Business Intelligence and Data Analytics Journal
(2024-2025)

VidyaVikas Education Society's

VIKAS COLLEGE OF ARTS, SCIENCE & COMMERCE

Affiliated to University of Mumbai
RE-ACCREDITED 'A' GRADE BY NAAC
ISO 9001 : 2008 CERTIFIED

Vikas High School Marg, Kannamwar Nagar No 2, Vikhroli (E), Mumbai – 400083

Dr. R. K. Patra
Principal

Hon' ble: **Shri P. M. Raut**
Chairman. V. V. Edu. Society

This is to certify that, _____

student of T.Y.B.Sc. (Information Technology) (Semester-V) with college enrolled Roll no._____ / University Seat _____ has satisfactorily completed the Project Dissertation work for the Subject Software Project Development in the program of Information Technology from the UNIVERSITY OF MUMBAI for the academic year 2024-2025.

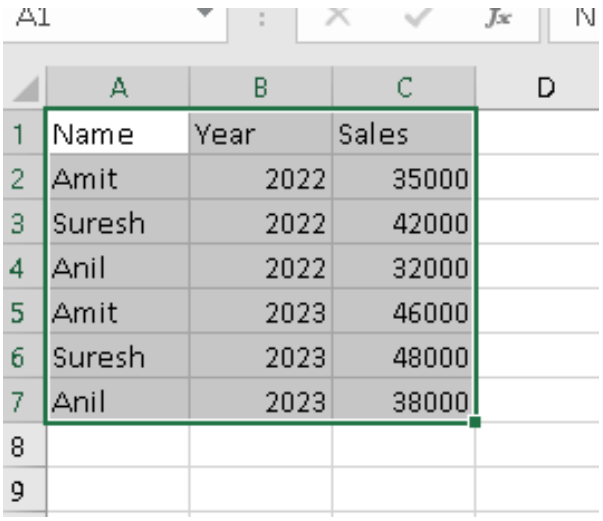
Guided By College Seal Head Of Department

External Examiner

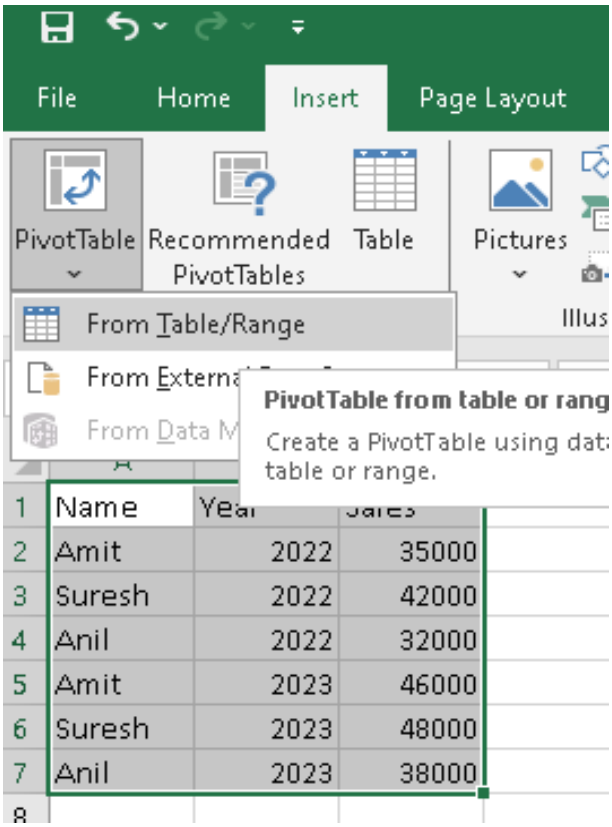
Index

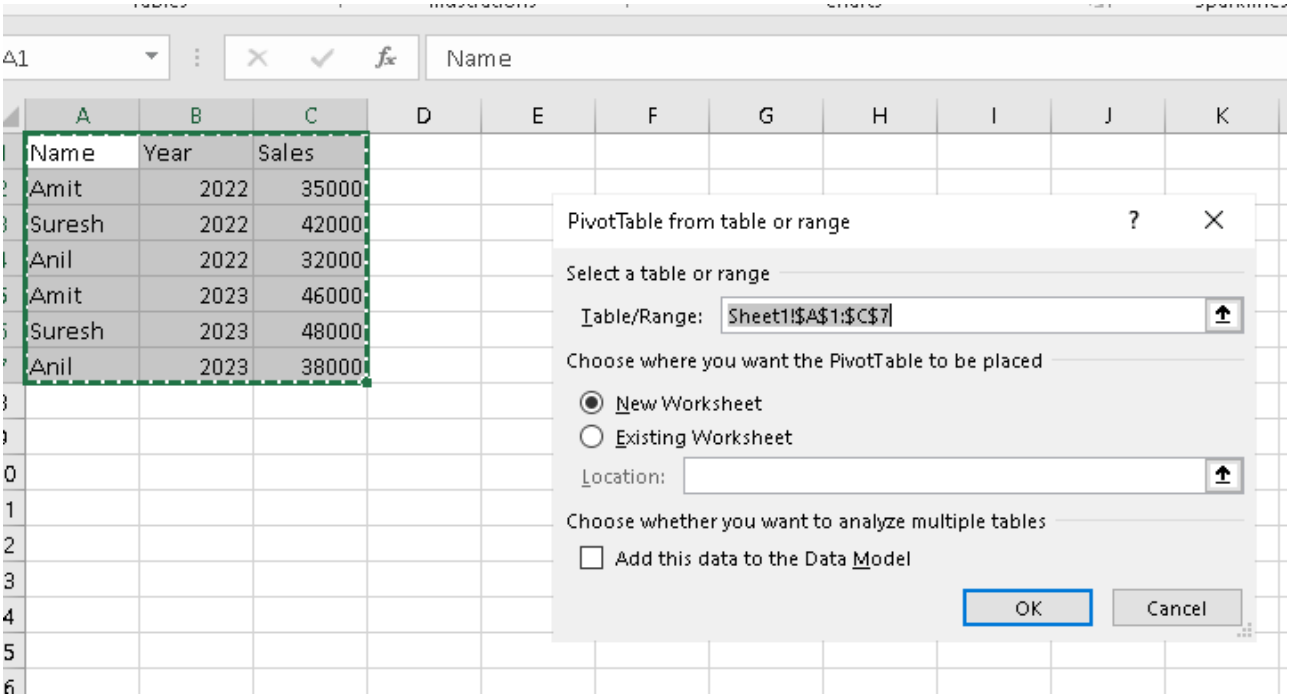
Practical: 1a

Aim: Import the data warehouse data in Microsoft Excel and create the pivot table and pivot chart.

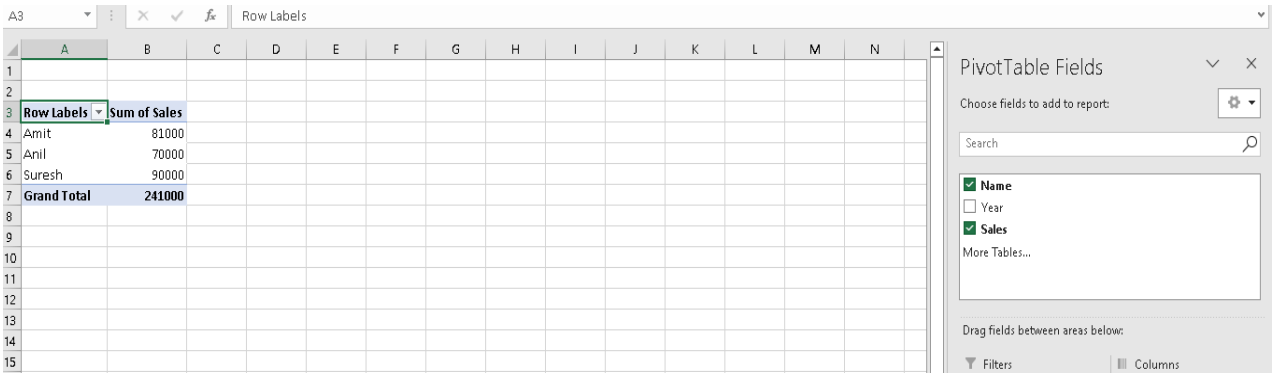


	A	B	C	D
1	Name	Year	Sales	
2	Amit	2022	35000	
3	Suresh	2022	42000	
4	Anil	2022	32000	
5	Amit	2023	46000	
6	Suresh	2023	48000	
7	Anil	2023	38000	
8				
9				

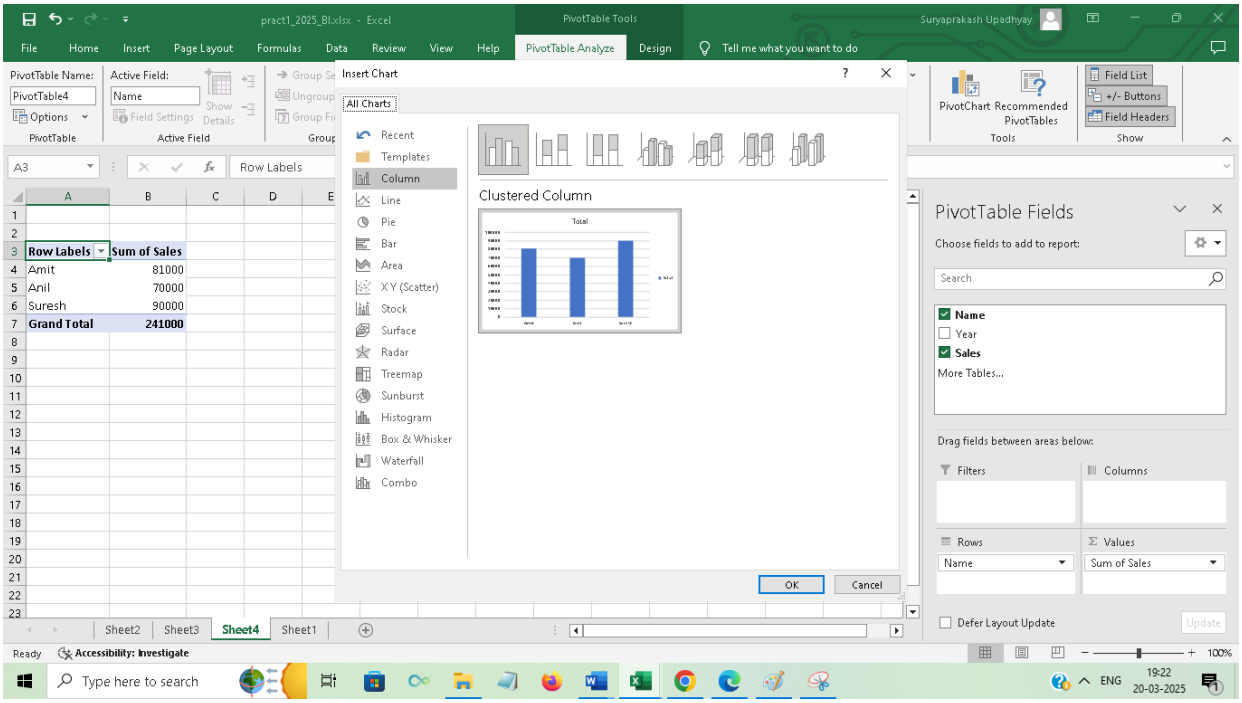




Click on ok button.



Click on Piovt Chart and select any chart.


Insert Chart
All Charts
Recent
Templates
Column
Line
Pie
Bar
Area
X Y (Scatter)
Stock
Surface
Radar
Treemap
Sunburst
Histogram
Box & Whisker
Waterfall
Combo
Clustered Column
OK
Cancel
PivotTable Fields
Choose fields to add to report:
Search
Name
Year
Sales
More Tables...
Drag fields between areas below:
Filters
Columns
Rows
Values
Name
Sum of Sales
Defer Layout Update
Update
Row Labels
Sum of Sales
Amit 81000
Anil 70000
Suresh 90000
Grand Total 241000
Sheet2
Sheet3
Sheet4
Sheet1
Ready
Accessibility: Investigate
Type here to search
ENG
19:22
20-03-2025

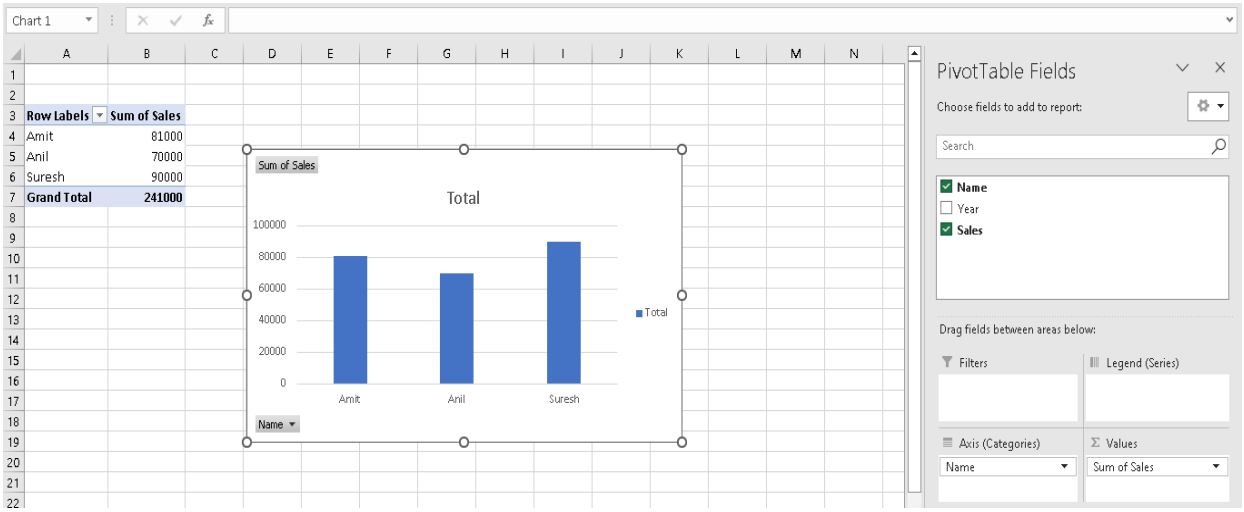


Chart 1
Row Labels
Sum of Sales
Amit 81000
Anil 70000
Suresh 90000
Grand Total 241000
Sum of Sales
Total
100000
80000
60000
40000
20000
0
Amit
Anil
Suresh
Total
Name
PivotTable Fields
Choose fields to add to report:
Search
Name
Year
Sales
Drag fields between areas below:
Filters
Legend (Series)
Axis (Categories)
Values
Name
Sum of Sales

Practical: 1b

Aim: Import the cube in Microsoft Excel and create the Pivot table and Pivot Chart to perform data analysis.

Step 1: Import Data from Data Warehouse to Excel

1. **Open Microsoft Excel**.

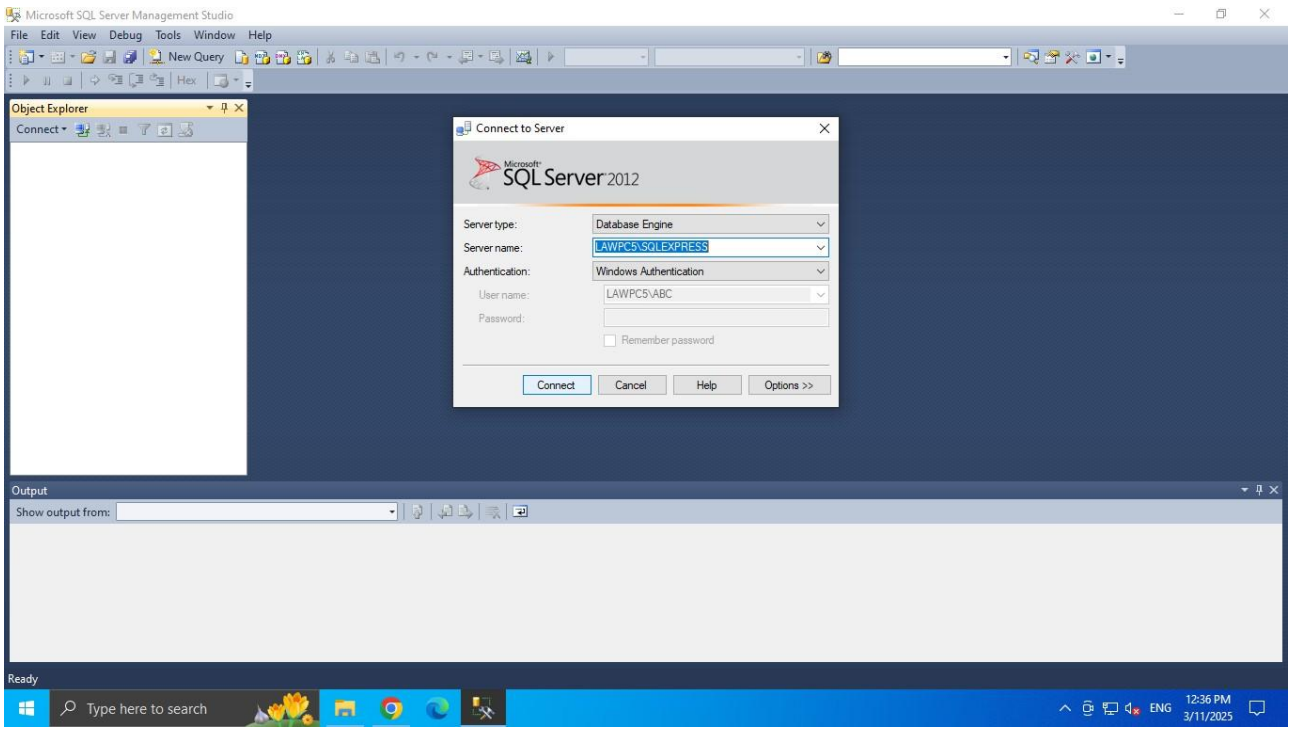
2. **Go to the Data tab**.

