     16.339      3.713      0.000
         28.381      92.943
310 x7             2.9749      0.805      3.697      0.000
          1.385       4.565
311 ==============================================================
    ====================
312 Omnibus:                       17.915   Durbin-Watson
    :                    0.991
313 Prob(Omnibus):                  0.000   Jarque-Bera (JB
    ):                71.425
314 Skew:                          -0.878   Prob(JB
    ):                     3.09e-16
315 Kurtosis:                       6.269   Cond. No
    .                      2.48e+05
316 ==============================================================
    ====================
317
318 Warnings:
319 [1] Standard Errors assume that the covariance matrix of
```

File - unknown

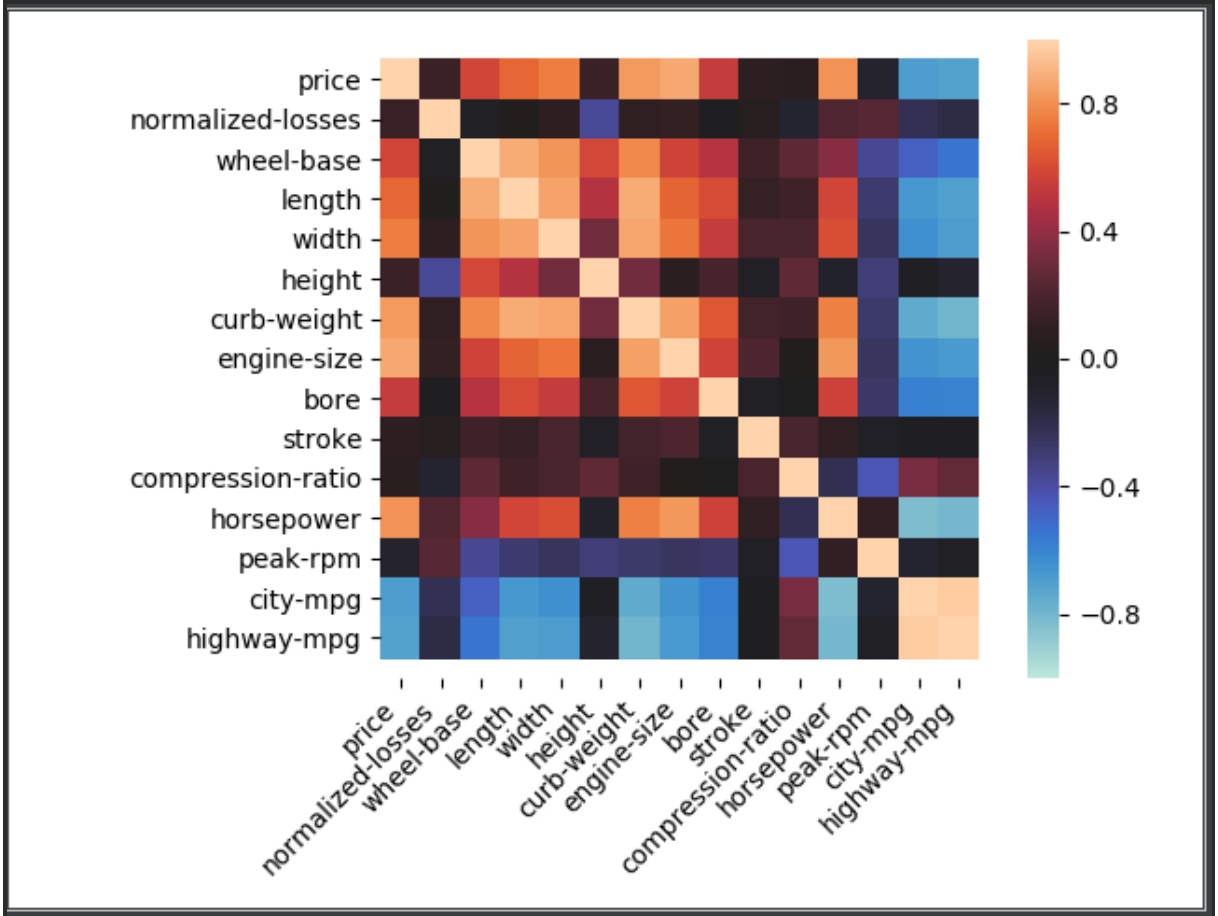
```
the errors is correctly specified.
[2] The condition number is large, 2.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.865
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     163.0
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           7.81e-64
Time:                        13:49:16   Log-Likelihood:                -1516.9
No. Observations:                 160   AIC:                             3048.
Df Residuals:                     153   BIC:                             3069.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.164e+04   6932.165     -6.006      0.000   -5.53e+04   -2.79e+04
x1           250.5779     53.941      4.645      0.000     144.012     357.144
x2           113.9081     15.092      7.548      0.000      84.093     143.723
x3         -2945.3206    812.205     -3.626      0.000   -4549.905   -1340.736
x4           321.4083     83.994      3.827      0.000     155.471     487.346
x5            66.9018     15.595      4.290      0.000      36.091      97.712
x6             3.0111      0.806      3.737      0.000       1.419       4.603
```

File - unknown

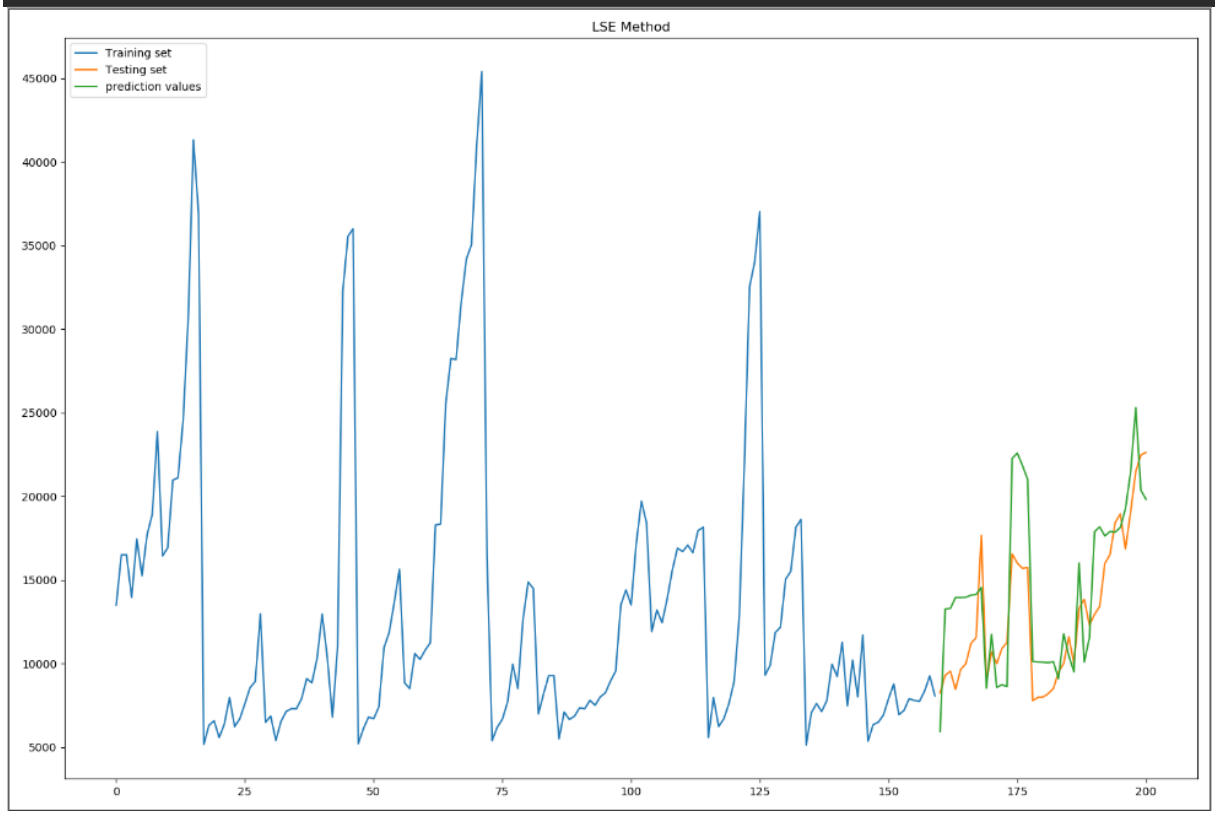
```
343 ==============================================================
    ====================
344 Omnibus:                       18.109   Durbin-Watson
    :                    1.032
345 Prob(Omnibus):                  0.000   Jarque-Bera (JB
    ):                70.826
346 Skew:                          -0.128   Prob(JB
    ):                     4.17e-16
347 Kurtosis:                       6.249   Cond. No
    .                      1.39e+05
348 ==============================================================
    ====================
349
350 Warnings:
351 [1] Standard Errors assume that the covariance matrix of
    the errors is correctly specified.
352 [2] The condition number is large, 1.39e+05. This might
    indicate that there are
353 strong multicollinearity or other numerical problems.
354
```

Page 16 of 16

