

PROJECT WRITE-UP:

PREDICTING RESTAURANT TIPS USING EXCEL.

Objectives:

The goal of this project is to predict the tips left by customers at a restaurant based on several influencing factors. By analyzing the provided dataset, we aim to identify the relationship between these factors and tips, build a predictive model, and evaluate its performance using statistical metrics. Microsoft Excel will be the primary tool for performing the tasks, leveraging its built-in functions, formulas, and the Data Analysis Add-in.

Dataset Description:

The dataset, *Restaurant Tips Dataset*, contains the following features:

1. Independent Variables:

- sex: Gender of the customer (Male/Female).
- smoker: Indicates whether the customer is a smoker (Yes/No).
- day: The day of the week the customer visited (e.g., Thursday, Friday).
- time: Indicates whether the meal was lunch or dinner.
- size: Number of members in the dining party.
- total bill: The total bill amount in USD.

2. Dependent Variable:

- tip: The tip amount in USD provided by the customer.
-

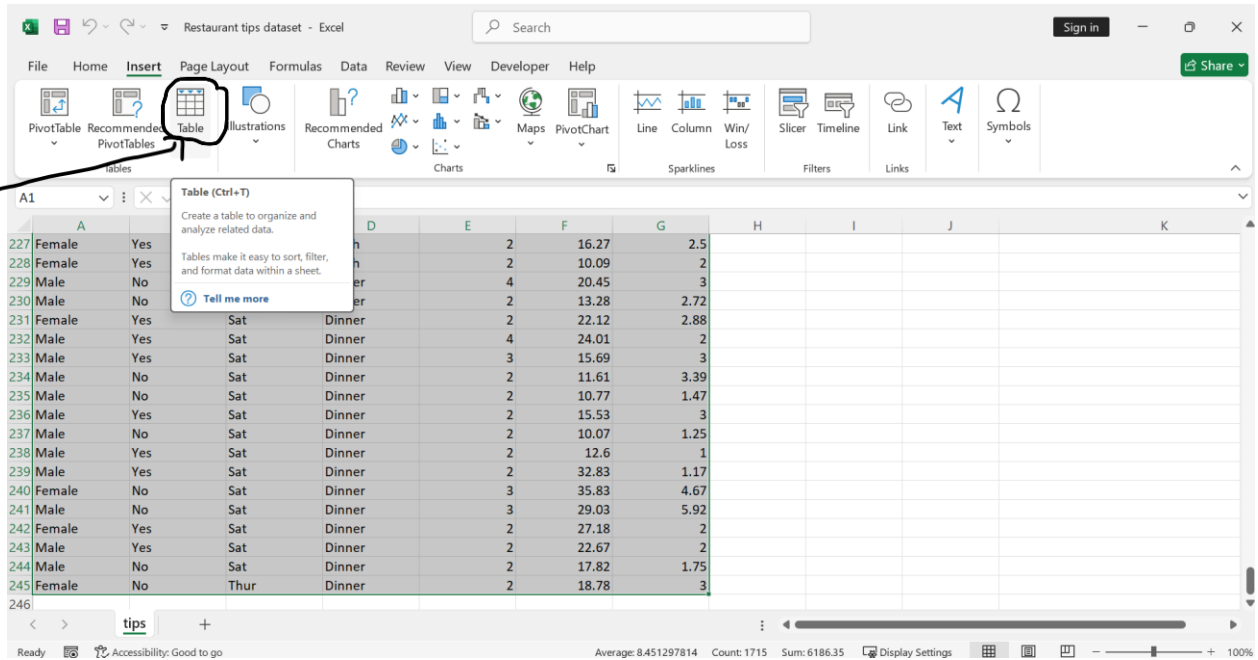
PROJECT TASKS:

-  **Download the data:** Provided by the organization.

Find out if there are any missing values and clean the data.

Solution:

- Open the data set in excel. Select the entire data and convert it into a **Table** from **Insert > Table**.
- Or
- Ctrl+T**.



Restaurant tips dataset - Excel

File Home Insert Page Layout Formulas Data Review View Developer Help

PivotTable Recommended PivotTables Illustrations Recommended Charts Maps PivotChart Line Column Win/Loss Slicer Timeline Link Text Symbols

A1 sex

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

tips

Point Accessibility: Good to go Average: 8.451297814 Count: 1715 Sum: 6186.35 Display Settings

Create Table

Where is the data for your table?

\$A\$1:\$G\$24\$

☒ My table has headers

OK Cancel

tip

➤ Name the tables as **raw**. Select the

tabledesign > Properties > Table Name.

Restaurant tips dataset - Excel

File Home Insert Page Layout Formulas Data Review View Developer Help **Table Design** Share

Table Name: raw

Summarize with PivotTable Remove Duplicates Insert Slicer Export Refresh Open in Browser Properties Convert to Range Unlink External Table Data

☒ Header Row ☐ First Column ☒ Filter Button

☐ Total Row ☐ Last Column

☒ Banded Rows ☐ Banded Columns

Table Style Options

Table Styles

C2 Sun

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

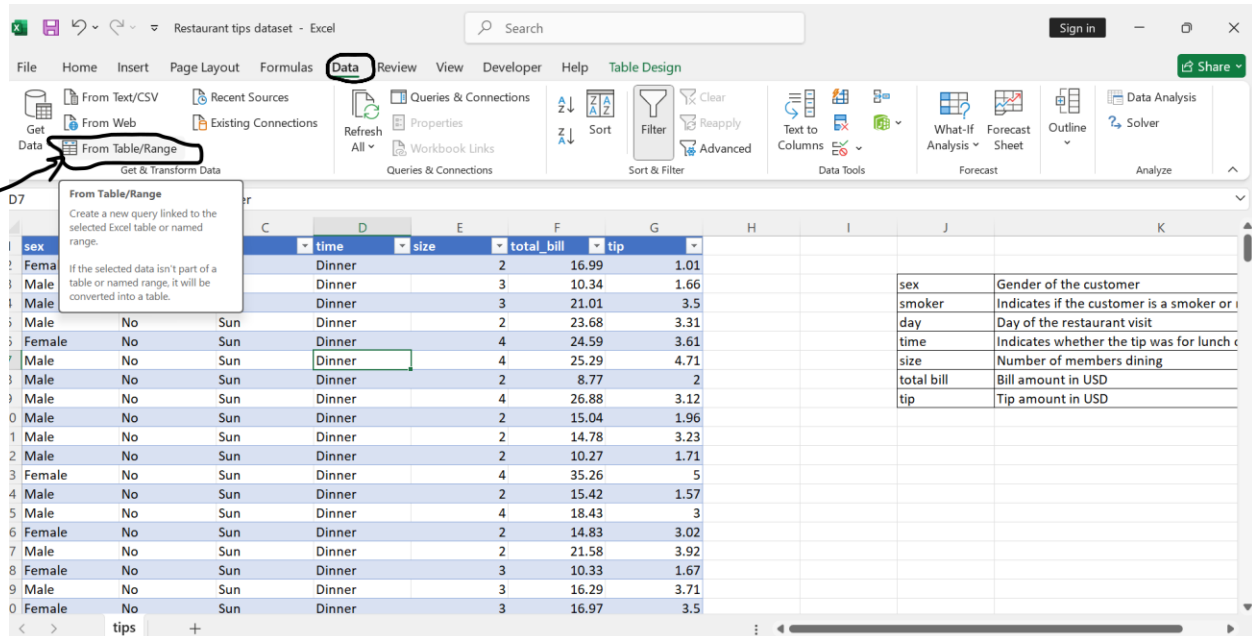
tips

Ready Accessibility: Good to go Display Settings 100%

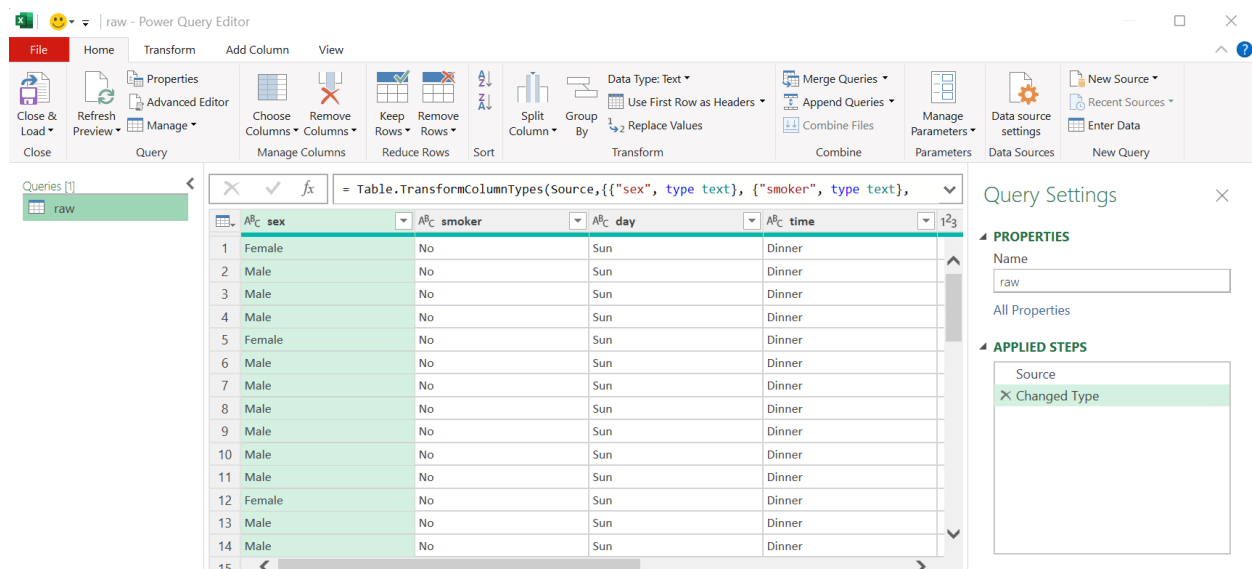
sex	Gender of the customer
smoker	Indicates if the customer is a smoker or not
day	Day of the restaurant visit
time	Indicates whether the tip was for lunch or dinner
size	Number of members dining
total bill	Bill amount in USD
tip	Tip amount in USD

