Predicting restaurant tips using predictive analytics on Excel

Course-end Project 1

By- Sujoybrata Dasgupta

Predicting restaurant tips using predictive analytics on Excel

Course-end Project 1

Description- Use excel to predict restaurant tips.

Description:

The dataset in file ***Restaurant tips dataset.xlsx*** contains tips data for different customers. The following are the features in the dataset:

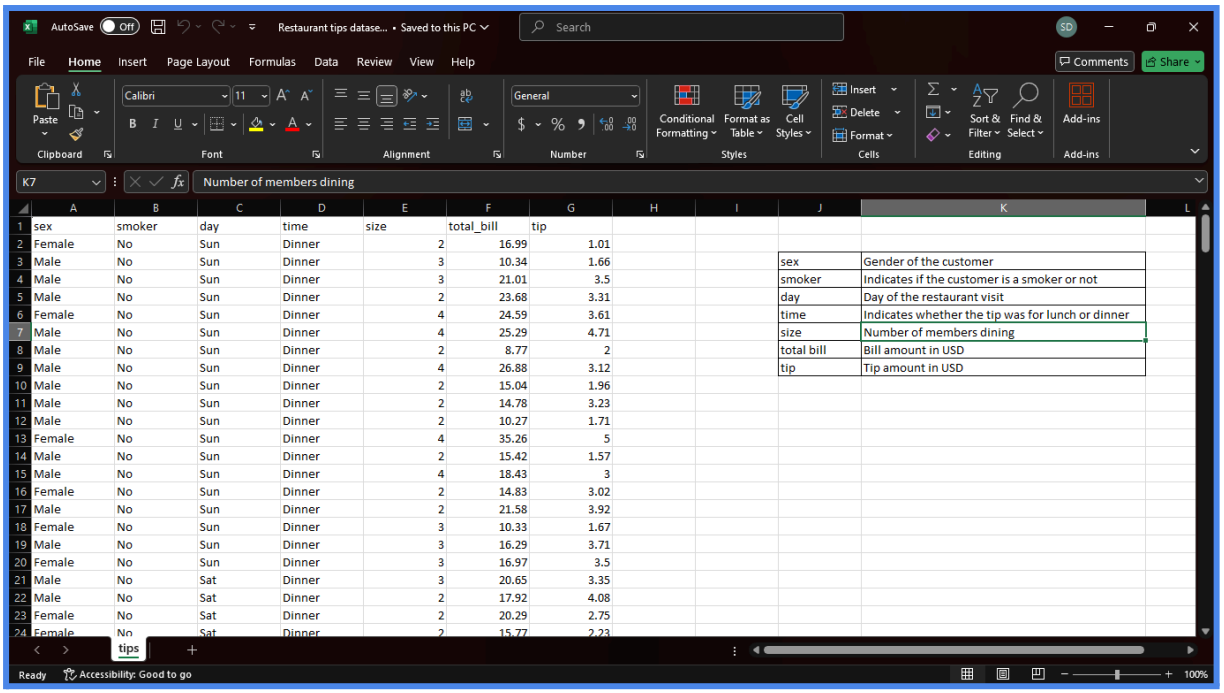
sex	Gender of the customer
smoker	Indicates if the customer is a smoker or not
day	Day of the restaurant visit
time	Indicates whether the tip was for lunch or dinner
size	Number of members dining
total bill	Bill amount in USD
tip	Tip amount in USD

The following project tasks are required to be performed in excel:

- Use the restaurant tips file for the analytics using Excel
- Find out if there are any missing values and clean the data
- Find the features that are independent and dependent
- Identify which predictive problem is needed.
- Encode the categorical variables to numeric values using IF conditions
- Build an appropriate model with the dataset.
- Calculate the predicted and actual tips values.
- Calculate the RMSE(Root Mean Square Error) of the model. RMSE is root of mean of square errors.

Tools required: Microsoft Excel, Data Analysis Add-in.
Expected Deliverables: Model to predict restaurant tips given input values with the mathematical equation for predicting the tips value.

