Predicting restaurant tips using predictive analytics on Excel.

In this Project:

1. I selected all the data and filtered if there is any blank cell or not

2. Checked and removed duplicates

3. Found Independent and dependent variables

Independent Variables = Sex, Smoker, Day, Time, size, total bill

Depended variable = Tips

4. Used multiple regression analysis to obtain data model as shown in screenshots and excel files

5. Converted categorical values into numerical using IF condition

6. Then using predictive analysis multiple regression formula, found the predictive tips alongside actual ones

7. Then found errors in next column

8. Finally calculated RMSE

Encode the categorical variables to numeric values using IF conditions.

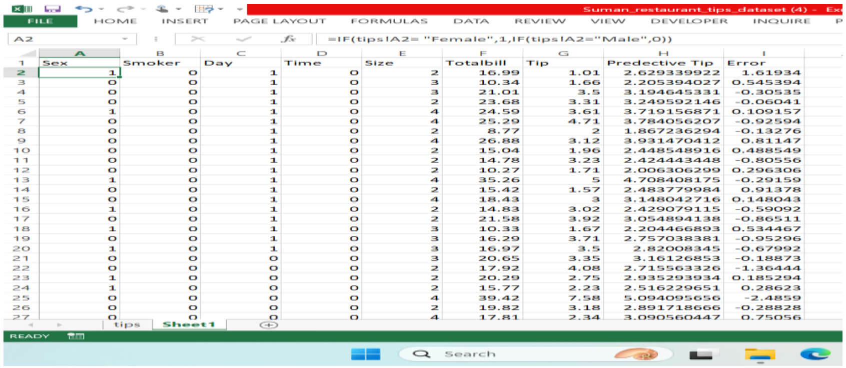
Problem 1:

Differentiate Gender by giving them the different values

Process:

Converting categorical variables to numeric values cab be done by adding dummy values to each column.

Output:



=IF(tips!A2= "Female",1,IF(tips!A2="Male",0))

	A	B	C	D	E	F	G	H	I
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip	Predective Tip	Error
2	1	0	1	0	2	16.99	1.01	2.629339922	1.61934
3	0	0	1	0	3	10.34	1.66	2.205394027	0.545394
4	0	0	1	0	3	21.01	3.5	3.194645331	-0.30535
5	0	0	1	0	2	23.68	3.31	3.249592146	-0.06041
6	1	0	1	0	4	24.59	3.61	3.719156871	0.109157
7	0	0	1	0	4	25.29	4.71	3.784056207	-0.92594
8	0	0	1	0	2	8.77	2	1.867236294	-0.13276
9	0	0	1	0	4	26.88	3.12	3.931470412	0.81147
10	0	0	1	0	2	15.04	1.96	2.448548916	0.488549
11	0	0	1	0	2	14.78	3.23	2.424443448	-0.80556
12	0	0	1	0	2	10.27	1.71	2.006306299	0.296306
13	1	0	1	0	4	35.26	5	4.708408175	-0.29159
14	0	0	1	0	2	15.42	1.57	2.483779984	0.91378
15	0	0	1	0	4	18.43	3	3.148042716	0.148043
16	1	0	1	0	2	14.83	3.02	2.429079115	-0.59092
17	0	0	1	0	2	21.58	3.92	3.054894138	-0.86511
18	1	0	1	0	3	10.33	1.67	2.204466893	0.534467
19	0	0	1	0	3	16.29	3.71	2.757038381	-0.95296
20	1	0	1	0	3	16.97	3.5	2.82008345	-0.67992
21	0	0	0	0	3	20.65	3.35	3.16126853	-0.18873
22	0	0	0	0	2	17.92	4.08	2.715563326	-1.36444
23	1	0	0	0	2	20.29	2.75	2.935293934	0.185294
24	1	0	0	0	2	15.77	2.23	2.516229651	0.28623
25	0	0	0	0	4	39.42	7.58	5.094095656	-2.4859
26	0	0	0	0	2	19.82	3.18	2.891718666	-0.28828
27	0	0	0	0	4	17.81	2.34	3.090560447	0.75056