- Data cleaning will be done in Power Query. Select any cell in the

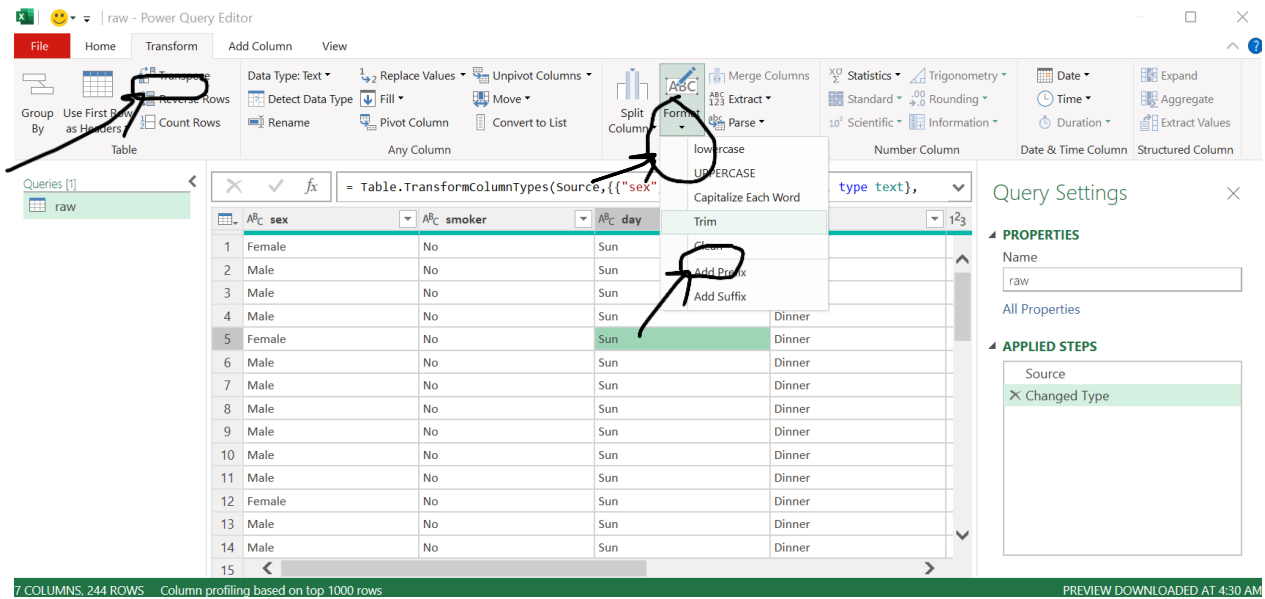
table>Data>From Table/Range



After click on From Table/Range open like below:

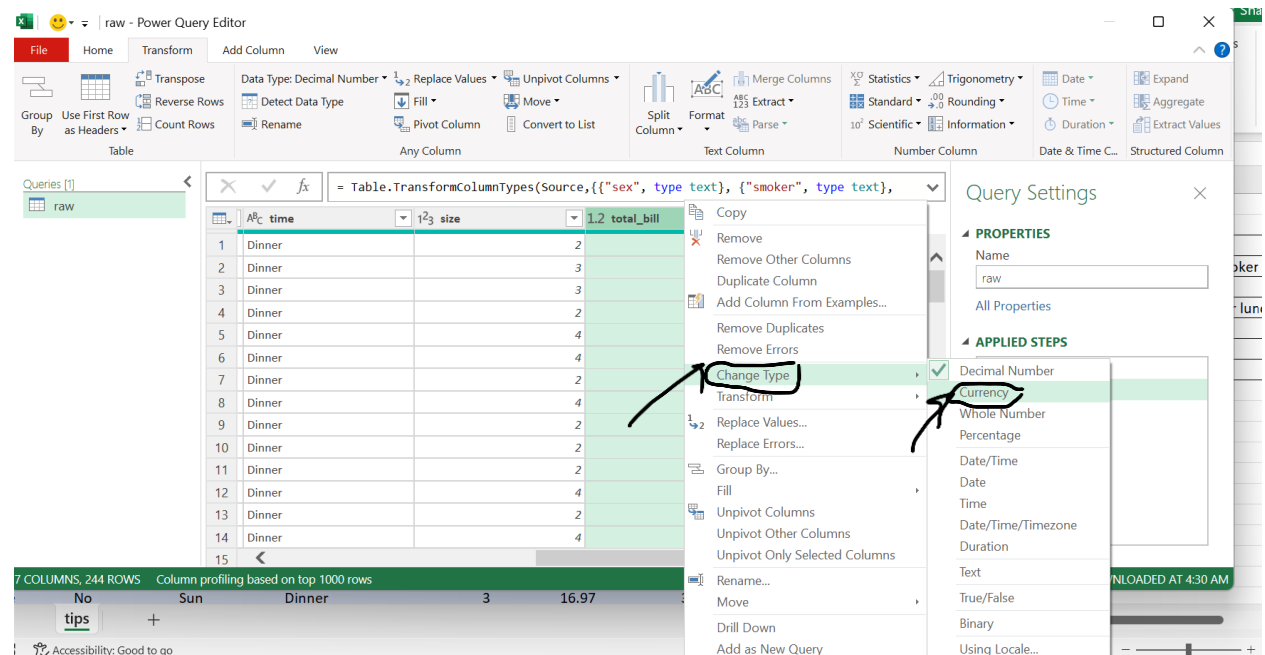


- Trimming extra spaces: **CTRL+A >Transform>Format>Trim**



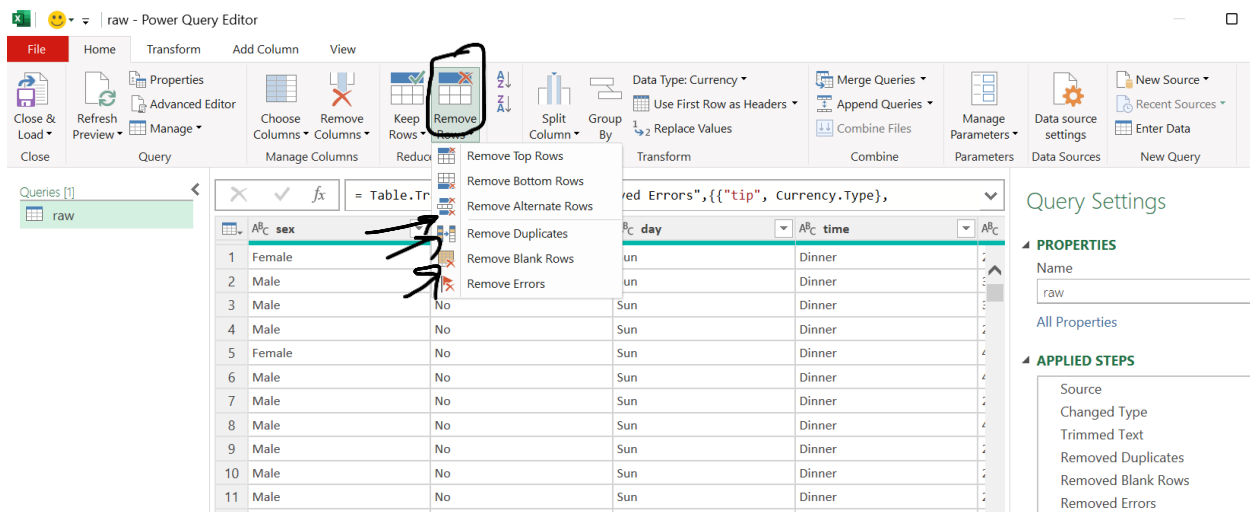
- Changing data types appropriately for columns:

Selecting the column for data type change -> Changes Type (total_bill and tips columns to Currency)



- Removing rows with **duplicates, errors, blank rows**:

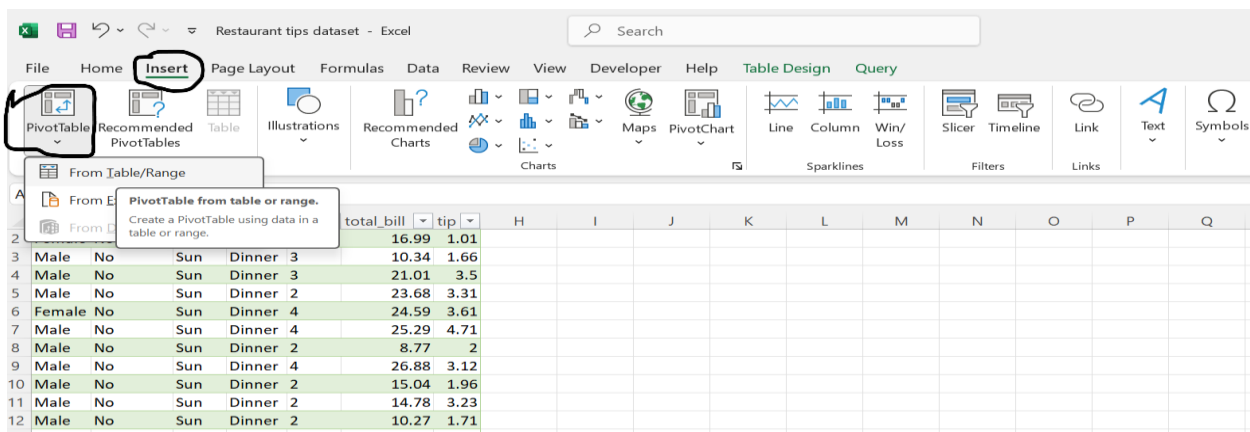
CTRL+A>Home>Remove Rows

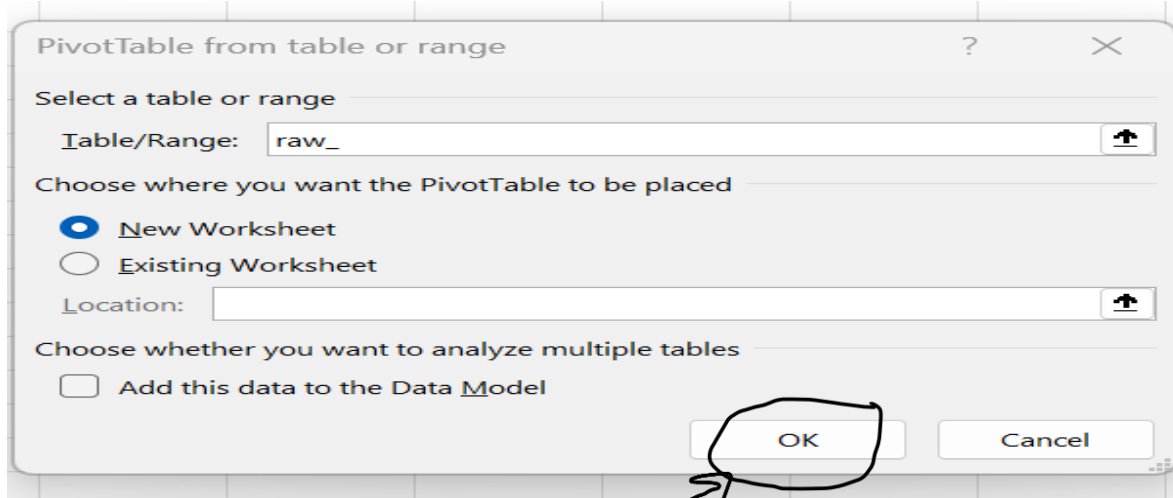


Exploratory Data Analysis:

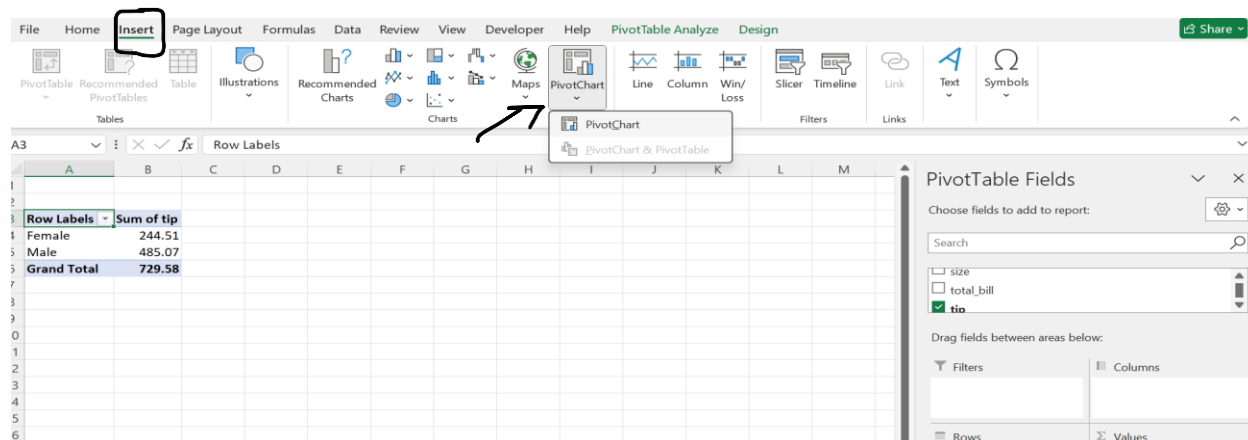
This step helps in understanding the data, finding patterns and anomalies through descriptive statistics and visualizations.

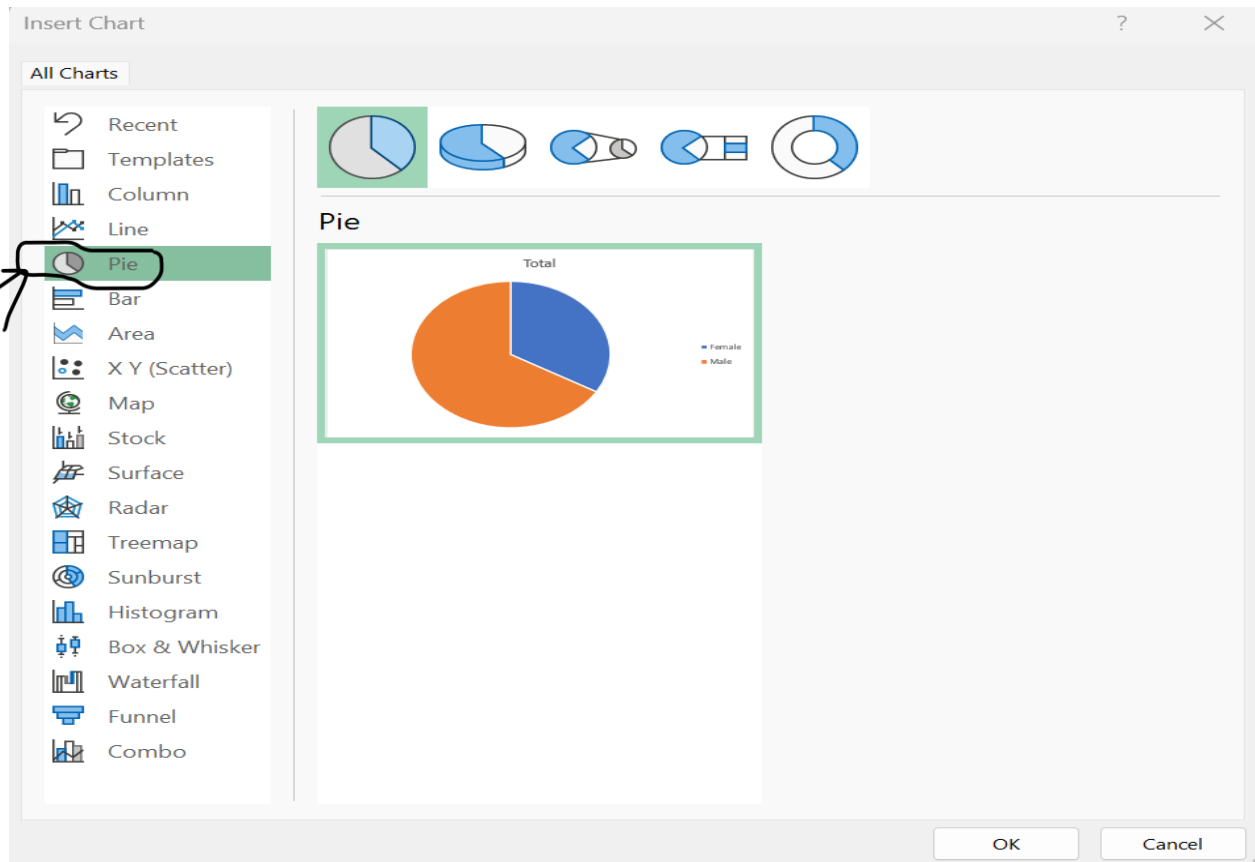
We'll use pivot tables to analyze and understand the patterns in our data. **Insert>Pivot tables>raw.**





1. Tips by Sex: sex in rows, tips in values. Add a pie chart; **click on insert -> PivotChart -> Pie chart.**



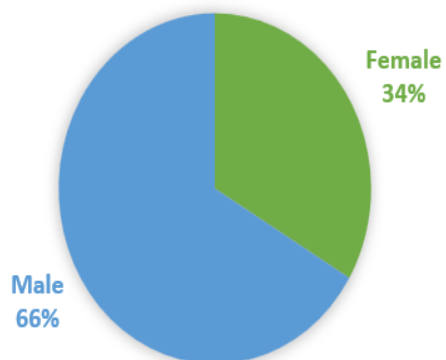


2. Tips by Time: time in rows, tips in values. Add a pie chart.

Output:

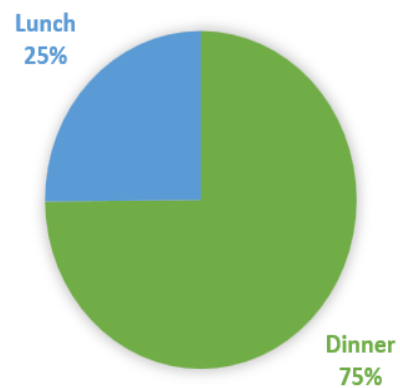
1.

TIPS BY SEX



2.

