PRACTICAL – 8

OBJECTIVE: To implement the K-Means Clustering algorithm in R using the Iris dataset

SOFTWARE/TOOL USED:

- RStudio / R Programming Language
- Built-in Iris Dataset
- Libraries: stats, ggplot2 (for visualization)

THEORY:

Clustering is an unsupervised machine learning technique that groups similar data points into clusters based on feature similarity.
Unlike classification, clustering does not use predefined labels.

The K-Means algorithm partitions the dataset into *K* clusters by minimizing the distance between data points and the cluster centroid (mean point).

Algorithm Steps:

1. Choose the number of clusters (*K*).
2. Initialize *K* centroids randomly.
3. Assign each data point to the nearest centroid.
4. Recalculate centroids based on assigned points.
5. Repeat steps 3–4 until centroids do not change significantly.

The **Iris dataset** contains 150 flower samples with four features:

- Sepal.Length
- Sepal.Width
- Petal.Length
- Petal.Width
  and three species: *Setosa*, *Versicolor*, *Virginica*.

PROCEDURE:

Load the Dataset

- Use R's built-in iris dataset.
- Display the first few records using head(iris).

Select Numerical Features

- K-Means works on numerical data, so select the first four columns (sepal & petal measurements).

Apply K-Means Clustering

- Set the number of clusters $K = 3$ (since there are 3 flower species).
- Use the kmeans() function to perform clustering.

Compare with Actual Species

- Compare the predicted cluster groups with actual species labels to evaluate clustering performance.

Visualize the Clusters

- Plot the clusters using ggplot2 or plot() functions.
- Mark the cluster centers using points() or geom_point().

RESULT:

The K-Means clustering algorithm successfully divided the Iris dataset into three clusters corresponding closely to the actual species categories.The plotted visualization clearly shows distinct clusters with separate centroids.

CONCLUSION:

K-Means clustering was effectively implemented on the Iris dataset in R.
The clustering results showed strong alignment with actual species labels, demonstrating that unsupervised learning can reveal natural groupings in data.
Visualization of clusters and centroids further validates the separation among the three flower species.

A data.frame: 6 × 5

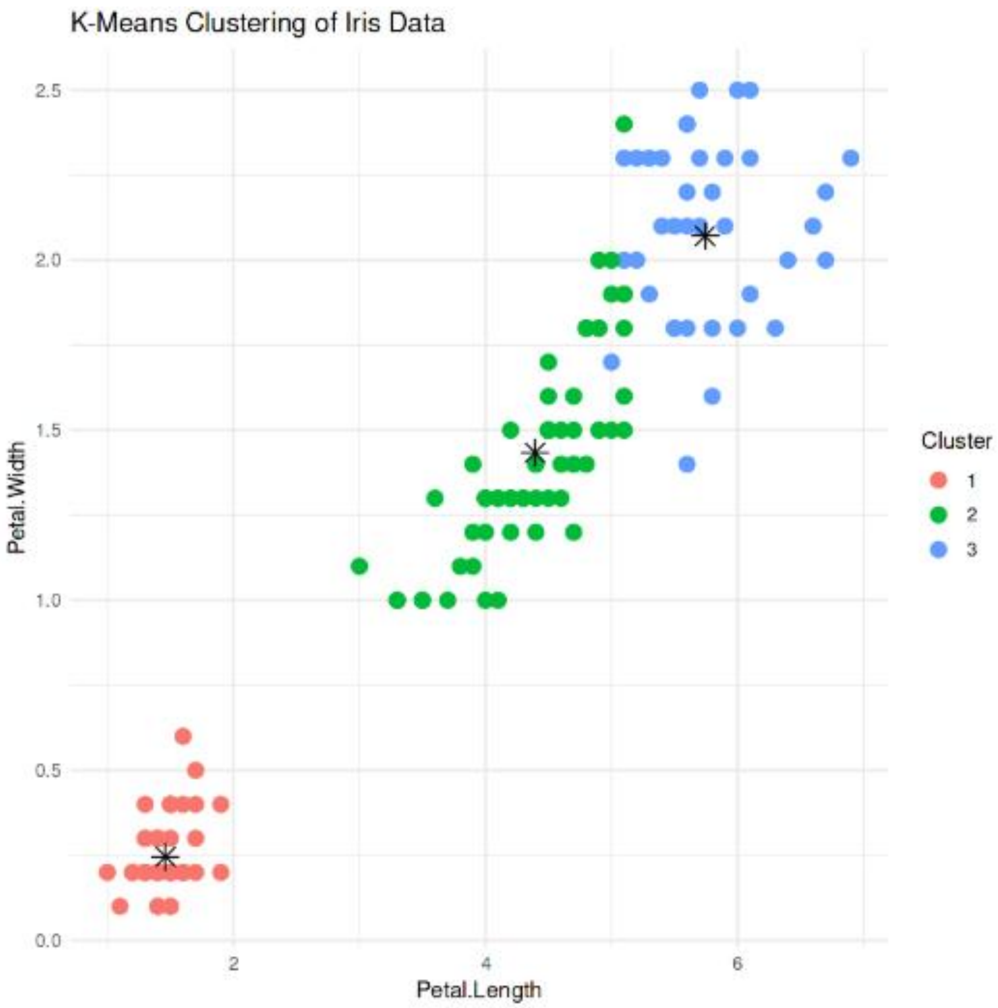
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

```
K-means clustering with 3 clusters of sizes 50, 62, 38
```

```
Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.006000    3.428000     1.462000    0.246000
2     5.901613    2.748387     4.393548    1.433871
3     6.850000    3.073684     5.742105    2.071053

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 3 3
[112] 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3 3 3 2 3
[149] 3 2

Within cluster sum of squares by cluster:
[1] 15.15100 39.82097 23.87947
 (between_SS / total_SS =  88.4 %)
```



