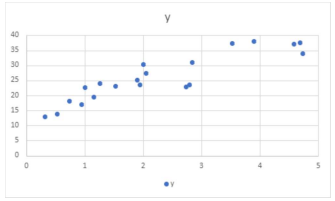
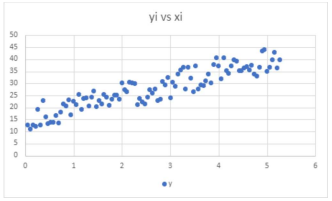
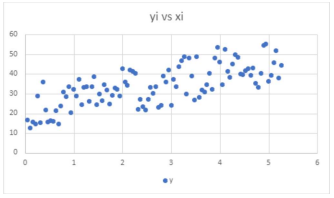
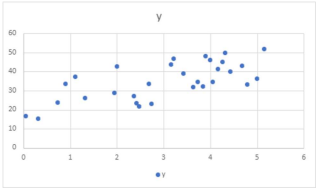
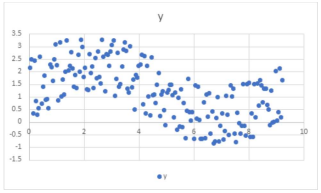
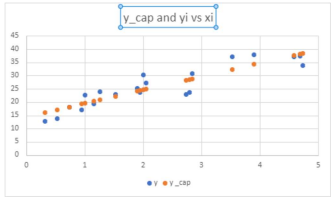
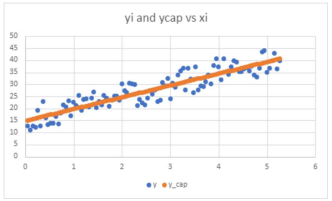
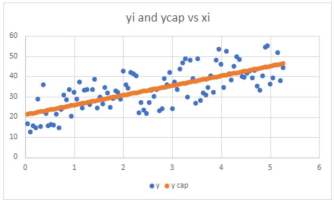
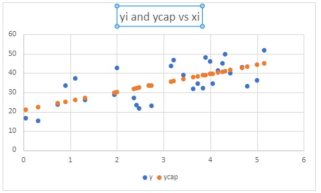
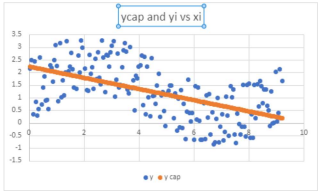
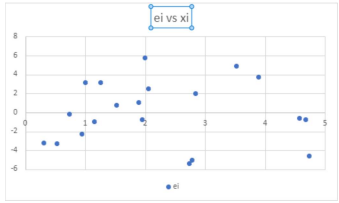
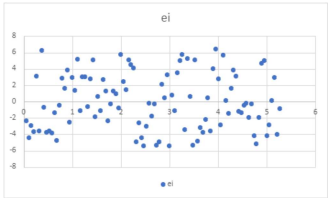
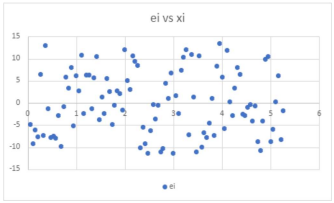
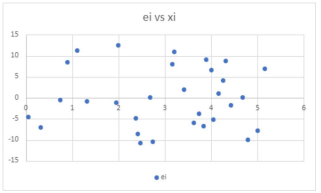
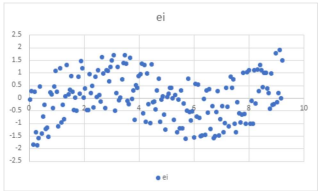
E 02 - Simple Linear Regression -

