SARTHAK MISHRA 22B0432
ASSIGNMENT #1
DS 203 - PROGRAMMING FOR DATA SCIENCE
19 JANUARY 2024

Train Data Statistics and Observations

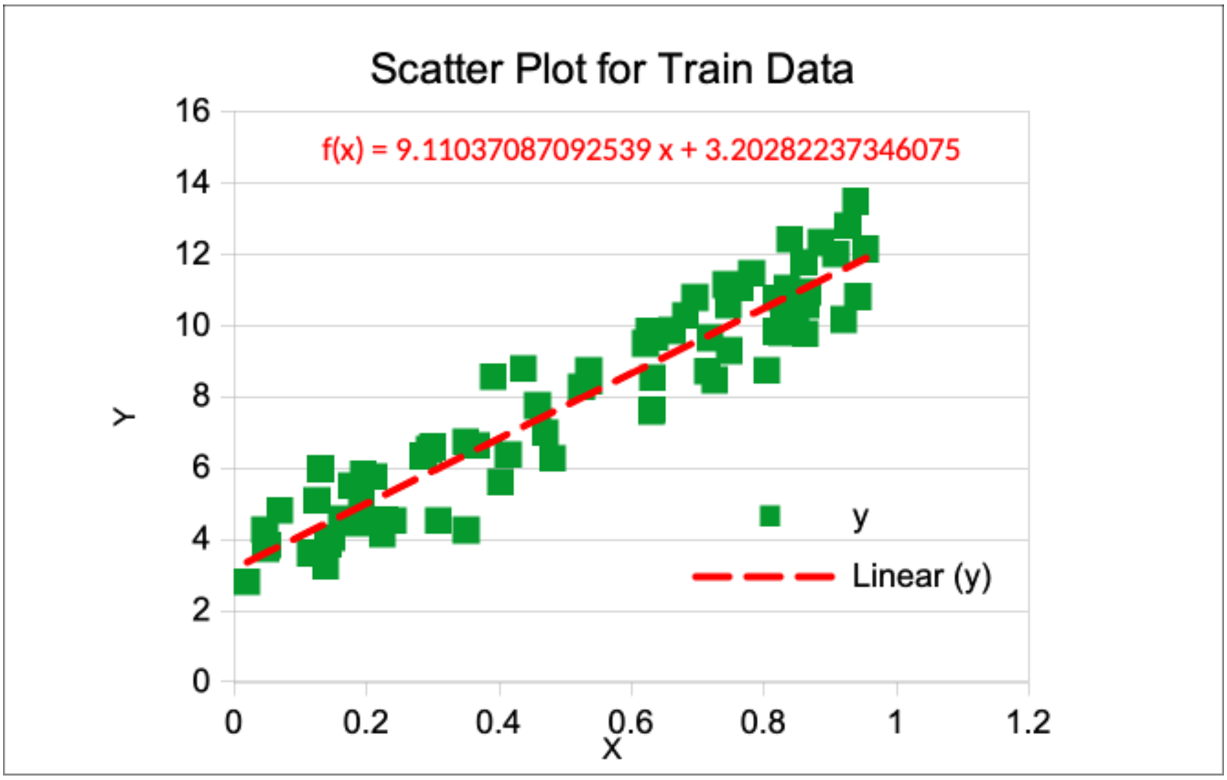

Figure 1- Scatter Plot Y v/s X

First upon Plotting the Data Points (x, y) we observe a sort of Linear Relationship between y and x with minimal visual (qualitative) variance. Thus we assume a SLR (Simple Linear Regression) as $y = ax + b$ now using the data we figure out the accurate values of a and b.

Now we use mathematically derived SLR formula for a and b.

$$a = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2},\ b = \frac{\bar{y}\overline{x^2} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2}$$ *. Where symbols have their usual meanings*

Now we need to calculate various parameters like $\overline{xy}$, $\bar{x}\,\bar{y}$, $\overline{x^2}$, $\bar{x}^2$, $\bar{y}$ to compute a and b.

y_bar	x_bar	xy_bar	x_sq_bar	x_bar_sq	x_bar_y_bar
7.951798	0.521271423143762	4.923991226	0.357225	0.271724	4.145045

Figure 2 - Computation of Parameters

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\ ,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i\ ,\bar{x}\,\bar{y} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right), \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2\ , \bar{x}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

In ***Figure 2*** We get computed parameters as follows from the Training data x and y using the above definitions and get the following results,

#1	Using Formula for *a* and *b*	
	a	9.110371
	b	3.202822

Figure 3 - a , b

Now we compare these results to the in-built Linear Regression / Regression functionality of LibreOffice Calc and get the Indicated Dashed Trend Line and the relevant equation as in ***Figure 1.***

If we look at the the data obtained in ***Figure 4*** after performing the in-built operation the values are we interested in are the ***Intercept*** and ***X1*** Coefficients which are for all our purposes very accurate and have high degree of precision.

