Predicting Restaurant Tips

Multiple Linear Regression Using Excel

SUBMISSION FOR

BUSINESS ANALYTICS WITH EXCEL

IITK PCP DATA ANALYTICS AND GENERATIVE AI

SUBMISSION BY

ANJALI HANSDA

Project Description

You are a data analyst working for a restaurant chain. The management has provided you with a dataset (Restaurant tips dataset.xlsx) containing information about customer tips. The dataset includes customer gender, smoking status, day of visit, time of visit, and total bill amount. You are tasked with building a predictive model to estimate tip amounts.

Expected deliverables as per the problem statement document :

1. Predictive Model
2. Pivot charts for EDA

Dataset

The dataset contains tips data for different customers. The following are the features in the dataset:

Sex: The gender of the customer

Smoker: indicate 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

Concepts of Multiple Linear Regression

Regression:

Regression is a statistical technique used to understand and model the relationship between a dependent variable and one or more independent variables. The model then helps in the predicting values of dependent variables for the new unseen independent variables.

Dependent Variable

Variable that you want to predict. In the regression model the dependent variable is a continuous numeric value.

Independent Variable

Variables that have some relationship with the dependent variable and affect the value of the dependent variable.

Regression Type: Multiple Linear Regression

There are several types of regression, you can check them out on your own time. Multiple Linear Regression models the relationship between multiple independent variables and one dependent variable.

The linear equation for MLR:

$$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon$$

- **y=** dependent variable
- **x1, x2.....,xn=** independent variables
- **β0=** intercept, point that crosses the y-axis when the linear equation depicting the relationship between x and y is plotted . It represents the value of the dependent variable(y) when the intercept variable(x) is 0.
- **β1,β2.....,βn**= slope, aka coefficient of independent variable, indicates the direction and steepness of the line. It is the ratio of the change in y to the change in x. It quantifies how much y will increase or decrease with each unit increase in x.
- **ϵ=** The error term

Covariance, Correlation and Multicollinearity:

These measure the relationship between two variables, but they do it in a slightly different way. These help in selecting the appropriate independent variables for the model.

Covariance

It measures the direction of the relationship between two variables. It indicates whether the variable increases or decreases together or in the opposite direction.

- Positive Covariance: Both variables tend to increase or decrease together.
- Negative Covariance: When one variable increases, the other tends to decrease.
- Zero Covariance: No relationship between the variables.

Limitations:

Doesn't give the strength of the relationship and it is sensitive to the scale of the variables. For instance changing the units of measurement (e.g. USD to INR) can change the covariance value.

Correlation

It measures both the strength and the direction of the linear relationship between two variables. It is a scaled version of variance, making it easier to interpret because it is dimensionless and falls in a fixed range -1 to +1.

- r = 1: Perfect positive correlation.
- r = -1: Perfect negative correlation.
- r = 0: No linear correlation (the variables are independent).

Multicollinearity

It occurs when two or more independent variables in a regression model are highly correlated with each other. Multicollinearity can cause problems in model interpretation and reduce the reliability of the estimated coefficients. It is present when two independent variables are highly correlated (typically above 0.8 or 0.9). Multicollinearity affects the feature selection for the model.

R-Squared

R-squared is a number between 0 and 1 that tells you how much of the variation in your data is explained by your model.

How does it help?

1. Measure of Fit: It helps to see how well the model fits the data. A higher R-squared value means better fit, meaning the model is better at predicting the outcomes.
    a. Good Fit: If your R-squared is high (close to 1), it suggests your model does a good job explaining the data.
    b. Poor Fit: If it's low (close to 0), it means the model doesn't explain much of the variation, and there might be other factors affecting the outcome that the model isn't capturing.
2. Percentage Explained: R-squared can be thought of as the percentage of the total variation in the dependent variable that your model is able to explain. For example, an R-squared of 0.70 means that 70% of the variation in the outcome can be explained by your model.

p-values

This is a statistical measure that helps in determining whether the results of the analysis are significant. A p-value tells the probability that the effect or relationship observed in the data could have happened just by random chance. It ranges between 0 and 1:

- A low P-value (typically < 0.05) suggests that the observed effect is unlikely to be due to chance, meaning it's statistically significant.
- A high P-value (> 0.05) indicates that the effect might have occurred by chance, meaning the result is not statistically significant.

Residuals

These are the differences between the actual values and the predicted values made by the model.

Residuals = Actual value - Predicted value

A positive residual means the model's prediction was too low (underestimated), while a negative residual means the prediction prediction was too high (overestimated).