PRACTICAL – 9

OBJECTIVE: To implement Linear Regression in R to predict continuous values.

SOFTWARE/TOOL USED:

- R Programming Language
- RStudio / Kaggle / Colab
- Dataset (created manually)

THEORY:

Linear Regression is a supervised machine learning method used to model the relationship between a dependent variable (Y) and one or more independent variables (X).

The Simple Linear Regression Formula is:

$$Y = a + bX$$

Where:

- Y = Dependent variable (value to be predicted)
- X = Independent variable (input feature)
- a = Intercept
- b = Slope (coefficient)

The goal is to find the best fit line that minimizes the difference between predicted and actual values (using Least Squares Method).

PROCEDURE:

- Create a dataset of hours studied vs marks scored.
- Load the dataset in R.
- Fit a Linear Regression model using lm() function.
- Print the regression summary.
- Use the model to predict marks for new values.
- Plot the best-fit regression line on a scatter plot.

DATASET:

Hours_Studied	Marks
2	40
3	45
4	50
5	55
6	60
7	63
8	68

RESULT:

Data from Excel and OData Feed was successfully imported into Power BI and loaded into the target data model for further analysis.

CONCLUSION:

This practical demonstrates how Power BI can connect to and import data from different legacy sources like Excel and OData Feed. The integrated data can then be used for creating BI dashboards, reports, and visual analytics.

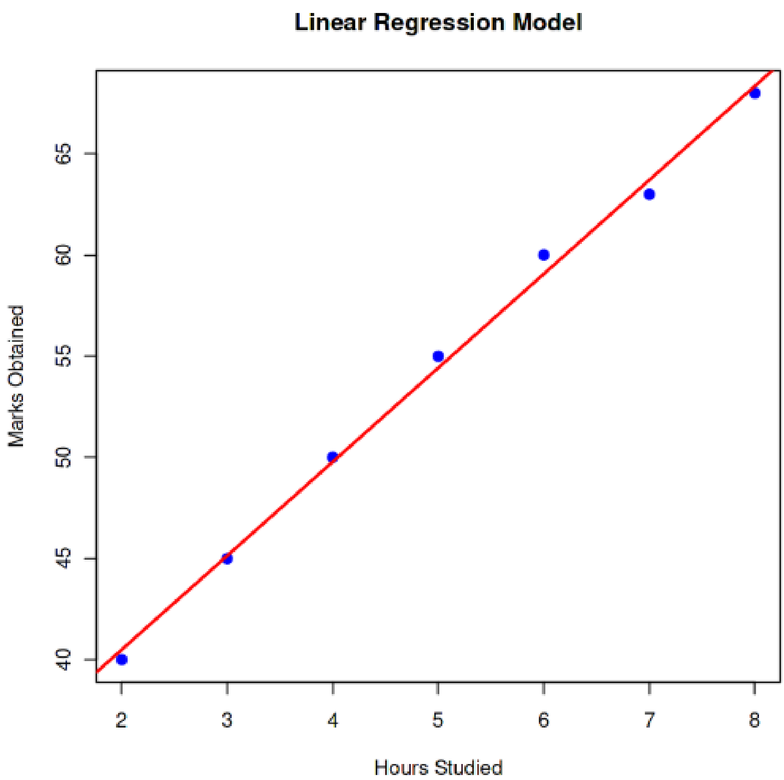
```
  hours marks
1     2    40
2     3    45
3     4    50
4     5    55
5     6    60
6     7    63
7     8    68

Call:
lm(formula = marks ~ hours, data = data)

Residuals:
      1       2       3       4       5       6       7 
-0.5000 -0.1429  0.2143  0.5714  0.9286 -0.7143 -0.3571 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.2143     0.6662   46.85 8.37e-08 ***
hours         4.6429     0.1237   37.53 2.53e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6547 on 5 degrees of freedom
Multiple R-squared:  0.9965,	Adjusted R-squared:  0.9958 
F-statistic:  1408 on 1 and 5 DF,  p-value: 2.531e-07
Predicted Marks for 6.5 hours of study: 61.39286
```


