K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Batch: B3 **Roll No.: 1811106**

Experiment / assignment / tutorial No. 2

Title : Exploratory data analysis - Data Pre-processing and Data Visualization

Aim: Data Pre-processing – Data integration (correlation analysis) , Data Reduction (PCA algorithm) and Graphic packages of Python to visualize and interpret data

Expected Outcome of Experiment:

CO2: Organize and Prepare the data needed for data mining using pre preprocessing techniques.

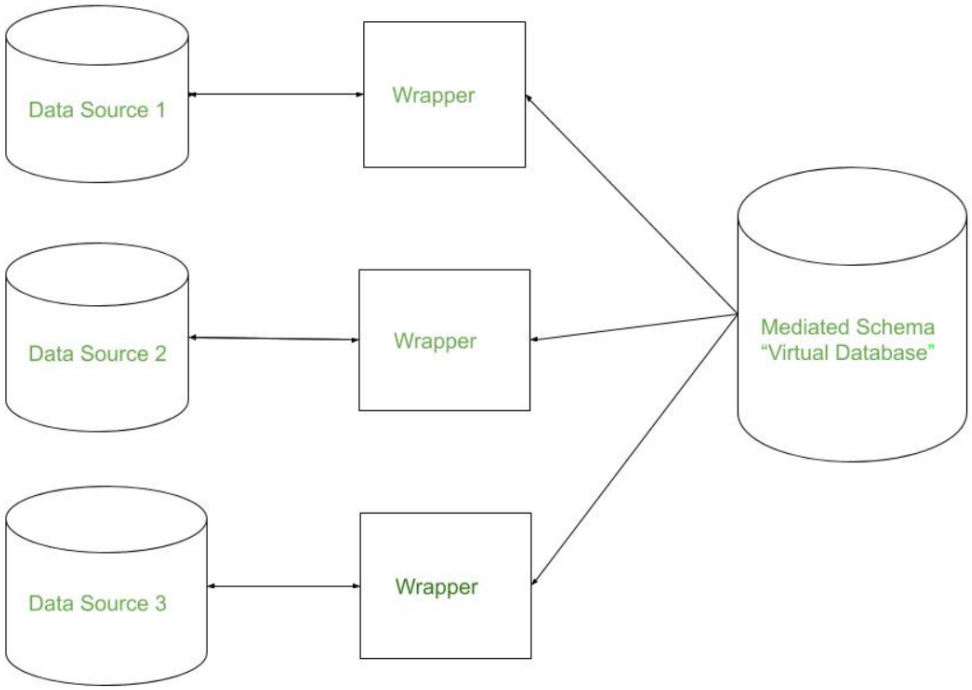
Books/ Journals/ Websites referred:

1. https://www.geeksforgeeks.org/data-integration-in-data-mining/
2. https://www.geeksforgeeks.org/data-reduction-in-data-mining/
3. https://www.jinfonet.com/resources/bi-defined/data-visualization/

Data Integration Tasks

Data Integration is a data preprocessing technique that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data. These sources may include multiple data cubes, databases or flat files.

The data integration approach are formally defined as triple <G, S, M> where,
G stand for the global schema,
S stand for heterogenous source of schema,
M stand for mapping between the queries of source and global schema.



There are mainly 2 major approaches for data integration – one is "tight coupling approach" and another is "loose coupling approach".

Tight Coupling:

Here, a data warehouse is treated as an information retrieval component.
In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation and Loading.

Loose Coupling:

Here, an interface is provided that takes the query from the user, transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.
And the data only remains in the actual source databases.

Issues in Data Integration:

There are no of issues to consider during data integration: Schema Integration, Redundancy, Detection and resolution of data value conflicts. These are explained in brief as following below.

1. Schema Integration: Integrate metadata from different sources. The real world entities from multiple source be matched referred to as the entity identification problem. For example, How can the data analyst and computer be sure that customer id in one data base and customer number in another reference to the same attribute.

2. Redundancy: An attribute may be redundant if it can be derived or obtaining from another attribute or set of attribute. Inconsistencies in attribute can also cause redundancies in the resulting data set. Some redundancies can be detected by correlation analysis.

3. Detection and resolution of data value conflicts: This is the third important issues in data integration. Attribute values from another different sources may differ for the same real world entity. An attribute in one system may be recorded at a lower level abstraction then the "same" attribute in another.

Data Reduction algorithms

1. Data Cube Aggregation:
This technique is used to aggregate data in a simpler form. For example, imagine that information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months. They involve you in the annual sales, rather than the quarterly average, So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

2. Dimension reduction:
Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates outdated or redundant features.

- Step-wise Forward Selection –
  The selection begins with an empty set of attributes later on we decide best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics.

