Use MS Excel to predict restaurant tips

Submitted by: Mohit Rohilla

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.

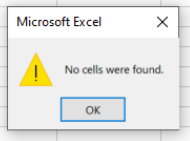
Tools required: Microsoft Excel, Data Analysis Add-in.

Expected Deliverables: Model to predict restaurant tips given input values with the mathematical equation for predicting the tips value.

Step1: To find if there are any missing values we can check for any blanks in the data by
Ctrl G >> special >> blanks
We have found no blank values in our data.

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
Male	No	Sat	Dinner	4	39.42	7.58
Male	No	Sat	Dinner	2	19.82	3.18

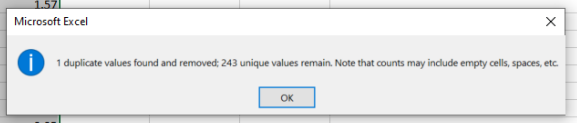
sex	Gender of the customer
smoker	Indicates if the custome
day	Day of the restaurant vis
time	Indicates whether the ti
size	Number of members dir
total bill	Bill amount in USD
tip	Tip amount in USD



Also to check for any duplicate values, we can select all the data and then go to:
Data >> Remove Duplicates

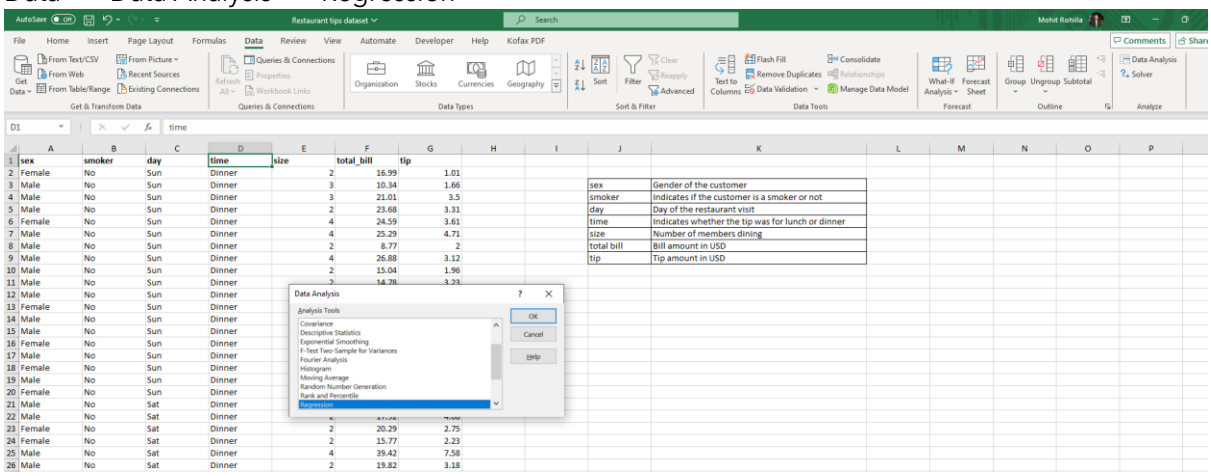
sex	smoker	day	time	size	total_bill	tip
Female	No	Sun	Dinner	2	16.99	1.01
Male	No	Sun	Dinner	3	10.34	1.66
Male	No	Sun	Dinner	3	21.01	3.5
Male	No	Sun	Dinner	2	23.68	3.31
Female	No	Sun	Dinner	4	24.59	3.61
Male	No	Sun	Dinner	4	25.29	4.71
Male	No	Sun	Dinner	2	8.77	2
Male	No	Sun	Dinner	4	26.88	3.12
Male	No	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	
Female	No	Sun	Dinner	2	14.83	
Male	No	Sun	Dinner	2	21.58	
Female	No	Sun	Dinner	3	10.33	
Male	No	Sun	Dinner	3	16.29	
Female	No	Sun	Dinner	3	16.97	
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
Male	No	Sat	Dinner	4	39.42	7.58
Male	No	Sat	Dinner	2	19.82	3.18
Male	No	Sat	Dinner	4	17.81	2.34
Male	No	Sat	Dinner	2	13.37	2
Male	No	Sat	Dinner	2	12.69	2
Male	No	Sat	Dinner	2	21.7	4.3
Female	No	Sat	Dinner	2	19.65	3
Male	No	Sat	Dinner	2	9.55	1.45
Male	No	Sat	Dinner	4	18.35	2.5