Linear Regression Model
Marks Obtained
Hours Studied
40
45
50
55
60
65
2
3
4
5
6
7
8

PRACTICAL – 10

OBJECTIVE: To perform Time Series Analysis in R and visualize the trend over a period of time.

SOFTWARE/TOOL USED:

- R Programming Language
- RStudio / Kaggle / Colab
- Time Series dataset (Monthly Sales Data)

THEORY:

Time Series Analysis is a statistical technique used to analyze data points collected or recorded at specific time intervals (daily, monthly, yearly, etc.).
It helps identify:

- Trend (overall increase/decrease),
- Seasonality (repeated patterns),
- Cyclic behavior, and
- Random variation.

In R, the ts() function is used to convert numeric data into a time-series object, which can then be visualized using plot().

PROCEDURE:

- Create a numeric vector representing the monthly data points.
- Convert the numeric vector into a time series object using ts().
- Plot the time series to visualize trends.
- Interpret the graph to understand data behavior over time.

DATASET:

Month	Sales
Jan	1200
Feb	1350
Mar	1280

Month	Sales
Apr	1490
May	1600
Jun	1550
Jul	1700
Aug	1650
Sep	1580
Oct	1720
Nov	1800
Dec	1900

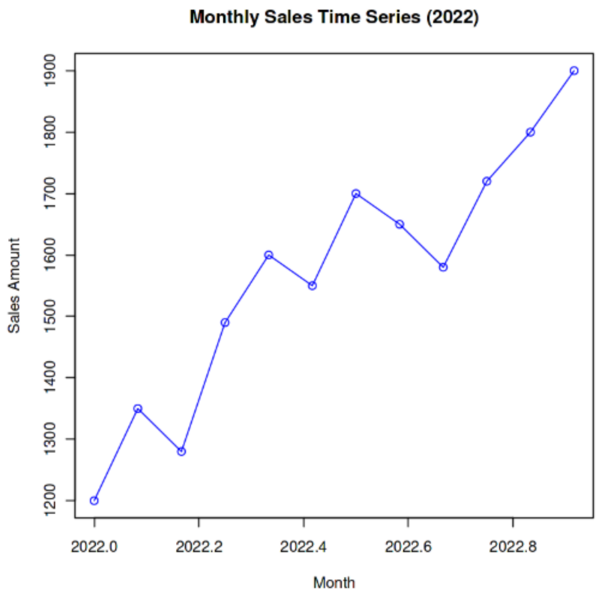
RESULT:

Time Series Analysis was successfully performed. The plotted time series clearly shows an overall increasing sales trend over the months.

CONCLUSION:

Time Series Analysis helps in understanding historical patterns and forecasting future values. The experiment demonstrated how monthly sales data can be analyzed to observe trend and seasonal patterns, which are useful in business planning and forecasting.

```
      Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
2022 1200 1350 1280 1490 1600 1550 1700 1650 1580 1720 1800 1900
```



PRACTICAL – 11

OBJECTIVE: To perform data modeling and analytical operations using Pivot Tables in Microsoft Excel.