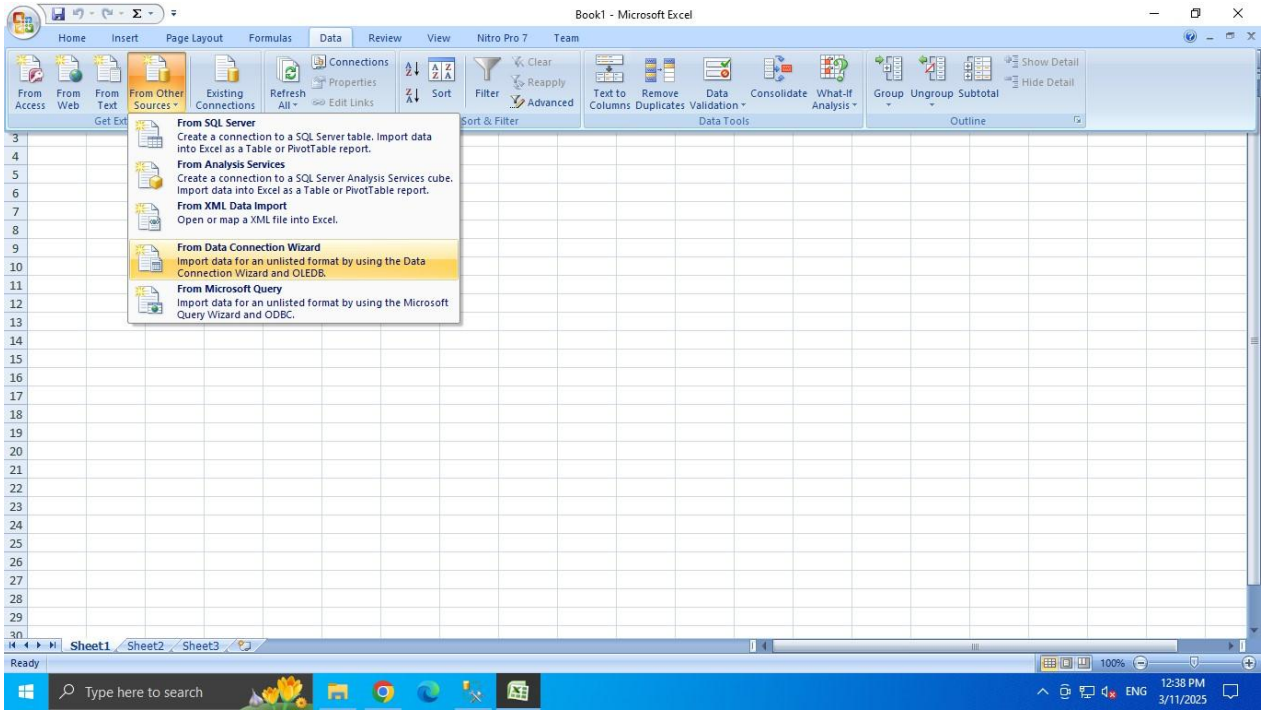
3. From the drop-down menu, choose **From Other Sources** and select **From Data Connection Wizard**.

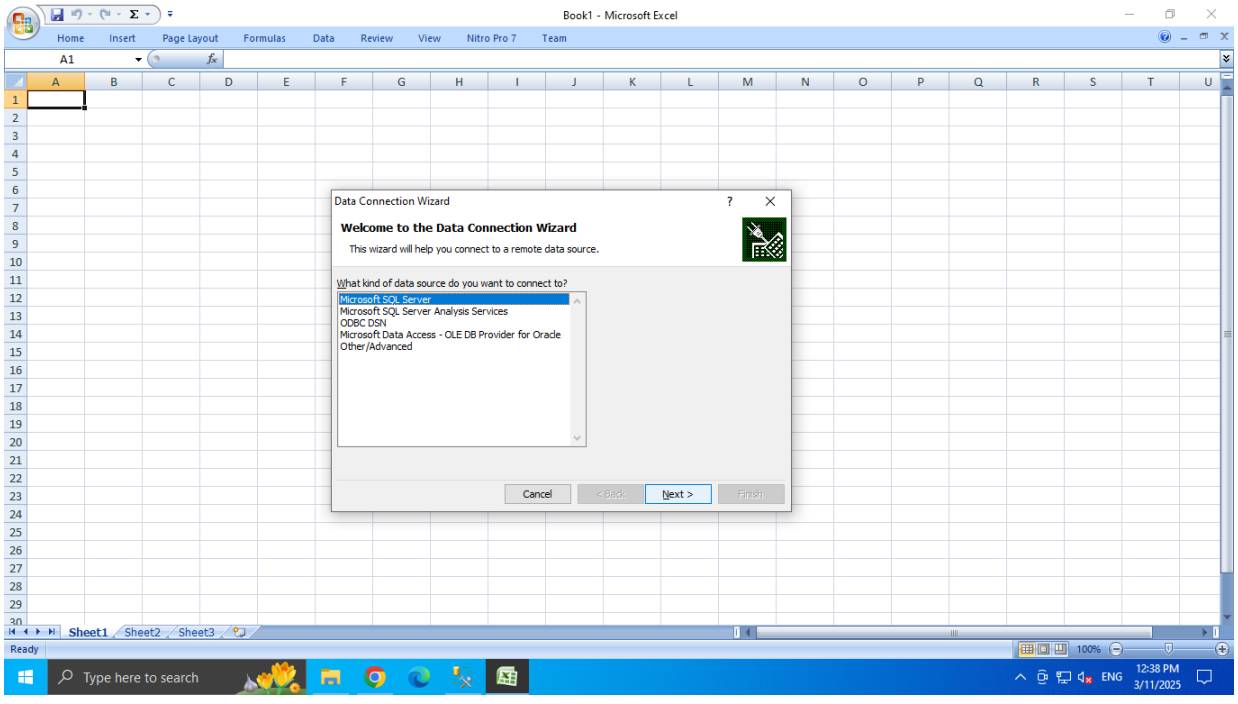
4. select **Microsoft From SQL Server** or choose the relevant source for your data warehouse.

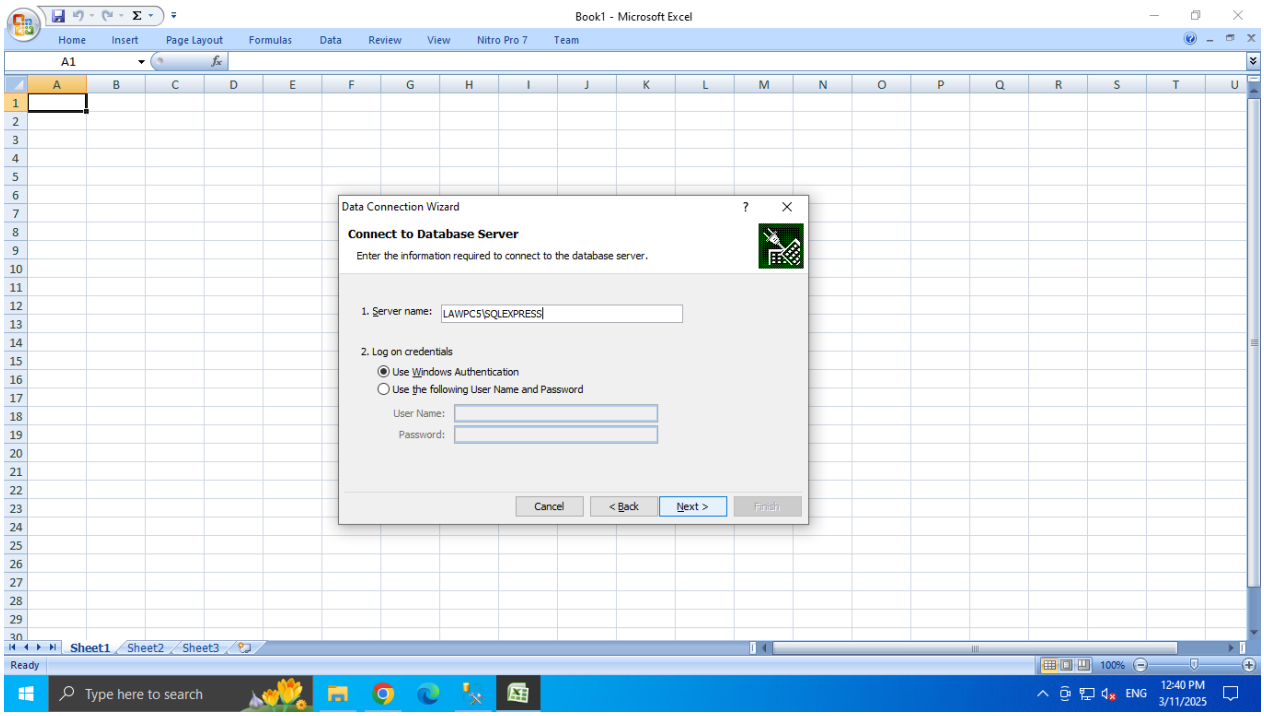
5. Connect to database server provide **connection** (Server Name).

6. Click **Next** and then **Finish**.



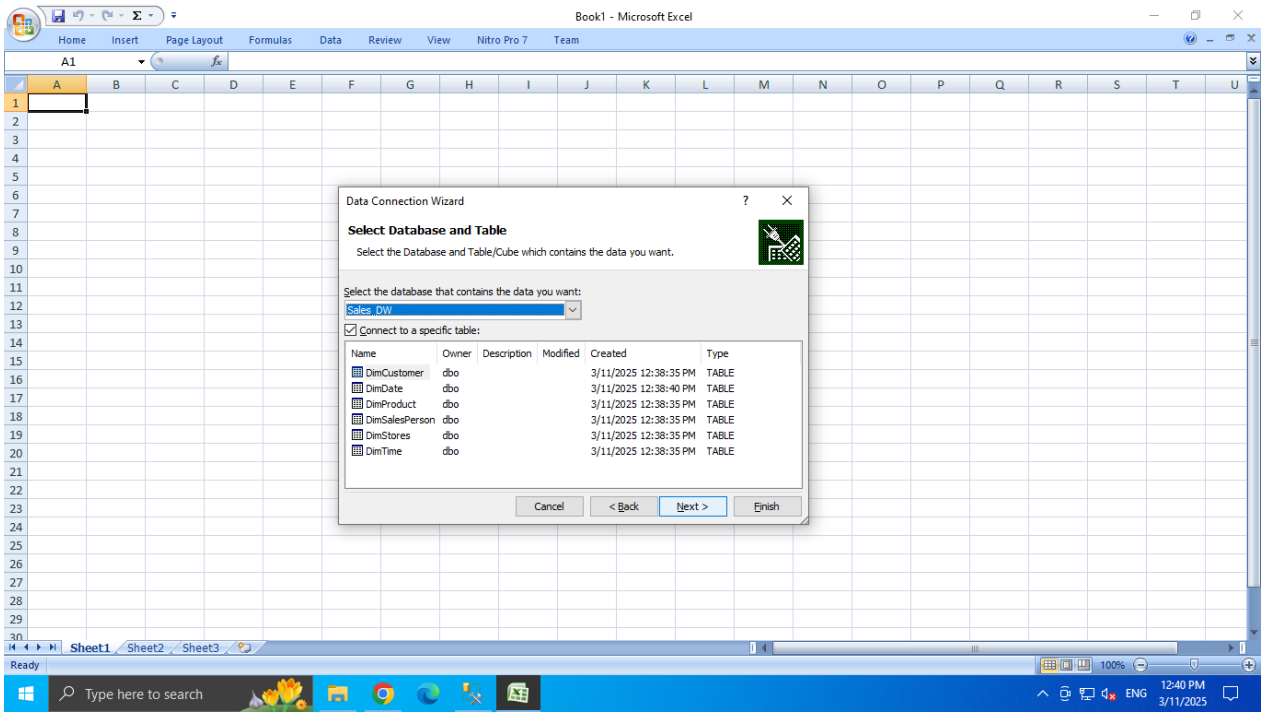

Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Nitro Pro 7
Team
From Access
From Web
From Text
From Other Sources
Existing Connections
Refresh All
Connections
Properties
Edit Links
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
Remove Duplicates
Data Validation
Consolidate
What-If Analysis
Group
Ungroup
Subtotal
Show Detail
Hide Detail
Get External Data
Sort & Filter
Data Tools
Outline
From SQL Server
Create a connection to a SQL Server table. Import data into Excel as a Table or PivotTable report.
From Analysis Services
Create a connection to a SQL Server Analysis Services cube. Import data into Excel as a Table or PivotTable report.
From XML Data Import
Open or map a XML file into Excel.
From Data Connection Wizard
Import data for an unlisted format by using the Data Connection Wizard and OLEDB.
From Microsoft Query
Import data for an unlisted format by using the Microsoft Query Wizard and ODBC.
Sheet1
Sheet2
Sheet3
Ready
100%
Type here to search
ENG
12:38 PM
3/11/2025

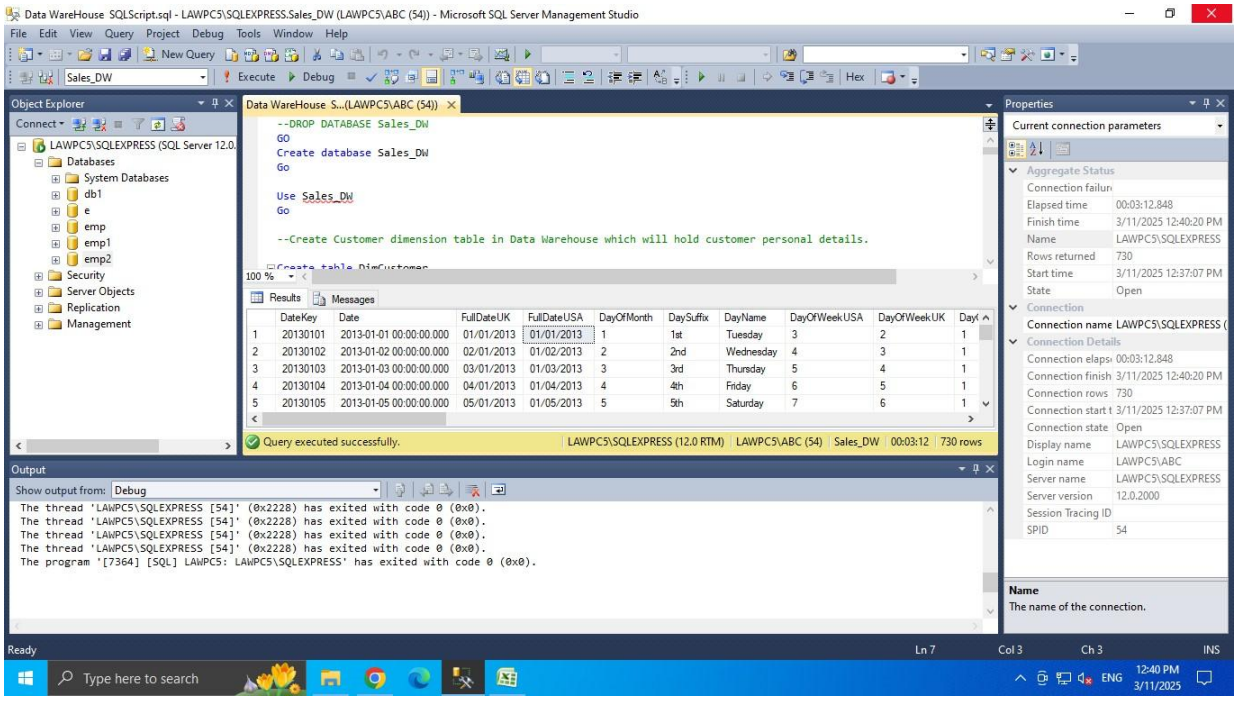

Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Nitro Pro 7
Team
A1
Data Connection Wizard
Welcome to the Data Connection Wizard
This wizard will help you connect to a remote data source.
What kind of data source do you want to connect to?
Microsoft SQL Server
Microsoft SQL Server Analysis Services
ODBC DSN
Microsoft Data Access - OLE DB Provider for Oracle
Other/Advanced
Cancel
< Back
Next >
Finish
Sheet1
Sheet2
Sheet3
Ready
100%
Type here to search
ENG
12:38 PM
3/11/2025