RSME

It stands for Root Mean Square Error. It is used in measuring the accuracy of the model in predicting the outcomes. It tells how far the model's predictions are from the actual values, on average.

- A lower RMSE means the model's predictions are closer to the actual values, indicating a more accurate model.
- RSME is often used to compare different models. The model with lowest RMSE is usually considered the best.
- RMSE is in the same units as the dependent variable, making it easier to interpret.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

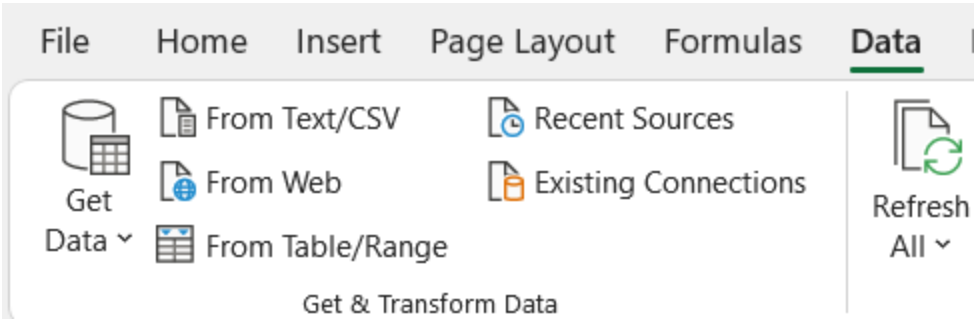
Project Workflow

Data Collection

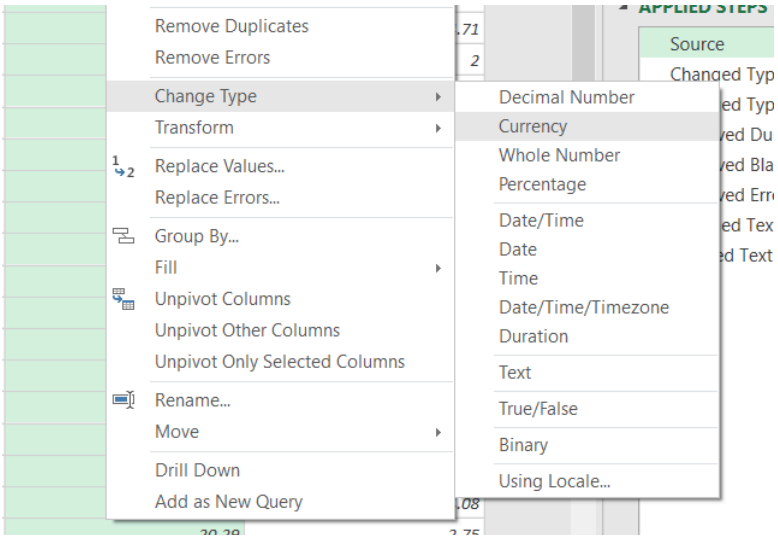
Provided by the organisation.

Data Cleaning

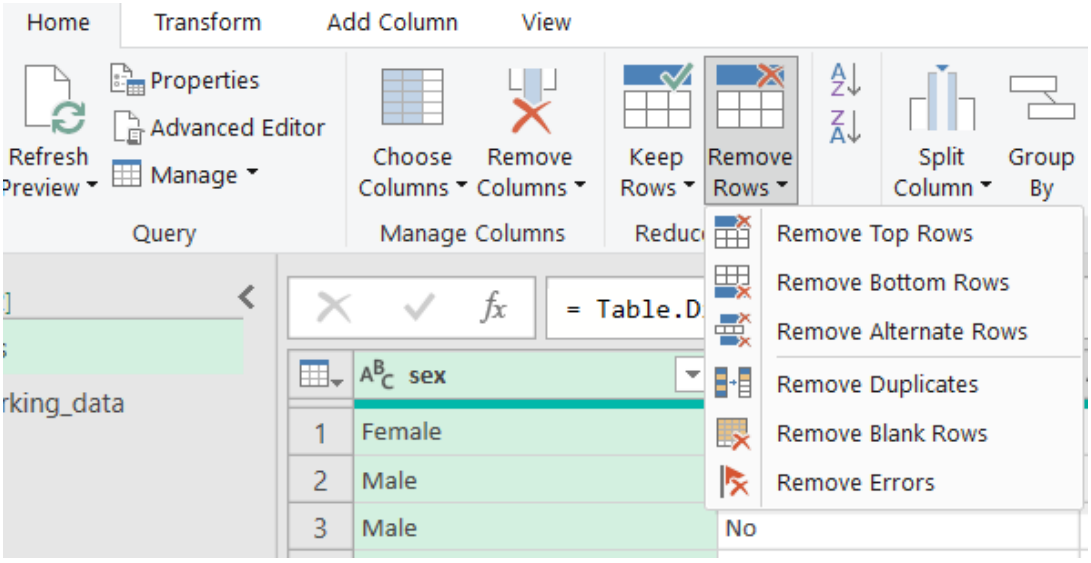
- Open the data set in excel. Select the entire data and convert it into a **Table** from **Insert>Table.**
- Name the tables as **`raw`. Select the table>Design>Properties>Table Name.**
- Data cleaning will be done in Power Query. **Select any cell in the table>Data>From Table/Range**



