**Simplilearn**

PC BA DEC 2022 Cohort 1

**Predicting Restaurant tips**

**Regression model using Excel**

**Analyst – Manohari Wijesooriya**

Contents

[1. Business case 2](#_Toc129463247)

[2. Cleaning Data 2](#_Toc129463248)

[3. Variable features 2](#_Toc129463249)

[4. Encoding data 3](#_Toc129463250)

[5. Exploratory Data Analysis 3](#_Toc129463251)

[6. Linear Regression Model 6](#_Toc129463252)

[7. Conclusion 9](#_Toc129463253)

# Business case

A local restaurant gathered amount of tip their customer make on the bill. Additional customer data is gathered about customers dining in the restaurant. The management wants to know what variables motivate customers to give a tip.

Proposed solution: Build a multiple linear regression model to predict restaurant tip.

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 |

First 6 records from the data file.



# Cleaning Data

Each column in Excel data file was examined to see if there are blank records. There were no blank records in the file.

# Variable features

To build the regression model, dependant and independent variables are identifies as follows.

Dependant variable: **tip**

Independent variables: **sex, smoker, day, time, size, total\_bill**

# Encoding data

Variables sex, smoker, day, time is categorical variables. We need to change those for numeric values to use these variables in regression analysis.

Formulas used for encoding:

sex\_M =IF(C2="Male",1,0) , sex\_F =IF(C2="Female",1,0)

smoker\_Y =IF(F2="Yes",1,0) , smoker\_N =IF(F2="No",1,0)

day\_THU =IF(K2="Thur",1,0)

day\_FRI =IF(K2="Fri",1,0)

day\_SAT =IF(K2="Sat",1,0)

day\_SUN =IF(K2="Sun",1,0)

time\_D =IF(K2="Dinner",1,0)

time\_L =IF(K2="Lunch",1,0)

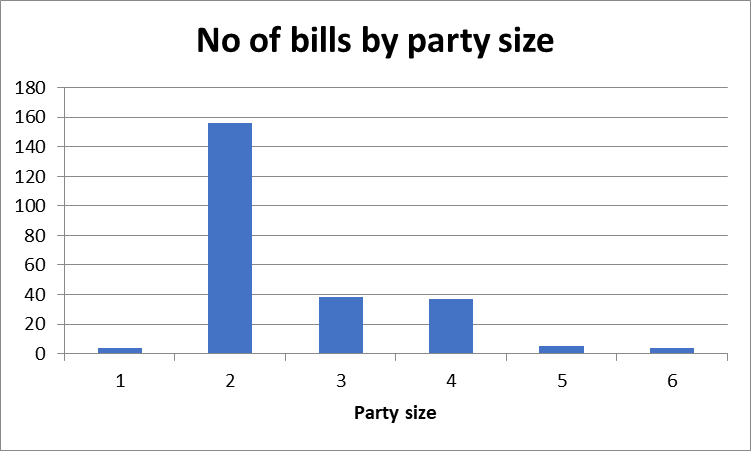
# Exploratory Data Analysis

These charts are prepared to understand data and to identify patterns in data.



* Tip has high correlation with size and total bill.
* Restaurant has more clients on Thursday for lunch. Also more clients on Saturday and Sunday for dinner.
* High correlation in total bill (as expected )
* no direct correlation with sex and smoker variables

Visualisations to understand data patterns –



Tip amount increases as the total bill increase

In most of the times, the bill amount lies in the range $11 - $20

# Linear Regression Model

A linear regression model is built to predict restaurant tip.

Dependant variable: **tip**

Independent variables: **sex, smoker, day, time, size, total\_bill**



R squared is Coefficient of determination. In this model 47% of values fit the model.



Significance F value is the significance associated P-value. Lower the p-value the better model.

Interpret Regression co-efficient



We see lower p-values for intercept and total bill, which is indication of good fit. T stat 65535 and #NUM for p-value indicates Excel couldn’t out put the extremely smaller p-value and it is showing as error.



In the regression co-efficient chart, Days THU,FRI and SAT shows higher p-value.

We can drop Day field from the independent variable list, run the regression model and check regression output.

Model2 –

Dependant variable: **tip**

Independent variables: **sex, smoker, time, size, total\_bill**



**Compare 2 models**



We select model 1 as the best model considering higher R squared and lower root mean square error.

# Conclusion

Independent variables sex, smoker, day, time, size and total\_bill can be used to predict tip with 47% accuracy using linear regression model.