- Step-wise Backward Selection –
  This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set.
  Suppose there are the following attributes in the data set in which few attributes are redundant.

3. Data Compression:

The data compression technique reduces the size of the files using different encoding mechanisms (Huffman Encoding & run-length Encoding). We can divide it into two types based on their compression techniques.

- Lossless Compression –
  Encoding techniques (Run Length Encoding) allows a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.

- Lossy Compression –
  Methods such as Discrete Wavelet transform technique, PCA (principal component analysis) are examples of this compression. For e.g., JPEG image format is a lossy compression, but we can find the meaning equivalent to the original the image. In lossy-data compression, the decompressed data may differ to the original data but are useful enough to retrieve information from them.

4. Numerosity Reduction:
In this reduction technique the actual data is replaced with mathematical models or smaller representation of the data instead of actual data, it is important to only store the model parameter. Or non-parametric method such as clustering, histogram, sampling. For More Information on Numerosity Reduction Visit the link below:

5. Discretization & Concept Hierarchy Operation:
Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes by labels of small intervals. This means that mining results are shown in a concise, and easily understandable way.

- Top-down discretization –
  If you first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat of this method up to the end, then the process is known as top-down discretization also known as splitting.

- Bottom-up discretization –

If you first consider all the constant values as split-points, some are discarded through a combination of the neighbourhood values in the interval, that process is called bottom-up discretization.

Concept Hierarchies: It reduces the data size by collecting and then replacing the low-level concepts (such as 43 for age) to high-level concepts (categorical variables such as middle age or Senior).

For numeric data following techniques can be followed:

- Binning –
  Binning is the process of changing numerical variables into categorical counterparts. The number of categorical counterparts depends on the number of bins specified by the user.
- Histogram analysis –
  Like the process of binning, the histogram is used to partition the value for the attribute X, into disjoint ranges called brackets. There are several partitioning rules:
- Equal Frequency partitioning: Partitioning the values based on their number of occurrences in the data set.
- Equal Width Partioning: Partioning the values in a fixed gap based on the number of bins i.e. a set of values ranging from 0-20.
- Clustering: Grouping the similar data together.

Data Visualization objective

Data visualization is the creation and study of the visual representation of data. Data visualization is sometimes referred to as visual communication or descriptive statistics, and includes the techniques to present data in a visual way so as to clearly communicate information and stimulate viewer engagement and attention.

Analyzing and reasoning about data through visualizations makes complex data more accessible, understandable and usable. For many people, seeing analytical results presented visually makes interpretation easier and makes it easier to see patterns trends and correlations. Some of the main objectives of data visualization are:

- to explore sources
- to tell stories
- to predict sales volumes

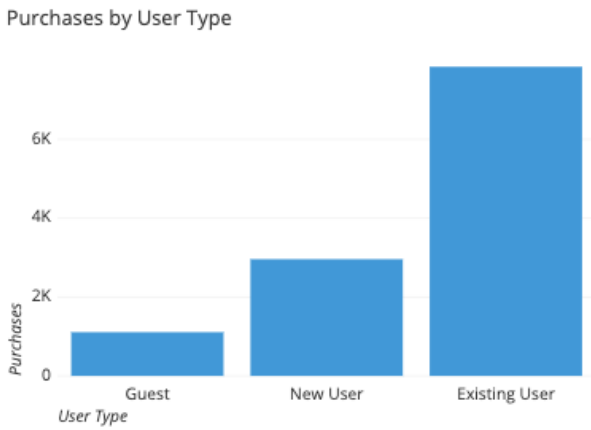
- to identify areas that need attention or improvement
- to understand what factors influence customers' behavior
- to know which products to place where
- to discover how to increase revenues or reduce expenses
- spreadsheets are hard to visualize
- patterns and trends can be spotted quickly and easily
- saves time and energy

Nowadays, advanced interactive data visualizations allow the user to not only put vast amounts of data into a pictorial or graphical format, but using computers and mobile devices allows users to drill down into charts and graphs for more details — changing what data is seen and how it is processed.

Overall, good data visualization should simplify data. It should make analytical tasks, such as making comparisons or understanding causality, providing insights into a data set, exposing and recognizing patterns or relationships easier and more effective. The best uses of data visualization will have a balance between form and function. This meaning the visual representation of data should not only be visually attractive, but also informative. Inversely, visual objects contained in graphics do not need to be boring or overly simple to convey useful information.