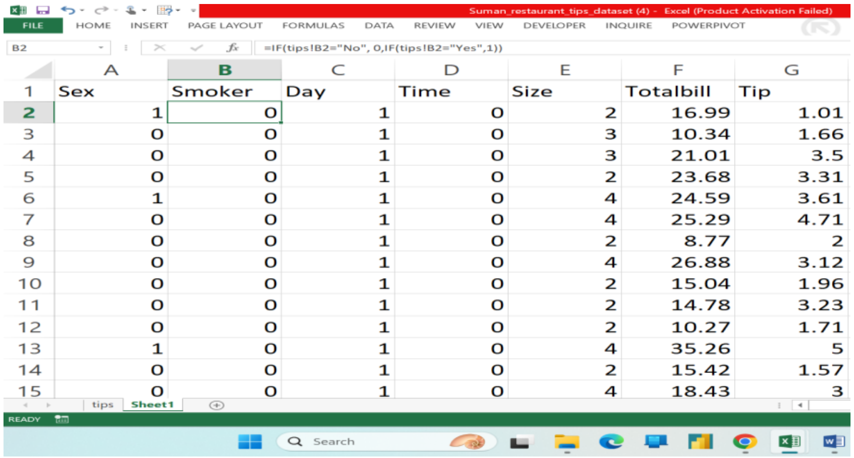
Problem 2:

Know whether the smoker is convertible or not? If convertible change them.

Process:

They are convertible because there is no varchar data type in the values of the column smoker.

Output:



Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)

B2 =IF(tips!B2="No", 0,IF(tips!B2="Yes",1))

	A	B	C	D	E	F	G
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip
2	1	0	1	0	2	16.99	1.01
3	0	0	1	0	3	10.34	1.66
4	0	0	1	0	3	21.01	3.5
5	0	0	1	0	2	23.68	3.31
6	1	0	1	0	4	24.59	3.61
7	0	0	1	0	4	25.29	4.71
8	0	0	1	0	2	8.77	2
9	0	0	1	0	4	26.88	3.12
10	0	0	1	0	2	15.04	1.96
11	0	0	1	0	2	14.78	3.23
12	0	0	1	0	2	10.27	1.71
13	1	0	1	0	4	35.26	5
14	0	0	1	0	2	15.42	1.57
15	0	0	1	0	4	18.43	3

Problem 3:

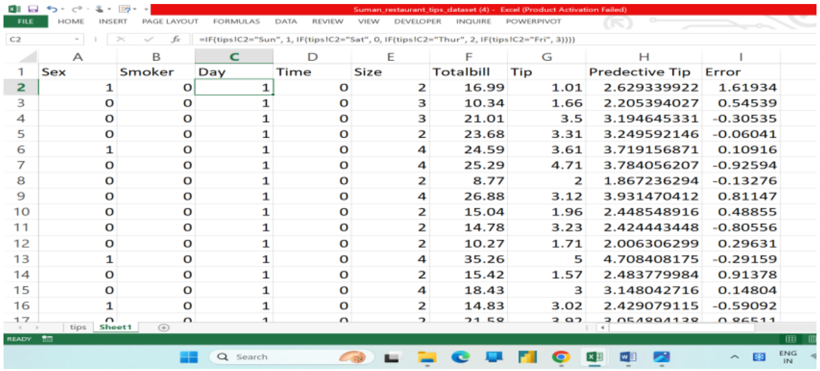
Analyses the days in the week based that are crowded.

Process:

According to the data the 3 week days are Thursday, Friday, Sunday and Saturday.

The columns are created and also variables are created based on the DAY column.

Output:



Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)

C2 =IF(tips!C2="Sun", 1, IF(tips!C2="Sat", 0, IF(tips!C2="Thur", 2, IF(tips!C2="Fri", 3))))

	A	B	C	D	E	F	G	H	I
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip	Predective Tip	Error
2	1	0	1	0	2	16.99	1.01	2.629339922	1.61934
3	0	0	1	0	3	10.34	1.66	2.205394027	0.54539
4	0	0	1	0	3	21.01	3.5	3.194645331	-0.30535
5	0	0	1	0	2	23.68	3.31	3.249592146	-0.06041
6	1	0	1	0	4	24.59	3.61	3.719156871	0.10916
7	0	0	1	0	4	25.29	4.71	3.784056207	-0.92594
8	0	0	1	0	2	8.77	2	1.867236294	-0.13276
9	0	0	1	0	4	26.88	3.12	3.931470412	0.81147
10	0	0	1	0	2	15.04	1.96	2.448548916	0.48855
11	0	0	1	0	2	14.78	3.23	2.424443448	-0.80556
12	0	0	1	0	2	10.27	1.71	2.006306299	0.29631
13	1	0	1	0	4	35.26	5	4.708408175	-0.29159
14	0	0	1	0	2	15.42	1.57	2.483779984	0.91378
15	0	0	1	0	4	18.43	3	3.148042716	0.14804
16	1	0	1	0	2	14.83	3.02	2.429079115	-0.59092
17	0	0	1	0	2	21.58	3.92	3.054894138	-0.86511

