Practical-1

Aim: To analyze sales data using central tendency measures, graphically represent findings, and summarize insights across payment methods, regions, and product categories.

- **PROGRAM:**

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

df = pd.read_csv("Online Sales Data.csv")
print(df.head(5))

#UNIT SOLD
unit_sold_data = df['Units Sold']
mean_unit_sold = unit_sold_data.mean()
median_unit_sold = unit_sold_data.median()
mode_unit_sold = unit_sold_data.mode()
var_unit_sold = unit_sold_data.var()
sd_unit_sold = unit_sold_data.std()

print("UNIT SOLD")
print("Mean : ", mean_unit_sold)
print("Median : ", median_unit_sold)
print("Mode : ", mode_unit_sold)
print("Variance : ", var_unit_sold)
print("Standard Deviation : ",sd_unit_sold)

#UNIT PRICE
unit_price_data = df['Unit Price']
mean_unit_price = unit_price_data.mean()
median_unit_price = unit_price_data.median()
mode_unit_price = unit_price_data.mode()
var_unit_price = unit_price_data.var()
sd_unit_price = unit_price_data.std()

print("UNIT PRICE")
print("Mean : ", mean_unit_price)
print("Median : ", median_unit_price)
print("Mode : ", mode_unit_price)
print("Variance : ", var_unit_price)
print("Standard Deviation : ",sd_unit_price)
```

```
#TOTAL REVENUE
total_revenue_data = df['Total Revenue']
mean_total_revenue = total_revenue_data.mean()
median_total_revenue = total_revenue_data.median()
mode_total_revenue = total_revenue_data.mode()
var_total_revenue = total_revenue_data.var()
sd_total_revenue = total_revenue_data.std()

print("TOTAL REVENUE")
print("Mean : ",mean_total_revenue)
print("Median : ",median_total_revenue)
print("Mode : ",mode_total_revenue)
print("Variance : ",var_total_revenue)
print("Standard Deviation : ",sd_total_revenue)

#SUMMARIZE REGION
region_data =df['Region']
na_data=region_data.value_counts()['North America']
e_data=region_data.value_counts()['Europe']
a_data=region_data.value_counts()['Asia']
data=[na_data,e_data,a_data]
labels=[ 'North America','Europe','Asia']

# Bar Chart
plt.figure(figsize=(10, 7))
plt.bar(labels, data)
plt.xlabel("Region")
plt.ylabel("Number of Sales")
plt.title("Sales Distribution by Region")
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()

# Pie Chart
plt.figure(figsize=(8, 6))
plt.pie(data, labels=labels, autopct='%1.1f%%', startangle=90)
plt.title("Sales Distribution by Region")
plt.axis('equal')
plt.show()

# SUMMARIZE PRODUCT CATEGORY

product_category_data = df['Product Category']
Electronics_data=product_category_data.value_counts()['Electronics']
Home_data=product_category_data.value_counts()['Home Appliances']
Clothing_data=product_category_data.value_counts()['Clothing']
```

```
Books_data=product_category_data.value_counts()['Books']
Beauty_Products=product_category_data.value_counts()['Beauty Products']
Sports_data=product_category_data.value_counts()['Sports']

data=[Electronics_data,Home_data,Clothing_data,Books_data,Beauty_Products,Sport
s_data]
labels=['Electronics','Home  Appliances','Clothing','Books','Beauty Products','Sports']
```

#Pie Chart

```
plt.figure(figsize=(8, 6))
plt.pie(data, labels=labels, autopct='%1.1f%%', startangle=90)
plt.title("Product Category Distribution")
plt.axis('equal')
plt.show()
```

#Bar Chart

```
plt.figure(figsize=(10, 7))
plt.bar(labels, data)
plt.xlabel("Product Category")
plt.ylabel("Number of Products")
plt.title("Product Category Distribution")
plt.show()
```

#SUMMARIZE PAYEMENT CATEGORY

```
payment_method_data=df['Payment Method']
debitcard_payment=payment_method_data.value_counts()["Debit Card"]
creditcard_payment=payment_method_data.value_counts()["Credit Card"]
paypal_payment=payment_method_data.value_counts()["PayPal"]

data=[debitcard_payment,creditcard_payment,paypal_payment]
label=[ "Debit Card","Credit Card","PayPal"]
```

#Pie Chart

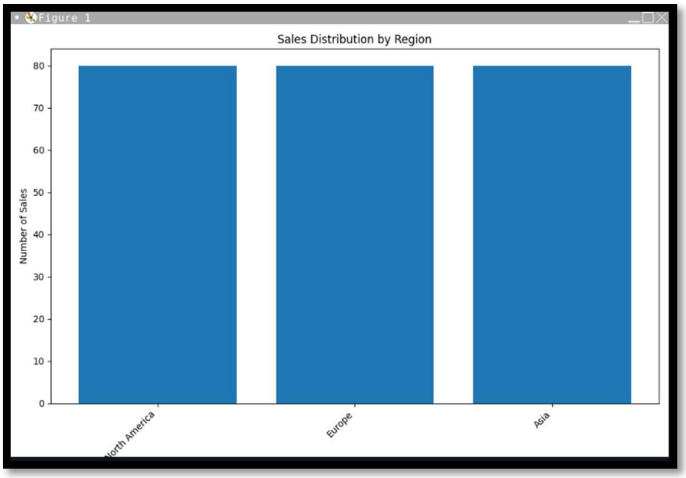
```
fig=plt.figure(figsize=(10,7))
plt.pie(data,labels=label)
plt.show()
```

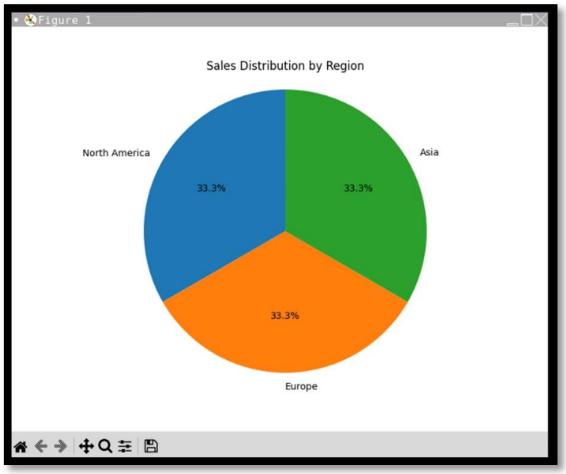
#Bar Chart

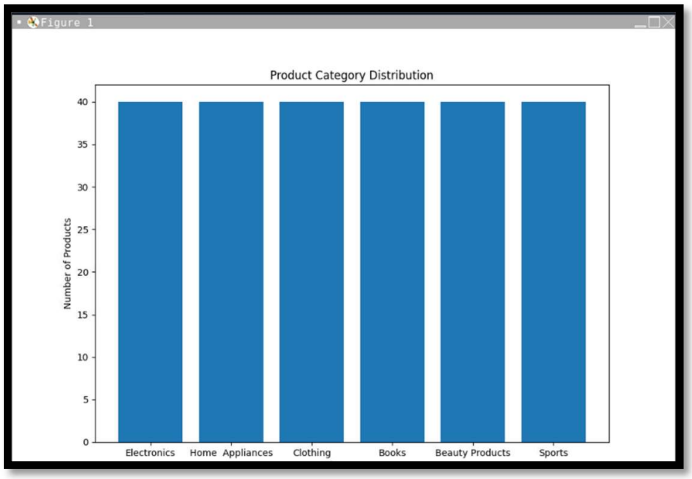
```
fig = plt.figure(figsize=(10, 7))
plt.bar(label, data)
plt.xlabel("Payment Method")
plt.ylabel("Number of Transactions")
plt.title("Payment Method Frequency")
plt.show()
```

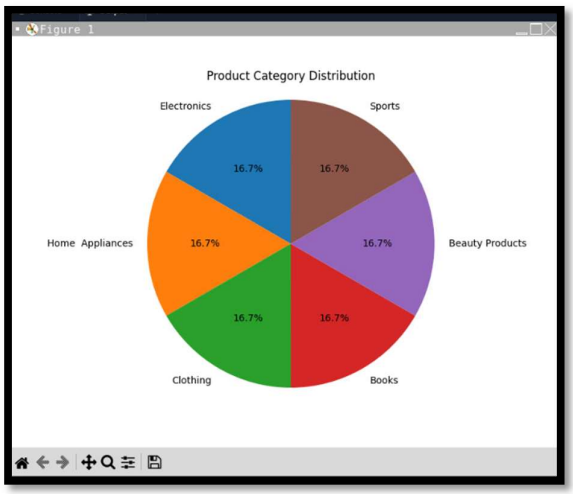
➢ OUTPUT:

```
   Transaction ID  ... Payment Method
0           10001  ...    Credit Card
1           10002  ...         PayPal
2           10003  ...     Debit Card
3           10004  ...    Credit Card
4           10005  ...         PayPal

[5 rows x 9 columns]
UNIT SOLD
Mean :  2.158333333333333
Median :  2.0
Mode :  0    1
Name: Units Sold, dtype: int64
Variance :  1.7488842398884237
Standard Deviation :  1.3224538706088858
UNIT PRICE
Mean :  236.3955833333333
Median :  89.99
Mode :  0    49.99
Name: Unit Price, dtype: float64
Variance :  184424.46376953277
Standard Deviation :  429.4466949104775
TOTAL REVENUE
Mean :  335.699375
Median :  179.97
Mode :  0    299.99
Name: Total Revenue, dtype: float64
Variance :  236005.9816778504
Standard Deviation :  485.8044685651321
```









Practical-2

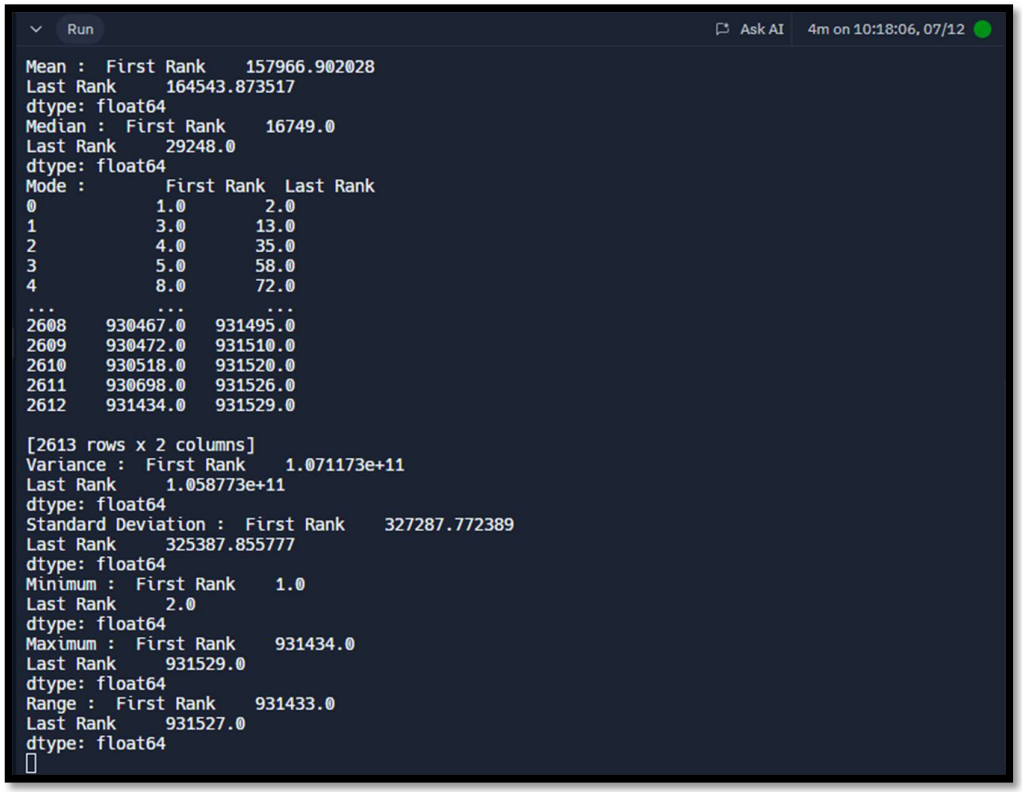
Aim: PDF to CSV Conversion and Perform Statistical Data Analysis

- **PROGRAM:**

```
import pdfplumber
import tabula
import csv

acpc_institute = "ACPC 1st Round Institute-wise allotment status.pdf"
acpc_cuttoff = "ACPC 1st Round Cut-off analysis.pdf"

def pdf_to_csv(pdf_file, csv_file):
  with pdfplumber.open(pdf_file) as pdf:
    for i in range(len(pdf.pages)):
      page = pdf.pages[i]
      table = page.extract_tables()[0]

      with open(csv_file, 'a', newline='') as f:
        writer = csv.writer(f)
        writer.writerows(table)

pdf_to_csv(acpc_institute, "acpc_institute.csv")
pdf_to_csv(acpc_cuttoff, "acpc_cutoff.csv")

import pandas as pd
import statistics

# Load the CSV file
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("acpc_cutoff.csv")

mean = df[["First Rank","Last Rank"]].mean()
median = df[["First Rank","Last Rank"]].median()
mode = df[["First Rank","Last Rank"]].mode()
var = df[["First Rank","Last Rank"]].var()
stdev = df[["First Rank","Last Rank"]].std()
min = df[["First Rank","Last Rank"]].min()
max = df[["First Rank","Last Rank"]].max()

print("Mean : ",mean)
print("Median : ",median)
print("Mode : ",mode)
print("Variance : ",var)
```

```
print("Standard Deviation : ",stdev)
print("Minimum : ",min)
print("Maximum : ",max)
print("Range : ",max-min)

gujcet_data = df[df['Quota'] == 'GUJCET-Based']
gujcet_count = gujcet_data['Quota'].count()
df['Is_GUJCET_Based'] = df['Quota'].apply(lambda x: 1 if x == 'GUJCET-Based' else 0)
grouped_data = df.groupby(by=['Institute Type'])['Quota'].count()