```
#%%==============================================================================
#1: Using Pandas library load the time series data from the BB.
# Split the dataset into training set and test set.
# Use 80% for training and 20% for testing. Display the training set and testing
set array as follow:
# Hint: This can be done using the following library in python.
# “from sklearn.model_selection import train_test_split” . Make sure the
“Shuffle=False”
# %%------------------------------------------------------------------------------

# See Appendix section for code.

#%%==============================================================================
#2: Plot the correlation matrix using the seaborn package and heatmap function.
#%%==============================================================================
```

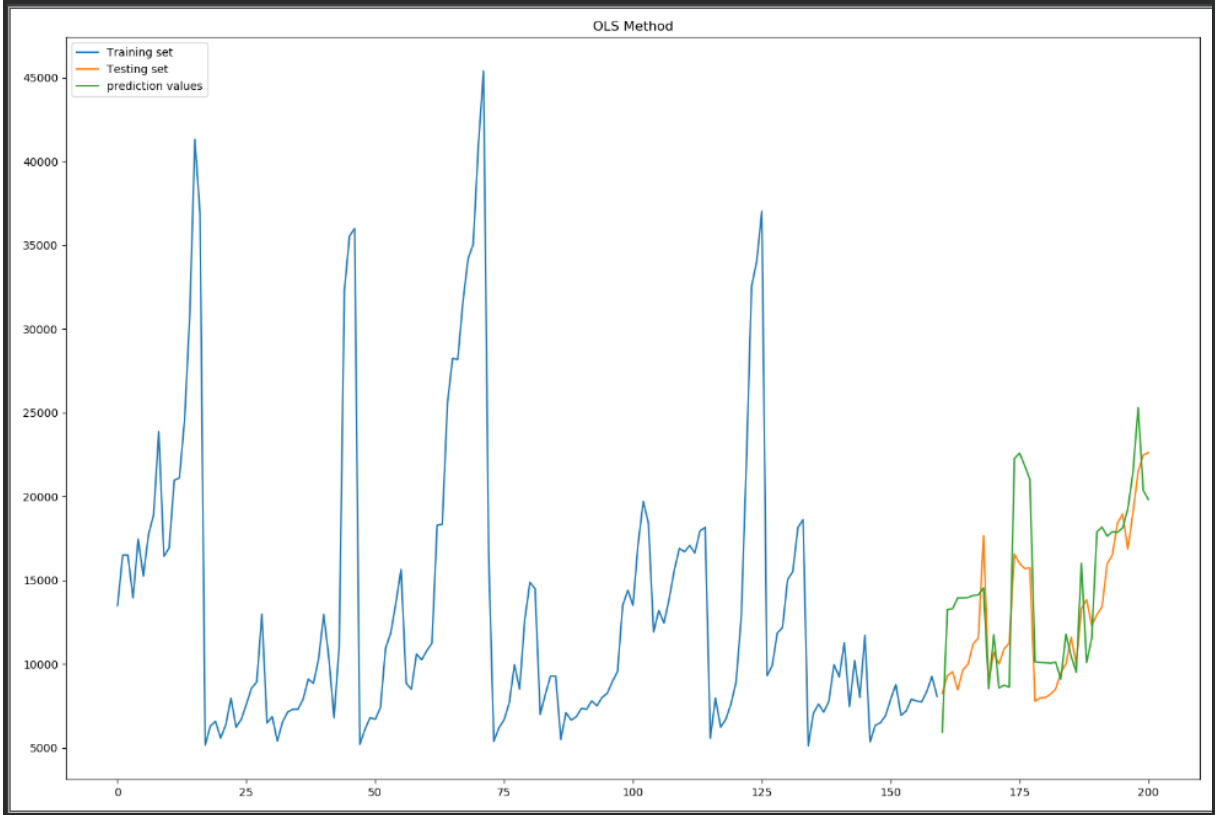