Part 1 -

The image below shows the calculation of the various statistical parameters related to linear regression of the five datasets given.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	y	x	xy	xsq	y_cap	e	yi_minus_bar	y_cap_minus_ybar	yi_minus_ybar_sq	y_cap_minus_ybar_sq	e_sq							
2	37.200879	3.526315789	131.182	12.4349	32.32404599	4.876833017	11.31033229	6.433499275	127.9236166	41.38991292	23.7835							
3	37.46053211	4.684210526	175.473	21.94183	38.19694989	-0.73641778	11.5699854	12.30640318	133.8645621	151.4475592	0.542311	sum_yi_minus_ybar_sq	SST	1185.221564		xbar	2.25789	
4	30.36517837	2	60.73036	4	24.58249084	5.782687523	4.474631653	-1.308055869	20.02232843	1.711010157	33.43947	sum_ycap_minus_ybar_sq	SSR	977.9908868		ybar	25.8905	
5	16.97950771	0.947368421	16.08585	0.897507	19.2434873	-2.26397959	-8.911039003	-6.647059417	79.40661611	44.18339889	5.125604	sum_ei_sq	SSE	207.2306769		xybar	68.0991	
6	30.8497319	2.842105263	87.67819	8.077562	28.85369368	1.996038219	4.959185188	2.963146969	24.59351773	8.78023996	3.984169	R_sq	SSR/SST	0.825154483		xsqbar	6.99889	
7	37.07110993	4.578947368	169.7467	20.96676	37.66304954	-0.5919396	11.18056322	11.77250282	125.0049939	138.5918227	0.350392							
8	19.35650899	1.157894737	22.4128	1.34072	20.31128801	-0.95477901	-6.53403772	-5.579258707	42.69364893	31.12812772	0.911603					a	5.07205	
9	22.96795621	1.526315789	35.05635	2.32964	22.17993925	0.788016968	-2.922590498	-3.710607466	8.541535219	13.76860776	0.620971					b	14.4384	
10	37.89454117	3.894736842	147.5893	15.16898	34.19269723	3.701843945	12.00399446	8.302150517	144.095883	68.9257032	13.70365							
11	18.01077771	0.736842105	13.2711	0.542936	18.17568659	-0.16490887	-7.879769	-7.714860127	62.09075949	59.51906677	0.027195							
12	27.31273343	2.052631579	56.06298	4.213296	24.84944102	2.463292409	1.422186717	-1.041105692	2.022615058	1.083901062	6.067809							
13	33.87169482	4.736842105	160.4449	22.43767	38.46390007	-4.59220525	7.981148103	12.57335336	63.69872504	158.0892146	21.08835							
14	22.69067344	1	22.69067	1	19.51043747	3.18023597	-3.199873269	-6.38010924	10.23918894	40.70579391	10.1139							
15	23.55654774	2.789473684	65.71037	7.781163	28.5867435	-5.03019577	-2.333998976	2.696196792	5.447551219	7.269477139	25.30287							
16	25.10304641	1.894736842	47.56367	3.590028	24.04859049	1.054455924	-0.7875003	-1.841956224	0.620156723	3.392802731	1.111877							
17	23.56896169	1.947368421	45.89745	3.792244	24.31554067	-0.74657897	-2.321585019	-1.575006047	5.389757	2.480644047	0.55738							
18	12.80130715	0.315789474	4.042518	0.099723	16.04008517	-3.23877801	-13.08923956	-9.850461546	171.3281923	97.03159266	10.48968							
19	24.01330076	1.263157895	30.33259	1.595568	20.84518836	3.168112402	-1.87724595	-5.045358353	3.524052358	25.45564091	10.03694							
20	13.81080111	0.526315789	7.268843	0.277008	17.10788588	-3.29708477	-12.0797456	-8.782660836	145.9202538	77.13513136	10.87077							
21	22.92514458	2.736842105	62.7425	7.490305	28.31979333	-5.39464875	-2.965402133	2.429246614	8.793609811	5.901239113	29.10224							
22																		
23	SUMMARY OUTPUT																	
24																		
25	*Regression Statistics*																	
26	Multiple R	0.908380142																
27	R Square	0.825154483																
28	Adjusted R Sq	0.815440843																
29	Standard Erro	3.39305399																
30	Observations	20																
31																		
32	ANOVA																	
33		*df*	*SS*	*MS*	*F*	*Significance F*												
34	Regression	1	977.9909	977.9909	84.94802135	3.08604E-08												
35	Residual	18	207.2307	11.51282														
36	Total	19	1185.222															
37																		
38		Coefficients	Standard E	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%									
39	Intercept	14.4383841	1.455867	9.917376	1.01457E-08	11.37972016	17.49704804	11.37972016	17.49704804									
40	X Variable 1	5.07205337	0.55031	9.216725	3.08604E-08	3.915895558	6.228211183	3.915895558	6.228211183									
41																		

Part 2 : Data Set v/s Regression Outcomes -

Data Set/ Parameters	Data 1	Data 2	Data 3	Data 4	Data 5
Scatter Plot y_i **vs** x_i					
Scatter Plot y_i **vs** x_i $\hat{y}_i$ **vs** x_i					
Scatter Plot e_i **vs** x_i					

Data Set/ Parameters	Data 1	Data 2	Data 3	Data 4	Data 5
Data Size (n)	20	100	100	30	180
Variance(y)	62.38008	67.50414	103.3773	98.113	1.15479
Standard Deviation(y)	13.999028575524	8.21609031109079	10.1674628103574	9.90511484032366	1.07461155772679
Corr(x ,y)	0.90838	0.911263	0.720646	0.689041	-0.55084
Coeff ‘a’	5.07205337	4.903347591	4.798640815	4.693246374	-0.22152735
Coeff ‘b’	14.4383841	14.81033235	21.10485906	20.79759598	2.232599232
p - value ‘a’	3.09E-08	1.54E-39	2.84E-17	2.55E-05	1.13E-15
p - value ‘b’	1.01E-08	4.6E-39	1.08E-26	3.15E-07	6.01E-38
R^2	0.825154483	0.830401026	0.51933107	0.474777689	0.303426178
1 - R^2	0.174845517	0.169598974	0.48066893	0.525222311	0.696573822
F - value	84.94802	479.8337	105.8825	25.31076	77.53645
Significance F	3.09E-08	1.54E-39	2.8446E-17	2.55E-05	1.13E-15
SSE	207.2307	1133.415	4919.334	1494.377	143.9867
MSE (SSE/n)	10.361535	11.33415	49.19334	49.8125666666667	0.799926111111111