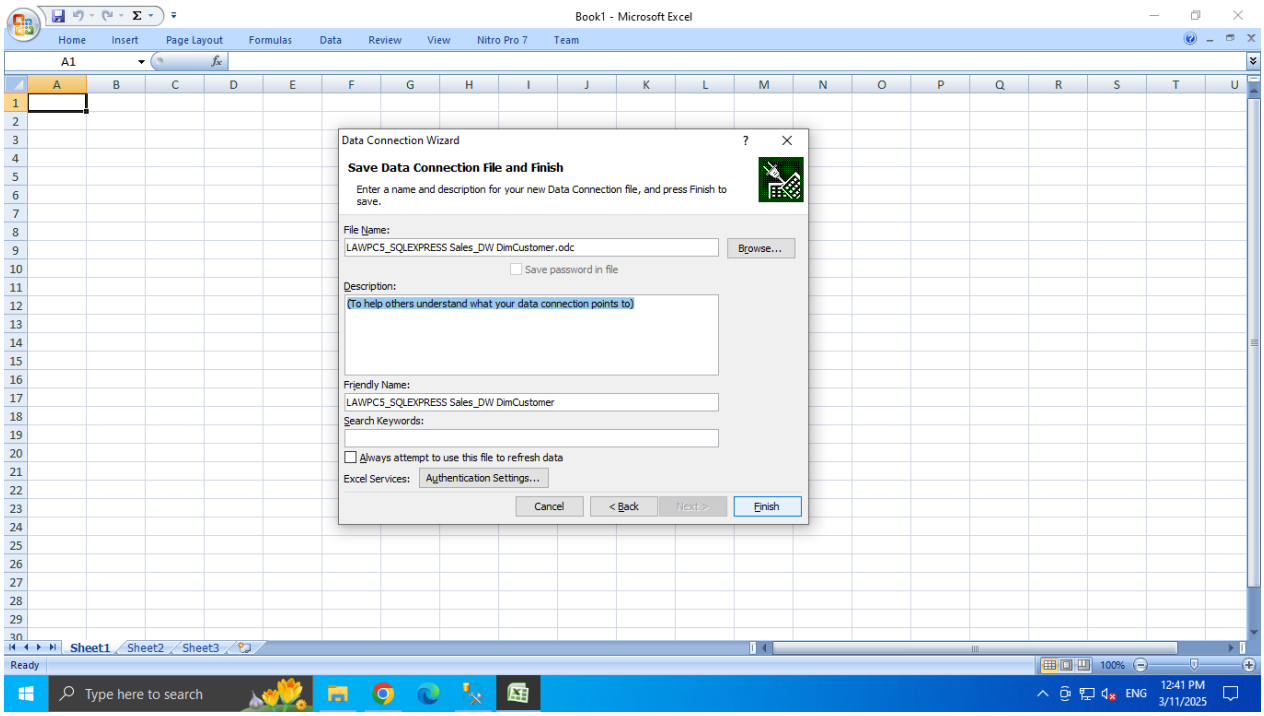


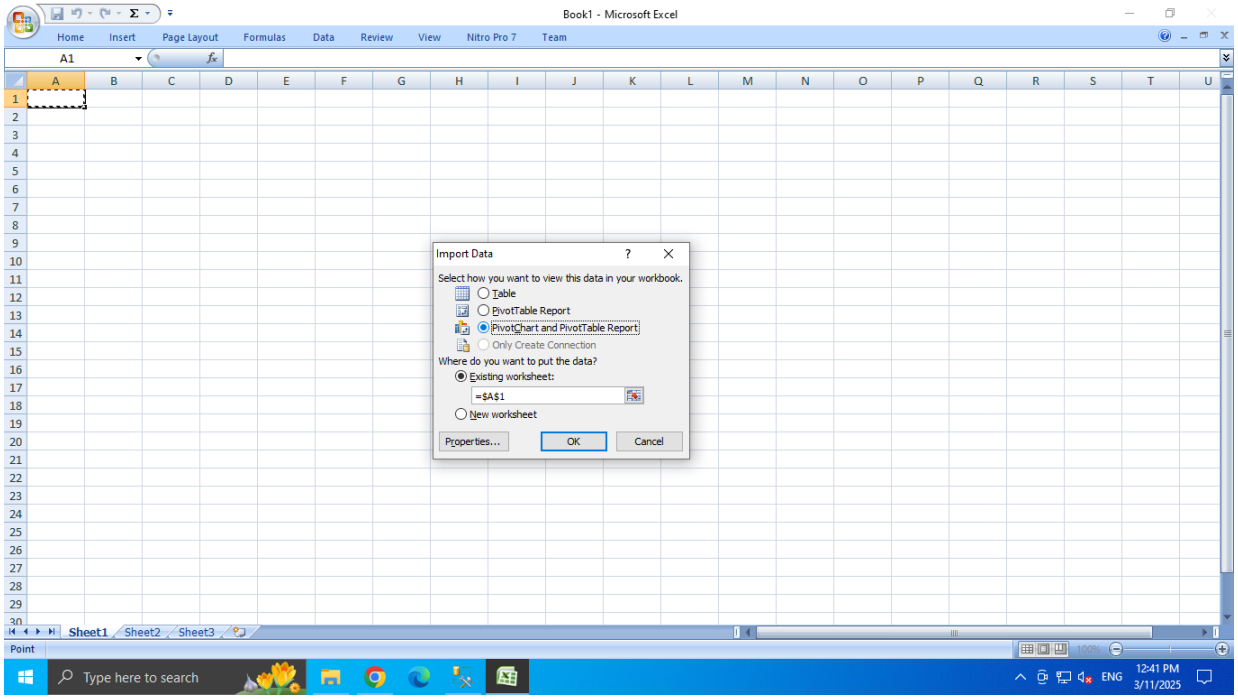
Step 2: Create a Pivot Chart

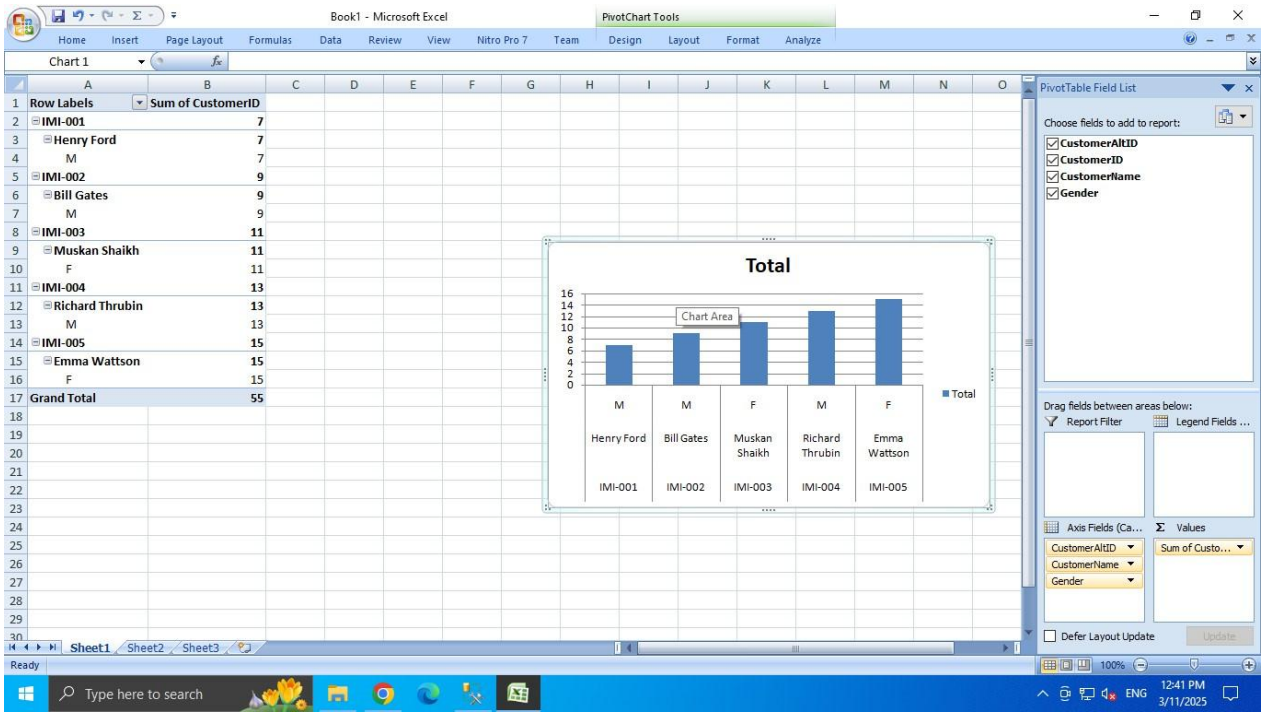
1. Once you have the PivotTable set up, click anywhere inside the PivotTable.
2. Select database and Table .
3. Save data connections file and finish .
4. Import data to select pivot chart and pivot table report.
5. Choose the chart type you want to visualize your data (e.g., Column, Line, Pie, etc.).
6. Excel will generate the PivotChart based on the data in your PivotTable.


Book1 - Microsoft Excel
Data Connection Wizard
Select Database and Table
Select the Database and Table/Cube which contains the data you want.
Select the database that contains the data you want:
Sales_DW
Connect to a specific table:
Name Owner Description Modified Created Type
DimCustomer dbo 3/11/2025 12:38:35 PM TABLE
DimDate dbo 3/11/2025 12:38:40 PM TABLE
DimProduct dbo 3/11/2025 12:38:35 PM TABLE
DimSalesPerson dbo 3/11/2025 12:38:35 PM TABLE
DimStores dbo 3/11/2025 12:38:35 PM TABLE
DimTime dbo 3/11/2025 12:38:35 PM TABLE
Cancel
< Back
Next >
Finish
Sheet1
Sheet2
Sheet3
Ready
Type here to search
ENG
12:40 PM
3/11/2025


Data WareHouse SQLScript.sql - LAWPC5\SQLEXPRESS.Sales_DW (LAWPC5\ABC (54)) - Microsoft SQL Server Management Studio
Object Explorer
Connect
LAWPC5\SQLEXPRESS (SQL Server 12.0.
Databases
System Databases
db1
e
emp
emp1
emp2
Security
Server Objects
Replication
Management
Data WareHouse S...(LAWPC5\ABC (54))
--DROP DATABASE Sales_DW
GO
Create database Sales_DW
Go
Use Sales_DW
Go
--Create Customer dimension table in Data Warehouse which will hold customer personal details.
Results
Messages
DateKey Date FullDateUK FullDateUSA DayOfMonth DaySuffix DayName DayOfWeekUSA DayOfWeekUK
1 20130101 2013-01-01 00:00:00.000 01/01/2013 01/01/2013 1 1st Tuesday 3 2 1
2 20130102 2013-01-02 00:00:00.000 02/01/2013 01/02/2013 2 2nd Wednesday 4 3 1
3 20130103 2013-01-03 00:00:00.000 03/01/2013 01/03/2013 3 3rd Thursday 5 4 1
4 20130104 2013-01-04 00:00:00.000 04/01/2013 01/04/2013 4 4th Friday 6 5 1
5 20130105 2013-01-05 00:00:00.000 05/01/2013 01/05/2013 5 5th Saturday 7 6 1
Query executed successfully.
LAWPC5\SQLEXPRESS (12.0 RTM)
LAWPC5\ABC (54)
Sales_DW
00:03:12
730 rows
Properties
Current connection parameters
Aggregate Status
Connection failur
Elapsed time 00:03:12.848
Finish time 3/11/2025 12:40:20 PM
Name LAWPC5\SQLEXPRESS
Rows returned 730
Start time 3/11/2025 12:37:07 PM
State Open
Connection
Connection name LAWPC5\SQLEXPRESS
Connection Details
Connection elaps 00:03:12.848
Connection finish 3/11/2025 12:40:20 PM
Connection rows 730
Connection start t 3/11/2025 12:37:07 PM
Connection state Open
Display name LAWPC5\SQLEXPRESS
Login name LAWPC5\ABC
Server name LAWPC5\SQLEXPRESS
Server version 12.0.2000
Session Tracing ID
SPID 54
Name
The name of the connection.
Output
Show output from: Debug
The thread 'LAWPC5\SQLEXPRESS [54]' (0x2228) has exited with code 0 (0x0).
The thread 'LAWPC5\SQLEXPRESS [54]' (0x2228) has exited with code 0 (0x0).
The thread 'LAWPC5\SQLEXPRESS [54]' (0x2228) has exited with code 0 (0x0).
The thread 'LAWPC5\SQLEXPRESS [54]' (0x2228) has exited with code 0 (0x0).
The program '[7364] [SQL] LAWPC5: LAWPC5\SQLEXPRESS' has exited with code 0 (0x0).
Ready
Ln 7
Col 3
Ch 3
INS
Type here to search
ENG
12:40 PM
3/11/2025


Book1 - Microsoft Excel
Data Connection Wizard
Save Data Connection File and Finish
Enter a name and description for your new Data Connection file, and press Finish to save.
File Name:
LAWPC5_SQLEXPRESS Sales_DW DimCustomer.odc
Browse...
Save password in file
Description:
(To help others understand what your data connection points to)
Friendly Name:
LAWPC5_SQLEXPRESS Sales_DW DimCustomer
Search Keywords:
Always attempt to use this file to refresh data
Excel Services:
Authentication Settings...
Cancel
< Back
Next >
Finish
Sheet1
Sheet2
Sheet3
Ready
Type here to search
ENG
12:41 PM
3/11/2025

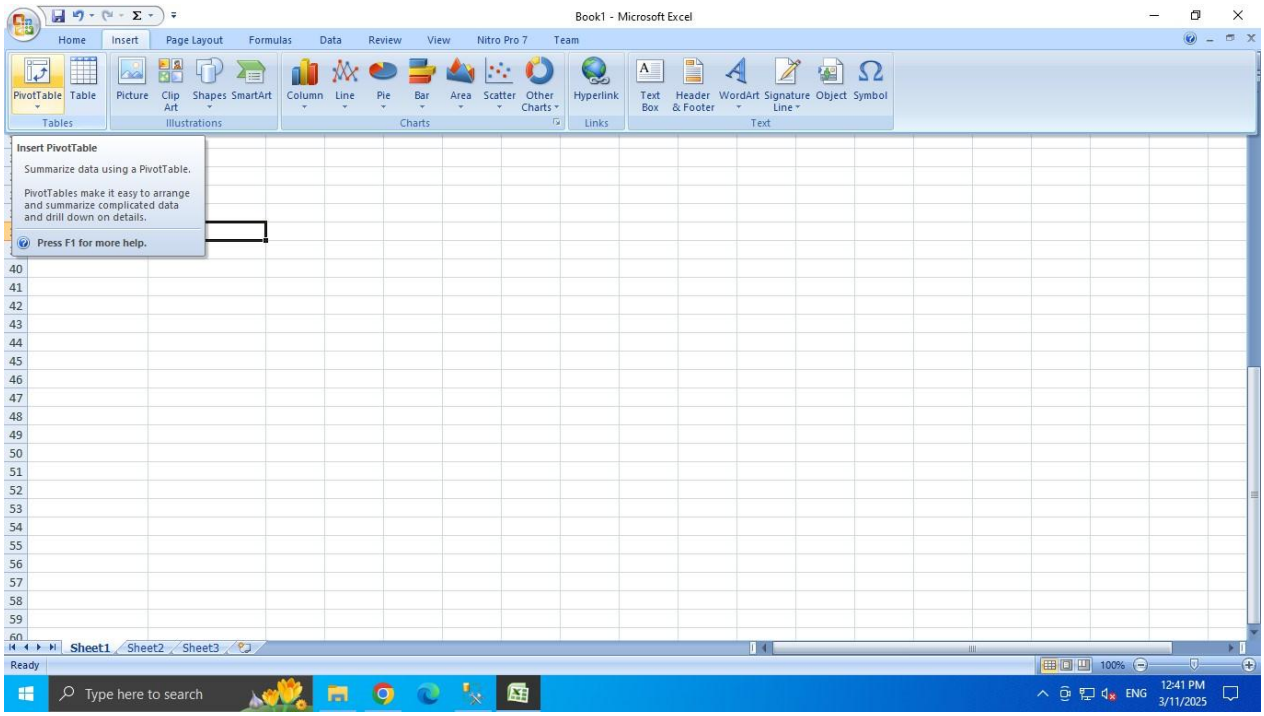

Book1 - Microsoft Excel
Import Data
Select how you want to view this data in your workbook.
Table
PivotTable Report
PivotChart and PivotTable Report
Only Create Connection
Where do you want to put the data?
Existing worksheet:
=$A$1
New worksheet
Properties...
OK
Cancel
Sheet1
Sheet2
Sheet3
Point
Type here to search
ENG
12:41 PM
3/11/2025