Step 1- Using the Restaurant tip dataset

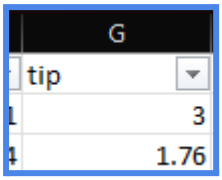


sex	smoker	day	time	size	total_bill	tip
Female	No	Sun	Dinner	2	16.99	1.01
Male	No	Sun	Dinner	3	10.34	1.66
Male	No	Sun	Dinner	3	21.01	3.5
Male	No	Sun	Dinner	2	23.68	3.31
Female	No	Sun	Dinner	4	24.59	3.61
Male	No	Sun	Dinner	4	25.29	4.71
Male	No	Sun	Dinner	2	8.77	2
Male	No	Sun	Dinner	4	26.88	3.12
Male	No	Sun	Dinner	2	15.04	1.96
Male	No	Sun	Dinner	2	14.78	3.23
Male	No	Sun	Dinner	2	10.27	1.71
Female	No	Sun	Dinner	4	35.26	5
Male	No	Sun	Dinner	2	15.42	1.57
Male	No	Sun	Dinner	4	18.43	3
Female	No	Sun	Dinner	2	14.83	3.02
Male	No	Sun	Dinner	2	21.58	3.92
Female	No	Sun	Dinner	3	10.33	1.67
Male	No	Sun	Dinner	3	16.29	3.71
Female	No	Sun	Dinner	3	16.97	3.5
Male	No	Sat	Dinner	3	20.65	3.35
Male	No	Sat	Dinner	2	17.92	4.08
Female	No	Sat	Dinner	2	20.29	2.75
Female	No	Sat	Dinner	2	15.77	2.23

sex	Gender of the customer
smoker	Indicates if the customer is a smoker or not
day	Day of the restaurant visit
time	Indicates whether the tip was for lunch or dinner
size	Number of members dining
total bill	Bill amount in USD
tip	Tip amount in USD

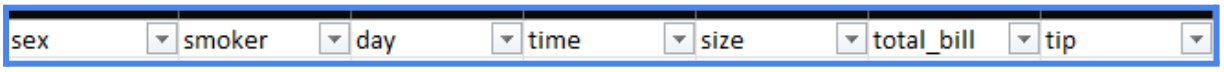
Step 2- Checking for Missing values, No missing values found.

Step3-
Dependent Variable -The target variable that we want to predict which is the TIP column.



G
tip
3
1.76

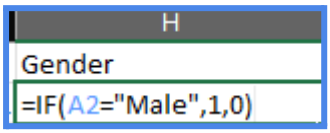
Independent Variable- The other variable columns that may influence the tip amount which are Sex, Smoker, Day, Time, Size, Total Bill columns.



sex	smoker	day	time	size	total_bill	tip

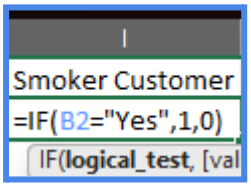
Step 4- Encoding Categorical Variables to Numeric Values by creating new columns

Sex to Gender- If "Male" with 1 and "Female" with 0

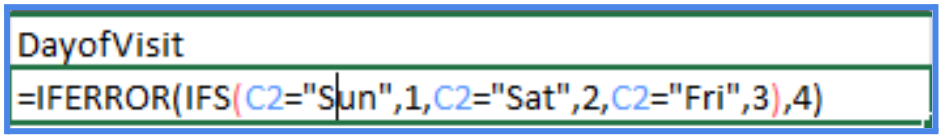


H
Gender
=IF(A2="Male",1,0)

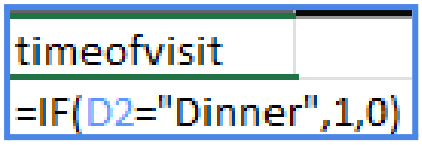
Smoker to Smoker Customer- If smoker, "Yes" with 1 and "No" with 0