Regression						
Regression Model	Linear					
LINEST raw output						
9.1103708709254	3.20282237					
0.3695899099068	0.22089759					
0.8914350750686	0.94213371					
607.61977772075	74					
539.33299199381	65.6835786					
Regression Statistics						
R^2	0.89143508					
Standard Error	0.94213371					
Count of X variable	1					
Observations	76					
Adjusted R^2	0.88996798					
Analysis of Variance	(ANOVA)					
	df	SS	MS	F	Significance F	
Regression	1	539.333	539.333	607.62	2.04E-37	
Residual	74	65.6836	0.88762			
Total	75	605.017				
Confidence level	0.95					
	Coefficients	Standard	t-Statistic	P-value	Lower 95	Upper 95%
Intercept	3.20282237	0.2209	14.4991	2.54E-23	2.762674	3.64297
X1	9.11037087	0.36959	24.6499	2E-37	8.373947	9.84679
X1	Predicted Y	Y	Residual			
0.2847939220642	5.79740063	6.34441	0.54701			
0.1413754792933	4.49080542	3.8735	-0.61731			
0.3492018245964	6.3841805	6.74298	0.35879			

Figure 4 - Linear Regression Stats

	A	B	E	F	G	H
1	y	x	y_cap	e	e_sqr	e_abs
2	6.344414	0.284794	5.797401	0.547014	0.299224	0.547014
3	3.8735	0.141375	4.490805	-0.61731	0.381066	0.617306
4	6.742975	0.349202	6.384181	0.358795	0.128734	0.358795
5	9.65571	0.718019	9.744241	-0.08853	0.007838	0.088531
6	9.495458	0.620738	8.85798	0.637478	0.406378	0.637478
7	4.254618	0.350113	6.39248	-2.13786	4.570456	2.137863
8	11.99909	0.908791	11.48224	0.516844	0.267127	0.516844
9	12.13666	0.95346	11.88919	0.24747	0.061241	0.24747
10	9.808154	0.826002	10.72801	-0.91985	0.846126	0.919851
11	10.80593	0.941933	11.78418	-0.97825	0.956975	0.978251
12	9.293594	0.747598	10.01372	-0.72012	0.518579	0.720124
13	11.13967	0.741946	9.962229	1.177445	1.386378	1.177445
14	8.716356	0.714306	9.710416	-0.99406	0.988155	0.99406
15	13.4821	0.93826	11.75072	1.731377	2.997667	1.731377
16	8.448424	0.535071	8.077515	0.370909	0.137573	0.370909
17	3.251278	0.137558	4.456031	-1.20475	1.451429	1.204753
18	10.57491	0.86304	11.06543	-0.49053	0.240619	0.490529
19	9.663281	0.634861	8.986646	0.676635	0.457835	0.676635
20	7.595873	0.630015	8.942496	-1.34662	1.813394	1.346623
21	4.274093	0.045782	3.619915	0.654177	0.427948	0.654177

Figure 5.1 - Calculated error parameters

Now we perform various error calculations on our data using the Linear Relationship we have obtained we apply it on the train data x and get $\hat{y} = ax + b$ and eventually get residual error as $e = y - \hat{y}$. We are also interested in some other parameters to calculate our final error metrics like $e^2 = (y - \hat{y})^2$ and $|e| = |y - \hat{y}|$.

If we take a look at ***Figure 5.1*** we have obtained exactly that for all the data points.

We define some error metrics as follows and their results obtained in ***Figure5.2***,

- **Mean Absolute error (MAE) :** $\frac{1}{n}\sum_{i=1}^{n}|e_i|$
- **Sum of Squared errors (SSE):** $\sum_{i=1}^{n} e_i^2$
- **Mean Squared error (MSE):** $\frac{1}{n}\sum_{i=1}^{n} e_i^2$
- **Root Mean Squared error (RMSE):** $\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$

MAE	0.80180299248412
SSE	65.683578564956
MSE	0.86425761269679
RMSE	0.92965456632923

Figure 5.2 - Calculated error metrics

Analysis of the Error Metrics

We can easily recognise that these values in some way or shape depends on the original residual error hence a small value of these metrics directly ensures the robustness of our model.