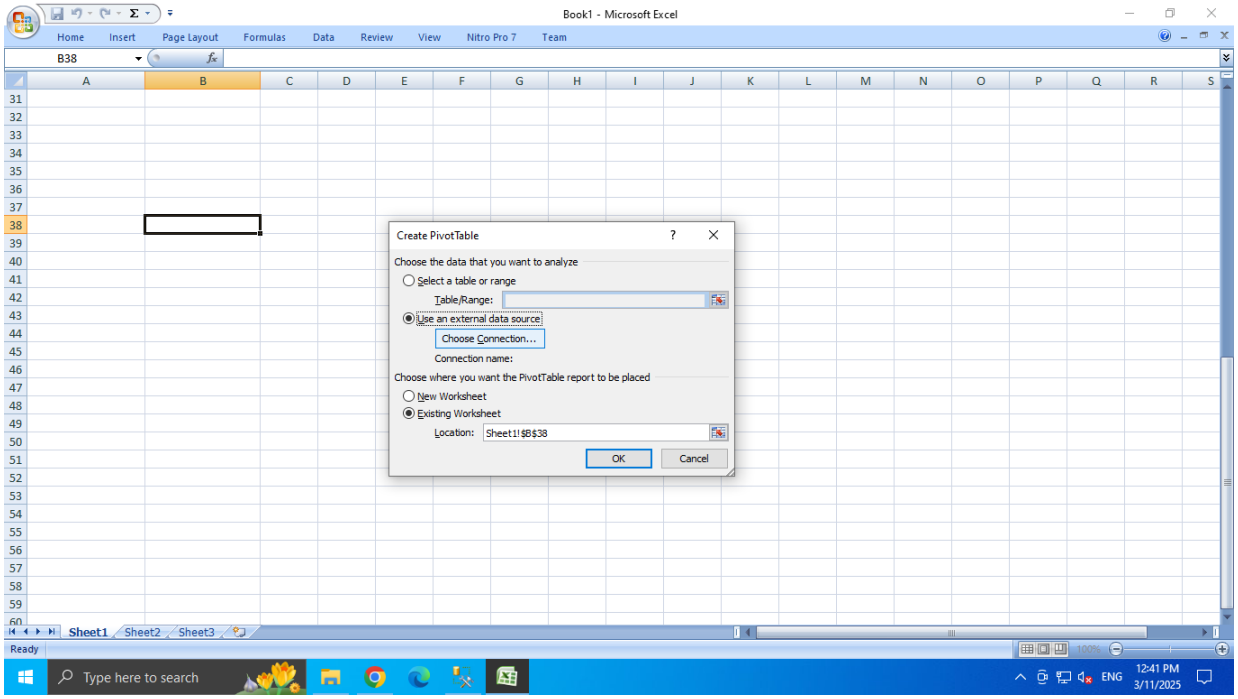


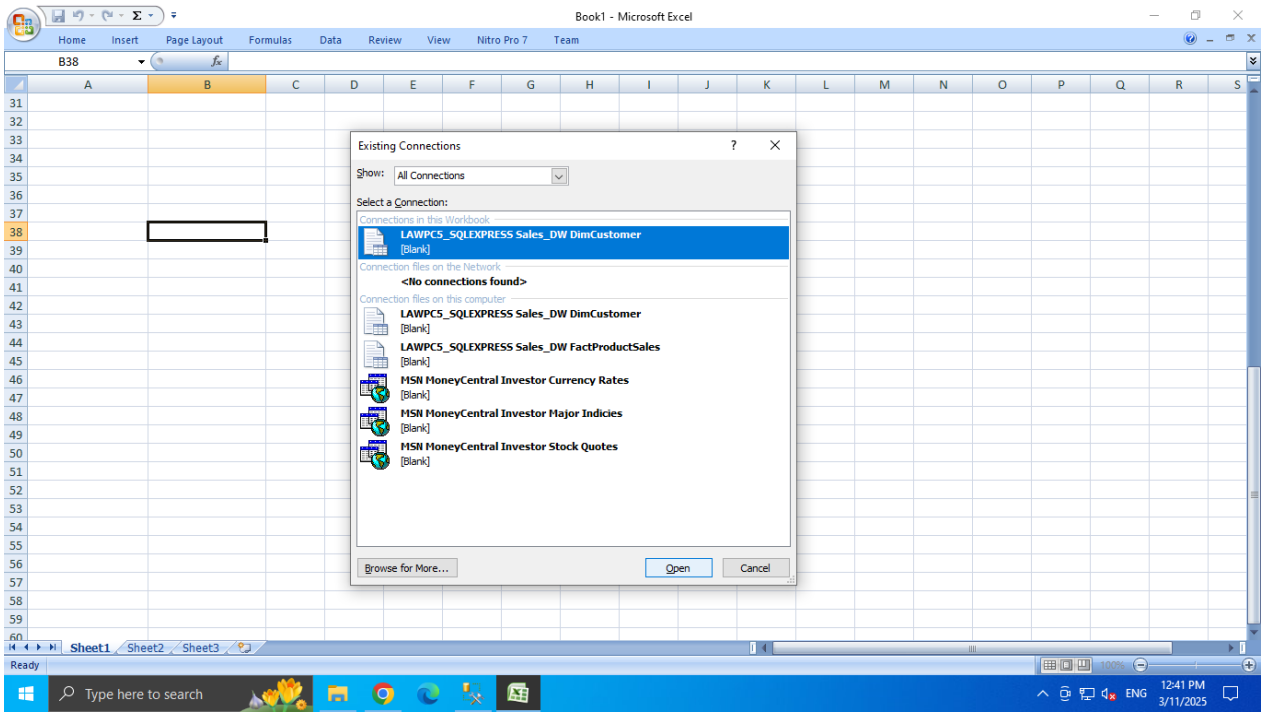
Row Labels	Sum of CustomerID
IMI-001	7
Henry Ford	7
M	7
IMI-002	9
Bill Gates	9
M	9
IMI-003	11
Muskan Shaikh	11
F	11
IMI-004	13
Richard Thrubin	13
M	13
IMI-005	15
Emma Wattson	15
F	15
Grand Total	55

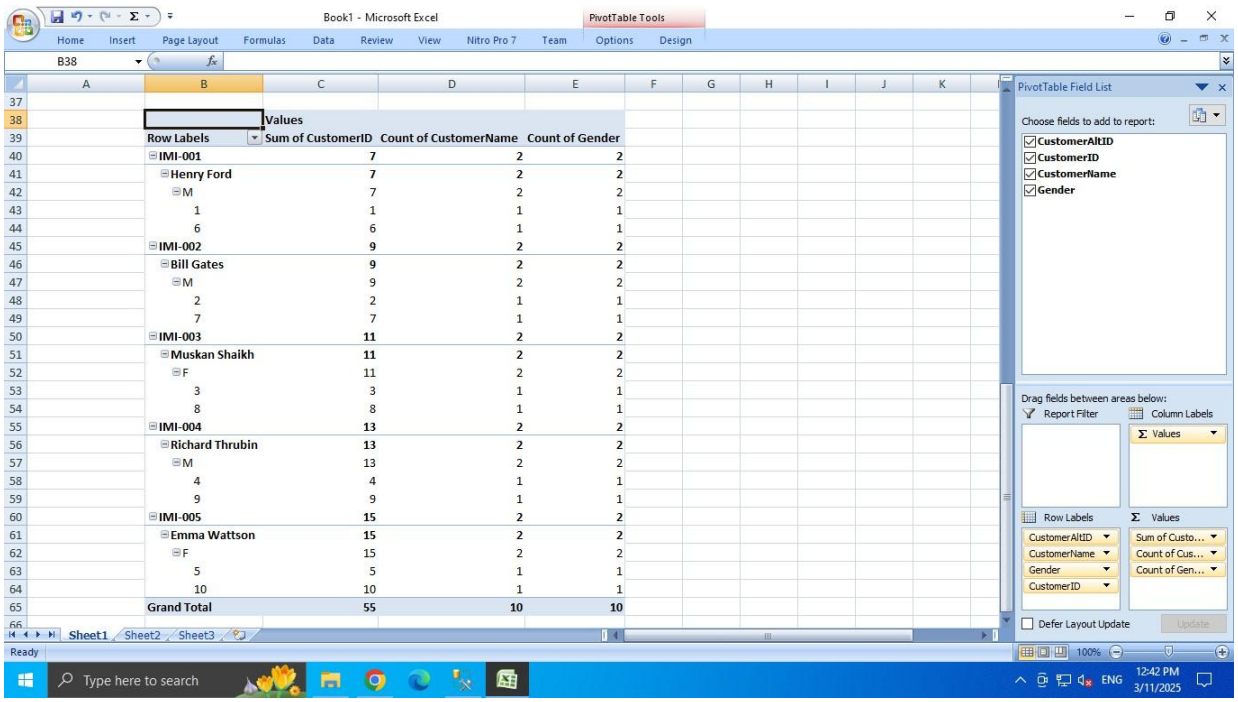
Step 3: Create a Pivot Table

1. Once the data is imported, click anywhere inside the data range.
2. Go to the **Insert tab** on the Ribbon.
3. Click on **PivotTable** in the Tables group.
4. You can also choose to place the PivotTable in a new worksheet or an existing one.
5. Click **OK** to create the PivotTable.
6. In the **PivotTable Field List** pane, drag fields into the **Rows**, **Columns**, **Values**, and **Filters** areas to organize the data. For example:
    a. **Rows**: Place dimensions like product names or regions.
    b. **Columns**: Place time periods (e.g., months or years).
    c. **Values**: Put numerical values like sales or revenue.


Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Nitro Pro 7
Team
PivotTable
Table
Picture
Clip Art
Shapes
SmartArt
Column
Line
Pie
Bar
Area
Scatter
Other Charts
Hyperlink
Text Box
Header & Footer
WordArt
Signature Line
Object
Symbol
Tables
Illustrations
Charts
Links
Text
Insert PivotTable
Summarize data using a PivotTable.
PivotTables make it easy to arrange and summarize complicated data and drill down on details.
Press F1 for more help.
Sheet1
Sheet2
Sheet3
Ready
100%
Type here to search
ENG
12:41 PM
3/11/2025


Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Nitro Pro 7
Team
B38
Create PivotTable
Choose the data that you want to analyze
Select a table or range
Table/Range:
Use an external data source
Choose Connection...
Connection name:
Choose where you want the PivotTable report to be placed
New Worksheet
Existing Worksheet
Location:
Sheet1!$B$38
OK
Cancel
Sheet1
Sheet2
Sheet3
Ready
Type here to search
ENG
12:41 PM
3/11/2025


Book1 - Microsoft Excel
Existing Connections
Show: All Connections
Select a Connection:
Connections in this Workbook
LAWPC5_SQLEXPRESS Sales_DW DimCustomer
[Blank]
Connection files on the Network
<No connections found>
Connection files on this computer
LAWPC5_SQLEXPRESS Sales_DW DimCustomer
[Blank]
LAWPC5_SQLEXPRESS Sales_DW FactProductSales
[Blank]
MSN MoneyCentral Investor Currency Rates
[Blank]
MSN MoneyCentral Investor Major Indicies
[Blank]
MSN MoneyCentral Investor Stock Quotes
[Blank]
Browse for More...
Open
Cancel
Sheet1
Sheet2
Sheet3
Ready
Type here to search
ENG
12:41 PM
3/11/2025


Book1 - Microsoft Excel
PivotTable Tools
Values
Row Labels
Sum of CustomerID
Count of CustomerName
Count of Gender
IMI-001 7 2 2
Henry Ford 7 2 2
M 7 2 2
1 1 1 1
6 6 1 1
IMI-002 9 2 2
Bill Gates 9 2 2
M 9 2 2
2 2 1 1
7 7 1 1
IMI-003 11 2 2
Muskan Shaikh 11 2 2
F 11 2 2
3 3 1 1
8 8 1 1
IMI-004 13 2 2
Richard Thrubin 13 2 2
M 13 2 2
4 4 1 1
9 9 1 1
IMI-005 15 2 2
Emma Wattson 15 2 2
F 15 2 2
5 5 1 1
10 10 1 1
Grand Total 55 10 10
PivotTable Field List
Choose fields to add to report:
CustomerAltID
CustomerID
CustomerName
Gender
Drag fields between areas below:
Report Filter
Column Labels
Values
Row Labels
Values
Defer Layout Update
Update
Sheet1
Sheet2
Sheet3
Ready
Type here to search
ENG
12:42 PM
3/11/2025

Practical: 2

Aim: Apply the what –if Analysis for data visualization. Design and generate necessary reports based on the data warehouse data. Use Excel.

Step 1: Apply What-If Analysis in Excel

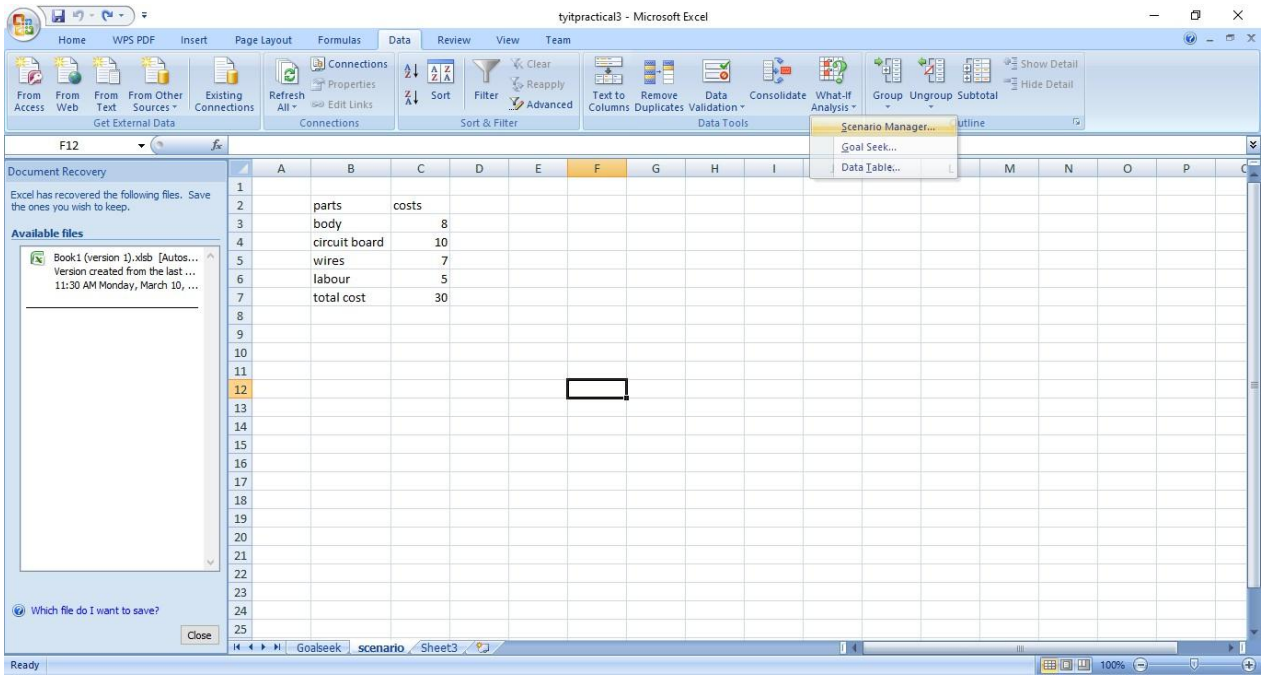
Microsoft Excel offers several tools for **What-If Analysis** that help simulate different scenarios and outcomes based on changes to input data. These tools are:

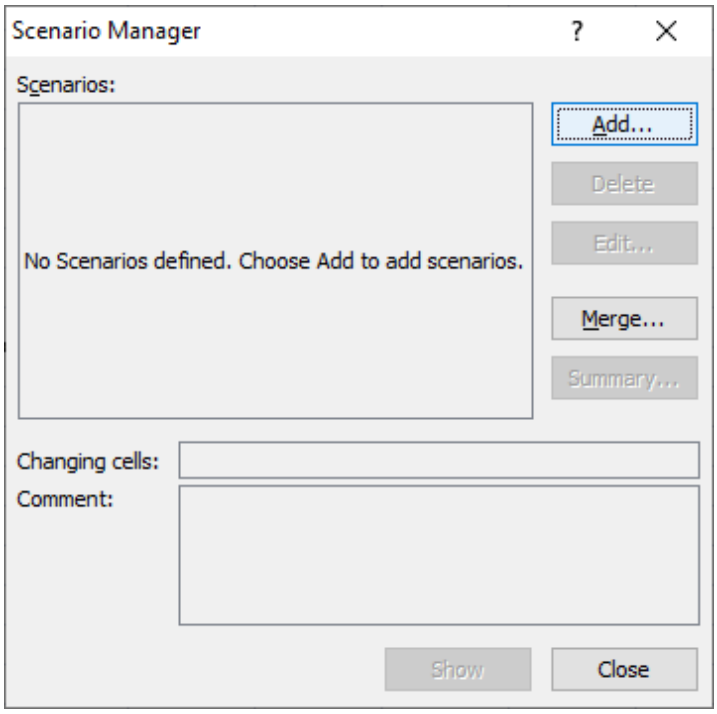
1. **Scenario Manager**
2. **Goal Seek**

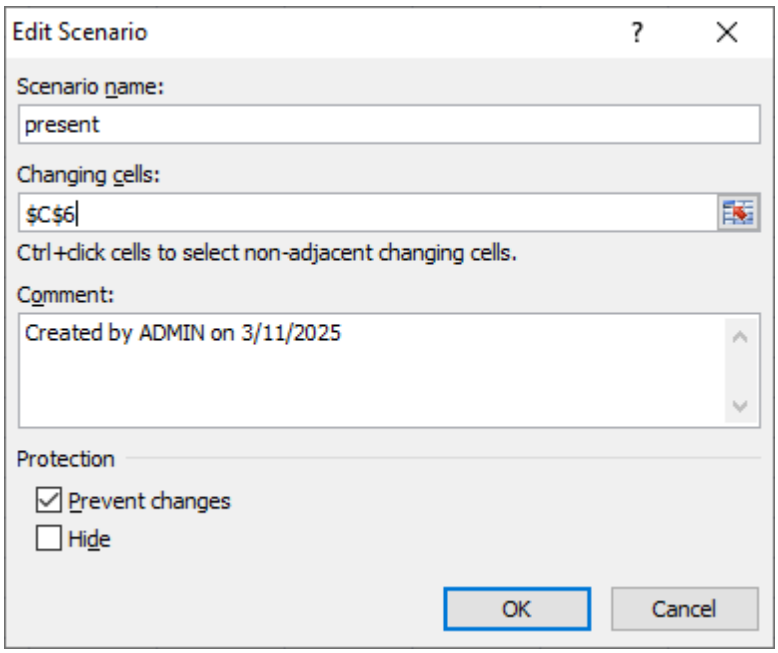
2.1. Scenario Manager

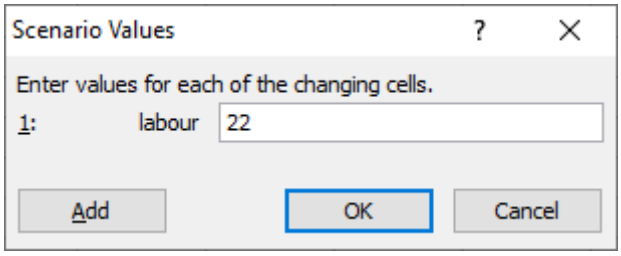
The Scenario Manager lets you create and compare different sets of input values to see how changes affect the output.

