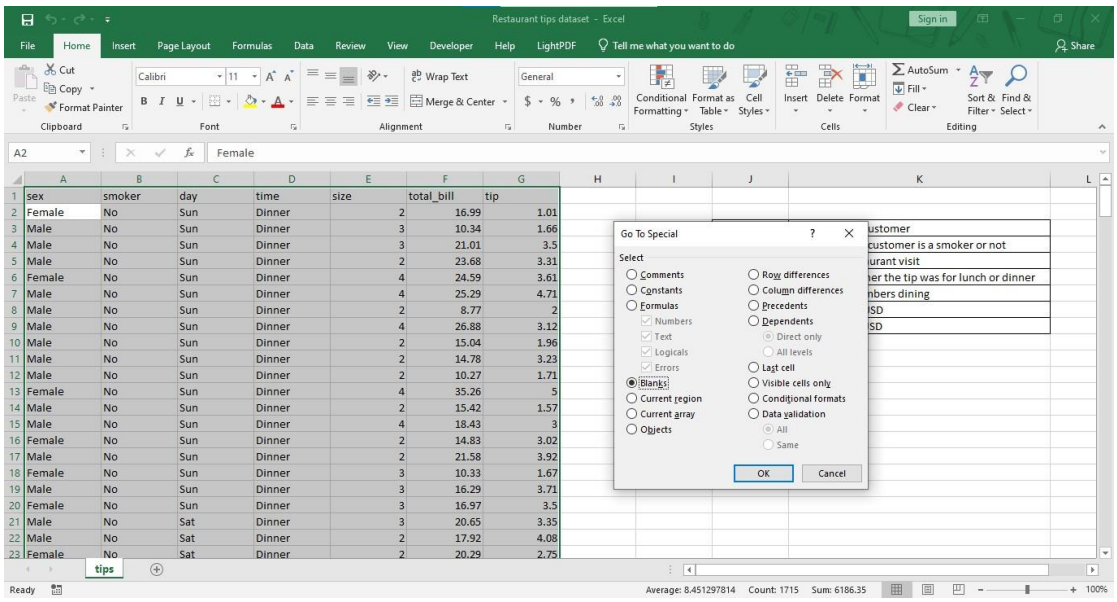
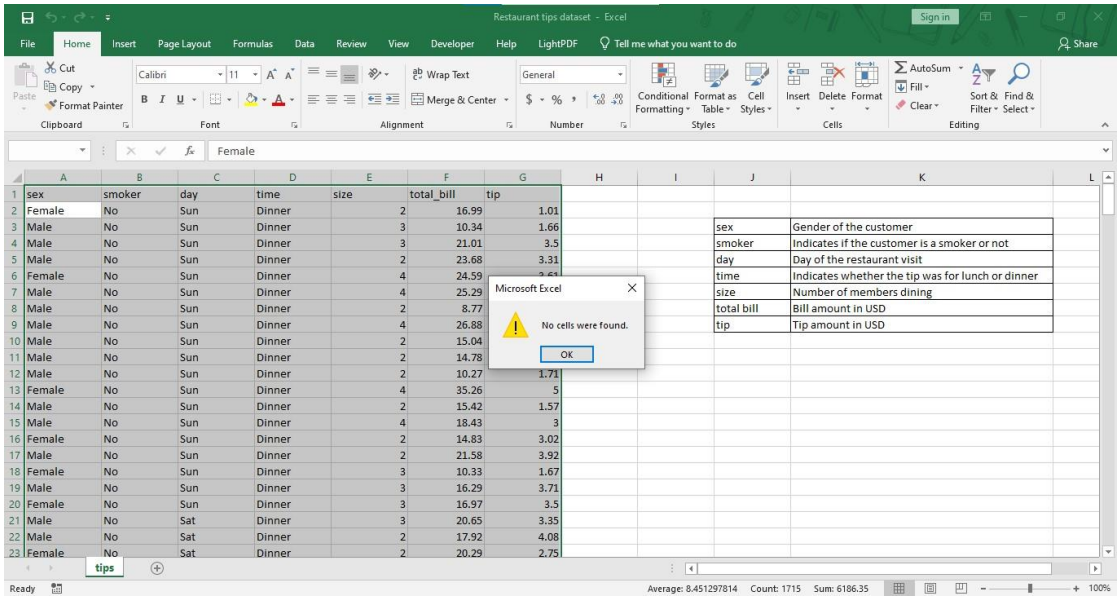
Predicting Restaurant Tips Using Predictive Analytics on Excel

1. **<u>Use the restaurant tips file for the analytics using Excel</u>**
   Open 'Restaurant tips dataset.xlsx' file in your computer.

2. **<u>Find out if there are any missing values and clean the data</u>**
   a. Remove rows with at least one empty cell; choose all data using Control+A and in the home tab, click on Find & Select and go to Special.

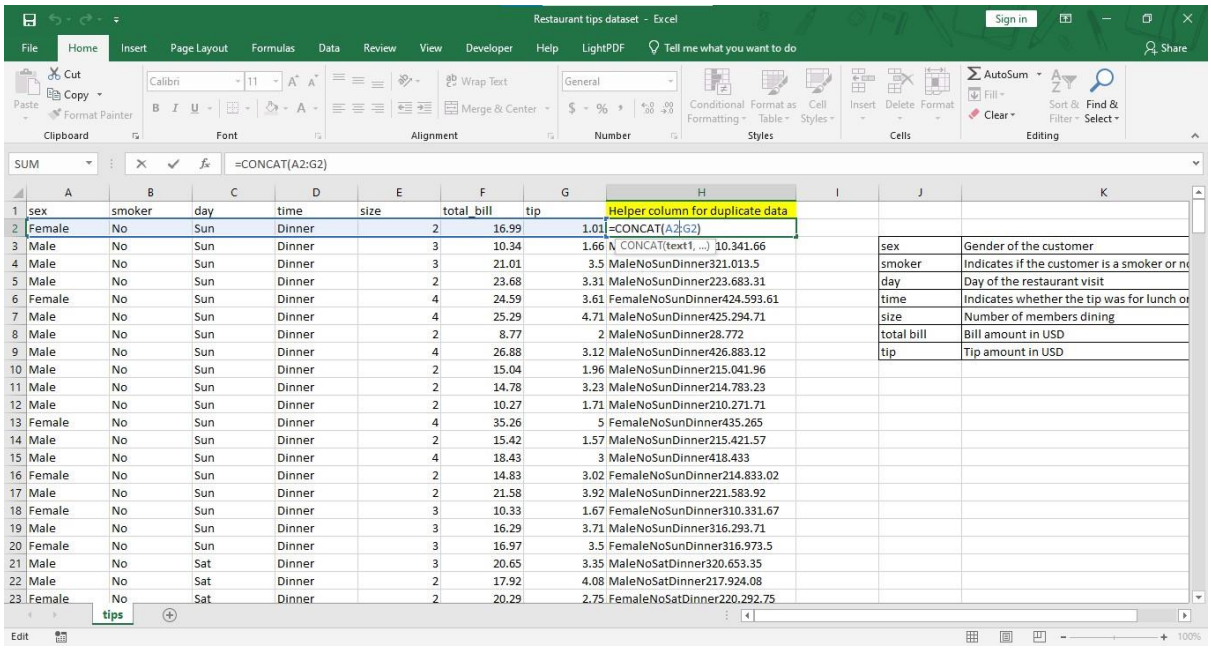


There is no missing value.

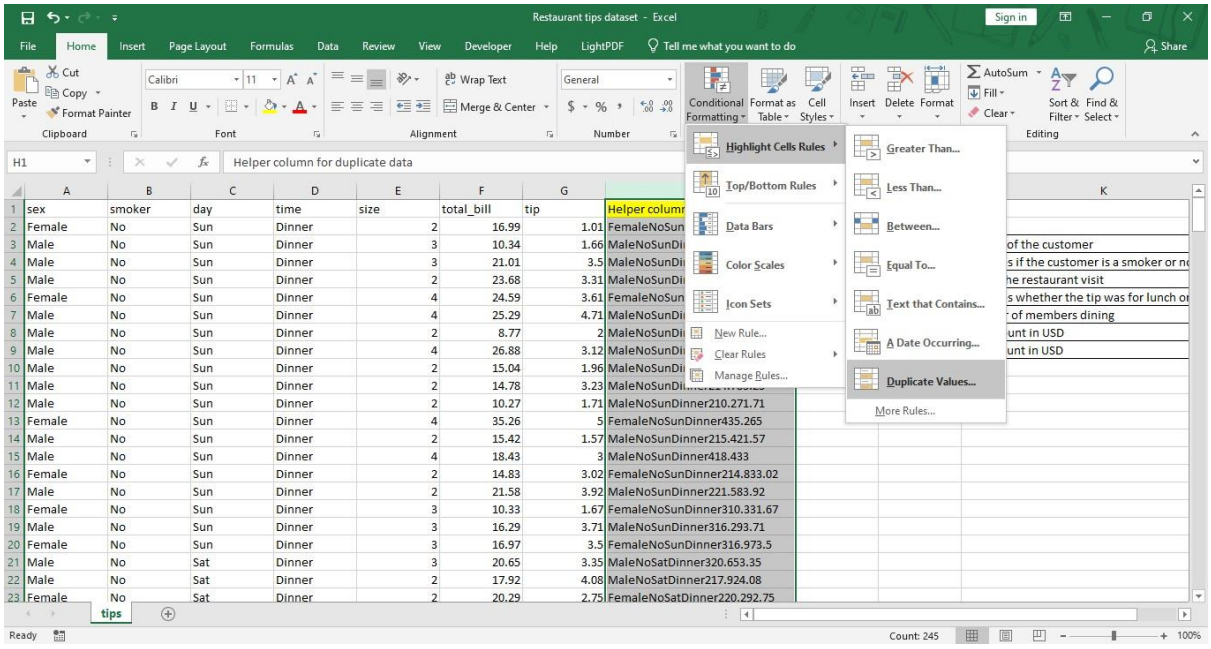


b. Check for duplicates: Choose the dataset using Control+A; In the Data tab, click on Remove duplicates.