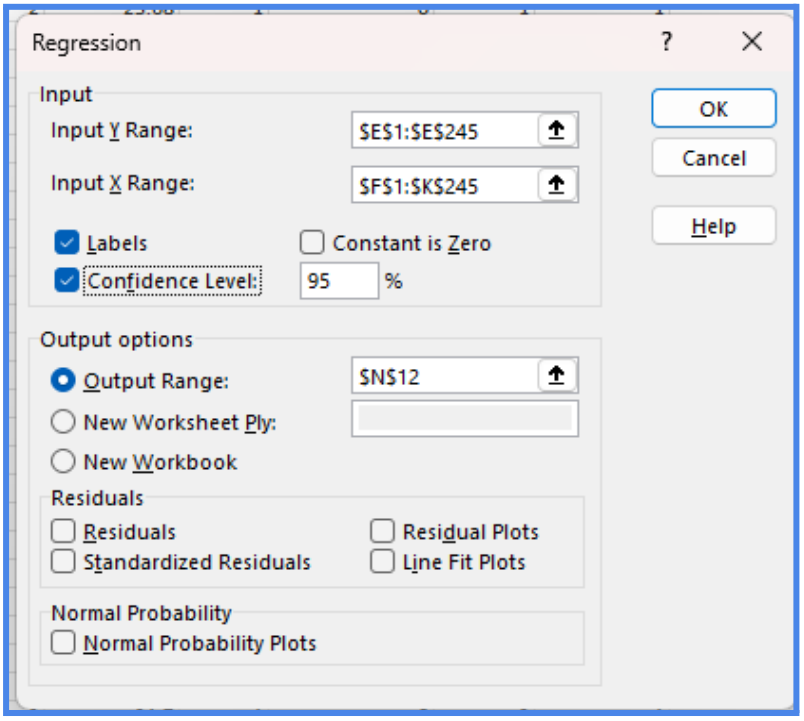
Day to Day of Visit- IFERROR(IFS(C3="Sun",1,C3="Sat",2,C3="Fri",3),4)
Here- SUN- 1, SAT- 2, FRI- 3, THUR- 4



Time to Time of Visit- Replace "Dinner" with 1 and "Lunch" with 0



Step 6: Multi linear regression were performed using the data analysis



SUMMARY OUTPUT

Regression Statistics	Column1
Multiple R	0.684980787
R Square	0.469198679
Adjusted R Square	0.455760671
Standard Error	1.020745565
Observations	244

ANOVA

Column1	df	SS	MS	F	Significance F
Regression	6	218.2770796	36.37951327	34.91579067	4.09922E-30
Residual	237	246.9353974	1.041921508		
Total	243	465.212477			

Column1	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.934962642	0.539857653	1.731868831	0.084598072	-0.128569896	1.998495181	-0.128569896	1.998495181
size	0.174819618	0.089187194	1.960142606	0.051150876	-0.000881295	0.350520531	-0.000881295	0.350520531
total_bill	0.094325088	0.009538173	9.889219168	1.57818E-19	0.075534657	0.113115518	0.075534657	0.113115518
Gender	-0.034644964	0.141081963	-0.245566218	0.806230561	-0.312579818	0.24328989	-0.312579818	0.24328989
Smoker Customer	-0.075663089	0.140198277	-0.539686293	0.589920088	-0.351857061	0.200530884	-0.351857061	0.200530884
DayofVisit	-0.05273982	0.120334639	-0.4382763	0.661585219	-0.289801948	0.184322308	-0.289801948	0.184322308
timeofvisit	-0.112477769	0.307526134	-0.365750277	0.714877667	-0.718311636	0.493356099	-0.718311636	0.493356099

Step 7: After analysing the Mulple linear regression summary table-

Column1	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.934962642	0.539857653	1.731868831	0.084598072	-0.128569896	1.998495181	-0.128569896	1.998495181
size	0.174819618	0.089187194	1.960142606	0.051150876	-0.000881295	0.350520531	-0.000881295	0.350520531
total_bill	0.094325088	0.009538173	9.889219168	1.57818E-19	0.075534657	0.113115518	0.075534657	0.113115518
Gender	-0.034644964	0.141081963	-0.245566218	0.806230561	-0.312579818	0.24328989	-0.312579818	0.24328989
Smoker Customer	-0.075663089	0.140198277	-0.539686293	0.589920088	-0.351857061	0.200530884	-0.351857061	0.200530884
DayofVisit	-0.05273982	0.120334639	-0.4382763	0.661585219	-0.289801948	0.184322308	-0.289801948	0.184322308
timeofvisit	-0.112477769	0.307526134	-0.365750277	0.714877667	-0.718311636	0.493356099	-0.718311636	0.493356099

Column1	Coefficients	P-value
Intercept	0.934962642	0.084598072
size	0.174819618	0.051150876
total_bill	0.094325088	1.57818E-19

After analyzing the MLR table -
total bill = 1.578185+19
Size = 0.0511
Where, 0.0511 > 0.05 (threshold)
And other variables p>0.05 can be ignored.
sex, smoker, day, time were excluded from the analysis as their p value was larger than 5% which could be explained by randomness.