Basic plots and their purpose

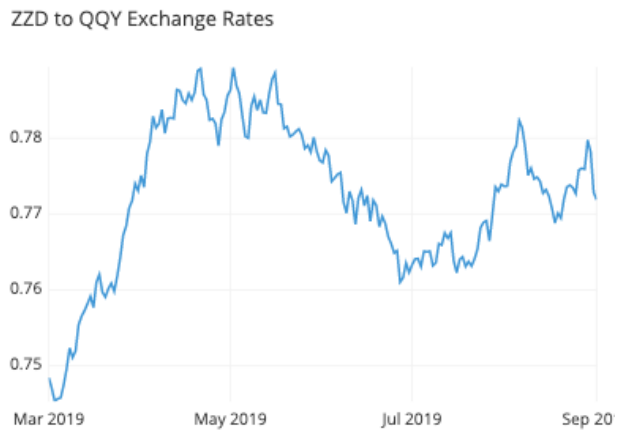
Bar chart:



In a bar chart, values are indicated by the length of bars, each of which corresponds with a measured group. Bar charts can be oriented vertically or horizontally; vertical

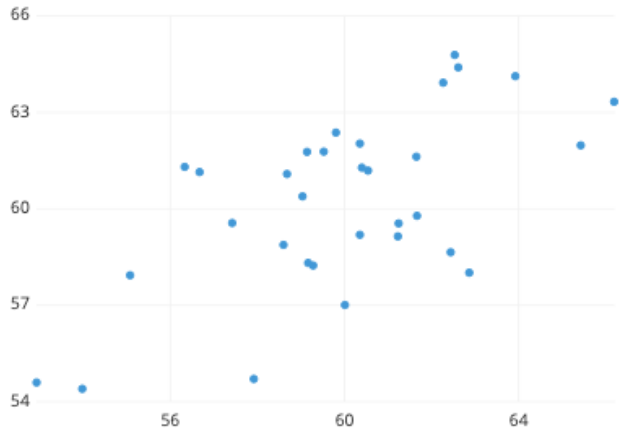
bar charts are sometimes called column charts. Horizontal bar charts are a good option when you have a lot of bars to plot, or the labels on them require additional space to be legible.

Line chart:



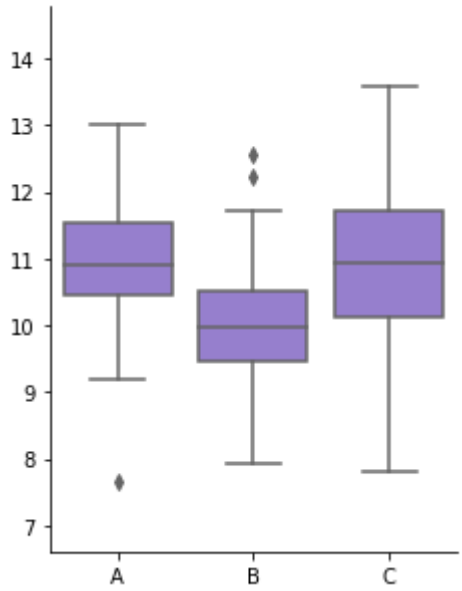
Line charts show changes in value across continuous measurements, such as those made over time. Movement of the line up or down helps bring out positive and negative changes, respectively. It can also expose overall trends, to help the reader make predictions or projections for future outcomes. Multiple line charts can also give rise to other related charts like the sparkline or ridgeline plot.

Scatter plot:



A scatter plot displays values on two numeric variables using points positioned on two axes: one for each variable. Scatter plots are a versatile demonstration of the relationship between the plotted variables—whether that correlation is strong or weak, positive or negative, linear or non-linear. Scatter plots are also great for identifying outlier points and possible gaps in the data.

Box plot:



A box plot uses boxes and whiskers to summarize the distribution of values within measured groups. The positions of the box and whisker ends show the regions where the majority of the data lies. We most commonly see box plots when we have multiple groups to compare to one another; other charts with more detail are preferred when we have only one group to plot.