# Create the pie chart
plt.figure(figsize=(8, 6))
plt.pie(grouped_data.values, labels=grouped_data.index, autopct='%1.1f%%', startangle=90)
plt.title("Pie Chart of Institute Type by Quota")
plt.axis('equal')
plt.show()
```

➢ OUTPUT:



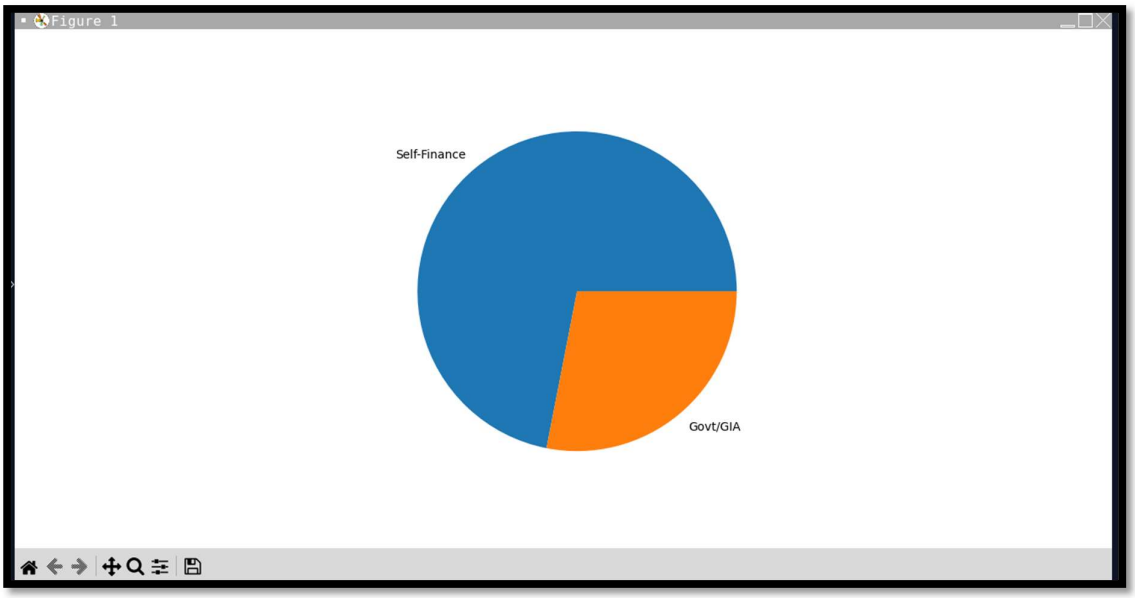


Figure 1
Self-Finance
Govt/GIA

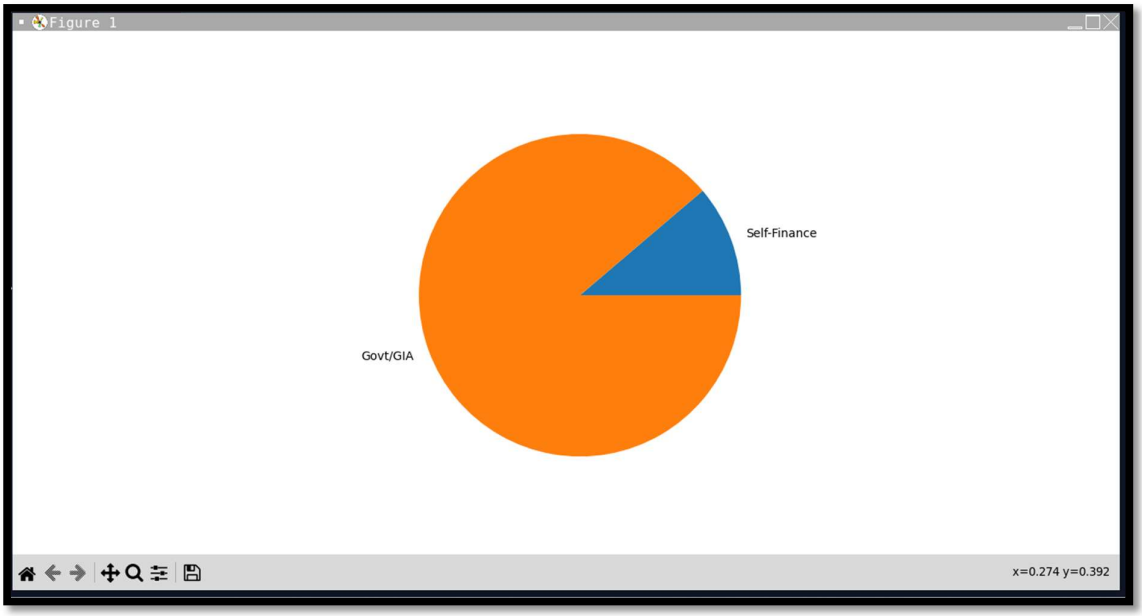


Figure 1
Self-Finance
Govt/GIA
x=0.274 y=0.392

Practical-3

Aim: Perform Data Analysis in Excel

➢ Central Tendency

The following Central Tendency is performed on acpc_cutoff.csv on "First Rank"& "Last Rank" Columns :

- **Average:**
  - To calculate the average (mean) of the provided columns:

    ```
    Eg) =AVERAGE(Data[First Rank]), =AVERAGE(Data[Last Rank])
    ```

- **Median:**
  - To find the median (middle value) of the provided columns:

    ```
    Eg) =MEDIAN(Data[First Rank]), =MEDIAN(Data[Last Rank])
    ```

- **Minimum:**
  - To determine the minimum value in the provided columns:

    ```
    Eg) =MIN(Data[First Rank]), =MIN(Data[Last Rank])
    ```

- **Maximum:**
  - To find the maximum value in the provided columns:

    ```
    Eg) =MAX(Data[First Rank]), =MAX(Data[Last Rank])
    ```

- **Range:**
  - To calculate the range (difference between maximum and minimum) of the provided columns:

    ```
    Eg) =MAX(Data[First Rank]) - MIN(Data[First Rank])
    ```

- **1st Quartile (Q1):**
  - To find the first quartile (25th percentile) of the provided columns:

    Eg) =QUARTILE.INC(Data[First Rank], 1)
    = QUARTILE.INC(Data[Last Rank], 1)

- **3rd Quartile (Q3):**
  - To calculate the third quartile (75th percentile) of the provided columns:

```
Eg) =QUARTILE.INC(Data[First Rank], 3)
     = QUARTILE.INC(Data[Last Rank], 3)
```

	A	B	C	D	E
1					
2		STATISTICAL DATA ANALYSIS			
3			FIRST RANK	LASTRANK	
4		AVERAGE	157966.902	164543.8735	
5		MEDIAN	16749	29248	
6		MIN	1	2	
7		MAX	931434	931529	
8		RANGE	931433	931527	
9		1ST Q	6406.5	12659	
10		3RD Q	32086.5	38312.5	
11					

➢ Conditional Formatting

- **Select the Range**: First, select the range of cells where you want to apply conditional formatting. You can select a single cell, a range of cells, or an entire column or row.

- **Open Conditional Formatting Menu**: Go to the "Home" tab on the Excel ribbon. In the "Styles" group, you'll find the "Conditional Formatting" button. Click on it to open a dropdown menu with various options.

- **Choose a Rule**: In the dropdown menu, you'll see different types of conditional formatting rules such as Highlight Cells Rules, Top/Bottom Rules, Data Bars, Color Scales, and Icon Sets. Select the type of rule that best suits your needs. Here are some common examples:

  - **Highlight Cell Rules**: Allows you to highlight cells that are greater than, less than, equal to, between certain values, etc.
  - **Top/Bottom Rules**: Highlights cells that are in the top or bottom values within a range.
  - **Color Scales**: Applies a gradient of colors based on cell values.
  - **Icon Sets**: Applies icons (like arrows, flags, etc.) based on cell values.

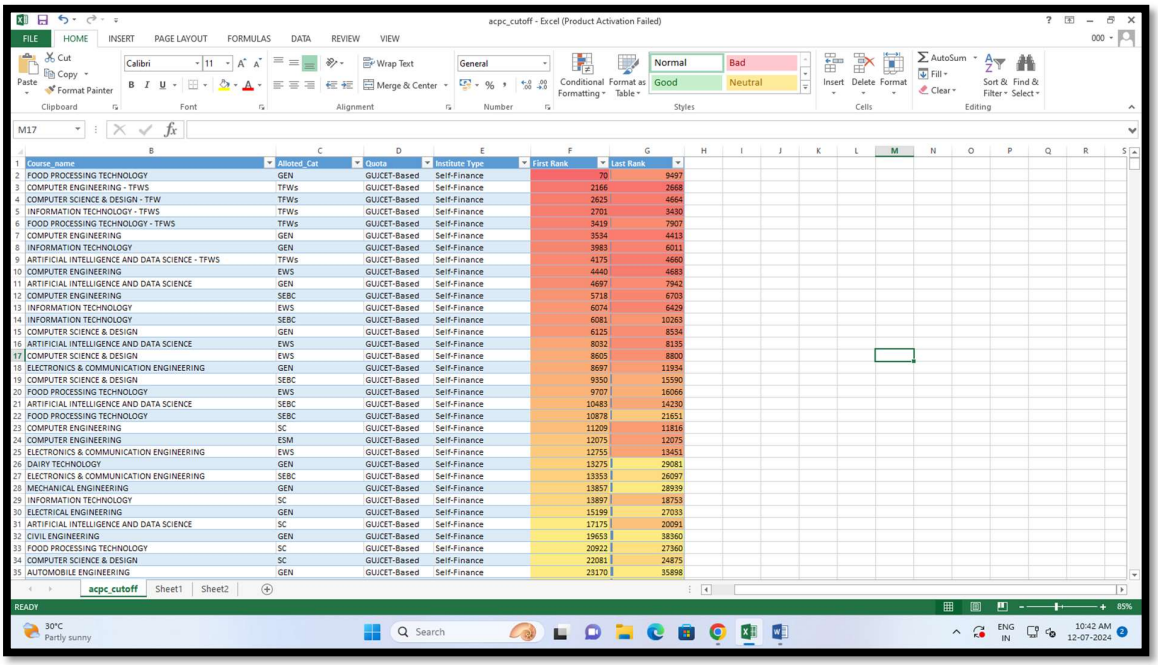
- **Set the Rule Parameters**: After choosing a rule type, a dialog box will appear where you can set the parameters for that rule. For example, if you choose "Greater Than" in Highlight Cell Rules, you'll specify the value and the formatting (like font color, fill color, etc.) for cells greater than that value.

- **Manage Rules**: You can add multiple rules to the same set of cells by repeating steps 2-4. Excel applies conditional formatting rules in the order they appear in the Conditional Formatting Rules Manager (accessible from the "Conditional Formatting" dropdown menu). You can also edit or delete existing rules from this manager.

- **Preview and Apply**: As you set up your rules, Excel provides a live preview of how the formatting will look on your selected range of cells. Once you're satisfied with the rules, click "OK" to apply them.

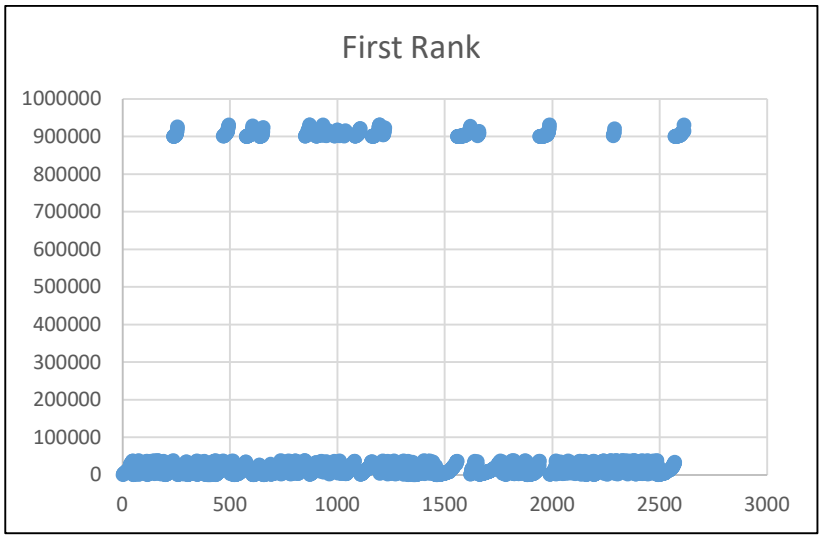
- **Modify Rules**: If you need to modify a rule after applying it, select the cells with the conditional formatting, go back to the "Conditional Formatting" dropdown, and choose "Manage Rules." From there, you can select the rule you want to edit and adjust its parameters.

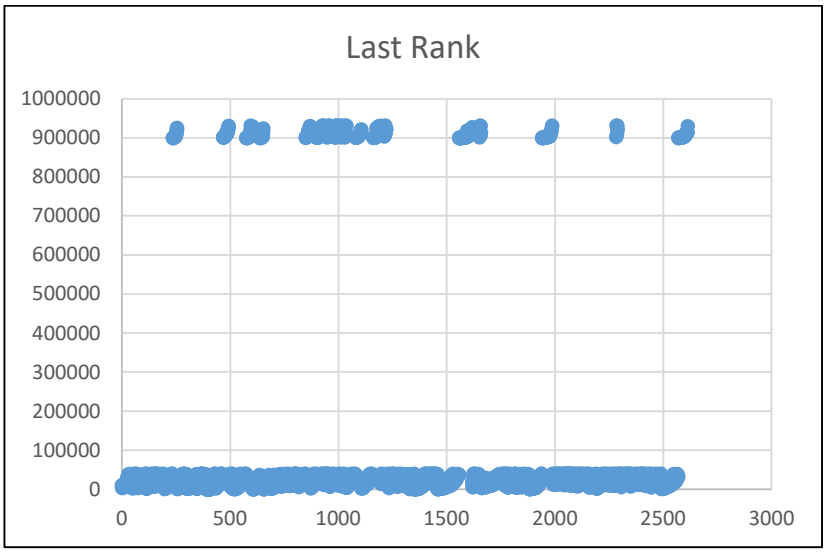
- **Clear Rules**: To remove conditional formatting from a range of cells, select the cells, go to the "Conditional Formatting" dropdown, choose "Clear Rules," and then select "Clear Rules from Selected Cells."



➢ Graphical Representation

Creating a scatter graph (or scatter plot) in Excel is a straightforward way to visualize relationships between two sets of data points.





Practical-4

Aim: Study of Exploratory Data Analysis using Excel. Task: Download any sales data for performing analysis.

To conduct Exploratory Data Analysis (EDA) using Excel on the dataset, you can follow these structured steps without referencing any specific source. Here's how you can approach the task:

Step 1: Download the Dataset

1. **Identify the Dataset**: Ensure you have access to the "Awesome Chocolates" dataset, which may be available from educational resources or data repositories.
2. **Download the Dataset**:
    - Click on the download link to obtain the dataset, typically available in CSV or Excel format.
    - Save the file to a known location on your computer for easy access.

Step 2: Load Data into Excel

1. **Open Excel**: Launch Microsoft Excel on your computer.
2. **Import the Dataset**:
    - Navigate to the **Data** tab.
    - Select **Get Data** (or **From Text/CSV** if it's a CSV file).
    - Locate the "Awesome Chocolates" file and select it.
    - Follow the prompts to import the data into a new worksheet.

Step 3: Clean the Data

1. **Check for Missing Values**:
    - Use the formula **=COUNTBLANK(range)** to identify any empty cells in the dataset.
    - Decide how to handle missing values, such as removing rows or filling them with averages.
2. **Format Data Types**:
    - Ensure that date columns are formatted correctly by right-clicking on the column and selecting **Format Cells**.
    - Convert any numerical values stored as text into proper numbers using the **VALUE()** function if necessary.
3. **Remove Duplicates**:

- Go to the **Data** tab and utilize the **Remove Duplicates** feature to clean up any duplicate entries in the dataset.

Step 4: Analyze the Data

1. **Descriptive Statistics**:
    - Calculate basic statistics using functions like **AVERAGE()**, **SUM()**, **COUNT()**, and **MAX()** to gain insights into sales performance.
2. **Create Pivot Tables**:
    - Insert a Pivot Table to summarize data by categories such as product type, month, or sales region.
    - Drag and drop fields to analyze various aspects of the data.
3. **Visualize Data**:
    - Create visual representations of the data using charts (bar, line, pie) to illustrate trends, such as sales over time or comparisons between different chocolate products.

Step 5: Interpret Findings

1. **Identify Trends**: Look for patterns in the data, such as peak sales periods, best-selling products, and customer preferences.
2. **Make Recommendations**: Based on your analysis, suggest strategies for improving sales, such as targeted marketing efforts or product promotions.

Step 6: Document Your Analysis

1. **Create a Summary Report**:
    - Compile your findings, insights, and visualizations into a summary report within Excel.
2. **Save Your Work**: Ensure all your work is saved and consider exporting your findings as a PDF for easy sharing.

- Below are the images illustrating the key steps and findings of the Exploratory Data Analysis (EDA) on the "Awesome Chocolates" dataset.

Pivot Tables

Sales Person	Geography	Product	Amou	Custome
Ram Mahesh	New Zealand	70% Dark Bites	$1,624	114
Brien Boise	USA	Choco Coated Almonds	$6,706	459
Husein Augar	USA	Almond Choco	$959	147
Carla Molina	Canada	Drinking Coco	$9,632	288
Curtice Advani	UK	White Choc	$2,100	414
Ram Mahesh	USA	Peanut Butter Cubes	$8,869	432
Curtice Advani	Australia	Smooth Sliky Salty	$2,681	54
Brien Boise	USA	After Nines	$5,012	210
Ches Bonnell	Australia	50% Dark Bites	$1,281	75
Gigi Bohling	New Zealand	50% Dark Bites	$4,991	12
Barr Faughny	UK	White Choc	$1,785	462
Gunar Cockshoot	New Zealand	Eclairs	$3,983	144
Husein Augar	Australia	Mint Chip Choco	$2,646	120
Barr Faughny	India	Milk Bars	$252	54
Gunar Cockshoot	USA	White Choc	$2,464	234
Gunar Cockshoot	USA	Manuka Honey Choco	$2,114	66
Curtice Advani	New Zealand	Smooth Sliky Salty	$7,693	87
Gigi Bohling	India	Orange Choco	$15,610	339
Carla Molina	India	After Nines	$336	144
Barr Faughny	UK	Orange Choco	$9,443	162
Husein Augar	India	Fruit & Nut Bars	$8,155	90
Brien Boise	Australia	Fruit & Nut Bars	$1,701	234
Oby Sorrel	Australia	After Nines	$2,205	141
Brien Boise	New Zealand	99% Dark & Pure	$1,771	204
Carla Molina	USA	Raspberry Choco	$2,114	186
Carla Molina	Canada	Milk Bars	$10,311	231
Gunar Cockshoot	UK	Mint Chip Choco	$21	168

Pivot | Report1 | Report2 | Report3 | Report4 | Report5 | Report6 | Report7 | Report8

Total Sales by Sales Person

Row Labels	Sum of Amount
Barr Faughny	₹ 2,70,914.00
Brien Boise	₹ 2,53,813.00
Carla Molina	₹ 2,53,078.00
Ches Bonnell	₹ 2,74,680.00
Curtice Advani	₹ 3,05,599.00
Gigi Bohling	₹ 2,94,280.00
Gunar Cockshoot	₹ 2,50,677.00
Husein Augar	₹ 2,37,412.00
Oby Sorrel	₹ 2,23,664.00
Ram Mahesh	₹ 2,71,719.00
Grand Total	₹ 26,35,836.00

Sales by Person & Geography

Sum of Amount	Column Labels						
Row Labels	Australia	Canada	India	New Zealand	UK	USA	Grand Total
Barr Faughny	42154	50267	33530	49196	73381	22386	270914
Brien Boise	47593	21525	43785	44114	50729	46067	253813
Carla Molina	17192	66920	27538	53032	17724	70672	253078