- 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

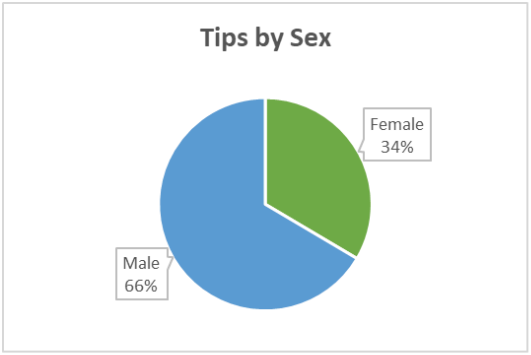


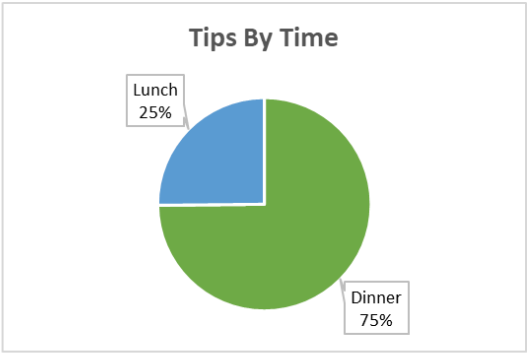
- Removing rows with duplicates, errors, blank rows. The bar below the column name shows the percentage of errors and blank rows. Although there aren't such rows but with influx of more data adding the step will automate the data cleaning task. **Ctrl+A>Home>Remove Rows**

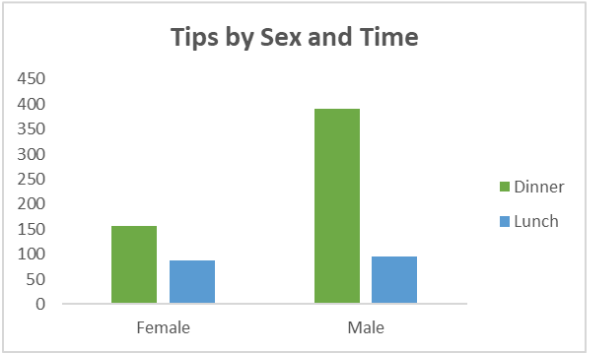


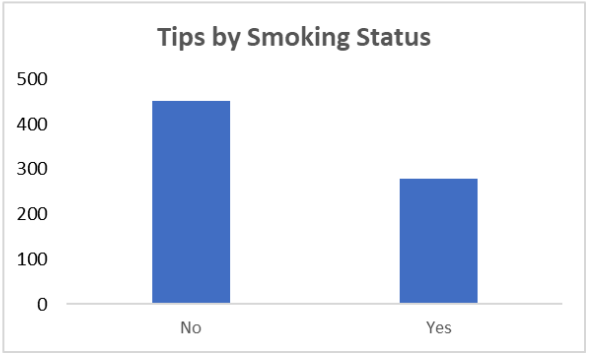
Exploratory Data Analysis

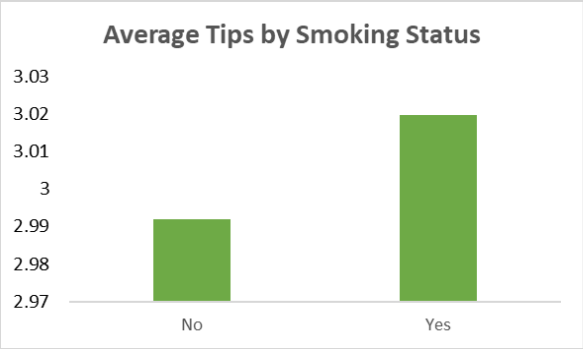
Helps in understanding the data, finding patterns and anomalies through descriptive statistics and visualisations.

