Course End Project

Predicting Restaurant Tips

By Parth Satish Shah, September 2024

Problem Statement:

In this project, you must analyse a dataset containing restaurant tips data. You aim to clean the data, identify the independent and dependent variables, and build a predictive model to estimate tip amounts. You will use regression analysis to encode categorical variables into numeric values. The project aims to provide insights that can help improve service and customer satisfaction in the restaurant industry.

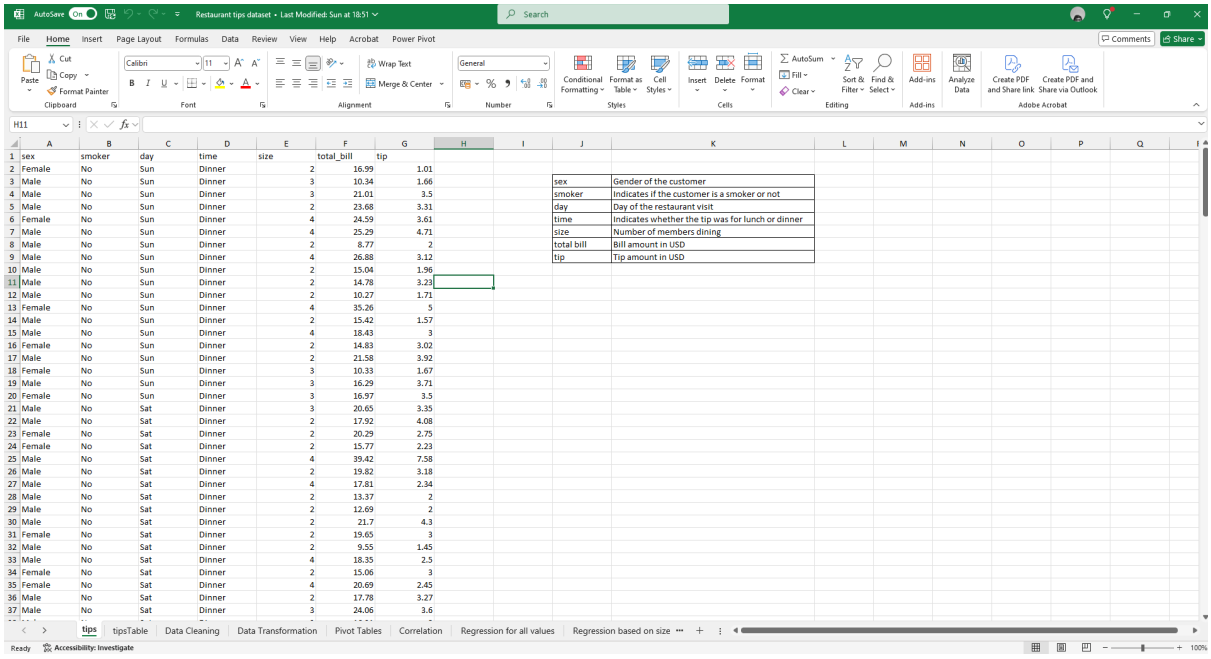
Input Dataset:

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

sex	The 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

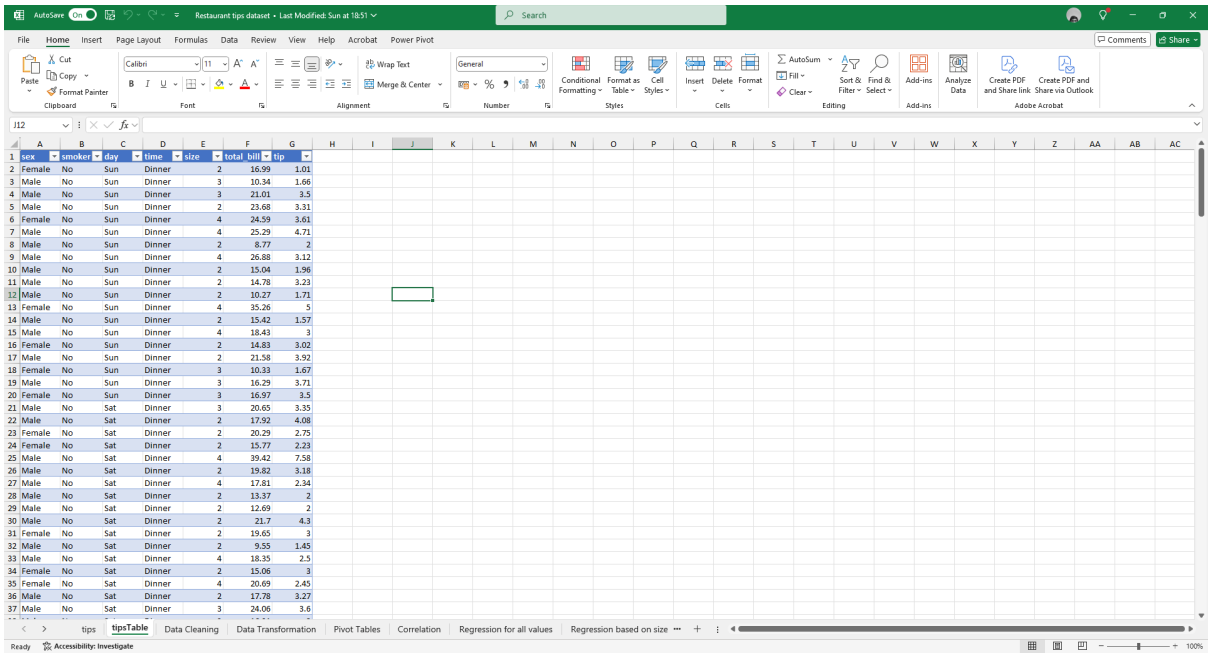
Actions

1. Open the dataset input file: **Restaurant tips dataset.xlsx**

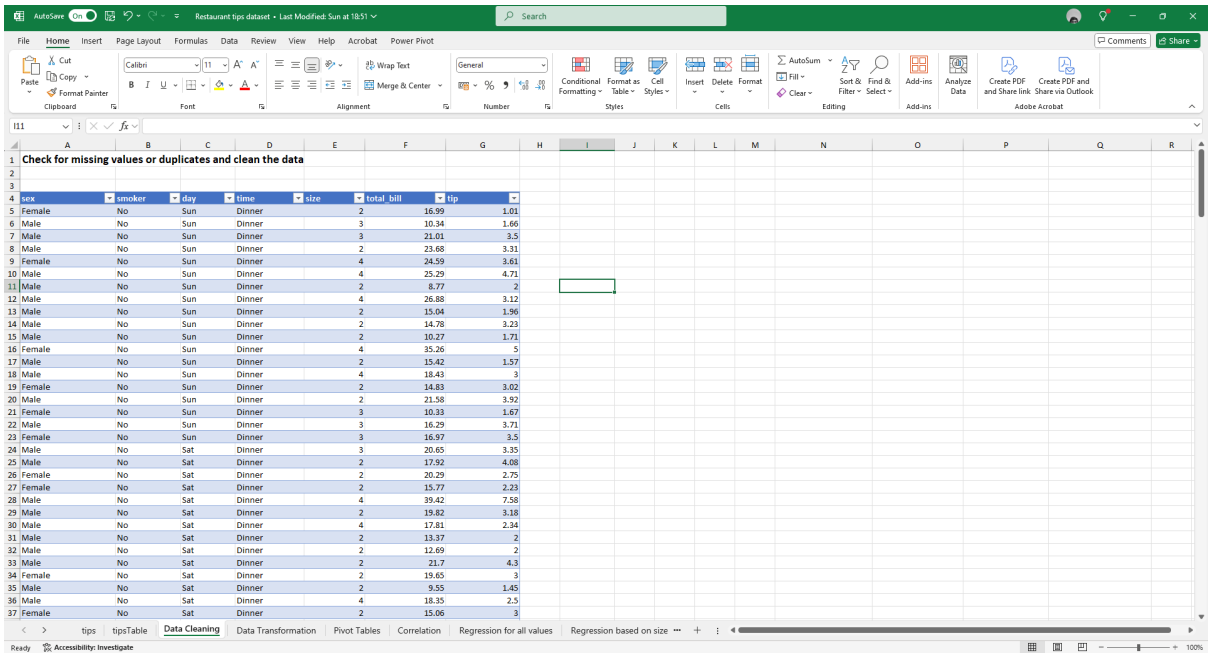


sex	Gender of the customer
smoker	Indicates if the customer is a smoker or not
day	Day of the restaurant visit
time	Indicates whether the tip was for lunch or dinner
size	Number of members dining
total bill	Bill amount in USD
tip	Tip amount in USD

The above data was copied and pasted in tipsTable sheet where it was converted into a table.