Ches Bonnell	40089	41636	72457	51142	19012	50344	274680
Curtice Advani	50897	60662	51681	40040	55881	46438	305599
Gigi Bohling	42644	67893	65044	31955	43393	43351	294280
Gunar Cockshoot	39347	57463	58359	28539	24311	42658	250677
Husein Augar	40320	50134	54110	29281	20412	43155	237412
Oby Sorrel	37226	50169	25494	38990	32613	39172	223664
Ram Mahesh	37968	48811	37513	55958	45395	46074	271719
Grand Total	395430	515480	469511	422247	382851	450317	2635836

Sales by Person with Geography Filter		
Geography	USA	
Row Labels	Sum of Amount	
Barr Faughny	₹ 22,386.00	
Brien Boise	₹ 46,067.00	
Carla Molina	₹ 70,672.00	
Ches Bonnell	₹ 50,344.00	
Curtice Advani	₹ 46,438.00	
Gigi Bohling	₹ 43,351.00	
Gunar Cockshoot	₹ 42,658.00	
Husein Augar	₹ 43,155.00	
Oby Sorrel	₹ 39,172.00	
Ram Mahesh	₹ 46,074.00	
Grand Total	₹ 4,50,317.00	

Top 10 Products with Sales - Value Filter	
Sales Person	Curtice Advani
Row Labels	Sum of Customers
Almond Choco	996
Caramel Stuffed Bars	885
Orange Choco	882
Manuka Honey Choco	858
Organic Choco Syrup	780
70% Dark Bites	771
Baker's Choco Chips	738
50% Dark Bites	708
Raspberry Choco	693
White Choc	576
Grand Total	7887

Interactive Reports with Pivots

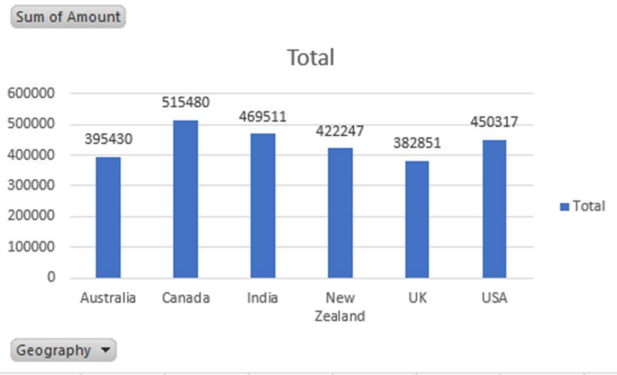
Row Labels	Sum of Amount
Barr Faughny	33530
Brien Boise	43785
Carla Molina	27538
Ches Bonnell	72457
Curtice Advani	51681
Gigi Bohling	65044
Gunar Cockshoot	58359
Husein Augar	54110
Oby Sorrel	25494
Ram Mahesh	37513
Grand Total	469511

Geography

- Australia
- Canada
- India
- New Zealand
- UK
- USA

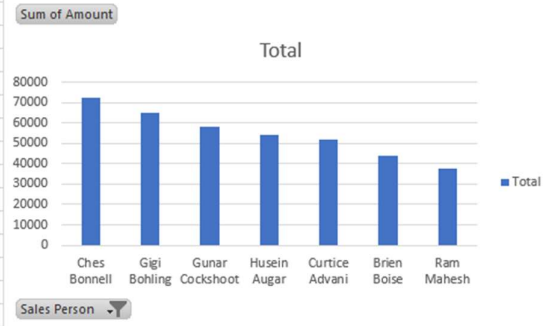
Pivot Charts

Row Labels	Sum of Amount
Australia	395430
Canada	515480
India	469511
New Zealand	422247
UK	382851
USA	450317
Grand Total	**2635836**



Interactive Pivot Charts

Row Labels	Sum of Amount
Ches Bonnell	72457
Gigi Bohling	65044
Gunar Cockshoot	58359
Husein Augar	54110
Curtice Advani	51681
Brien Boise	43785
Ram Mahesh	37513
Grand Total	**382949**



Sales as value & Percentages with bars

Row Labels	Sum of Amount	Sum of Amount2
Barr Faughny	270914	10.28%
Brien Boise	253813	9.63%
Carla Molina	253078	9.60%
Ches Bonnell	274680	10.42%
Curtice Advani	305599	11.59%
Gigi Bohling	294280	11.16%
Gunar Cockshoot	250677	9.51%
Husein Augar	237412	9.01%
Oby Sorrel	223664	8.49%
Ram Mahesh	271719	10.31%
Grand Total	**2635836**	**100.00%**

Practical-5

Aim: Learn creating dashboard using Excel. Task: Consider data from Exp.4 and prepare a dynamic dashboard.

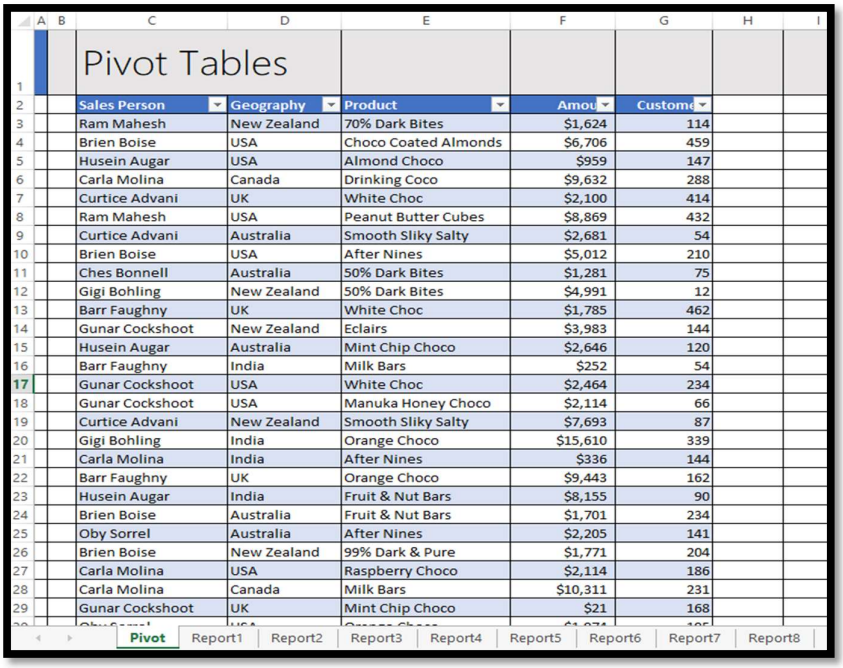
- To create a dynamic dashboard in Excel using data from your previous Exploratory Data Analysis (EDA) on the "Awesome Chocolates" dataset, follow these structured steps:

Step 1: Prepare Your Data

1. **Load the Data:** Ensure your dataset is loaded into Excel and cleaned as per the previous EDA steps.
2. **Organize Data**: Make sure your data is structured in a table format with headers for easy reference.

Step 2: Create Pivot Tables

1. **Insert a Pivot Table:**
   - Select your data range.
   - Go to the Insert tab and click on PivotTable.
   - Choose to place the Pivot Table in a new worksheet.
2. **Configure the Pivot Table:**
   - Drag and drop fields into the Rows, Columns, and Values areas to summarize your data. For example, you might want to analyze sales by product type or month.



Pivot Tables

Sales Person	Geography	Product	Amou	Custome
Ram Mahesh	New Zealand	70% Dark Bites	$1,624	114
Brien Boise	USA	Choco Coated Almonds	$6,706	459
Husein Augar	USA	Almond Choco	$959	147
Carla Molina	Canada	Drinking Coco	$9,632	288
Curtice Advani	UK	White Choc	$2,100	414
Ram Mahesh	USA	Peanut Butter Cubes	$8,869	432
Curtice Advani	Australia	Smooth Sliky Salty	$2,681	54
Brien Boise	USA	After Nines	$5,012	210
Ches Bonnell	Australia	50% Dark Bites	$1,281	75
Gigi Bohling	New Zealand	50% Dark Bites	$4,991	12
Barr Faughny	UK	White Choc	$1,785	462
Gunar Cockshoot	New Zealand	Eclairs	$3,983	144
Husein Augar	Australia	Mint Chip Choco	$2,646	120
Barr Faughny	India	Milk Bars	$252	54
Gunar Cockshoot	USA	White Choc	$2,464	234
Gunar Cockshoot	USA	Manuka Honey Choco	$2,114	66
Curtice Advani	New Zealand	Smooth Sliky Salty	$7,693	87
Gigi Bohling	India	Orange Choco	$15,610	339
Carla Molina	India	After Nines	$336	144
Barr Faughny	UK	Orange Choco	$9,443	162
Husein Augar	India	Fruit & Nut Bars	$8,155	90
Brien Boise	Australia	Fruit & Nut Bars	$1,701	234
Oby Sorrel	Australia	After Nines	$2,205	141
Brien Boise	New Zealand	99% Dark & Pure	$1,771	204
Carla Molina	USA	Raspberry Choco	$2,114	186
Carla Molina	Canada	Milk Bars	$10,311	231
Gunar Cockshoot	UK	Mint Chip Choco	$21	168

Pivot | Report1 | Report2 | Report3 | Report4 | Report5 | Report6 | Report7 | Report8

Step 3: Create Dynamic Charts

1. **Insert Charts:**
    - Select your Pivot Table.
    - Go to the Insert tab and choose a chart type (e.g., Column, Line, Pie).
    - This chart will dynamically update based on the Pivot Table.
2. **Format Charts:**
    - Customize your charts by adding titles, adjusting colors, and modifying labels to enhance readability.

	A	B	C
1			
2	Total Sales by Sales Person		
3	Row Labels	Sum of Amount	
4	Barr Faughny	₹ 2,70,914.00	
5	Brien Boise	₹ 2,53,813.00	
6	Carla Molina	₹ 2,53,078.00	
7	Ches Bonnell	₹ 2,74,680.00	
8	Curtice Advani	₹ 3,05,599.00	
9	Gigi Bohling	₹ 2,94,280.00	
10	Gunar Cockshoot	₹ 2,50,677.00	
11	Husein Augar	₹ 2,37,412.00	
12	Oby Sorrel	₹ 2,23,664.00	
13	Ram Mahesh	₹ 2,71,719.00	
14	Grand Total	₹ 26,35,836.00	

	A	B	C	D
1	Sales as value & Percentages with bars			
2				
3	Row Labels	Sum of Amount	Sum of Amount2	
4	Barr Faughny	270914	10.28%	
5	Brien Boise	253813	9.63%	
6	Carla Molina	253078	9.60%	
7	Ches Bonnell	274680	10.42%	
8	Curtice Advani	305599	11.59%	
9	Gigi Bohling	294280	11.16%	
10	Gunar Cockshoot	250677	9.51%	
11	Husein Augar	237412	9.01%	
12	Oby Sorrel	223664	8.49%	
13	Ram Mahesh	271719	10.31%	
14	Grand Total	2635836	100.00%	

Step 4: Add Slicers for Interactivity

1. **Insert Slicers:**
    - Click on your Pivot Table.
    - Go to the PivotTable Analyze tab and click on Insert Slicer.
    - Choose the fields you want to filter by (e.g., product type, region).
2. **Position and Format Slicers:**
    - Place the slicers on your dashboard for easy access.
    - Format the slicers to improve visual appeal.

Step 5: Design the Dashboard Layout

1. **Arrange Components:**
    - Organize your Pivot Tables, charts, and slicers in a logical layout on a single worksheet.
    - Ensure that the dashboard is visually appealing and easy to navigate.
2. **Add Titles and Labels:**
    - Clearly label each section of your dashboard to indicate what data is being presented.

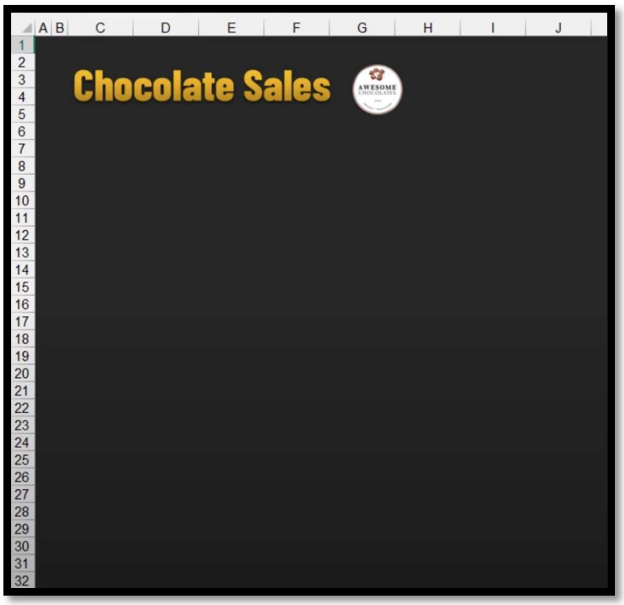
Step 6: Finalize and Test the Dashboard

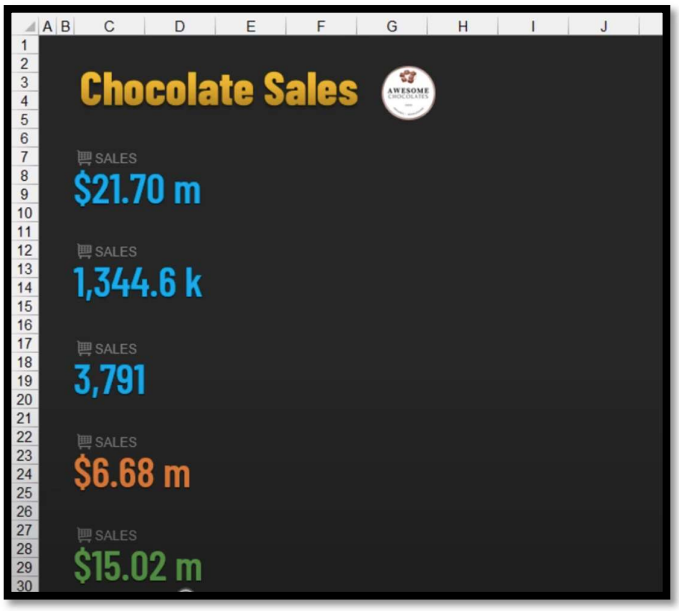
1. **Review Functionality:**
    - Test the slicers to ensure they correctly filter the data and update the charts.
    - Make adjustments as necessary to improve usability.
2. **Save Your Work:**
    - Save your Excel file to preserve your dynamic dashboard.

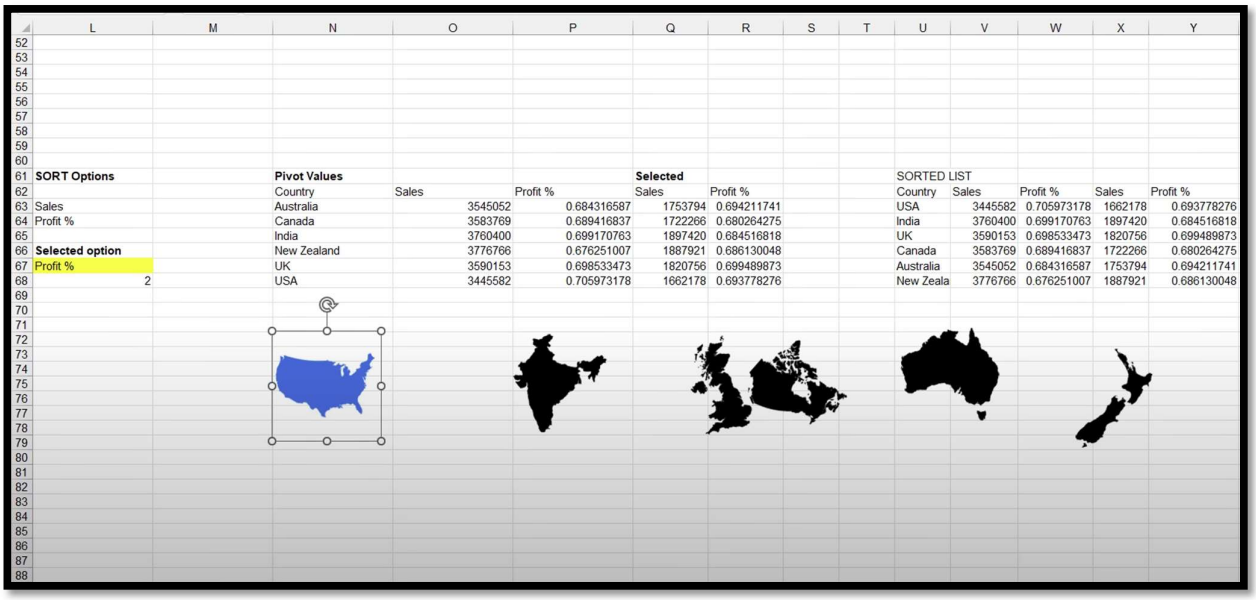
Step 7: Document Your Process

1. **Create a Summary:**
    - Write a brief summary of the dashboard's purpose and the insights it provides.
    - Consider adding instructions for users on how to interact with the dashboard.

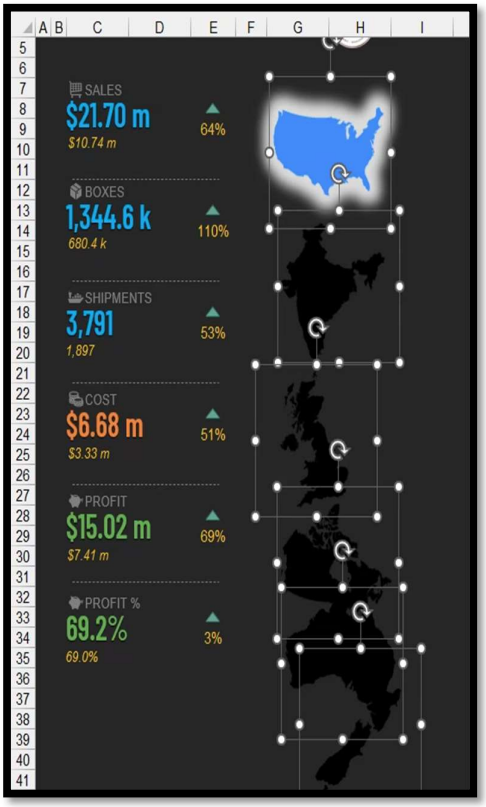
- **Below are the images highlighting the essential steps and results of the dynamic Excel dashboard created from the "Awesome Chocolates" dataset. This dashboard facilitates interactive data visualization and exploration.**







SORT Options		Pivot Values			Selected		SORTED LIST				
		Country	Sales	Profit %	Sales	Profit %	Country	Sales	Profit %	Sales	Profit %
Sales		Australia	3545052	0.684316587	1753794	0.694211741	USA	3445582	0.705973178	1662178	0.693778276
Profit %		Canada	3583769	0.689416837	1722266	0.680264275	India	3760400	0.699170763	1897420	0.684516818
		India	3760400	0.699170763	1897420	0.684516818	UK	3590153	0.698533473	1820756	0.699489873
Selected option		New Zealand	3776766	0.676251007	1887921	0.686130048	Canada	3583769	0.689416837	1722266	0.680264275