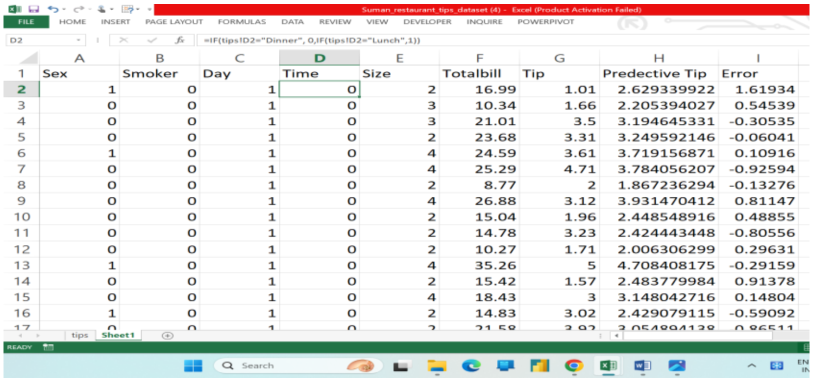
Problem 4:

Calculate the dinner column by converting categorical variables in to numerical to know the bill size

Process:

As predicted tips was calculated by dinner size, the column was created.

Output:



=IF(tips!D2="Dinner", 0,IF(tips!D2="Lunch",1))

	A	B	C	D	E	F	G	H	I
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip	Predective Tip	Error
2	1	0	1	0	2	16.99	1.01	2.629339922	1.61934
3	0	0	1	0	3	10.34	1.66	2.205394027	0.54539
4	0	0	1	0	3	21.01	3.5	3.194645331	-0.30535
5	0	0	1	0	2	23.68	3.31	3.249592146	-0.06041
6	1	0	1	0	4	24.59	3.61	3.719156871	0.10916
7	0	0	1	0	4	25.29	4.71	3.784056207	-0.92594
8	0	0	1	0	2	8.77	2	1.867236294	-0.13276
9	0	0	1	0	4	26.88	3.12	3.931470412	0.81147
10	0	0	1	0	2	15.04	1.96	2.448548916	0.48855
11	0	0	1	0	2	14.78	3.23	2.424443448	-0.80556
12	0	0	1	0	2	10.27	1.71	2.006306299	0.29631
13	1	0	1	0	4	35.26	5	4.708408175	-0.29159
14	0	0	1	0	2	15.42	1.57	2.483779984	0.91378
15	0	0	1	0	4	18.43	3	3.148042716	0.14804
16	1	0	1	0	2	14.83	3.02	2.429079115	-0.59092
17	0	0	1	0	2	21.58	3.92	3.054894138	0.86511

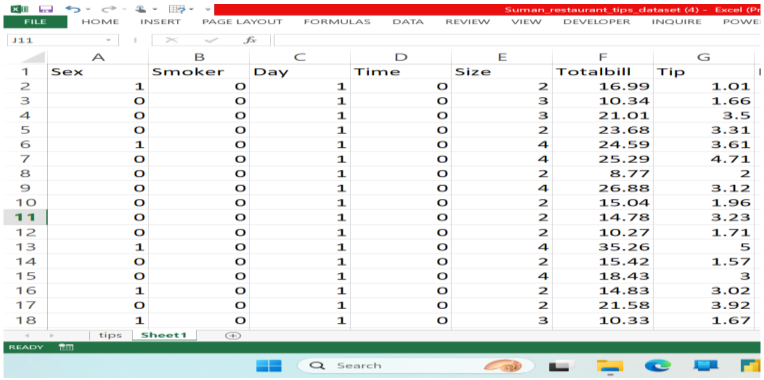
Problem 5:

Identify an appropriate model with the dataset.

Process:

To identify the right model the complete dataset should be cleaned and being formatted.

Output:



	A	B	C	D	E	F	G
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip
2	1	0	1	0	2	16.99	1.01
3	0	0	1	0	3	10.34	1.66
4	0	0	1	0	3	21.01	3.5
5	0	0	1	0	2	23.68	3.31
6	1	0	1	0	4	24.59	3.61
7	0	0	1	0	4	25.29	4.71
8	0	0	1	0	2	8.77	2
9	0	0	1	0	4	26.88	3.12
10	0	0	1	0	2	15.04	1.96
11	0	0	1	0	2	14.78	3.23
12	0	0	1	0	2	10.27	1.71
13	1	0	1	0	4	35.26	5
14	0	0	1	0	2	15.42	1.57
15	0	0	1	0	4	18.43	3
16	1	0	1	0	2	14.83	3.02
17	0	0	1	0	2	21.58	3.92
18	1	0	1	0	3	10.33	1.67

As we had completed the process of data cleaning.