Implementation (with Python code and snap shot of output with interpretation of graph)

Data set used:

Title: seaborn_tips_dataset
Source: https://www.kaggle.com/ranjeetjain3/seaborn-tips-dataset?select=tips.csv
Number of instances: 100
Number of attributes: 7

Attribute information:
Sr. No: Serial number
TotalBill: Cost of the meal, including tax, in US dollars
Tips: Tip (gratuity) in US dollars
Smoker: Is the customer a smoker or non-smoker?
Day: Day of the week visited
Time: Time of the meal
Size: Size of the group visited

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
C:\Users\ACER\Anaconda3\lib\site-packages\statsmodels\tools\_testing.py:19: FutureWarning: pandas.util.testing is deprecated. U
se the functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm
```

```
data = pd.read_csv('Tips.csv')
```

```
data
```

	SINO	TotalBill	Tips	Smoker	Day	Time	Size
0	1	16.99	1.01	No	Sun	Dinner	2.0
1	2	10.34	1.66	No	Sun	Dinner	3.0
2	3	21.01	3.50	No	Sun	Dinner	3.0
3	4	23.68	3.31	No	Sun	Dinner	2.0
4	5	24.59	3.61	No	Sun	Dinner	4.0
...	...	...	...	...	...	...	...
95	96	40.17	4.73	Yes	Fri	Dinner	4.0
96	97	27.28	4.00	Yes	Fri	Dinner	2.0
97	98	12.03	1.50	Yes	Fri	Dinner	2.0
98	99	21.01	3.00	Yes	Fri	Dinner	2.0
99	100	25.10	NaN	No	NaN	Lunch	2.0

```
data.isnull().sum()
```

```
SINO         0
TotalBill    2
Tips         6
Smoker       7
Day          7
Time         0
Size         1
dtype: int64
```

```
data = data.dropna()
```

```
data.columns
```

```
Index(['SINO', 'TotalBill', 'Tips', 'Smoker', 'Day', 'Time', 'Size'], dtype='object')
```

```
data = data[['TotalBill', 'Tips', 'Smoker', 'Day', 'Time', 'Size']]
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88 entries, 0 to 98
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   TotalBill  88 non-null     float64
 1   Tips       88 non-null     float64
 2   Smoker     88 non-null     object
 3   Day        88 non-null     object
 4   Time       88 non-null     object
 5   Size       88 non-null     float64
dtypes: float64(3), object(3)
memory usage: 4.8+ KB
```

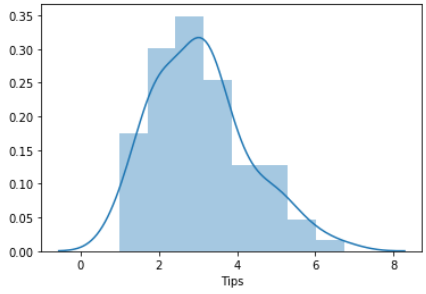
```
data.describe()
```

	TotalBill	Tips	Size
count	88.000000	88.000000	88.000000
mean	19.671136	3.065227	2.500000
std	7.861549	1.227688	0.830455
min	3.070000	1.000000	1.000000
25%	14.965000	2.025000	2.000000
50%	18.160000	3.000000	2.000000
75%	23.775000	3.635000	3.000000
max	48.270000	6.730000	4.000000

Histogram

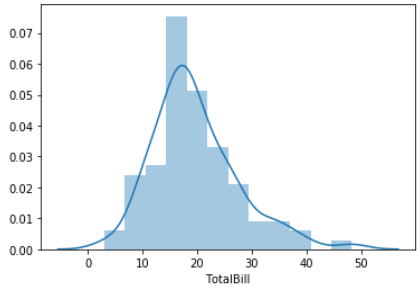
```
sns.distplot(data.Tips)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x119a9005388>
```




```
sns.distplot(data.TotalBill)
```

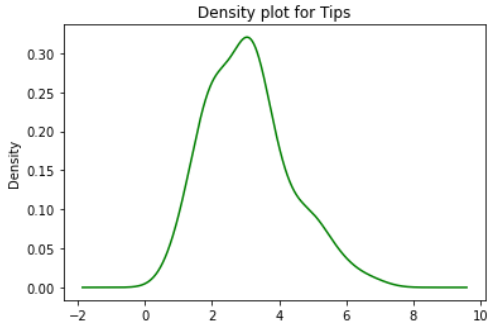
```
<matplotlib.axes._subplots.AxesSubplot at 0x119a9067e08>
```



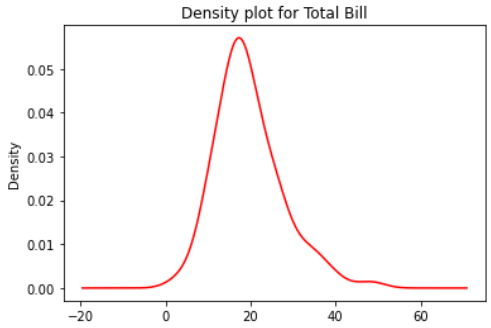
Most tips are around 3 dollars and most total bills are little less than 20 dollars.

Density plot