Profit %		UK	3590153	0.698533473	1820756	0.699489873	Australia	3545052	0.684316587	1753794	0.694211741
	2	USA	3445582	0.705973178	1662178	0.693778276	New Zeala	3776766	0.676251007	1887921	0.686130048


SALES
$21.70 m
$10.74 m
64%
BOXES
1,344.6 k
680.4 k
110%
SHIPMENTS
3,791
1,897
53%
COST
$6.68 m
$3.33 m
51%
PROFIT
$15.02 m
$7.41 m
69%
PROFIT %
69.2%
69.0%
3%


Chocolate Sales
SALES
$21.70 m
$10.74 m
64%
BOXES
1,344.6 k
680.4 k
110%
SHIPMENTS
3,791
1,897
53%
COST
$6.68 m
$3.33 m
51%
PROFIT
$15.02 m
$7.41 m
69%
PROFIT %
69.2%
69.0%
3%
COUNTRY SALES
USA
70.6%
69.4%
India
69.9%
68.5%
UK
69.9%
69.9%
Canada
68.9%
68.0%
Australia
68.4%
69.4%
New Zealand
67.6%
68.6%


Chocolate Sales
SALES
$21.70 m
$7.08 m
60%
BOXES
1,344.6 k
442.0 k
81%
SHIPMENTS
3,791
1,197
35%
COST
$6.68 m
$2.81 m
101%
PROFIT
$15.02 m
$4.27 m
33%
PROFIT %
69.2%
60.3%
-17%
COUNTRY PROFIT %
USA
70.6%
63.2%
India
69.9%
61.0%
UK
69.9%
61.1%
Canada
68.9%
61.3%
Australia
68.4%
57.2%
New Zealand
67.6%
58.0%
Yummies
$7.89m
$5.47m
Delish
$6.14m
$4.23m
Sales & Profits by TEAMS
Jucies
$4.24m
$2.91m
Tempo
$3.44m
$2.41m
DATA
Calculations
Assets
Dashboard

Practical-6

Aim: Study of Power BI Desktop tool. Task to be performed:

1. Download any suitable sale data from Kaggle.

2. Create dashboard on the given data using Power BI Desktop tool.

To study the Power BI Desktop tool and create a dashboard using sales data from Kaggle, follow these concise steps:

Step 1: Download Sales Data

1. **Visit Kaggle**: Go to the Kaggle website and sign in or create an account.
2. **Search for Sales Data**: Use the search bar to find suitable sales datasets. Examples include "Sales Data," "Retail Sales," or "E-commerce Sales."
3. **Download the Dataset**: Select a dataset that fits your needs and download it in CSV or Excel format.

Step 2: Load Data into Power BI

1. **Open Power BI Desktop**: Launch the Power BI Desktop application.
2. **Import Data**:
    - Click on **Get Data** in the Home tab.
    - Choose the appropriate data source (e.g., Excel or CSV).
    - Navigate to the downloaded file and load the data.

Step 3: Data Preparation

1. **Transform Data**:
    - Use the Power Query Editor to clean and transform your data (e.g., remove duplicates, change data types).
    - Ensure your data is structured correctly for analysis.

Step 4: Create Relationships (if necessary)

1. **Manage Relationships**:
    - If you have multiple tables, go to the **Model** view.
    - Create relationships between tables by dragging and dropping fields.

Step 5: Build the Dashboard

1. **Create Visualizations**:

- Use the **Report** view to add visualizations (e.g., bar charts, line graphs, pie charts).
- Drag fields onto the canvas to create visuals that represent your data insights.

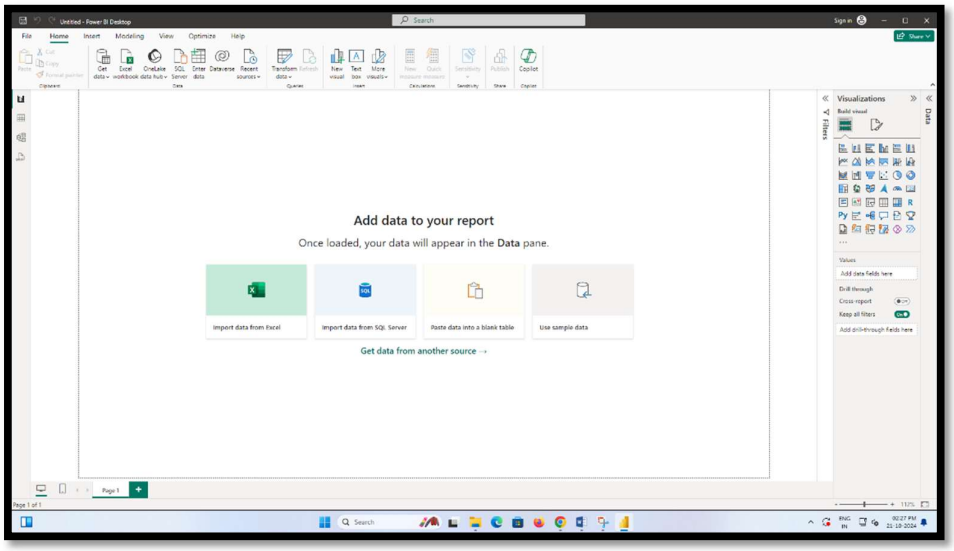
2. **Customize Visuals**:
   - Format your visuals with titles, colors, and labels for clarity.
   - Use slicers for interactivity, allowing users to filter data dynamically.

Step 6: Arrange the Dashboard

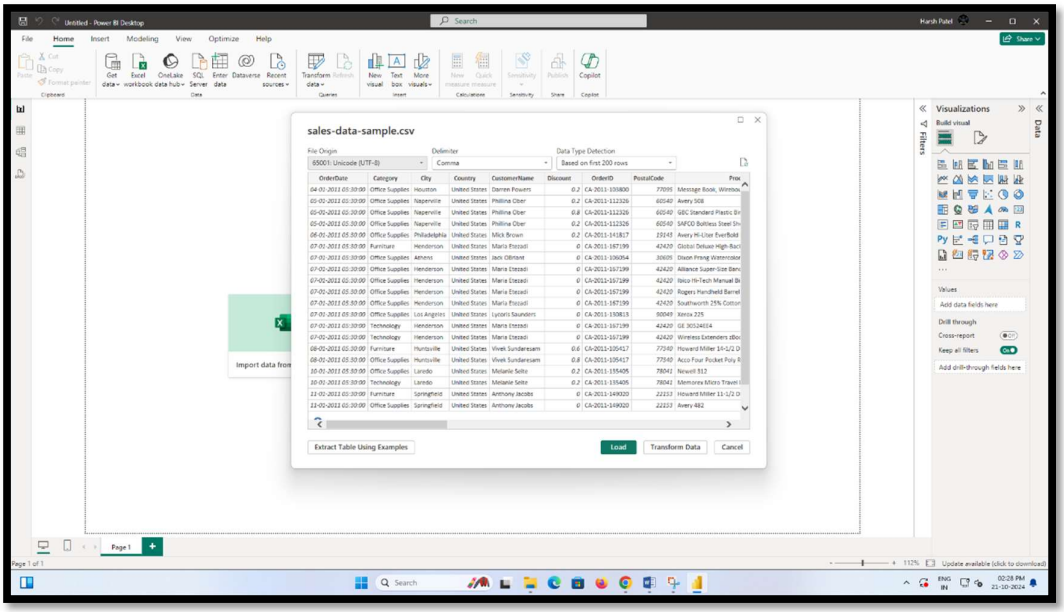
1. **Layout Design**:
   - Organize your visuals in a logical and visually appealing manner.
   - Ensure key insights are easily accessible and comprehensible.

Step 7: Publish and Share

1. **Publish Dashboard**:
   - Click on the **Publish** button to upload your dashboard to the Power BI service.
   - Share the dashboard with stakeholders or colleagues as needed.

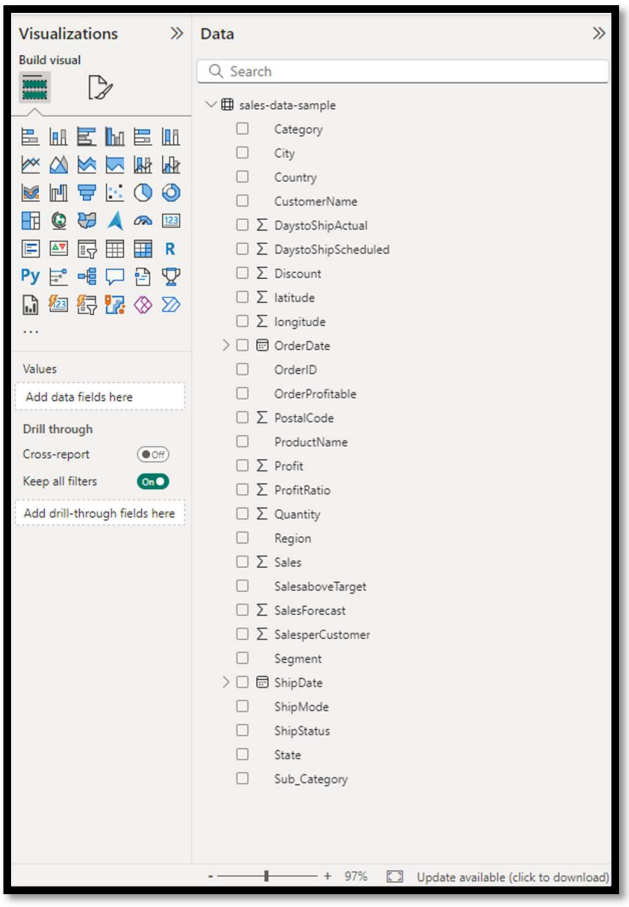


Open The Power BI Desktop App

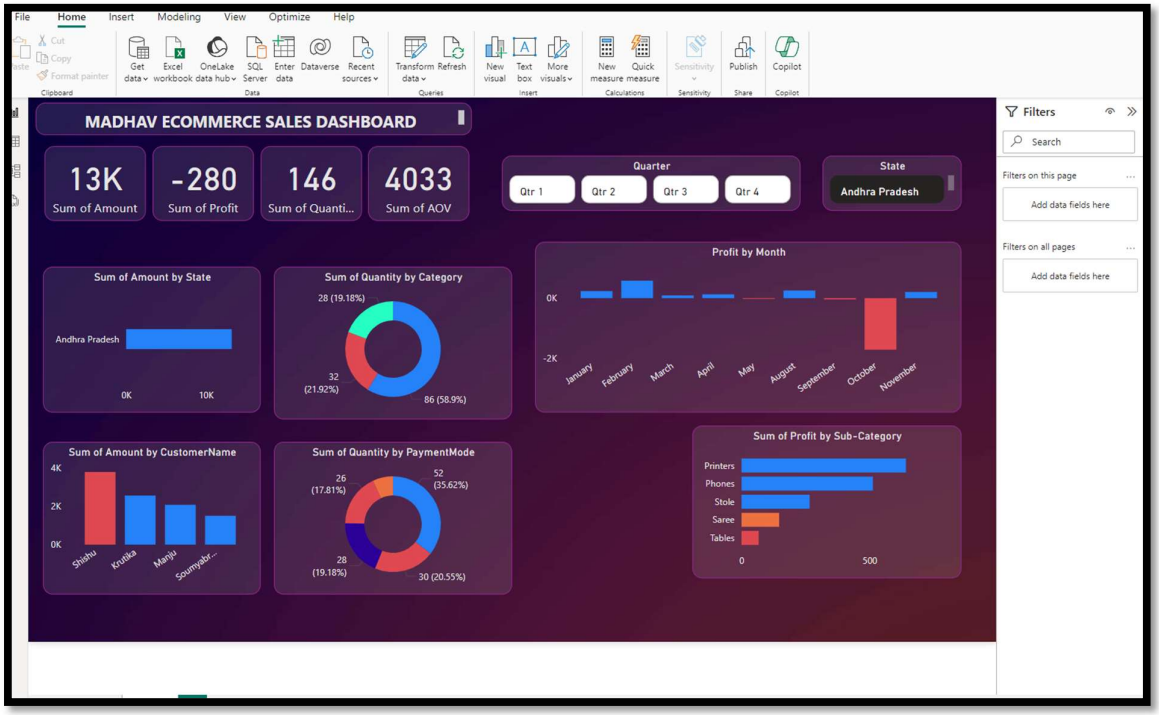


Insert The Data

Load The Data



Data Visualization


MADHAV ECOMMERCE SALES DASHBOARD
13K
Sum of Amount
-280
Sum of Profit
146
Sum of Quanti...
4033
Sum of AOV
Quarter
Qtr 1
Qtr 2
Qtr 3
Qtr 4
State
Andhra Pradesh
Sum of Amount by State
Andhra Pradesh
0K
10K
Sum of Quantity by Category
28 (19.18%)
32
(21.92%)
86 (58.9%)
Profit by Month
0K
-2K
January
February
March
April
May
August
September
October
November
Sum of Amount by CustomerName
4K
2K
0K
Shishu
Krutika
Manju
Soumyabr...
Sum of Quantity by PaymentMode
26
(17.81%)
52
(35.62%)
28
(19.18%)
30 (20.55%)
Sum of Profit by Sub-Category
Printers
Phones
Stole
Saree
Tables
0
500
Filters
Search
Filters on this page
Add data fields here
Filters on all pages
Add data fields here

Dashboard

Practical-7

Aim: Demonstrate various sampling methods on any given dataset. Tasks to be performed: 1. Show results of Random Sampling, Systematic and Stratified Sampling

Certainly! Below is a theoretical overview of the three sampling methods: Random Sampling, Systematic Sampling, and Stratified Sampling.

1. Random Sampling

Definition: Random sampling is a technique where each member of a population has an equal chance of being selected. This method helps to eliminate bias and ensures that the sample is representative of the entire population.

Advantages:

- Unbiased Representation: Since every individual has an equal chance of selection, the sample is less likely to be biased.
- Simplicity: The method is straightforward to implement and understand.

Disadvantages:

- Variability: Random samples can sometimes lead to samples that are not representative of the population, especially in smaller samples.
- Requires a Complete List: A complete list of the population is necessary to ensure true randomness.

2. Systematic Sampling

Definition: Systematic sampling involves selecting every k-th member from a list of the population. The starting point is usually chosen randomly, and then every k-th individual is selected thereafter.

Advantages:

- Ease of Use: This method is easy to implement and can be more efficient than random sampling.
- Structured Approach: It provides a structured way to sample, which can be useful in certain scenarios.

Disadvantages:

- Potential for Bias: If there is an underlying pattern in the population, systematic sampling can introduce bias. For example, if every k-th individual shares a common characteristic, the sample may not be representative.

- Less Randomness: The systematic approach may not capture the randomness that is inherent in random sampling.

3. Stratified Sampling

Definition: Stratified sampling involves dividing the population into distinct subgroups (strata) that share similar characteristics. Samples are then drawn from each stratum, either proportionally or equally.

Advantages:

- Increased Precision: By ensuring that all subgroups are represented, stratified sampling can lead to more accurate and reliable results.
- Focus on Specific Groups: This method allows researchers to focus on specific subgroups, which can be beneficial for analyses that require insights from different segments of the population.

Disadvantages:

- Complexity: Stratified sampling is more complex to administer compared to random or systematic sampling.
- Requires Knowledge of Population: Researchers need to have prior knowledge about the population to define appropriate strata.

Conclusion

Each sampling method has its strengths and weaknesses, and the choice of method depends on the research objectives, the nature of the population, and the resources available. Understanding these methods is crucial for conducting effective and reliable research.

Practical-8

Aim: Perform Hypothesis Testing through Excel.

Tasks: 1. Consider suitable dataset.

2. Perform Z-test and t-test on the dataset

Hypothesis Testing Theory

Definition: Hypothesis testing is a statistical method used to make decisions or inferences about population parameters based on sample data. It involves formulating a hypothesis, collecting data, and using statistical analysis to determine whether to accept or reject the hypothesis.

Key Concepts

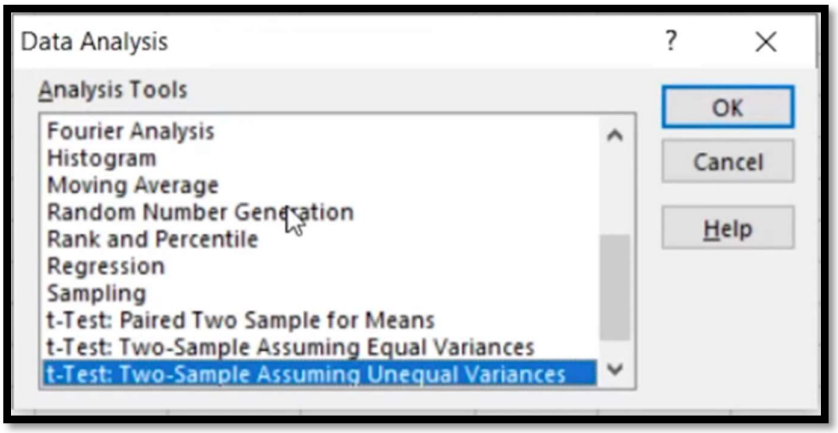
1. Hypothesis:
    - Null Hypothesis (H0): This is the statement being tested, which typically asserts that there is no effect or no difference. It serves as the default or starting assumption.
    - Alternative Hypothesis (H1 or Ha): This statement contradicts the null hypothesis and represents the effect or difference that the researcher aims to demonstrate.
