**Predicting Restaurant Tips: Multiple Linear Regression in Excel**

Problem Statement

You are a data analyst working for a restaurant chain. The management has provided you with a dataset 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.

Project Workflow

* Data was provided by the organization.
* The data was cleaned by removing duplicate rows, errors, and blank rows. The bar beneath the column name displays the percentage of errors and blank rows in the power query. While there currently aren't such rows, this step will automate data cleaning as more data comes in.
* Exploratory Data Analysis (EDA) is used to understand the data, identify patterns, and detect anomalies through descriptive statistics and visualizations.
* We'll utilize pivot tables to analyze and interpret the patterns in our data and also draw insights through data visualization

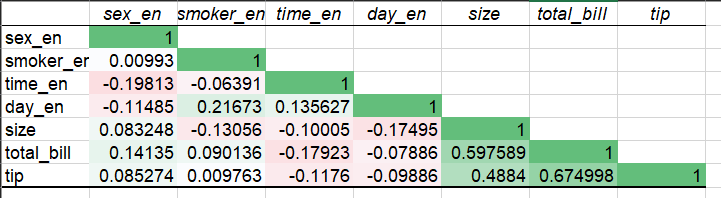
**Insights**

* According to the "Tips by gender" graph, men seem to give more tips than women. However, the number of men visiting to the restaurant is double than that of 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.
* With the increase in size i.e. guests in a table, the increase in the tips is not as high as the change in the amount of tip.

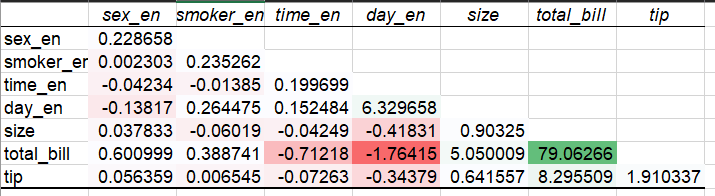
Next, finding the covariance, correlation and multicollinearity will help in selecting independent variables with strong relationship with the target variable but not with one another.

Firstly, we’ll decide the independent variables and dependent 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.

**Correlation**



**Covariance**



Insights

* Two columns smoker and sex have correlation 0.009 and 0.08 respectively, which shows 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 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.
* day, size and total bill are positively correlated to the target variable, tip. While time is negatively correlated to tip.

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.