An
INTERNSHIP PROJECT REPORT ON

Analysis and prediction of waiter's Tips Amount in a Restaurant

A Project Report for Internship Programme

Submitted by

MAHASWETA CHAKRABORTY
ANWESHA MANNA
JINIA GHOSH
BAIDURYAY NARAYAN BASU

In partial fulfillment for the award of the degree of

Master of Administration

in

Business Analytics

MCKV Institute of Engineering

At

Euphoria GenX

DECLARATION

I undersigned, hereby declare that the project titled "***Analysis and prediction of waiter's Tips Amount in a Restaurant***" submitted in partial fulfilment for the award of Degree of ***Master of Business Administration*** of ***Maulana Abul Kalam Azad University of Technology (MAKAUT)*** is a bonafide record of work done by me under the guidance of ***Animesh Ojha***. This report has not previously formed the basis for the award of any degree, diploma, or similar title of any University.

09/10/2021 **Mahasweta Chakraborty**

Euphoria
genx
EXPLORE YOURSELF LIKE NEVER BEFORE

BONAFIDE CERTIFICATE

Certified that this project work was carried out under my supervision

“Development **of a feature-rich, Analysis and prediction of waiter’s Tips Amount in a Restaurant**” is the bonafide work of

Name of the student	***Signature***
MAHASWETA CHAKRABORTY	Mahasweta Chakraborty
ANWESHA MANNA	Anwesha Manna.
JINIA GHOSH	Jinia Ghosh
BAIDURYAY NARAYAN BASU	Baiduryay Narayan Basu

SIGNATURE

Name :
ANIMESH OJHA

MCKV Institute of Engineering
MBA in Business Analytics
(Institute Emblem)

CERTIFICATE

*This is to certify that the report titled "**Analysis and prediction of waiter's Tips Amount in a Restaurant**" being submitted by **Anwesha Manna**, (1161010003), in partial fulfilment of the requirements for the award of the Degree of **Master of Business Administration**, is a bonafide record of the project work done by Anwesha Manna of MBA in Business Analytics, MCKVIE.*

Name of guide – ANIMESH OJHA
Designation –
Director

ACKNOWLEDGEMENT

Through this acknowledgement we express our heartfelt gratitude towards all those who have helped us in this project, which has been a learning experience.

Firstly, we would sincerely thank the **Lord Almighty** for his adherent grace, mental strength, blessings and unconditional love bestowed on us throughout our study.

Next, we would like to express our special thanks of gratitude to **Prof. (Dr.) Arghya Sarkar**, Principal, MCKV Institute of Engineering and **Prof. (Dr.) Debapriya De**, the Board of Controller, MCKV Institute of Engineering for providing us this opportunity.

This space would not be enough to extend our warm gratitude towards **Mr. Anirban Banerjee**, the esteemed CEO of Euphoria GenX and **Mr. Tridib Kr Singha**, respected HR of Euphoria GenX who have provided us with this internship opportunity and proper coordination to keep us going.

Our project was successfully carried out under the guidance of **Mr. Animesh Ojha** who has extended his valuable guidance and knowledge sharing to keep us motivated and rooted towards this project.

We hereby take this opportunity to express our deepest regards to our Department Head, **Prof. Diptayan Bhattacheryya** and our Educational Officer **Prof. Ambarish Chatterjee** for continuously boosting up our confidence to embark in this journey.

It would be injustice to proceed without acknowledging those additional yet essential supports we received from our faculties **Mr. Partha Pratim Saha, Mr. Nilay Kr. Nag, Mr. Subir Bhadra, Mr. Abhisek Saha, Mr. Avijit Bose, Dr. Arijit Ghosh, Dr. Sudip Mukherjee, Dr. Shampa Sengupta**. We are really fortunate and blessed to have their cooperation and thorough guidance.

We would further like to thank all our **fellow classmates and friends**, without whom we would not have the right amount of confidence, courage and dedication to complete the project on time.

We are very much thankful to all **fellow members of the project team** who have worked hand in hand and showcased an excellent team work and perfect coordination in finishing the project.

Finally, last but not the least, we would like to use this space to offer our sincere love from the core of our hearts to our **parents, siblings and relatives** who have always been there by our side, living with us through this project, showering us with their endless cooperation, love, care and support that altogether has given a shape and form to our study.

Mahasweta Chakraborty
Anwesha Manna
Jinia Ghosh
Baiduryay Narayan Basu

List of Chapters and Tables

List of Figures

List of Symbols/ Abbreviations

Symbol/ Abbreviation	Explanation	Page No.
np	Alias name for python package Numpy (for taking array of data as required, for mathematical computation etc.)	**14**
pd	Alias name for python package Pandas (for reading dataset, storing data into dataframes, showing charts and diagrams etc.)	**14**
plt	Alias name for python package Matplotlib.pyploy (for plotting graphs, charts etc.)	**14**
sns	Alias name for python package Seaborn (for generating various plots, invoking in-built dataset)	**14**
Fri	Short form for Friday; used in Figure 8	**13,18**
Sat	Short form for Saturday; used in Figure 8	**13,18**
Sun	Short form for Sunday; used in Figure 8	**13,18**
Thur	Short form for Thursday; used in Figure 8	**13,18**
Mon	Implies Monday; used to frame questionnaire	**13**
Tue	Implies Tuesday; used to frame questionnaire	**13**
Wed	Implies Wednesday; used to frame questionnaire	**13**

TABLE OF CONTENTS

CHAPTER 1

ABSTRACT

With the greater development of technology and automation human history is predominantly updated. The technology movement shifted from large mainframes to PCs to cloud when computing the available data for a larger period. This has happened only due to the advent of many tools and practices, that elevated the next generation in computing. A large number of techniques has been developed so far to automate such computing. Research dragged towards training the computers to behave similar to human intelligence. Here the diversity of machine learning came into play for knowledge discovery. Machine Learning (ML) is applied in many areas such as medical, marketing, telecommunications, stock, health care and so on.

Here, an attempt has been made to study and most importantly make the machine learn the data based on “tips” dataset that showcases the amount of tip received by a restaurant’s waiters based on meal timings, day of week and from different category of customer like male and female, smoker or non-smoker. The data would be fit into a suitable model whereby, after training, it will be able to predict the amount of tips in future.

This paper presents reviews about machine learning algorithm foundations, its types and flavours together with enriched visualization using Python scripts for each machine learning techniques.

Keywords: Machine Learning, tips, visualization, Python, automation

CHAPTER 2

INTRODUCTION

ML refers to the methods tangled in distributing through massive facts in the greatest intellectual way to arise better understandings. ML algorithms are defined to be culturing an objective function (f) which better draws input identifier (g) to an output identifier (h) as in equation:

$$h = f(g)$$

This future output prediction is not that much easier to do manually. Hence an automated system is expected to do the process. Thus, use of machine learning algorithms come into the scene. For every new input (g) the output (h) is predicted genuinely using machine learning algorithms. This state is said to be predictive analytics. The major operation is to assess he most possible predictions with the present data. Each data is segregated as training set and testing set as in figure:

Five basic steps for an ML task:

1. **Data accumulation**: Data gathered from various sources are used for analysis.
2. **Data pre-processing**: Before getting into the actual processing of data, pre-processing is mandatory. This step is used to noise or other unwanted data from the gathered data.
3. **Prototype training**: This step contains selecting the suitable algorithm and depiction of data in a pattern (model) format. The pre-processed data is often divided into two parts namely training and testing data.
4. **Pattern evaluation**: In this step, the resultant pattern is validated for its correctness. However substantial time is required in data gathering and training.
5. **Performance enrichment**: This step involves picking another different pattern with better efficiency.

Supervised learning (SL) or prognostic models

This is used to assess the upcoming result with the help of chronological data. These models are instructive as much concentration is emphasized in training phase. For instance, SL is applied if a selling firm wishes to find its customers list. It could also be used in prediction of earthquakes, cyclones etc. to determine the Insurance credit. Few examples of these prediction algorithms are: multiple linear regression logistic regression Naïve Bayes, Decision Trees

Unsupervised learning (UL) or evocative models UL is suitable to train vivid models with no target and no sole feature is significant compared to one another. For instance, UL is applied in case if a vender desires to find which product does the customer buys frequently. Moreover, in medicinal business, UL may be applied to envisage the diseases that may prone to occur laterally with diabetes. Few examples of UL based algorithms are, Simple K- means clustering.

Reinforcement learning (RL) RL is applied when the system is trained to yield decisions automatically with the business requirements only with a sole motto to exploit better effectiveness (performance). The underlying idea is a software agent is trained in an environment for problem solving. This repeated learning procedure promises lower human proficiency thus saving human effort. One best example of RL algorithm is Markov Decision Process.

2.1 Background of the study

The dataset in this section is the so-called Food servers' tips in restaurants. In one restaurant, a food server recorded the following data on all customer they served during an interval of two and a half months in early1990.The restaurant, located in Albany is the capital city of New York State and served a varied menu. In observance of local law, the restaurant offered to seat in a non-smoking section to patrons who requested it. Each record includes a day and time, and taken together, they show the server's work schedule.

2.2 Need and Significance of the study

Food servers' tips in restaurants may be influenced by many factors, including the nature of the restaurant, size of the party, and table locations in the restaurant. Restaurant managers need to know which factors matter when they assign tables to food servers. For the sake of staff morale, they usually want to avoid either the substance or the appearance of unfair treatment of the servers, for whom tips (at least in restaurants in the United States) are a major component of pay.

2.3 Statement of Problem

Analysis of the amount of tips received by a restaurant based on some factors and prediction of the tips amount based on those factors.

2.4 Objective of the study

The objectives of the study of Tips dataset are as follows: -

- **GENERAL OBJECTIVE**
  - To find meaning in data so that the derived knowledge can be used to make informed decisions and help business organization to run more profitably.
  - To analyse the overall dataset and find out how the factors are affecting the tipping practice of customers.

- **SPECIFIC OBJECTIVE**
  - To examine which type of customer are highly correlated with high tips payment through Correlation analysis of the dataset.
  - To fit the data in a suitable model in order to make the machine learn the pattern.
  - To predict amount of tips to be received by the restaurant in future.

2.5 Scope of the study

This study aims to analyse the accuracy of predicting tipping amount using Machine Learning, Python Visualization & linear regression algorithms. Thus, the purpose of this study is to deepen the knowledge in regression methods in machine learning. In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods to eliminate the unwanted variables, keeping only the relevant ones. These features may or may not be shared with all tips, which means they do not have the same influence on the tipping resulting in inaccurate output.

2.6 Limitation of the study

- The Results generated from the Questionnaire are done on the assumption that the respondents have revealed the correct information.
- Our study report confined to sample size of 244 respondents only.
- The period of study was not sufficient to study all aspects.
- The study was restricted to a specific restaurant only.
- Due to the brevity of time and resource, all the possible factors might not have been taken into consideration.
- Some vital factors like quality of food, quality of service etc. were not taken into account while collecting data which could have otherwise provided better results.

2.7 Organization of the Report

The dataset used here is the **“tips”** dataset which is available as an in-built dataset under the seaborn package in python. Hence, all the reports and visualizations were generated based on **seaborn, pandas** and **matplotlib** packages.

In this study, we have also referred to the dataset from a third-party website for better usage of the **pandas** package.
All the reports generated, were stored in local directory in the form of **.jpg image files** for ease of retrieval while working with the research paper.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Objectives

Research is a careful and detailed study of a particular problem or concern, using scientific methods. An in-depth analysis of information creates space for generating new questions, concepts and understandings. The main objective of research is to explore the unknown and unlock new possibilities. It is an essential component of success.