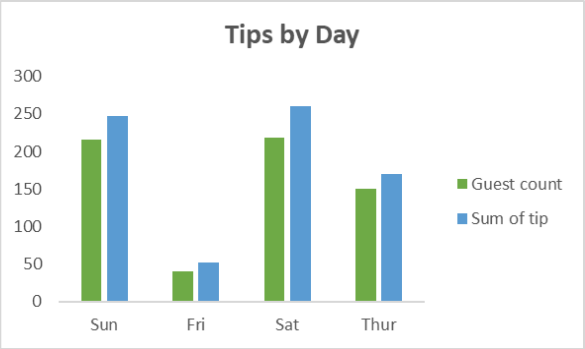












Row Labels	Sum of size	Average of size
Dinner	463	2.63
Lunch	162	2.42

Row Labels	Count of sex
Female	86
Male	157

EDA Visualisations

Insights:

- From the **`Tips by Sex`** graph, it seems like Men tend to give more tips than women but the number of men visiting is twice as much as 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.

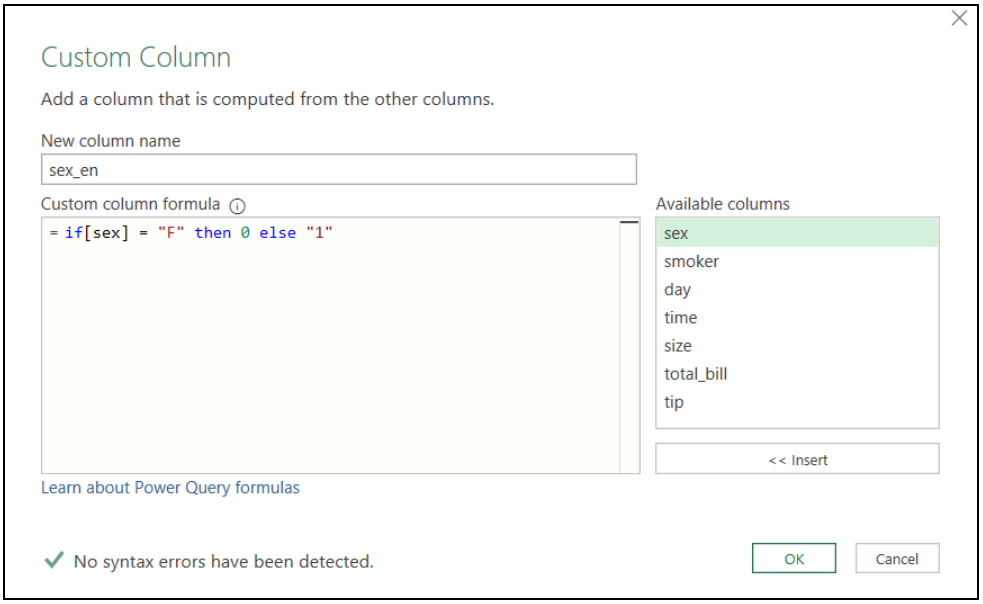
Feature Engineering

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'll encode the categorical columns using Power Query. Since model training can't be done with categorical values.
  Open Power Query>Right click on the existing query>Reference
  A reference sheet (or referencing a query) refers to the process of creating a new query that is based on an existing query. Instead of duplicating or copying the

query, the new query references the original one. Any changes you make to the original query will automatically reflect in the reference query.