TIPS BY TIME



3. Tips by Sex and Time: sex in rows, time in columns, tips in values. Add a column chart.

4. Tips by Smoking Status: smoker in rows, tips in values. Add a column chart.

Output:

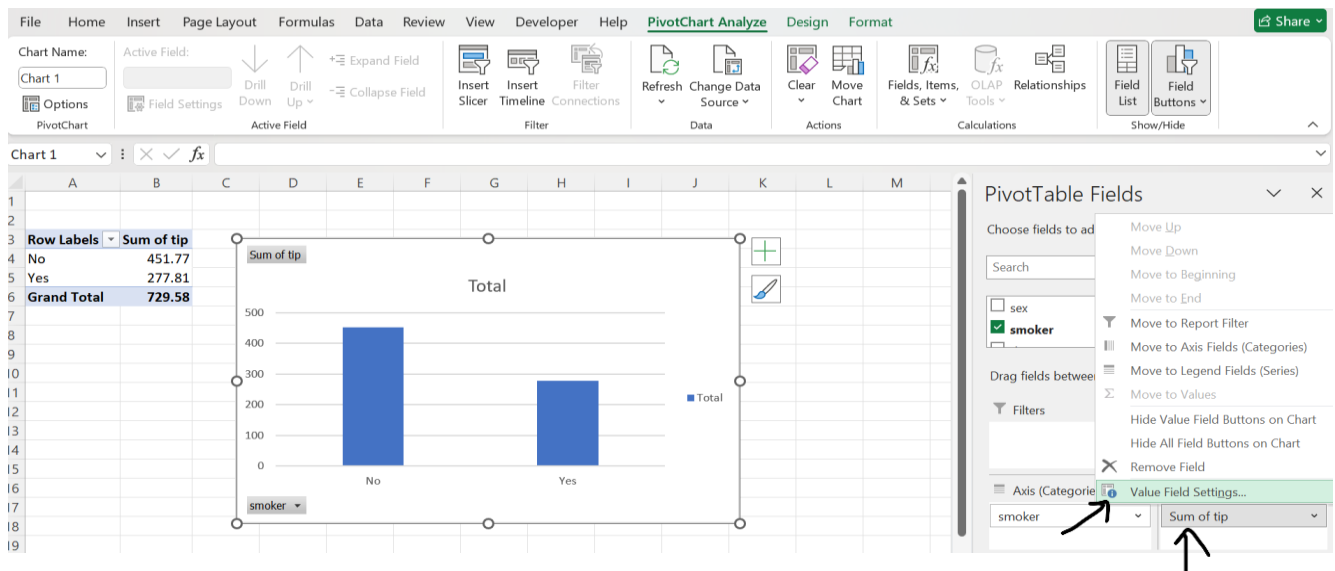
3.

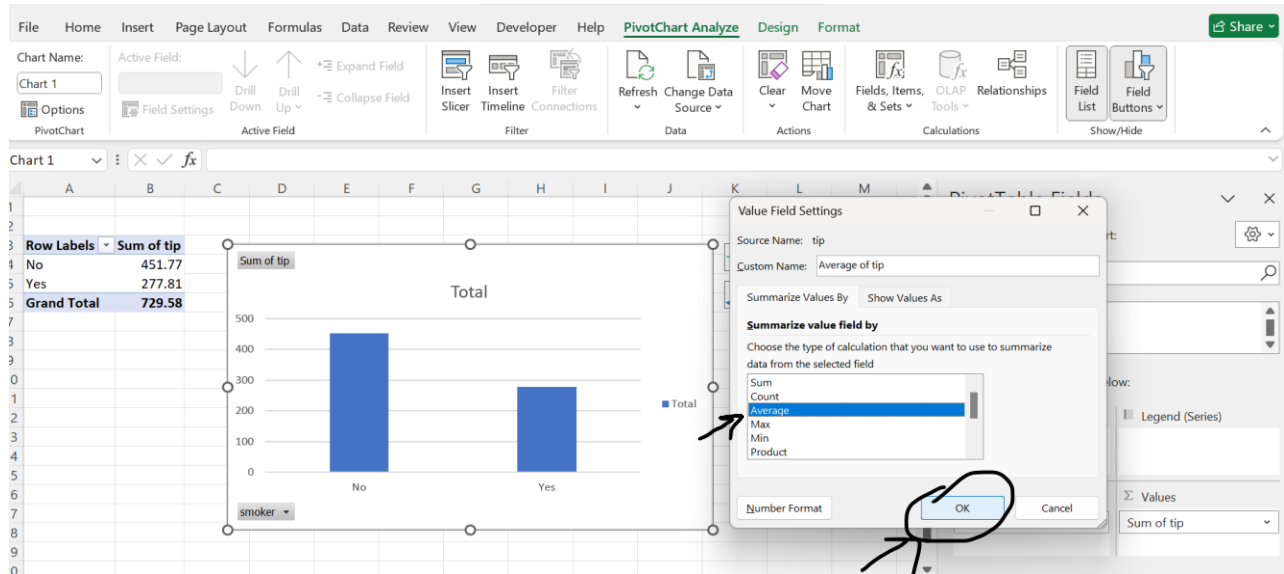


4.



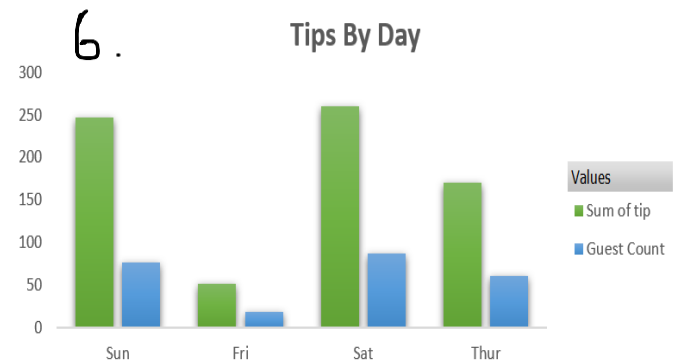
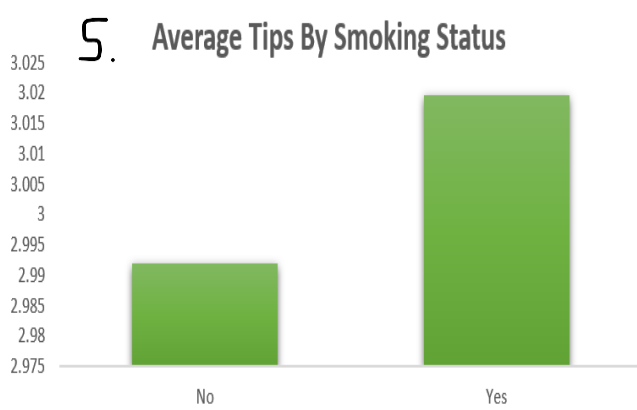
5. Average Tips by Smoking Status: smoker in rows, tips in values. Add a column chart. **PivotTable Fields>Values->Click Sum of Tip->Value Field Settings->Summarize value fields by>Average->OK.**





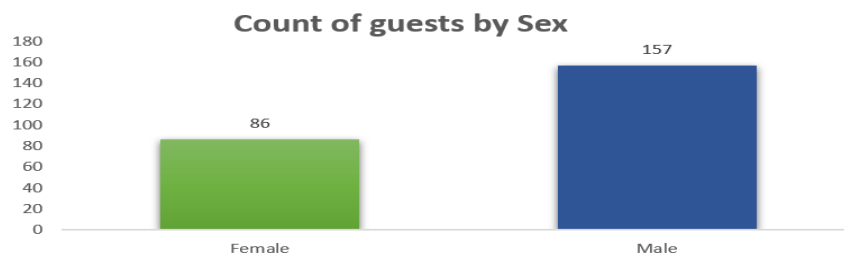
6. Tips By Day: day in rows, sum of tip and count of in values. Add a column chart.

Output:



7. Count of guests by Sex: sex in rows, sex in values.

Output:



Insights:

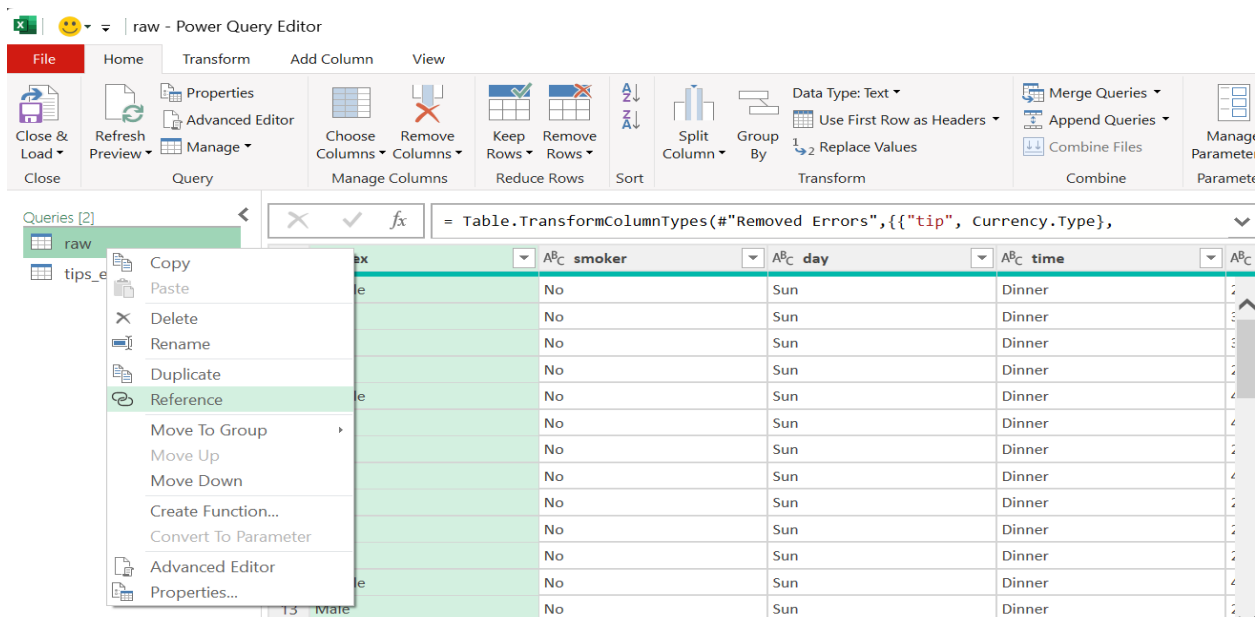
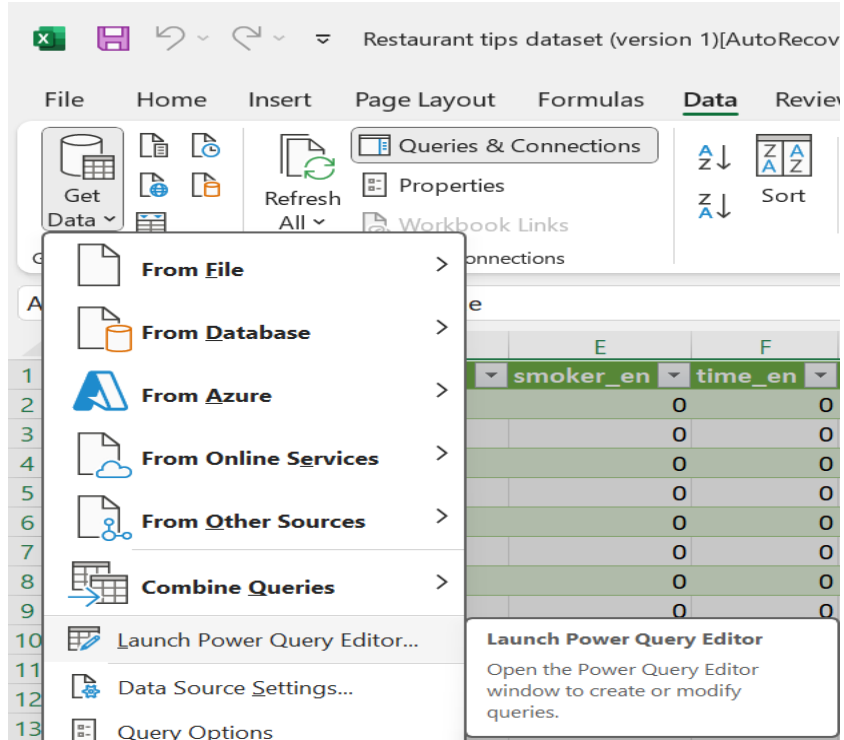
- From the Tips by Sex graph, it seems like men tend to give more tips than women.
- Tips received during dinner time are higher than lunch time but really, more people are visiting during dinner time. The size of guests at a table is almost the same for dinner and lunch time.
- The total tips given is higher by non-smokers but the average tips given by smokers is higher. Indicating that the number of non-smoker guests is higher but smokers tend to tip more.
- Guests tend to tip more during weekends.
- As the total bill of the table increases the tips also increases.

Feature Engineering:

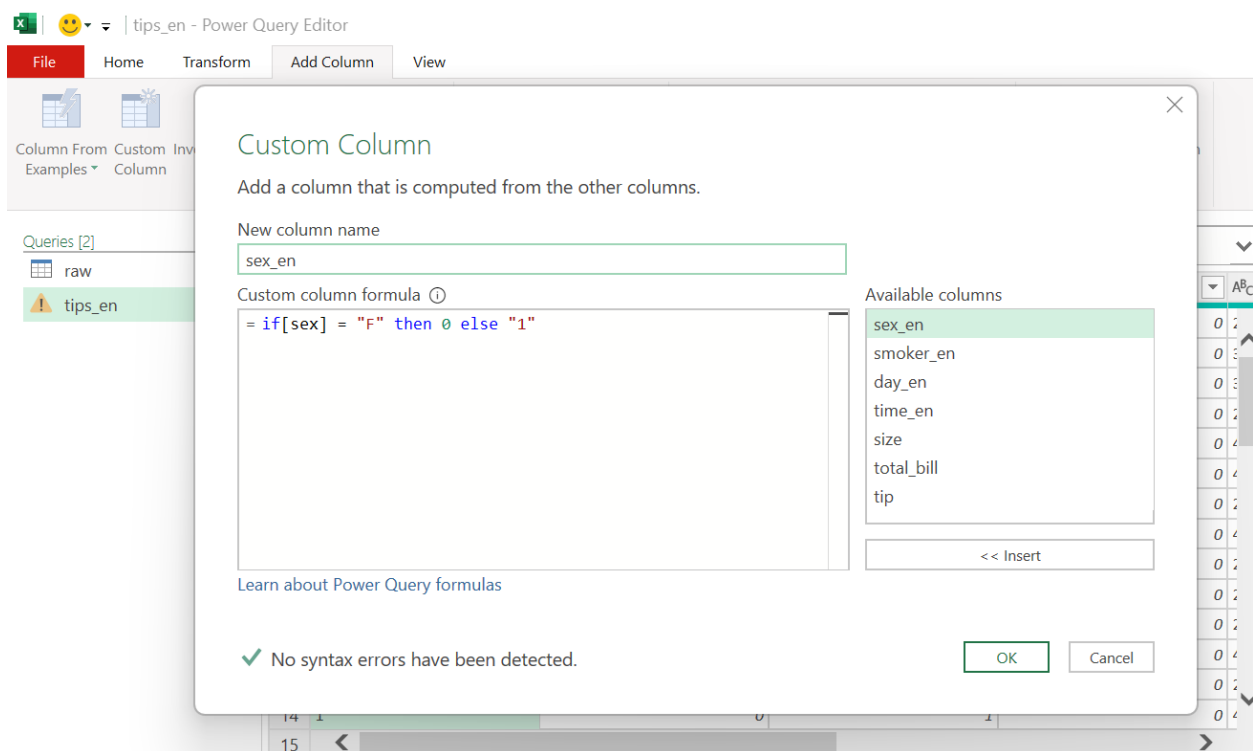
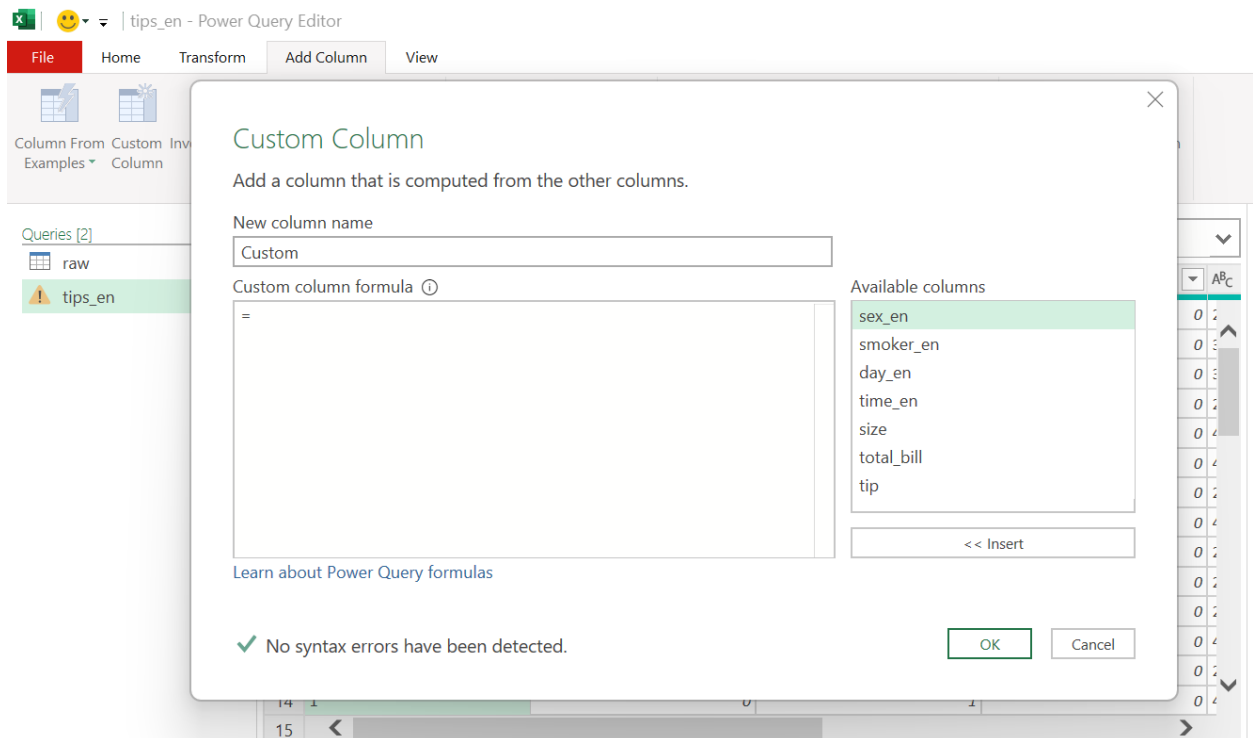
- It's a process of creating new features or modifying existing ones to improve the performance of models. Features are nothing but independent variables.
- In this step, we will use Power Query to convert categorical columns into a numeric format, ensuring compatibility for model training. because model training can't be done with categorical values.
- Instead of modifying the original query, we will create a reference query. This allows the new query to link to the original, ensuring that any updates made to the original query are automatically reflected in the reference query.