2. Significance Level (α): This is the probability of rejecting the null hypothesis when it is true, commonly set at 0.05 (5%). It defines the threshold for determining whether the observed data is statistically significant.
3. Test Statistic: A standardized value calculated from sample data during a hypothesis test. It measures how far the sample statistic is from the null hypothesis, expressed in terms of standard errors.
4. P-Value: The probability of obtaining a test statistic at least as extreme as the one observed, under the assumption that the null hypothesis is true. A small p-value (typically $\leq \alpha$) indicates strong evidence against the null hypothesis.
5. Decision Rule: Based on the p-value or the test statistic, a decision is made:
    - If the p-value $\leq \alpha$, reject the null hypothesis (evidence suggests a significant effect).

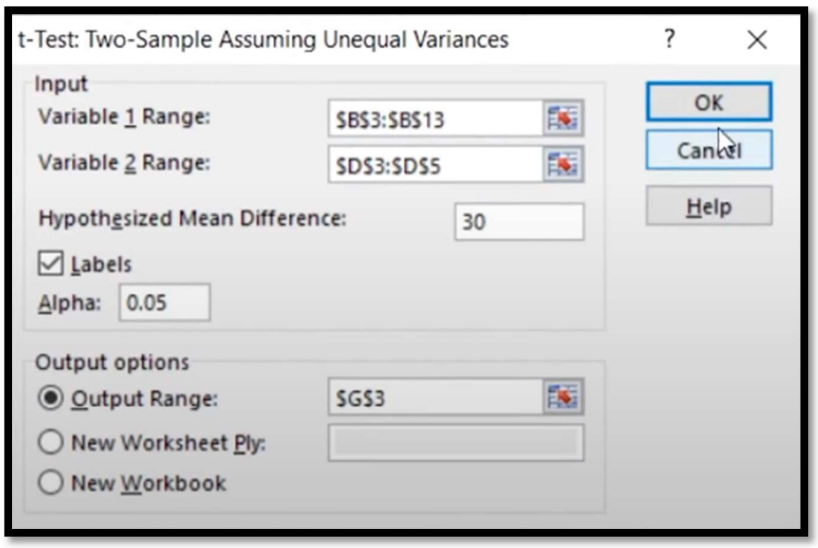
- If the p-value > α, fail to reject the null hypothesis (insufficient **evidence to suggest a significant effect).**

Types of Tests

1. Z-Test:
    - Used when the sample size is large (n > 30) or when the population variance is known.
    - Assumes that the sampling distribution of the sample mean is normally distributed.

2. T-Test:
    - Used when the sample size is small (n < 30) and the population variance is unknown.
    - There are different types of t-tests:
        - Independent t-test: Compares means from two different groups.
        - Paired t-test: Compares means from the same group at different times.
        - One-sample t-test: Compares the sample mean to a known value or population mean.





T-TEST

Hypothesis:

Ho: there is no significace difference between the mean age of male & female
H1: there is significace difference between the mean age of male & female

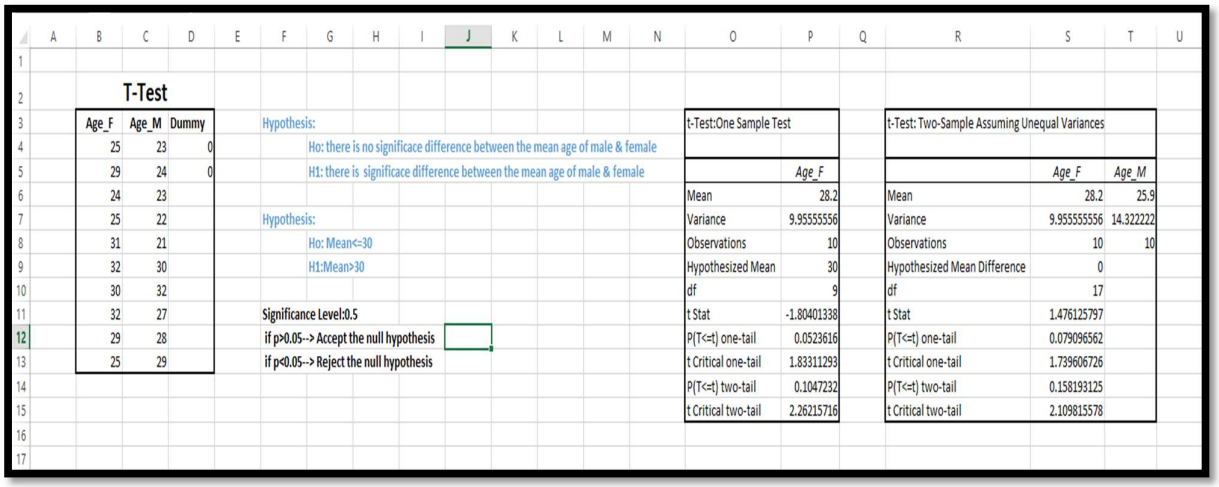
Hypothesis:

Ho: Mean<=30
H1:Mean>30

Significance Level:0.5

if p>0.05--> Accept the null hypothesis

if p<0.05--> Reject the null hypothesis



T-Test

Age_F	Age_M	Dummy
25	23	0
29	24	0
24	23	
25	22	
31	21	
32	30	
30	32	
32	27	
29	28	
25	29	

Hypothesis:
Ho: there is no significace difference between the mean age of male & female
H1: there is significace difference between the mean age of male & female

Hypothesis:
Ho: Mean<=30
H1:Mean>30

Significance Level:0.5
if p>0.05--> Accept the null hypothesis
if p<0.05--> Reject the null hypothesis

t-Test:One Sample Test

	Age_F
Mean	28.2
Variance	9.95555556
Observations	10
Hypothesized Mean	30
df	9
t Stat	-1.80401338
P(T<=t) one-tail	0.0523616
t Critical one-tail	1.83311293
P(T<=t) two-tail	0.1047232
t Critical two-tail	2.26215716

t-Test: Two-Sample Assuming Unequal Variances

	Age_F	Age_M
Mean	28.2	25.9
Variance	9.955555556	14.322222
Observations	10	10
Hypothesized Mean Difference	0	
df	17	
t Stat	1.476125797	
P(T<=t) one-tail	0.079096562	
t Critical one-tail	1.739606726	
P(T<=t) two-tail	0.158193125	
t Critical two-tail	2.109815578	

OUTPUT

Practical-9

Aim: Perform One way & Two way ANOVA through Excel. Consider suitable dataset.

To perform One-way and Two-way ANOVA in Excel, you can follow these structured steps. Below, I will outline the procedures along with a suitable example dataset.

Let's consider a dataset that examines the effect of different fertilizers on plant growth. The dataset includes three types of fertilizers and their corresponding plant growth measurements (in cm) over a fixed period.

One-Way ANOVA in Excel

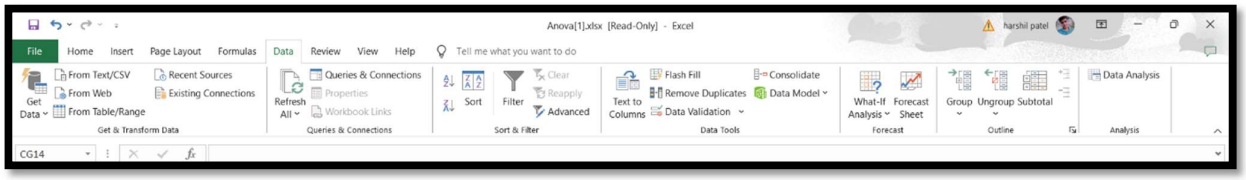
One-way ANOVA is used to compare means among three or more groups based on one independent variable.

Steps to Perform One-Way ANOVA:

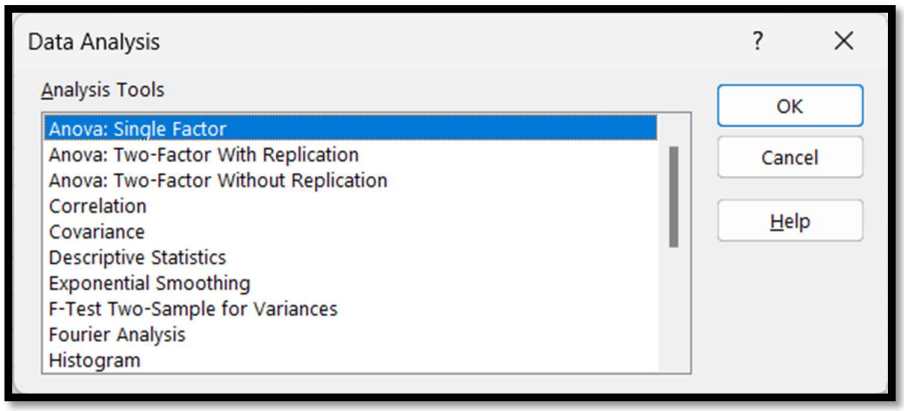
1. Input Data: Enter your data in Excel in a format similar to the example above.
2. Data Analysis Tool:
    - Go to the Data tab.
    - Click on Data Analysis (If you don't see this option, you may need to enable the Analysis ToolPak add-in).
3. Select ANOVA:
    - Choose ANOVA: Single Factor from the list and click OK.
4. Input Range:
    - Select the range of your dataset (including headers).
    - Specify the grouping option (Columns or Rows) based on how your data is organized.
5. Output Options:
    - Choose where you want the output to be displayed (New Worksheet or Existing Worksheet).
6. Run Analysis: Click OK to run the analysis.
7. Interpret Results:
    - Look for the F-statistic and p-value in the output. If the p-value is less than your significance level (commonly 0.05), you reject the null hypothesis, indicating that there are significant differences among group means.

Group 1	Group2	Group3
48	49	47
51	55	51
49	53	55
45	54	47
53	48	57
46	47	52
51	54	48
45	55	52
53	48	48
53	45	52
50	47	47
45	46	51
48	48	58
45	51	47
45	50	49

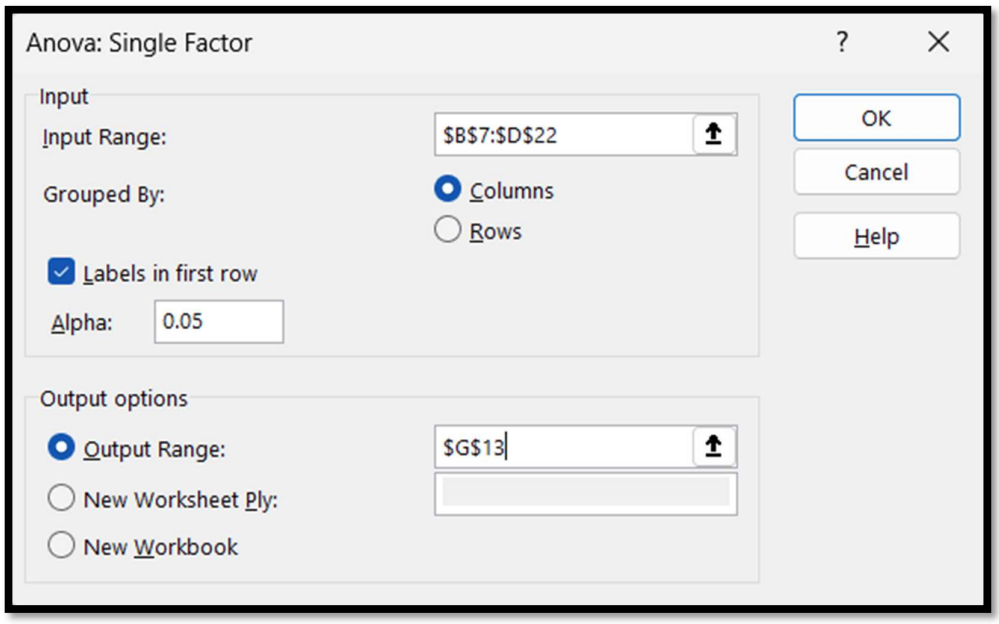
DATASET



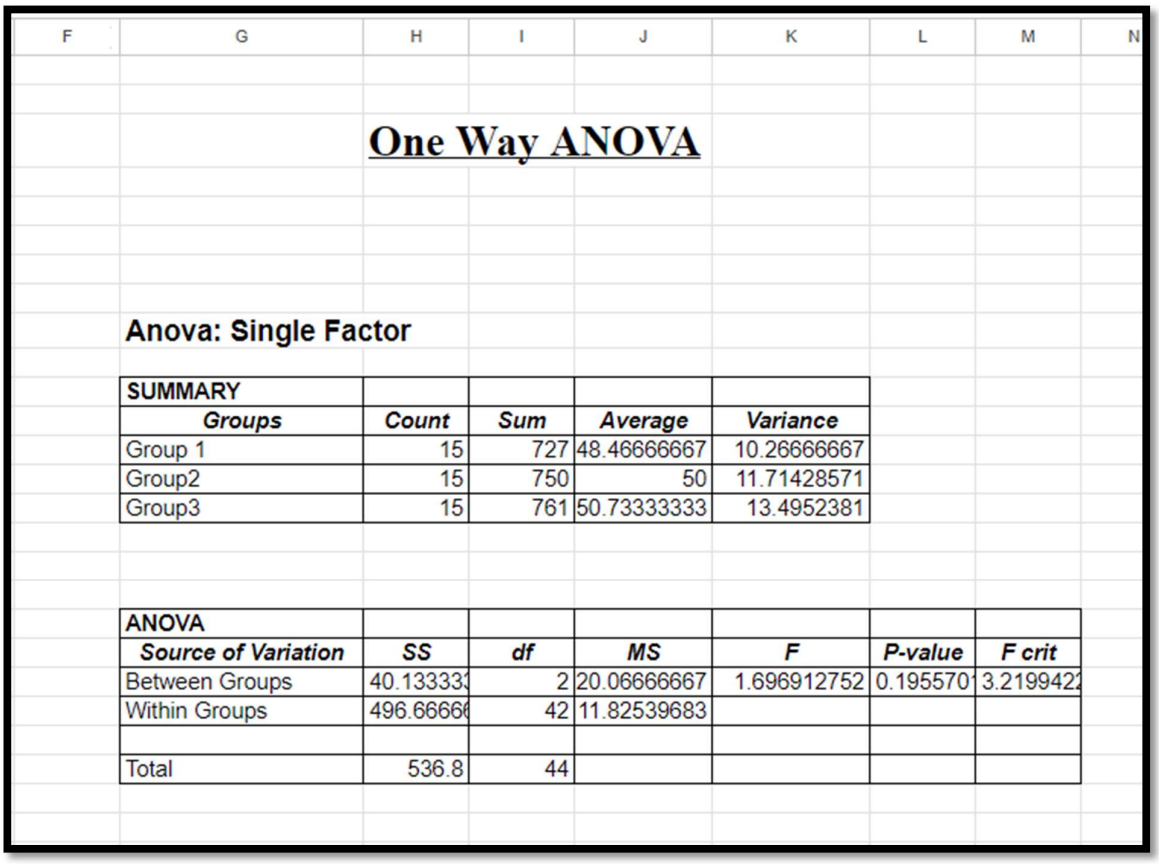
Excel's data analysis tools are a powerhouse for interpreting data. From PivotTables to Data Analysis Toolpak, they simplify complex data into actionable insights.



Excel's ANOVA Single Factor tool to analyze variance across multiple groups, testing hypotheses and determining significance levels to identify meaningful differences between group means



Input and output range of single factor Anova with alpha value 0.05



One Way ANOVA

Anova: Single Factor

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Group 1	15	727	48.46666667	10.26666667
Group2	15	750	50	11.71428571
Group3	15	761	50.73333333	13.4952381

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	40.13333	2	20.06666667	1.696912752	0.195570	3.219942
Within Groups	496.6666	42	11.82539683			
Total	536.8	44				

OUTPUT

Two-Way ANOVA in Excel

Two-way ANOVA is used when you have two independent variables and want to understand their effect on a dependent variable.

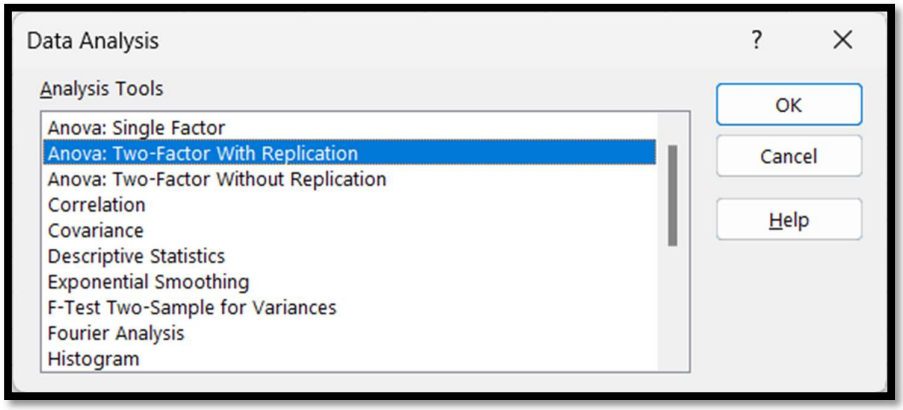
Let's expand our dataset to include two factors: Fertilizer Type and Watering Frequency (Low, Medium, High).

Steps to Perform Two-Way ANOVA:

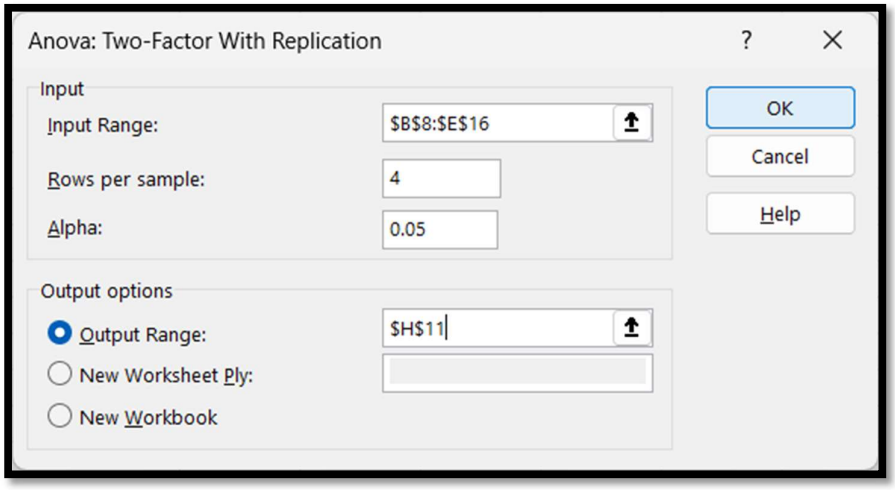
1. Input Data: Enter your data in Excel in a format similar to the example above.
2. Data Analysis Tool:
    - Go to the Data tab.
    - Click on Data Analysis.
3. Select ANOVA:
    - Choose ANOVA: Two-Factor With Replication from the list and click OK.
4. Input Range:
    - Select the range of your dataset (including headers).
    - Specify the number of rows per sample (the number of observations for each combination of factors).
5. Output Options:
    - Choose where you want the output to be displayed (New Worksheet or Existing Worksheet).
6. Run Analysis: Click OK to run the analysis.
7. Interpret Results:
    - Check the output for the F-statistics and p-values for each factor and their interaction. If any p-value is less than 0.05, you can conclude that there is a significant effect.

	Cold	Warm	Hot
Super	4	7	10
	5	9	12
	6	8	11
	5	12	9
Best	6	13	12
	6	15	13
	4	12	10
	4	12	13

DATSET



Excel's two-factor ANOVA with replication analyzes two factors' impact on continuous data, accounting for repeated measures. Evaluate factor interactions, main effects, and within-group variability.



Two factor Anova with input and output range with alpha value 0.05

Two Way ANOVA

Anova: Two-Factor With Replication

SUMMARY	Cold	Warm	Hot	Total
Super				
Count	4	4	4	12
Sum	20	36	42	98
Average	5	9	10.5	8.166666
Variance	0.666666	4.666666	1.666666	7.787878
Best				
Count	4	4	4	12
Sum	20	52	48	120
Average	5	13	12	10
Variance	1.333333	2	2	15.27272
Total				
Count	8	8	8	
Sum	40	88	90	
Average	5	11	11.25	
Variance	0.857142	7.428571	2.214285	

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	20.16666	1	20.16666	9.810810	0.005758	4.413873
Columns	200.3333	2	100.1666	48.72972	0.000000	3.554557
Interaction	16.33333	2	8.166666	3.972972	0.037224	3.554557
Within	37	18	2.055555			
Total	273.8333	23				

OUTPUT

Practical-10

Aim: Perform Regression Analysis through Excel or Python .

Tasks:

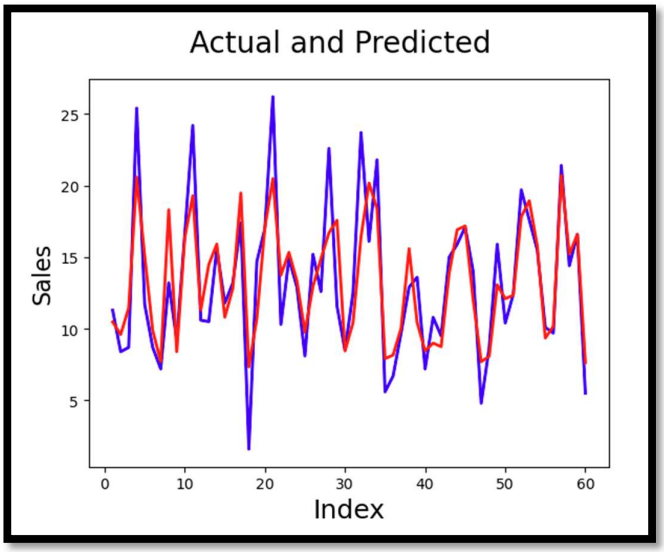
1. Perform Simple linear regression on suitable dataset from Kaggle or UCI.

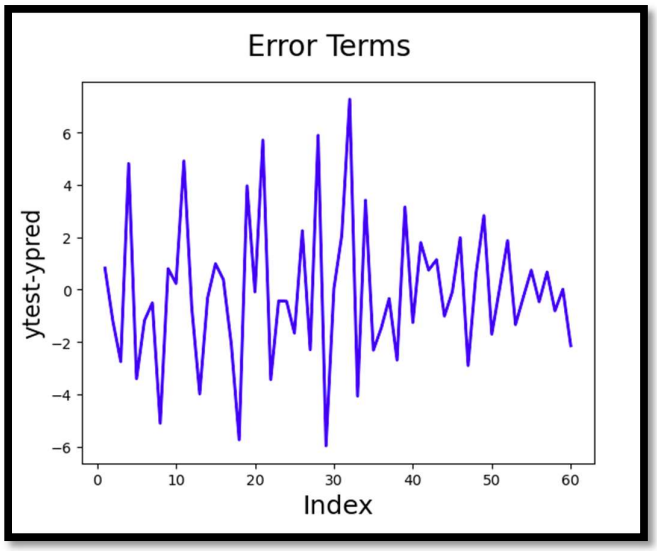
- PROGRAM:

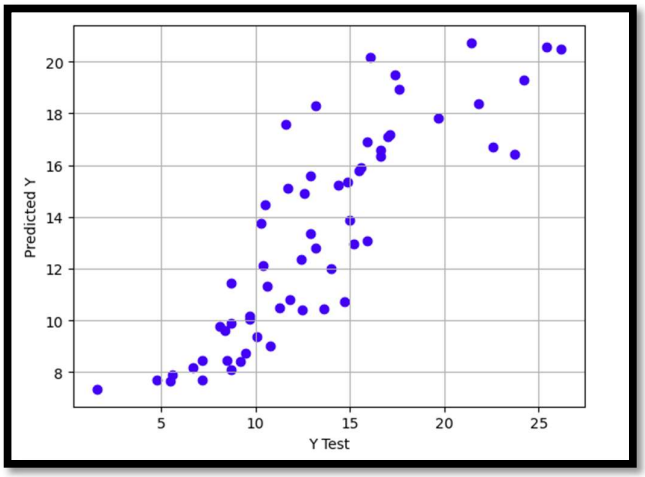
```
import pandas as pd