❖ GENERAL OBJECTIVE

Also known as secondary objectives, general objectives provide a detailed view of the aim of a study. In other words, we get a general overview of what we want to achieve by the end of our study. Here our general objective is –

- to analyse how much tips various categories of people give.
- to identify if all the identified factors affect the amount of tips.

❖ SPECIFIC OBJECTIVE

Specific objectives define the primary aim of the study. Typically, general objectives provide the foundation for identifying specific objectives. In other words, when general objectives are broken down into smaller and logically connected objectives, they are known as specific objectives. They help define the who, what, why, when and how aspects of a project. Once we identify the main objective of research, it is easier to develop and pursue a plan of action. Here we are trying to find out answers to the following –

- “Does the total bill affect the amount of tip received?”
- “Does the size of the customer per table affect the tipping amount?”
- “How does the other demographics influence tip amount?”

- “What can be the possible model to fit this data?”
- “How can the future tip amount be predicted?”

3.2 Hypothesis

The hypothesis of the “tips” data analysis are as follows –

- ❖ **NULL HYPOTHESIS**

  The factors considered in the study do not affect and are not correlated with the amount of tips received by the restaurant.

- ❖ **RESEARCH HYPOTHESIS**

  The factors considered in the study significantly affect and are correlated with the amount of tips received by the restaurant.

3.3 Research Design

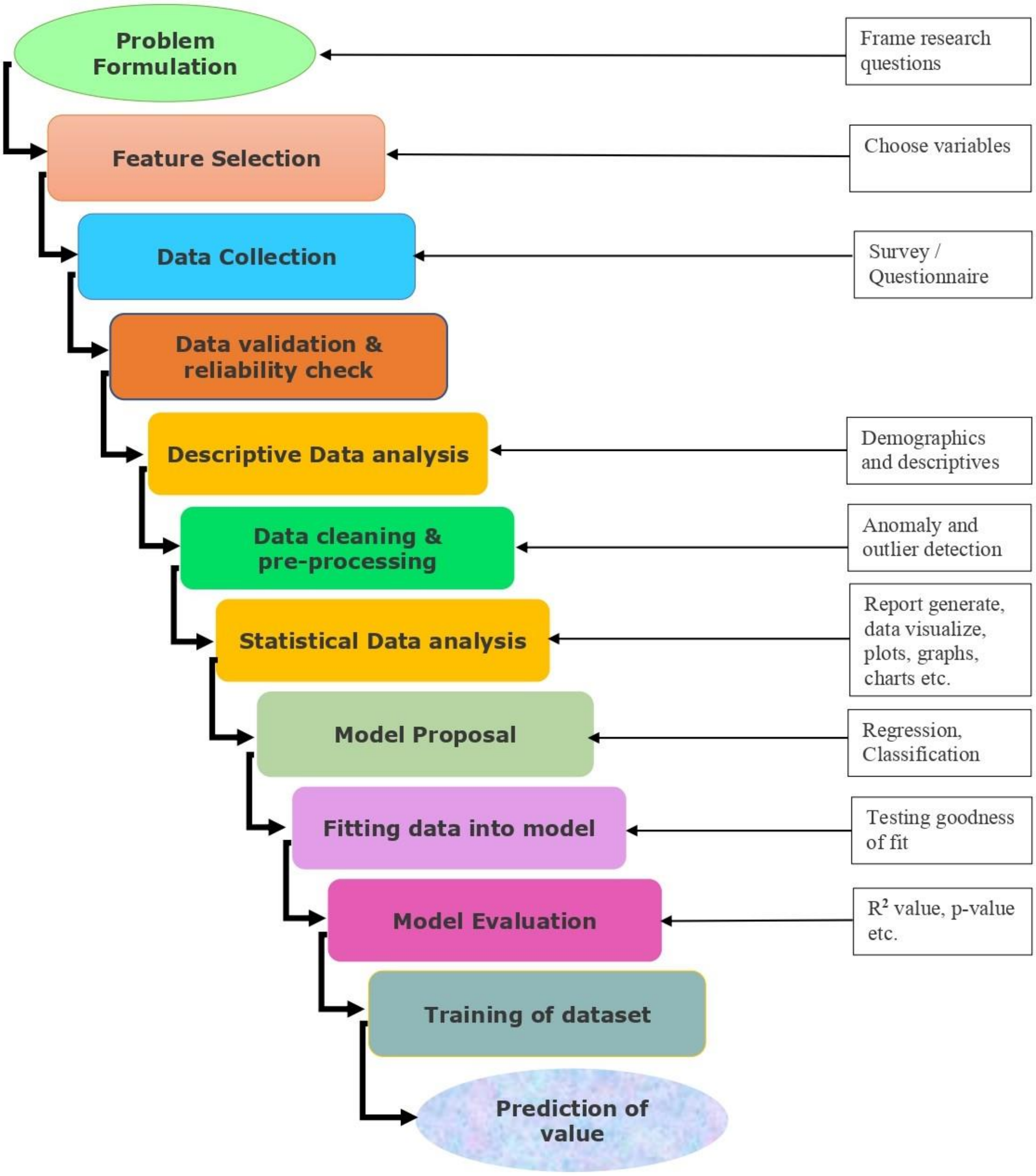

Problem Formulation
Feature Selection
Data Collection
Data validation & reliability check
Descriptive Data analysis
Data cleaning & pre-processing
Statistical Data analysis
Model Proposal
Fitting data into model
Model Evaluation
Training of dataset
Prediction of value
Frame research questions
Choose variables
Survey / Questionnaire
Demographics and descriptives
Anomaly and outlier detection
Report generate, data visualize, plots, graphs, charts etc.
Regression, Classification
Testing goodness of fit
R² value, p-value etc.

Fig 1: Research design for study of tips dataset

3.4 Software Requirement

The main software used in this project is as follows: -

1) **Python**:

   Python is an interpreted, high level programming language created by Guido ban Rossum in 1991 it emphasizes on code readability by using whitespace to terminate statements and blocks. Its ease of use and syntax makes it a language of preference for data analysis.

2) **Spyder**:

   It is an application that provides an integrated development environment for python scripting. Using this one can share documents that contains equation, visualizations such as graphs, test as a live, for this reason it is highly used tool for the purposes of data analysis.

3.5 Tools Description

The tools that helped in achieving the required data visualization and calculations are basically the ***python packages***. The following packages used in this project are described below: -

1) **NumPy**:

   NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

2) **Pandas**:

   Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays.

3) **Matplotlib**:

   Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension.

4) **Seaborn:**

   Seaborn is a package under python which provides a considerable number of datasets to perform machine learning, data visualization etc. It also chiefly provides room for exhibiting statistical nature of data through wide varieties of plotting techniques like pairplot, distplot, jointplot etc.

5) **Scikit-Learn**:

   It is a python package used for data modelling. It provides a number of supervised and unsupervised machine learning models. Scikit-learn makes it extremely simple to train models with simple function calls on the input data being all that is needed.

CHAPTER 4

DATASET (RESTAURANT TIPS)

4.1 Source of data

- **seaborn.load_dataset** : This function provides quick access to a small number of example datasets that are useful for documenting seaborn or generating reproducible reports.
- **Origin**: “tips” dataset is based on the study of Food servers’ or waiter’s tips in restaurants. In one restaurant, a food server recorded the following data on all customer they served during an interval of two and a half months in early 1990.
- **Features of the dataset**: There are 7 attributes or features in the above dataset –

  "total_bill", "tip", "sex", "smoker/non-smoker", "day", "time", "size".

Total_bill: total bill which is collected by the different customer during this whole time period of recorded

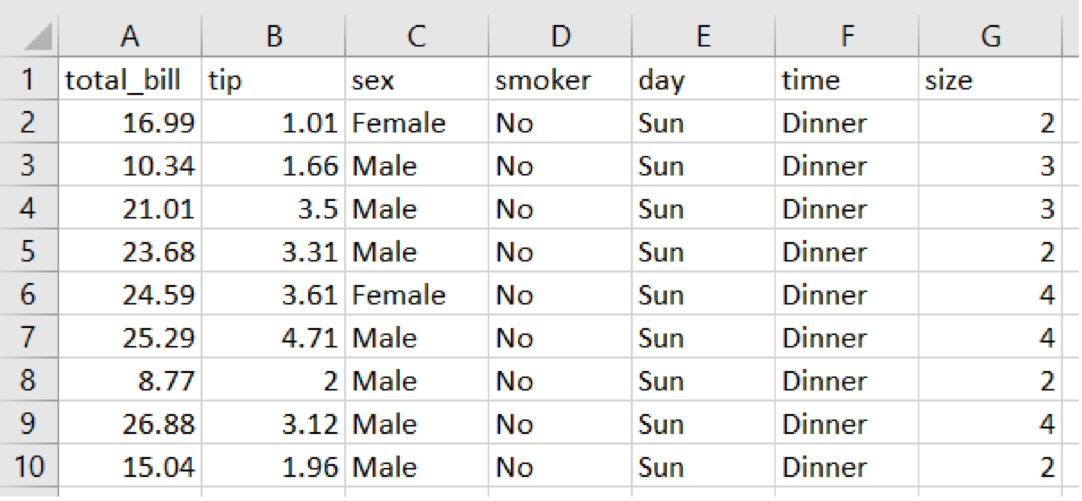
Tip: total tip awarded by customer based on sex or gender

Sex: represent the number of male and female during this whole period of study of record

Smoker/Non-smoker: the customer who involved in smoking and paying tips

Day: day represent the particular week days in which customer used to visit restaurant

Time: time represent the particular hour like (lunch or dinner)



	A	B	C	D	E	F	G
1	total_bill	tip	sex	smoker	day	time	size
2	16.99	1.01	Female	No	Sun	Dinner	2
3	10.34	1.66	Male	No	Sun	Dinner	3
4	21.01	3.5	Male	No	Sun	Dinner	3
5	23.68	3.31	Male	No	Sun	Dinner	2
6	24.59	3.61	Female	No	Sun	Dinner	4
7	25.29	4.71	Male	No	Sun	Dinner	4
8	8.77	2	Male	No	Sun	Dinner	2
9	26.88	3.12	Male	No	Sun	Dinner	4
10	15.04	1.96	Male	No	Sun	Dinner	2

Fig 2: A snapshot of the tips dataset

4.2 Primary and Secondary Data

- **PRIMARY DATA**

  The data was primarily collected by a group of researchers from an US restaurant where a food server recorded the above data on all customers they served during an interval of two and a half months in early 1990.

  The data was originally mentioned in the following –

  (2007) Datasets. In: Interactive and Dynamic Graphics for Data Analysis. Use R!. Springer, New York, NY. https://doi.org/10.1007/978-0-387-71762-3_7

- **SECONDARY DATA**

  We have mainly collected the dataset from a website: https://www.kaggle.com/**,**
  in order to use the properties of pandas package.
  For sanity check, we have also retrieved the dataset from **seaborn.load_dataset** and verified the data.

4.3 Population

Population is the pool of individuals from which a statistical sample is drawn for study. Here in one restaurant, a food server recorded the following data on all customer they served during an interval of two and a half months in early 1990. Hence, the population denotes the entire restaurants' customers during the mentioned time period.

4.4 Sample size and sample design

- **Sample Size**

Sample size measures the number of individual samples measured or observations that are actually used in a study, a survey or an experiment.
In this case, the Sample size **(n) = 244**

- **Sample design:**

Sampling is a mathematical function that gives the probability of any given sample being drawn. Here we've used Random Sampling design.

4.5 Method of data collection

Here, secondary data has been used to conduct the research study. The dataset was taken from the following website → **https://www.kaggle.com/**

4.6 Drafting a questionnaire

Some of the basic practices that are needed to be followed while preparing a questionnaire are –

- Keep the questionnaire as short as possible.
- Ask short, simple, and clearly worded questions.
- Start with demographic questions to help respondents get started comfortably.
- Use dichotomous (yes | no) and multiple-choice questions.
- Use open-ended questions cautiously.
- Avoid using leading-questions.
- Pretest a questionnaire on a small number of people.
- Think about the way you intend to use the collected data when preparing the questionnaire.