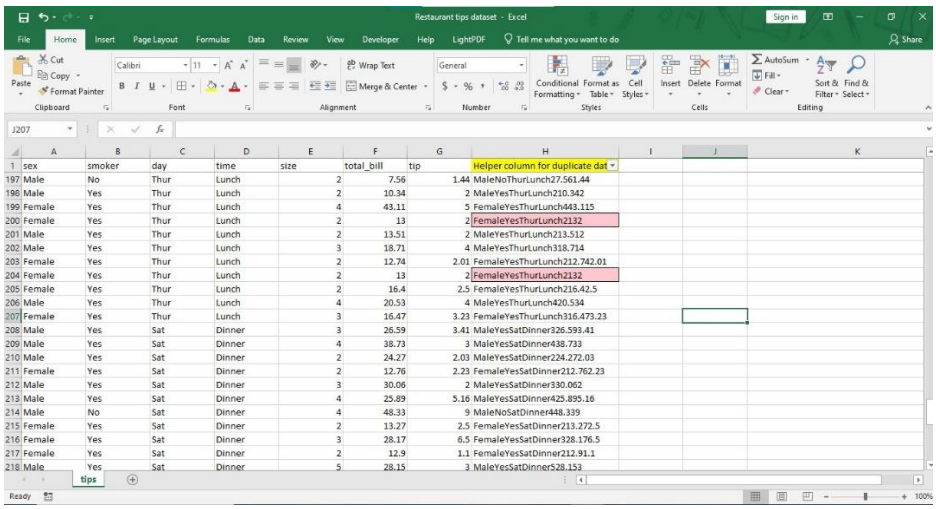
To check for duplicate values create helper column using CONCAT() function.



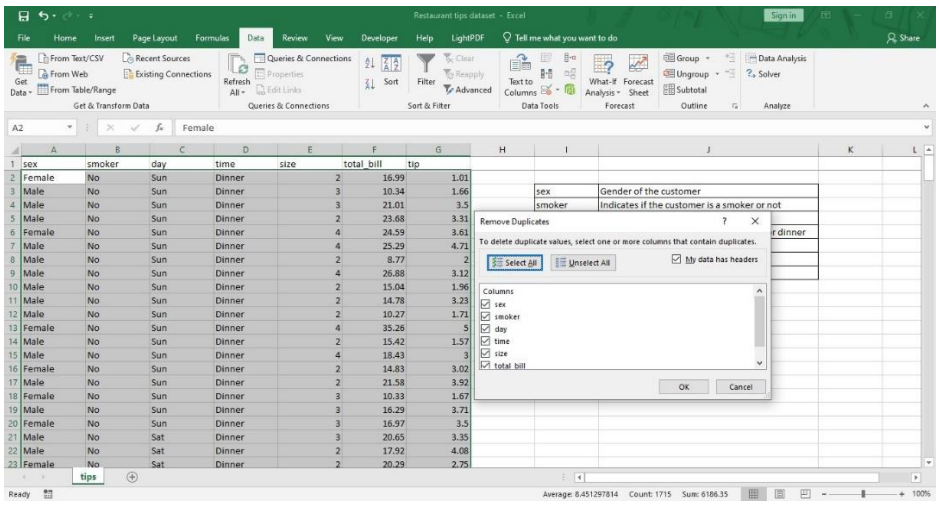
Check for similar data in helper column using conditional formatting.



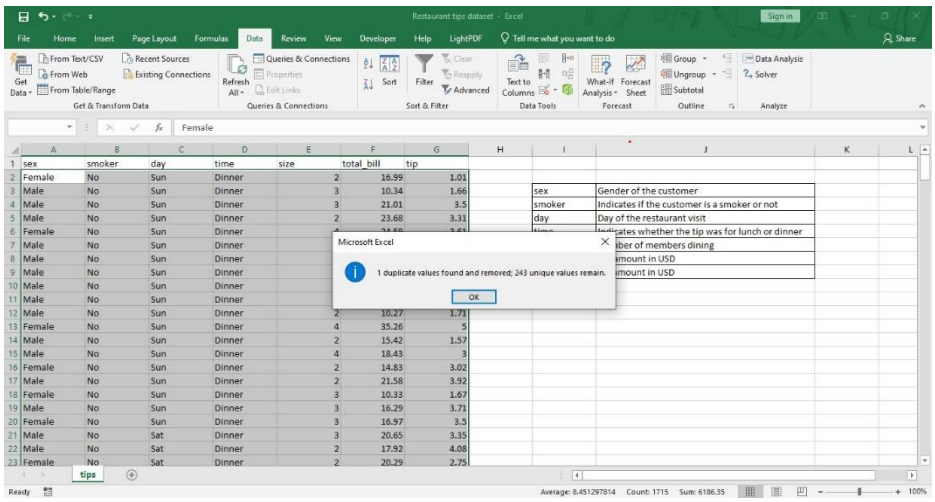
There are only two rows with similar data i.e. only one duplicate data.



Remove duplicate data.



One record is deleted.

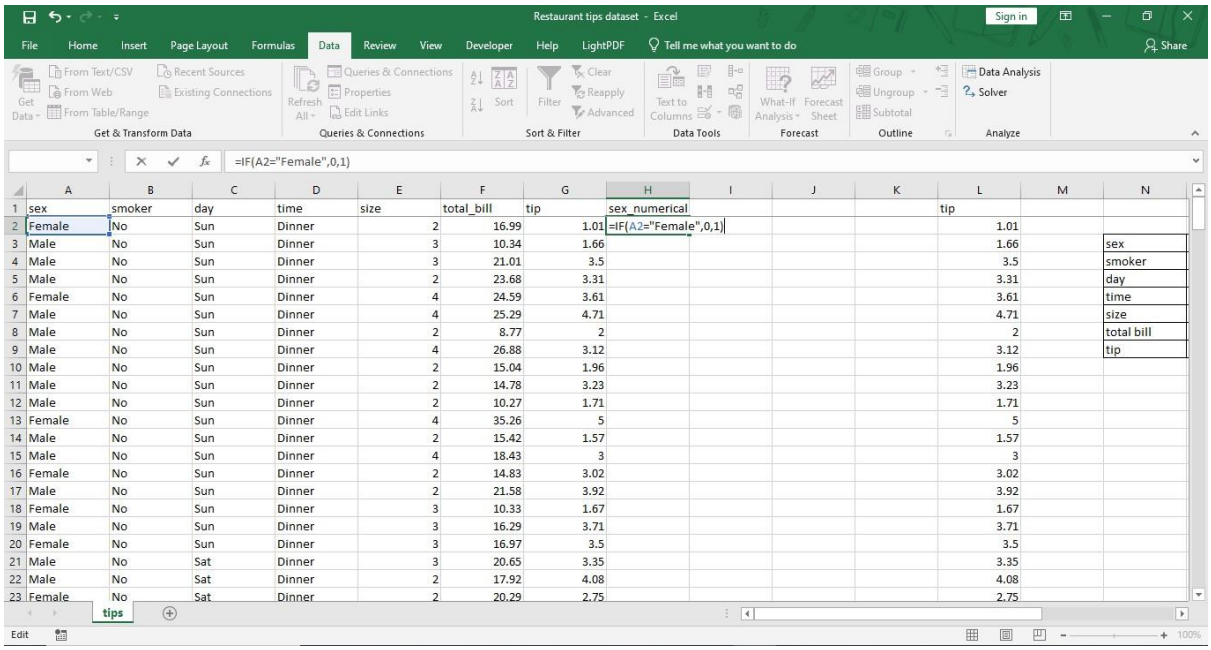


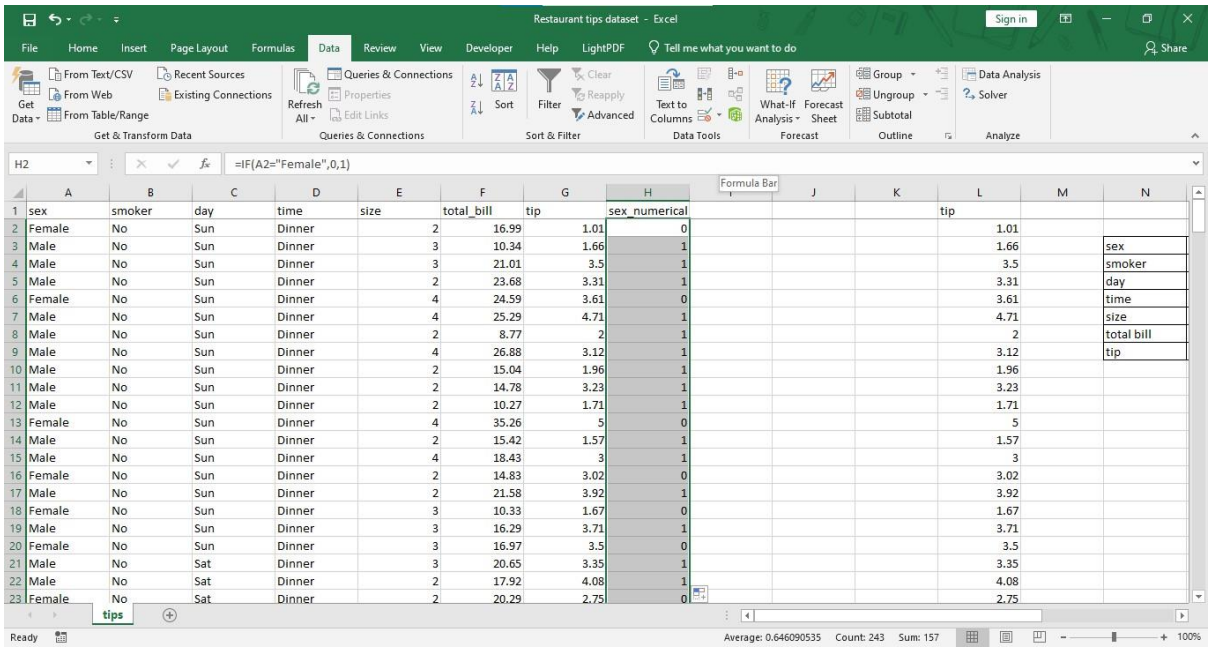
3. Encode the categorical variables to numeric values using IF conditions

There are four categorical columns – Sex, Smoker, Day and Time.

a. New column ‘Sex_numerical’ is added to covert ‘Sex’ into numerical values.