# Reading csv file from github repo
url = "https://raw.githubusercontent.com/devzohaib/Simple-Linear-Regression/master/tvmarketing.csv"
advertising = pd.read_csv(url)

# Putting feature variable to X
X = advertising['TV']

# Putting response variable to y
y = advertising['Sales']

#random_state is the seed used by the random number generator, it can be any integer.

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7 , random_state=0000)

# import LinearRegression from sklearn
from sklearn.linear_model import LinearRegression

# Representing LinearRegression as lr(Creating LinearRegression Object)
lr = LinearRegression()

# Fit the model using lr.fit()
lr.fit(X_train, y_train)

# Print the intercept and coefficients
print(lr.intercept_)
```

```
print(lr.coef_)

# Making predictions on the testing set
y_pred = lr.predict(X_test)

# Actual vs Predicted
import matplotlib.pyplot as plt
c = [i for i in range(1,61,1)]         # generating index
fig = plt.figure()
plt.plot(c,y_test, color="blue", linewidth=2, linestyle="-")
plt.plot(c,y_pred, color="red",  linewidth=2, linestyle="-")
fig.suptitle('Actual and Predicted', fontsize=20)          # Plot heading
plt.xlabel('Index', fontsize=18)                           # X-label
plt.ylabel('Sales', fontsize=16)                           # Y-label


# Error terms
c = [i for i in range(1,61,1)]
fig = plt.figure()
plt.plot(c,y_test-y_pred, color="blue", linewidth=2, linestyle="-")
fig.suptitle('Error Terms', fontsize=20)              # Plot heading
plt.xlabel('Index', fontsize=18)                      # X-label
plt.ylabel('ytest-ypred', fontsize=16)                # Y-label

from sklearn.metrics import mean_squared_error, r2_score
mse = mean_squared_error(y_test, y_pred)

r_squared = r2_score(y_test, y_pred)

print('Mean_Squared_Error :' ,mse)
print('r_square_value :',r_squared)

import matplotlib.pyplot as plt
plt.scatter(y_test,y_pred,c='blue')
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
plt.grid()
```

➢ Output:


Actual and Predicted
Sales
Index
25
20
15
10
5
0
10
20
30
40
50
60


Error Terms
ytest-ypred
Index
6
4
2
0
−2
−4
−6
0
10
20
30
40
50
60


Predicted Y
Y Test
20
18
16
14
12
10
8
5
10
15
20
25

2. Perform Multiple linear regression on suitable dataset from Kaggle or UCI.

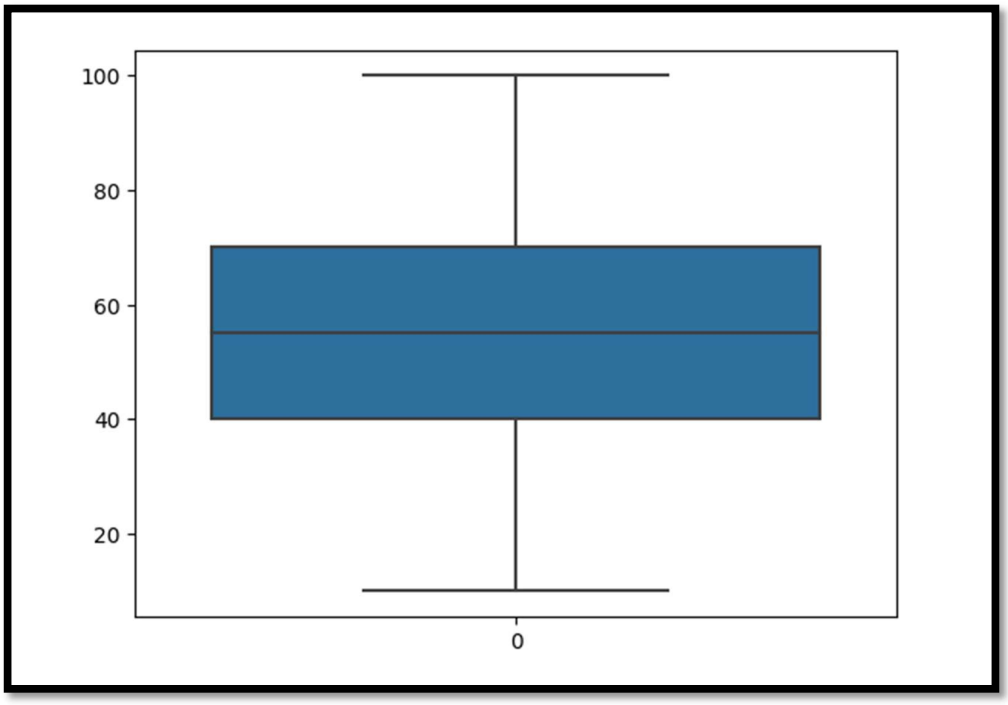
- **Program:**

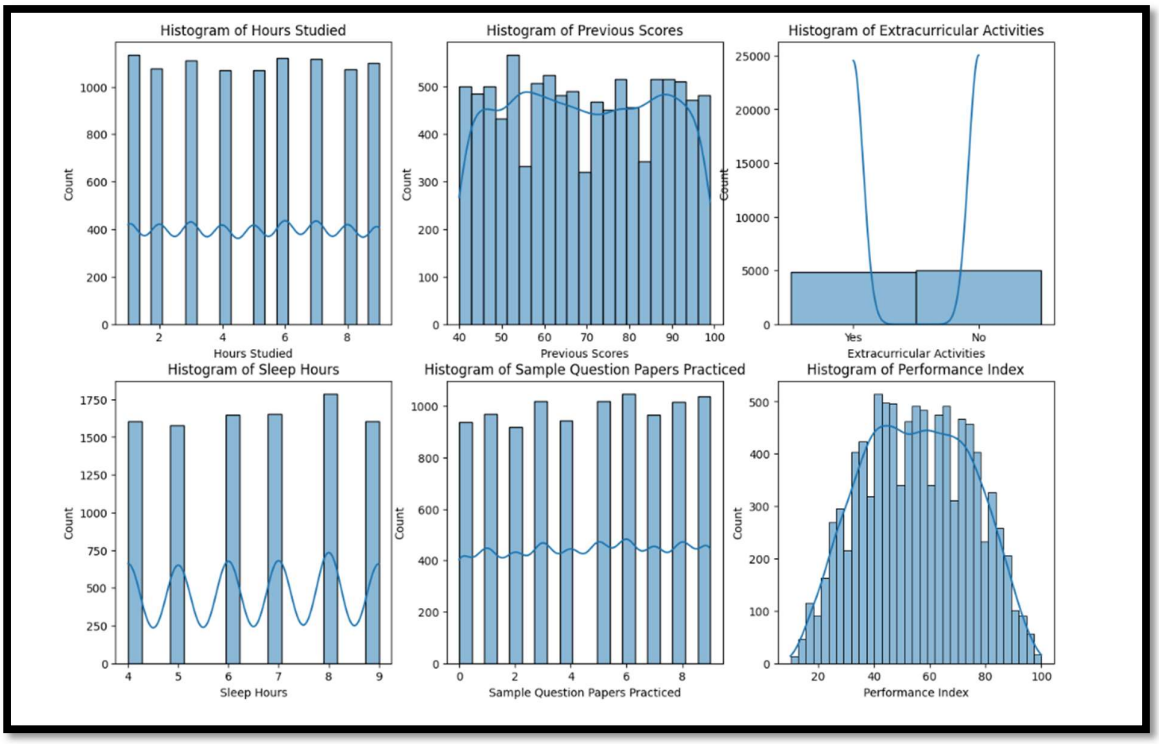
```
# Import necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

