**Use MS Excel to predict restaurant tips**

Submitted by: Deepanshu Verma

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.

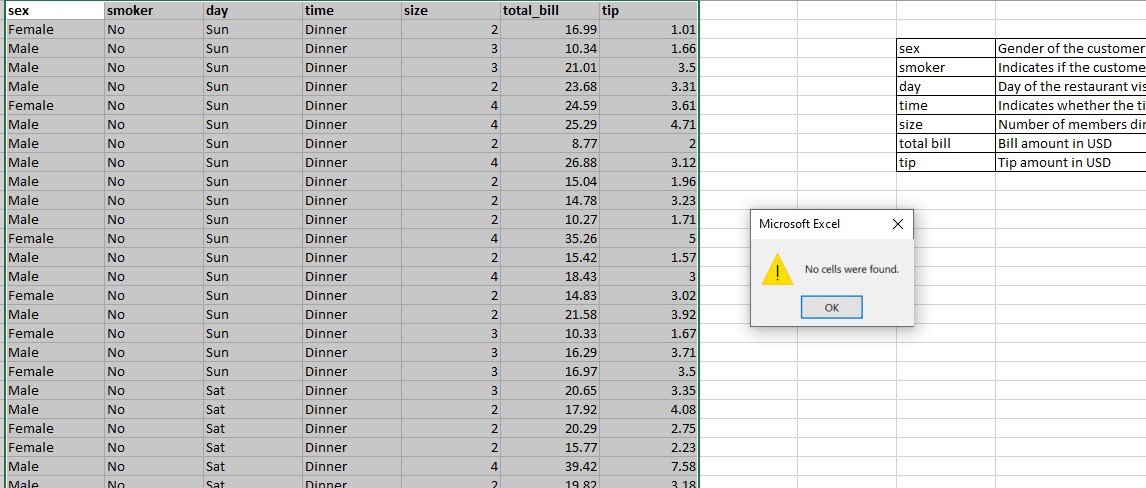
Tools required: Microsoft Excel, Data Analysis Add-in.

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

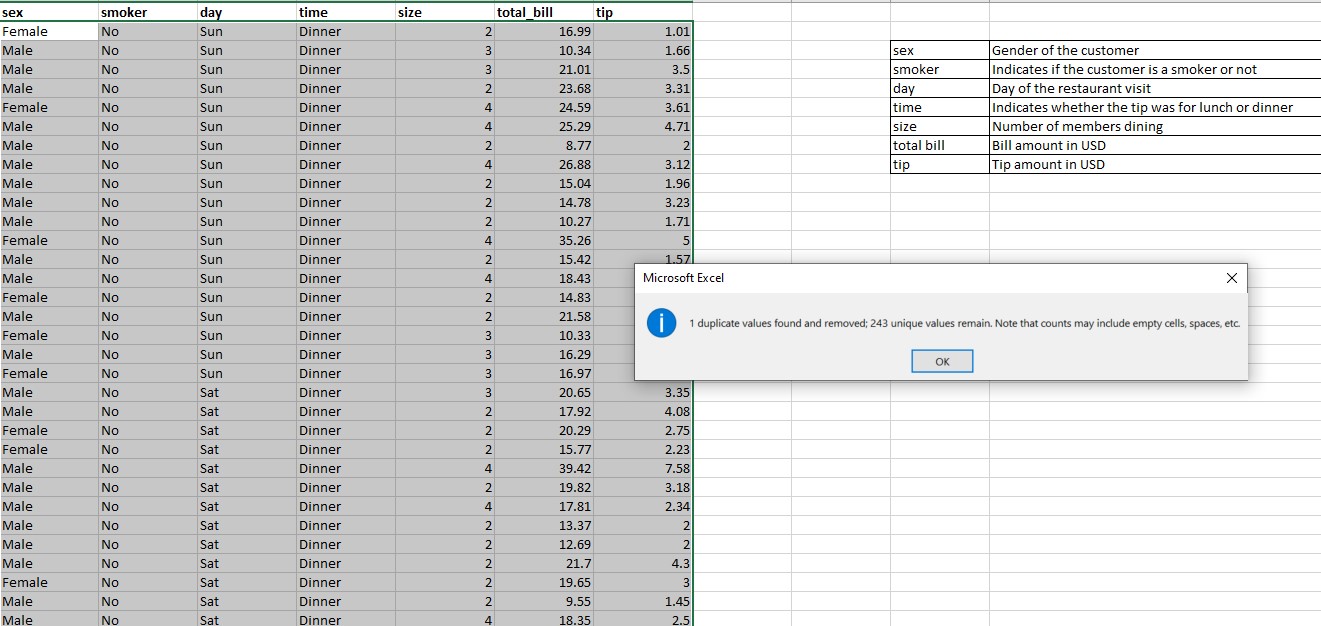
Step1: To find if there are any missing values we can check for any blanks in the data by