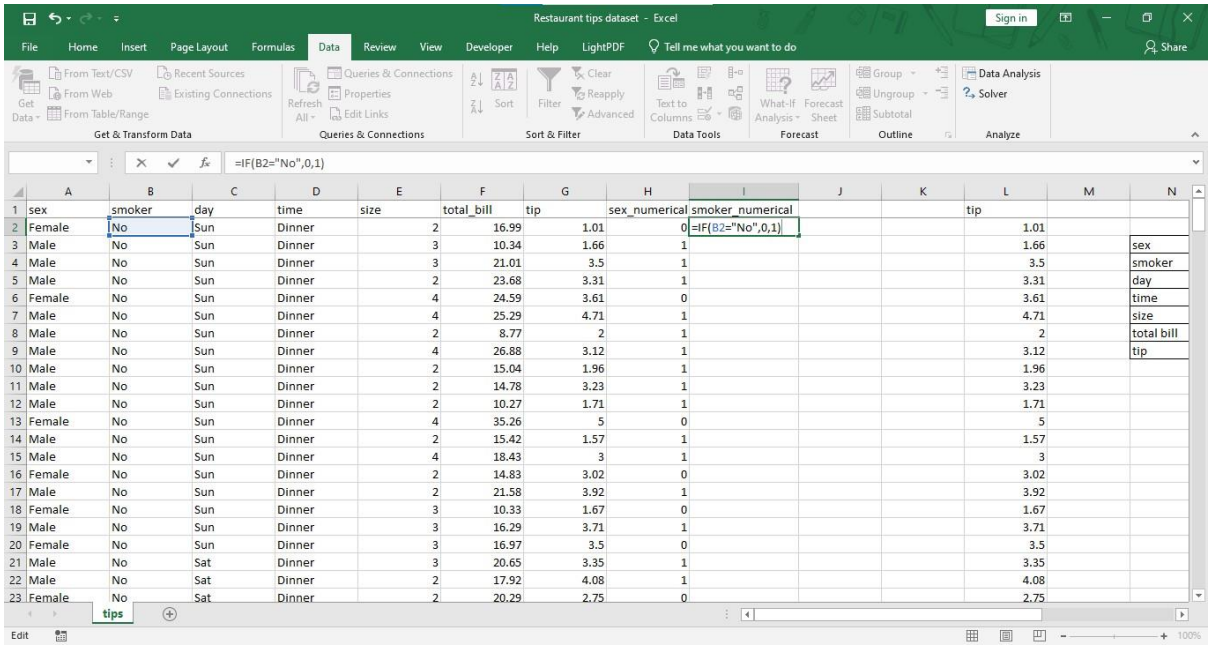
Male = 1
Female=0



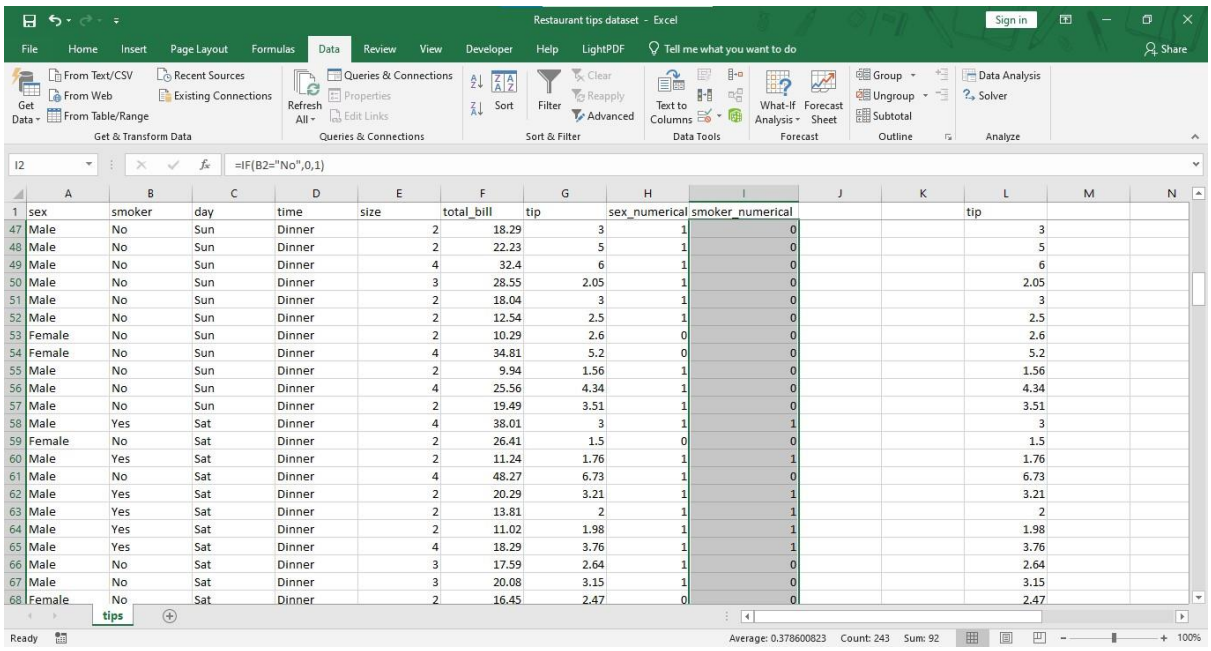


b. New column ‘Smoker_numerical’ is added to covert ‘Smoker’ into numerical values.

Yes = 1
No = 0



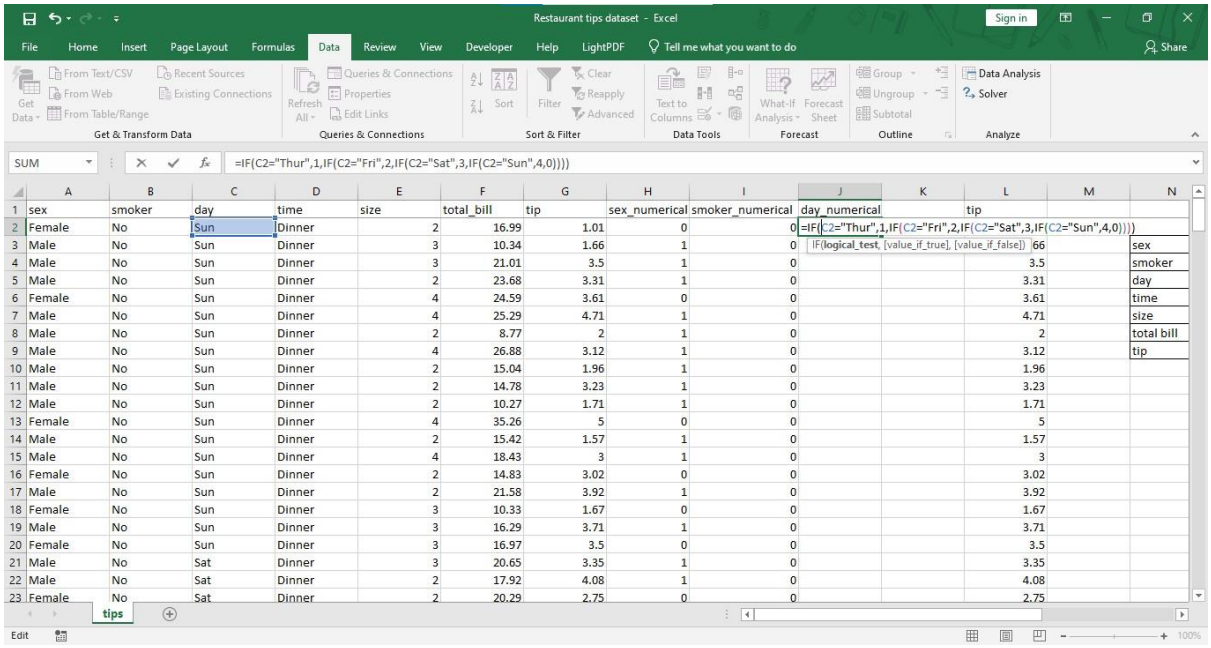
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	sex	smoker	day	time	size	total_bill	tip	sex_numerical	smoker_numerical			tip		
2	Female	No	Sun	Dinner	2	16.99	1.01	0	=IF(B2="No",0,1)			1.01		
3	Male	No	Sun	Dinner	3	10.34	1.66	1				1.66		sex
4	Male	No	Sun	Dinner	3	21.01	3.5	1				3.5		smoker
5	Male	No	Sun	Dinner	2	23.68	3.31	1				3.31		day
6	Female	No	Sun	Dinner	4	24.59	3.61	0				3.61		time
7	Male	No	Sun	Dinner	4	25.29	4.71	1				4.71		size
8	Male	No	Sun	Dinner	2	8.77	2	1				2		total bill
9	Male	No	Sun	Dinner	4	26.88	3.12	1				3.12		tip
10	Male	No	Sun	Dinner	2	15.04	1.96	1				1.96		
11	Male	No	Sun	Dinner	2	14.78	3.23	1				3.23		
12	Male	No	Sun	Dinner	2	10.27	1.71	1				1.71		
13	Female	No	Sun	Dinner	4	35.26	5	0				5		
14	Male	No	Sun	Dinner	2	15.42	1.57	1				1.57		
15	Male	No	Sun	Dinner	4	18.43	3	1				3		
16	Female	No	Sun	Dinner	2	14.83	3.02	0				3.02		
17	Male	No	Sun	Dinner	2	21.58	3.92	1				3.92		
18	Female	No	Sun	Dinner	3	10.33	1.67	0				1.67		
19	Male	No	Sun	Dinner	3	16.29	3.71	1				3.71		
20	Female	No	Sun	Dinner	3	16.97	3.5	0				3.5		
21	Male	No	Sat	Dinner	3	20.65	3.35	1				3.35		
22	Male	No	Sat	Dinner	2	17.92	4.08	1				4.08		
23	Female	No	Sat	Dinner	2	20.29	2.75	0				2.75		