Q1. Data sets Data1 and Data2 are samples of different sizes randomly drawn from the same population. Care has been taken to ensure that the samples are true representative of the population. What can you say about the impact of sample size on regression quality? Explain by comparing the following: regression coefficients and their p-values, R2, MSE, F-value, Significance F. Which of these is majorly impacted, and what can you conclude from this? Can you explain why?

Ans:- The p - value and significance F of data set 2 are quite less than that of data set 1 and also its F - value is greater than that of data 1.
The other factors such as regression coefficients ,R2,MSE are quite similar.
Hence , regression quality of set 2 is more accurate than data set 1 (due to p - value and significance F).

Q2. Compare the data sets Data2 and Data3 by focusing on the plots, variances, and correlation coefficients. What differences do you observe? What do they indicate?

Ans:- By focusing on plots we can see that data set 2 is more close to linearity while data set 3 is more scattered vertically.
This means the variance of data set 3 is greater than that of data set 2 which in turn means that data set 3 is less representative of the whole population while data set 2 is more reliable for predicting .

Q3. Compare the regression outputs from Data2 and Data3. Focus on the following: correlation coefficients, regression coefficients and their p-values, R 2 , MSE, F-value, Significance F. What major differences do you observe? What do they indicate about the relative quality of these regressions?

Ans:- The p-value, Variance, Significance F, MSE of data 2 are much smaller than that of data 3 while the F-value, R 2 are much greater than data 3 which implies data 2 has a better linear regression quality.

Q4. Compare the data sets Data1 and Data4, and their regression outputs. What are your observations about the relative quality of the data sets themselves? How do the two regression outcomes compare? What conclusions can you make from this analysis – particularly related to data quality and regression quality?

Ans:- Data 4 has the lowest quality data and the poorest regression quality, with a low R^2, p-value and a higher amount of F-value , variance , SSE and MSE than data 1 resulting in more errors.
We can conclude that data 1 is much reliable than data 4.

Q5. Consider Data5, in this case you have fitted a linear model over non-linear data. What is the impact on the regression metrics and how do they correlate with the data quality – variances, correlation coefficient, regression coefficients and their p-values, R 2 , MSE, F-value, and Significance F? In your analysis also compare these metrics with those observed in the previous 4 data sets – and state your observations and conclusions.

Ans:- Data5 is different from all the other data sets because of its non-linear relationship, negative correlation, and inverse relationship between x and y. This results in the low R^2 value, weak F-value, and low MSE, implying that the linear model doesn't recognize the underlying pattern well. In contrast, all the other data sets have positive correlations, implying a better fit for linear models. Their higher R^2 values, higher MSE values and higher F-values suggest better linear fits.
Hence, fitting a linear model to non-linear data (Data5) leads to poor regression metrics and a weak fit.

Q6. Across the data sets, what relationship do you observe between the correlation coefficients and the quality of regressions? Can you detect any mathematical relationship between the correlation coefficient and R2?

Ans:- Generally, a greater correlation coefficient indicates a stronger linear relationship between variables. A strong linear relationship (high correlation) generally results in a higher R2 value because more of the variance in the dependent variable is explained by the independent variable. So there is a positive mathematical relationship between correlation and R2, where higher correlation coefficients are associated with higher R2 values. Data2 and Data1 have the highest correlation coefficients (0.911 and 0.908, respectively) and also the highest R2 values (0.83 and 0.825, respectively). Data4 and Data5 have lower correlation coefficients (0.689 and -0.551, respectively) and also lower R2 values (0.475 and 0.303, respectively).

Q7. Error plots and error metrics are an important consideration while assessing regression quality. Observe all the error plots. Qualitatively, in what way does the error plot for Data5 differ from the other error plots? Why do such error plots indicate an incorrect regression?

Ans:- Error plot of 5 follows a pattern and is clustered while the error plots of other data sets are scattered and stochastic . As errors are clustered around , it suggests that the model either overestimates or underestimates certain regions of the data and is not understanding the data pattern.

Q8. Based on all the observations, analyses, and conclusions you have made so far, now list down the specific criteria and steps you will take to decide upon the quality of a Linear Regression model, and to decide whether you will use the regression model for prediction!

Ans:- We can decide whether a model is good by checking whether the error is clustered in some region (bad model) or if it is random and does not follow any pattern (good model). We can also check for the missing values in the data and ensure that even for their absence the data set provided is sufficient.
Another way to test a model is to check the values of statistical coefficients and numbers.