- **Add Column>Custom Column>**



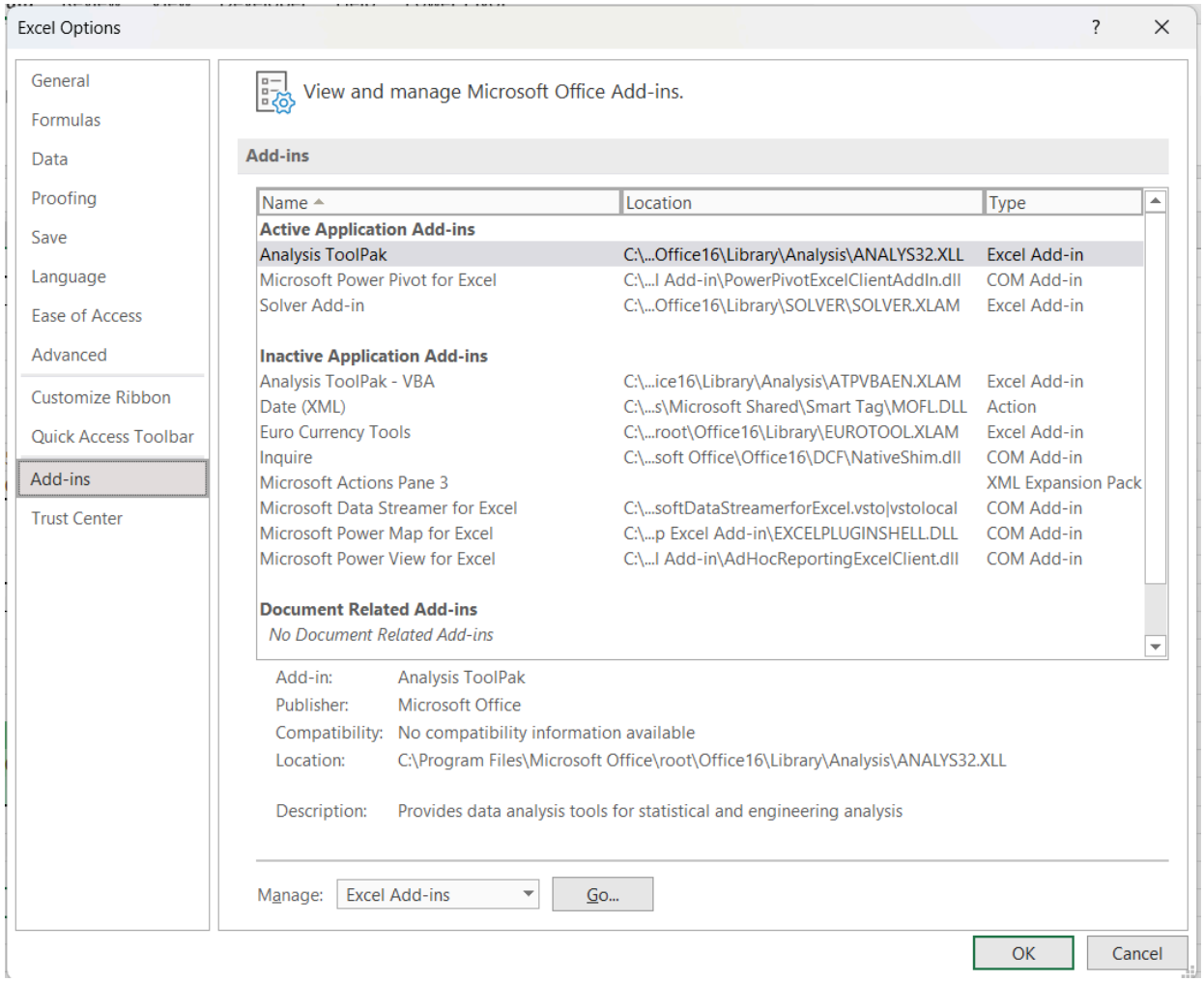
- Repeat for all the categorical columns
  - Smoker: No = 0, Yes = 1
  - Time: Dinner = 0, Lunch = 1
  - Day: For Sunday to Saturday use 1 to 7 respectively
- Remove the original columns `sex`, `smoker`, `time`, `day` columns
- Reorder the column similar to the original table. **Rename the table >Close and Load**

1	sex_en	smoker_en	time_en	day_en	size	total_bill	tip
2	0	0	0	1	2	16.99	1.01
3	1	0	0	1	3	10.34	1.66
4	1	0	0	1	3	21.01	3.5
5	1	0	0	1	2	23.68	3.31
6	0	0	0	1	4	24.59	3.61
7	1	0	0	1	4	25.29	4.71
8	1	0	0	1	2	8.77	2
9	1	0	0	1	4	26.88	3.12
10	1	0	0	1	2	15.04	1.96