Column1	Coefficients	P-value
Intercept	0.934962642	0.084598072
size	0.174819618	0.051150876
total_bill	0.094325088	1.57818E-19

Step 8: Calculating Predicted TIP, ERROR, MSE, & RMSE from MLR Summary

Actual_tip	total_bill	size	predicted_tip	ERROR	MSE	RMSE
1.01	16.99	2	2.89	-1.88	3.52	1.88
1.66	10.34	3	1.67	-0.01	0.00	0.01
3.5	21.01	3	2.60	0.90	0.81	0.90
3.31	23.68	2	2.58	0.73	0.53	0.73
3.61	24.59	4	3.02	0.59	0.35	0.59
4.71	25.29	4	3.08	1.63	2.64	1.63

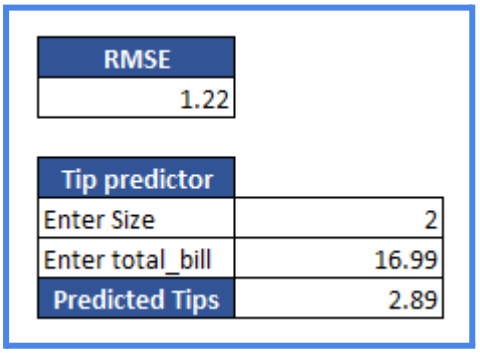
predicted Tip =
Intercept coefficient + (total bill_Coeff * total bill) + (size_coeff * size)

Error =
Actual tip - predicted tip

MSE (Squared Error)=
Error^2

RMSE=
SQRT(MSE)
SQRT(Mean(MSE))

Step 9: Prediction calculator was built to calculate tips based on size and total_bill.



RMSE
1.22

Tip predictor	
Enter Size	2
Enter total_bill	16.99
Predicted Tips	2.89

Predicted Tips= intercept +(coefficient of size * size) + (coefficient of total_bill* total_bill)

Y=Constant +B1*(X1)+B2*(X2)+....BnXn.

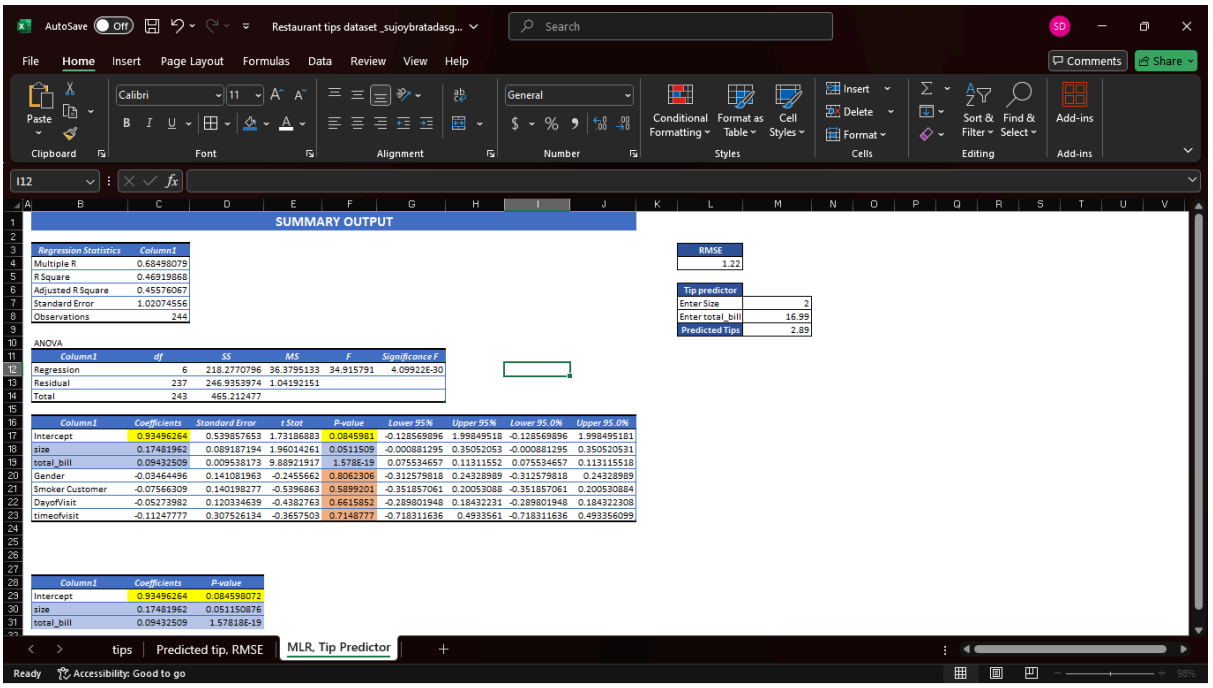
Root mean square error was calculated which could be used to evaluate the quality of predictions. The RMSE value provided a measure of the model's prediction accuracy, indicating how closely the predicted tips matched the actual tips.