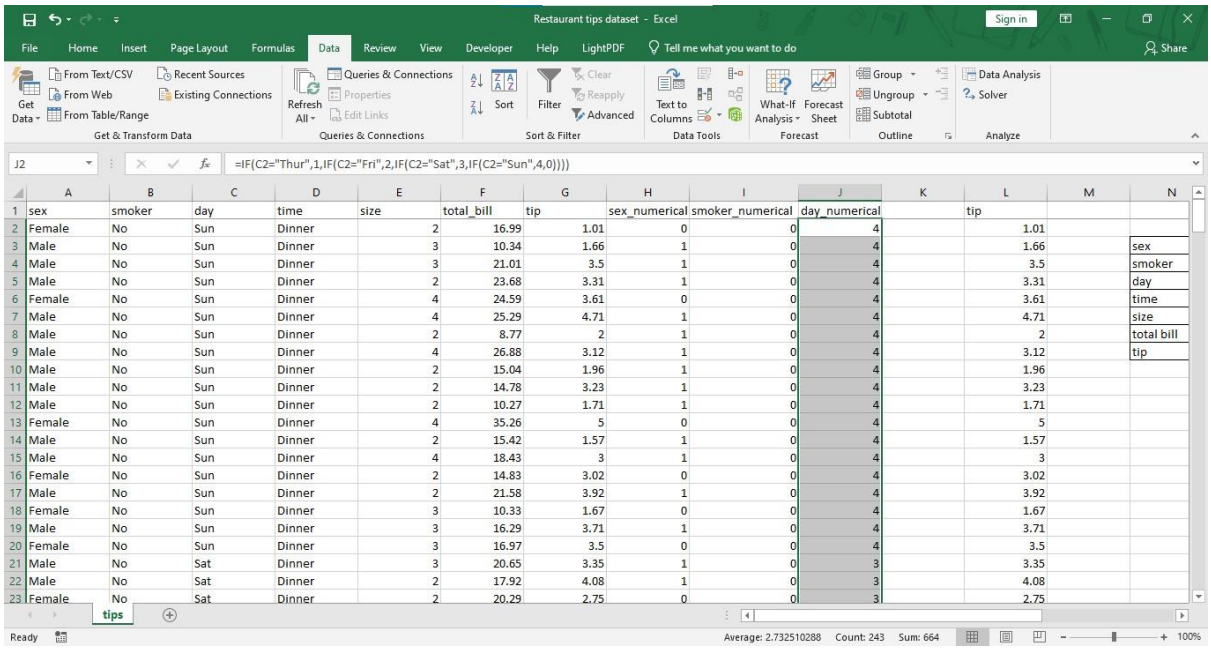


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	sex	smoker	day	time	size	total_bill	tip	sex_numerical	smoker_numerical			tip		
47	Male	No	Sun	Dinner	2	18.29	3	1	0			3		
48	Male	No	Sun	Dinner	2	22.23	5	1	0			5		
49	Male	No	Sun	Dinner	4	32.4	6	1	0			6		
50	Male	No	Sun	Dinner	3	28.55	2.05	1	0			2.05		
51	Male	No	Sun	Dinner	2	18.04	3	1	0			3		
52	Male	No	Sun	Dinner	2	12.54	2.5	1	0			2.5		
53	Female	No	Sun	Dinner	2	10.29	2.6	0	0			2.6		
54	Female	No	Sun	Dinner	4	34.81	5.2	0	0			5.2		
55	Male	No	Sun	Dinner	2	9.94	1.56	1	0			1.56		
56	Male	No	Sun	Dinner	4	25.56	4.34	1	0			4.34		
57	Male	No	Sun	Dinner	2	19.49	3.51	1	0			3.51		
58	Male	Yes	Sat	Dinner	4	38.01	3	1	1			3		
59	Female	No	Sat	Dinner	2	26.41	1.5	0	0			1.5		
60	Male	Yes	Sat	Dinner	2	11.24	1.76	1	1			1.76		
61	Male	No	Sat	Dinner	4	48.27	6.73	1	0			6.73		
62	Male	Yes	Sat	Dinner	2	20.29	3.21	1	1			3.21		
63	Male	Yes	Sat	Dinner	2	13.81	2	1	1			2		
64	Male	Yes	Sat	Dinner	2	11.02	1.98	1	1			1.98		
65	Male	Yes	Sat	Dinner	4	18.29	3.76	1	1			3.76		
66	Male	No	Sat	Dinner	3	17.59	2.64	1	0			2.64		
67	Male	No	Sat	Dinner	3	20.08	3.15	1	0			3.15		
68	Female	No	Sat	Dinner	2	16.45	2.47	0	0			2.47		

c. New column ‘Day_numerical’ is added to covert ‘Day’ into numerical values using label encoding.