Now the question is how small of a value is acceptable?
Well it depends entirely on the domain of the data and the domain knowledge associated with it for example in *medical field* where each value of measurement data is highly sensitive so even a small amount of error could be highly dangerous, meanwhile on other domains like *estimating weather patterns* these values are some what acceptable due to the inherent complex nature of the environment and the variables associated with it

Superimposing $\hat{y}$ v/s x data onto ***Figure 1*** we get ,

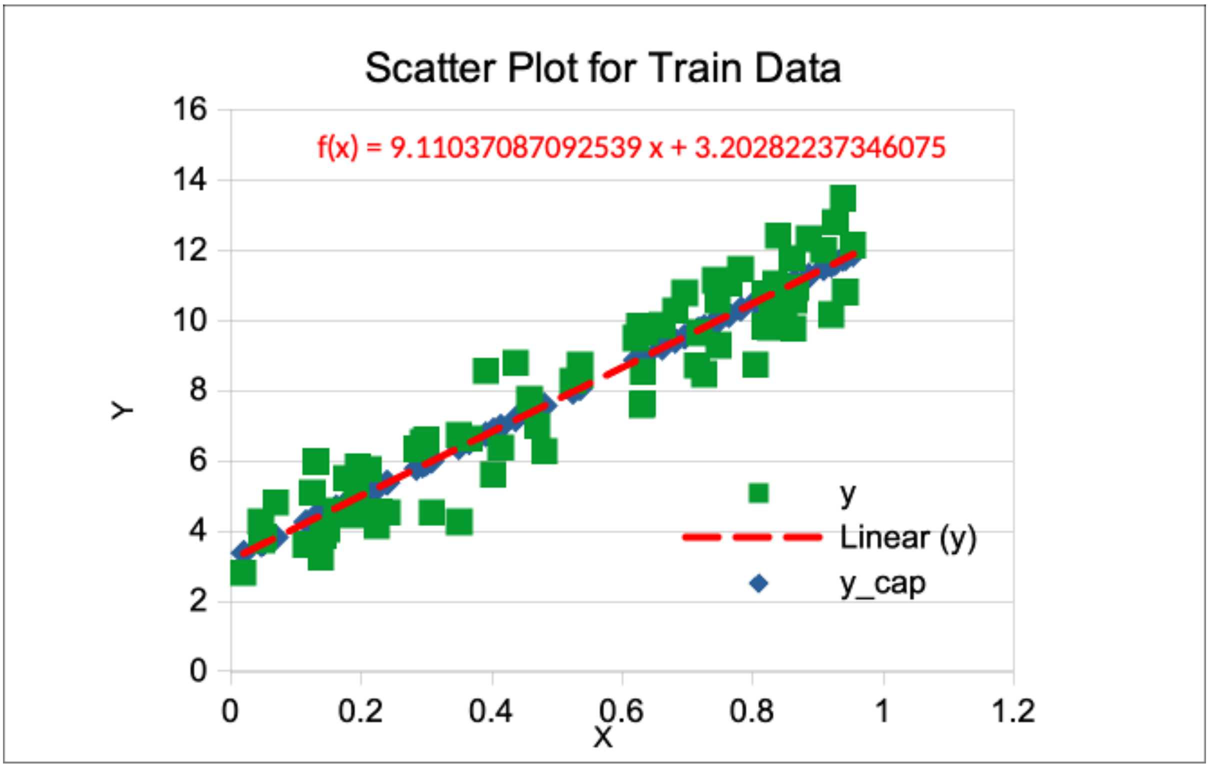

Figure 6 - Superimposed Figure 1

We clearly observe that the calculated $\hat{y}$ follows the trend line perfectly thus our computed variables a, b are accurate.

Qualitative & Quantitative Error Analysis of e v/s x

Lets Plot e v/s x and try to infer something from it,

We observe in ***Figure 7*** that the e is randomly scattered across all ranges of x with some values having outliers but with no clear pattern, for most part points are scattered uniformly about 0. However this is just a qualitative analysis.