```
#%%=============================================================================
#3: Using python, construct matrix X and Y using x-train and y-train dataset and
estimate the regression model unknown coefficients using the Normal equation (LSE
method, above equation).
Display the unknown coefficients on the console.
Note: You are not allowed to use OLS for this part.
# %%-----------------------------------------------------------------------------

Unknown Coefficients: [-58079.71740468      1.65121682    164.25166205    -
59.57672822
    393.49354813    161.54990942      1.00056352    117.72013471
   -152.30810365  -3019.31506977    319.07356781     48.53290032
      3.07146384   -281.5442421     224.5818695 ]
```

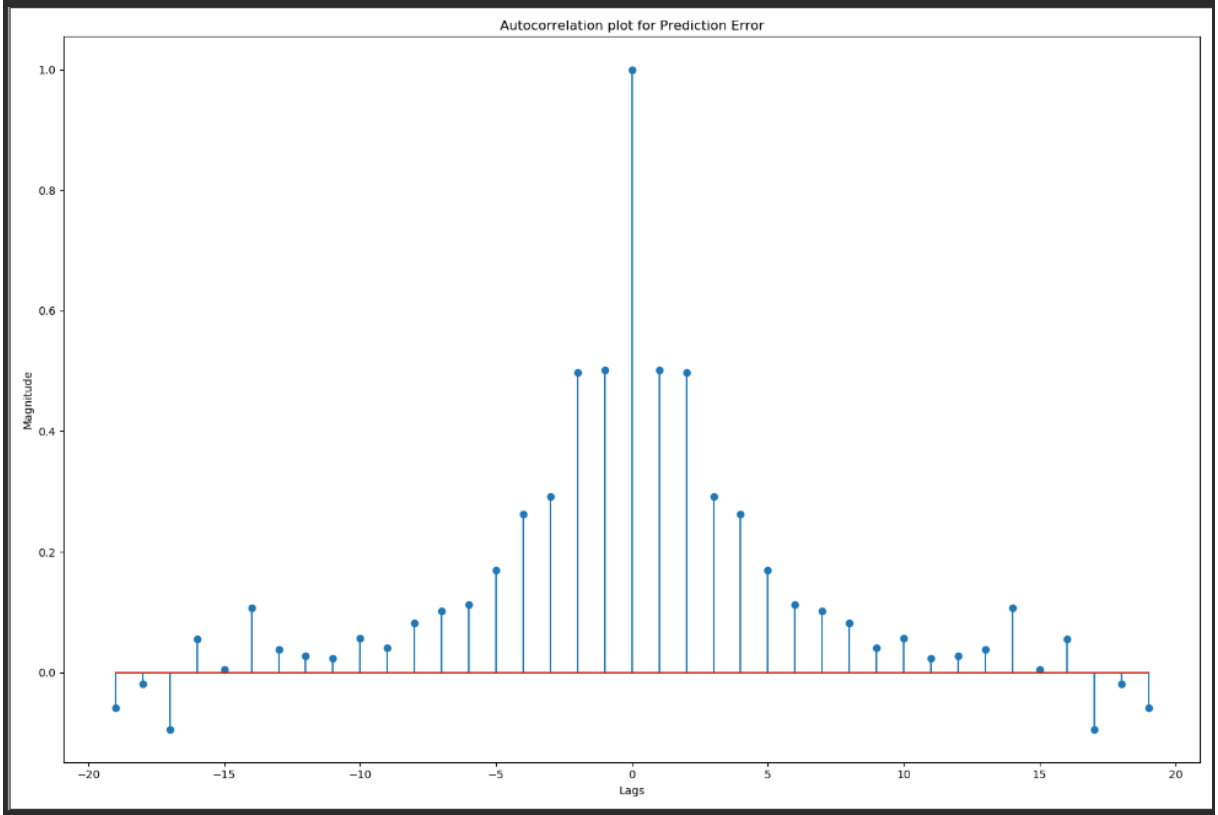


```
#%%=============================================================================
#4: Using python, statsmodels package and OLS function, find the unknown
# coefficients. Compare the results with the step 3.
# Display the result on the console.
# Are the unknown coefficient calculated using step 3 and step 4 the same?
# %%-----------------------------------------------------------------------------


OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.856
Method:                 Least Squares   F-statistic:                     68.63
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           1.25e-56
Time:                        13:31:46   Log-Likelihood:                -1514.4
No. Observations:                 160   AIC:                             3059.
Df Residuals:                     145   BIC:                             3105.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.808e+04   1.85e+04     -3.138      0.002   -9.47e+04   -2.15e+04
x1             1.6512     10.580      0.156      0.876     -19.259      22.561
x2           164.2517    115.175      1.426      0.156     -63.388     391.891
x3           -59.5767     62.746     -0.949      0.344    -183.592      64.438
x4           393.4935    285.660      1.377      0.170    -171.101     958.088
x5           161.5499    162.238      0.996      0.321    -159.107     482.207
x6             1.0006      1.997      0.501      0.617      -2.947       4.948
x7           117.7201     17.209      6.841      0.000      83.707     151.733
x8          -152.3081   1521.939     -0.100      0.920   -3160.359    2855.742
x9         -3019.3151    849.671     -3.554      0.001   -4698.655   -1339.975
x10          319.0736    103.795      3.074      0.003     113.927     524.220
x11           48.5329     22.888      2.120      0.036       3.296      93.770
x12            3.0715      0.936      3.283      0.001       1.222       4.921
x13         -281.5442    214.788     -1.311      0.192    -706.064     142.976
x14          224.5819    196.248      1.144      0.254    -163.293     612.457
==============================================================================
Omnibus:                       18.174   Durbin-Watson:                   0.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               73.988
Skew:                          -0.070   Prob(JB):                     8.58e-17
Kurtosis:                       6.328   Cond. No.                     4.10e+05
==============================================================================

'''
Yes, the unknown coefficient calculated using step 3 and step 4 are same.