2. Check for missing values or duplicates and clean the data

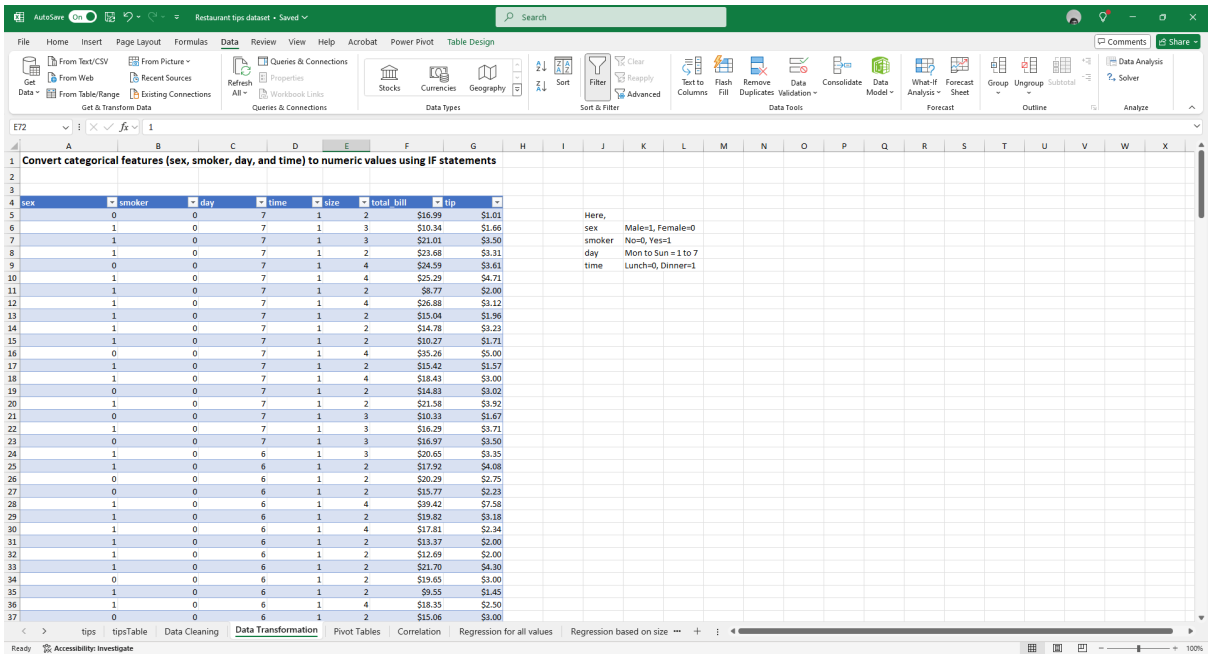


Check for missing values or duplicates and clean the data

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
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
Female	No	Sat	Dinner	2	15.06	3

Duplicate values were removed by using **Alt + A + M** which is a shortcut for going to **Data Tab > Data Tools Section > Remove Duplicates**

3. Convert categorical features (sex, smoker, day, and time) to numeric values using IF statements



Categorical values were transformed to numerical values using IF function.

a. sex

=IF(tipsTable[sex]="Female",0,1)

b. smoker

=IF(tipsTable[smoker]="No",0,1)

c. day

=IF(tipsTable[day]="Mon",1,IF(tipsTable[day]="Tue",2,IF(tipsTable[day]="Wed",3,IF(tipsTable[day]="Thur",4,IF(tipsTable[day]="Fri",5,IF(tipsTable[day]="Sat",6,IF(tipsTable[day]="Sun",7)))))))

d. time

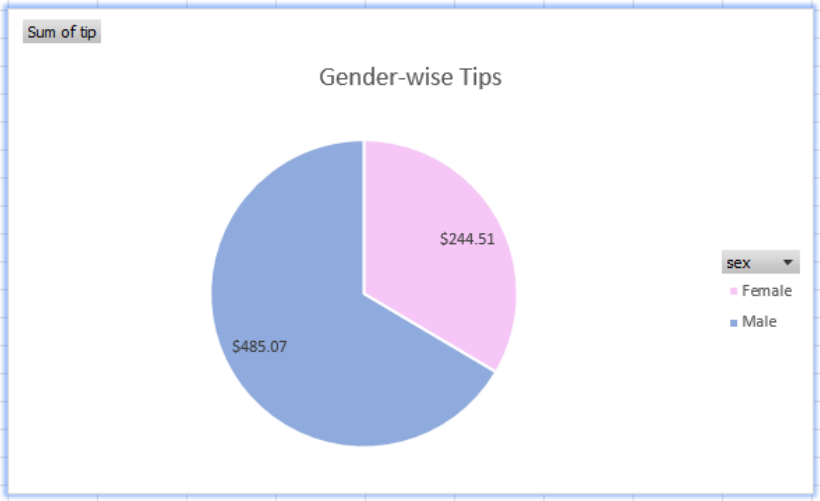
=IF(tipsTable[time]="Lunch",0,1)

total_bill and tip values were also converted to currency format with $ sign and commas.