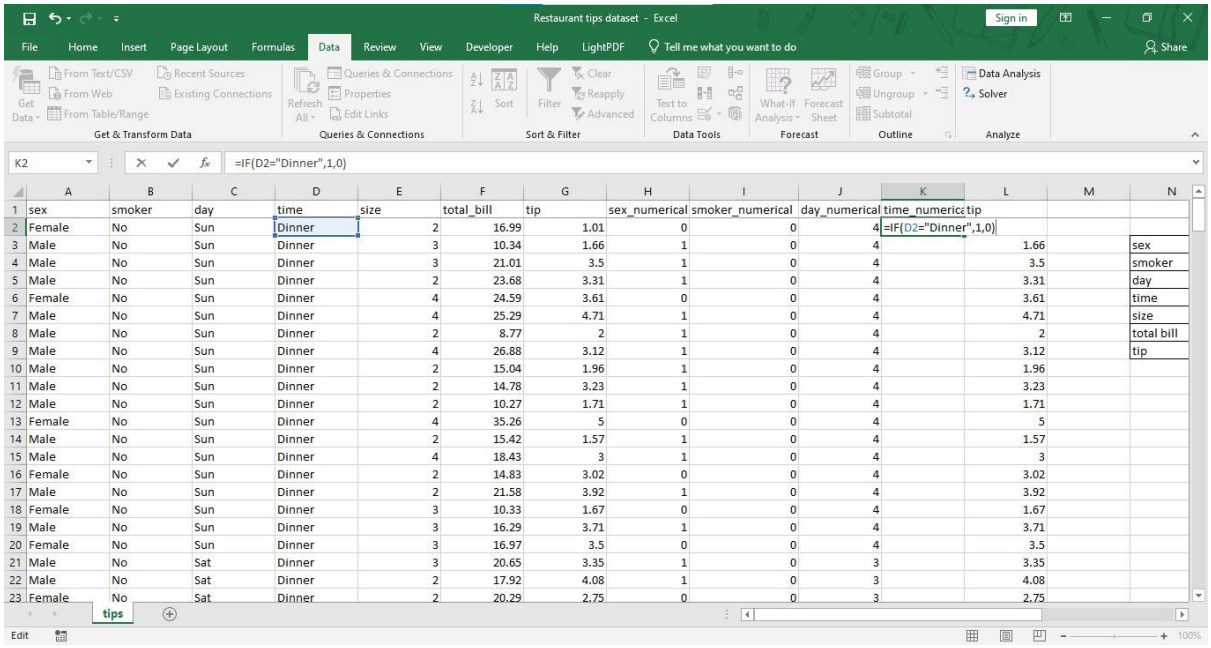
Thur = 1
Fri = 2
Sat = 3
Sun = 4

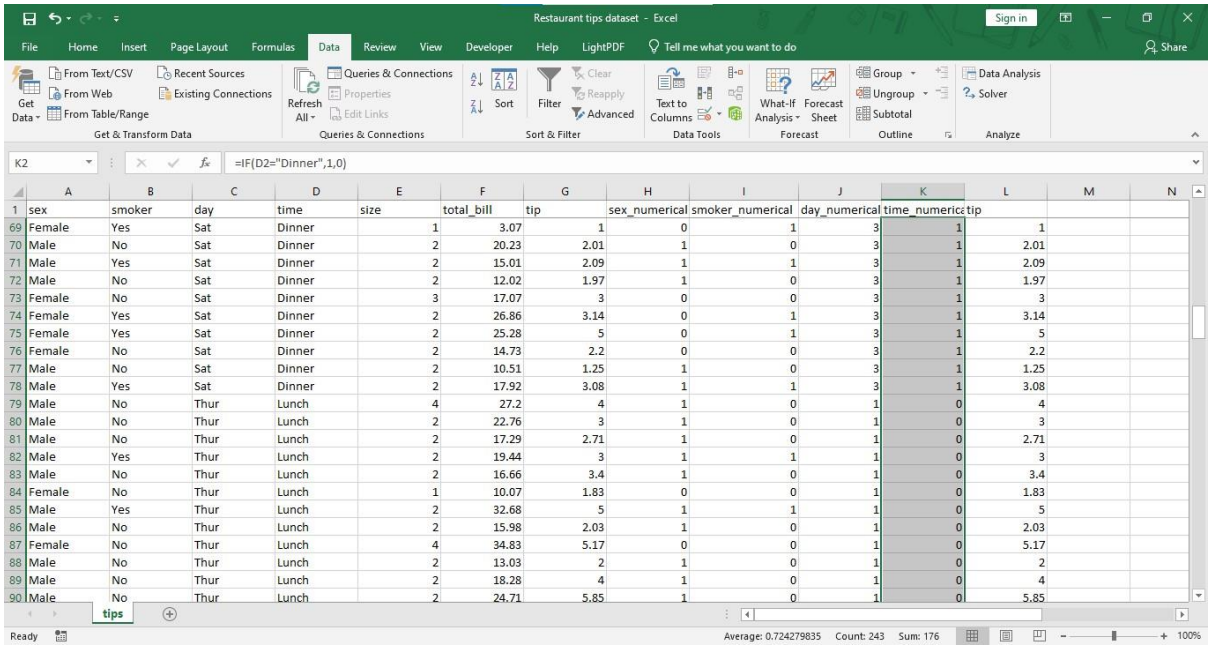




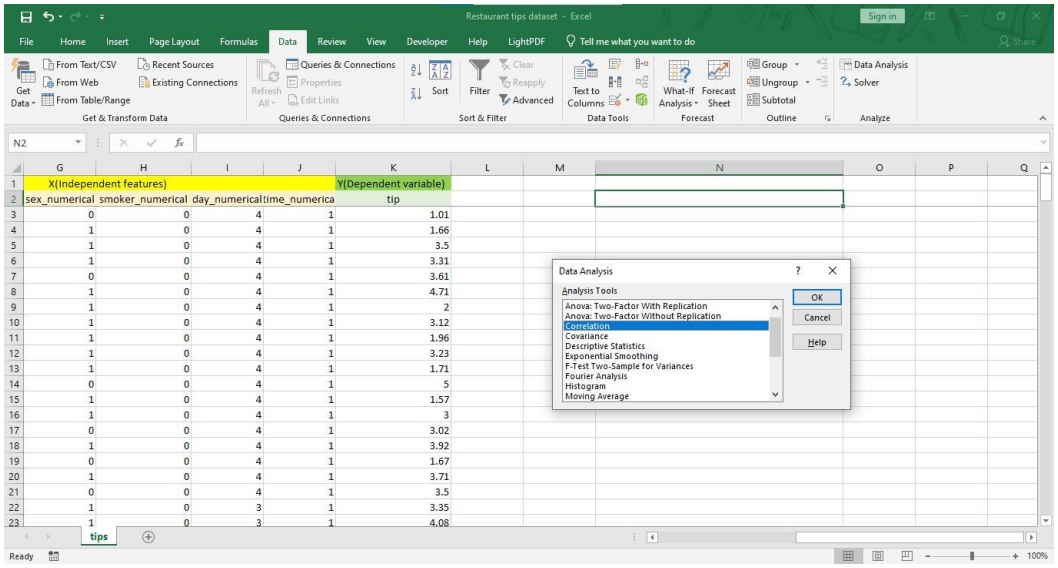
d. New column ‘Time_numerical’ is added to covert ‘Time’ into numerical values.

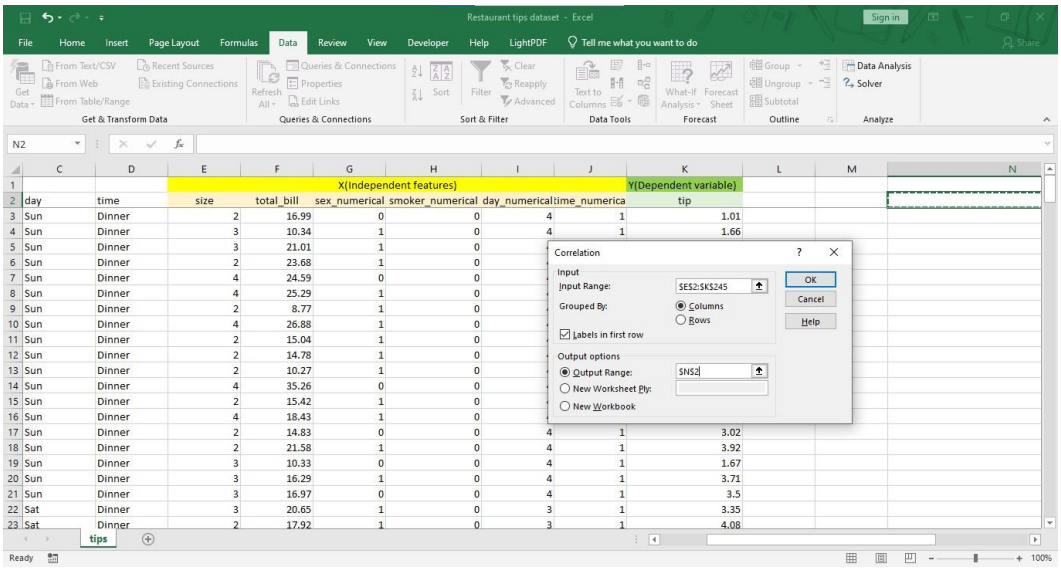
Dinner = 1
Lunch = 0

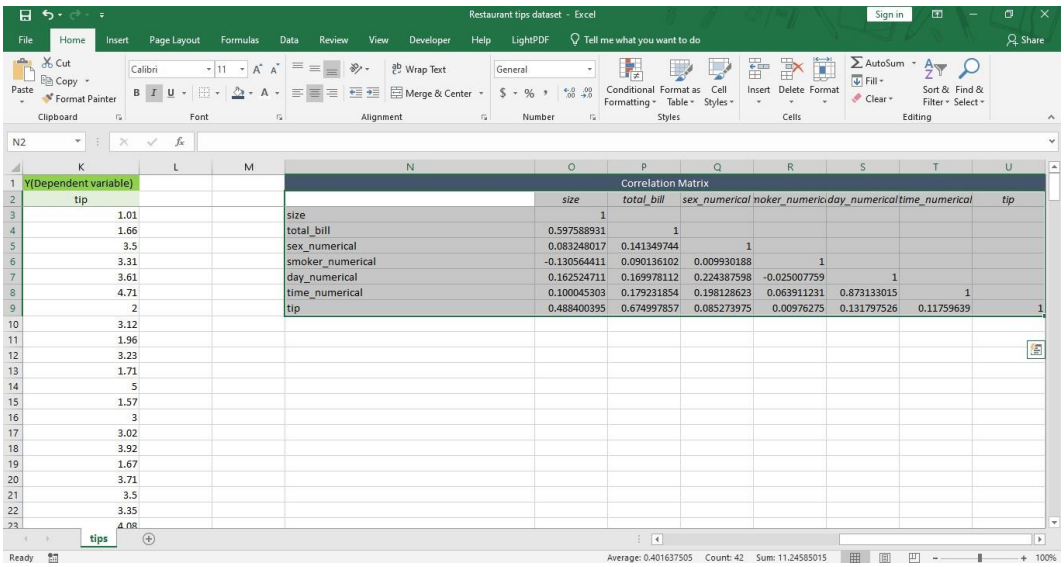




4. **For each independent numeric value, find its correlation coefficient with respect to the tip.**





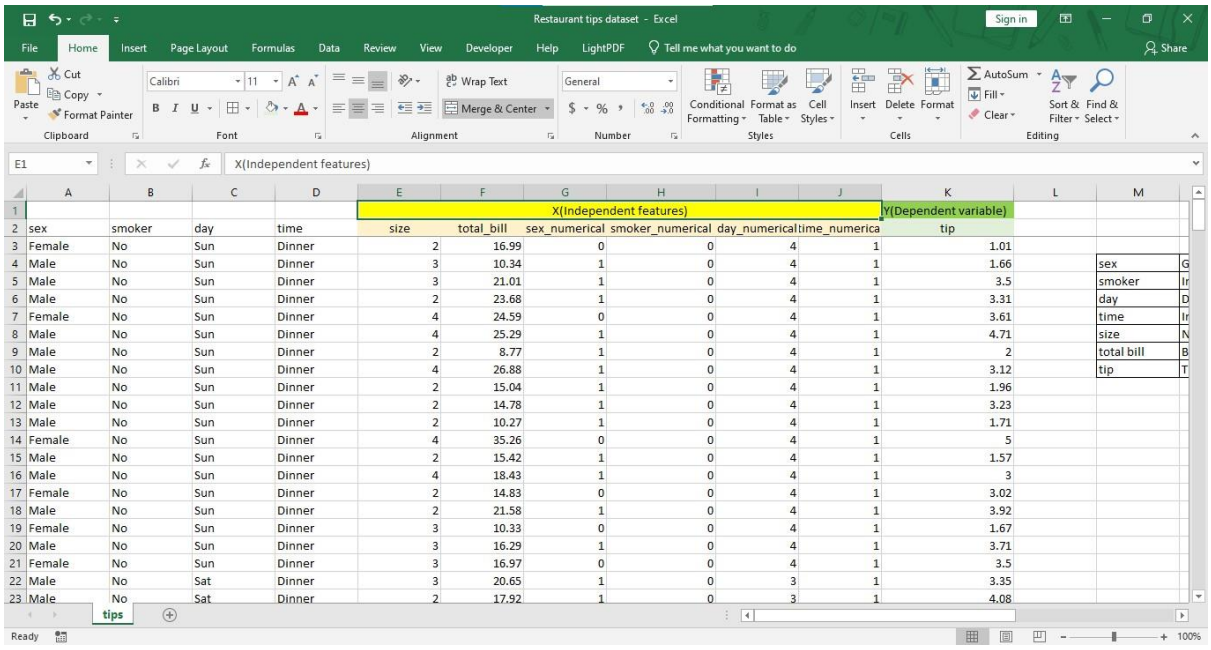


5. Find the features that are independent and dependent.

As per objective of the problem, we need to prepare a model to predict restaurant tips given input values with the mathematical equation for predicting the tips value.