Ctrl G >> special >> blanks

We have found no blank values in our data.



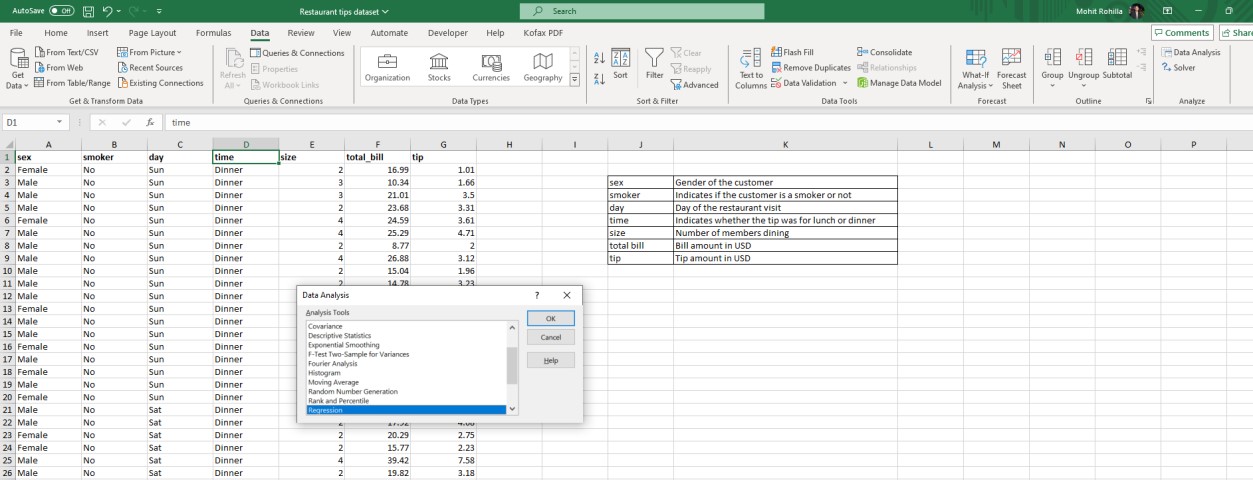
Also to check for any duplicate values, we can select all the data and then go to: Data >> Remove Duplicates