'''
```

```
#%%=========================================================================================
# 5: Perform a prediction for the length of test-set and plot the train, test and
predicted values in one graph.
#Add appropriate x-label, y-label, title, and legend to your graph.
# %%-------------------------------------------------------------------------------------
```

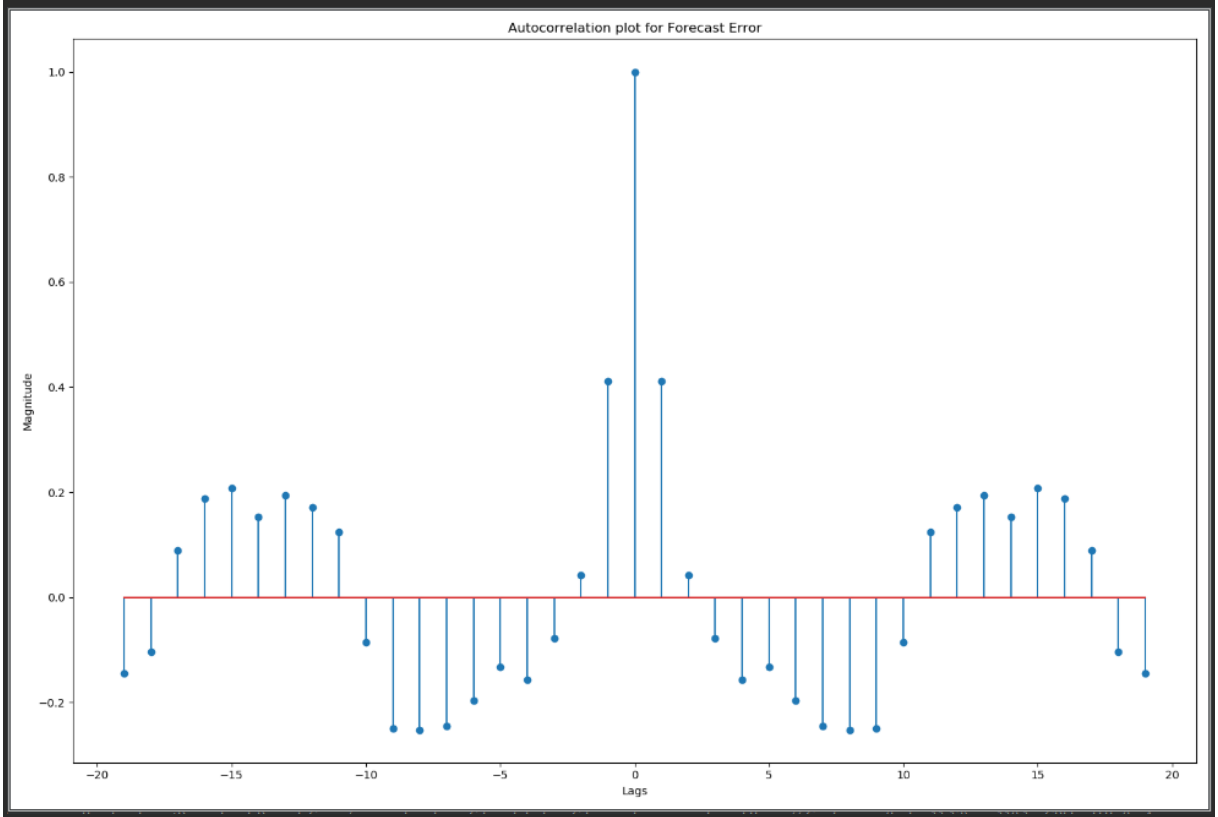


```
#%%=============================================================================
# 6: Calculate the prediction errors and plot the ACF of prediction errors.
# Write down your observation
# %%-----------------------------------------------------------------------------
```



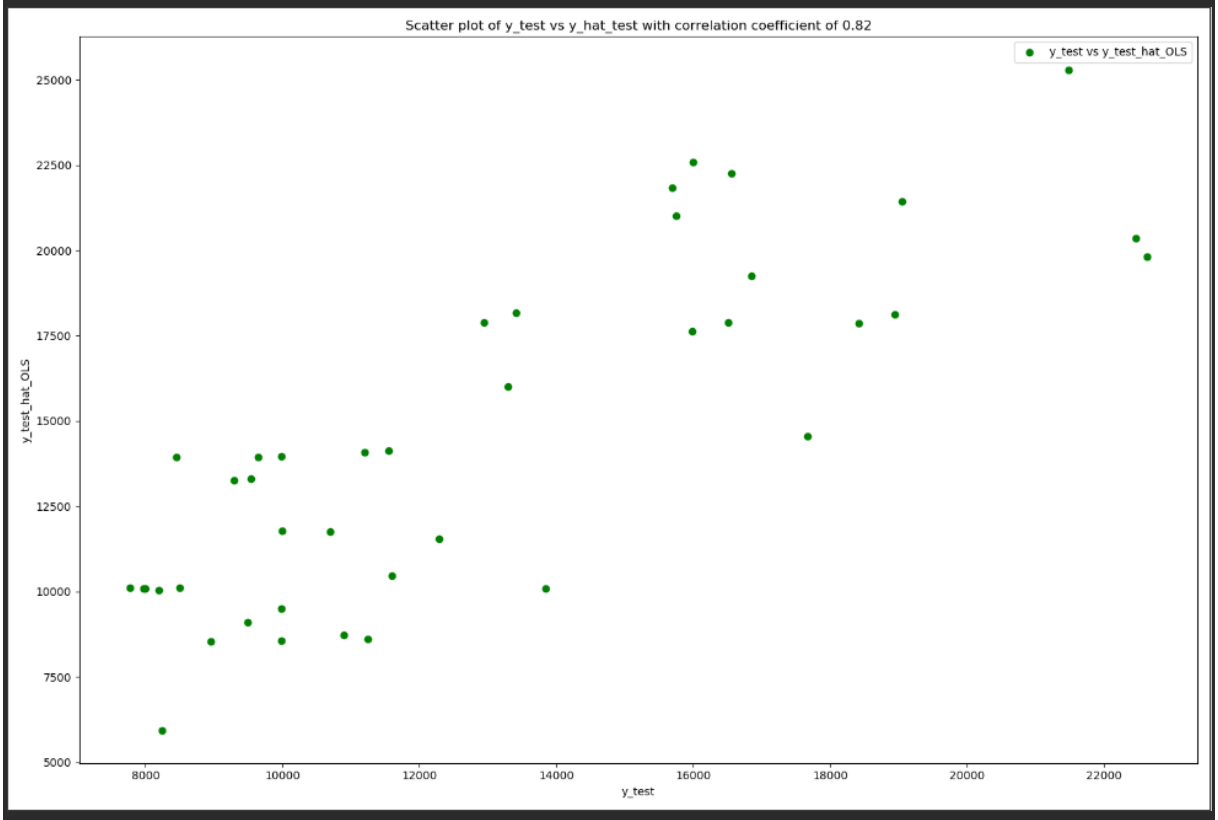
Autocorrelation of have a strong peak at ”0”. The values are distributed identically with mean of zero and variance of σ 2.