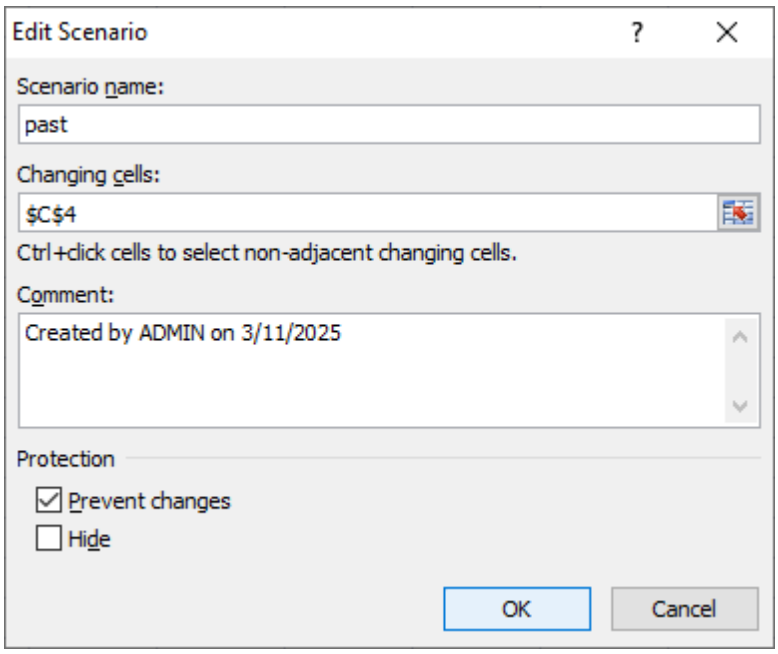
- **Step-by-Step**:
    1. Go to the **Data tab** on the Ribbon.
    2. Click on **What-If Analysis**, then select **Scenario Manager**.
    3. In the **Scenario Manager** dialog box, click **Add** to create a new scenario.
    4. Name the scenario (e.g., "Best Case", "Worst Case", "Most Likely Case").
    5. Define the changing cells (these are the input values you want to experiment with, such as sales volume, price, etc.).
    6. Enter the values for each scenario you want to test.
    7. After setting up all scenarios, click **OK**.
    8. To view the results of each scenario, click **Show** in the Scenario Manager dialog box.

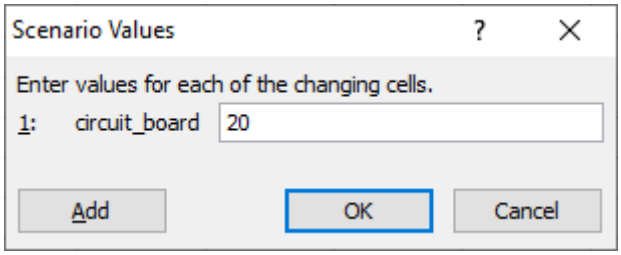

tyitpractical3 - Microsoft Excel
Home
WPS PDF
Insert
Page Layout
Formulas
Data
Review
View
Team
Scenario Manager...
Goal Seek...
Data Table...
Document Recovery
Excel has recovered the following files. Save the ones you wish to keep.
Available files
Book1 (version 1).xlsb [Autos...
Version created from the last ...
11:30 AM Monday, March 10, ...
Which file do I want to save?
Close
parts
costs
body
8
circuit board
10
wires
7
labour
5
total cost
30
Goalseek
scenario
Sheet3
Ready

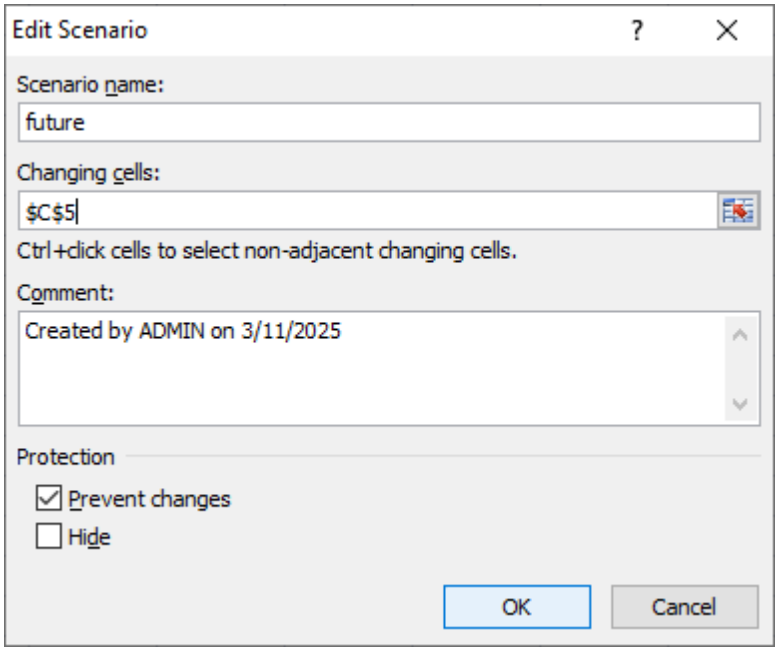

Scenario Manager
Scenarios:
No Scenarios defined. Choose Add to add scenarios.
Add...
Delete
Edit...
Merge...
Summary...
Changing cells:
Comment:
Show
Close

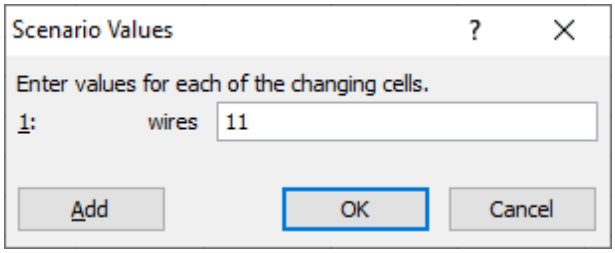

Edit Scenario
Scenario name:
present
Changing cells:
$C$6
Ctrl+click cells to select non-adjacent changing cells.
Comment:
Created by ADMIN on 3/11/2025
Protection
Prevent changes
Hide
OK
Cancel

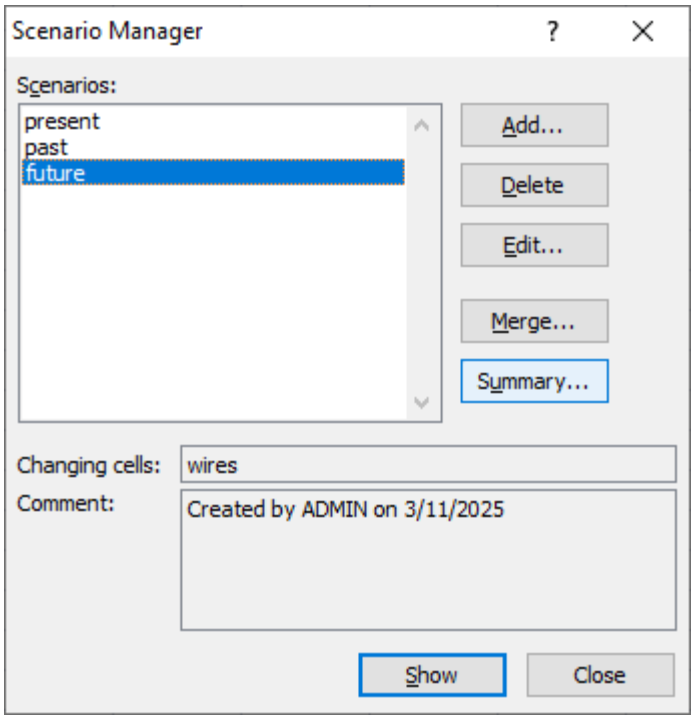

Scenario Values
Enter values for each of the changing cells.
1:
labour
22
Add
OK
Cancel

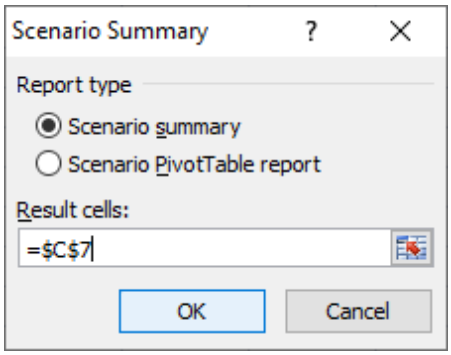

Edit Scenario
Scenario name:
past
Changing cells:
$C$4
Ctrl+click cells to select non-adjacent changing cells.
Comment:
Created by ADMIN on 3/11/2025
Protection
Prevent changes
Hide
OK
Cancel

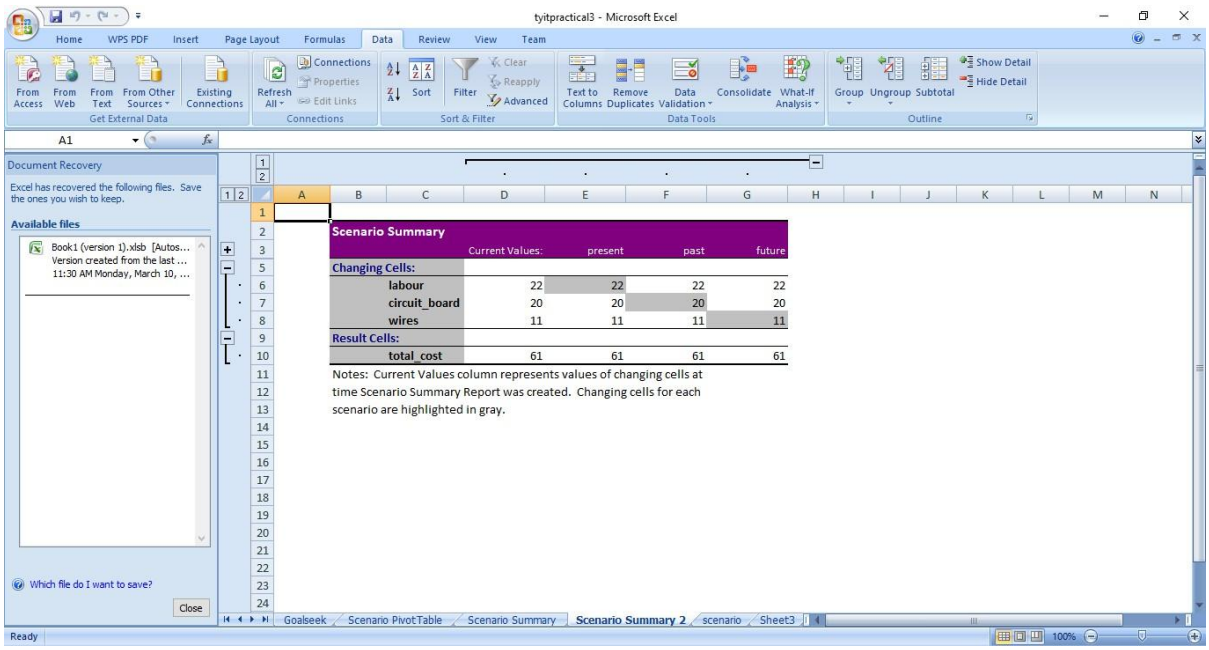

Scenario Values
Enter values for each of the changing cells.
1: circuit_board 20
Add
OK
Cancel


Edit Scenario
Scenario name:
future
Changing cells:
$C$5
Ctrl+click cells to select non-adjacent changing cells.
Comment:
Created by ADMIN on 3/11/2025
Protection
Prevent changes
Hide
OK
Cancel


Scenario Values
Enter values for each of the changing cells.
1: wires 11
Add
OK
Cancel


Scenario Manager
Scenarios:
present
past
future
Add...
Delete
Edit...
Merge...
Summary...
Changing cells:
wires
Comment:
Created by ADMIN on 3/11/2025
Show
Close

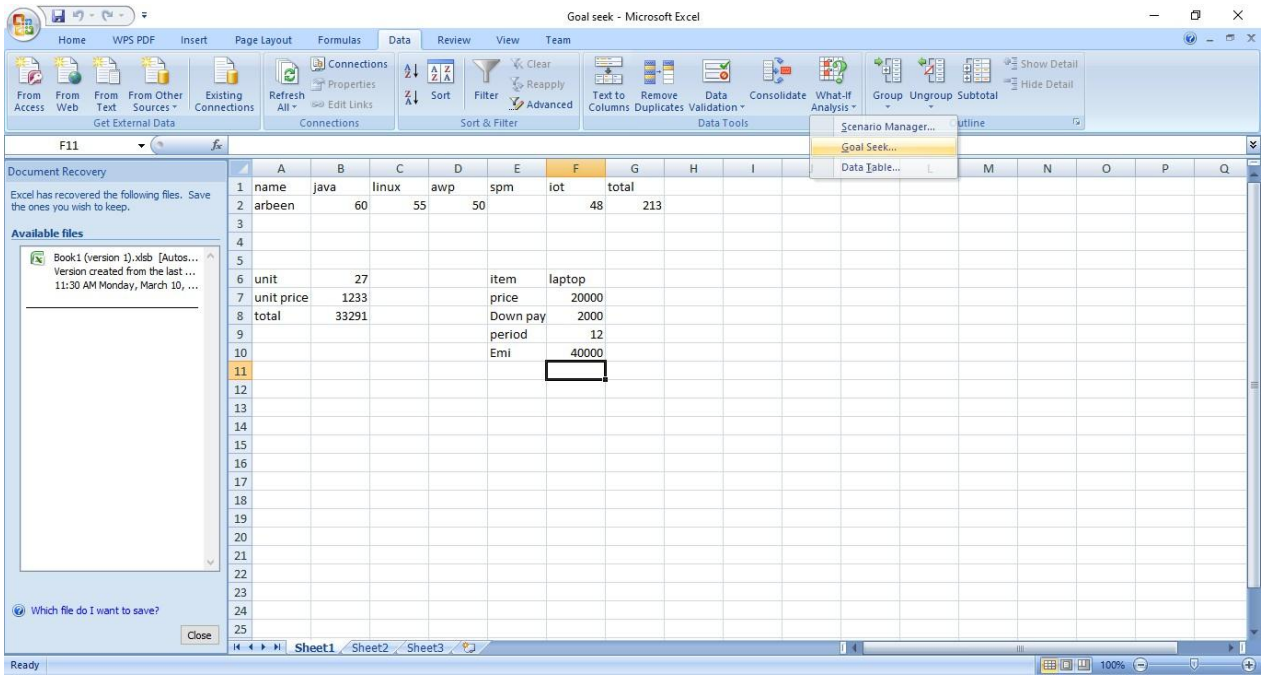

Scenario Summary
Report type
Scenario summary
Scenario PivotTable report
Result cells:
=$C$7
OK
Cancel



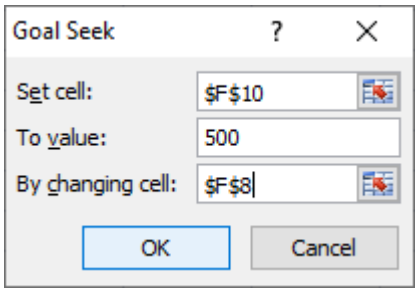
2.2. Goal Seek

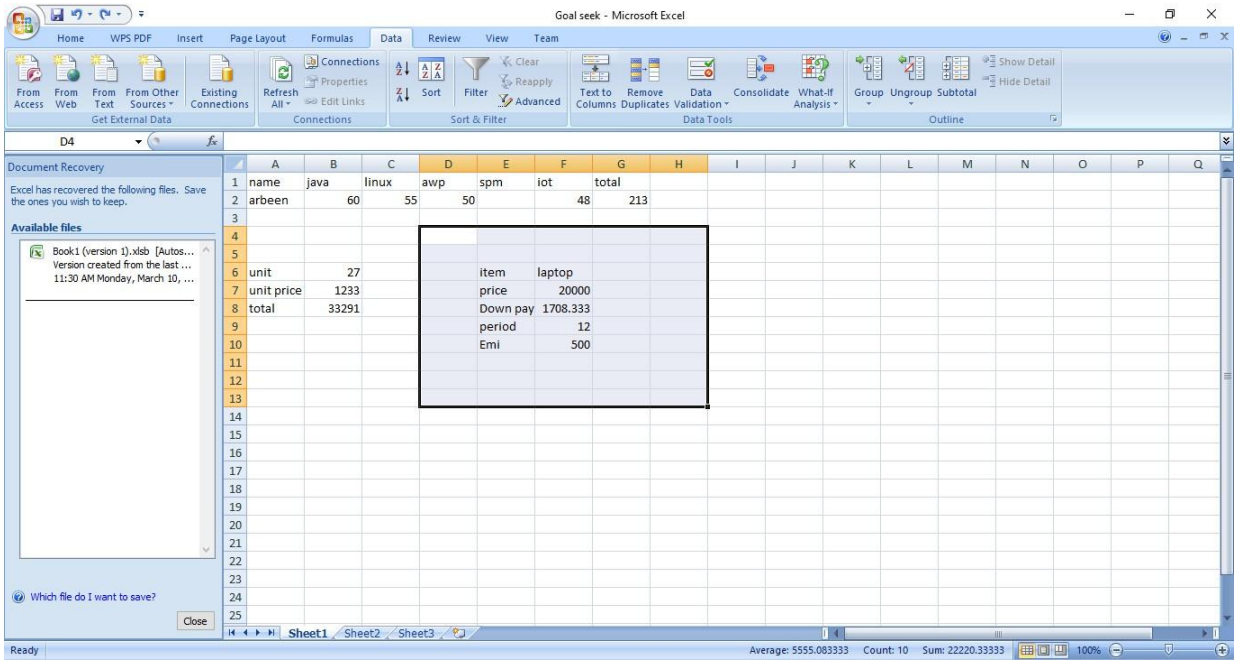
Goal Seek helps you find the input value needed to reach a specific result.