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

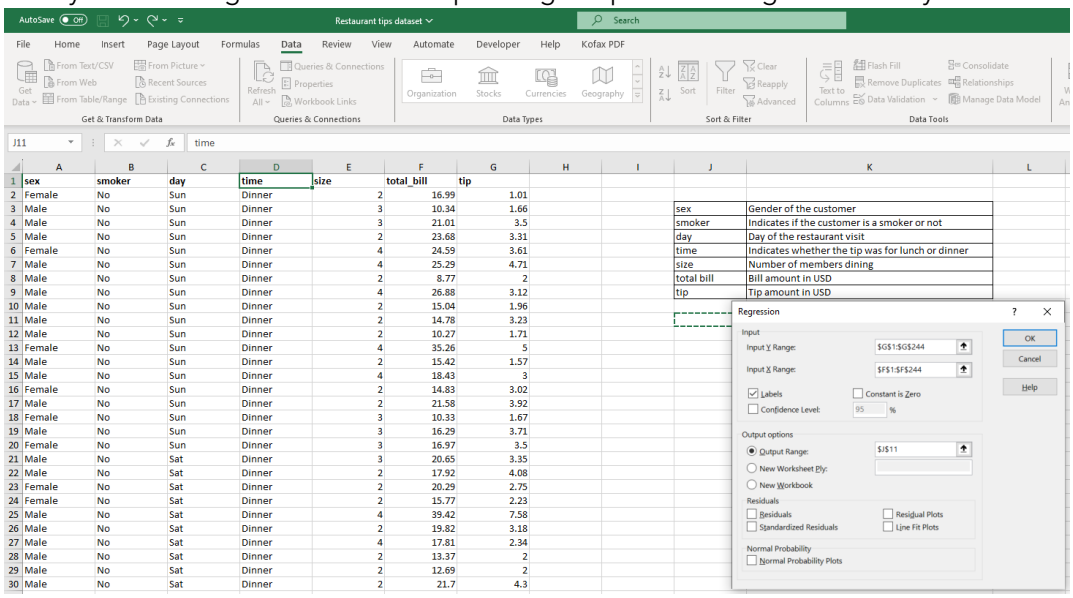


We have found 1 duplicate value which we have removed and 243 unique values. Now with this data we can move to our analysis

Step 2: Now we can create a regression model by got to:
Data >> Data Analysis >> Regression



A regression dialogue box will open, under this dialogue box we can enter Tip Data i.e. G1:G244 under Input Y Range, and Bill Data i.e. F1:F244 under Input X range, and tick under the checkbox label as we are all selecting the labels.
Finally we can assign a cell under output range to perform a regression analysis:



We got the result in which we got coefficient values:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.674997857
R Square	0.455622106
Adjusted R Squa	0.453363277
Standard Error	1.023999665
Observations	243

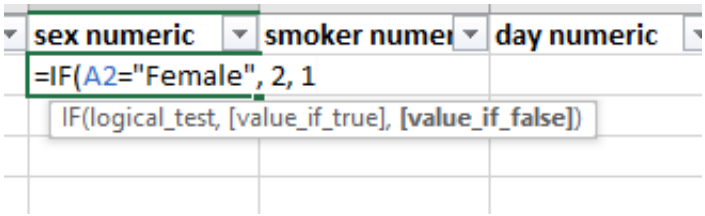
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	211.5051652	211.5051652	201.707176	1.13521E-33
Residual	241	252.7066504	1.048575313		
Total	242	464.2118156			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.923451775	0.160443391	5.755623649	2.61235E-08	0.607401364	1.239502185	0.607401364	1.239502185
total_bill	0.10492323	0.007387729	14.20236516	1.13521E-33	0.090370465	0.119475994	0.090370465	0.119475994

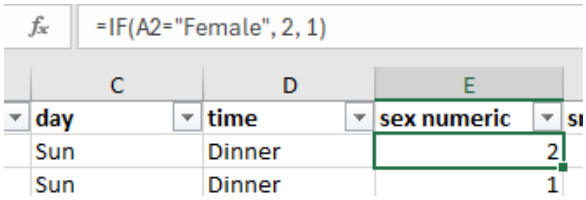
However, the challenge is we have other independent variables as well. As Sex, Smoker, Day, and Time. The challenge is these values are non-numeric values and to prepare a good model we must consider these independent variables as well.
Therefore, we consider:
Under Sex Category: We consider Male as 1, and female as 2.
Under Smoker category: We consider Smoker as 1, and Non-smoker as 2
Under Day Category: We consider Thursday as 1, Friday as 2, Saturday as 3, and Sunday as 4
Under Time Category: We consider Lunch as 1, and Dinner as 2
And the size will remain the same as this is not a non-numeric category.