```
data.Tips.plot.density(color='green')
plt.title('Density plot for Tips')
plt.show()
```



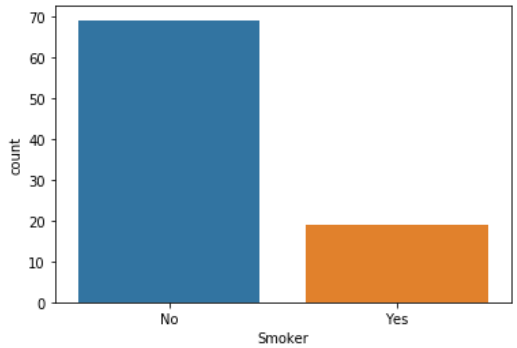

```
data.TotalBill.plot.density(color='red')
plt.title('Density plot for Total Bill')
plt.show()
```



Bar graph and its types

```
sns.countplot(x = 'Smoker', data = data)
```

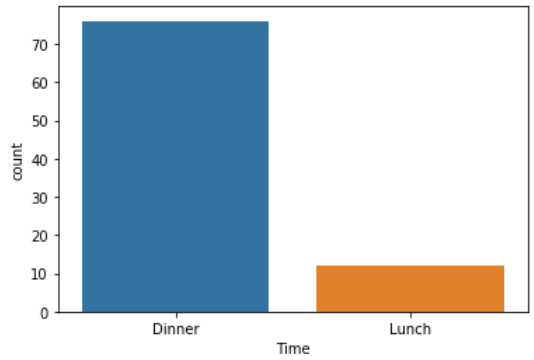
```
<matplotlib.axes._subplots.AxesSubplot at 0x119a9108f48>
```



Most customers are non smokers.

```
sns.countplot(x = 'Time', data = data)
```

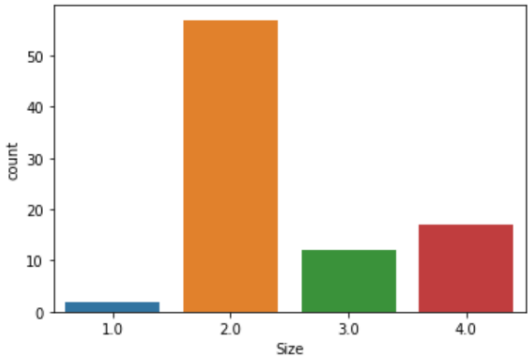
```
<matplotlib.axes._subplots.AxesSubplot at 0x119a916dc48>
```



Most customers come for dinner.

```
sns.countplot(x = 'Size', data = data)
```

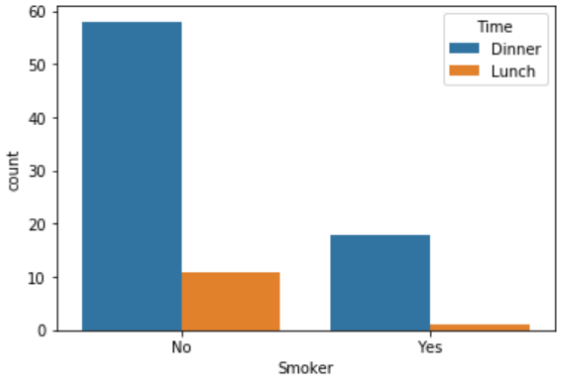
```
<matplotlib.axes._subplots.AxesSubplot at 0x119a7138748>
```



2 is the most common group size.

```
sns.countplot(x = 'Smoker', data = data, hue = 'Time')
```

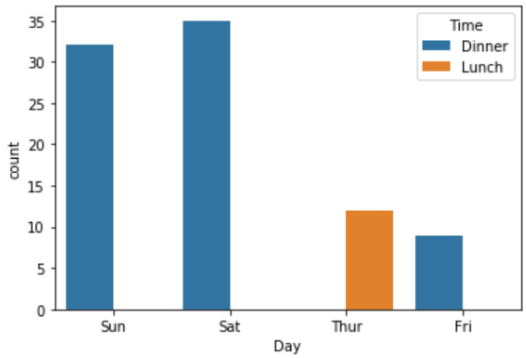
```
<matplotlib.axes._subplots.AxesSubplot at 0x119a930b3c8>
```



Greater number of non smokers at both dinner and lunch.