So, Our **Dependent Variable (Target)** is 'tip', which we want to predict.

And rest of the features including 'size', 'total bill', 'sex_numerical', 'smoker_numerical', 'day_numerical' and 'time_numerical' are our **Independent Variables**.



Restaurant tips dataset - Excel

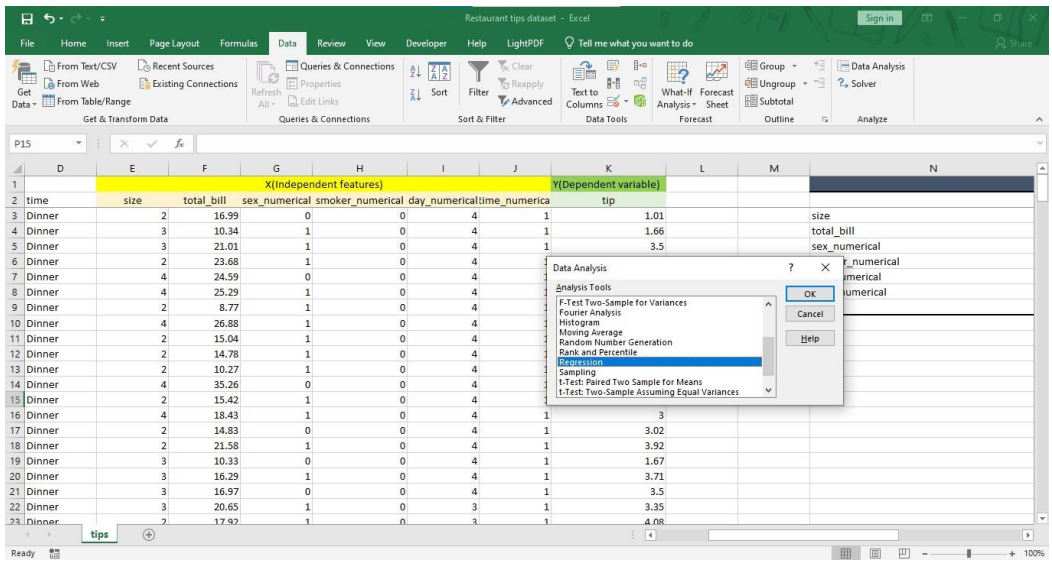
	A	B	C	D	E	F	G	H	I	J	K	L	M
1					X(Independent features)						Y(Dependent variable)		
2	sex	smoker	day	time	size	total_bill	sex_numerical	smoker_numerical	day_numerical	time_numerica	tip		
3	Female	No	Sun	Dinner	2	16.99	0	0	4	1	1.01		
4	Male	No	Sun	Dinner	3	10.34	1	0	4	1	1.66		sex
5	Male	No	Sun	Dinner	3	21.01	1	0	4	1	3.5		smoker
6	Male	No	Sun	Dinner	2	23.68	1	0	4	1	3.31		day
7	Female	No	Sun	Dinner	4	24.59	0	0	4	1	3.61		time
8	Male	No	Sun	Dinner	4	25.29	1	0	4	1	4.71		size
9	Male	No	Sun	Dinner	2	8.77	1	0	4	1	2		total bill
10	Male	No	Sun	Dinner	4	26.88	1	0	4	1	3.12		tip
11	Male	No	Sun	Dinner	2	15.04	1	0	4	1	1.96		
12	Male	No	Sun	Dinner	2	14.78	1	0	4	1	3.23		
13	Male	No	Sun	Dinner	2	10.27	1	0	4	1	1.71		
14	Female	No	Sun	Dinner	4	35.26	0	0	4	1	5		
15	Male	No	Sun	Dinner	2	15.42	1	0	4	1	1.57		
16	Male	No	Sun	Dinner	4	18.43	1	0	4	1	3		
17	Female	No	Sun	Dinner	2	14.83	0	0	4	1	3.02		
18	Male	No	Sun	Dinner	2	21.58	1	0	4	1	3.92		
19	Female	No	Sun	Dinner	3	10.33	0	0	4	1	1.67		
20	Male	No	Sun	Dinner	3	16.29	1	0	4	1	3.71		
21	Female	No	Sun	Dinner	3	16.97	0	0	4	1	3.5		
22	Male	No	Sat	Dinner	3	20.65	1	0	3	1	3.35		
23	Male	No	Sat	Dinner	2	17.92	1	0	3	1	4.08		

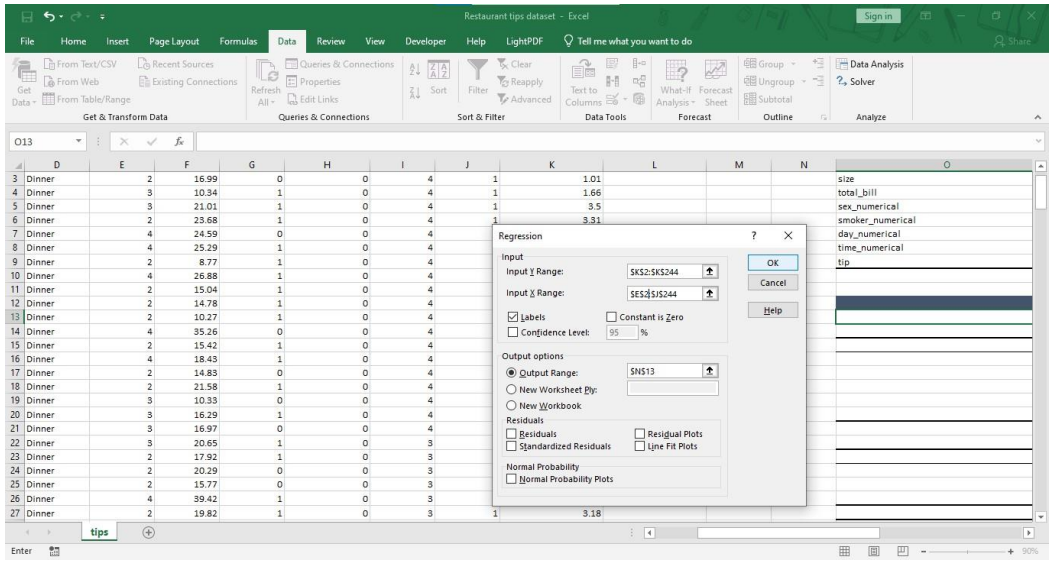
6. Identify which predictive problem is needed.

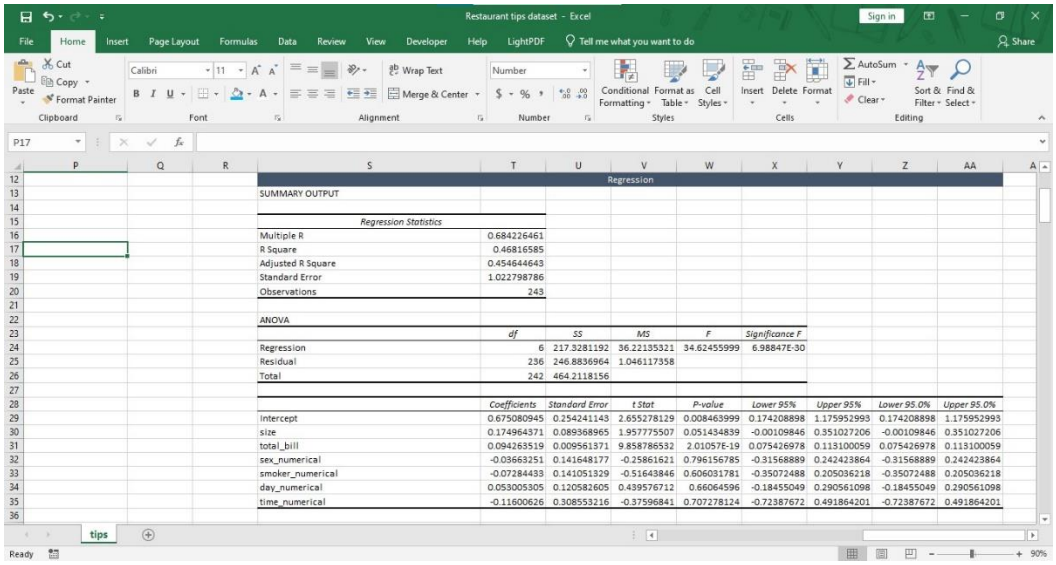
This is a **regression problem** because we are predicting a continuous numerical value ('tip' amount).