SOFTWARE/TOOL USED:

- Microsoft Excel (Any version that supports Pivot Tables)

THEORY:

A Pivot Table is a powerful data summarization and reporting tool in Excel.
It allows users to:

- Summarize large datasets
- Group data by categories
- Calculate totals, averages, and percentages
- Filter and slice data dynamically
- Create reports and dashboards easily

Pivot tables convert raw data into meaningful insights by arranging data fields into:

- Rows
- Columns
- Values
- Filters

PROCEDURE:

1. Open Excel and enter or import the dataset into a worksheet.
2. Select the entire table including headers.
3. Go to the Ribbon → Click Insert → PivotTable.
4. Choose New Worksheet and click OK.
5. The PivotTable Field List appears on the right.
6. Drag Category to the Rows area.
7. Drag Region to the Columns area.
8. Drag Sales to the Values area.
    - Ensure it shows as Sum of Sales.

9. (Optional) Drag Product to Filters area to analyze product-wise.
10. Format the Pivot Table for readability:
    - Use Design → Report Layout → Show in Tabular Form
    - Apply border + bold headers
11. Insert a Pivot Chart:
    - Go to PivotTable Analyze → PivotChart
    - Select Column Chart or Pie Chart

DATASET:

Order ID	Product	Category	Region	Sales
101	Laptop	Electronics	North	55000
102	Keyboard	Electronics	West	1200
103	Chair	Furniture	East	2500
104	Desk	Furniture	North	7500
105	Mouse	Electronics	South	900
106	Monitor	Electronics	East	14500
107	Sofa	Furniture	South	22000
108	Table	Furniture	West	6500

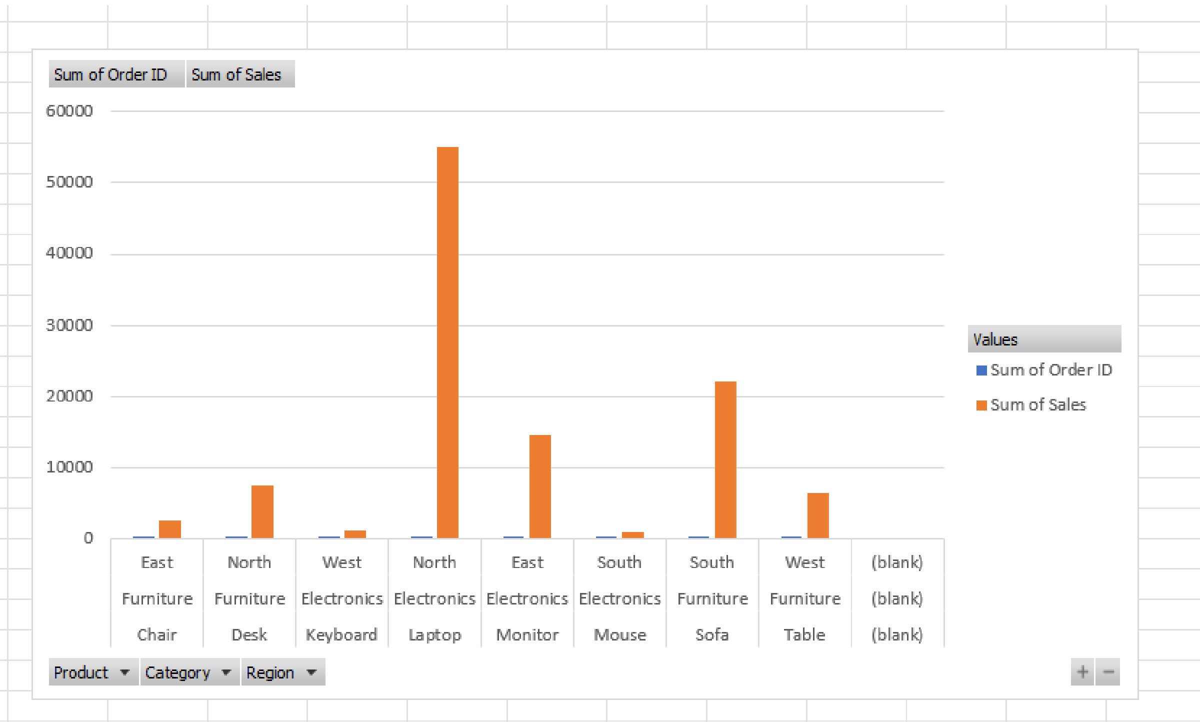
RESULT:

Data was successfully summarized and analyzed using a pivot table.The pivot chart visually represented sales distribution across different regions and categories.

CONCLUSION:

Pivot Tables in Excel provide an easy and effective method to analyze and interpret large datasets.
They support summarization, comparison, and reporting without requiring complex formulas making them essential for business data analytics.

