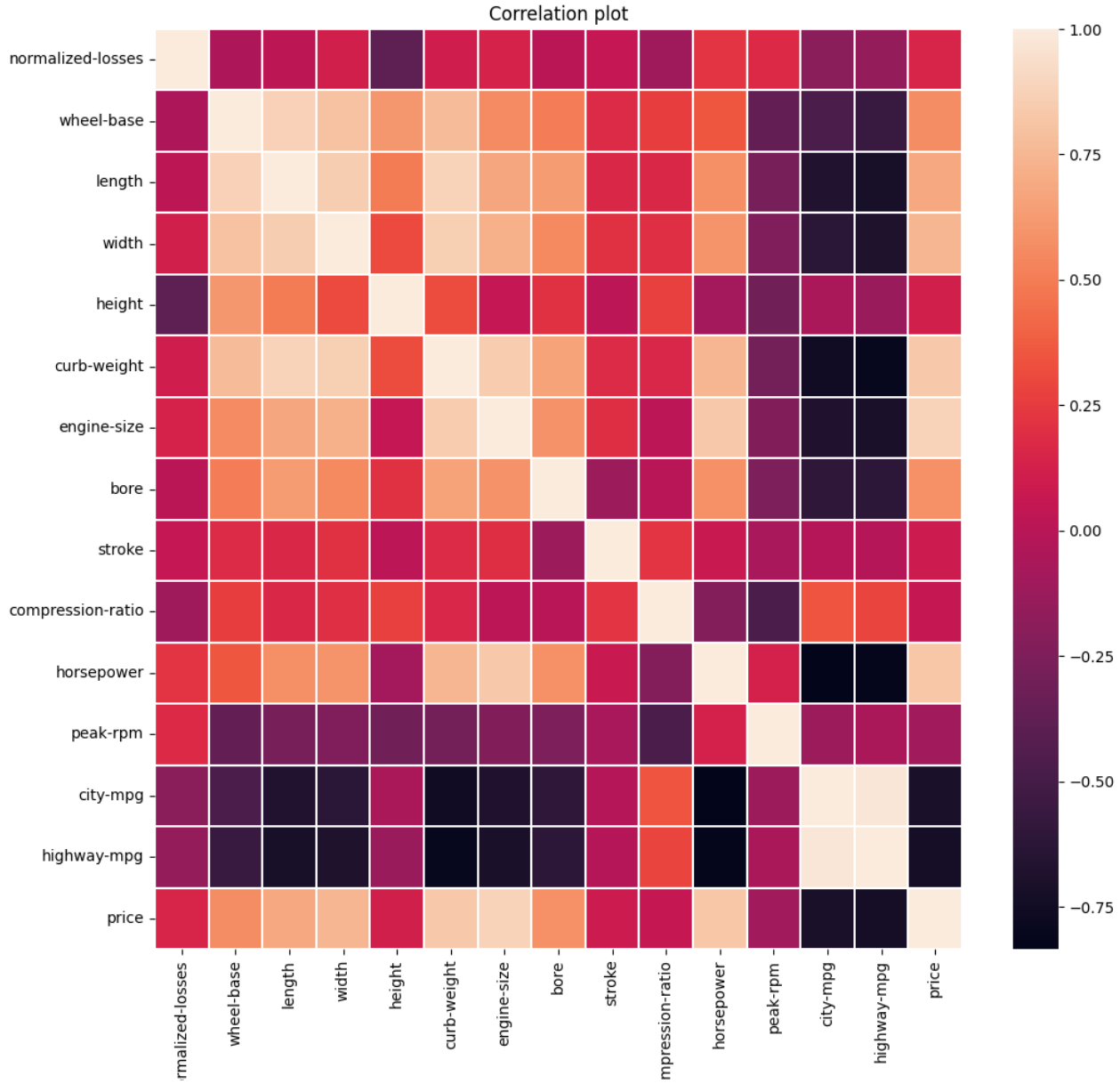
Time Series Lab3

Rajkumar Conjeevaram Mohan

Question 2 – Correlation plot



Question 3

a. SVD analysis

Original singular values

	Variable	Singular Value
0	normalized-losses	6.743324
1	wheel-base	2.521845
2	length	1.305017
3	width	0.947848
4	height	0.749920
5	curb-weight	0.481518
6	engine-size	0.380131
7	bore	0.278539
8	stroke	0.245097
9	compression-ratio	0.121098
10	horsepower	0.094341
11	peak-rpm	0.060107
12	city-mpg	0.052764
13	highway-mpg	0.018452