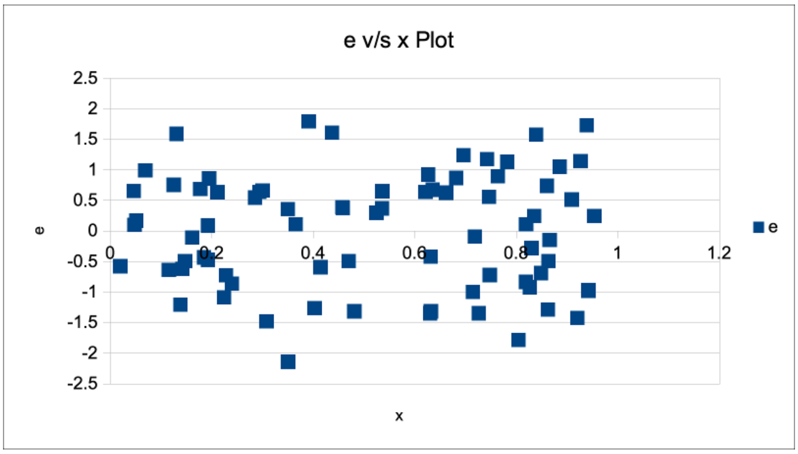

Figure 7 - e v/s x Scatter Plot

Now lets get some statistical data regarding ***Figure 7*** **,** Metrics are already computed in ***Figure 5.2*** **,** Some other useful ones are in ***Figure 8.1 and 8.2 .***

e_max	e_min	count	e_c_pos	e_c_neg
1.792635	-2.13786	76	41	35

Figure 8.2

	e
Mean	-8.5896E-16
Standard Error	0.107347263
Mode	#VALUE!
Median	0.111542815
First Quartile	-0.72048547
Third Quartile	0.667318083
Variance	0.875781048
Standard Deviation	0.935831741
Kurtosis	-0.8514819
Skewness	-0.11398741
Range	3.930497878
Minimum	-2.13786256
Maximum	1.792635317
Sum	-6.5281E-14
Count	76

Figure 8.1

Thus, we can conclude that the data sample we are working with has a very little bias and thus can be used to create a robust model.

Histogram of e for Train Data

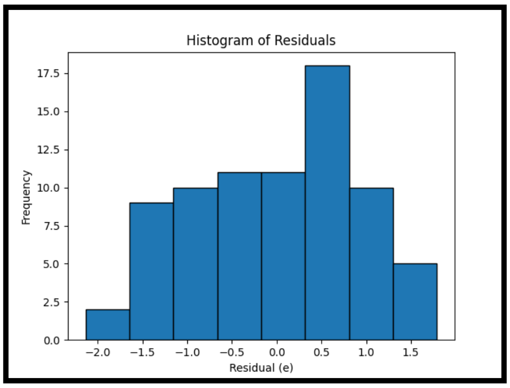

Figure 9.1 - Histogram of Residuals for Train Data

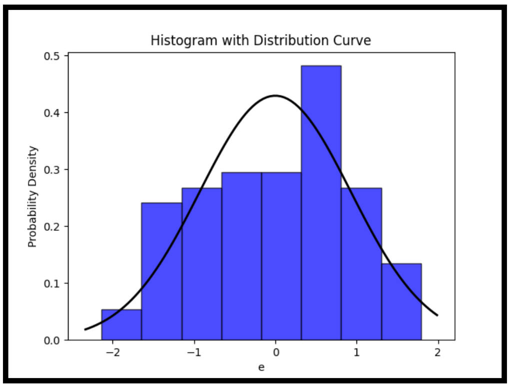

Figure 9.2 - Distribution curve

Now we wish to plot the residuals e for some interval value to analyse its distribution from e_{min} to e_{max} . We observe that $\bar{e}$ (mean) is very close to 0 slightly to the negative side and the And with negative skewness as per the ***Figure 8.2 .***

We shall also see the distribution curve superimposed onto the histogram with ***Figure 9.2*** which verifies our quantitative analysis qualitatively.