We can do all the task by using formula if. For eg: to change the Sex in numeric values we can use the formula:
=IF(A2="Female", 2, 1)

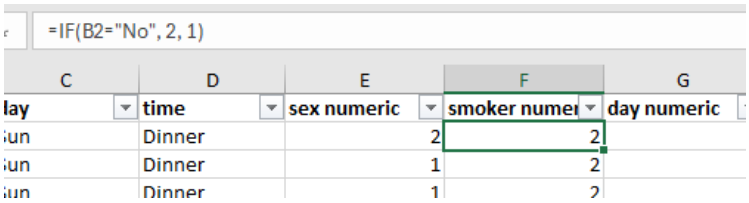


The same concept can be used to convert all the independent values mention above into the numeric form.
The formulas for each category can be used as:

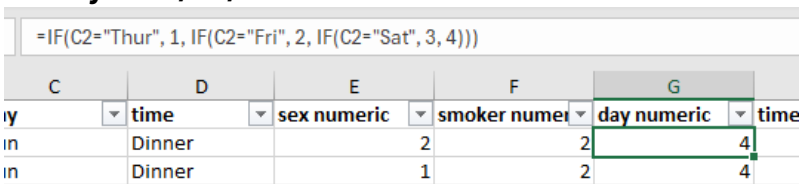
For sex: Male or Female



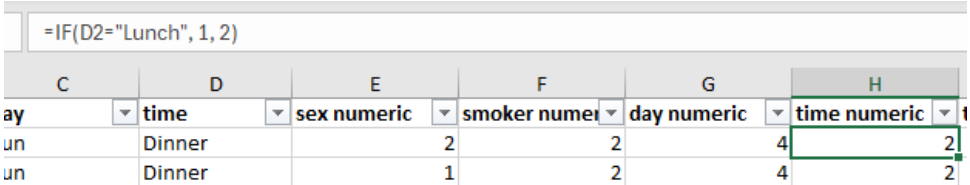
For smoker: Yes or No



For Day: Thur, Fri, Sat or Sun



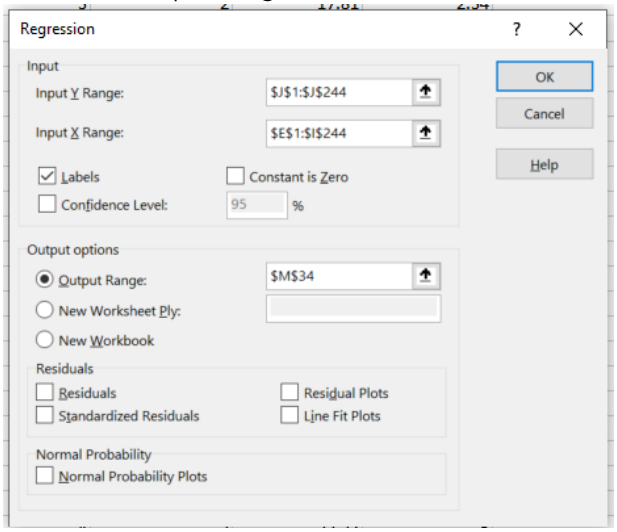
For time: Lunch or Dinner



Step 3: Now we can create a regression model by consider all the independent values i.e. sex, smoker, day, time, number and bill amount.