	A	B	C	D	E	F
3	**Product**	**Category**	**Region**	**Sum of Order ID**	**Sum of Sales**	
4	**Chair**	**Furniture**	East	103	2500	
5		**Furniture Total**		**103**	**2500**	
6	**Chair Total**			**103**	**2500**	
7	**Desk**	**Furniture**	North	104	7500	
8		**Furniture Total**		**104**	**7500**	
9	**Desk Total**			**104**	**7500**	
10	**Keyboard**	**Electronics**	West	102	1200	
11		**Electronics Total**		**102**	**1200**	
12	**Keyboard Total**			**102**	**1200**	
13	**Laptop**	**Electronics**	North	101	55000	
14		**Electronics Total**		**101**	**55000**	
15	**Laptop Total**			**101**	**55000**	
16	**Monitor**	**Electronics**	East	106	14500	
17		**Electronics Total**		**106**	**14500**	
18	**Monitor Total**			**106**	**14500**	
19	**Mouse**	**Electronics**	South	105	900	
20		**Electronics Total**		**105**	**900**	
21	**Mouse Total**			**105**	**900**	
22	**Sofa**	**Furniture**	South	107	22000	
23		**Furniture Total**		**107**	**22000**	
24	**Sofa Total**			**107**	**22000**	
25	**Table**	**Furniture**	West	108	6500	
26		**Furniture Total**		**108**	**6500**	
27	**Table Total**			**108**	**6500**	
28	**(blank)**	**(blank)**	(blank)			
29		**(blank) Total**				
30	**(blank) Total**					
31	**Grand Total**			**836**	**110100**	
32						
33						



PRACTICAL – 12

OBJECTIVE: To perform data analysis and create visualizations in Advanced Excel using functions, filters, pivot charts, slicers, and conditional formatting.

SOFTWARE/TOOL USED:

- Microsoft Excel (Advanced Features)

THEORY:

Advanced Excel tools help analyze and interpret datasets effectively.
Some commonly used features include:

Feature	Purpose
Sorting & Filtering	Organize and extract specific data quickly
Conditional Formatting	Highlight data patterns (e.g., highest values, duplicates)
Pivot Tables	Summarize and group large datasets
Pivot Charts	Visualize pivot table results
Slicers	Create interactive filtering for tables and charts
Advanced Formulas	Perform calculations and derive insights

Frequently used formulas:

Function	Meaning	Example
SUM()	Computes total	=SUM(B2:B20)
AVERAGE()	Finds mean	=AVERAGE(C2:C20)
COUNTIF()	Counts matching values	=COUNTIF(A2:A20,"North")
IF()	Conditional result	=IF(D2>500,"High","Low")

These tools together help convert raw data into meaningful analytics.

PROCEDURE:

1. Enter / Import the dataset into Excel.
2. Select the entire data table and apply Format as Table.
3. Use Sorting & Filtering to view sales results by department.

4. Apply Conditional Formatting to highlight:

- Top 3 Sales values
- Ratings greater than 4.5

5. Create a Pivot Table:

- Insert → PivotTable → New Worksheet
- Rows → Department
- Values → Sales (Sum)
- Values → Rating (Average)

6. Insert Pivot Chart:

- PivotTable → PivotChart → Choose **Column Chart**

7. Add **Slicer** for Department:

- PivotTable Analyze → Insert Slicer → Select Department

8. Format chart for readability:

- Add title, axis labels, bold headings.

DATASET:

Emp ID	Name	Department	Sales	Rating
101	Rahul	Sales	45000	4.2
102	Anita	HR	12000	3.5
103	Mohit	Sales	62000	4.8
104	Neha	Finance	38000	4.0
105	Vivek	Sales	54000	4.6
106	Sonam	HR	15000	3.8
107	Tarun	Finance	41000	4.1
108	Raj	Sales	47000	4.3

RESULT:

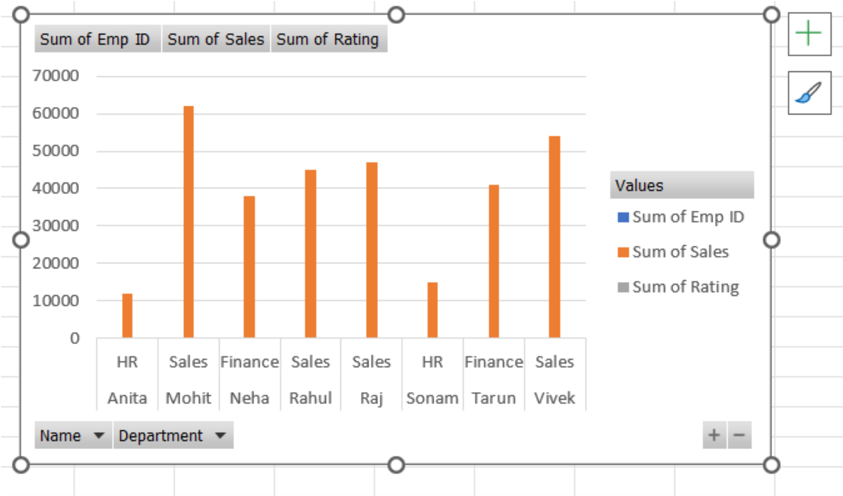
Data from Excel and OData Feed was successfully imported into Power BI and loaded into the target data model for further analysis.

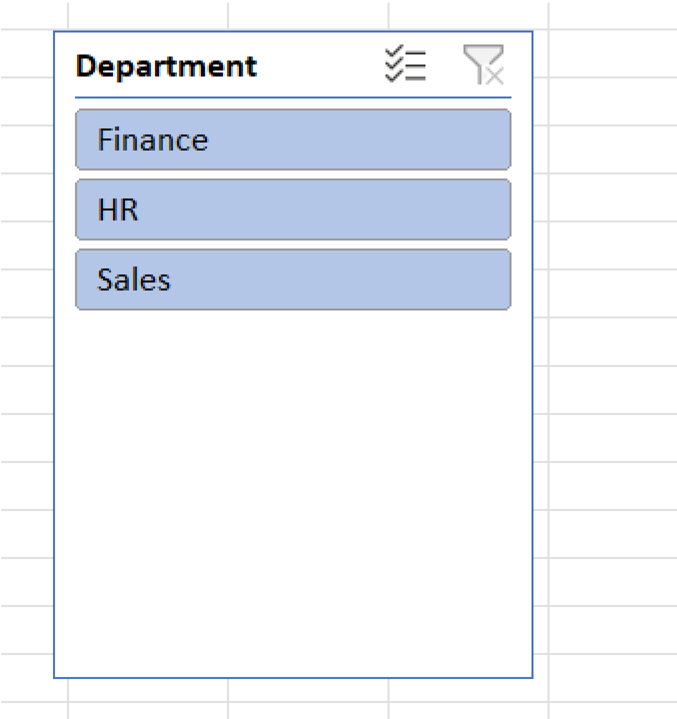
CONCLUSION:

This practical demonstrates how Power BI can connect to and import data from different legacy sources like Excel and OData Feed. The integrated data can then be used for creating BI dashboards, reports, and visual analytics.

Emp ID	Name	Department	Sales	Rating
103	Mohit	Sales	62000	4.8
105	Vivek	Sales	54000	4.6
108	Raj	Sales	47000	4.3
101	Rahul	Sales	45000	4.2
107	Tarun	Finance	41000	4.1
104	Neha	Finance	38000	4
106	Sonam	HR	15000	3.8
102	Anita	HR	12000	3.5

Row Labels	Sum of Emp ID	Sum of Sales	Sum of Rating
Anita	**102**	**12000**	**3.5**
HR	102	12000	3.5
Mohit	**103**	**62000**	**4.8**
Sales	103	62000	4.8
Neha	**104**	**38000**	**4**
Finance	104	38000	4
Rahul	**101**	**45000**	**4.2**
Sales	101	45000	4.2
Raj	**108**	**47000**	**4.3**
Sales	108	47000	4.3
Sonam	**106**	**15000**	**3.8**
HR	106	15000	3.8
Tarun	**107**	**41000**	**4.1**
Finance	107	41000	4.1
Vivek	**105**	**54000**	**4.6**
Sales	105	54000	4.6
Grand Total	**836**	**314000**	**33.3**


Sum of Emp ID
Sum of Sales
Sum of Rating
70000
60000
50000
40000
30000
20000
10000
0
HR
Anita
Sales
Mohit
Finance
Neha
Sales
Rahul
Sales
Raj
HR
Sonam
Finance
Tarun
Sales
Vivek
Values
Sum of Emp ID
Sum of Sales
Sum of Rating
Name
Department


Department
Finance
HR
Sales