The regression model successfully predicted tip amounts based on the given features.

Conclusion-

This project demonstrated the use of Excel for data cleaning, exploratory analysis, and regression modeling. By encoding categorical variables and performing regression analysis, we built a predictive model to estimate restaurant tips. The RMSE value helped evaluate the model's performance, ensuring its reliability for future predictions. This project is a great example of how Excel can be used for basic predictive analytics, even by beginners.

Screenshots Attached-



SUMMARY OUTPUT

Regression Statistics	Column1
Multiple R	0.68498079
R Square	0.46919868
Adjusted R Square	0.45576067
Standard Error	1.02074556
Observations	244

ANOVA

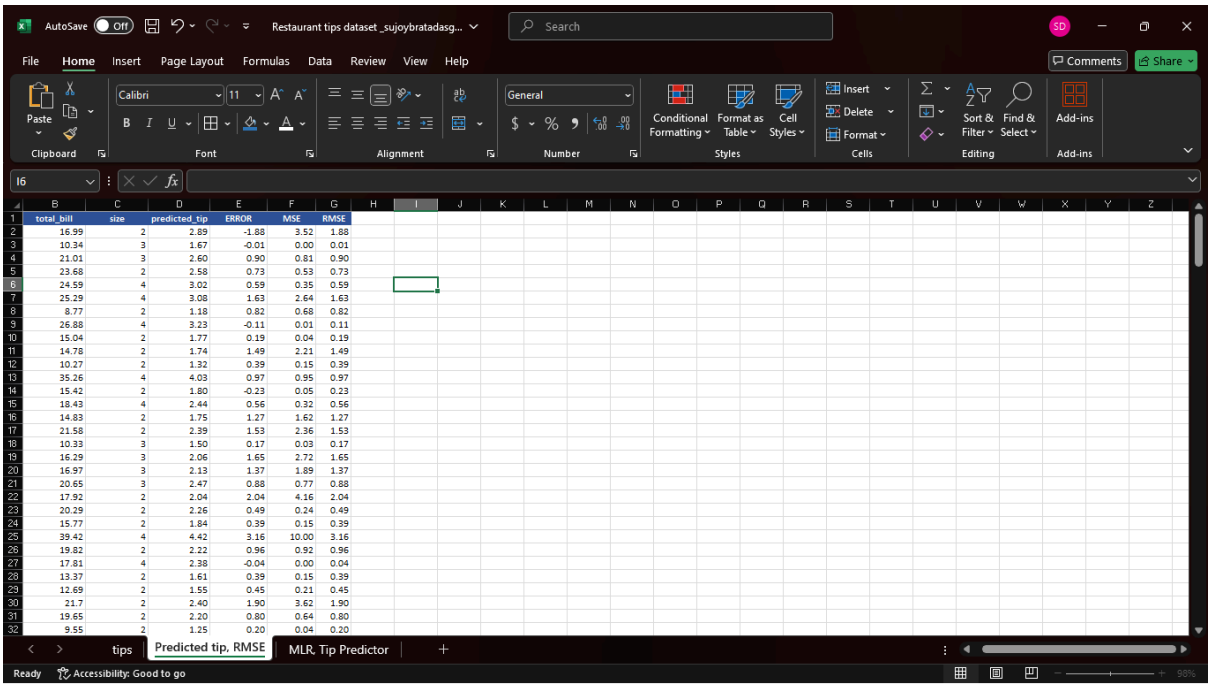
Column1	df	SS	MS	F	Significance F
Regression	6	218.2770796	36.3795133	34.915791	4.09922E-30
Residual	237	246.9353974	1.04192151		
Total	243	465.212477			