The desired model would be REGRESSION according to the data obtained.

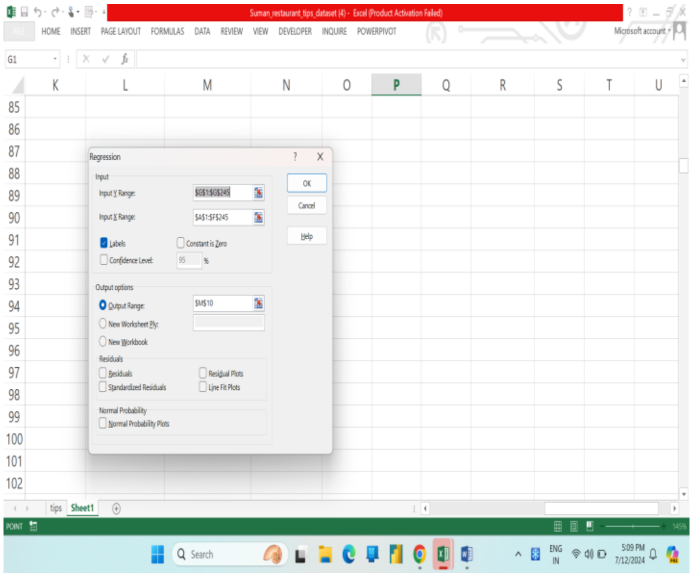
Problem 6:

Build an appropriate model with the new table.

Process:

We have created the Regression Model

Output:


Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER INQUIRE POWERPIVOT
Microsoft account
G1
tips Sheet1
POINT

Regression

Input

Input Y Range: $G$1:$G$245

Input X Range: $A$1:$F$245

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $M$10

○ New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

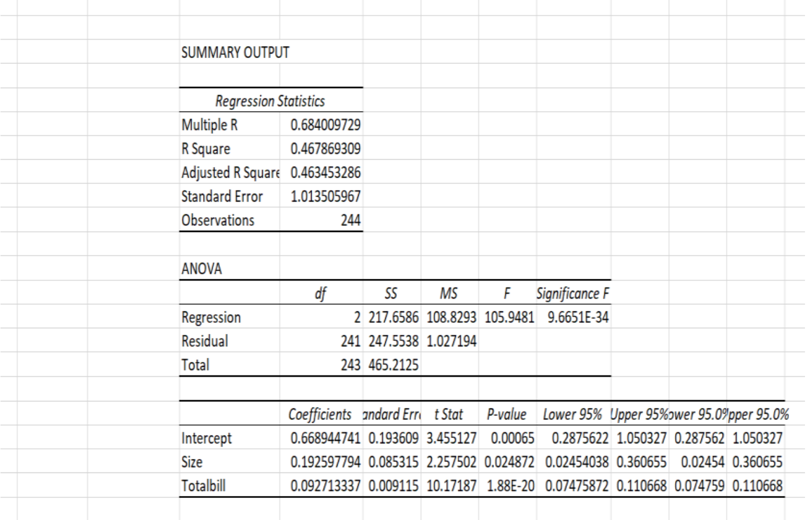
Normal Probability

☐ Normal Probability Plots

OK Cancel Help

After applying the Regression Model.

Output:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.684009729
R Square	0.467869309
Adjusted R Square	0.463453286
Standard Error	1.013505967
Observations	244

ANOVA

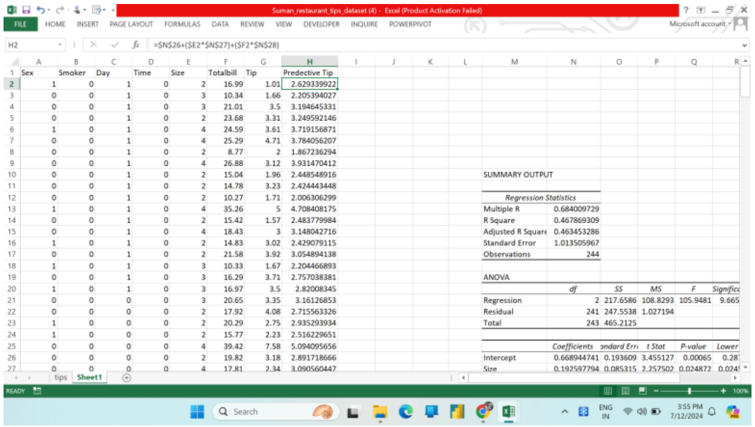
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	217.6586	108.8293	105.9481	9.6651E-34
Residual	241	247.5538	1.027194		
Total	243	465.2125			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.668944741	0.193609	3.455127	0.00065	0.2875622	1.050327	0.287562	1.050327
Size	0.192597794	0.085315	2.257502	0.024872	0.02454038	0.360655	0.02454	0.360655
Totalbill	0.092713337	0.009115	10.17187	1.88E-20	0.07475872	0.110668	0.074759	0.110668

Problem 7:

Calculate the Predicted Tips.