The possible questions that were formulated for collection of data were –

1) What is the gender of the customer? [Male / Female]

2) Is the customer a smoker? [Yes / No]

3) On which day the customer visited? [Sun/Mon/Tue/Wed/Thur/Fri/Sat]

4) At what timing did the customer visit the restaurant? [Lunch/Dinner/Breakfast]

5) How many customers were sitting in the table? [1 / 2 / 3 / 4 / 5 / 6 / more than 6]

6) What is the total bill incurred by the customer? -------------------------------------

7) What amount of tip did the customer award to the waiter? ------------------------

CHAPTER 5

DATA ANALYSIS

5.1 Data analysis techniques

Data analysis, visualization and interpretation are the most essential parts of a machine learning study. The dataset used here is “Tips” containing N = 244 sample size.
The following techniques were used for **data analysis**:

- Descriptive statistics
- Outlier detection
- Comparative data plotting
- Correlation analysis
- Demographics for Gender (sex), smoker/non-smoker, time of day, day of week etc.
- Regression scatter plots for all columns

The following techniques were applied in python to obtain **data visualization**:

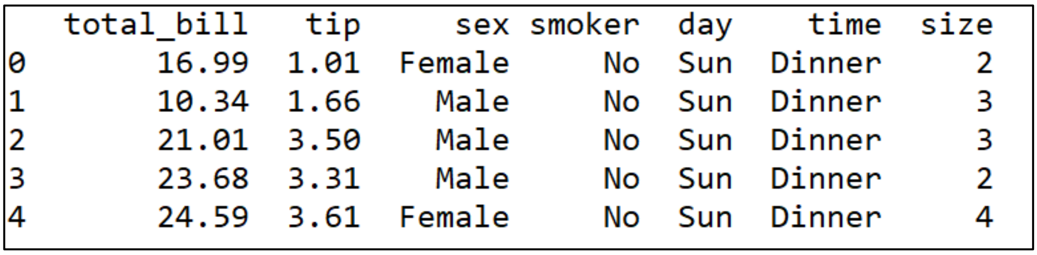
- Piechart, barchart and histograms for demographic data
- Boxplots for outlier detection
- Lineplots and subplots for overall data plotting
- Heatmap for correlation analysis
- Scatter plots for understanding relation among variables
- Jointplot, Distplot and pairplots for comparative regression analysis.

5.2 Project Coding – Visualization – Interpretation

1. Firstly, we included some necessary python packages and called the entire dataset from the saved location. Then we viewed the head( ) data:

```
#py packages
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
#Calling dataset
tips = pd.read_csv("C://NEW//Tips.csv")
tips.head()
```

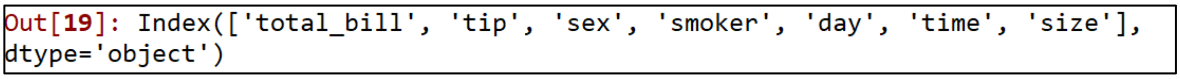


```
   total_bill   tip     sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun  Dinner     2
1       10.34  1.66    Male     No  Sun  Dinner     3
2       21.01  3.50    Male     No  Sun  Dinner     3
3       23.68  3.31    Male     No  Sun  Dinner     2
4       24.59  3.61  Female     No  Sun  Dinner     4
```

Fig 3: First 5 rows of the dataset displayed

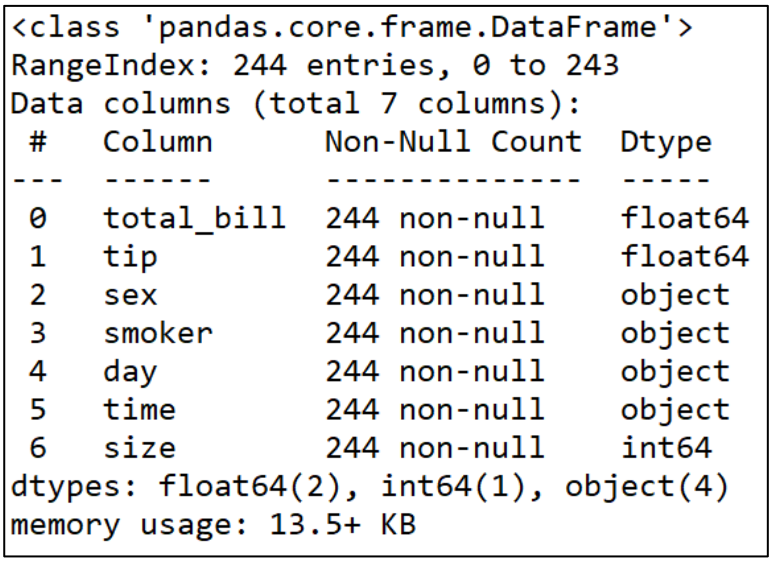
2. Then, we fetched some basic information like the columns, datatypes and the descriptive statistics of the data:

```
#Basic information
tips.columns
tips.info()

#Descriptive data
tips.tip.describe()
```



```
Out[19]: Index(['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size'],
dtype='object')
```

Fig 4: Columns of the dataset displayed



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    object 
 3   smoker      244 non-null    object 
 4   day         244 non-null    object 
 5   time        244 non-null    object 
 6   size        244 non-null    int64  
dtypes: float64(2), int64(1), object(4)
memory usage: 13.5+ KB
```

Fig 5: Column description of the dataset displayed

From the above data we can see what are all the columns of the dataset. Also, we can see the datatype of each data in each column.

Another important thing that becomes clear from the above output is that there is no-null value in any of the column which gives a get go towards statistical analysis.

Table 1 : Descriptive statistics for Tips

Count (N)	244.000000
mean	2.998279
std	1.383638
min	1.000000
25%	2.000000
50%	2.900000
75%	3.562500
Max	10.000000
Name: tip, dtype: float64	

The above table depicts the descriptive summary of our main variable of concern that is the **'Amount of tips received'**. As we can see, here sample size is 244, the average tip received is almost 3 with a standard deviation of 1.38. The minimum tip being 1 while the maximum being 10.

The values at the 25th quantile is 2, that at 50% quantile is the median which is 2.9, approximately 3. And, the value for the 75th quantile is around 3.56. This signifies that the data has very less to no anomaly and all the data falls within interquartile range.

3. Now, we come to the demographic data analysis part where we plot the gender percentages using piechart and then the variation of tip amount with gender, time of day, day of week, size of customers per table, smoker-non smoker criteria using bar charts.

```
#Demographics - % male and female pie chart
freqs = tips.sex.value_counts(normalize = True )
freqs.plot.pie(y='sex',figsize=(5,5),autopct='%1.1f%%')
```

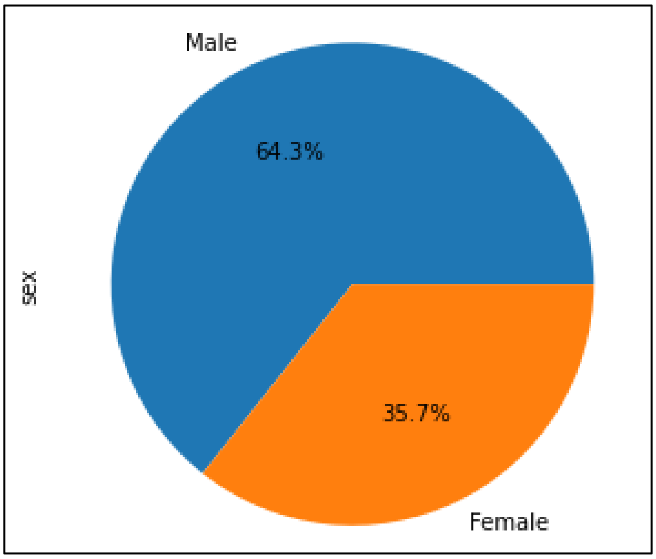


Fig 6: Male VS Female (in percentage)

The above figure shows the percentage of male and females among the customers giving tips. Here, 35.75 of the customers are females while 64.3% that is the majority are the males.

```
#Does the average tip differ by gender? Does one gender tip more than
the other?
tips.groupby(['sex'])['tip'].mean().plot.bar(color="violet",width=0.4)
```

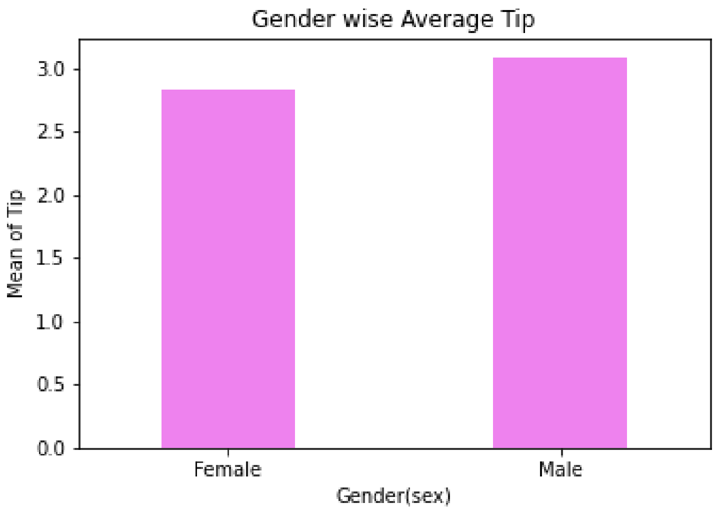
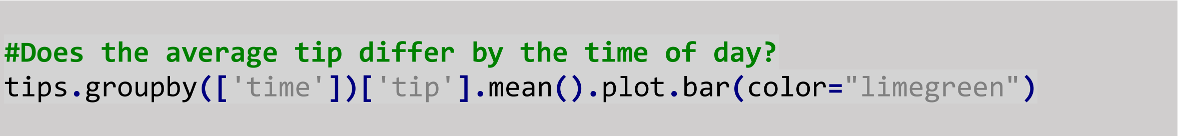


Fig 7: Variation of tip with Gender

The above figure captures the amount of tip contributed by each of the male and female groups. Here, average mean is contributed mostly by the male customers than the female customers.



```
#Does the average tip differ by the time of day?
tips.groupby(['time'])['tip'].mean().plot.bar(color="limegreen")
```

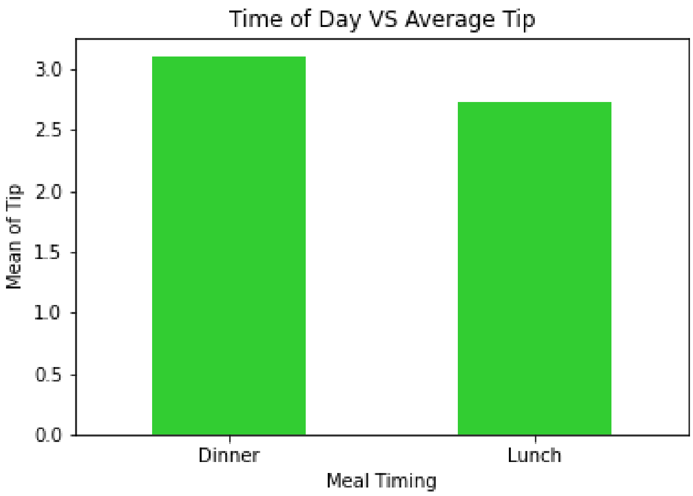
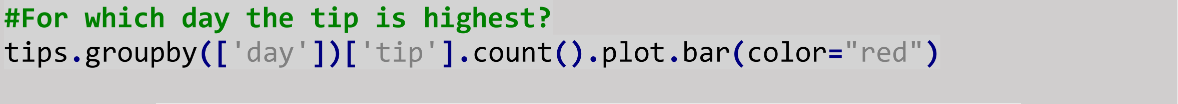


Fig 8: Variation of tip with time of day

From the bar graph above, we see that people pay more amount of tip at the time of dinner and during lunch hours, average amount of tips received is lesser as compared to dinner.