4. Create a pivot table to sum the tips for males and females separately, then create a pie chart

Create a pivot table to sum the tips for males and females separately, then create a pie chart

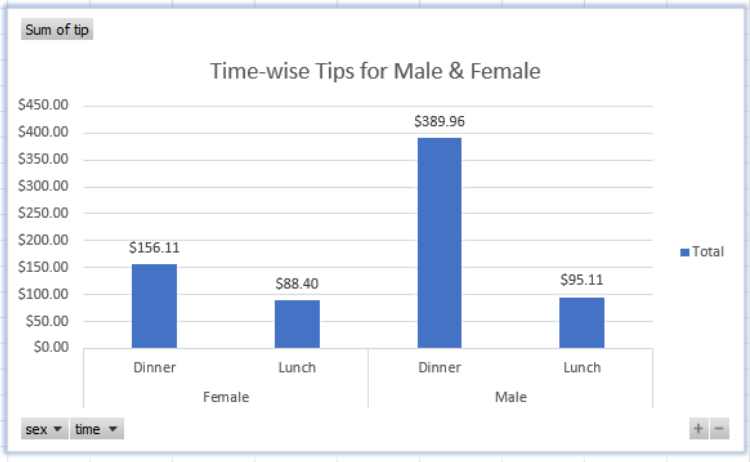
Row Labels	Sum of tip
Female	$244.51
Male	$485.07
Grand Total	**$729.58**



5. Create another pivot table to sum the tips for males and females at various times, and then create a bar chart

Create another pivot table to sum the tips for males and females at various times, and then create a bar chart

Row Labels	Sum of tip
Female	**$244.51**
Dinner	$156.11
Lunch	$88.40
Male	**$485.07**
Dinner	$389.96
Lunch	$95.11
Grand Total	**$729.58**



Outcome: We can see, sum of tip by men is more than sum of tip by women. And overall tips are higher during the dinner time.

6. The tip feature is the dependent variable, while the rest are independent variables. Find the correlation between the tips and the other variables.

The tip feature is the dependent variable, while the rest are independent variables. Find the correlation between the tips and the other variables.

Correlation between tip and sex

	sex	*tip*
sex	1	
tip	8.53%	1

Correlation between tip and smoker

	smoker	*tip*
smoker	1	
tip	0.98%	1

Correlation between tip and day

	day	*tip*
day	1	
tip	13.18%	1

Correlation between tip and time

	time	*tip*
time	1	
tip	11.76%	1

Correlation between tip and size

	size	*tip*
size	1	
tip	48.84%	1

Correlation between tip and total_bill

	total_bill	*tip*
total_bill	1	
tip	67.50%	1

Correlations between all values of the table

	sex	*smoker*	*day*	*time*	*size*	*total_bill*	*tip*
sex	100.00%						
smoker	0.99%	100.00%					
day	22.44%	-2.50%	100.00%				
time	19.81%	6.39%	87.31%	100.00%			
size	8.32%	-13.06%	16.25%	10.00%	100.00%		
total_bill	14.13%	9.01%	17.00%	17.92%	59.76%	100.00%	
tip	8.53%	0.98%	13.18%	11.76%	48.84%	67.50%	100.00%

Here I used the Data Analysis add-in in Excel for finding the correlation between tip and other variables. I also created a common correlation table showing relations between all variables.

Outcome: There wasn't much correlation between most variables but the relation between **total_bill** and **tip** showed some promise (although just 67.50%, it was still better compared to other relations). **size** variable is also significantly better than sex, smoker, day and time variables.