Output:


Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER INQUIRE POWERPIVOT
Microsoft account
tips Sheet1
READY
100%
Search
ENG IN
3:55 PM 7/12/2024

H2 =$N$26+($E2*$N$27)+($F2*$N$28)

	A	B	C	D	E	F	G	H
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip	Predective Tip
2	1	0	1	0	2	16.99	1.01	2.629339922
3	0	0	1	0	3	10.34	1.66	2.205394027
4	0	0	1	0	3	21.01	3.5	3.194645331
5	0	0	1	0	2	23.68	3.31	3.249592146
6	1	0	1	0	4	24.59	3.61	3.719156871
7	0	0	1	0	4	25.29	4.71	3.784056207
8	0	0	1	0	2	8.77	2	1.867236294
9	0	0	1	0	4	26.88	3.12	3.931470412
10	0	0	1	0	2	15.04	1.96	2.448548916
11	0	0	1	0	2	14.78	3.23	2.424443448
12	0	0	1	0	2	10.27	1.71	2.006306299
13	1	0	1	0	4	35.26	5	4.708408175
14	0	0	1	0	2	15.42	1.57	2.483779984
15	0	0	1	0	4	18.43	3	3.148042716
16	1	0	1	0	2	14.83	3.02	2.429079115
17	0	0	1	0	2	21.58	3.92	3.054894138
18	1	0	1	0	3	10.33	1.67	2.204466893
19	0	0	1	0	3	16.29	3.71	2.757038381
20	1	0	1	0	3	16.97	3.5	2.82008345
21	0	0	0	0	3	20.65	3.35	3.16126853
22	0	0	0	0	2	17.92	4.08	2.715563326
23	1	0	0	0	2	20.29	2.75	2.935293934
24	1	0	0	0	2	15.77	2.23	2.516229651
25	0	0	0	0	4	39.42	7.58	5.094095656
26	0	0	0	0	2	19.82	3.18	2.891718666
27	0	0	0	0	4	17.81	2.34	3.090560447

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.684009729
R Square	0.467869309
Adjusted R Square	0.463453286
Standard Error	1.013505967
Observations	244

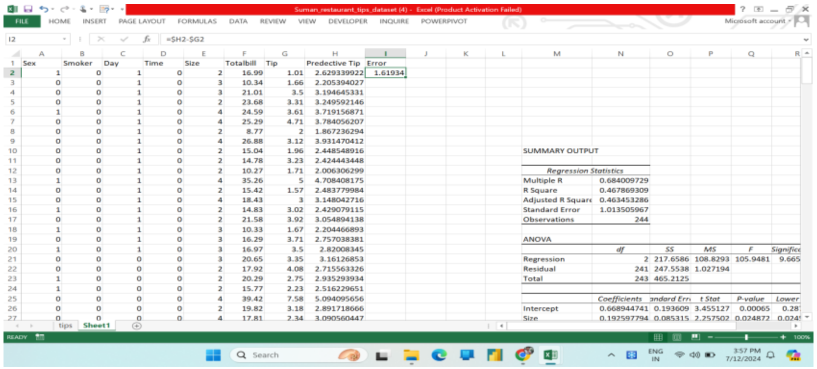
ANOVA

	df	SS	MS	F	Significa
Regression	2	217.6586	108.8293	105.9481	9.665
Residual	241	247.5538	1.027194		
Total	243	465.2125			

	Coefficients	andard Err	t Stat	P-value	Lower
Intercept	0.668944741	0.193609	3.455127	0.00065	0.28
Size	0.192597794	0.085315	2.257502	0.024872	0.024

Problem 8:

Find the Error in the tips according to the Predicted tips.


Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER INQUIRE POWERPIVOT
Microsoft account
tips Sheet1
READY
100%
Search
ENG IN
3:57 PM 7/12/2024

I2 =$H2-$G2

	A	B	C	D	E	F	G	H	I
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip	Predective Tip	Error
2	1	0	1	0	2	16.99	1.01	2.629339922	1.61934
3	0	0	1	0	3	10.34	1.66	2.205394027	
4	0	0	1	0	3	21.01	3.5	3.194645331	
5	0	0	1	0	2	23.68	3.31	3.249592146	
6	1	0	1	0	4	24.59	3.61	3.719156871	
7	0	0	1	0	4	25.29	4.71	3.784056207	
8	0	0	1	0	2	8.77	2	1.867236294	
9	0	0	1	0	4	26.88	3.12	3.931470412	
10	0	0	1	0	2	15.04	1.96	2.448548916	
11	0	0	1	0	2	14.78	3.23	2.424443448	
12	0	0	1	0	2	10.27	1.71	2.006306299	
13	1	0	1	0	4	35.26	5	4.708408175	
14	0	0	1	0	2	15.42	1.57	2.483779984	
15	0	0	1	0	4	18.43	3	3.148042716	
16	1	0	1	0	2	14.83	3.02	2.429079115	
17	0	0	1	0	2	21.58	3.92	3.054894138	
18	1	0	1	0	3	10.33	1.67	2.204466893	
19	0	0	1	0	3	16.29	3.71	2.757038381	
20	1	0	1	0	3	16.97	3.5	2.82008345	
21	0	0	0	0	3	20.65	3.35	3.16126853	
22	0	0	0	0	2	17.92	4.08	2.715563326	
23	1	0	0	0	2	20.29	2.75	2.935293934	
24	1	0	0	0	2	15.77	2.23	2.516229651	
25	0	0	0	0	4	39.42	7.58	5.094095656	
26	0	0	0	0	2	19.82	3.18	2.891718666	
27	0	0	0	0	4	17.81	2.34	3.090560447	

SUMMARY OUTPUT

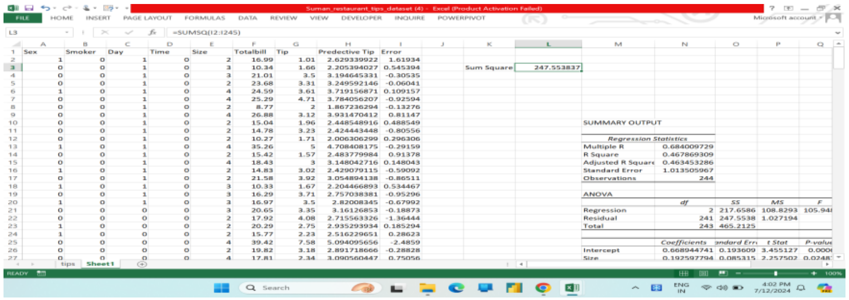
Regression Statistics	
Multiple R	0.684009729
R Square	0.467869309
Adjusted R Square	0.463453286
Standard Error	1.013505967
Observations	244

ANOVA

	df	SS	MS	F	Significa
Regression	2	217.6586	108.8293	105.9481	9.665
Residual	241	247.5538	1.027194		
Total	243	465.2125			

	Coefficients	andard Err	t Stat	P-value	Lower
Intercept	0.668944741	0.193609	3.455127	0.00065	0.28
Size	0.192597794	0.085315	2.257502	0.024872	0.024

Find the Sum Square:


Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER INQUIRE POWERPIVOT
Microsoft account
tips Sheet1
READY
100%
Search
ENG IN
4:02 PM 7/12/2024

L3 =SUMSQ(I2:I245)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip	Predective Tip	Error			
2	1	0	1	0	2	16.99	1.01	2.629339922	1.61934			
3	0	0	1	0	3	10.34	1.66	2.205394027	0.545394		Sum Square	247.553837
4	0	0	1	0	3	21.01	3.5	3.194645331	-0.30535			
5	0	0	1	0	2	23.68	3.31	3.249592146	-0.06041			
6	1	0	1	0	4	24.59	3.61	3.719156871	0.109157			
7	0	0	1	0	4	25.29	4.71	3.784056207	-0.92594			
8	0	0	1	0	2	8.77	2	1.867236294	-0.13276			
9	0	0	1	0	4	26.88	3.12	3.931470412	0.81147			
10	0	0	1	0	2	15.04	1.96	2.448548916	0.488549			
11	0	0	1	0	2	14.78	3.23	2.424443448	-0.80556			
12	0	0	1	0	2	10.27	1.71	2.006306299	0.296306			
13	1	0	1	0	4	35.26	5	4.708408175	-0.29159			
14	0	0	1	0	2	15.42	1.57	2.483779984	0.91378			
15	0	0	1	0	4	18.43	3	3.148042716	0.148043			
16	1	0	1	0	2	14.83	3.02	2.429079115	-0.59092			
17	0	0	1	0	2	21.58	3.92	3.054894138	-0.86511			
18	1	0	1	0	3	10.33	1.67	2.204466893	0.534467			
19	0	0	1	0	3	16.29	3.71	2.757038381	-0.95296			
20	1	0	1	0	3	16.97	3.5	2.82008345	-0.67992			
21	0	0	0	0	3	20.65	3.35	3.16126853	-0.18873			
22	0	0	0	0	2	17.92	4.08	2.715563326	-1.36444			
23	1	0	0	0	2	20.29	2.75	2.935293934	0.185294			
24	1	0	0	0	2	15.77	2.23	2.516229651	0.28623			
25	0	0	0	0	4	39.42	7.58	5.094095656	-2.4859			
26	0	0	0	0	2	19.82	3.18	2.891718666	-0.28828			
27	0	0	0	0	4	17.81	2.34	3.090560447	0.75056			