# Check the files in the Kaggle input directory (if using Kaggle environment)
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# Load the dataset
df=pd.read_csv('/kaggle/input/student-performance-multiple-linear-
regression/Student_Performance.csv')

# Show the first few rows, descriptive statistics, and info about the dataset
print(df.head())
print(df.describe())
print(df.info())
print(f'Duplicated rows: {df.duplicated().sum()}')

# Drop missing and duplicate values
df.dropna(inplace=True)
df.drop_duplicates(inplace=True)

# Visualize the Performance Index distribution using a boxplot
sns.boxplot(df['Performance Index'])
plt.show()

# Define the numerical variables for histogram plotting
num_variables = ['Hours Studied', 'Previous Scores', 'Extracurricular Activities',
                 'Sleep Hours', 'Sample Question Papers Practiced', 'Performance Index']

plt.figure(figsize=(15,10))
for i, var in enumerate(num_variables):
    plt.subplot(2, 3, i + 1)
    sns.histplot(df[var], kde=True)
    plt.title(f'Histogram of {var}')
plt.tight_layout()
plt.show()
```

```
# Convert 'Extracurricular Activities' categorical data into numerical using get_dummies
extracurricular_activitie s= pd.get_dummies(df['Extracurricular Activities'], drop_first=True, dtype=int)
df['extracurricular_activities'] = extracurricular_activities['Yes']
df.drop('Extracurricular Activities', axis=1, inplace=True)

# Import necessary libraries for model creation and evaluation
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Define features (X) and target variable (y)
X = df.drop('Performance Index', axis=1)
y = df['Performance Index']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

# Create and train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Output model coefficients and intercept
coef = model.coef_
intercept = model.intercept_
print('Regression Line coefficients are:', coef)
print('Regression Line intercept is:', intercept)

# Compare actual vs predicted values
compare = pd.DataFrame({'y_actual': y_test, 'y_predicted': y_pred})
print(compare.head())

# Calculate and print evaluation metrics
r_square = r2_score(y_test, y_pred)
print(f'R_Square is: {round(r_square * 100, 1)}')
print(f'Mean Squared Error is: {round(mean_squared_error(y_test, y_pred), 2)}')
print(f'Mean Absolute Error is: {round(mean_absolute_error(y_test, y_pred), 2)}')

# Calculate Adjusted R-Square
N = len(y_test)
P = X_test.shape[1]
adj_rsquare = 1 - ((1 - r_square) * (N - 1) / (N - P - 1))
```

```
print(f"Adjusted R_Square is: {adj_rsquare}")
print(f"Difference between R_Square and Adjusted R_Square is: {r_square - adj_rsquare}")
```

- Output:





Practical-11

Aim: Time series forecasting using Excel or Python

Task:

1. Explore methods for time series analysis(Moving Average, Exponential smoothing AR, ARMA, ARIMA, etc.)

2. Demonstrate time series forecasting using any one of the method using either excel or python.

Time Series Forecasting Theory

Time series forecasting is a statistical technique used to predict future values based on previously observed values. It involves analyzing historical data points collected at consistent time intervals to identify patterns, trends, and seasonal variations.

Key Concepts

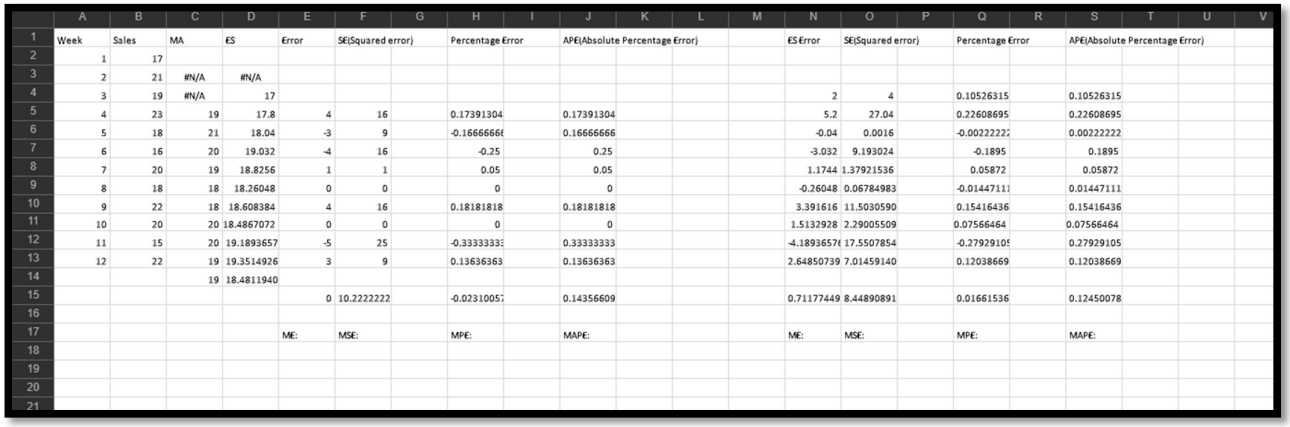
1. **Time Series Data**:
    - A sequence of data points indexed in time order, often collected at regular intervals (e.g., daily, monthly, yearly).
    - Examples include stock prices, weather data, sales figures, and economic indicators.
2. **Components of Time Series**:
    - **Trend**: The long-term movement in the data, indicating a general direction (upward or downward).
    - **Seasonality**: Regular, periodic fluctuations in the data that occur at specific intervals (e.g., increased sales during holidays).
    - **Cyclic Patterns**: Long-term fluctuations that are not fixed in length, often influenced by economic or business cycles.
    - **Irregular/Random Variations**: Unpredictable fluctuations caused by unforeseen events.
3. **Stationarity**:
    - A time series is stationary if its statistical properties (mean, variance, autocorrelation) do not change over time.
    - Many forecasting methods, such as ARIMA, require stationary data. Techniques like differencing or transformation are used to achieve stationarity.

Common Forecasting Methods

1. **Moving Average (MA)**:
    - Averages a fixed number of past observations to smooth out short-term fluctuations.
    - Simple Moving Average (SMA) gives equal weight to all observations, while Weighted Moving Average (WMA) assigns different weights.
2. **Exponential Smoothing**:
    - A forecasting method that applies decreasing weights to older observations, allowing more recent data to have a greater influence.
    - Variants include:
        - **Simple Exponential Smoothing**: For data without trend or seasonality.
        - **Holt's Linear Trend Method**: For data with a linear trend.
        - **Holt-Winters Seasonal Method**: For data with both trend and seasonality.
3. **Autoregressive (AR) Models**:
    - Models that express the current value of a time series as a linear combination of its past values.
    - Suitable for stationary data.
4. **Autoregressive Moving Average (ARMA)**:
    - Combines AR and MA models to capture both autoregressive and moving average components.
    - Effective for stationary time series data.
5. **Autoregressive Integrated Moving Average (ARIMA)**:
    - An extension of ARMA that includes differencing to make the data stationary.
    - ARIMA models are characterized by three parameters: (p, d, q), where:
        - **p**: Number of lag observations included in the model (AR part).
        - **d**: Number of times that the raw observations are differenced (I part).
        - **q**: Size of the moving average window (MA part).
6. **Seasonal Decomposition of Time Series (STL)**:
    - Decomposes a time series into seasonal, trend, and residual components.
    - Useful for understanding and visualizing the underlying patterns in the data.

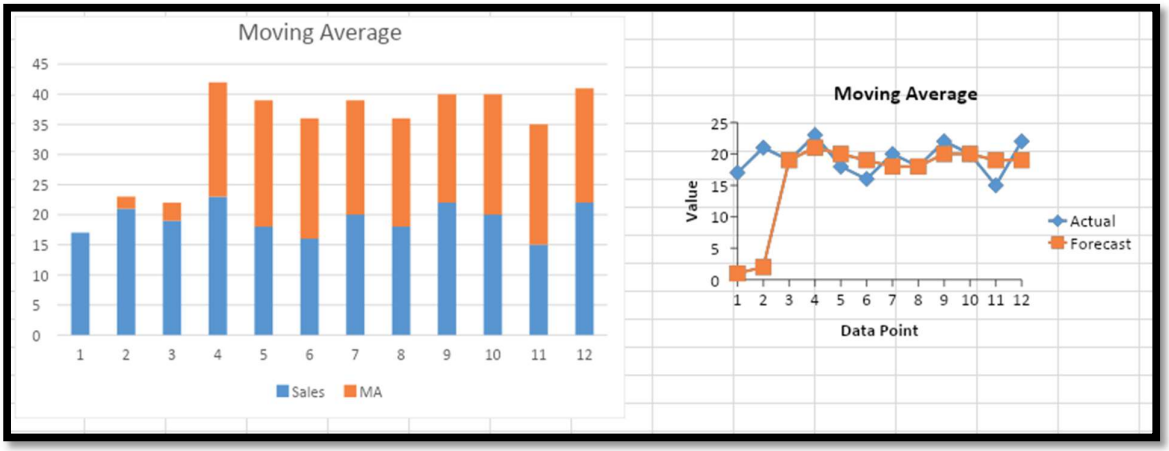
Evaluation Metrics

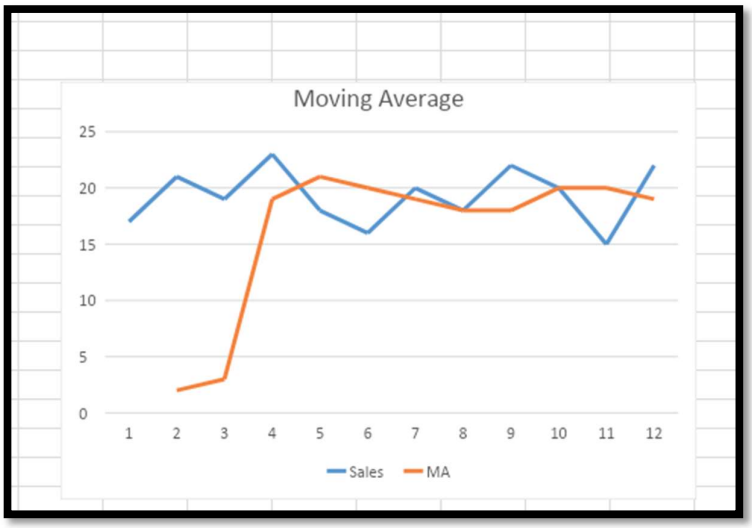
- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in a set of predictions, without considering their direction.
- **Mean Squared Error (MSE)**: Measures the average of the squares of the errors, giving more weight to larger errors.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, providing an error metric in the same units as the data.
- **Mean Absolute Percentage Error (MAPE)**: Expresses the accuracy of a forecasting method as a percentage, making it easier to interpret.



Week	Sales	MA	ES	Error	SE(Squared error)	Percentage Error	APE(Absolute Percentage Error)	ES Error	SE(Squared error)	Percentage Error	APE(Absolute Percentage Error)
1	17										
2	21	#N/A	#N/A								
3	19	#N/A	17					2	4	0.10526315	0.10526315
4	23	19	17.8	4	16	0.17391304	0.17391304	5.2	27.04	0.22608695	0.22608695
5	18	21	18.04	-3	9	-0.1666666	0.16666666	-0.04	0.0016	-0.0022222	0.00222222
6	16	20	19.032	-4	16	-0.25	0.25	-3.032	9.193024	-0.1895	0.1895
7	20	19	18.8256	1	1	0.05	0.05	1.1744	1.37921536	0.05872	0.05872
8	18	18	18.26048	0	0	0	0	-0.26048	0.06784983	-0.0144711	0.01447111
9	22	18	18.608384	4	16	0.18181818	0.18181818	3.391616	11.5030590	0.15416436	0.15416436
10	20	20	18.4867072	0	0	0	0	1.5132928	2.29005509	0.07566464	0.07566464
11	15	20	19.1893657	-5	25	-0.3333333	0.33333333	-4.1893657	17.5507854	-0.2792910	0.27929105
12	22	19	19.3514926	3	9	0.13636363	0.13636363	2.64850739	7.01459140	0.12038669	0.12038669
		19	18.4811940								
				0	10.2222222	-0.0231005	0.14356609	0.71177449	8.44890891	0.01661536	0.12450078
				ME:	MSE:	MPE:	MAPE:	ME:	MSE:	MPE:	MAPE:

DATASET FOR MOVING AVERAGE





OUTPUT FOR MOVING AVERAGE