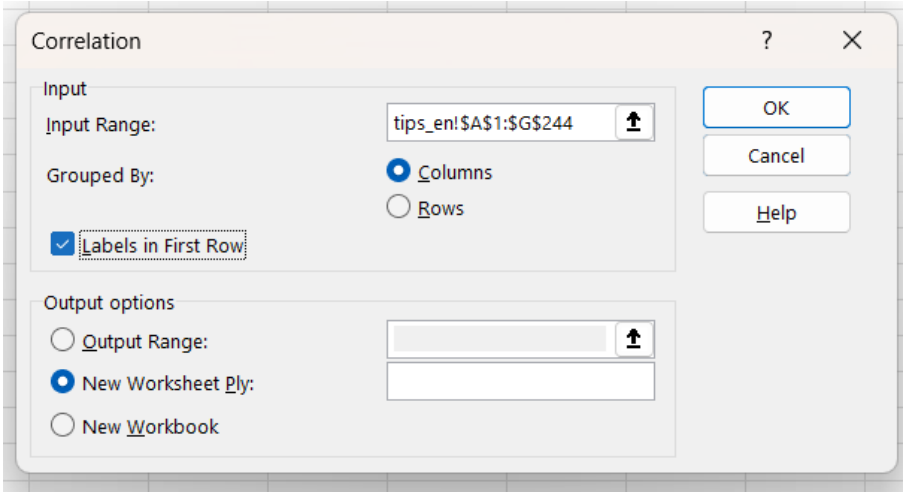
working_table

Feature Selection

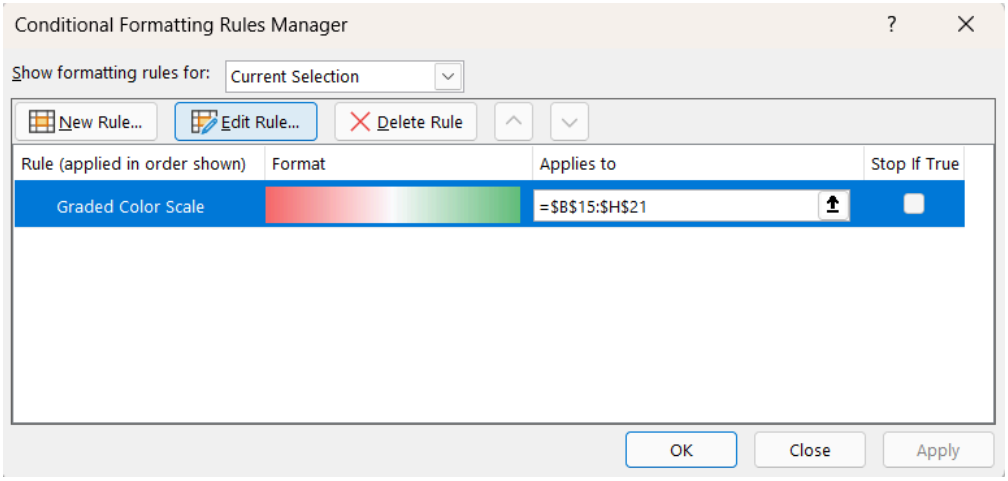
- Finding independent features and target variables for the regression model. Since we want to predict tips, `tips` will be dependent variable aka target variable and rest of the variables are independent variables or predictors.
- Finding covariance and correlation and multicollinearity to select independent variables with strong relationship with the target variable but not with one another. For further steps make sure the Data Analysis Toolpak is added. If not **Files>Options>Add-ins>Click on Analysis ToolPak> Manage**(at the bottom)**>Go>Ok**



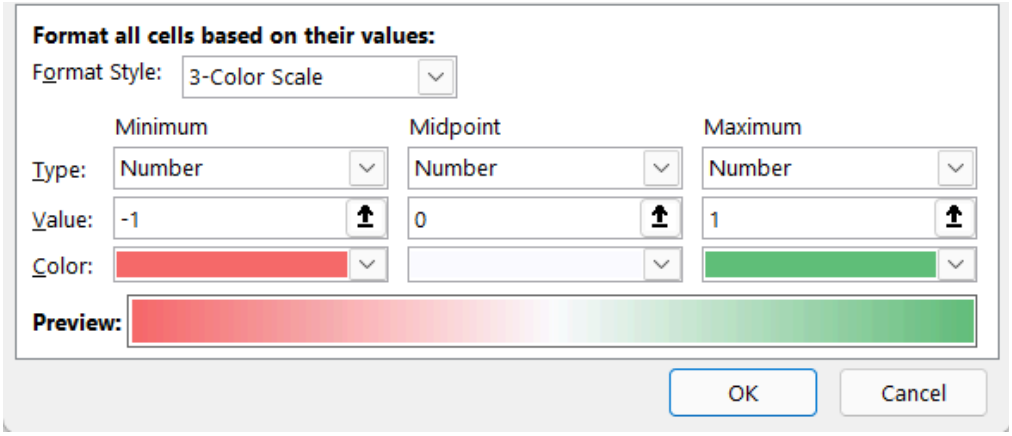
- Now go to the **Data tab>Analyze>Data Analysis>Correlation.** Input range is the entire table.