Test Data Statistics and Observations

Now we shall apply the Linear Relationship obtained on the Train Data onto the Test one to evaluate how good our model is. We get,

	A	B	C	D	E	F
1	y	x	y_cap	e	e_sq	e_abs
2	3.397539	0.0272056	3.45067549	-0.05314	0.002823	0.053137
3	7.467513	0.3962519	6.81282439	0.654689	0.428617	0.654689
4	8.315798	0.7133492	9.70169814	-1.3859	1.920718	1.3859
5	8.309891	0.6196321	8.84790092	-0.53801	0.289455	0.53801
6	12.43841	0.8906064	11.316577	1.121835	1.258513	1.121835
7	6.948992	0.2484275	5.46608944	1.482902	2.198999	1.482902
8	2.351169	0.2015937	5.03941538	-2.68825	7.226666	2.688246
9	3.522185	0.0555367	3.70878264	-0.1866	0.034819	0.186597
10	9.35813	0.5473743	8.18960564	1.168525	1.36545	1.168525
11	11.11345	0.8703586	11.1321119	-0.01866	0.000348	0.018664
12	5.033997	0.2693613	5.65680411	-0.62281	0.387889	0.622808
13	6.137698	0.3323311	6.23048176	-0.09278	0.008609	0.092784
14	3.08212	0.0461489	3.62325626	-0.54114	0.292828	0.541136
15	5.170329	0.2884631	5.83082779	-0.6605	0.436259	0.660499
16	4.062765	0.2596986	5.56877318	-1.50601	2.26806	1.506008
17	11.30703	0.7219349	9.77991694	1.527113	2.332073	1.527113
18	8.095555	0.522158	7.95987508	0.13568	0.018409	0.13568
19	9.359504	0.7703658	10.2211409	-0.86164	0.742418	0.861637
20	9.90627	0.7503464	10.038756	-0.13249	0.017553	0.132486
21	5.164621	0.2233095	5.23725435	-0.07263	0.005276	0.072633
22	4.130518	0.1210497	4.30562967	-0.17511	0.030664	0.175111
23	12.2971	0.9802748	12.1334894	0.163609	0.026768	0.163609
24	9.703482	0.9627632	11.9739519	-2.27047	5.155031	2.270469
25	4.982304	0.2922458	5.8652898	-0.88299	0.779664	0.882986

Figure 10.1

MAE	0.789311
SSE	27.22791
MSE	1.134496
RMSE	1.065127

Figure 10.2

We observe a slightly higher residual error for our test data using the Linear Fit of our train data which as expected due to our model having not seen this data but the difference in error is not high hence our model can be said to be quite robust.

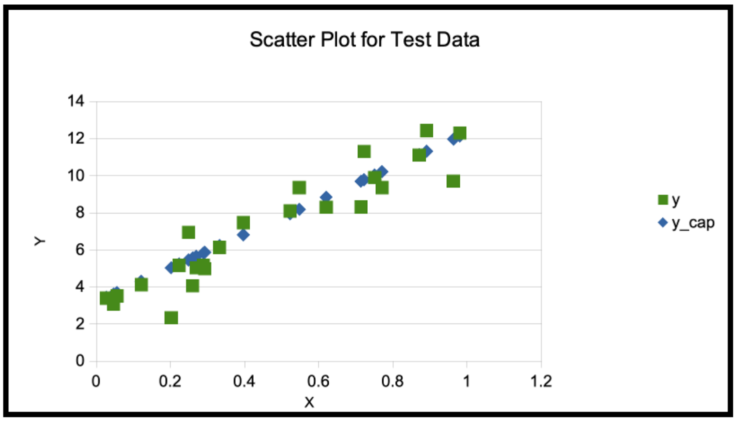

Figure 11

Upon Plotting the X-Y Scatter Plot for both the predicted $\hat{y}$ and y for the same x using the Linear Fit of train model we observe that the model fits very well with minimal outliers. Thus, Establishing a Linear Relationship between X and Y is a good way of creating a generalised model.

Now we shall see what effects does it have on the residual error v/s x plot.