```
#For which day the tip is highest?
tips.groupby(['day'])['tip'].count().plot.bar(color="red")
```

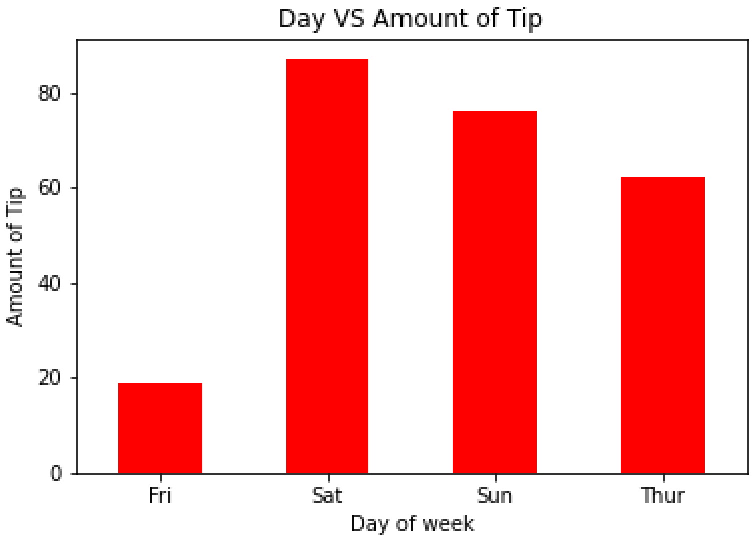


Fig 9: Variation of tip with day of week

Since, there is no null value in the Tip section, we can conclude that the restaurant receives tip on all days of the week. Again, since, Saturday and Sunday are mostly holidays, so naturally more people visit the restaurant and hence more amount of tip is received on those two days.

```
#Visualize the number of customers based on day of week
res=pd.pivot_table(data=tips,index="day",columns="size",values="tip")
sns.heatmap(res, annot=True, cmap="RdYlGn")
```

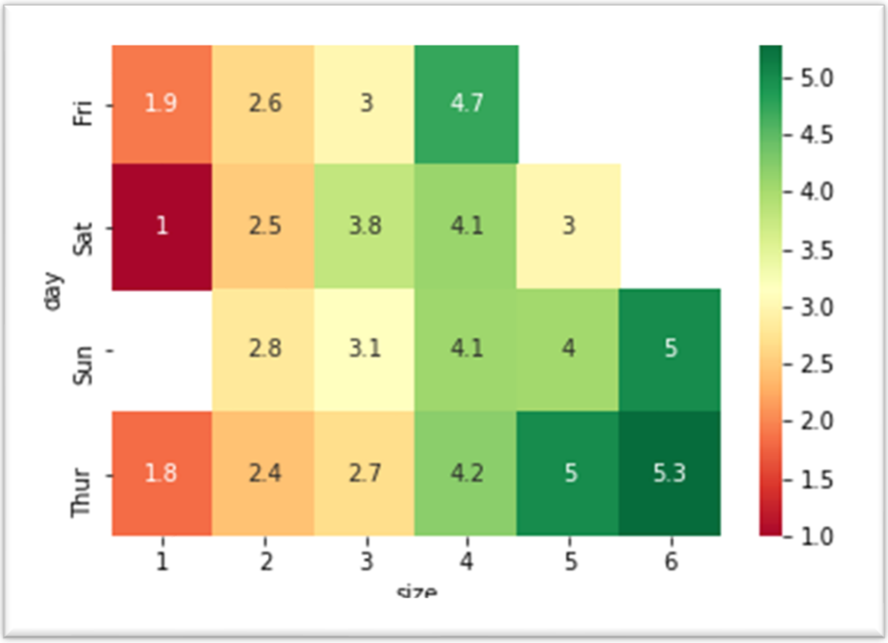


Fig 10: Heatmap of day VS size of customers per table

The above heatmap shows a clearer picture of how many customers visit together for each table on which day of the week. Highest customer size has been noticed in a table on Thursday. Overall, Saturdays and Sundays have higher number of customers visiting the restaurants.

```
#Does the average tip differ by size (number of people at the table)?
tips.groupby(['size'])['tip'].mean().plot.bar(color="orange")
```

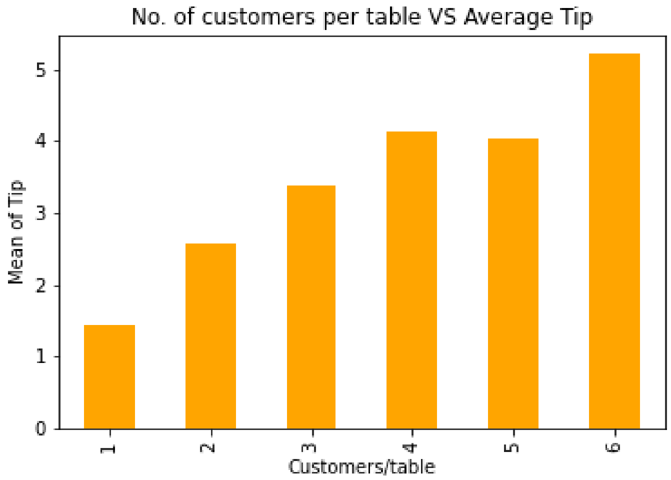


Fig 11: Variation of tip with the size of customers per table

Average amount of tip definitely varies with the number of customers on each table. As per our data, we see that the maximum amount of tip is received when there are most people in a table, that is, 6 persons per table; whereas, when there is only 1 person in a table, amount of tip decreases drastically.

```
#Does the tip amount depend on whether a person is smoker or non smoker
tips.groupby(['smoker'])['tip'].count().plot.bar(color="darkgreen")
tips.groupby(['smoker','sex'])['tip'].mean().plot.bar(color="skyblue")
```

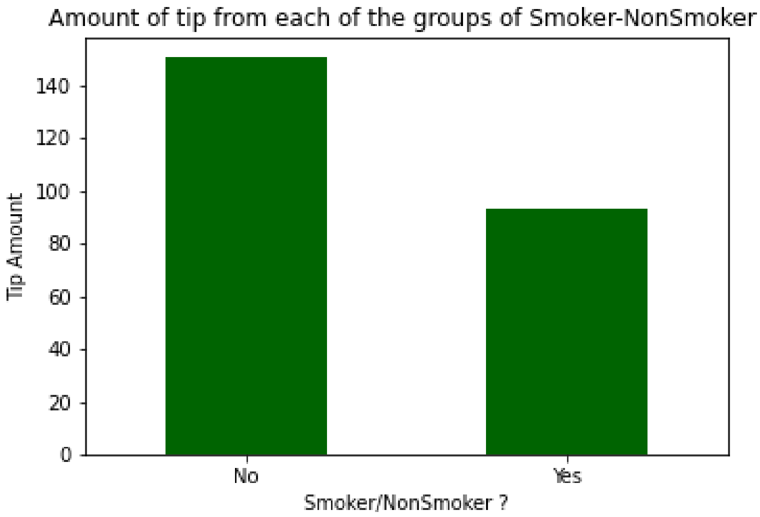


Fig 12: Variation of tip with customer being smoker/non smoker

The above bar chart shows that the non smokers pay more tip as compared to the smokers, irrespective of which gender they belong to.

```
#pivot table of above figure
tips.groupby(['smoker','sex'])['tip'].mean().unstack()
```

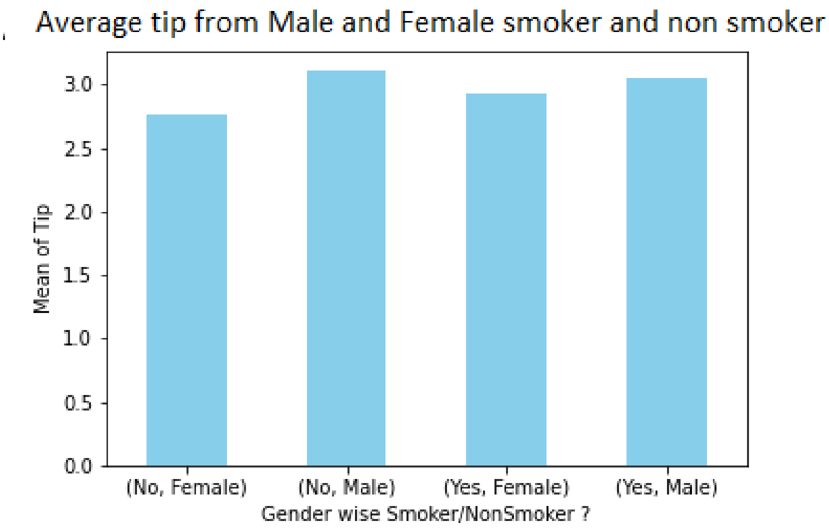


Fig 13: Variation of tip depending upon male and female smoker/non smoker

Here, we observe that the lowest tip is given by female non smokers, followed by female smokers. The second highest tip amount is received from the male smokers while, the male customers who are non smokers pay the highest amount of tip.

4. Now, we applied the boxplot technique to detect if there was any anomaly in the concerned dataset.

```
#Outlier detection
figg = tips.boxplot()
tips.tip.mean()
tips.tip.median()
```

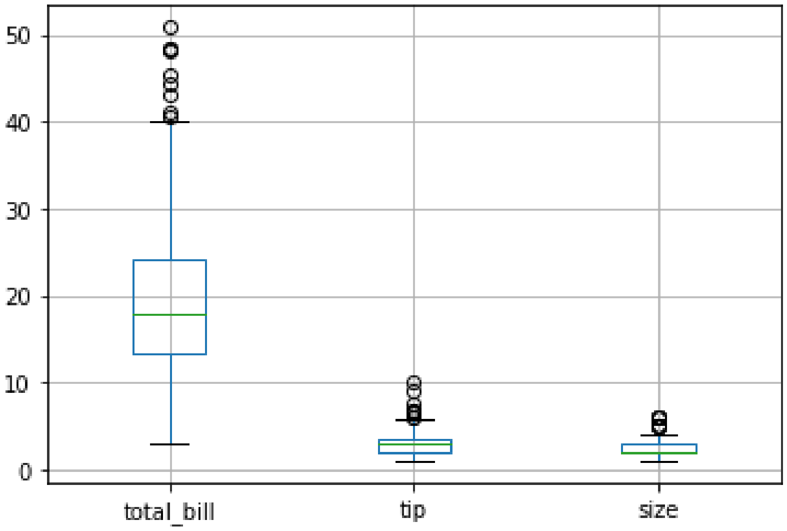
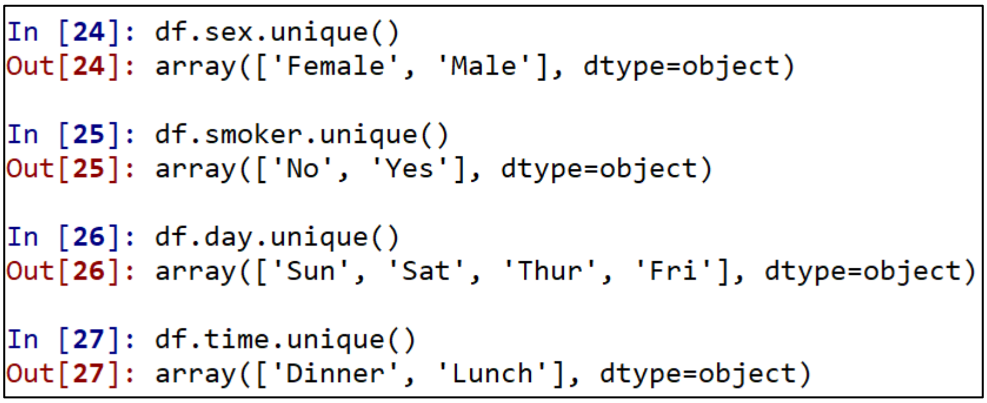


Fig 14: Outlier detection using boxplot

The above figure shows a boxplot for all the variables under consideration. This can identify any outlier or discrepancy in the dataset. Each one of the box in the above figure represent each attribute of the dataset and the bubbles beyond their interquartile range shows outliers. Some of the data here are outliers and should be avoided for further data analysis.

5. Now, since this is a mixed dataset with numeric as well as categorical data, hence, we applied some recoding techniques and converted the categorical data into numeric data:

```
#Calling dataset
df = tips

df.sex.unique()
df.smoker.unique()
df.day.unique()
df.time.unique()
```



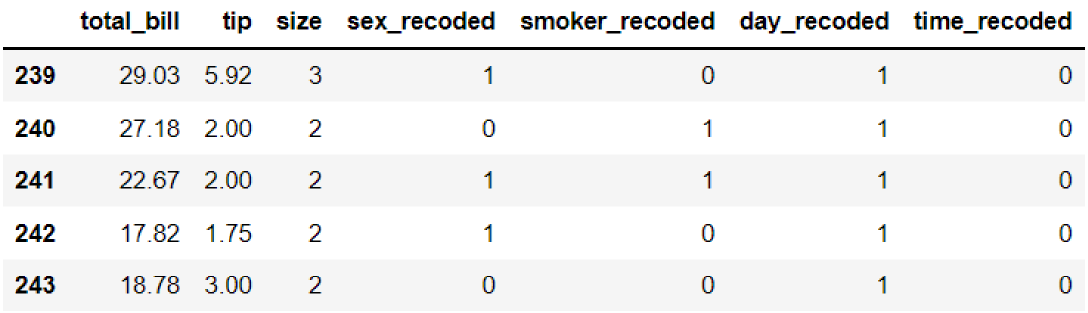
```
In [24]: df.sex.unique()
Out[24]: array(['Female', 'Male'], dtype=object)

In [25]: df.smoker.unique()
Out[25]: array(['No', 'Yes'], dtype=object)

In [26]: df.day.unique()
Out[26]: array(['Sun', 'Sat', 'Thur', 'Fri'], dtype=object)

In [27]: df.time.unique()
Out[27]: array(['Dinner', 'Lunch'], dtype=object)
```

In the above output, we have found out what are all the categories or classes of data under each categorical column of data. The most notable data here is for ‘day’ as all the days are not present in the dataset. Now, we proceed with the recoding part.
Here three types of recoding methods have been applied:

1. **One-hot encoding:** Here we get the dummy values for a particular column and remove any of the column as the values are divided into two columns with same complementary values.
2. **Mapping:** Here, a column is taken and its categories are stored as ‘keys’ of a dictionary and the recoded data are stored as ‘values’. So, a python dictionary is created with key-value pairs.
3. **Label-encoding:** Using the function `LabelEncoder(),`the entire data of a column is transformed into 0,1,2 etc. arranged alphabetically.

```
##one hot encoding
sex_recode = pd.get_dummies(tips['sex'], drop_first=True)
smoker_recode = pd.get_dummies(tips['smoker'], drop_first=True)

##mapping
recode = tips.day.map({'Sun':0, 'Thur':1, 'Fri':2, 'Sat':1})
recode = pd.DataFrame(recode).rename(columns = {'day':'day_recoded'})

##label encoding : follows alphabetical ordering
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
recode['time_recoded'] = le.fit_transform(tips['time'])
#time_recode_series = pd.Series(recode['time_recode'])
#display(time_recode_series)

##concat above 4 obj data types into a single df
new_tips = pd.concat([tips,sex_recode,smoker_recode,recode], axis=1)
new_tips.head()
```

```
##renaming new columns and dropping unwanted columns
new_tips = new_tips.rename(columns = {'Male':'sex_recoded',
'Yes':'smoker_recoded', 0:'time_recoded'})
new_tips.drop(['sex','smoker','time','day'],axis=1,inplace=True)

##viewing the new recoded table
new_tips.tail()
```



	total_bill	tip	size	sex_recoded	smoker_recoded	day_recoded	time_recoded
239	29.03	5.92	3	1	0	1	0
240	27.18	2.00	2	0	1	1	0
241	22.67	2.00	2	1	1	1	0
242	17.82	1.75	2	1	0	1	0
243	18.78	3.00	2	0	0	1	0

Fig 15: Tail data of the newly recoded dataset

The above table shows the last 5 data rows of the dataset after being recoded. After we apply recoding (encoding) techniques, we concat them into a new dataset called “**new_tips**”.

6. Now, we tried performing correlation amongst the entire dataset to find out if all the factors can be considered for further statistical analysis:

```
corr = new_tips.corr(method='spearman')
sns.heatmap(corr, vmin=-1, vmax=1, center=0, linewidths=.5)
```

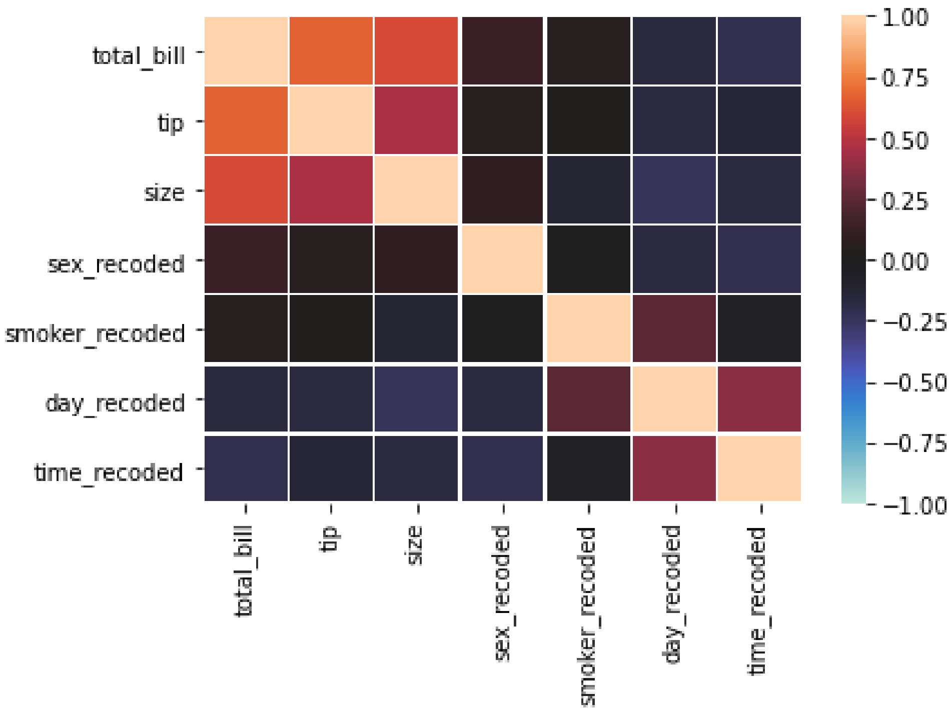


Fig 16: Correlation heatmap amongst all attributes of the tips dataset

We know that the coefficient of correlation must be very near to either +1 or -1 in order to find dependency or relation among variables. Here, we are mainly concerned with the tips variable. From the above figure, we can understand that expect for the three variables total_bill, tip and size; there exists almost no correlation of other variables with the tips. Hence, we drop the idea of taking the other variables and only proceed with the the three variables for our further data analysis, namely **total_bill, tip** and **size**.

7. Then, we tried mapping the entire numerical data through linecharts and subplots.

```
#Plotting all data into subplots
tips1=tips.loc[:,["total_bill","tip","size"]]
tips1.plot()
plt.show()

tips1.plot(subplots = True)
plt.show()
```

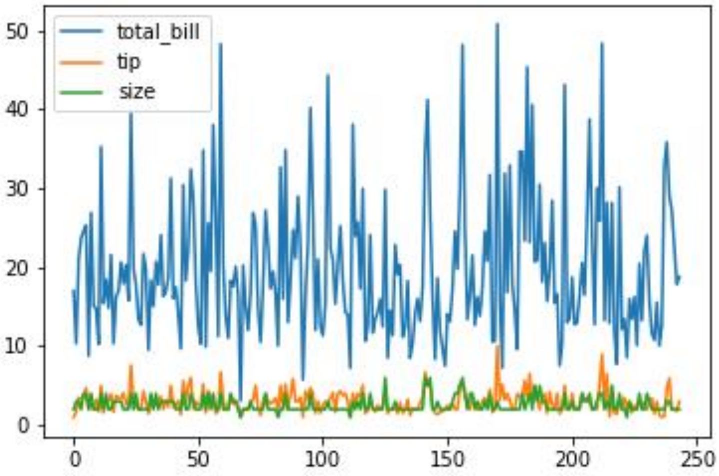


Fig 17: Lineplots with the data of all three attributes total_bill, tip and size

The above figure displays a comparative view of the entire dataset. Even without individually knowing the actual data from the dataset, we can understand the value of all three attributes at a given value in the scale.

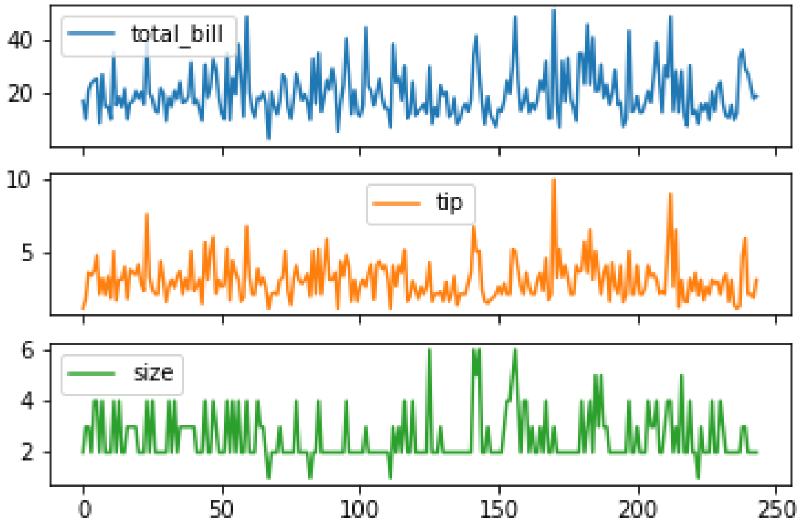


Fig 18: Subplots with the data of all three attributes total_bill, tip and size

This figures is a better version of fig 2. When the values of a certain attribute are of smaller range while another attribute have bigger values, plotting them altogether can make it difficult for us to read the values. Hence subplot has been introduced which make the job much easier and accurate in terms of data representation.

8. Next, we find the correlation among the variables taken into consideration and generate correlation table and heatmap for this.