7. Build an appropriate model with the dataset.

Building the Regression Model



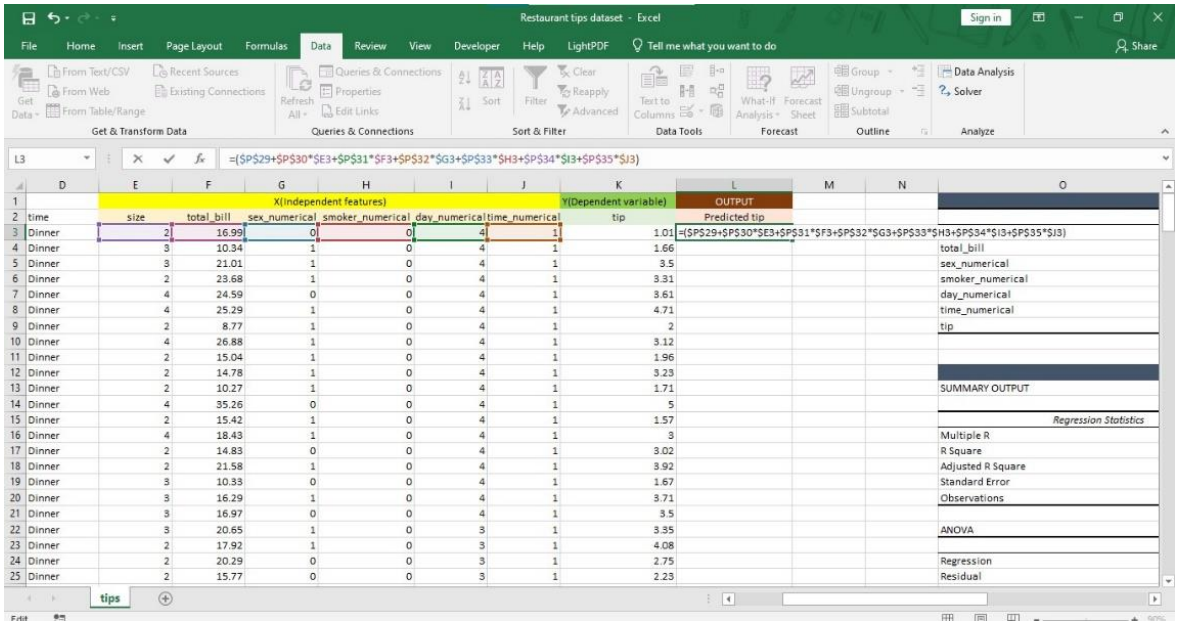


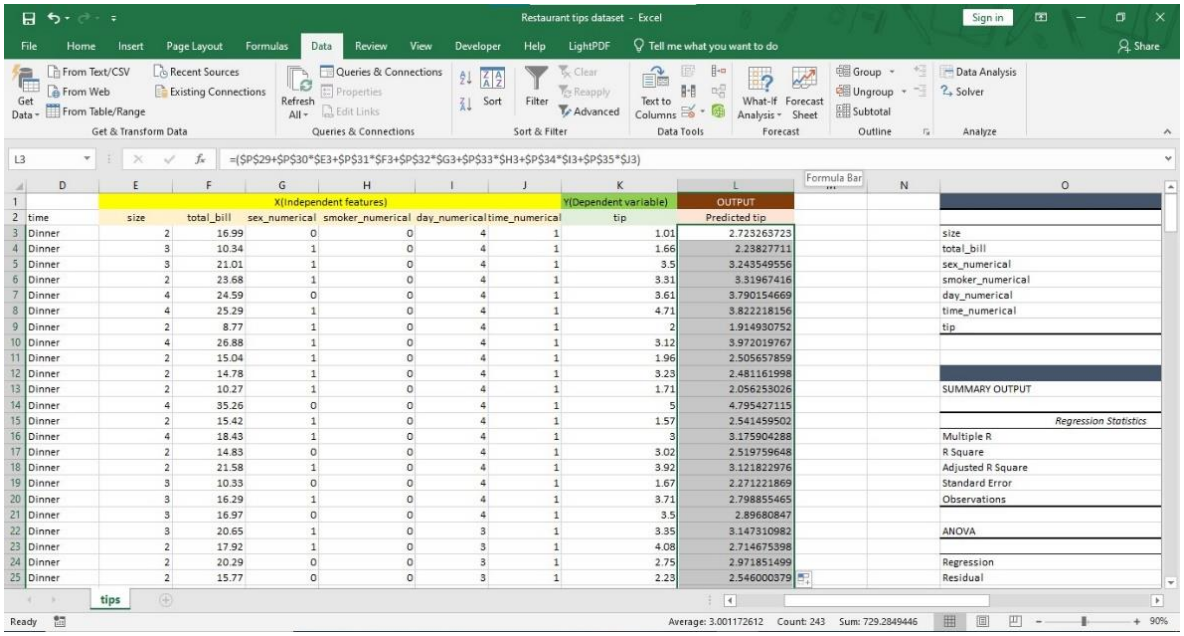


8. <u>Calculate the predicted tips values.</u>

Create a new column named predicted_tip, which is calculated using the coefficients obtained from the regression model.

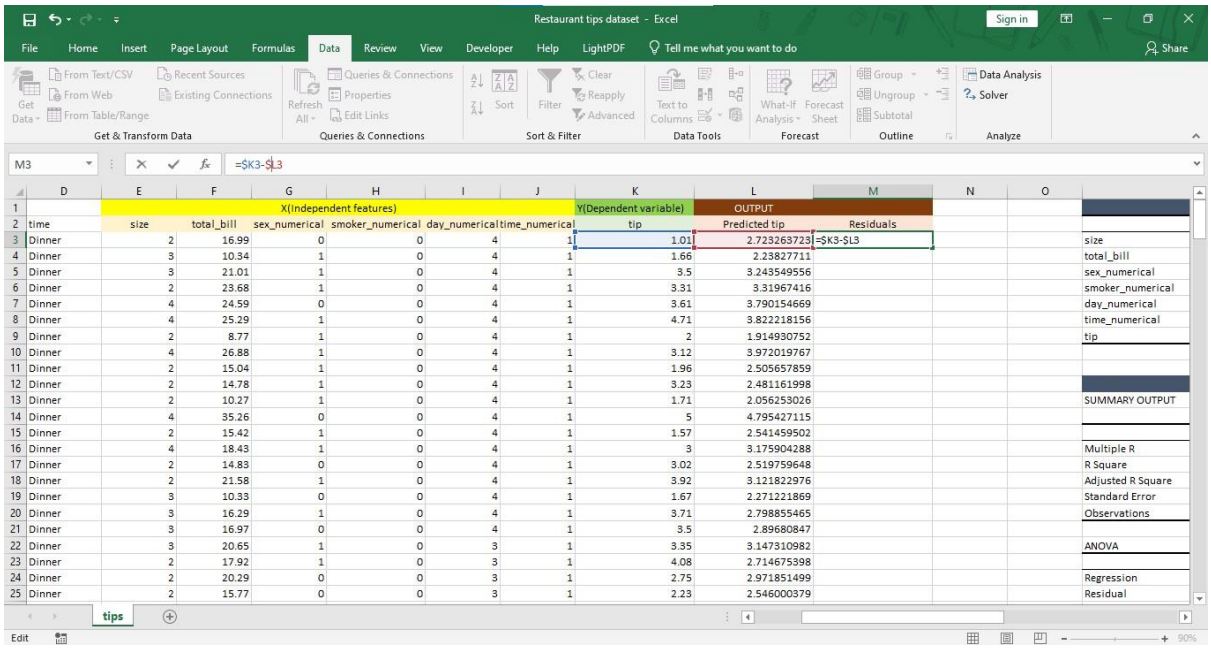
= { Intercept_coeff (P29) +
size_coeff(P30) * size(E3) +
total bill_coeff(P31) * total bill(F3) +
Sex_numerical_coeff(P32) * Sex_numerical(G3) +
Smoker_numerical_coeff(P33) * Smoker_numerical(H3) +
Day_numerical_coeff(P34) * Day_numerical(I3) +
Time_numerical_coeff(P35) * Time_numerical(J3) }