```
#%%========================================================================================
# 7: Calculate the forecast errors and plot the ACF of forecast errors.
# Write down your observation.
# %%----------------------------------------------------------------------------------------
```




```
Autocorrelation of  have a strong peak at ”0”. The values are distributed
identically with mean of zero and variance of σ 2.
```

```
#%%=========================================================================
#8: Calculate the estimated variance of the prediction errors and the forecast
errors. Compare the results?
# What is your observation?
Hint: Use need to the following equation to estimate the variance:
# %%-------------------------------------------------------------------------

variance of prediction error:  10752385.280026225
variance of forecast error:  16120350.653554574.

The condition number for forecast error is large. This might indicate that there
strong multicollinearity.


#%%=========================================================================
#9: Plot the scatter plot between y-test and ŷt+h and display the correlation
coefficient between them on the title.
# Justify the accuracy of this predictor by observing the correlation coefficient
between y-test and ŷt.
# %%-------------------------------------------------------------------------
```




```
Accuracy of the predictor kind of matches with correlation coefficient which is
positive in nature. Here we can see the extend of correlation between y_test v/s
y_test_hat_ols
```

```
#%%========================================================================
#10: Using a stepwise regression, try to reduce the feature space dimension.
# You need to use the AIC, BIC and Adjusted R2 as a predictive accuracy for your
analysis.
# If your analysis recommends an elimination, which feature(s) would you
eliminate?
# You can use the backward or forward or hybrid stepwise regression for feature
selection.
# %%------------------------------------------------------------------------

# Removing bore
OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.857
Method:                 Least Squares   F-statistic:                     74.41
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           1.43e-57
Time:                        15:42:51   Log-Likelihood:                -1514.4
No. Observations:                 160   AIC:                             3057.
Df Residuals:                     146   BIC:                             3100.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.88e+04    1.7e+04     -3.457      0.001   -9.24e+04   -2.52e+04
x1             1.7897     10.453      0.171      0.864     -18.869      22.449
x2           163.3571    114.438      1.427      0.156     -62.812     389.526
x3           -60.2266     62.197     -0.968      0.334    -183.150      62.697
x4           394.0880    284.628      1.385      0.168    -168.435     956.611
x5           163.2132    160.836      1.015      0.312    -154.655     481.082
x6             1.0186      1.982      0.514      0.608      -2.899       4.937
x7           117.8979     17.059      6.911      0.000      84.183     151.612
x8         -3005.5437    835.606     -3.597      0.000   -4656.990   -1354.098
x9           317.1114    101.580      3.122      0.002     116.354     517.869
x10           47.9373     22.026      2.176      0.031       4.407      91.467
x11            3.1064      0.865      3.590      0.000       1.396       4.817
x12         -279.5013    213.090     -1.312      0.192    -700.640     141.638
x13          224.7105    195.577      1.149      0.252    -161.817     611.238
==============================================================================
Omnibus:                       17.965   Durbin-Watson:                   0.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               72.315
Skew:                          -0.064   Prob(JB):                     1.98e-16
Kurtosis:                       6.291   Cond. No.                     3.78e+05
==============================================================================
```

```
#Removing normalized losses
OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.858
Method:                 Least Squares   F-statistic:                     81.15
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           1.57e-58
Time:                        15:42:51   Log-Likelihood:                -1514.4
No. Observations:                 160   AIC:                             3055.
Df Residuals:                     147   BIC:                             3095.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.871e+04   1.69e+04     -3.465      0.001   -9.22e+04   -2.52e+04
x1           166.0377    112.987      1.470      0.144     -57.251     389.327
x2           -60.6161     61.950     -0.978      0.329    -183.044      61.812
x3           398.8909    282.306      1.413      0.160    -159.011     956.792
x4           153.9118    150.883      1.020      0.309    -144.269     452.092
x5             1.0483      1.968      0.533      0.595      -2.842       4.938
x6           117.5226     16.862      6.970      0.000      84.200     150.845
x7         -3007.6949    832.748     -3.612      0.000   -4653.400   -1361.990
x8           316.9487    101.240      3.131      0.002     116.876     517.022
x9            48.0568     21.942      2.190      0.030       4.695      91.419
x10            3.1160      0.861      3.621      0.000       1.415       4.817
x11         -286.6455    208.273     -1.376      0.171    -698.242     124.951
x12          231.9652    190.300      1.219      0.225    -144.113     608.043
==============================================================================
Omnibus:                       17.991   Durbin-Watson:                   0.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               72.485
Skew:                          -0.066   Prob(JB):                     1.82e-16
Kurtosis:                       6.295   Cond. No.                     3.77e+05
==============================================================================
```