- **Step-by-Step**:
    1. Go to the **Data tab** on the Ribbon.
    2. Click on **What-If Analysis**, then select **Goal Seek**.
    3. In the **Goal Seek** dialog box, specify the following:
        - **Set cell**: The cell that contains the formula you want to solve.
        - **To value**: The result you want to achieve.
        - **By changing cell**: The input value that will be adjusted to reach the desired result.
    4. Click **OK**. Excel will calculate the value needed to achieve the target.



	A	B	C	D	E	F	G
1	name	java	linux	awp	spm	iot	total
2	arbeen	60	55	50		48	213
3							
4							
5							
6	unit	27			item	laptop	
7	unit price	1233			price	20000	
8	total	33291			Down pay	2000	
9					period	12	
10					Emi	40000	





	A	B	C	D	E	F	G
1	name	java	linux	awp	spm	iot	total
2	arbeen	60	55	50		48	213
3							
4							
5							
6	unit	27			item	laptop	
7	unit price	1233			price	20000	
8	total	33291			Down pay	1708.333	
9					period	12	
10					Emi	500	

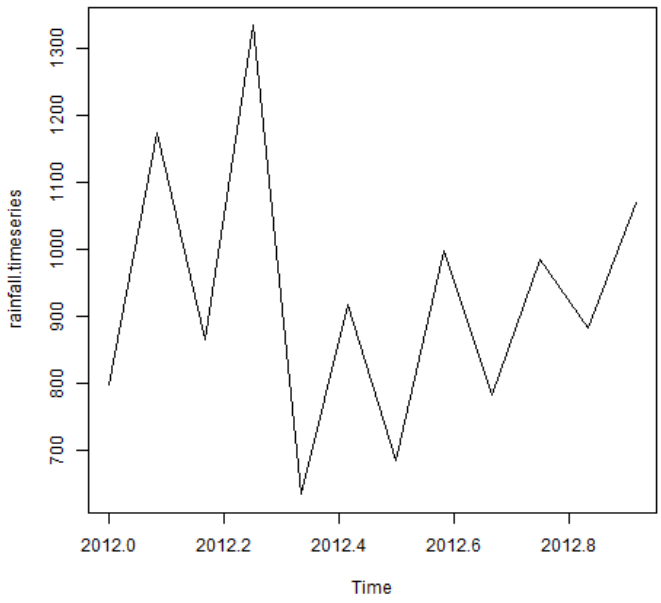
Practical: 3

Aim: Perform the data classification using classification algorithm using R/Python.

Code:

```
rainfall <-c(799,1174.8,865.1,1334.6,635.4,918.5,685.5,998.6,784.2,985,882.8,1071)

rainfall.timeseries <- ts(rainfall,start = c(2012,1),frequency = 12)

print(rainfall.timeseries)

png(file = "rainfall.png")

plot(rainfall.timeseries)

dev.off()
```

Output:



Practical: 4

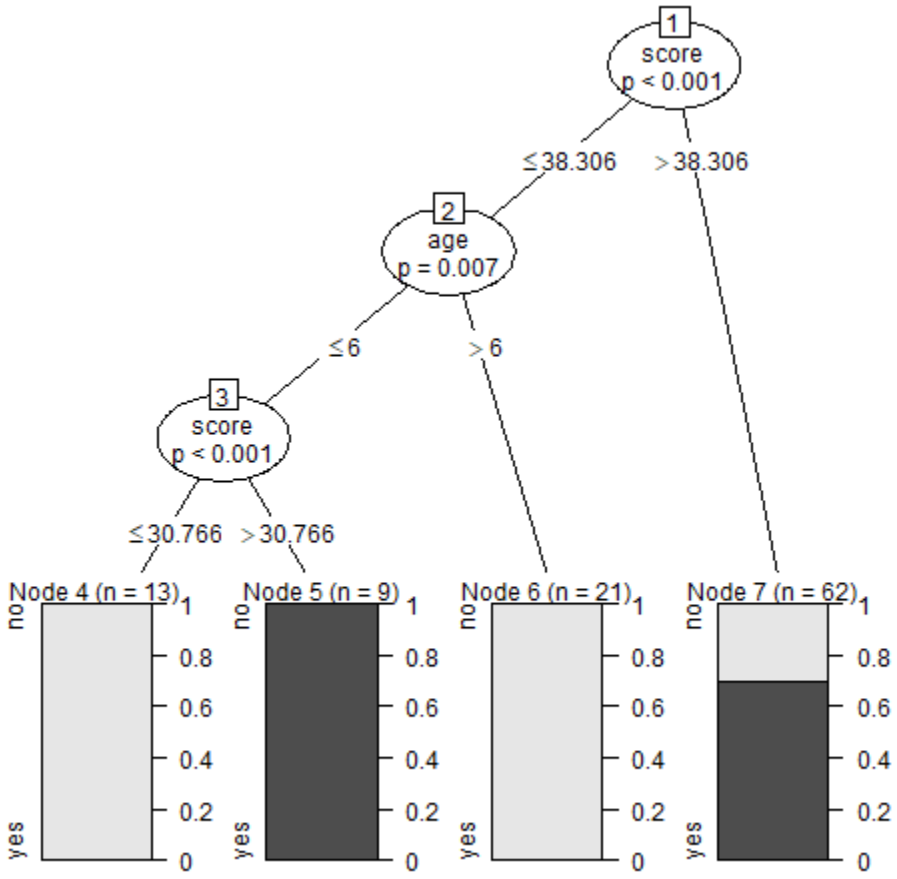
Aim: Perform the data clustering using clustering algorithm using R/Python.

Code:

```
library(party)

print(head(readingSkills))

input.dat <- readingSkills[c(1:105),]

png(file = "decision_tree.png")

output.tree <- ctree(nativeSpeaker ~ age + shoeSize + score,data = input.dat)

plot(output.tree)

dev.off()
```

Output:

```
  nativeSpeaker age shoeSize    score
1           yes   5 24.83189 32.29385
2           yes   6 25.95238 36.63105
3            no  11 30.42170 49.60593
4           yes   7 28.66450 40.28456
5           yes  11 31.88207 55.46085
6           yes  10 30.07843 52.83124
> |
```


1
score
p < 0.001
≤ 38.306
> 38.306
2
age
p = 0.007
≤ 6
> 6
3
score
p < 0.001
≤ 30.766
> 30.766
Node 4 (n = 13)
Node 5 (n = 9)
Node 6 (n = 21)
Node 7 (n = 62)
no
yes
1
0.8
0.6
0.4
0.2
0

Practical: 5

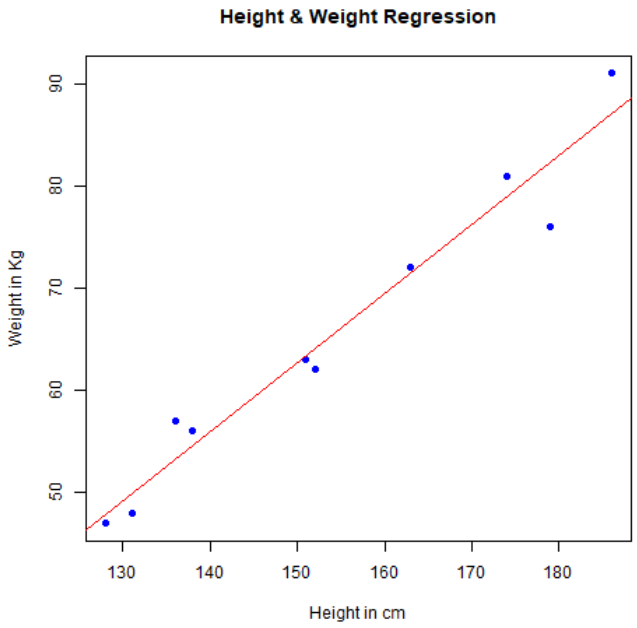
Aim: Perform the Linear regression on the given data warehouse data using R/Python.

Code:

```
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
# Apply the lm() function to create a linear regression model
relation <- lm(y ~ x)
# Print the model summary
print(summary(relation))
# Predict the weight of a person with height 170
a <- data.frame(x = 170)
result <- predict(relation, a)
print(result)
# Visualizing the Regression Graphically
png(file = "linearregression.png")
plot(x, y, col = "blue", main = "Height & Weight Regression",
    xlab = "Height in cm", ylab = "Weight in Kg", pch = 16)
abline(relation, col = "red")
dev.off()
```

Output:

```
> source("C:/Users/Lenovo/OneDrive/Desktop/SEM 6 practicals/BI PRACTICALS/p7.
R")

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3002 -1.6629  0.0412  1.8944  3.9775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.45509    8.04901  -4.778  0.00139 **
x             0.67461    0.05191  12.997 1.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.253 on 8 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06

       1
76.22869
>
```



Practical :6

Aim: Perform the logistic regression on the given data warehouse data using R/Python.

Code:

```
# Load necessary libraries
library(dplyr)
library(titanic)
library(pROC)

# Load Titanic dataset
data("titanic_train")

# Check dataset structure
head(titanic_train)

# Data Cleaning: Removing rows with missing values
titanic_clean <- titanic_train %>%
  filter(!is.na(Age), !is.na(Embarked), !is.na(Sex), !is.na(Pclass))

# Convert categorical variables to factors
titanic_clean$Survived <- as.factor(titanic_clean$Survived)
titanic_clean$Pclass <- as.factor(titanic_clean$Pclass)
titanic_clean$Sex <- as.factor(titanic_clean$Sex)
titanic_clean$Embarked <- as.factor(titanic_clean$Embarked)

# Build Logistic Regression Model
model <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,
          data = titanic_clean, family = binomial)

# Print model summary
summary(model)
```

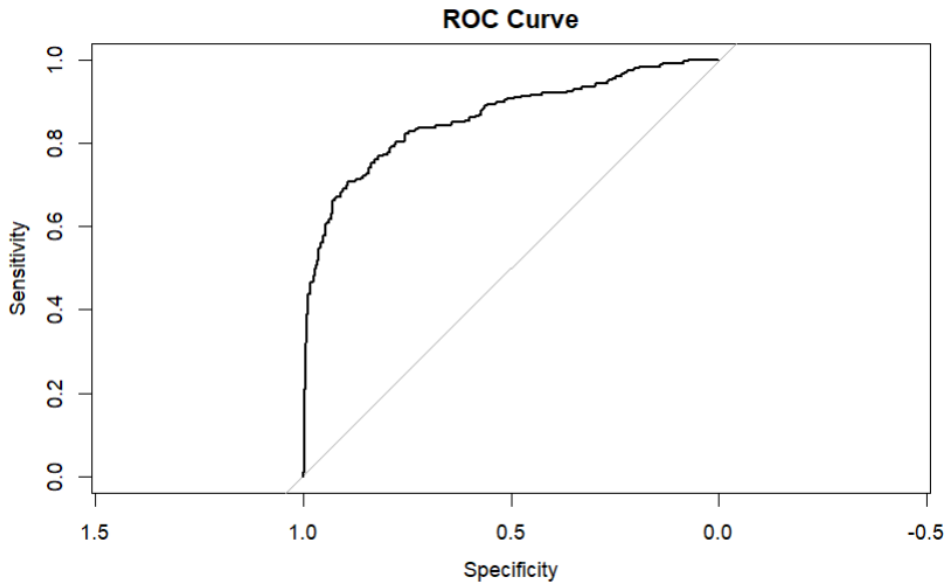
```
# Predict probabilities
predictions <- predict(model, type = "response")

# Convert probabilities to binary classification
predictions_class <- ifelse(predictions > 0.5, 1, 0)

# Evaluate Model Accuracy
confusion_matrix <- table(Predicted = predictions_class, Actual =
titanic_clean$Survived)
print(confusion_matrix)

accuracy <- mean(predictions_class == as.numeric(titanic_clean$Survived) - 1)
print(paste("Accuracy:", accuracy))

# ROC Curve
roc_curve <- roc(as.numeric(titanic_clean$Survived) - 1, predictions)
plot(roc_curve, main = "ROC Curve")
```

Output:

```
> # Load necessary libraries
> library(dplyr)
> library(titanic)
> library(pROC)
>
> # Load Titanic dataset
> data("titanic_train")
>
> # Check dataset structure
> head(titanic_train)
  PassengerId Survived Pclass                                                 Name    Sex Age
1           1        0      3                              Braund, Mr. Owen Harris   male  22
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38
3           3        1      3                               Heikkinen, Miss. Laina female  26
4           4        1      1       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35
5           5        0      3                             Allen, Mr. William Henry   male  35
6           6        0      3                                     Moran, Mr. James   male  NA
  SibSp Parch         Ticket    Fare Cabin Embarked
1     1     0      A/5 21171  7.2500           S
2     1     0       PC 17599 71.2833   C85       C
3     0     0 STON/O2. 3101282  7.9250           S
```

```
4    1    0         113803 53.1000  C123       S
5    0    0         373450  8.0500           S
6    0    0         330877  8.4583           Q
>
> # Data Cleaning: Removing rows with missing values
> titanic_clean <- titanic_train %>%
+   filter(!is.na(Age), !is.na(Embarked), !is.na(Sex), !is.na(Pclass))
>
> # Convert categorical variables to factors
> titanic_clean$Survived <- as.factor(titanic_clean$Survived)
> titanic_clean$Pclass <- as.factor(titanic_clean$Pclass)
> titanic_clean$Sex <- as.factor(titanic_clean$Sex)
> titanic_clean$Embarked <- as.factor(titanic_clean$Embarked)
>
> # Build Logistic Regression Model
> model <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,
+              data = titanic_clean, family = binomial)
>
> # Print model summary
> summary(model)

Call:
glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
    Fare + Embarked, family = binomial, data = titanic_clean)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  16.691979 607.920015   0.027 0.978095
Pclass2      -1.189637   0.329197  -3.614 0.000302 ***
Pclass3      -2.395220   0.343356  -6.976 3.04e-12 ***
Sexmale      -2.637859   0.223006 -11.829  < 2e-16 ***
Age          -0.043308   0.008322  -5.204 1.95e-07 ***
SibSp        -0.362925   0.129290  -2.807 0.005000 **
Parch        -0.060365   0.123944  -0.487 0.626233
Fare          0.001451   0.002595   0.559 0.576143
EmbarkedC   -12.259048 607.919885  -0.020 0.983911
EmbarkedQ   -13.082427 607.920088  -0.022 0.982831
EmbarkedS   -12.661895 607.919868  -0.021 0.983383
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 632.34  on 703  degrees of freedom
AIC: 654.34
```

```
Number of Fisher Scoring iterations: 13

>
> # Predict probabilities
> predictions <- predict(model, type = "response")
>
> # Convert probabilities to binary classification
> predictions_class <- ifelse(predictions > 0.5, 1, 0)
>
> # Evaluate Model Accuracy
> confusion_matrix <- table(Predicted = predictions_class, Actual = titanic_clean$Survived)
> print(confusion_matrix)
         Actual
Predicted   0   1
        0 365  83
        1  59 207
>
> accuracy <- mean(predictions_class == as.numeric(titanic_clean$Survived) - 1)
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.801120448179272"
>
> # ROC Curve
> roc_curve <- roc(as.numeric(titanic_clean$Survived) - 1, predictions)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot(roc_curve, main = "ROC Curve")
```



Practical: 7

Aim: Write a Python program to read data from a CSV file, perform simple data analysis, and generate basic insights. (Use Pandas is a Python library).