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

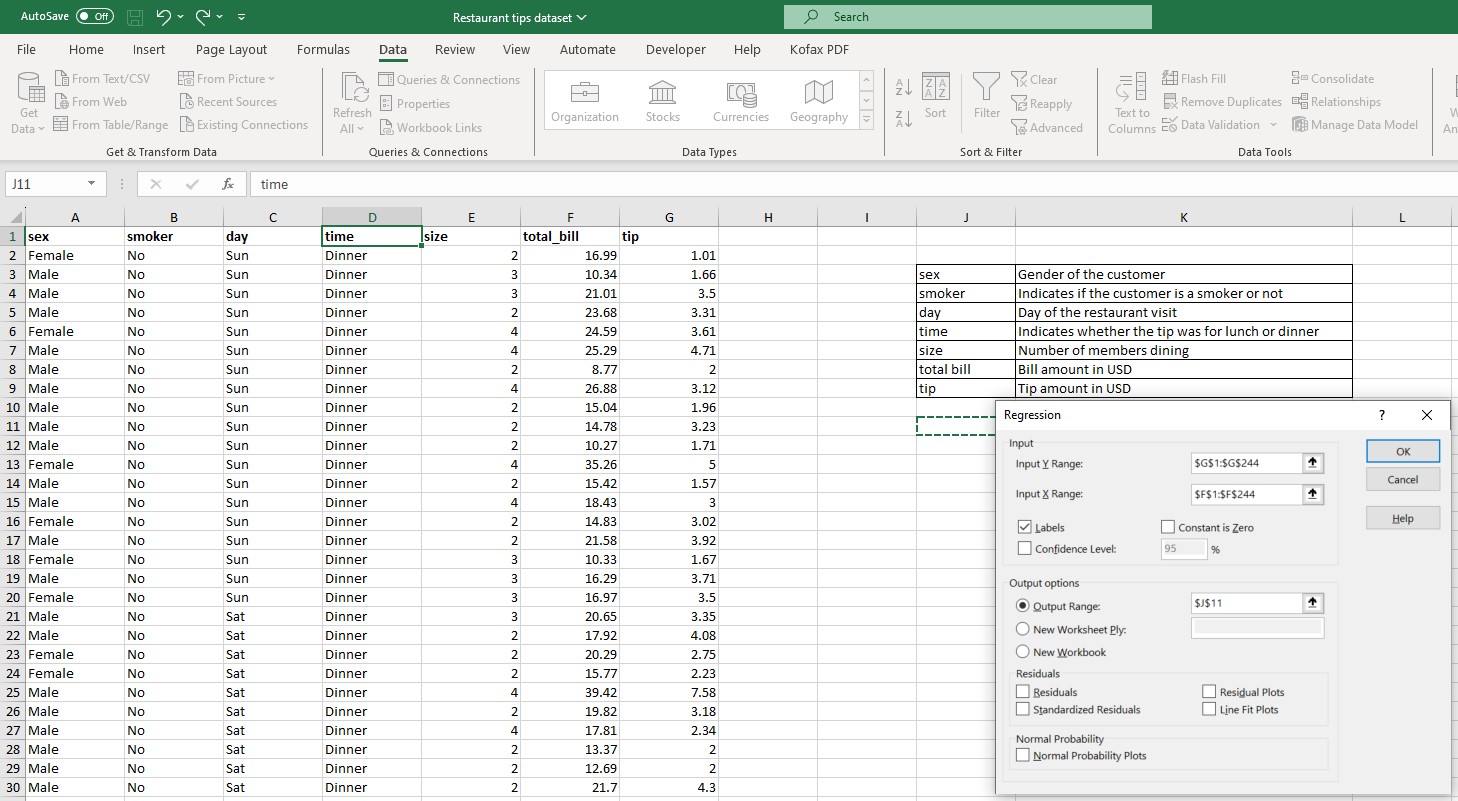
**Step 2:** Now we can create a regression model by got to:

Data >> Data Analysis >> Regression

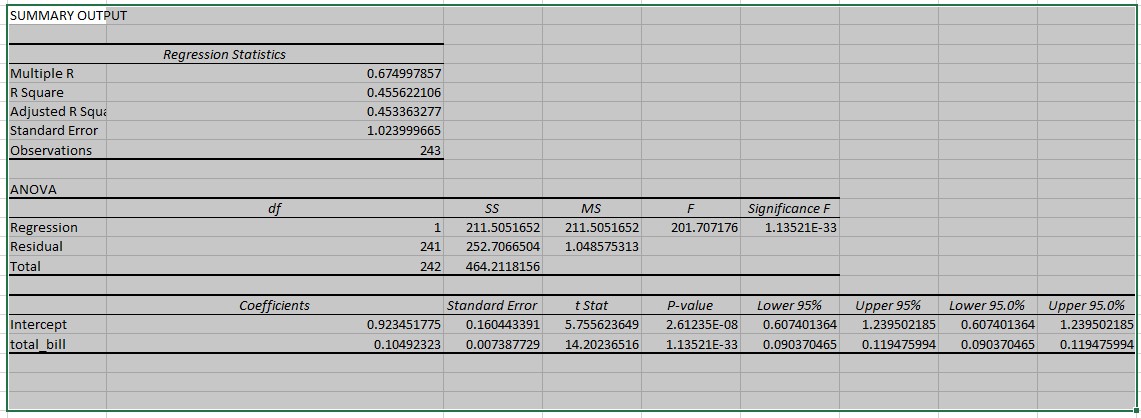


A regression dialogue box will open, under this dialogue box we can enter Tip Data i.e. G1:G244 under Input Y Range, and Bill Data i.e. F1:F244 under Input X range, and tick under the checkbox label as we are all selecting the labels.