```
#Correlation
tips.corr()

#correlation map
f,ax = plt.subplots(figsize=(5,5))
sns.heatmap(tips.corr(), annot=True, linewidths=.5, cmap="RdYlGn",
fmt='.1f',ax=ax)
plt.show()
```

Table 2 : Correlation Analysis

	total_bill	tip	size
total_bill	1.000000	**0.675734**	**0.598315**
tip	0.675734	1.000000	**0.489299**
size	0.598315	0.489299	1.000000

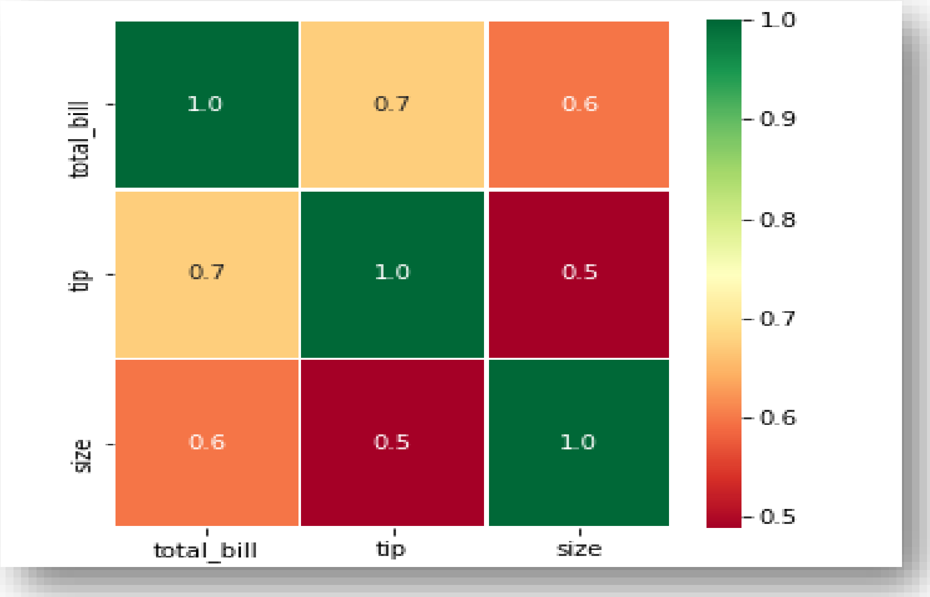


Fig 19: Correlation Heatmap

The above table and figure both demonstrate the correlation amongst the attributes total_bill, tip and size of customers. Correlation coefficient value ideally ranges between +1 to -1 where +1 implies very **strong and perfectly positive correlation**, while -1 implies very **strong and perfectly negative correlation**. 0 signifies that there is no correlation between two variables.

From the projected value, we can see that there is ***positive and moderate correlation*** of size of customers and total_bill with ‘tip’. Thus we proceed with further analysis.

9. Finally, with the help of scatterplot, jointplot, pairplot from the seaborn package, we demonstrate the relationship between the variables.

- **<u>Tip VS Total_Bill</u>**

```
#Tip vs Total_Bill scatterplot
sns.scatterplot(x='total_bill', y='tip', data = tips)
```

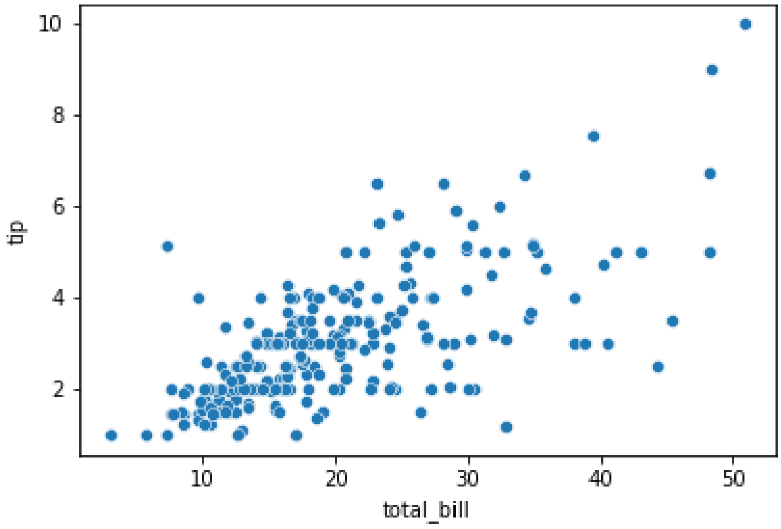


Fig 20: total_bill VS Tip scatter-plot

The above scatterplot shows that there is a **moderate positive correlation** of the amount of tip received with the total_bill of the restaurant. When the total_bill amount increases, if we consider for a particular food item, its price remains constant at all servings, but even then if the total_bill increases, we can conclude that the increase in the tip is contributing to the increase in total_bill and vice-versa.
However, studying the pattern of the scatterplot, we see that initially when the total_bill increases, the tip amount also increases but after some point of time, even if bill increase, no change in tip amount is observed.

- **Tip VS Size**

```
#Tip vs Size scatterplot
sns.scatterplot(x='size', y='tip', data = tips)
```

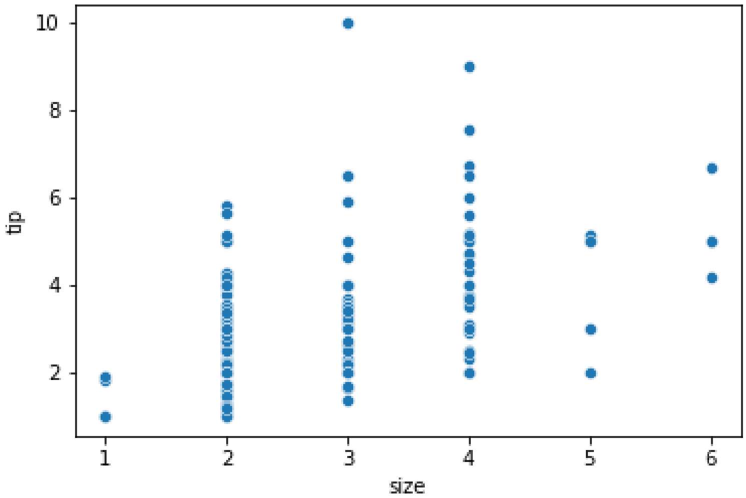


Fig 21: size VS Tip scatter-plot

The above scatterplot shows that among the size of customers per table and the amount of tips received, there is a almost no correlation or at least it seems to be weaker than the other correlation we already observed. The amount of tip received by the restaurant does not really depend on the number of customers sitting at a table. Only 2 customers can also tip an amount upto 6 while 6 customers in the table also do not go beyond 6. However, 4 customers together can tip an amount as high as 8. Thus, the dependency is not thorough and constant for all values.

- **<u>Distribution plot of tips data</u>**

```
#data distribution plot
import warnings
warnings.filterwarnings("ignore")
sns.distplot(tips.tip)
```

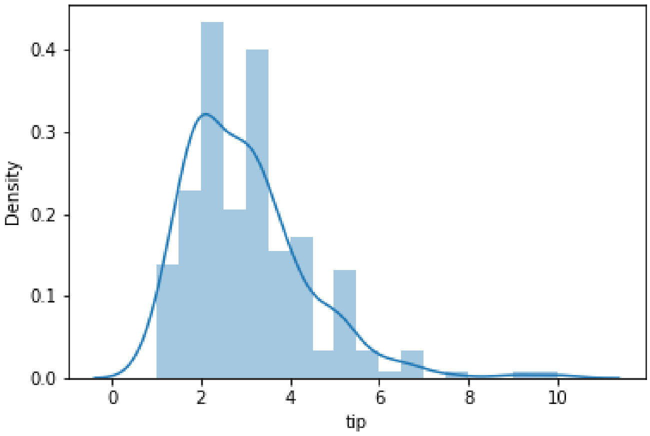


Fig 22: Data Distribution of tip

From the above distplot which is basically a frequency polygon, we observe that the tip data is **normally distributed** with a bit of skewness at the left due to presence of much higher values of tip for 2 and 4 customers per table.

- **<u>Jointplot of total_bill VS tips</u>**

```
sns.jointplot(data=tips, x='total_bill', y='tip', kind="reg")
```

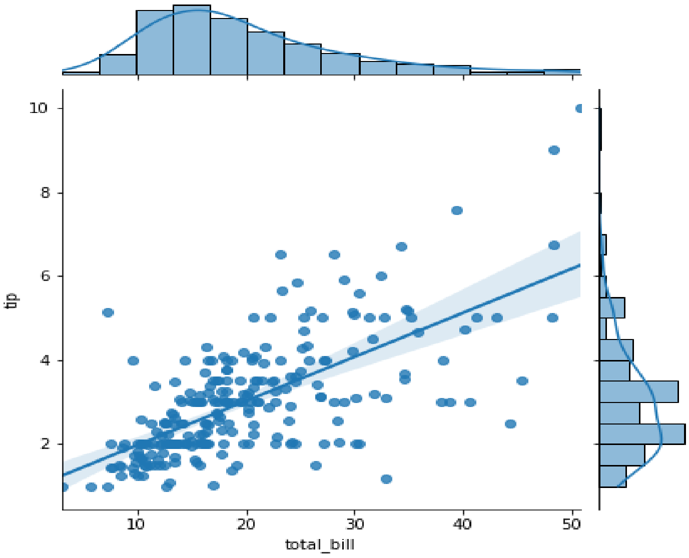


Fig 23: Jointplot of tip VS total_bill

The jointplot of tip with total_bill again demonstrates a positive correlation among the two with the distribution taken into consideration. Both the attributes are normally distributed, with the data slightly skewed in some cases.

- **Regression scatter-plots**

```
#regression-scatterplots
sns.pairplot(data=tips, kind="reg", size=2)
```

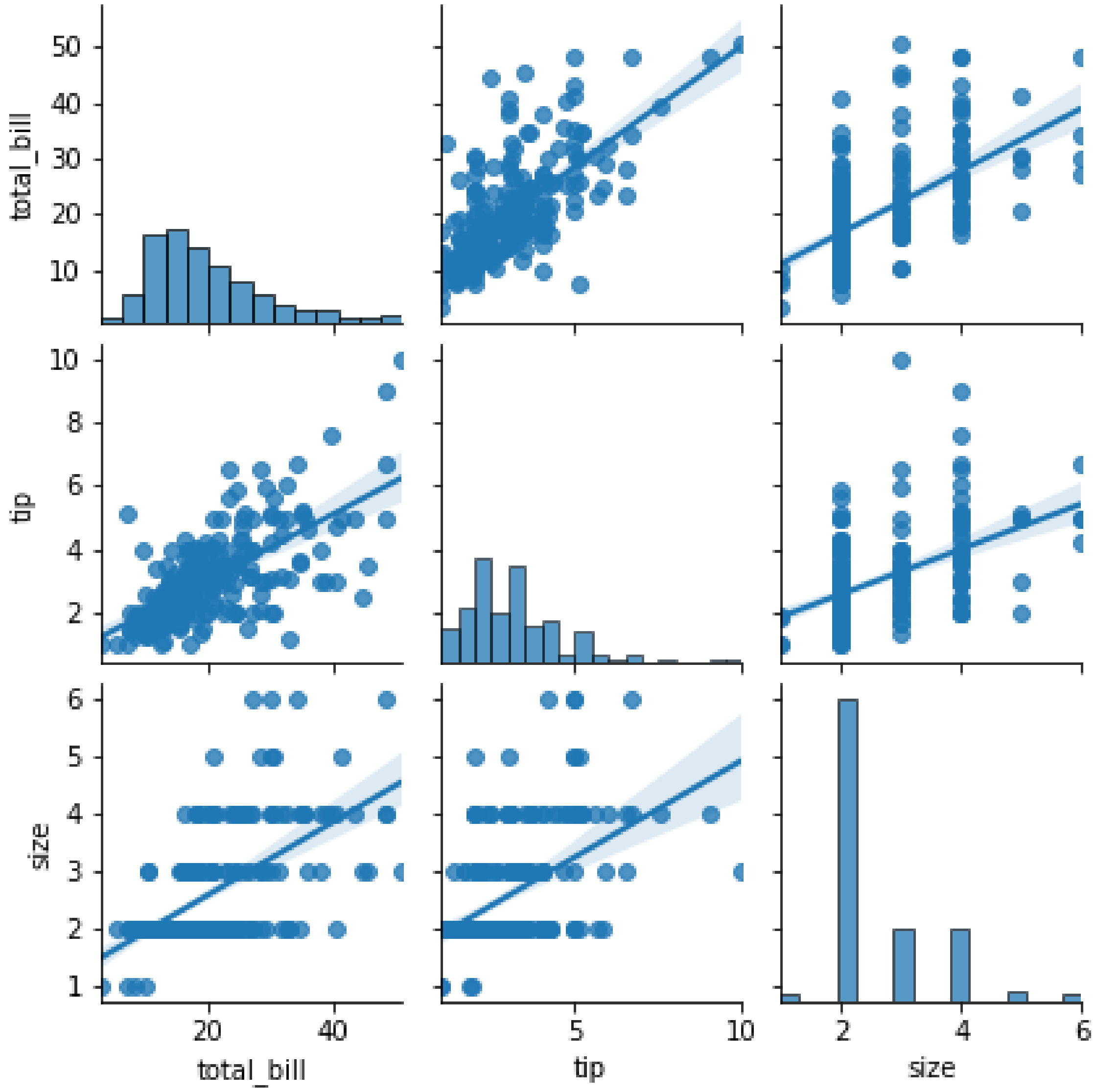


Fig 24: Pairplot of total_bill, tip and size

The above figure plots pairwise relationships in a dataset. By default, this function will create a grid of Axes such that each numeric variable in the **Tips** dataset will be shared across the y-axes across a single row and the x-axes across a single column. The diagonal plots are treated differently: a univariate distribution plot is drawn to show the marginal distribution of the data in each column.

The simplest invocation uses **scatterplot()** for each pairing of the variables and **histplot()** for the marginal plots along the diagonal.
From the above plots, we can identify a strong positive correlation and can thus conclude that the amount of tip and total_bill are interdependent.