Normalized Singular values

	Variable	Singular Value
0	normalized-losses	1.000000
1	wheel-base	0.373977
2	length	0.193527
3	width	0.140561
4	height	0.111209
5	curb-weight	0.071407
6	engine-size	0.056371
7	bore	0.041306
8	stroke	0.036347
9	compression-ratio	0.017958
10	horsepower	0.013990
11	peak-rpm	0.008914
12	city-mpg	0.007825
13	highway-mpg	0.002736

The singular value from SVD provides information on the amount of variance present in each axis. Axis/variables with the least variance indicate the least informative variables, thus not adding much value to the overall modeling process.

b. The condition number 365.45799943149837 was obtained which means moderate to strong co-linearity exists.

c. Looking at the normalized version of the singular values, I would remove features that have variance below 5%, and we have about 6 variables that fit the criteria.

Question 6.

Both the OLS and LSE method produced identical coefficients for all features except intercept which is still approximately the same.

Question 7

For the backward stepwise elimination, I wish to remove normalized-losses, bore, curb-weight, height, length, highway-mpg, and city-mpg since these having these variables did not produce any produce additional value in the metrics such as Adjusted R-square or AIC or BIC.

I wish to retain compression-ratio, engine-size, horsepower, peak-rpm, stroke, wheel-base, and width.

Question 8

From the feature selection using the VIF method, the program has detected features – city-mpg, curb-weight, horsepower, length, wheel-base, and highway-mpg as a set of features that have potential co-linearity in them. I wish to retain features – bore, compression-ratio, engine-size, height, normalized-losses, peak-rpm, stroke, and width.

Question 9

There are intersecting features among the two sets of elimination features produced by both methods, and both final models produce somewhat similar Adjusted R-square value. However, I believe the latter method VIF is more computationally expensive than the former method – backward step wise elimination.

Question 10

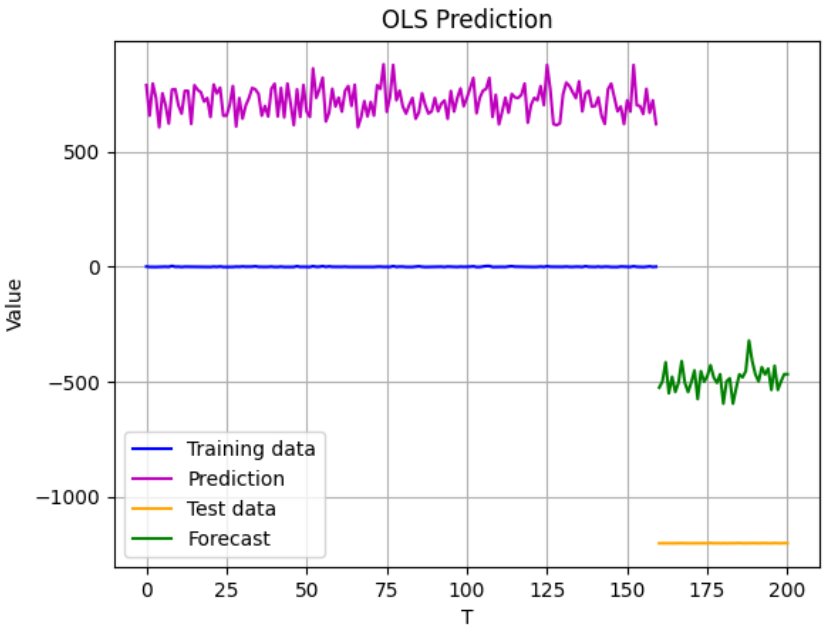
OLS Regression Results

Dep. Variable:	y	R-squared:	0.850
Model:	OLS	Adj. R-squared:	0.843
Method:	Least Squares	F-statistic:	123.0
Date:	Sat, 22 Oct 2022	Prob (F-statistic):	2.65e-59
Time:	23:37:00	Log-Likelihood:	-75.284
No. Observations:	160	AIC:	166.6
Df Residuals:	152	BIC:	191.2
Df Model:	7		
Covariance Type:	nonrobust		