To create the regression model go to:
Data >> Data Analysis >> Regression
Add tip values in the Input Y range i.e. J1:J244
Add tip values in the Input X range i.e. E1:I244

Check the output range:



The results came out to be:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.677885176
R Square	0.459528311
Adjusted R Squ	0.448125955
Standard Error	1.028893437
Observations	243

ANOVA

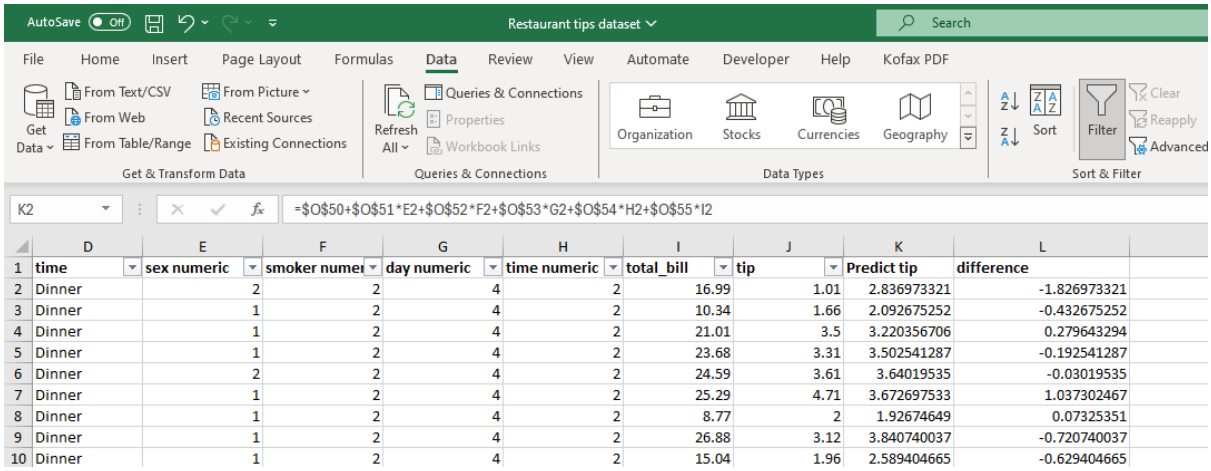
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	213.3184718	42.66369436	40.30117104	6.71789E-30
Residual	237	250.8933439	1.058621704		
Total	242	464.2118156			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.731285889	0.480954464	1.520488827	0.12972115	-0.216205953	1.67877773	-0.216205953	1.67877773
sex numeric	0.041478793	0.14247047	0.291139583	0.771199201	-0.239191453	0.32214904	-0.239191453	0.32214904
smoker numeri	0.130211847	0.13879624	0.93815111	0.349121831	-0.143220079	0.403643773	-0.143220079	0.403643773
day numeric	0.08592815	0.120115741	0.71537793	0.475079703	-0.150702743	0.322559044	-0.150702743	0.322559044
time numeric	-0.188515216	0.308147679	-0.611769062	0.541276972	-0.795573543	0.41854311	-0.795573543	0.41854311
total_bill	0.105687109	0.007619616	13.87039751	2.00082E-32	0.090676282	0.120697936	0.090676282	0.120697936

We have obtained intercept coefficient and independent variable coefficients in the regression model:

	Coefficients
Intercept	0.731285889
sex numeric	0.041478793
smoker numeri	0.130211847
day numeric	0.08592815
time numeric	-0.188515216
total_bill	0.105687109

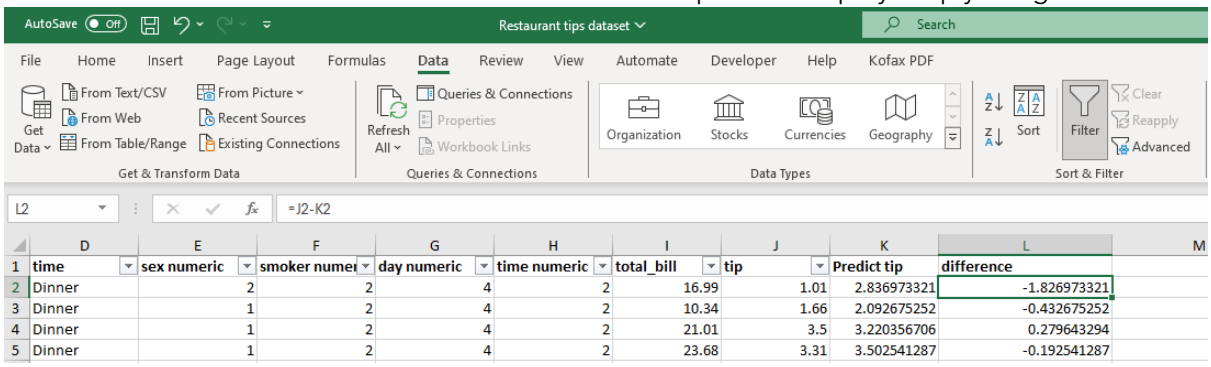
Now we can calculate the predicted tip by using the formula:
=$N$50+$N$51*E2+$N$52*F2+$N$53*G2+$N$54*H2+$N$55*I2