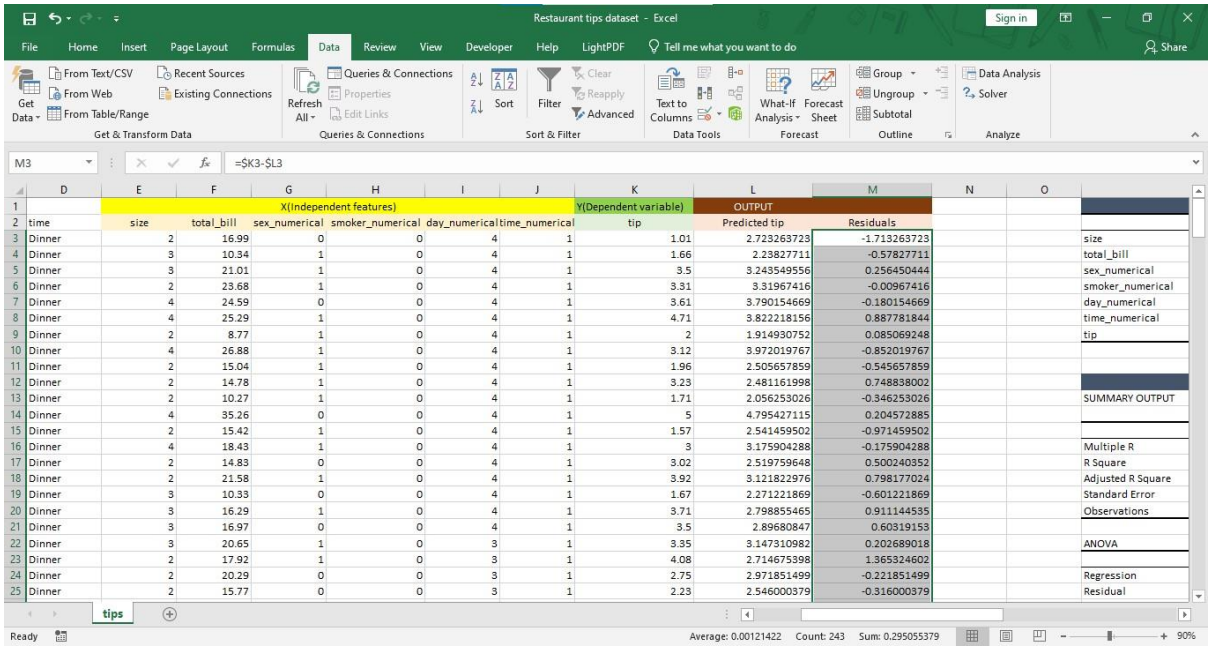




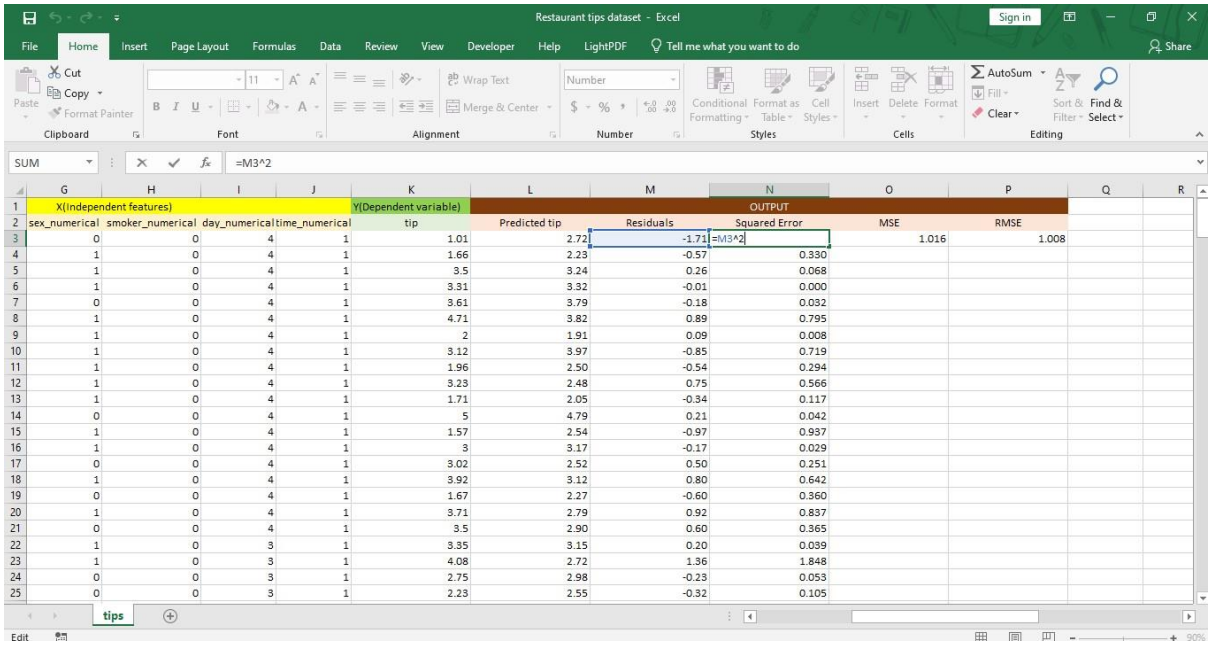
9. <u>Calculate the RMSE (Root Mean Square Error) of the model.</u>

To caluculate RMSE, first we need to calculate errors (Residuals).

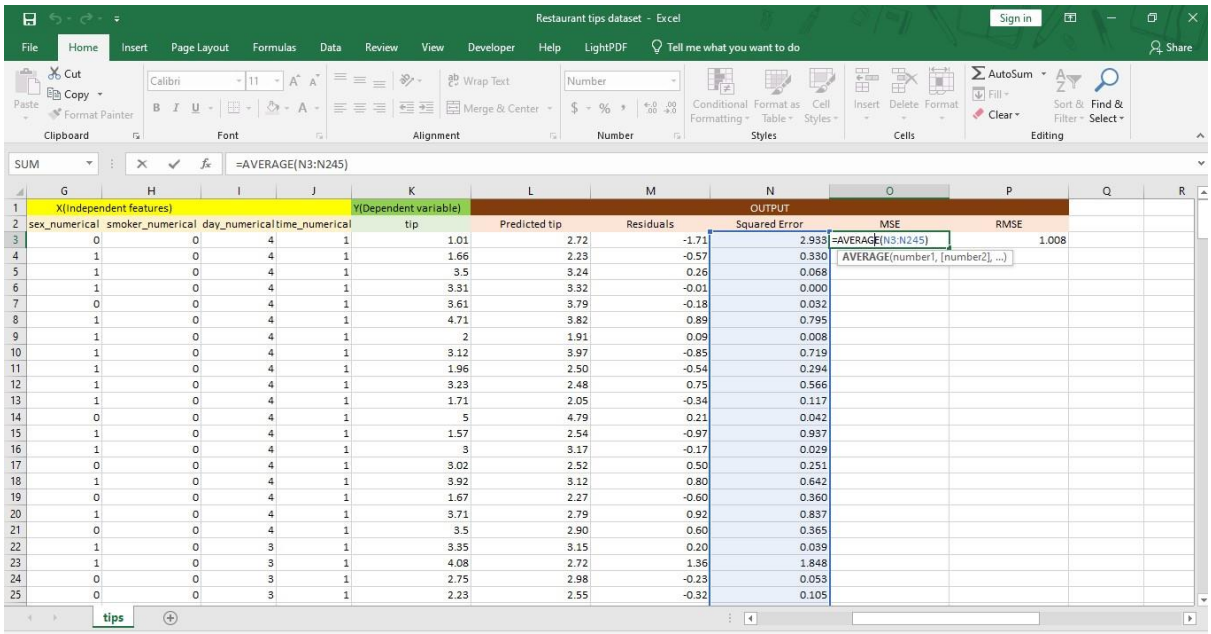




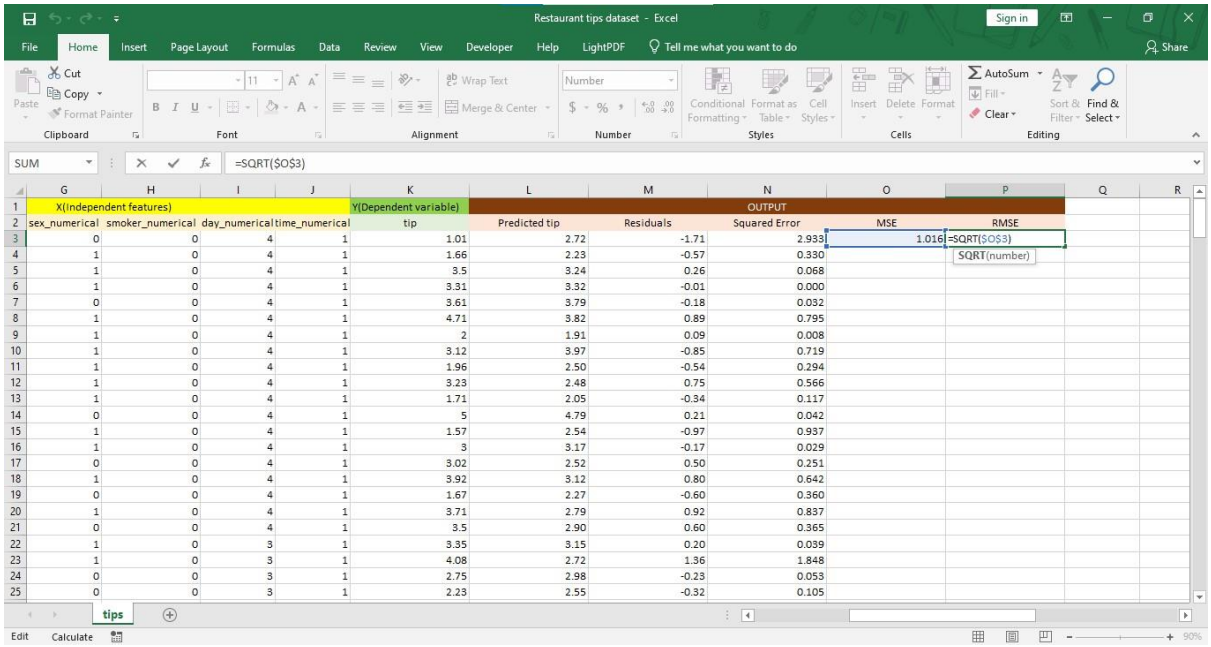
Now, we need to find square of these errors.



Now, we need to find average of these squared error to find Mean Squared Error (MSE).



Ultimately, RMSE is found taking square root of MSE.



	G	H	I	J	K	L	M	N	O	P	Q	R
1	X(Independent features)				Y(Dependent variable)	OUTPUT						
2	sex_numerical	smoker_numerical	day_numerical	time_numerical	tip	Predicted tip	Residuals	Squared Error	MSE	RMSE		
3	0	0	4	1	1.01	2.72	-1.71	2.933	1.016	=SQRT(O3)		
4	1	0	4	1	1.66	2.23	-0.57	0.330				
5	1	0	4	1	3.5	3.24	0.26	0.068				
6	1	0	4	1	3.31	3.32	-0.01	0.000				
7	0	0	4	1	3.61	3.79	-0.18	0.032				
8	1	0	4	1	4.71	3.82	0.89	0.795				
9	1	0	4	1	2	1.91	0.09	0.008				
10	1	0	4	1	3.12	3.97	-0.85	0.719				
11	1	0	4	1	1.96	2.50	-0.54	0.294				
12	1	0	4	1	3.23	2.48	0.75	0.566				
13	1	0	4	1	1.71	2.05	-0.34	0.117				
14	0	0	4	1	5	4.79	0.21	0.042				
15	1	0	4	1	1.57	2.54	-0.97	0.937				
16	1	0	4	1	3	3.17	-0.17	0.029				
17	0	0	4	1	3.02	2.52	0.50	0.251				
18	1	0	4	1	3.92	3.12	0.80	0.642				
19	0	0	4	1	1.67	2.27	-0.60	0.360				
20	1	0	4	1	3.71	2.79	0.92	0.837				
21	0	0	4	1	3.5	2.90	0.60	0.365				
22	1	0	3	1	3.35	3.15	0.20	0.039				
23	1	0	3	1	4.08	2.72	1.36	1.848				
24	0	0	3	1	2.75	2.98	-0.23	0.053				
25	0	0	3	1	2.23	2.55	-0.32	0.105				