```
#Removing curb-weight
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     88.93
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           1.87e-59
Time:                        15:42:52   Log-Likelihood:                -1514.6
No. Observations:                 160   AIC:                             3053.
Df Residuals:                     148   BIC:                             3090.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -6.006e+04   1.67e+04     -3.594      0.000   -9.31e+04    -2.7e+04
x1           178.8338    110.135      1.624      0.107     -38.807     396.475
x2           -53.8332     60.480     -0.890      0.375    -173.348      65.682
x3           425.3626    277.222      1.534      0.127    -122.461     973.186
x4           158.2753    150.296      1.053      0.294    -138.727     455.278
x5           119.7368     16.301      7.345      0.000      87.523     151.950
x6         -2973.4431    828.249     -3.590      0.000   -4610.164   -1336.722
x7           337.4087     93.440      3.611      0.000     152.760     522.057
x8            51.3240     21.016      2.442      0.016       9.795      92.853
x9             2.9923      0.827      3.620      0.000       1.359       4.626
x10         -285.6199    207.760     -1.375      0.171    -696.179     124.939
x11          208.6321    184.739      1.129      0.261    -156.436     573.700
==============================================================================
Omnibus:                       17.907   Durbin-Watson:                   1.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               72.325
Skew:                          -0.043   Prob(JB):                     1.97e-16
Kurtosis:                       6.293   Cond. No.                     3.35e+05
==============================================================================
```

```
#Removing length
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     97.88
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           2.72e-60
Time:                        15:42:52   Log-Likelihood:                -1515.0
No. Observations:                 160   AIC:                             3052.
Df Residuals:                     149   BIC:                             3086.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.875e+04   1.66e+04     -3.531      0.001   -9.16e+04   -2.59e+04
x1           133.0737     97.332      1.367      0.174     -59.256     325.404
x2           348.9836    263.423      1.325      0.187    -171.544     869.511
x3           120.7793    144.171      0.838      0.404    -164.104     405.662
x4           118.9571     16.266      7.313      0.000      86.814     151.100
x5         -2943.7248    826.999     -3.560      0.000   -4577.885   -1309.564
x6           328.8318     92.877      3.541      0.001     145.306     512.358
x7            50.6306     20.986      2.413      0.017       9.161      92.100
x8             3.0961      0.818      3.786      0.000       1.480       4.712
x9          -218.9137    193.637     -1.131      0.260    -601.544     163.716
x10          171.0471    179.724      0.952      0.343    -184.089     526.183
==============================================================================
Omnibus:                       17.374   Durbin-Watson:                   1.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               66.791
Skew:                          -0.078   Prob(JB):                     3.14e-15
Kurtosis:                       6.161   Cond. No.                     3.34e+05
==============================================================================
```

```
#Removing height
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.867
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     108.9
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           3.58e-61
Time:                        15:42:52   Log-Likelihood:                -1515.4
No. Observations:                 160   AIC:                             3051.
Df Residuals:                     150   BIC:                             3082.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.297e+04   1.51e+04     -3.502      0.001   -8.29e+04   -2.31e+04
x1           178.8177     80.493      2.222      0.028      19.771     337.865
x2           305.6295    258.033      1.184      0.238    -204.219     815.478
x3           117.4667     16.153      7.272      0.000      85.550     149.383
x4         -3048.7616    816.626     -3.733      0.000   -4662.338   -1435.185
x5           331.0819     92.746      3.570      0.000     147.825     514.339
x6            50.5417     20.965      2.411      0.017       9.116      91.967
x7             3.0132      0.811      3.715      0.000       1.411       4.616
x8          -215.8108    193.409     -1.116      0.266    -597.970     166.348
x9           168.0934    179.510      0.936      0.351    -186.602     522.789
==============================================================================
Omnibus:                       17.709   Durbin-Watson:                   1.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               70.213
Skew:                          -0.059   Prob(JB):                     5.67e-16
Kurtosis:                       6.243   Cond. No.                     3.04e+05
==============================================================================
```

```
#Removing highway-mpg

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.866
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     122.5
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           4.83e-62
Time:                        15:42:52   Log-Likelihood:                -1515.9
No. Observations:                 160   AIC:                             3050.
Df Residuals:                     151   BIC:                             3077.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.952e+04   1.47e+04     -3.377      0.001   -7.85e+04   -2.06e+04
x1           168.3419     79.679      2.113      0.036      10.911     325.772
x2           281.3419    256.621      1.096      0.275    -225.690     788.374
x3           115.4832     16.007      7.215      0.000      83.857     147.109
x4         -2979.3949    812.928     -3.665      0.000   -4585.577   -1373.213
x5           328.8373     92.677      3.548      0.001     145.726     511.948
x6            52.5411     20.848      2.520      0.013      11.350      93.732
x7             3.0272      0.811      3.735      0.000       1.426       4.629
x8           -52.2924     83.114     -0.629      0.530    -216.508     111.923
==============================================================================
Omnibus:                       17.933   Durbin-Watson:                   0.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               72.250
Skew:                          -0.056   Prob(JB):                     2.05e-16
Kurtosis:                       6.290   Cond. No.                     2.94e+05
==============================================================================
```