SUMMARY OUTPUT

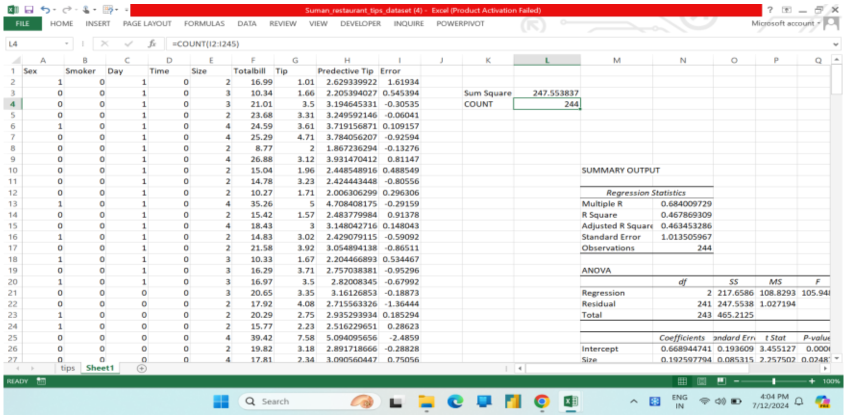
Regression Statistics	
Multiple R	0.684009729
R Square	0.467869309
Adjusted R Square	0.463453286
Standard Error	1.013505967
Observations	244

ANOVA

	df	SS	MS	F
Regression	2	217.6586	108.8293	105.94
Residual	241	247.5538	1.027194	
Total	243	465.2125		

	Coefficients	andard Err	t Stat	P-value
Intercept	0.668944741	0.193609	3.455127	0.000
Size	0.192597794	0.085315	2.257502	0.0248

Find the COUNT:



Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)

L4 =COUNT(I2:I245)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sex	Smoker	Day	Time	Size	Totalbill	Tip	Predective Tip	Error			
2	1	0	1	0	2	16.99	1.01	2.629339922	1.61934			
3	0	0	1	0	3	10.34	1.66	2.205394027	0.545394		Sum Square	247.553837
4	0	0	1	0	3	21.01	3.5	3.194645331	-0.30535		COUNT	244
5	0	0	1	0	2	23.68	3.31	3.249592146	-0.06041			
6	1	0	1	0	4	24.59	3.61	3.719156871	0.109157			
7	0	0	1	0	4	25.29	4.71	3.784056207	-0.92594			
8	0	0	1	0	2	8.77	2	1.867236294	-0.13276			
9	0	0	1	0	4	26.88	3.12	3.931470412	0.81147			
10	0	0	1	0	2	15.04	1.96	2.448548916	0.488549			
11	0	0	1	0	2	14.78	3.23	2.424443448	-0.80556			
12	0	0	1	0	2	10.27	1.71	2.006306299	0.296306			
13	1	0	1	0	4	35.26	5	4.708408175	-0.29159			
14	0	0	1	0	2	15.42	1.57	2.483779984	0.91378			
15	0	0	1	0	4	18.43	3	3.148042716	0.148043			
16	1	0	1	0	2	14.83	3.02	2.429079115	-0.59092			
17	0	0	1	0	2	21.58	3.92	3.054894138	-0.86511			
18	1	0	1	0	3	10.33	1.67	2.204466893	0.534467			
19	0	0	1	0	3	16.29	3.71	2.757038381	-0.95296			
20	1	0	1	0	3	16.97	3.5	2.82008345	-0.67992			
21	0	0	0	0	3	20.65	3.35	3.16126853	-0.18873			
22	0	0	0	0	2	17.92	4.08	2.715563326	-1.36444			
23	1	0	0	0	2	20.29	2.75	2.935293934	0.185294			
24	1	0	0	0	2	15.77	2.23	2.516229651	0.28623			
25	0	0	0	0	4	39.42	7.58	5.094095656	-2.4859			
26	0	0	0	0	2	19.82	3.18	2.891718666	-0.28828			
27	0	0	0	0	4	17.81	2.34	3.090560447	0.75056			