Column1	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.93496264	0.539857653	1.73186883	0.0845981	-0.128569896	1.99849518	-0.128569896	1.998495181
size	0.17481962	0.089187194	1.96014261	0.0511509	-0.000881295	0.35052053	-0.000881295	0.350520531
total_bill	0.09432509	0.009538173	9.88921917	1.578E-19	0.075534657	0.11311552	0.075534657	0.113115518
Gender	-0.03464496	0.141081963	-0.2455662	0.8062306	-0.312579818	0.24328989	-0.312579818	0.24328989
Smoker Customer	-0.07566309	0.140198277	-0.5396863	0.5899201	-0.351857061	0.20053088	-0.351857061	0.200530884
DayofVisit	-0.05273982	0.120334639	-0.4382763	0.6615852	-0.289801948	0.18432231	-0.289801948	0.184322308
timeofvisit	-0.11247777	0.307526134	-0.3657503	0.7148777	-0.718311636	0.4933561	-0.718311636	0.493356099

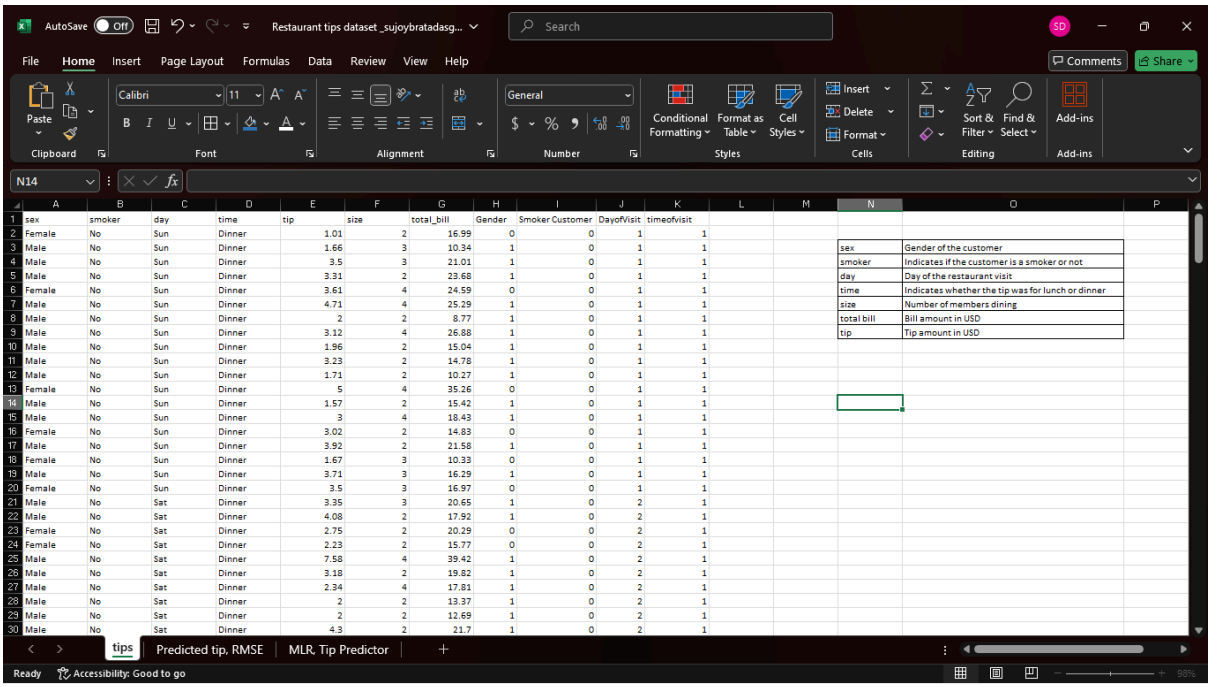
Column1	Coefficients	P-value
Intercept	0.93496264	0.084598072
size	0.17481962	0.051150876
total_bill	0.09432509	1.57818E-19