```
#Removing city-mpg

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.866
Model:                            OLS   Adj. R-squared:                  0.860
Method:                 Least Squares   F-statistic:                     140.5
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           4.77e-63
Time:                        15:42:52   Log-Likelihood:                -1516.1
No. Observations:                 160   AIC:                             3048.
Df Residuals:                     152   BIC:                             3073.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.448e+04   1.23e+04     -4.414      0.000   -7.89e+04   -3.01e+04
x1           183.9395     75.574      2.434      0.016      34.628     333.251
x2           314.8122    250.547      1.256      0.211    -180.193     809.817
x3           112.2363     15.122      7.422      0.000      82.360     142.113
x4         -2984.4499    811.271     -3.679      0.000   -4587.273   -1381.627
x5           305.5303     84.783      3.604      0.000     138.026     473.035
x6            60.6620     16.339      3.713      0.000      28.381      92.943
x7             2.9749      0.805      3.697      0.000       1.385       4.565
==============================================================================
Omnibus:                       17.915   Durbin-Watson:                   0.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               71.425
Skew:                          -0.078   Prob(JB):                     3.09e-16
Kurtosis:                       6.269   Cond. No.                     2.48e+05
==============================================================================
```

```
#Removing width

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.865
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     163.0
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           7.81e-64
Time:                        15:42:52   Log-Likelihood:                -1516.9
No. Observations:                 160   AIC:                             3048.
Df Residuals:                     153   BIC:                             3069.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.164e+04   6932.165     -6.006      0.000   -5.53e+04   -2.79e+04
x1           250.5779     53.941      4.645      0.000     144.012     357.144
x2           113.9081     15.092      7.548      0.000      84.093     143.723
x3         -2945.3206    812.205     -3.626      0.000   -4549.905   -1340.736
x4           321.4083     83.994      3.827      0.000     155.471     487.346
x5            66.9018     15.595      4.290      0.000      36.091      97.712
x6             3.0111      0.806      3.737      0.000       1.419       4.603
==============================================================================
Omnibus:                       18.109   Durbin-Watson:                   1.032
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               70.826
Skew:                          -0.128   Prob(JB):                     4.17e-16
Kurtosis:                       6.249   Cond. No.                     1.39e+05
==============================================================================
```

```
#%%========================================================================
#11: Perform a complete t-test and F-test analysis on the final model and write
down your observations
# %%------------------------------------------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.865
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     163.0
Date:                Fri, 16 Oct 2020   Prob (F-statistic):           7.81e-64
Time:                        15:42:52   Log-Likelihood:                -1516.9
No. Observations:                 160   AIC:                             3048.
Df Residuals:                     153   BIC:                             3069.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.164e+04   6932.165     -6.006      0.000   -5.53e+04   -2.79e+04
x1           250.5779     53.941      4.645      0.000     144.012     357.144
x2           113.9081     15.092      7.548      0.000      84.093     143.723
x3         -2945.3206    812.205     -3.626      0.000   -4549.905   -1340.736
x4           321.4083     83.994      3.827      0.000     155.471     487.346
x5            66.9018     15.595      4.290      0.000      36.091      97.712
x6             3.0111      0.806      3.737      0.000       1.419       4.603
==============================================================================
Omnibus:                       18.109   Durbin-Watson:                   1.032
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               70.826
Skew:                          -0.128   Prob(JB):                     4.17e-16
Kurtosis:                       6.249   Cond. No.                     1.39e+05
==============================================================================

An F-test is a type of statistical test that is very flexible. You can use them in
a wide variety of settings. F-tests can evaluate multiple model terms
simultaneously, which allows them to compare the fits of different linear models.
In contrast, t-tests can evaluate just one term at a time.
In statistical output, we can find the overall ols result in the above table.
Comparing the p-value for the F-test to your significance level.
Here the p-value is less than the significance level, the sample data provided has
sufficient evidence to conclude that your regression model fits the data better
than the model with no independent variables.