- Do the same for covariance
- I have applied the conditional formatting for correlation.
  Home>Style>Conditional Formatting>Colour Scales>Pick whichever you like
  Home>Style>Conditional Formatting>Manage Rules>Select Rule>Edit Rule



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



Results:

Covariance							
	sex_en	*smoker_en*	*time_en*	*day_en*	*size*	*total_bill*	*tip*
sex_en	0.228658						
smoker_en	0.002303	0.235262					
time_en	-0.04234	-0.01385	0.199699				
day_en	-0.13286	0.267591	0.154753	6.305746			
size	0.037833	-0.06019	-0.04249	-0.41361	0.90325		
total_bill	0.600999	0.388741	-0.71218	-1.75564	5.050009	79.06266	
tip	0.056359	0.006545	-0.07263	-0.34377	0.641557	8.295509	1.910337
Correlation							
	sex_en	*smoker_en*	*time_en*	*day_en*	*size*	*total_bill*	*tip*
sex_en	1						
smoker_en	0.00993	1					
time_en	-0.19813	-0.06391	1				
day_en	-0.11064	0.219699	0.137906	1			
size	0.083248	-0.13056	-0.10005	-0.17331	1		
total_bill	0.14135	0.090136	-0.17923	-0.07863	0.597589	1	
tip	0.085274	0.009763	-0.1176	-0.09905	0.4884	0.674998	1

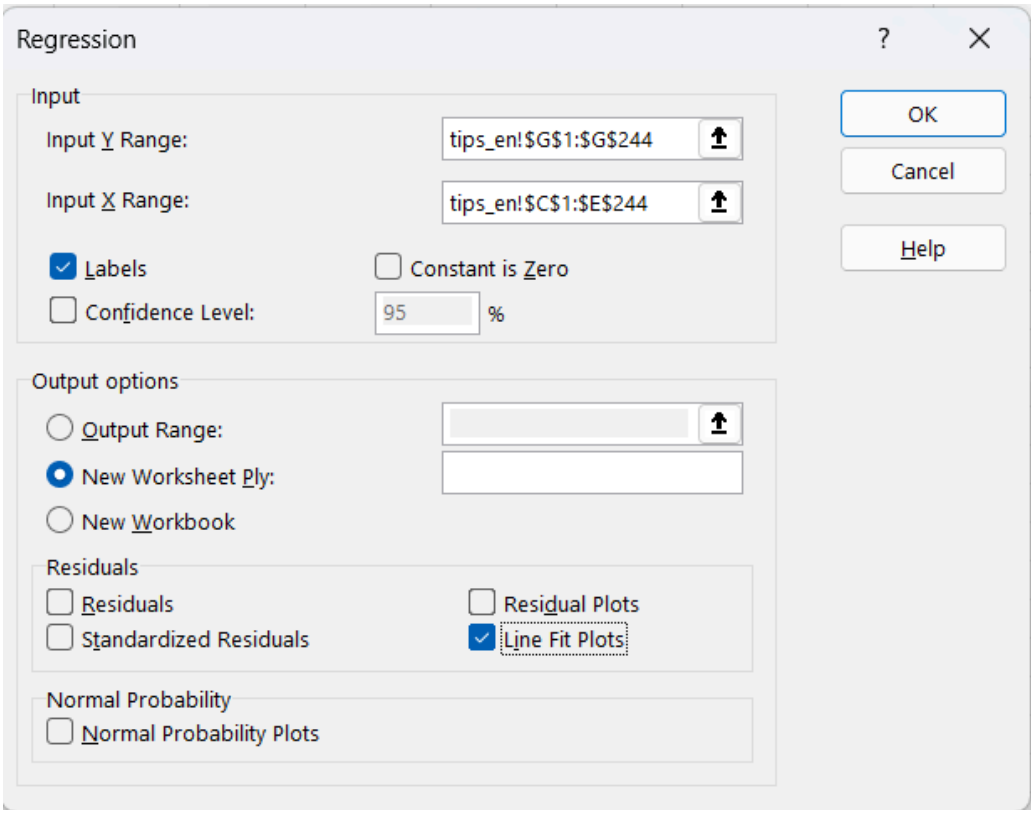
Covariance and correlation table

Insights:

- Two columns `smoker` and `sex` have correlation 0.009 and 0.08 respectively, show a weak relationship with **tip**, so they may not contribute significantly to the model or have a nonlinear relationship with tip (since correlation coefficient only checks for linear relationship, in case of mentioned cases, 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.
- `day`, `size` and `total_bill` are positively correlated to the target variable, `tip`. While `time` is negatively correlated to `tip`.