Finally we can assign a cell under output range to perform a regression analysis:



We got the result in which we got coefficient values:



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.

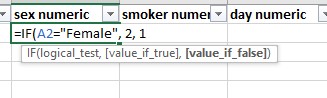
Under Smoker category: We consider Smoker as 1, and Non-smoker as 2

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:

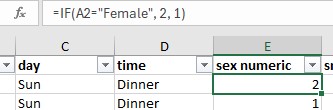
=IF(A2="Female", 2, 1)



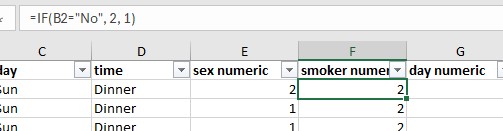
The same concept can be used to convert all the independent values mention above into the numeric form.

The formulas for each category can be used as:

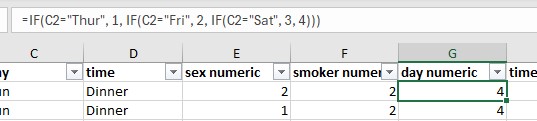
# For sex: Male or Female



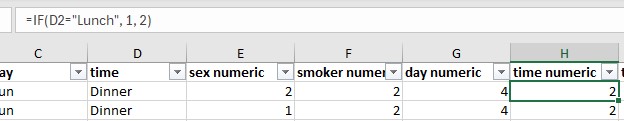
# For smoker: Yes or No



## 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.

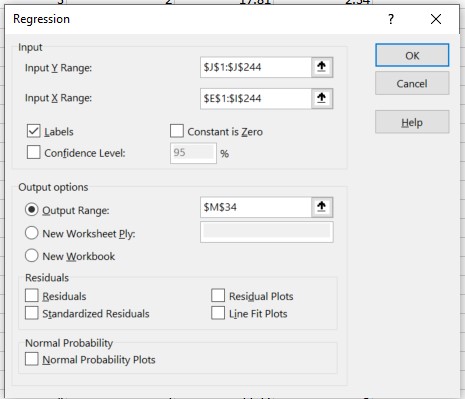
To create the regression model go to:

Data >> Data Analysis >> Regression

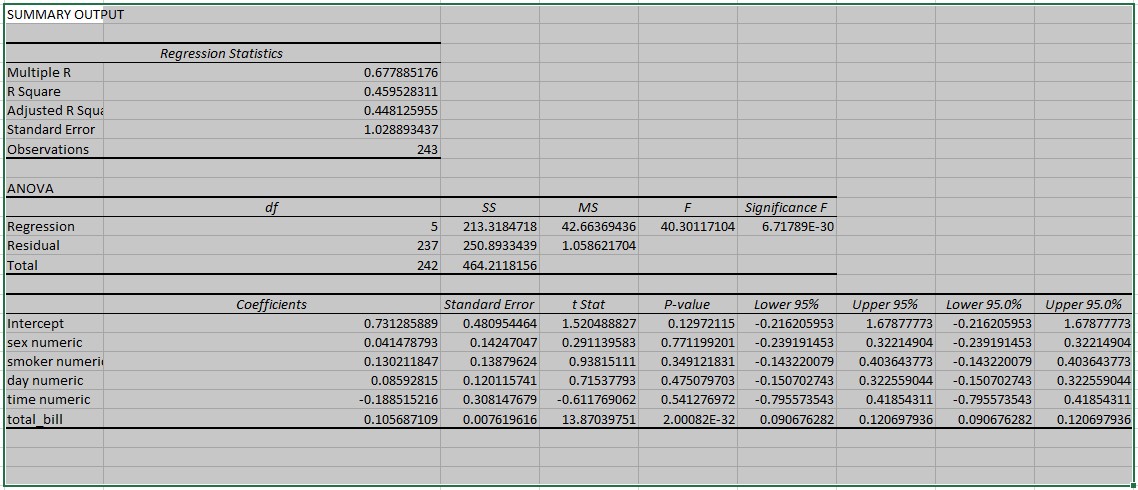
Add tip values in the Input Y range i.e. J1:J244