Code:

```
import pandas as pd
file_path = 'data.csv'
data = pd.read_csv(file_path)
print("First 5 rows of the dataset:")
print(data.head())
print("\nDataset Information:")
print(data.info())
print("\nSummary Statistical:")
print(data.describe())
if 'Category' in data.columns:
    print("\nUnique values in 'Category' column:")
    print(data['Category'].value_counts())
```

	A	B	C
1	Category	Sales	Profit
2	Electronics	300	40
3	Furniture	400	50
4	Clothing	350	60
5	Electronics	250	80
6	Clothing	200	70
7	Furniture	500	100

Output:

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

  warnings.warn(
PS C:\Users\Rashmi Pandey\Desktop\Logistic> & "C:/Program Files/Python313/python.exe" "c:/Users/Rashmi Pandey/Desktop/Logistic/data.py"
First 5 rows of the dataset:
      Category   Sales   Profit
0  Electronics     300       40
1    Furniture     400       50
2     Clothing     350       60
3  Electronics     250       80
4     Clothing     200       70

Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Category  6 non-null      object
 1    Sales    6 non-null      int64
 2    Profit   6 non-null      int64
dtypes: int64(2), object(1)
memory usage: 276.0+ bytes
None

Summary Statistical:
            Sales      Profit
count    6.000000    6.000000
mean   333.333333   66.666667
std    108.012345   21.602469
min    200.000000   40.000000
25%    262.500000   52.500000
50%    325.000000   65.000000
75%    387.500000   77.500000
max    500.000000  100.000000

Unique values in 'Category' column:
Category
Electronics    2
Furniture      2
Clothing       2
Name: count, dtype: int64
```

Practical: 8a

Aim: Perform data visualization using python on any sales data.

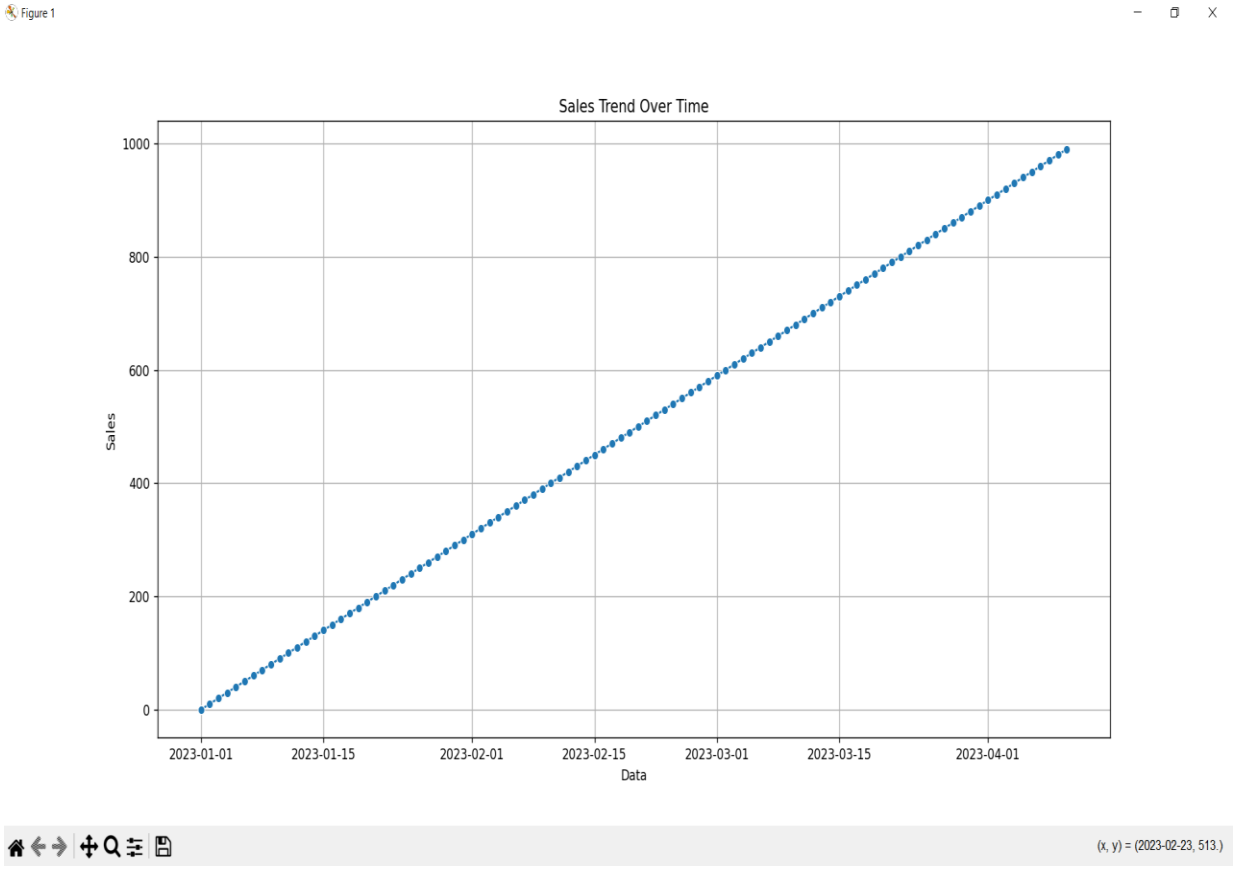
Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data = {
    'Data': pd.date_range(start='2023-01-01', periods=100, freq='D'),
    'Product': ['Product A', 'Product B', 'Product C', 'Product D']* 25,
    'Category': ['Electronics', 'Furniture', 'Clothing', 'Books']* 25,
    'Region': ['North', 'South', 'East', 'West']* 25,
    'Sales': [x * 10 for x in range(100)],
    'Profit': [x * 2 for x in range(100)],
}
df = pd.DataFrame(data)
print(df.head())
plt.figure(figsize=(10, 5))
sns.lineplot(data=df, x='Data', y='Sales', marker='o')
plt.title('Sales Trend Over Time')
plt.xlabel('Data')
plt.ylabel('Sales')
plt.grid(True)
plt.show()
plt.figure(figsize=(8, 5))
sns.barplot(data=df, x='Category', y='Sales', ci=None, palette='viridis')
plt.title('Sales by Category')
plt.xlabel('Category')
plt.ylabel('Total Sales')
plt.show()
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='Region', y='Profit', palette='coolwarm')
plt.title('Profit Distribution by Region')
plt.xlabel('Region')
plt.ylabel('Profit')
plt.show()
plt.figure(figsize=(8, 5))
product_sales = df.groupby('Product')['Sales'].sum().reset_index()
sns.barplot(data=product_sales, x='Product', y='Sales', palette='magma')
plt.title('Total Sales by Product')
plt.xlabel('Product')
```

```
plt.ylabel('Sales')
plt.show()
```

Output:

```
        Data    Product     Category Region  Sales  Profit
0 2023-01-01  Product A  Electronics  North      0       0
1 2023-01-02  Product B    Furniture  South     10       2
2 2023-01-03  Product C     Clothing   East     20       4
3 2023-01-04  Product D        Books   West     30       6
4 2023-01-05  Product A  Electronics  North     40       8
```



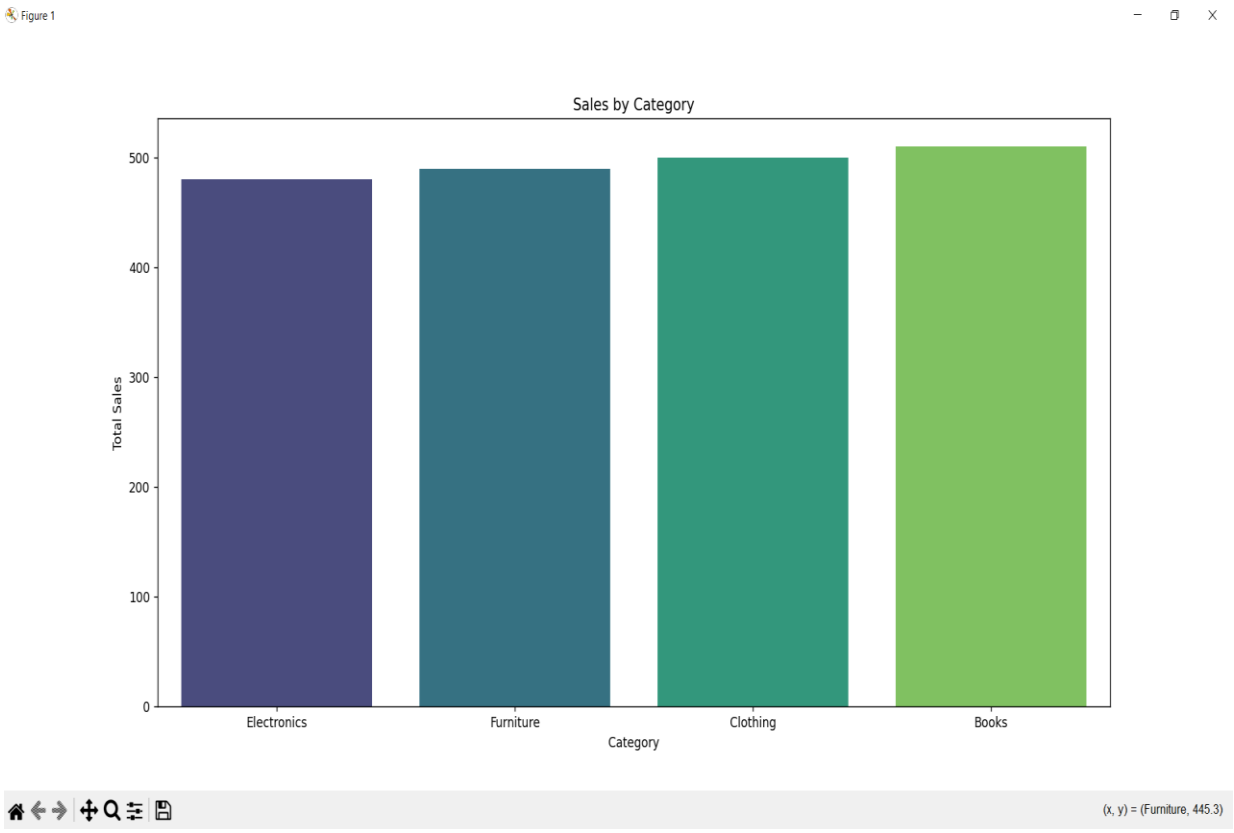


Figure 1
Sales by Category
Total Sales
0
100
200
300
400
500
Electronics
Furniture
Clothing
Books
Category
(x, y) = (Furniture, 445.3)

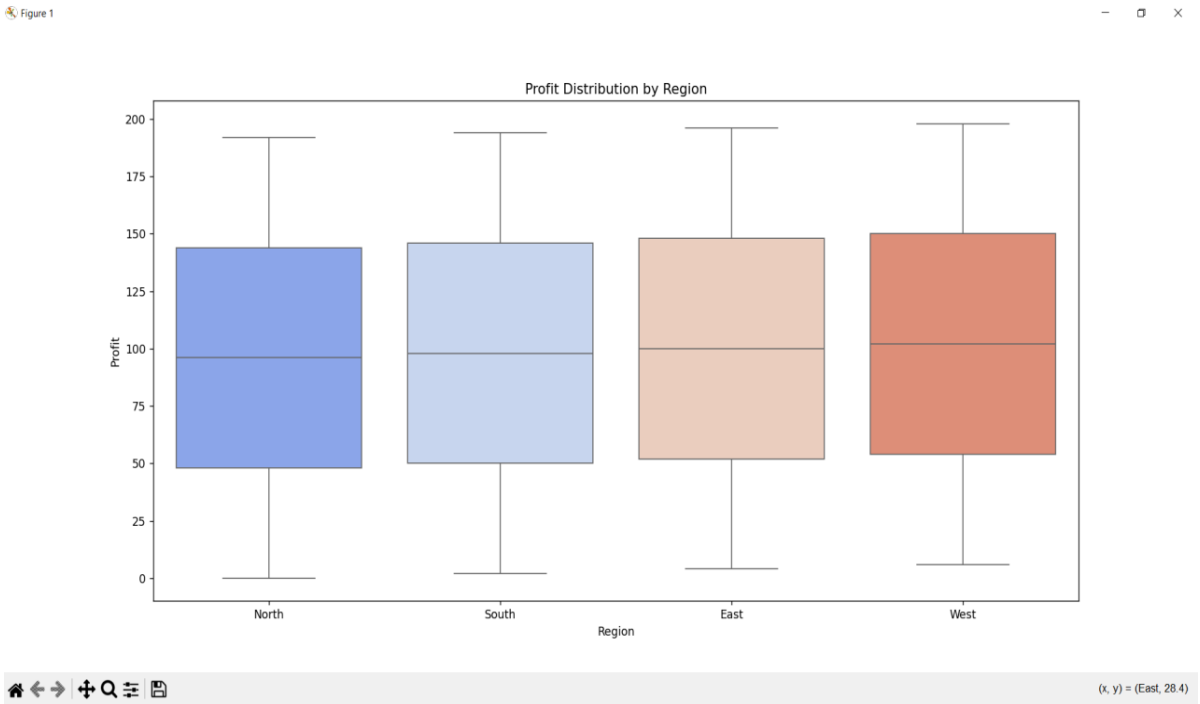


Figure 1
Profit Distribution by Region
Profit
0
25
50
75
100
125
150
175
200
North
South
East
West
Region
(x, y) = (East, 28.4)

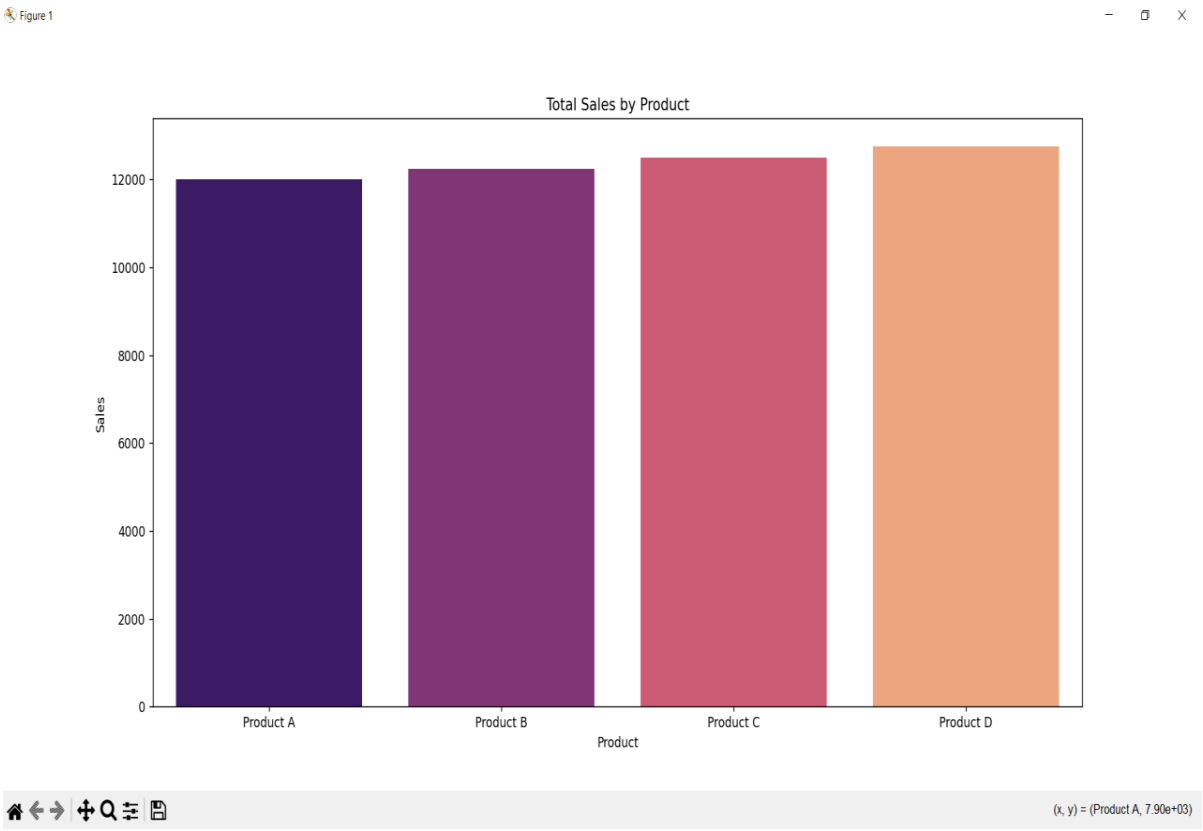


Figure 1
Total Sales by Product
Sales
12000
10000
8000
6000
4000
2000
0
Product A
Product B
Product C
Product D
Product
(x, y) = (Product A, 7.90e+03)

Practical: 8b

Aim: Perform data visualization using PowerBI on any sales data.