Steps: to add reference query

- **Open Power Query:** Go to the **data** tab -> click on **get data** -> click on **launch power query** -> Right click on the existing query (**raw**) -> click on **Reference**



Right click on **reference (copy raw)**-> click on **Add Column** -> click on **Custom Column**



Sex: if[sex] = "F" then 0 else "1" (write this query in custom column formula) then change custom (new column name) in to “sex_en”

Repeat for all the categorical columns

Smoker: No = 0, Yes = 1. **Custom column formula:** if[smoker] = " No" then 0 else "1"

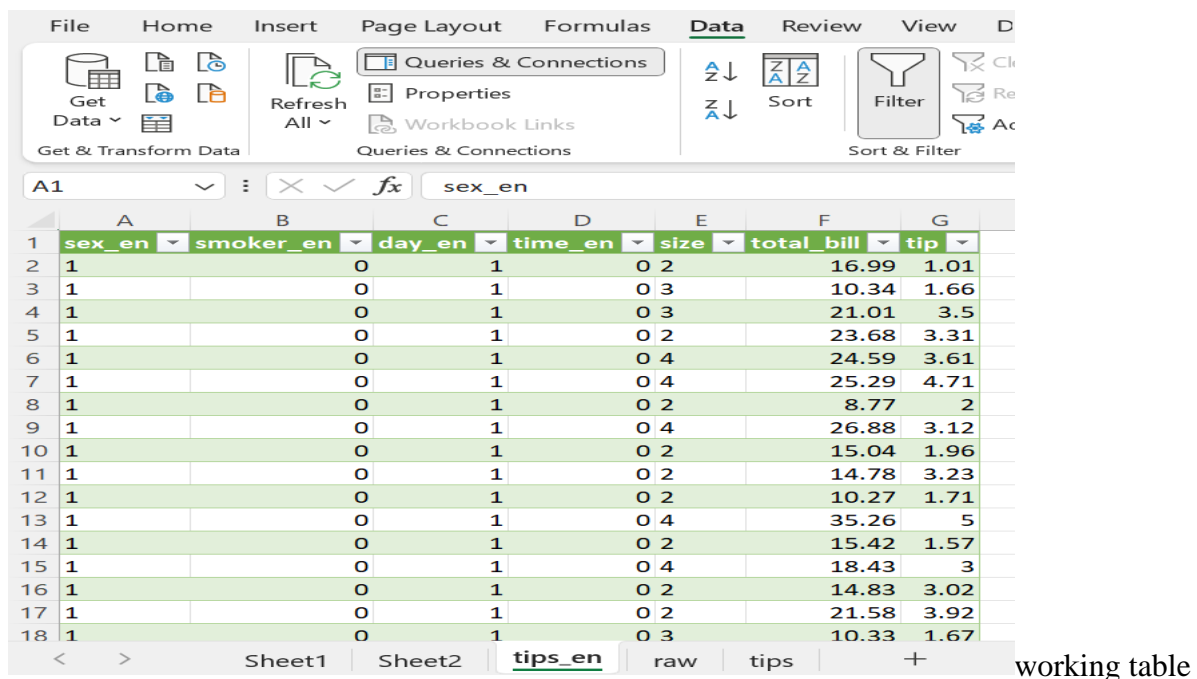
Time: Dinner = 0, Launch = 1. **Custom column formula:** if[time] = " Dinner" then 0 else "1"

Day: For Sunday to Saturday use 1 to 7 respectively. **Custom column formula:**
If[day] = “sun” then 1 else if [day] = “Mon” then 2 else if [day] = “Tue” then 3 else if [day] =
"Wed" then 4 else if [day] = "Thu" then 5 else if [day] = "Fri" then 6 else if [day] = "Sat" then 7
else null

Remove the original columns sex, smoker, time, day columns (right click on column then -> click on **remove**)

Reorder the columns similar to the original table. Rename the table as “tips_en” -> Close and Load(keep).

Final look of “tips_en” will be:



	A	B	C	D	E	F	G
1	sex_en	smoker_en	day_en	time_en	size	total_bill	tip
2	1	0	1	0	2	16.99	1.01
3	1	0	1	0	3	10.34	1.66
4	1	0	1	0	3	21.01	3.5
5	1	0	1	0	2	23.68	3.31
6	1	0	1	0	4	24.59	3.61
7	1	0	1	0	4	25.29	4.71
8	1	0	1	0	2	8.77	2
9	1	0	1	0	4	26.88	3.12
10	1	0	1	0	2	15.04	1.96
11	1	0	1	0	2	14.78	3.23
12	1	0	1	0	2	10.27	1.71
13	1	0	1	0	4	35.26	5
14	1	0	1	0	2	15.42	1.57
15	1	0	1	0	4	18.43	3
16	1	0	1	0	2	14.83	3.02
17	1	0	1	0	2	21.58	3.92
18	1	0	1	0	3	10.33	1.67

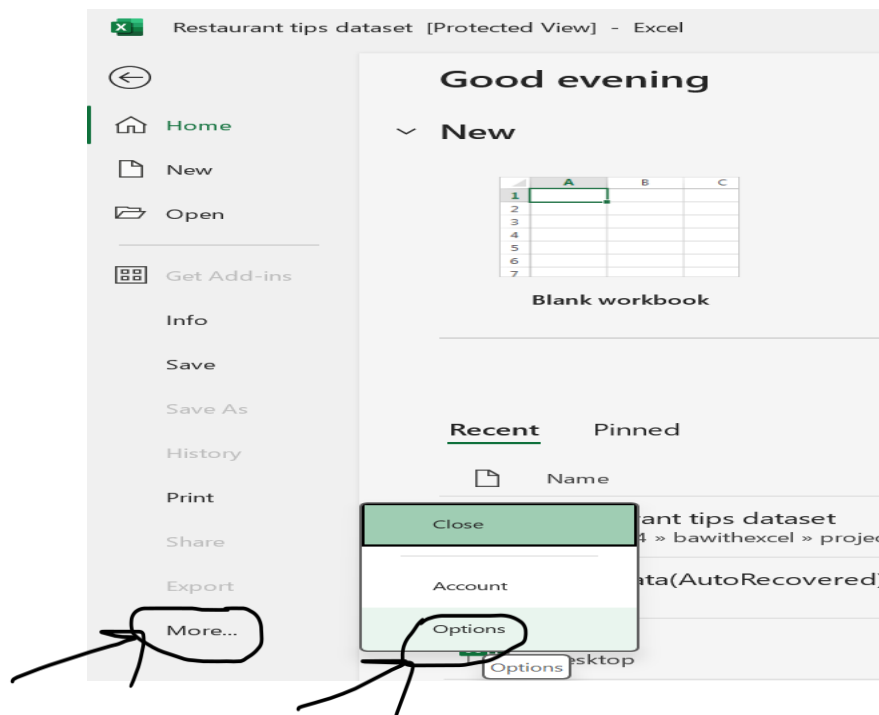
Note: From the next step forward, understanding of certain statistics concepts are necessary. Check out “[Understanding Multiple Linear Regression in Simple Terms](#)”

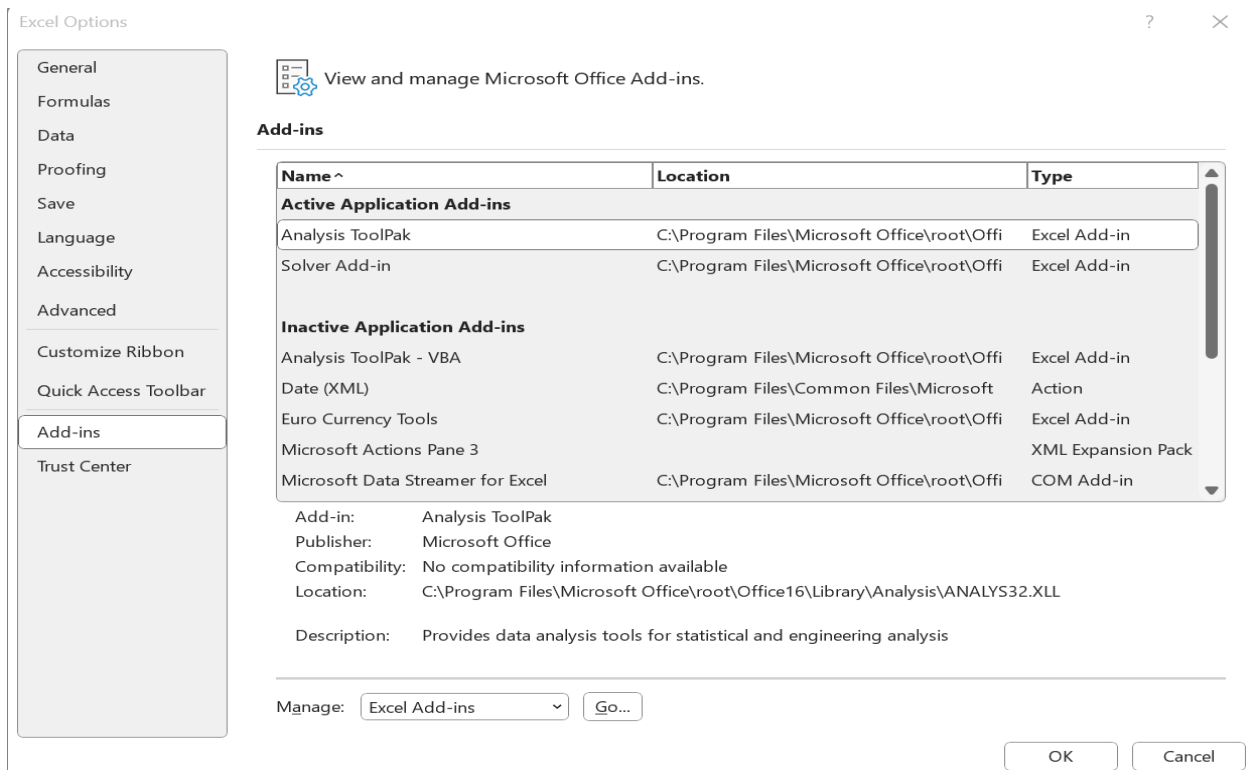
Feature Selection:

First, we will identify the independent and dependent variables for the regression model. Since our goal is to predict tips, the 'tips' column will be the dependent (target) variable, while all other columns will be independent variables (predictors)

Next, we will analyze covariance, correlation, and multicollinearity to help in selecting independent variables that have a strong relationship with the target variable while minimizing relationships with one another.