```
sns.countplot(x = 'Day', data = data, hue = 'Time')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x119a935c208>
```

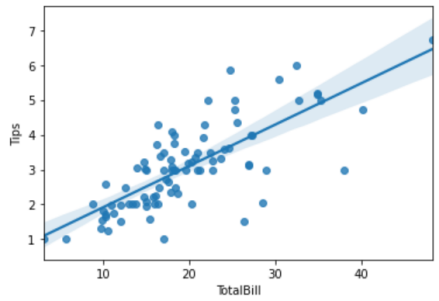


Most customers come to dinner only and for lunch only on Thursday.

Scatter Plot

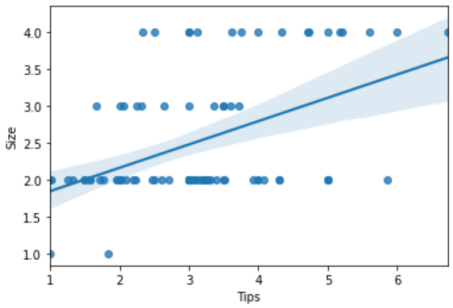
```
sns.regplot(x = data.TotalBill, y = data.Tips)
```

```
<AxesSubplot:xlabel='TotalBill', ylabel='Tips'>
```



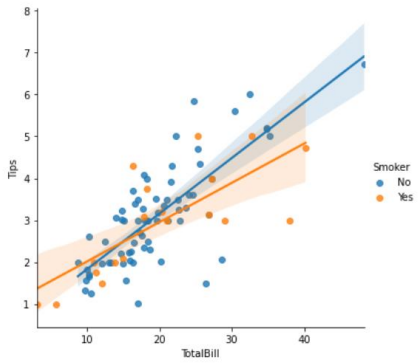

```
sns.regplot(x = data.Tips, y = data.Size)
```

```
<AxesSubplot:xlabel='Tips', ylabel='Size'>
```



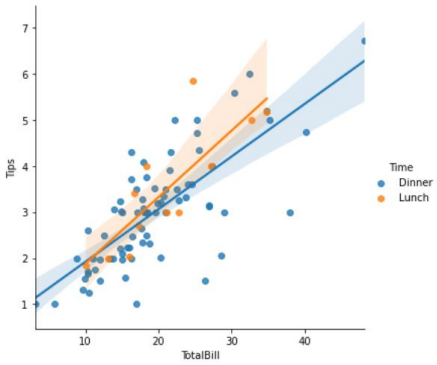

```
sns.lmplot(x = 'TotalBill', y = 'Tips', hue = 'Smoker', data = data)
```

```
<seaborn.axisgrid.FacetGrid at 0x21d6c8eca30>
```




```
sns.lmplot(x = 'TotalBill', y = 'Tips', hue = 'Time', data = data)
```

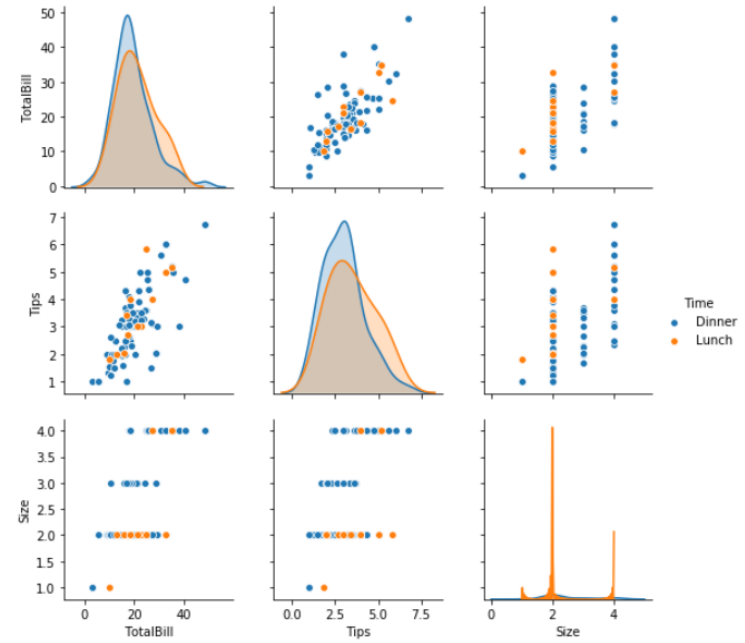
```
<seaborn.axisgrid.FacetGrid at 0x21d6c9406d0>
```



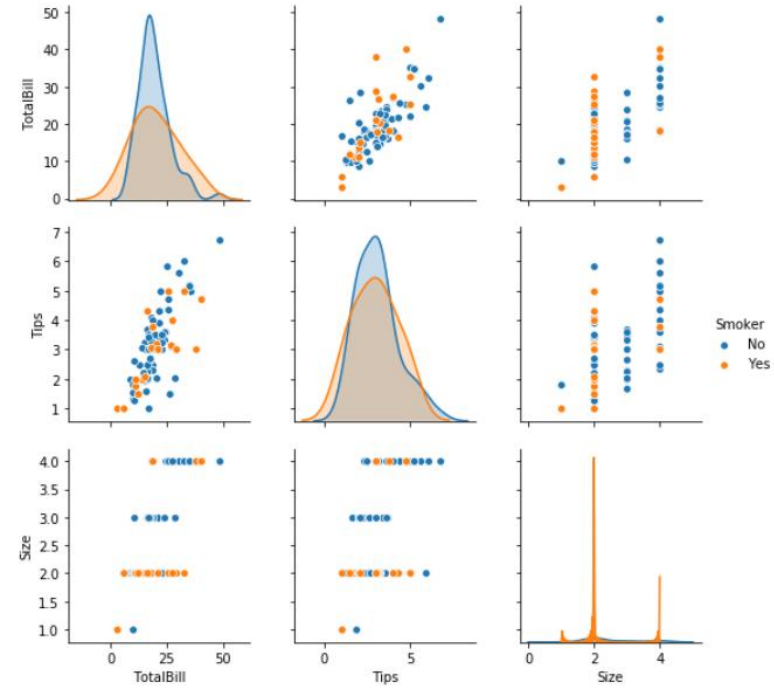
Non smokers tip slightly above smokers but their total bill is a little less than smokers.

Scatter Plot Matrix