Add tip values in the Input X range i.e. E1:I244

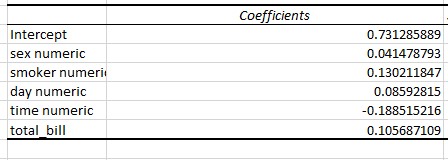
Check the output range:



The results came out to be:

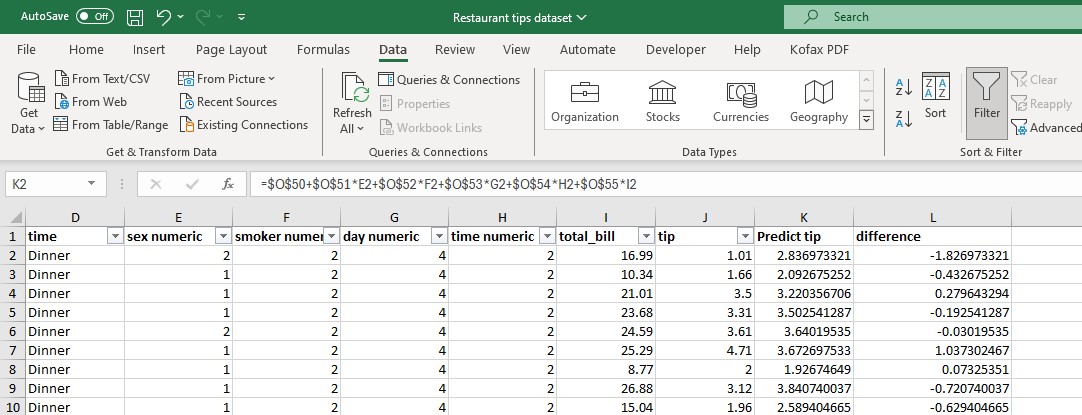


We have obtained intercept coefficient and independent variable coefficients in the regression model:



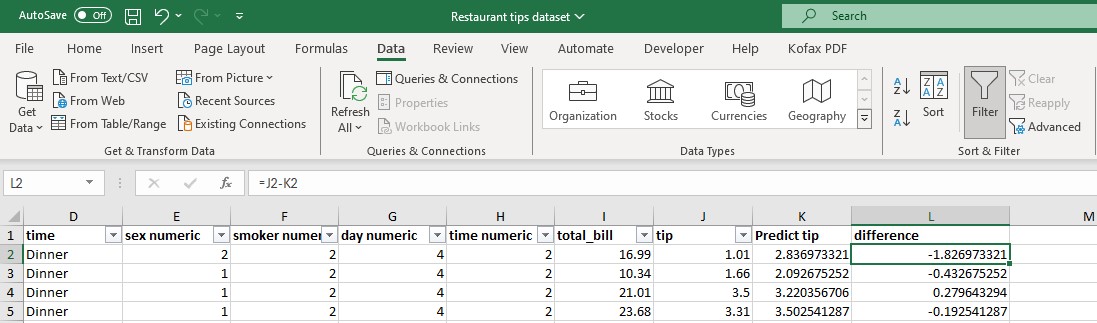
Now we can calculate the predicted tip by using the formula:

=$N$50+$N$51\*E2+$N$52\*F2+$N$53\*G2+$N$54\*H2+$N$55\*I2



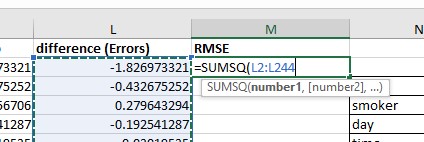
Where we have fixed the values of coefficient.

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

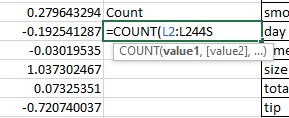


**Step 4:** The RMSE is root of mean of square errors. The formula for the same is under route of sum of square of all the errors divided by the number of errors. To calculate the RMSE, we can use the formula:

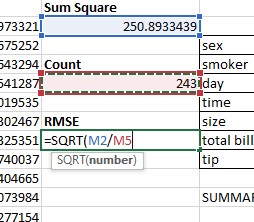
SUMSQ to calculate the sum squares of all the errors i.e. =SUMSQ(L2:L244)



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**