RMSE
1.22

Tip predictor	
Enter Size	2
Enter total_bill	16.99
Predicted Tips	2.89



total_bill	size	predicted_tip	ERROR	MSE	RMSE
16.99	2	2.89	-1.88	3.52	1.88
10.34	3	1.67	-0.01	0.00	0.01
21.01	3	2.60	0.90	0.81	0.90
23.68	2	2.58	0.73	0.53	0.73
24.59	4	3.02	0.59	0.35	0.59
25.29	4	3.08	1.63	2.64	1.63
8.77	2	1.18	0.82	0.68	0.82
26.88	4	3.23	-0.11	0.01	0.11
15.04	2	1.77	0.19	0.04	0.19
14.78	2	1.74	1.49	2.21	1.49
10.27	2	1.32	0.39	0.15	0.39
35.26	4	4.03	0.97	0.95	0.97
15.42	2	1.80	-0.23	0.05	0.23
18.43	4	2.44	0.56	0.32	0.56
14.83	2	1.75	1.27	1.62	1.27
21.58	2	2.39	1.53	2.36	1.53
10.33	3	1.50	0.17	0.03	0.17
16.29	3	2.06	1.65	2.72	1.65
16.97	3	2.13	1.37	1.89	1.37
20.65	3	2.47	0.88	0.77	0.88
17.92	2	2.04	2.04	4.16	2.04
20.29	2	2.26	0.49	0.24	0.49
15.77	2	1.84	0.39	0.15	0.39
39.42	4	4.42	3.16	10.00	3.16
19.82	2	2.22	0.96	0.92	0.96
17.81	4	2.38	-0.04	0.00	0.04
13.37	2	1.61	0.39	0.15	0.39
12.69	2	1.55	0.45	0.21	0.45
21.7	2	2.40	1.90	3.62	1.90
19.65	2	2.20	0.80	0.64	0.80
9.55	2	1.25	0.20	0.04	0.20



sex	smoker	day	time	tip	size	total_bill	Gender	Smoker Customer	DayofVisit	timeofvisit
Female	No	Sun	Dinner	1.01	2	16.99	0	0	1	1
Male	No	Sun	Dinner	1.66	3	10.34	1	0	1	1
Male	No	Sun	Dinner	3.5	3	21.01	1	0	1	1
Male	No	Sun	Dinner	3.31	2	23.68	1	0	1	1
Female	No	Sun	Dinner	3.61	4	24.59	0	0	1	1
Male	No	Sun	Dinner	4.71	4	25.29	1	0	1	1
Male	No	Sun	Dinner	2	2	8.77	1	0	1	1
Male	No	Sun	Dinner	3.12	4	26.88	1	0	1	1
Male	No	Sun	Dinner	1.96	2	15.04	1	0	1	1
Male	No	Sun	Dinner	3.23	2	14.78	1	0	1	1
Male	No	Sun	Dinner	1.71	2	10.27	1	0	1	1
Female	No	Sun	Dinner	5	4	35.26	0	0	1	1
Male	No	Sun	Dinner	1.57	2	15.42	1	0	1	1
Male	No	Sun	Dinner	3	4	18.43	1	0	1	1
Female	No	Sun	Dinner	3.02	2	14.83	0	0	1	1
Male	No	Sun	Dinner	3.92	2	21.58	1	0	1	1
Female	No	Sun	Dinner	1.67	3	10.33	0	0	1	1
Male	No	Sun	Dinner	3.71	3	16.29	1	0	1	1
Female	No	Sun	Dinner	3.5	3	16.97	0	0	1	1
Male	No	Sat	Dinner	3.35	3	20.65	1	0	2	1
Male	No	Sat	Dinner	4.08	2	17.92	1	0	2	1
Female	No	Sat	Dinner	2.75	2	20.29	0	0	2	1
Female	No	Sat	Dinner	2.23	2	15.77	0	0	2	1
Male	No	Sat	Dinner	7.58	4	39.42	1	0	2	1
Male	No	Sat	Dinner	3.18	2	19.82	1	0	2	1
Male	No	Sat	Dinner	2.34	4	17.81	1	0	2	1
Male	No	Sat	Dinner	2	2	13.37	1	0	2	1
Male	No	Sat	Dinner	2	2	12.69	1	0	2	1
Male	No	Sat	Dinner	4.3	2	21.7	1	0	2	1

sex	Gender of the customer
smoker	Indicates if the customer is a smoker or not
day	Day of the restaurant visit
time	Indicates whether the tip was for lunch or dinner
size	Number of members dining
total bill	Bill amount in USD
tip	Tip amount in USD