SUMMARY OUTPUT

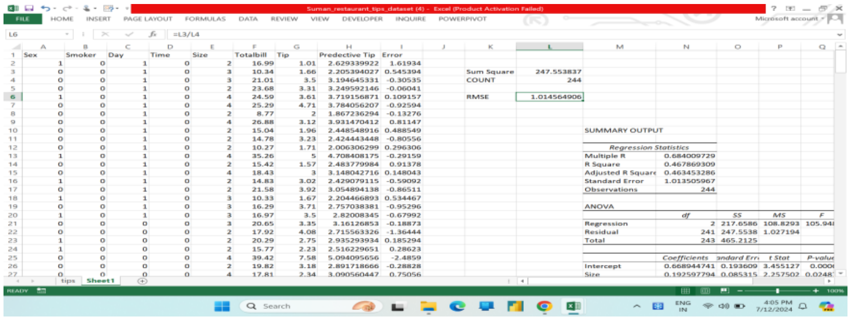
Regression Statistics	
Multiple R	0.684009729
R Square	0.467869309
Adjusted R Square	0.463453286
Standard Error	1.013505967
Observations	244

ANOVA

	df	SS	MS	F
Regression	2	217.6586	108.8293	105.94
Residual	241	247.5538	1.027194	
Total	243	465.2125		

	Coefficients	andard Err	t Stat	P-value
Intercept	0.668944741	0.193609	3.455127	0.000
Size	0.192597794	0.085315	2.257502	0.0248

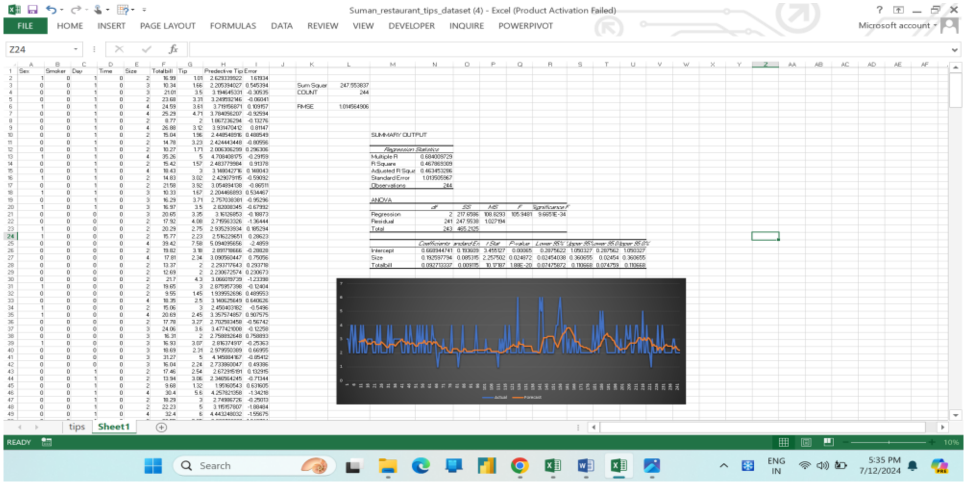
Find the RMSE:



Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)

L6 =L3/L4

	K	L
3	Sum Square	247.553837
4	COUNT	244
6	RMSE	1.014564906



Suman_restaurant_tips_dataset (4) - Excel (Product Activation Failed)

Z24

	K	L
3	Sum Squar	247.553837
4	COUNT	244
6	RMSE	1.014564906

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.684009729
R Square	0.467869309
Adjusted R Squ	0.463453286
Standard Error	1.013505967
Observations	244

ANOVA

	df	SS	MS	F	Significance F
Regression	2	217.6586	108.8293	105.9481	9.6651E-34
Residual	241	247.5538	1.027194		
Total	243	465.2125			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	0.668944741	0.193609	3.455127	0.00065	0.2875622	1.050327	0.287562	1.050327
Size	0.192597794	0.085315	2.257502	0.024872	0.02454038	0.360655	0.02454	0.360655
Totalbill	0.092713337	0.009115	10.17187	1.88E-20	0.07475872	0.110668	0.074759	0.110668