- ✓ For further steps make sure the “**Data Analysis Toolpak**” is added. If not go to **Files -> more -> Options -> Add-ins -> Click on Analysis ToolPak -> Manage**(at the bottom side) -> **Go -> Ok**





Now go to the **Data tab -> Data Analysis -> click on Correlation**. Input range is the entire table.
 (Make sure table is “**numeric**” to check: Select your dataset and press **Ctrl + G** → Click **Special**
 → Choose **Constants** → Uncheck everything except **Text** → Click **OK**.)

❖ If Excel highlights any non-numeric values, either remove them or convert them to numbers.

NOTE: If show non-numeric values then Go to the -> **get data -> click on launch power query editor -> click on reference table (tips_en) then right click on cell -> click on change type -> whole number. Repeat For all the columns.**

tips_en - Power Query Editor

File Home Transform Add Column View

Close & Load Refresh Advanced Editor Manage Query

Choose Columns Remove Columns Keep Rows Remove Rows Split Column Group By Replace Values

Data Type: Whole Number Merge Queries Append Queries Combine Files

Use First Row as Headers Combine Parameters Data source settings New Source Recent Sources Enter Data

Queries [2] raw tips_en

1 sex_en 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Copy Remove Remove Other Columns Duplicate Column Add Column From Examples... Remove Duplicates Remove Errors Change Type Transform Replace Values... Replace Errors... Group By... Fill Unpivot Columns Unpivot Other Columns Unpivot Only Selected Columns Rename... Move Drill Down Add as New Query

Reordered Columns1, {{sex_en, Int64.Type}, time_en, day_en}

Decimal Number Currency Whole Number Percentage Date/Time Date Time Date/Time/Timezone Duration Text True/False Binary Using Locale

Query Settings

PROPERTIES Name tips_en All Properties

APPLIED STEPS Source Added Custom Added Custom1 Added Custom2 Added Custom3 Removed Columns Reordered Columns Removed Errors Removed Errors1

7 COLUMNS, 243 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 2:22 PM

Correlation

Restaurant tips dataset(AutoRecovered) - Excel

File Home Insert Page Layout Formulas Data Review View Developer Help

Get Data Refresh Properties Workbook Links

Sort Filter Clear Reapply Text to Columns What-If Analysis Forecast Sheet Outline Solver

Get & Transform Data Queries & Connections Sort & Filter Data Tools Forecast Analyze

112

	A	B	C	D
	sex_en	smoker_en	time_en	day_en
1	0	0	0	1
2	1	0	0	1
3	1	0	0	1
4	1	0	0	1
5	1	0	0	1
6	0	0	0	1
7	1	0	0	1
8	1	0	0	1
9	1	0	0	1
10	1	0	0	1
11	1	0	0	1
12	1	0	0	1
13	0	0	0	1
14	1	0	0	1
15	1	0	0	1
16	0	0	0	1

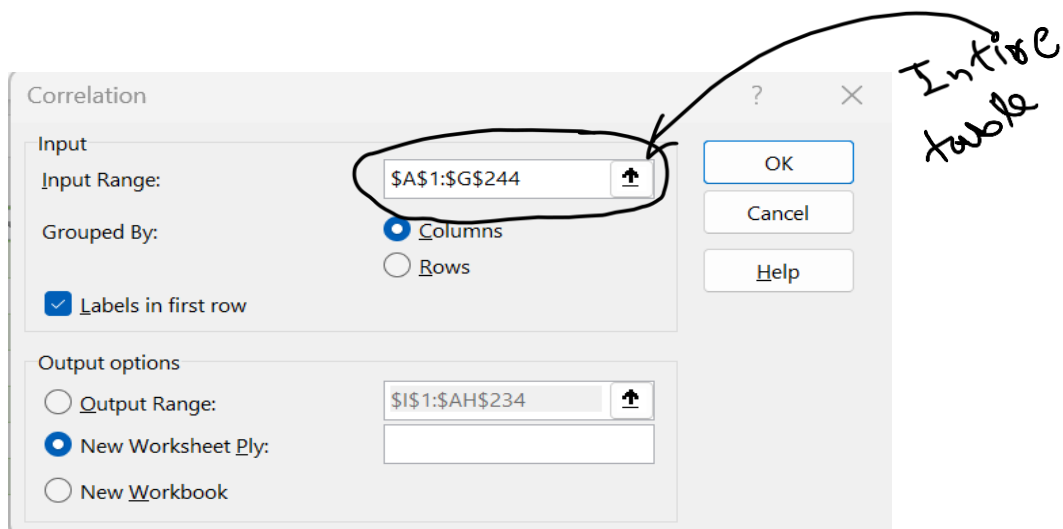
Data Analysis

Analysis Tools

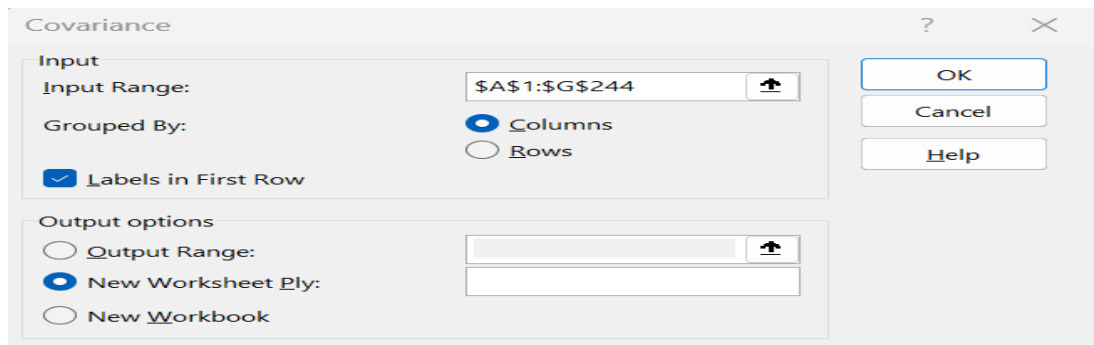
- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

OK Cancel Help

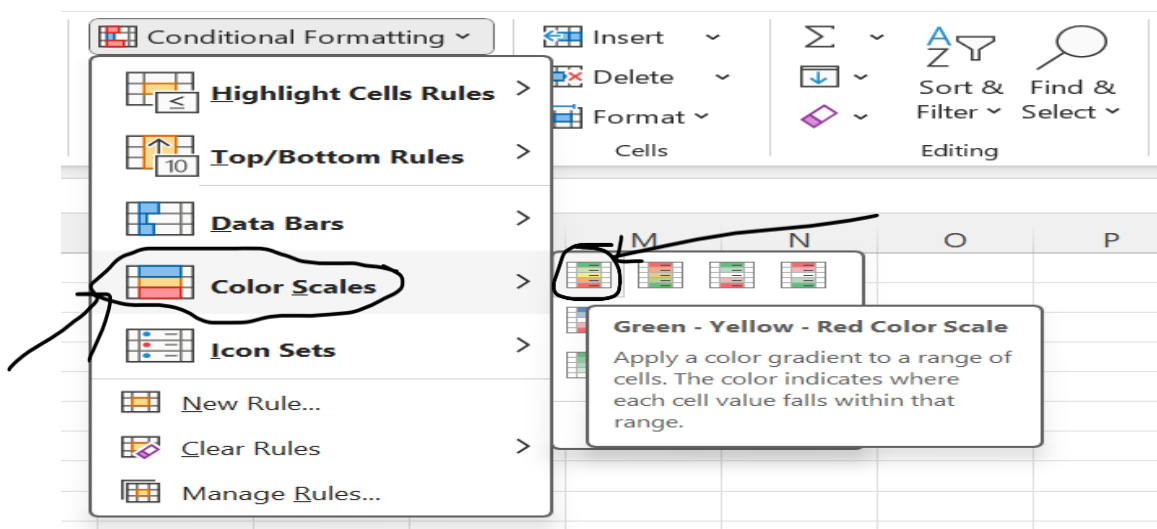
	K	L	M	N
9	4	26.88	3.12	
10	2	15.04	1.96	
11	2	14.78	3.23	
12	2	10.27	1.71	
13	4	35.26	5	
14	2	15.42	1.57	
15	4	18.43	3	
16	2	14.83	3.02	



Do the same for covariance:



- Applying the conditional formatting for correlation **Home** -> click on **Conditional Formatting** -> click on **Color_Scales** -> Pick whichever you like




- **Home -> Conditional Formatting -> click on Manage Rules -> Select Rule Edit Rule**

Conditional Formatting Rules Manager

Show formatting rules for: Current Selection

New Rule Edit Rule... Delete Rule

Rule (applied in order shown)	Format	Applies to	Stop If True
Graded Color Scale		=A\$1:\$H\$8	<input type="checkbox"/>