```
sns.pairplot(data=tips, kind="reg", size=1.8, hue="smoker")
```

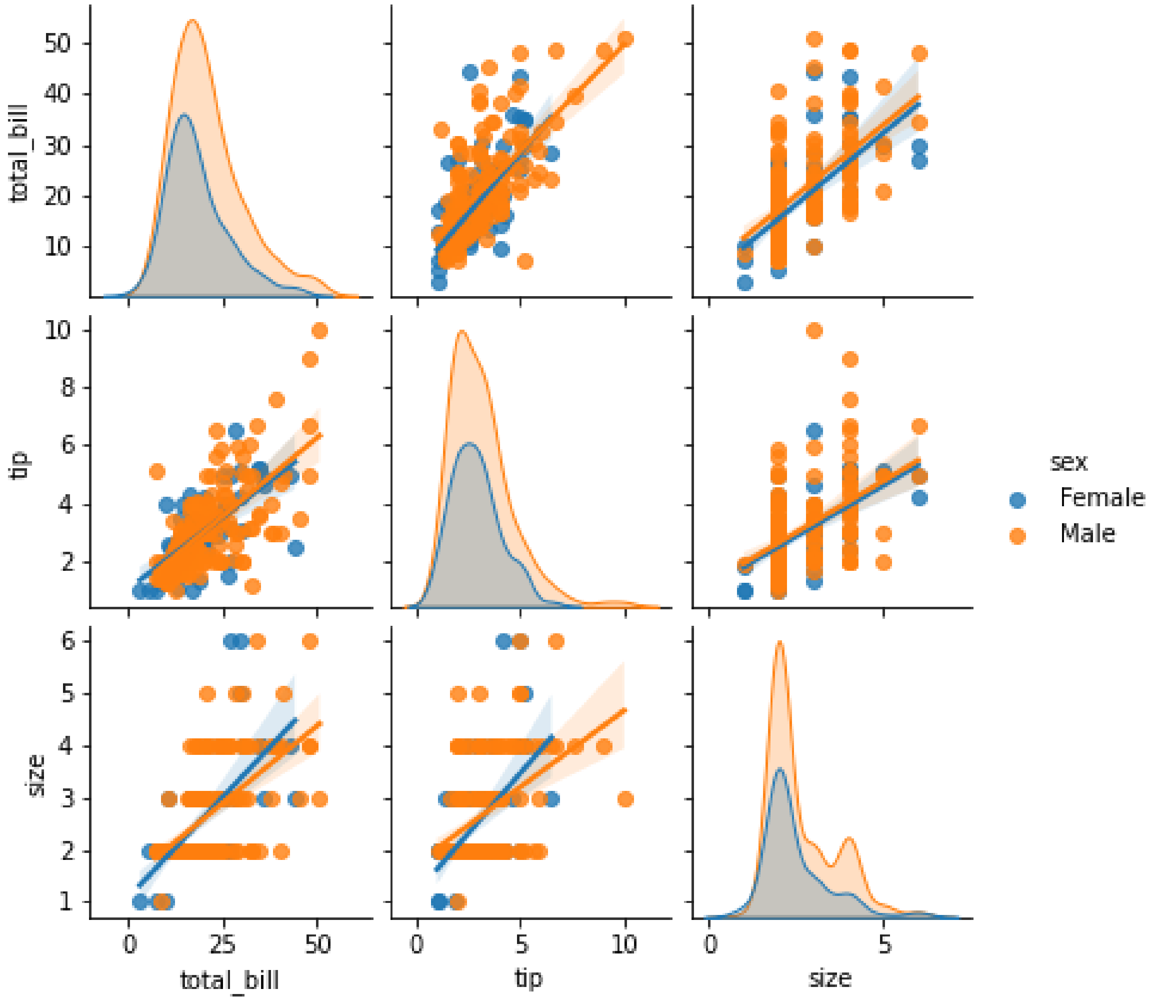


Fig 25: Pairplot of total_bill, tip and size based on gender (sex)

Similar to the previous plot, this figure also depicts a pairplot, but the diagonals are replaced with bell-shaped curves. Assigning a hue variable adds a semantic mapping and changes the default marginal plot to a layered kernel density estimate (KDE). This plot helps us to study the variation or the relation among the attributes total_bill, tip and size and shows which of the group under gender (sex) influences these relations the most. Clearly, we can see that the male customers here contribute more to all the underlying relationships.

```
sns.pairplot(data=tips, kind="reg", size=2, hue="sex")
```

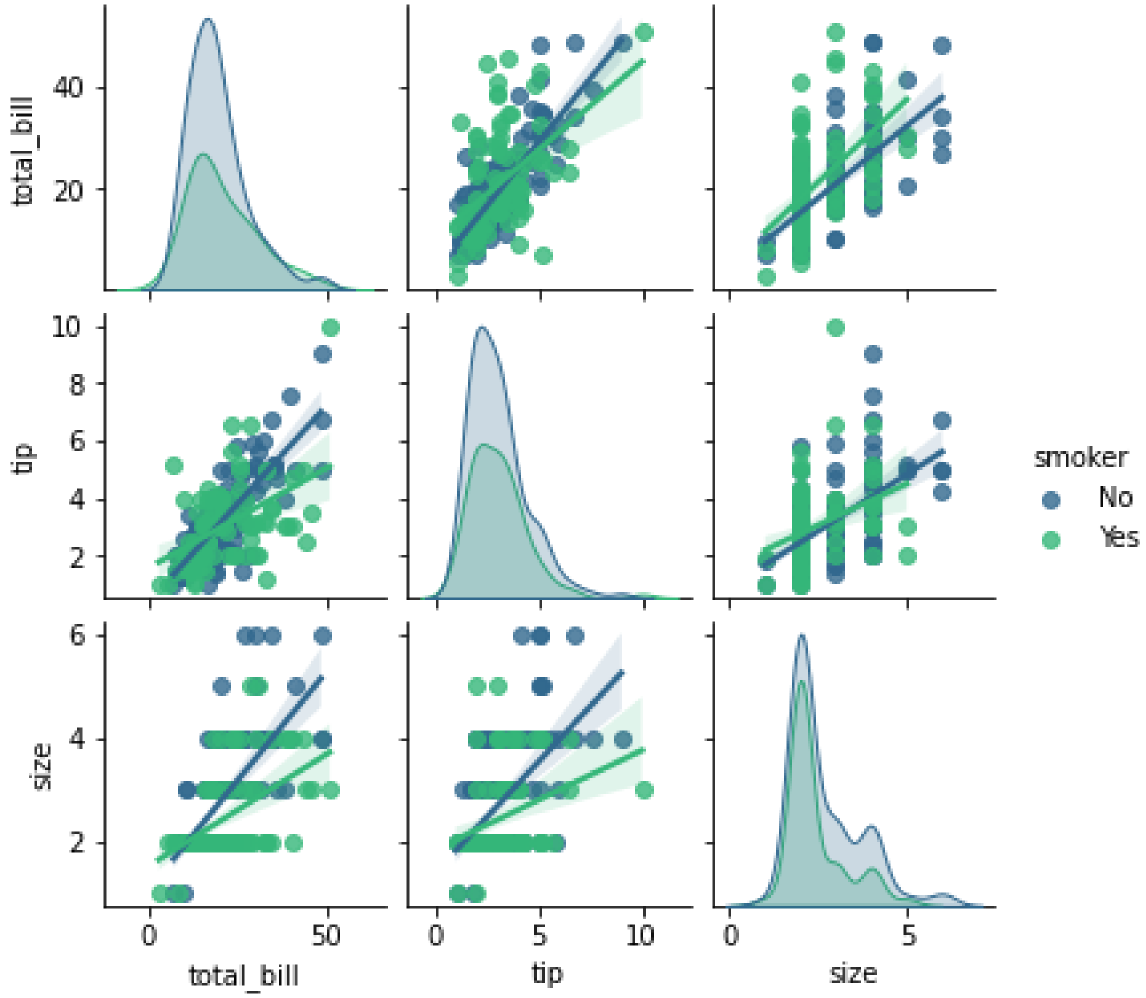


Fig 26: Pairplot of total_bill, tip and size based on smoker/non smoker

Similarly, the above pairplot shows the relations among total_bill, tip and size categorizing the customers as smokers and non-smokers. Since the number of smokers and non-smokers are almost equal, we see almost similar contribution of both smoker and non-smokers in the above relationships.

5.3 Model development and Evaluation

After finding out the correlation among all the variables and plotting them in various scatter plots, we have observed that the total_bill seems to have influenced the tip amount the most ; given that the size of the customers per table is not very much correlated with the tip amount. Also the other attributes are considered as uncorrelated with respect to tip amount. Hence, we propose the 'Simple Linear Regression' model in this scenario. Here, the variable to be computed or predicted, y = 'tip' and the variable on which it depends is x = 'total_bill'.

Thus, it should follow the equation :- $y = b_0 + b_1x$

In order to fit the data into linear regression model, we perform the following: -

1. Importing necessary model packages

```
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

2. Storing the independent and dependent variables into x and y

```
X = np.array(df1['total_bill']).reshape((-1,1))
Y = np.array(df1['tip'])
```

3. Splitting the data and training with 0.2 part of the data

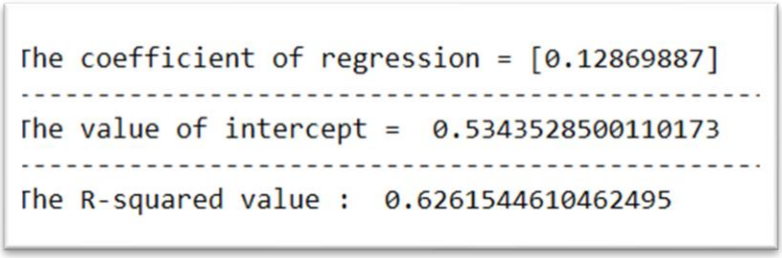
```
X_train,X_test, Y_train, Y_test =
train_test_split(X,Y,test_size=0.2,random_state=4)
```

4. Fitting the trained data into the model

```
##Fitting Linear Regression to the dataset
lin_reg = LinearRegression()
lin_reg.fit(X,Y)
```

5. Finding the regression coefficient, intercept and the R2 value

```
print("The coefficient of regression =",lin_reg.coef_)
print("The value of intercept = ",lin_reg.intercept_)
print("The R-squared value : ",lin_reg.score(X,Y))
```



```
The coefficient of regression = [0.12869887]
---------------------------------------------
The value of intercept =  0.5343528500110173
---------------------------------------------
The R-squared value :  0.6261544610462495
```

The above data can be interpreted as follows: -

If the linear regression equation is given by : $y = b_0 + b_1x$

then, coefficient of regression = coefficient of independent variable x = **b_1 =0.128**

and, value of intercept = **b_0 = 0.534**

Hence, the equation turns out to be: $y = 0.534 + 0.128x$

The value of $R^2 = 0.626$ signifies that the model is able to explain 62.6% of the variability of the variable y with respect to x. A score closer to 1 and above 0.5 is considered moderately good fit. Hence, the fitting the model is good.