Data Tab > Analyze Section > Data Analysis > Correlation

7. The problem is a multiple regression problem. Use the data analysis add-in for regression.

	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.684226461								
5	R Square	0.46816585								
6	Adjusted R Square	0.454644643								
7	Standard Error	1.022798786								
8	Observations	243								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	6	217.3281192	36.22135321	34.62455999	6.98847E-30				
13	Residual	236	246.8836964	1.046117358						
14	Total	242	464.2118156							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
17	Intercept	0.51606503	0.525825377	0.981438045	0.327381815	-0.519846116	1.551976177	-0.519846116	1.551976177	
18	sex	-0.036632515	0.141648177	-0.25861621	0.796156785	-0.315688894	0.242423864	-0.315688894	0.242423864	
19	smoker	-0.072844331	0.141051329	-0.516438456	0.606031781	-0.350724879	0.205036218	-0.350724879	0.205036218	
20	day	0.053005305	0.120582605	0.439576712	0.66064596	-0.184550488	0.290561098	-0.184550488	0.290561098	
21	time	-0.116006262	0.308553216	-0.375968409	0.707278124	-0.723876724	0.491864201	-0.723876724	0.491864201	
22	size	0.174964371	0.089368965	1.957775507	0.051434839	-0.001098465	0.351027206	-0.001098465	0.351027206	
23	total_bill	0.094263519	0.009561371	9.858786532	2.01E-19	0.075426978	0.113100059	0.075426978	0.113100059	
24										
25										
26										
27	RESIDUAL OUTPUT				PROBABILITY OUTPUT					
28										
29	*Observation*	*Predicted tip*	*Residuals*		*Percentile*	*tip*				
30	1	2.722561826	-1.712561826		0.205761317	1				
31	2	2.234041283	-0.574041283		0.617283951	1				
32	3	3.239833026	0.260166974		1.028806584	1				
33	4	3.316552251	-0.006552251		1.440329218	1				
34	5	3.788893308	-0.178893308		1.851851852	1.01				
35	6	3.818245257	0.891754743		2.263374486	1.1				
36	7	1.911083188	0.088916812		2.674897119	1.17				
37	8	3.968124251	-0.848124251		3.086419753	1.25				

I used the data analysis add-in in Excel to get more insights from the table.

Outcome: As seen in the image above, the **p-value** of only 2 variables seem to be **<= 0.05** and hence I decided to go ahead with these 2 variables for my further analysis.

Usually, we only take values less than 0.05 but here, since we don't have a large test data, I have decided to include the 'size' variable as well which has the p-value of 0.0514.

8. Use coefficients and intercepts to form the regression equation

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.683277523							
5	R Square	0.466868174							
6	Adjusted R Square	0.462425409							
7	Standard Error	1.01547627							
8	Observations	243							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	216.7257226	108.3628613	105.085043	1.66017E-33			
13	Residual	240	247.486093	1.031192054					
14	Total	242	464.2118156						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	0.672163792	0.194391852	3.4577776	0.000644182	0.289231742	1.055095841	0.289231742	1.055095841
18	size	0.192346316	0.085486043	2.250031813	0.025353509	0.023947562	0.36074507	0.023947562	0.36074507
19	total_bill	0.092637395	0.009137207	10.13848061	2.46028E-20	0.074638033	0.110636757	0.074638033	0.110636757
20									
21									
22									
23	RESIDUAL OUTPUT					PROBABILITY OUTPUT			
24									
25	*Observation*	*Predicted tip*	*Residuals*	*Standard Residuals*		*Percentile*	*tip*		
26	1	2.630765768	-1.620765768	-1.602701113		0.205761317	1		
27	2	2.207073406	-0.547073406	-0.540975861		0.617283951	1		
28	3	3.195514412	0.304485588	0.301091867		1.028806584	1		
29	4	3.250509941	0.059490059	0.058826997		1.440329218	1		
30	5	3.719502603	-0.109502603	-0.108282114		1.851851852	1.01		
31	6	3.784348779	0.925651221	0.915334141		2.263374486	1.1		
32	7	1.869286379	0.130713621	0.129256719		2.674897119	1.17		
33	8	3.931642238	-0.811642238	-0.802595874		3.086419753	1.25		
34	9	2.450122847	-0.490122847	-0.484660059		3.497942387	1.25		
35	10	2.426037124	0.803962876	0.795002104		3.909465021	1.25		
36	11	2.008242472	-0.298242472	-0.294918335		4.320987654	1.32		
37	12	4.707943609	0.292056391	0.288801203		4.732510288	1.36		

I further went ahead with another regression round, this time only including '**size**' and '**total_bill**' variables as their p-values were <=0.05.

Outcome: As seen above, the P-value is further reduced, which makes our model even more accurate.

Regression Equation:

Based on our analysis, the regression equation is:

$$0.192346316 * \text{size} + 0.092637395 * \text{total\_bill} + 0.672163792$$

9. Run the regression equation for all input values to get the predicted y

Tip Predictor		
Size	Total Bill	Tip (in $)
3	21.01	3.195514412

Outcome: Here, **Tip** cell is calculated based on the equation which we got in the previous step and plugging in the **Size** cell and **Total Bill** cell values in the equation.