Steps:-

Step 1:- Download & Install Power BI

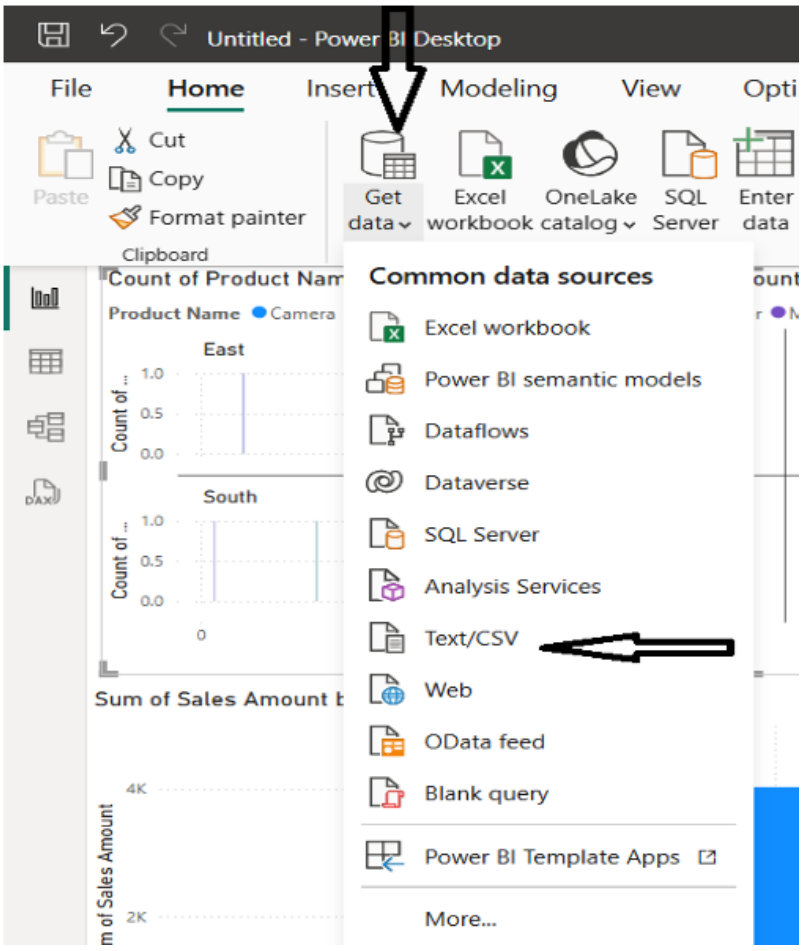
Step 2:- create a sample data

	A	B	C	D	E	F	G	H	I	J	K
1	**Order ID**	**Date**	**Product Name**	**Category**	**Sales Amount**	**Quantity Sold**	**Customer Name**	**Customer Region**	**Payment Method**	**Discount (%)**	**Profit**
2	ORD1001	01-01-2024	Laptop	Electronics	1200	2	John Doe	New York	Credit Card	10	300
3	ORD1002	02-01-2024	Smartphone	Electronics	800	1	Alice Smith	California	Debit Card	5	150
4	ORD1003	03-01-2024	Shoes	Fashion	100	3	Robert Brown	Texas	Cash	15	30
5	ORD1004	04-01-2024	Washing Machine	Home Appliances	500	1	Emily Davis	Florida	Net Banking	8	80
6	ORD1005	05-01-2024	Tablet	Electronics	300	2	Michael Lee	Illinois	UPI	10	60
7	ORD1006	06-01-2024	Headphones	Accessories	50	4	Sophia Wilson	New York	Credit Card	5	15
8	ORD1007	07-01-2024	Refrigerator	Home Appliances	900	1	Daniel Scott	Texas	Debit Card	12	180

Step 3 :- Import the Sales Dataset

- Click "Home" > "Get Data".
- Choose your data source:
    - o Excel (XLSX/CSV)
    - o SQL Server
    - o Online Services (Google Sheets, SharePoint, etc.)
- Browse and select the dataset.

- Click "Load" to import.



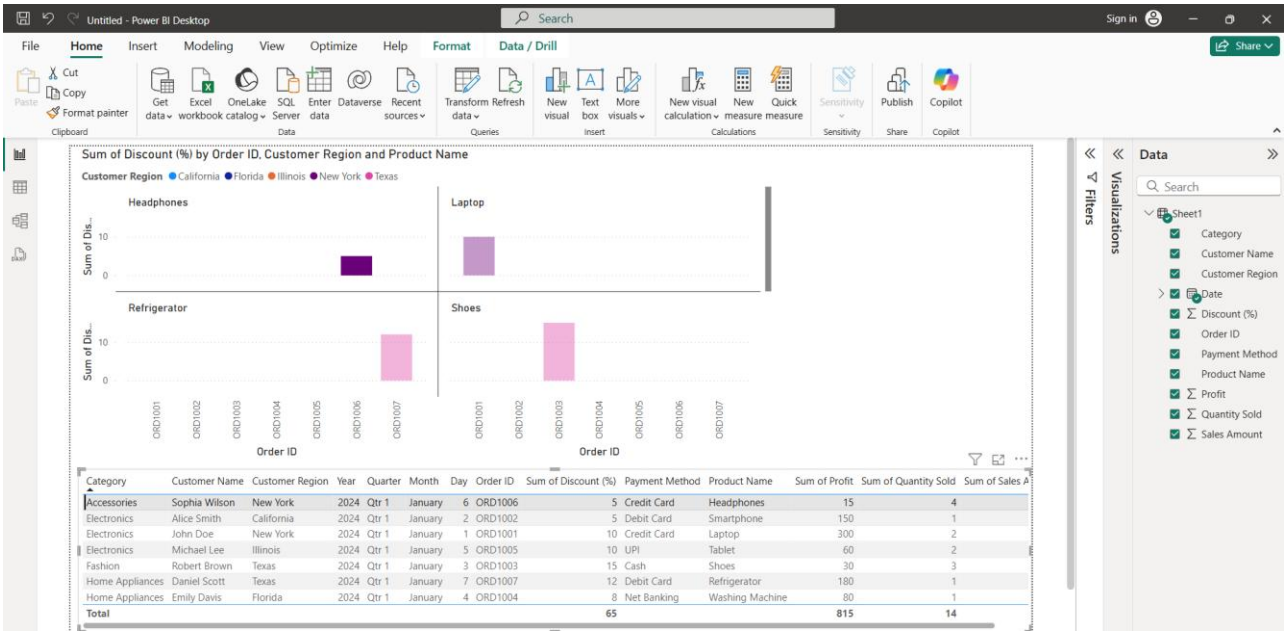
Step 4 :- Clean & Transform Data (Power Query)

Step 5 :- Create Data Visualizations

- Click on "Line Chart" in the "Visualizations" pane.
- Drag Order Date to the X-axis.
- Drag Sales Amount to the Y-axis.
- Customize the chart (format labels, add title, etc.).

Step 6:- Add Filters & Interactivity

Output:-



Category	Customer Name	Customer Region	Year	Quarter	Month	Day	Order ID	Sum of Discount (%)	Payment Method	Product Name	Sum of Profit	Sum of Quantity Sold	Sum of Sales A
Accessories	Sophia Wilson	New York	2024	Qtr 1	January	6	ORD1006	5	Credit Card	Headphones	15	4	
Electronics	Alice Smith	California	2024	Qtr 1	January	2	ORD1002	5	Debit Card	Smartphone	150	1	
Electronics	John Doe	New York	2024	Qtr 1	January	1	ORD1001	10	Credit Card	Laptop	300	2	
Electronics	Michael Lee	Illinois	2024	Qtr 1	January	5	ORD1005	10	UPI	Tablet	60	2	
Fashion	Robert Brown	Texas	2024	Qtr 1	January	3	ORD1003	15	Cash	Shoes	30	3	
Home Appliances	Daniel Scott	Texas	2024	Qtr 1	January	7	ORD1007	12	Debit Card	Refrigerator	180	1	
Home Appliances	Emily Davis	Florida	2024	Qtr 1	January	4	ORD1004	8	Net Banking	Washing Machine	80	1	
Total								65			815	14	

Practical: 9

Aim: Create the Data staging area for the selected database using SQL

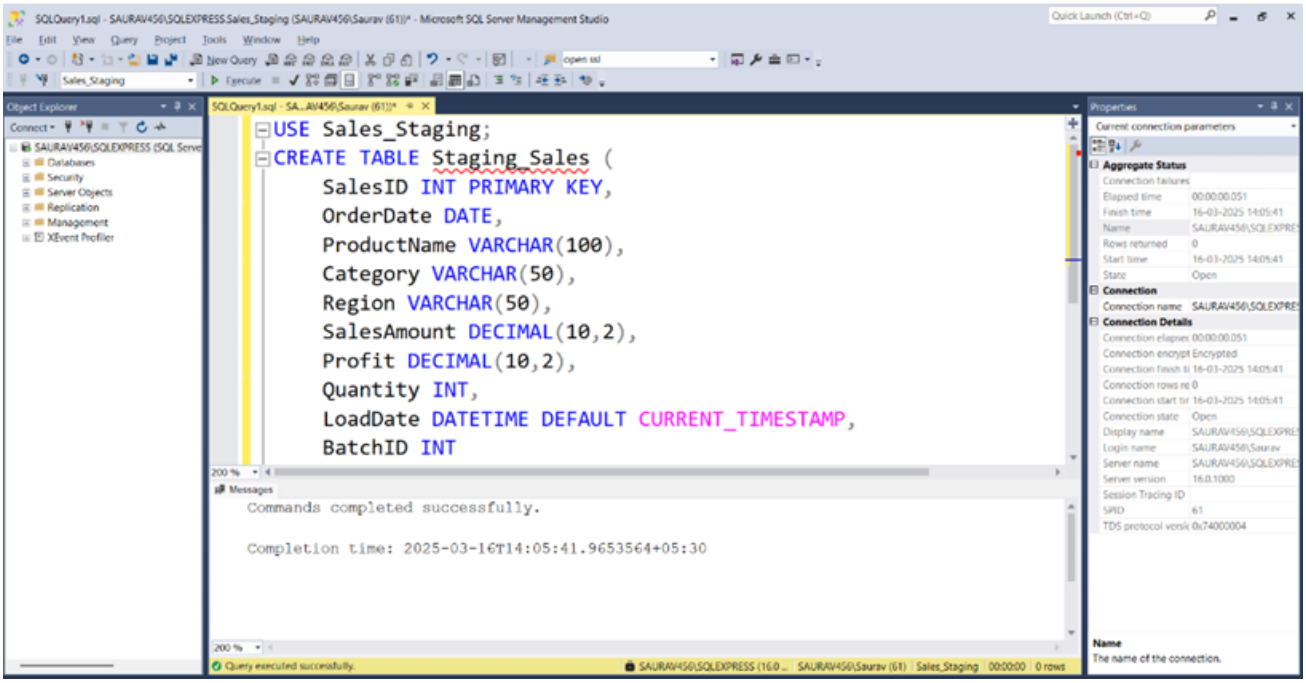
STEPS:-

Step 1:- Create a Staging Database First, create a staging database to store raw sales data.

```
CREATE DATABASE Sales_Staging;

USE Sales_Staging;
```

Step 2:- Create Staging Tables Create tables that match the structure of raw sales data but include additional fields like load date and batch ID.

```
CREATE TABLE Staging_Sales (

SalesID INT PRIMARY KEY,

OrderDate DATE,

ProductName VARCHAR(100),

Category VARCHAR(50),

Region VARCHAR(50),

SalesAmount DECIMAL(10,2),

Profit DECIMAL(10,2),

Quantity INT,

LoadDate DATETIME DEFAULT CURRENT_TIMESTAMP,

BatchID INT );
```

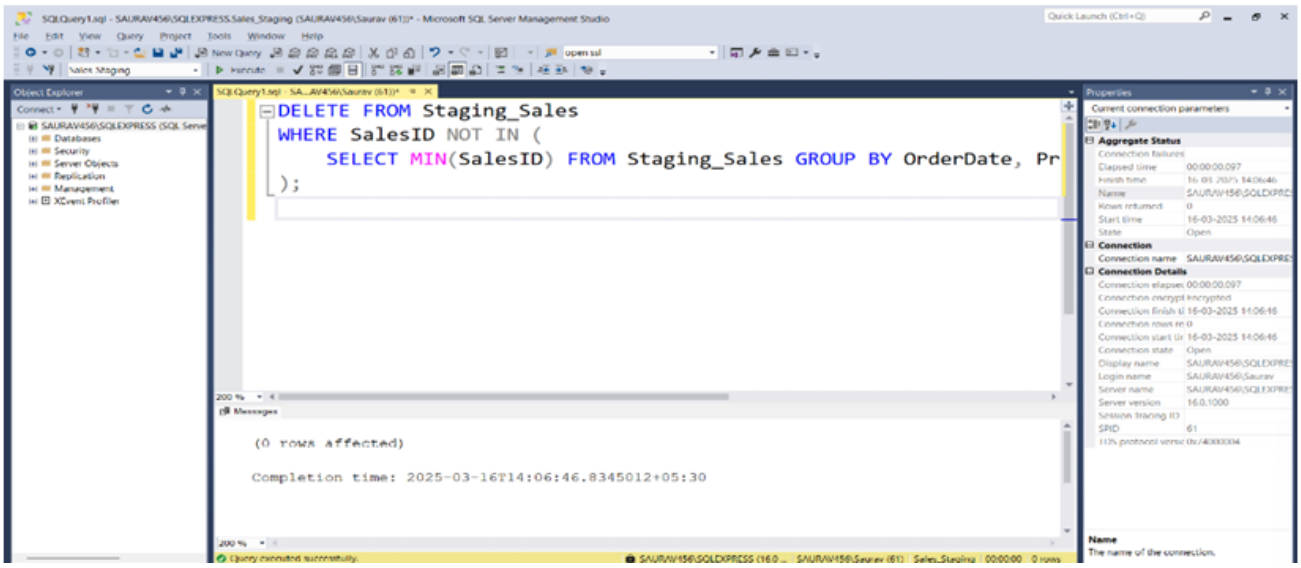


Step 4:- Perform Data Cleansing & Transformation

• Remove Duplicates DELETE FROM Staging_Sales

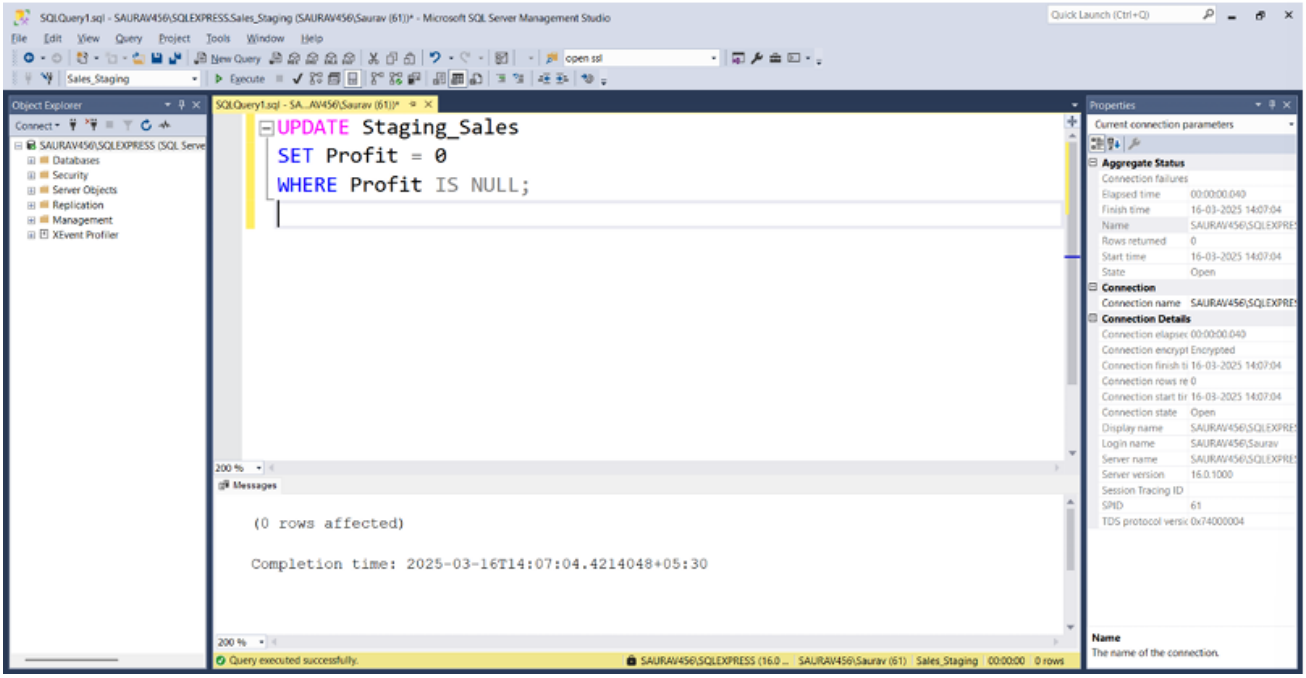
WHERE SalesID NOT IN (

SELECT MIN(SalesID) FROM Staging_Sales GROUP BY OrderDate, ProductName, Region);



- **Handle Null Values**

```
UPDATE Staging_Sales
SET Profit = 0
WHERE Profit IS NULL;
```



Output:

Results | Messages

	SalesID	OrderDate	ProductName	Category	Region	SalesAmount	Profit	Quantity	LoadDate	BatchID
1	1	2024-01-01	Laptop	Electronics	North	1200.00	200.00	3	2025-03-21 13:49:51.297	101
2	2	2024-01-02	Smartphone	Electronics	South	800.00	150.00	2	2025-03-21 13:49:51.297	101
3	3	2024-01-03	Tablet	Electronics	East	600.00	100.00	5	2025-03-21 13:49:51.297	101