	D	E	F	G	H	I	J	K	L
1	time	sex numeric	smoker numer	day numeric	time numeric	total_bill	tip	Predict tip	difference
2	Dinner	2	2	4	2	16.99	1.01	2.836973321	-1.826973321
3	Dinner	1	2	4	2	10.34	1.66	2.092675252	-0.432675252
4	Dinner	1	2	4	2	21.01	3.5	3.220356706	0.279643294
5	Dinner	1	2	4	2	23.68	3.31	3.502541287	-0.192541287
6	Dinner	2	2	4	2	24.59	3.61	3.64019535	-0.03019535
7	Dinner	1	2	4	2	25.29	4.71	3.672697533	1.037302467
8	Dinner	1	2	4	2	8.77	2	1.92674649	0.07325351
9	Dinner	1	2	4	2	26.88	3.12	3.840740037	-0.720740037
10	Dinner	1	2	4	2	15.04	1.96	2.589404665	-0.629404665

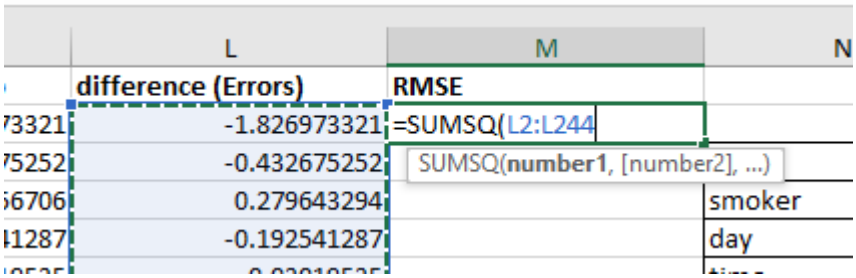
Where we have fixed the values of coefficient.

We can now calculate the difference between actual and predicted tip by simply using the formula:

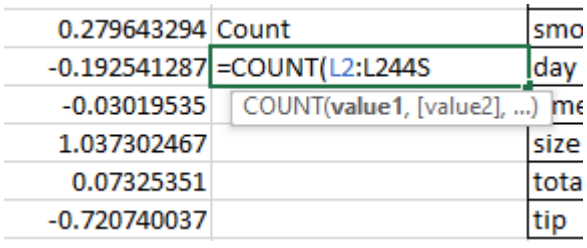


	D	E	F	G	H	I	J	K	L	M
1	time	sex numeric	smoker numer	day numeric	time numeric	total_bill	tip	Predict tip	difference	
2	Dinner	2	2	4	2	16.99	1.01	2.836973321	-1.826973321	
3	Dinner	1	2	4	2	10.34	1.66	2.092675252	-0.432675252	
4	Dinner	1	2	4	2	21.01	3.5	3.220356706	0.279643294	
5	Dinner	1	2	4	2	23.68	3.31	3.502541287	-0.192541287	

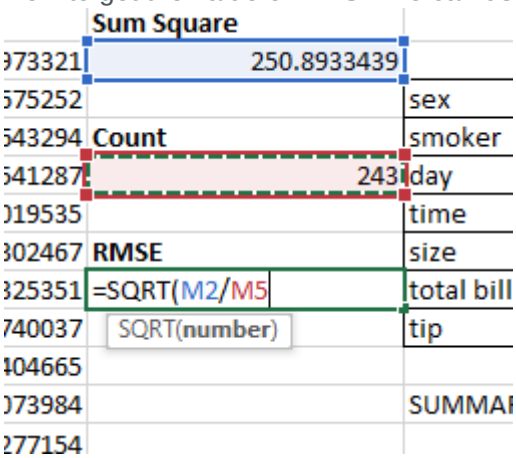
Step 4: The RMSE is root of mean of square errors. The formula for the same is under route of sum of square of all the errors divided by the number of errors. To calculate the RMSE, we can use the formula:
SUMSQ to calculate the sum squares of all the errors i.e. =SUMSQ(L2:L244)



Count function to calculate all the errors. i.e. =COUNT(L2:L244S), which is obviously 243



Now to get the value of RMSE we can use the formula SQRT i.e. =SQRT(M2/M5)



We get the value of **RMSE as 1.016111656**