Model Building

- Since there are more than one feature affecting the tip, multiple linear regression models will be used(which is different from multivariate regression).
  Data>Analyze>Data Analysis>Regression



Output and Model Evaluation Measures

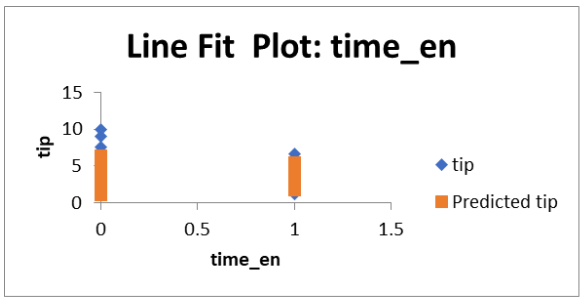
R-squared

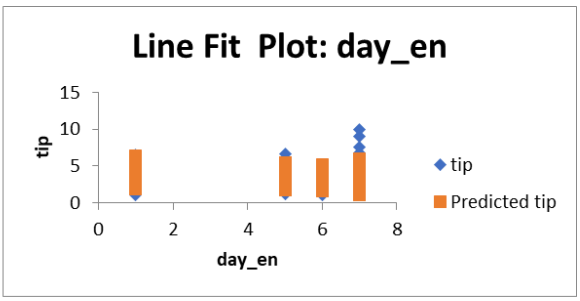
Regression Statistics	
Multiple R	0.68395383
R Square	**0.46779284**
Adjusted R Square	0.45884818
Standard Error	1.01884935
Observations	243

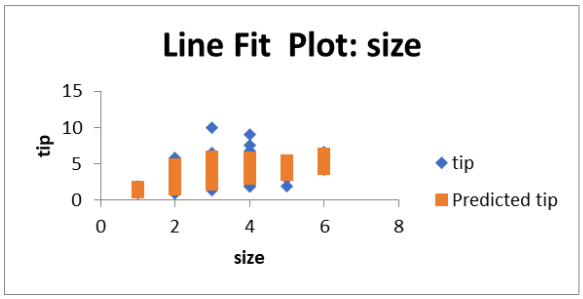
p-values

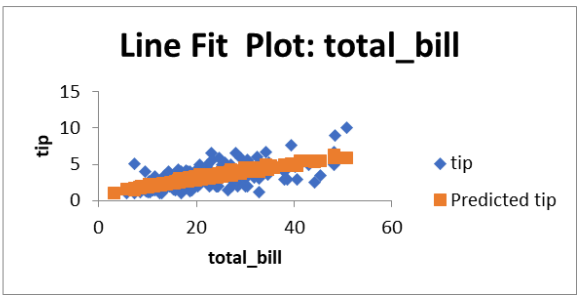
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.75984875	0.250420973	3.03428559	0.00267865	0.266524058	1.25317344	0.266524058	1.25317344
time_en	0.0202549	0.149929543	0.135096133	0.89265005	-0.27510353	0.31561333	-0.27510353	0.31561333
day_en	-0.017089	0.02666475	-0.64088243	0.52221568	-0.06961804	0.0354401	-0.06961804	0.0354401
size	0.18336989	0.086899346	2.110141191	0.03589024	0.012179783	0.35456	0.012179783	0.35456
total_bill	0.09301373	0.009284134	10.0185678	6.1329E-20	0.074724157	0.1113033	0.074724157	0.1113033

Line Fit









Residuals

Observation	*Predicted tip*	*Residuals*
1	2.68980283	-1.679802833
2	2.25463142	-0.594631417
3	3.24708792	0.252912082
4	3.31206469	-0.002064688
5	3.76344696	-0.153446961
6	3.82855657	0.881443428

RSME

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

RMSE	1.00831288

Conclusion

The model can predict restaurant tips with moderate accuracy, as indicated by a Root Mean Square Error (RMSE) of 1.0083 and an R-squared of 0.46, which means the model explains 46% of the variability in tips. However, only the variables `size` and `total_bill` are statistically significant, suggesting that these two variables have a meaningful impact on predicting tips,other variables do not show statistical significance and may not have a strong influence on predicting tips.