Qualitative & Quantitative Error Analysis of e v/s x

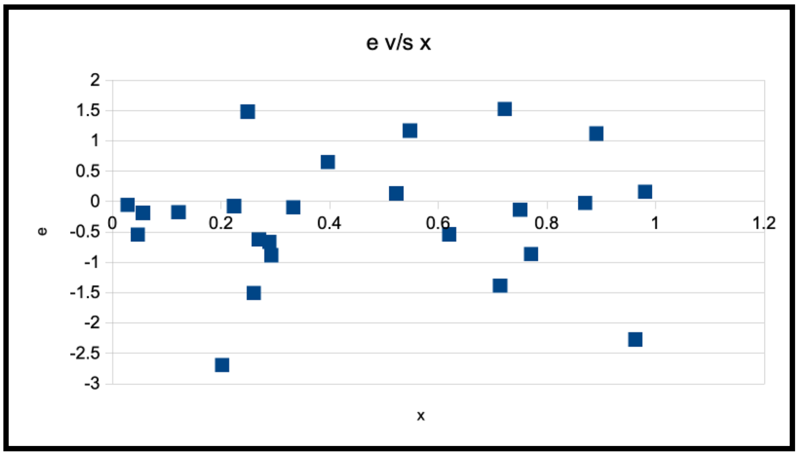

Figure 12 - e v/s x Scatter Plot

Upon observing the plot in ***Figure 12*** we see that the error for the most part behaves the same way as the train data in Figure 7 but with a higher variance from the 0 position with a few outliers but not a clear pattern hence describing a lower bias dataset.

e_max	e_min	e_count	e_c_pos	e_c_neg
1.527112623	-2.68825	24	7	17

Figure 13.1

Quantitative estimation in ***Figure 13.1 , 13.2*** tells us that the model is performing optimally its just that for a significant chunk of the test data is getting under-predicted than the actual value.

	e
Mean	-0.26811
Standard Err	0.214943
Mode	#VALUE!
Median	-0.1538
First Quartile	-0.71078
Third Quartil	0.142662
Variance	1.108811
Standard Dev	1.053001
Kurtosis	0.385225
Skewness	-0.33733
Range	4.215359
Minimum	-2.68825
Maximum	1.527113
Sum	-6.43476
Count	24

Figure 13.2

Histogram of e for Test Data

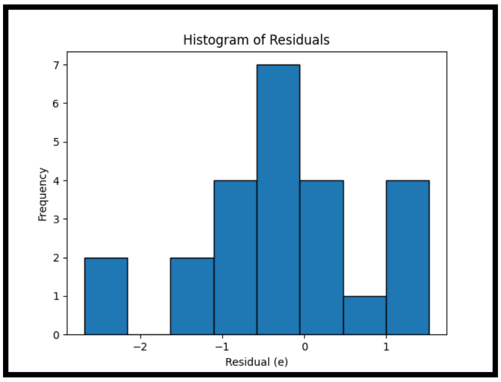

Figure 14.1 - Histogram of Residuals for Test Data

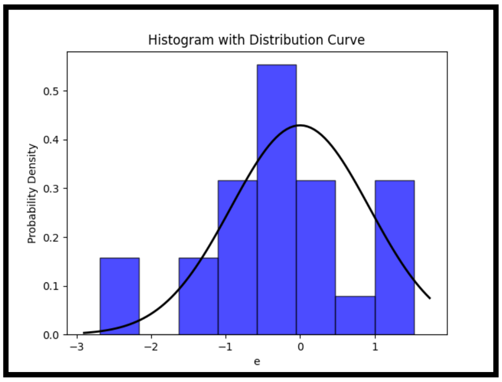

Figure 14.2 - Distribution curve

In this case too we observe that the mean is still negative by ***Figure 13.2*** and is close to zero while the skewness is also negative same case as in for Train Data indicating some biases baked into our model but if that is too small too consider or not? It entirely depends on the domain of data and its eventual applications in real world scenarios.

Conclusion

- **Comparing Train and Test Metrics:** The model's performance on the test data is generally similar to its performance on the train data, which is a positive sign. It suggests that the model is not overfitting to the training data.
- **MAE:** The MAE values are relatively low, indicating that, on average, the model's predictions are close to the actual values. Lower MAE values are desirable.
- **SSE and MSE:** Both SSE and MSE provide insights into the overall error in the model predictions. Lower values are preferred, indicating better model performance.
- **RMSE:** The RMSE values are relatively low, suggesting that the model's predictions have a small average magnitude of error. Lower RMSE values are desirable.

Some Comments :

- ***Further*** *data unless its not overfitting the model may improve the results even more.*
- ***Domain Knowledge*** *of the data will help in identifying if the data analysis is Good for practical applications or not.*
- *The* ***consistency*** *between the train and test metrics suggests that the model is generalising well to new, unseen data.*
- *The model seems to provide reasonable predictions, given the low MAE, SSE, MSE, and RMSE values.*