```
sns.pairplot(data, kind = "scatter", hue = "Time" )
plt.show()
```



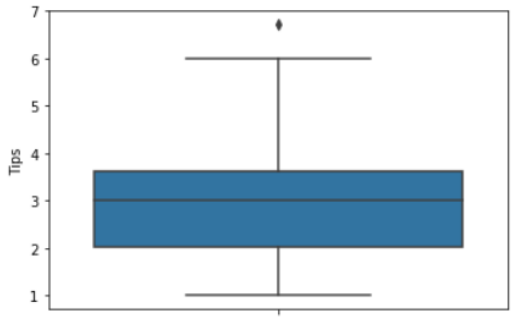

```
sns.pairplot(data, kind = "scatter", hue = "Smoker" )
plt.show()
```



Box plot analysis

```
sns.boxplot(y = data.Tips)
```

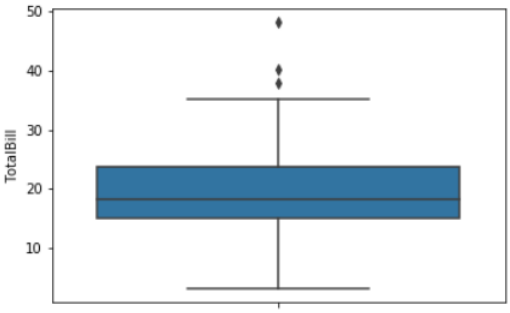
```
<matplotlib.axes._subplots.AxesSubplot at 0x119a93dfcc8>
```



Median value is around 3 and there is a single outlier at 7.

```
sns.boxplot(y = data.TotalBill)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x119a9454788>
```

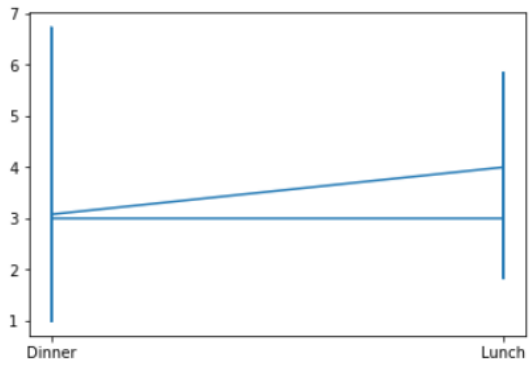


Median value is a little less than 20 and there are few outliers above 40.

Line graph