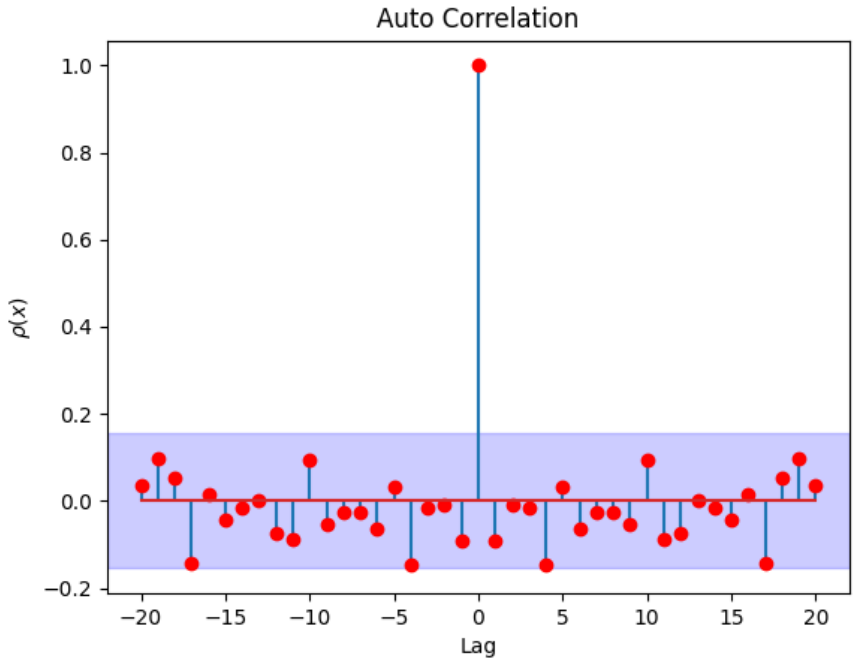
	coef	std err	t	P>|t|	[0.025	0.975]
bias_c	2.429e-17	0.031	7.73e-16	1.000	-0.062	0.062
compression-ratio	0.1380	0.039	3.552	0.001	0.061	0.215
engine-size	0.5602	0.079	7.051	0.000	0.403	0.717
horsepower	0.2718	0.075	3.631	0.000	0.124	0.420
peak-rpm	0.1177	0.044	2.693	0.008	0.031	0.204
stroke	-0.1039	0.034	-3.099	0.002	-0.170	-0.038
wheel-base	0.0804	0.057	1.415	0.159	-0.032	0.193
width	0.1376	0.067	2.039	0.043	0.004	0.271

Omnibus:	13.744	Durbin-Watson:	2.242
Prob(Omnibus):	0.001	Jarque-Bera (JB):	41.388
Skew:	0.058	Prob(JB):	1.03e-09
Kurtosis:	5.489	Cond. No.	5.98

Question 11



Question 12



Question 13

T-test

Test for Constraints

	coef	std err	t	P>|t|	[0.025	0.975]
c0	1.2018	0.087	13.735	0.000	1.029	1.375

F-test

<F test: F=188.6590579149282, p=1.9960835561351837e-28, df_denom=152, df_num=1>

I have performed t-test and f-test on the final model with constraints that coefficients equal the intercept, which is approximately 0, and both the tests produced a p value of 0 that shows all coefficients are statistically significant from the null model.

Extra

While both the training and the test set produced an MSE of 522982.84557965735, and 514486.77330305695 respectively, doing PCA decorrelated the data and produced a comparatively minimal MSE value of 0.15421711 on the training set.