This finding is good news because it means that the independent variables in your
model improve the fit!
```

CONCLUSION

For the data set provided we applied multi linear regression. For the very model using Least squares estimation, the sum square of errors was minimized. using datasets, we plotted various time series plots and made a comparison with respect to training, testing and different regression methods. using Coefficient of determination or R-squared measured we assessed how well a model explains and predicts future outcome. Below table summarizes the regression model and their accuracy after stepwise regression.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.865
Model:	OLS	Adj. R-squared:	0.859
Method:	Least Squares	F-statistic:	163.0
Date:	Fri, 16 Oct 2020	Prob (F-statistic):	7.81e-64
Time:	17:22:49	Log-Likelihood:	-1516.9
No. Observations:	160	AIC:	3048.
Df Residuals:	153	BIC:	3069.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.164e+04	6932.165	-6.006	0.000	-5.53e+04	-2.79e+04
x1	250.5779	53.941	4.645	0.000	144.012	357.144
x2	113.9081	15.092	7.548	0.000	84.093	143.723
x3	-2945.3206	812.205	-3.626	0.000	-4549.905	-1340.736
x4	321.4083	83.994	3.827	0.000	155.471	487.346
x5	66.9018	15.595	4.290	0.000	36.091	97.712
x6	3.0111	0.806	3.737	0.000	1.419	4.603

Omnibus:	18.109	Durbin-Watson:	1.032
Prob(Omnibus):	0.000	Jarque-Bera (JB):	70.826
Skew:	-0.128	Prob(JB):	4.17e-16
Kurtosis:	6.249	Cond. No.	1.39e+05

CHALLENGE

Calculations was little tricky to understand in the beginning, after lot of clarifications it provided clarity.

APPENDIX

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import statsmodels.api as sm
import warnings
import seaborn as sns

from Autocorrelation import cal_auto_corr
from Pearson_Correlation_Coefficient import correlation_coefficent_cal

warnings.filterwarnings("ignore")
np.set_printoptions(suppress=True)

df = pd.read_csv('autos.clean_corr.csv')
df = df[['price','normalized-losses', 'wheel-base', 'length', 'width', 'height',
'curb-weight', 'engine-size',
         'bore', 'stroke','compression-ratio', 'horsepower', 'peak-rpm', 'city-
mpg','highway-mpg']]

features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()

features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)

corr = df.corr()
ax = sns.heatmap(corr, vmin=-1, vmax=1, center=0, square=True)
bottom, top = ax.get_ylim()
ax.set_ylim(bottom + 0.5, top - 0.5)
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.figure(figsize=(15,10))
plt.show()

def plot_fun(train, test, predicted, title):
    plt.figure(figsize=(15,10))
    plt.plot(range(0, len(train)), train, label='Training set')
    plt.plot(range(len(train), len(train)+len(test)), test, label='Testing set')
    plt.plot(range(len(train), len(train)+ len(predicted)), predicted,
label='prediction values')
    plt.xlabel('')
    plt.title(title)
    plt.legend(loc='upper left')
    plt.show()

x_transpose = np.transpose(x_train)
beta_hat = np.matmul(np.matmul(np.linalg.inv(np.matmul(x_transpose, x_train)),
x_transpose), y_train)
print('\n')
```

```
print('Unknown Coefficients:',beta_hat)

y_hat = np.matmul(x_train, beta_hat)
eps = y_train - y_hat
y_test_hat = np.matmul(x_test, beta_hat)
plot_fun(y_train, y_test, y_test_hat, 'LSE Method')

print('\n')
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

y_hat_OLS = model.predict(x_train)
eps_OLS = y_train - y_hat_OLS
y_test_hat_OLS = model.predict(x_test)
plot_fun(y_train, y_test, y_test_hat_OLS, 'OLS Method')

prediction_error = y_train - y_hat_OLS
forecast_error = y_test - y_test_hat_OLS
lags = 20
prediction_error_acf = cal_auto_corr(prediction_error, lags)
forecast_error_acf = cal_auto_corr(forecast_error, lags)


plt.figure(figsize=(15, 10))
plt.stem(range(-(lags-1),lags), prediction_error_acf, use_line_collection=True)
plt.xlabel('Lags')
plt.ylabel('Magnitude')
plt.title('Autocorrelation plot for Prediction Error')
plt.show()


plt.figure(figsize=(15, 10))
plt.stem(range(-(lags-1),lags), forecast_error_acf, use_line_collection=True)
plt.xlabel('Lags')
plt.ylabel('Magnitude')
plt.title('Autocorrelation plot for Forecast Error')
plt.show()

print('\n')
T = len(x_train)
K = len(x_train[0, 1:])
pred_var = (1/(T-K-1)) * (np.sum((prediction_error)**2))
print("variance of prediction error: ", pred_var)

T = len(x_test)
K = len(x_test[0, 1:])
forecast_var = (1/(T-K-1)) * (np.sum((forecast_error)**2))
print("variance of forecast error: ", forecast_var)
print('\n')


corr_coeff = round(correlation_coefficent_cal(y_test, y_test_hat_OLS),2)
print('correlation coefficient between y-test and ŷt:',corr_coeff)
print('\n')

plt.figure(figsize=(15, 10))
plt.scatter(y_test, y_test_hat_OLS, c='green', alpha=1, label='y_test vs
y_test_hat_OLS')
```

```
plt.xlabel('y_test')
plt.ylabel('y_test_hat_OLS')
plt.title("Scatter plot of y_test vs y_hat_test with correlation coefficient of
{}".format(corr_coeff))
plt.legend()
plt.show()

# Removing bore
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price','normalized-losses', 'wheel-base', 'length', 'width', 'height',
'curb-weight', 'engine-size',
         'stroke','compression-ratio', 'horsepower', 'peak-rpm', 'city-
mpg','highway-mpg']]

features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

#Removing normalized losses
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price', 'wheel-base', 'length', 'width', 'height', 'curb-weight',
'engine-size', 'stroke',
         'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg','highway-mpg']]
features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

#Removing curb-weight
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price', 'wheel-base', 'length', 'width', 'height', 'engine-size',
'stroke',
         'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg','highway-mpg']]
features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

#Removing length
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price', 'wheel-base', 'width', 'height', 'engine-size',
'stroke','compression-ratio',
         'horsepower', 'peak-rpm', 'city-mpg','highway-mpg']]
features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
```

```
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

#Removing height
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price', 'wheel-base', 'width', 'engine-size', 'stroke','compression-
ratio', 'horsepower', 'peak-rpm',
         'city-mpg','highway-mpg']]
features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

#Removing highway-mpg
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price', 'wheel-base', 'width', 'engine-size', 'stroke','compression-
ratio', 'horsepower', 'peak-rpm',
         'city-mpg']]
features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

#Removing city-mpg
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price', 'wheel-base', 'width', 'engine-size', 'stroke','compression-
ratio', 'horsepower', 'peak-rpm']]
features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
model = sm.OLS(y_train, x_train).fit()
print(model.summary())

#Removing width
df = pd.read_csv('autos.clean_corr.csv')
df = df[['price', 'wheel-base', 'engine-size', 'stroke','compression-ratio',
'horsepower', 'peak-rpm']]
features = df.drop(columns='price').to_numpy()
target = df['price'].to_numpy()
features = sm.add_constant(features)
x_train, x_test, y_train, y_test = train_test_split(features, target,
shuffle=False, test_size=0.2)
model = sm.OLS(y_train, x_train).fit()
print(model.summary())
```

```
#AUTO CORRELATION
import numpy as np

def correlation_coefficent_cal(x, y):
    result = 0
    cov_res = 0
    var_res1 = 0
    var_res2 = 0
    mean_x = np.mean(x)
    mean_y = np.mean(y)
    if len(x) == len(y):
        for i in range(0, len(x)):
            cov_res += ((x[i]-mean_x)*(y[i]-mean_y))
            var_res1 += (x[i]-mean_x)**2
            var_res2 += (y[i]-mean_y)**2
    result += cov_res/(np.sqrt(var_res1)*np.sqrt(var_res2))
    return result

#PEARSON CORRELATION COEFFICIENT
import numpy as np
def correlation_coefficent_cal(x, y):
    result = 0
    cov_res = 0
    var_res1 = 0
    var_res2 = 0
    mean_x = np.mean(x)
    mean_y = np.mean(y)
    if len(x) == len(y):
        for i in range(0, len(x)):
            cov_res += ((x[i]-mean_x)*(y[i]-mean_y))
            var_res1 += (x[i]-mean_x)**2
            var_res2 += (y[i]-mean_y)**2
    result += cov_res/(np.sqrt(var_res1)*np.sqrt(var_res2))
    return result
```

REFERENCES

https://otexts.com/fpp2/#