```
plt.plot(data['Time'], data['Tips'])
```

```
[<matplotlib.lines.Line2D at 0x21d6fb8be50>]
```



Pie graph

```
data.Smoker.value_counts().plot(kind = 'pie')
```

```
<AxesSubplot:ylabel='Smoker'>
```

Nearly 75% customers are non smokers.

Identifying outliers

The boxplot for Tips has an outlier with value of 7, and for the boxplot of TotalBill there are a couple outliers with value around 40 to 50.

Data Reduction (Using PCA algorithm)

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
sc = StandardScaler()
X = sc.fit_transform(data.iloc[:, [0, 1, 5]].values)
pca = PCA(n_components = 1)
X = pca.fit_transform(X)
```

```
X
```

```
array([[-1.51468052],
       [-1.09436359],
       [ 0.63188989],
       [ 0.1170134 ],
       [ 1.60309114],
       [ 2.18563594],
       [-1.69083215],
       [ 1.54951736],
       [-1.21375566],
       [-0.6257235 ],
       [-1.71108714],
       [ 3.11369536],
       [-1.3705762 ],
       [ 0.82322733],
       [-0.72240252],
       [ 0.24313138],
       [ 0.35895717],
       [ 0.31214012],
       [ 0.53151425],
       [ 0.030132371
```

Correlation analysis

```
data.corr()
```

	TotalBill	Tips	Size
TotalBill	1.000000	0.760541	0.595077
Tips	0.760541	1.000000	0.467644
Size	0.595077	0.467644	1.000000

```
data.cov()
```

	TotalBill	Tips	Size
TotalBill	61.803960	7.340389	3.885057
Tips	7.340389	1.507218	0.476782
Size	3.885057	0.476782	0.689655

```
data.columns
```

```
Index(['TotalBill', 'Tips', 'Smoker', 'Day', 'Time', 'Size'], dtype='object')
```

```
mat = pd.crosstab(index = data.Day, columns = data.Time)
```

```
sq, p, f, exp = chi2_contingency(mat)
```

```
print("Chi square value", sq)
```

```
Chi square value 87.99999999999999
```

```
print("p value", p)
```

```
p value 5.889492383190853e-19
```

```
print("Degrees of freedom", f)
```

```
Degrees of freedom 3
```

Day and time are correlated.

Conclusion:

Hence successfully completed EDA and visualisation on the Tips dataset.

Post lab Questions:

Use mtcars.csv, Tips.csv, diamond.csv to compelte the following post lab questions**.**

1. Which of the following plot is a visual representation of the statistical five-number summary of a data?
    a. BoxPlot
    b. BarPlot
    c. Histogram
    d. ScatterPlot

Ans: a. BoxPlot

2. Which of the following statement is not true about histograms?
    a. Represent the frequency distribution of categorical variables
    b. It is a graphical representation of data using bars of different heights
    c. Groups numbers into ranges and the height of each bar depicts the frequency of each range or bin
    d. Represent the frequency distribution of numerical variables

Ans: a. Represent the frequency distribution of categorical variables

3. Correlation between two variables X&Y is 0.85. Now, after adding the value 2 to all the values of X, the correlation co-efficient will be
    a. 0.85
    b. 0.87
    c. 0.65
    d. 0.82

Ans: a. 0.85

4. Which of the following can be inferred from scatter plot of ‘mpg’ (Miles per gallon) vs ‘wt’ (Weight of car) from the dataset **mtcars.csv**?
    a. As weight of the car increases, the mpg decreases
    b. As weight of the car increases, the mpg increases
    c. There is no relation between weight of the car and mpg

d. When weight increases, mpg increases exponentially

Ans: a. As weight of the car increases, the mpg decreases

5. Plot a boxplot for "price" vs "cut" from the dataset "**diamond.csv**". Which of the categories under "cut" have the highest median price?
   a. Good
   b. Very Good
   c. Premium
   d. Fair

Ans: d. Fair

6. The command used for line plot from the package Matplotlib?
   a. plot( )
   b. line( )
   c. join( )
   d. plt( )

Ans: a. plot( )

7. Read the given dataset "Tips.csv" as a dataframe "Data". For the given dataframe "Data" which reads plot a histogram for the variable 'TotalBill' to check which range has the highest frequency.
   a. 10-15
   b. 15-20
   c. 20-25
   d. 25-30

Ans: b. 15-20

8. For the given dataframe "Data" draw a bar chart for the variable "Day". Identify the category with the maximum count
   a. Friday
   b. Thursday
   c. Saturday
   d. Sunday

Ans: c. Saturday

9. On which day sum of the total bill was maximum?
   a. Friday
   b. Saturday
   c. Sunday
   d. Thursday

Ans: b. Saturday

10. In which of the following methods of data reduction are redundant attributes or dimensions or irrelevant data identified and removed
    a. Attribute subset selection
    b. Data cube aggregation
    c. Dimensionality reduction
    d. Numerosity reduction

Ans: c. Dimensionality reduction