6. Visualizing the model fitting through scatter plot

```
##Visualizing the Linear Regression results

plt.scatter(X,Y,color='red')
plt.plot(X, lin_reg.predict(X), color='blue')

plt.title('Linear Regression')

plt.xlabel('total_bill')
plt.ylabel('tip')

plt.show()
```

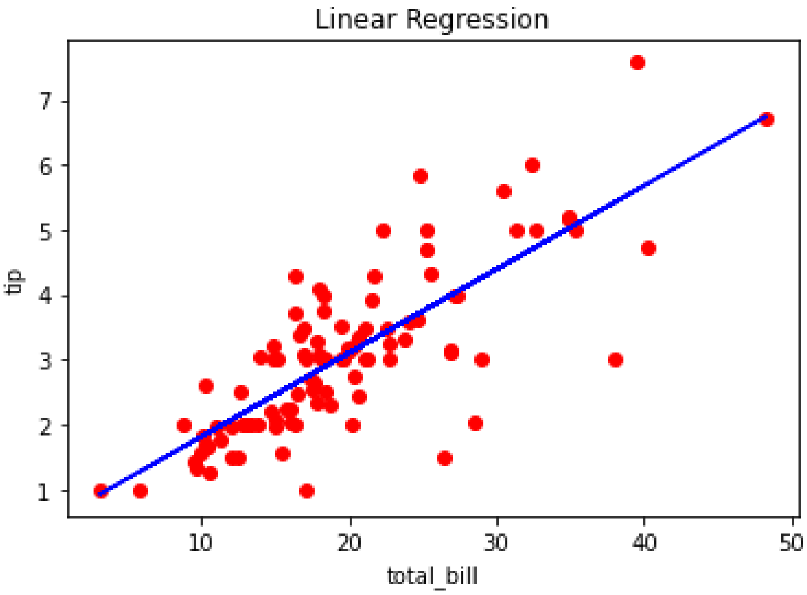
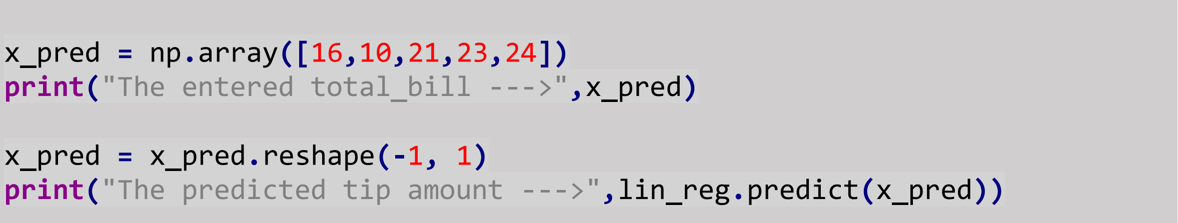


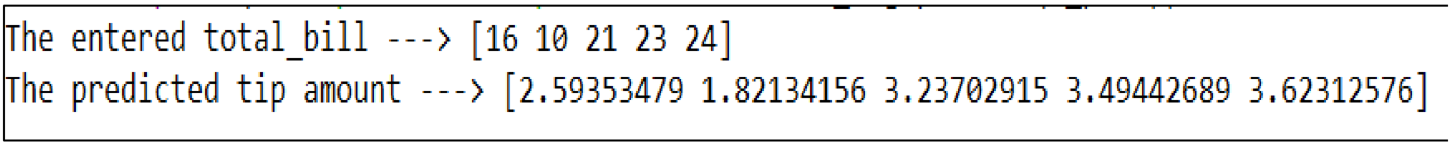
Fig 27: Linear Regression Model fitting scatter-plot

The above plot shows that the data fits into the linear regression model as we can see a **straight line** whose slope has been computed to be **0.128**. A positive slope implies that as x increase, y also increases.

7. Prediction of future values

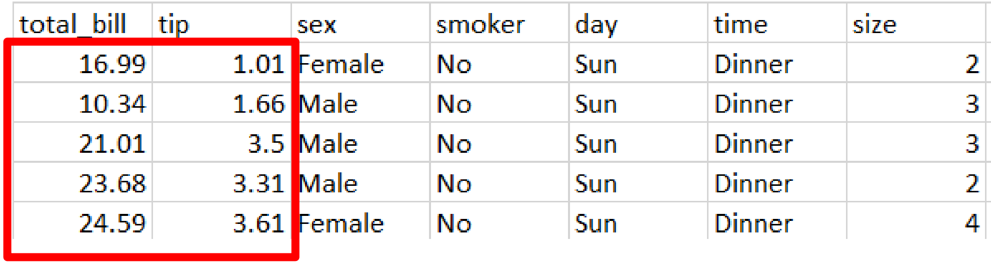


```
x_pred = np.array([16,10,21,23,24])
print("The entered total_bill --->",x_pred)

x_pred = x_pred.reshape(-1, 1)
print("The predicted tip amount --->",lin_reg.predict(x_pred))
```



```
The entered total_bill ---> [16 10 21 23 24]
The predicted tip amount ---> [2.59353479 1.82134156 3.23702915 3.49442689 3.62312576]
```

If we compare the above predicted values with the actual values present in the dataset, we can see that almost all the values were predicted nearly correct as can be tallied with following:



total_bill	tip	sex	smoker	day	time	size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.5	Male	No	Sun	Dinner	3
23.68	3.31	Male	No	Sun	Dinner	2
24.59	3.61	Female	No	Sun	Dinner	4

CHAPTER 6

FINDINGS & DISCUSSIONS

The key findings from the study are as follows:

- There was no missing value in the dataset, however, there were a few outliers detected.
- The total dataset size was 244.
- The maximum tip given was 10 while minimum was 1.
- 64.3% customers were male while the rest 35.7% were females.
- Male customers gave more tip than female customers.
- More tip was received during dinner time than lunch time.
- Considering all days of a week, most of the tips was received on Saturdays, followed by Sundays and the lowest tip was received on Fridays.
- Number of customers per table varied with the day of week; however, most no. of customers per table was on Thursday and lowest no. of customers was on Saturday.
- Amount of tip was highest when customers per table was 6 and lowest when there was only 1 customer in a table.
- Non-smoker customers gave more tips than smokers.
- Highest tip was given by non-smoker male customer while the lowest was given by non-smoker females, even lower than smokers in general.
- When acted upon together, sex, smoking criteria, day of week and meal timing had no effect on the amount of tip and there exists slight multi-collinearity amongst them.
- Total_bill was highly correlated with the amount of tip; size of customers per table was weakly correlated with amount of tip.
- The data distribution of the amount of tip is slightly skewed towards the left.
- The data is well-fitted into the Simple Linear Regression model with a R2 score of 0.626.
- The model is also able to predict almost all data correctly as per its learning.

CHAPTER 7

IMPLEMENTATION OF THE PROJECT

The above project was implemented using the following steps:

- Importing necessary python packages.
- Invoking required dataset ‘tips’.
- Computing the descriptive and demographic data.
- Visualizing all the above data.
- Identifying the statistical relations.
- Visualizing the above relations.
- Proposing suitable model.
- Fitting the data.
- Splitting the data and training the machine using ML algorithm.
- Testing with foreign dependent values and predict the dependent values.
- Computing the R2 score, intercept, coefficient and hence finding the regression equation.

CHAPTER 8

CONCLUSION

Restaurants are a popular source of income and business among hoteliers and similar other businessmen. For them, estimating and foreseeing the income is very much vital as well as keeping track of the income of the employees is also important. These collectively help them to set price for a particular menu. Tipping a waiter is also considered as a source of earning for the waiters.

The main motto of this study was to find out if the factors are all affecting the tipping amount in the restaurant. Also, another important side to this study was to propose a model that would efficiently predict or foresee the amount of tips based on the total_bill of the customers. Although there could have been other more relevant factors like – service quality, behaviour of waiters, quality of food etc. which could have yielded a better result.

Still, our model was able to yield an R2 score of more than 60% which is fine to predict the tipping amount with limited data.

Hence, our null hypothesis, that is, "***The factors considered in the study do not affect and are not correlated with the amount of tips received by the restaurant***" is rejected as we have observed from the study that at least one factor significantly affects the tips and they are strongly correlated to each other. Hence, our Research hypothesis is verified and can be accepted with the implementation of the above discussed model.

CHAPTER 9

REFERENCES

1. https://seaborn.pydata.org/tutorial/color_palettes.html

2. https://link.springer.com/chapter/10.1007%2F978-0-387-71762-3_7

3. https://seaborn.pydata.org/generated/seaborn.pairplot.html

4. https://www.google.com/search?q=scatter+plot+interpretation&rlz=1C1VDKB_enIN960IN960&sxsrf=AOaemvIleOWoPb7yIudYYbjA3LthUSVX_g:1633285232319&tbm=isch&source=iu&ictx=1&fir=cHoFwl3duKmRzM%252Ca7zpy0UXHa7bzM%252C_%253BiRFpFd44LDcTCM%252CSHEkFcCZXf_wAM%252C_%253BCW4vT5304UdbkM%252CMNHWxkGF2Z5_DM%252C_%253BnzeiBnequ-fwfM%252CTKDDKNQ5wpOYIM%252C_%253B8FhX6YwbeVVT_M%252CV4tsFca3JInYoM%252C_%253Bf0bLkSb9ebnqkM%252CgnnsHfk3h0AREM%252C_&vet=1&usg=AI4_-kQ22OZ3psc2PJJddqSJVt86LaNNNg&sa=X&ved=2ahUKEwiG6trQ7a7zAhXkB2MBHdKeDk0Q_h16BAgPEAE#imgrc=f0bLkSb9ebnqkM

5. https://www.youtube.com/watch?v=GdkUbZkF5bo

6. https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

7. https://realpython.com/linear-regression-in-python/#simple-linear-regression

8. https://www.datacamp.com/community/tutorials/python-rename-column?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=332602034358&utm_targetid=dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9300053&gclid=CjwKCAjwtfqKBhBoEiwAZuesiFjhc8yXnyfK0NhMRaUtzt_ku80jgciSqV4XpStVx-f4d3QcM-sz1RoC4DUQAvD_BwE

9. https://www.codegrepper.com/code-examples/python/map+column+names+pandas

10. https://www.machinelearningplus.com/pandas/pandas-drop-column-using-dataframe-drop/

11. https://www.w3schools.com/python/python_ml_multiple_regression.asp

12. https://aegis4048.github.io/mutiple_linear_regression_and_visualization_in_python

13. https://www.kaggle.com/ranjeetjain3/seaborn-tips-dataset

14. https://www.youtube.com/watch?v=slH8Xt-tqhw