OK Close Apply

- ❖ **Set Minimum, Midpoint, Maximum to -1, 0, 1, since range of correlation is -1 to 1.**

Edit Formatting Rule

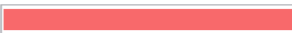
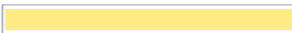

Select a Rule Type:


- Format all cells based on their values
- Format only cells that contain
- Format only top or bottom ranked values
- Format only values that are above or below average
- Format only unique or duplicate values
- Use a formula to determine which cells to format

Edit the Rule Description:

Format all cells based on their values:

Format Style: 3-Color Scale

	Minimum	Midpoint	Maximum
Type:	Number	Number	Number
Value:	-1	0	1
Color:			

Preview: 

OK Cancel

CORRELATION TABLE OUTPUT WILL BE:

	A	B	C	D	E	F	G	H
1		<i>sex_en</i>	<i>smoker_en</i>	<i>time_en</i>	<i>day_en</i>	<i>size</i>	<i>total_bill</i>	<i>tip</i>
2	<i>sex_en</i>	1						
3	<i>smoker_en</i>	0.009930188	1					
4	<i>time_en</i>	-0.198128623	-0.063911231	1				
5	<i>day_en</i>	-0.110641838	0.219698794	0.137905919	1			
6	<i>size</i>	0.083248017	-0.130564411	-0.100045303	-0.17330627	1		
7	<i>total_bill</i>	0.141349744	0.090136102	-0.179231854	-0.078628796	0.597588931	1	
8	<i>tip</i>	0.085273975	0.00976275	-0.11759639	-0.099046544	0.488400395	0.674997857	1

COVARIANCE TABLE OUTPUT WILL BE:

	<i>sex_en</i>	<i>smoker_en</i>	<i>time_en</i>	<i>day_en</i>	<i>size</i>	<i>total_bill</i>	<i>tip</i>
<i>sex_en</i>	0.228657556						
<i>smoker_en</i>	0.002303172	0.23526224					
<i>time_en</i>	-0.04233772	-0.013852902	0.199698555				
<i>day_en</i>	-0.132855764	0.267591322	0.154752832	6.305746075			
<i>size</i>	0.037832986	-0.060187302	-0.042490135	-0.41360565	0.903249843		
<i>total_bill</i>	0.600998662	0.388741215	-0.712177683	-1.7556404	5.050009484	79.06265664	
<i>tip</i>	0.056359125	0.006544903	-0.072633406	-0.343765686	0.641556504	8.295509286	1.91033669

Key Observations:

- **Tip & Total Bill (0.67):** This shows a strong positive correlation, meaning higher total bills generally result in higher tips.
- **Size & Total Bill (0.59):** A moderate positive correlation, suggesting that larger groups tend to spend more.
- **Time & Total Bill (-0.17):** A weak negative correlation, possibly indicating that meal timing affects spending habits.
- **Smoker & Tip (0.0097):** Almost no correlation, meaning smoking status does not significantly impact tipping. (or have a nonlinear relationship with tip (Since correlation coefficient only checks for linear relationship. In case of other type of relationship, coefficient will either 0 or closer to 0) but that is beyond the scope of this project. So, we can exclude sex and smoker features from the Regression Model.)

Model Building:

Using Excel's **Data Analysis Add-in**, we will build a **linear regression model** to predict the tip amount. Since multiple factors influence the tip amount, a multiple linear regression model will be used (Note: This is not the same as multivariate regression).

Now go to the **Data tab -> Data Analysis -> click on Regression**. **Input Y range** is the Select “tip cell” as the **dependent variable**. And **input X range** is Select the transformed/encoded **independent variables (time, day, size, total bill)** as input features.

The screenshot shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' set to '\$G\$1:\$G\$244' and 'Input X Range' set to '\$C\$1:\$F\$244'. The 'Labels' checkbox is checked, and 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected. The 'Residuals' section has 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' all unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. Buttons for 'OK', 'Cancel', and 'Help' are on the right.

For line fit plots:

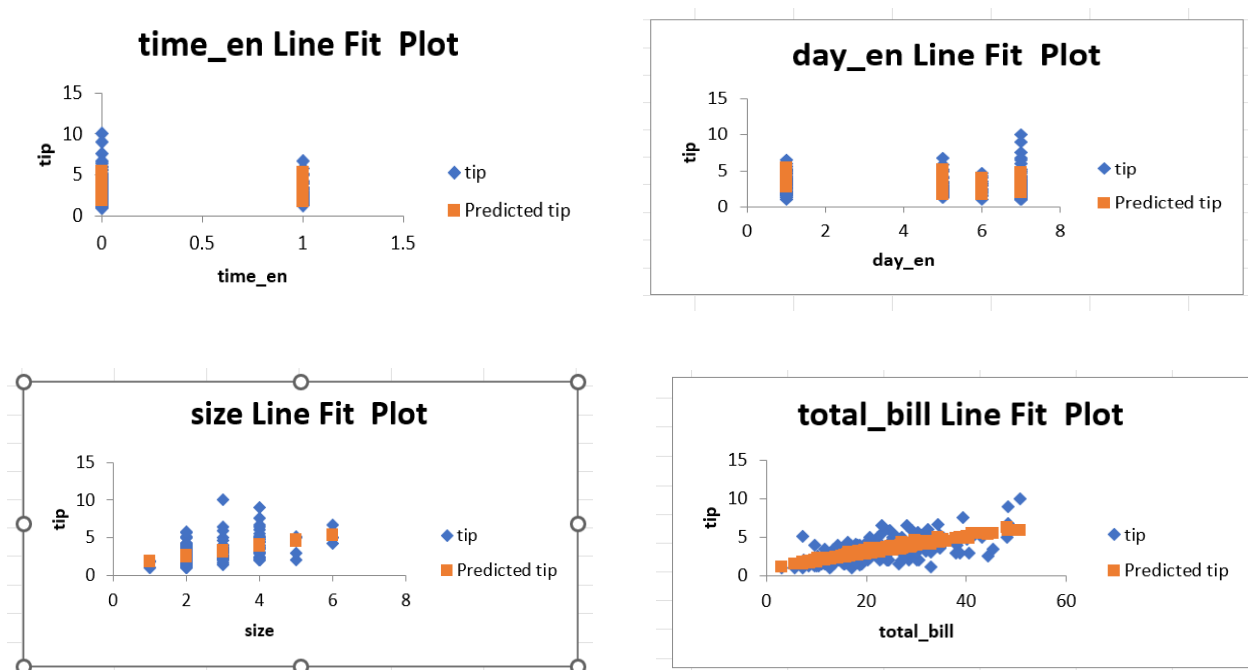
This screenshot is identical to the previous one, but with the 'Line Fit Plots' checkbox in the 'Residuals' section checked. A hand-drawn arrow points to this checkbox, and the handwritten text 'For plots' is written next to it. The 'OK', 'Cancel', and 'Help' buttons remain on the right.

Output and Model Evaluation Measures:

➤ R-squared and p-values

Regression Statistics									
Multiple R	0.683953829								
R Square	0.46779284								
Adjusted R Square	0.458848182								
Standard Error	1.018849351								
Observations	243								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	4	217.1549637	54.28874093	52.29857113	1.43528E-31				
Residual	238	247.0568519	1.038054						
Total	242	464.2118156							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	0.75984875	0.250420973	3.03428559	0.002678647	0.266524058	1.253173441	0.266524058	1.253173441	
time_en	0.020254901	0.149929543	0.135096133	0.892650052	-0.275103529	0.315613331	-0.275103529	0.315613331	
day_en	-0.01708897	0.02666475	-0.640882426	0.522215685	-0.069618036	0.035440096	-0.069618036	0.035440096	
size	0.183369889	0.086899346	2.110141191	0.035890244	0.012179783	0.354559995	0.012179783	0.354559995	
total_bill	0.09301373	0.009284134	10.0185678	6.13287E-20	0.074724157	0.111303303	0.074724157	0.111303303	

Line Fit Plots: Predicted vs. Actual Tips



Residual:

Regression

Input

Input Y Range: \$G\$1:\$G\$244

Input X Range: \$C\$1:\$F\$244

☒ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK

Cancel

Help

RESIDUAL OUTPUT:

RESIDUAL OUTPUT		
Observation	Predicted tip	Residuals
1	2.689802833	-1.679802833
2	2.254631417	-0.594631417
3	3.247087918	0.252912082
4	3.312064688	-0.002064688
5	3.763446961	-0.153446961
6	3.828556572	0.881443428
7	1.925229972	0.074770028
8	3.976448403	-0.856448403
9	2.50842606	-0.54842606
10	2.48424249	0.74575751
11	2.064750567	-0.354750567
12	4.755903461	0.244096539
13	2.543771277	-0.973771277
14	3.190482383	-0.190482383
15	2.488893176	0.531106824
16	3.116735855	0.803264145
17	2.25370128	-0.58370128
18	2.808063111	0.901936889
19	2.871312448	0.628687552
20	2.111060155	0.228020845

Calculate RMSE:

The **Root Mean Square Error (RMSE)** is calculated to evaluate the model's performance.

$$=SQRT(SUMSQ(C28:C270)/COUNTA(C28:C270))$$

RMSE	1.00831288
-------------	------------

CONCLUSIONS:

The model can predict restaurant tips with moderate accuracy. It has a Root Mean Square Error (**RMSE**) of 1.0083 and an R-squared value of 0.46, meaning it explains 46% of the variation in tips. However, only the **size** and **total_bill** variables are statistically significant, meaning they have a meaningful impact on predicting tips. The other variables do not show statistical significance and may not strongly influence tip predictions.