**SAJINE SANTHAKUMAR 240701459**



Department of Computer Science and Engineering

CS23334 Fundamentals of Data Science Lab III semester II Year (2023R)

Name of the Student : SAJINE SANTHAKUMAR

Register Number : 240701459

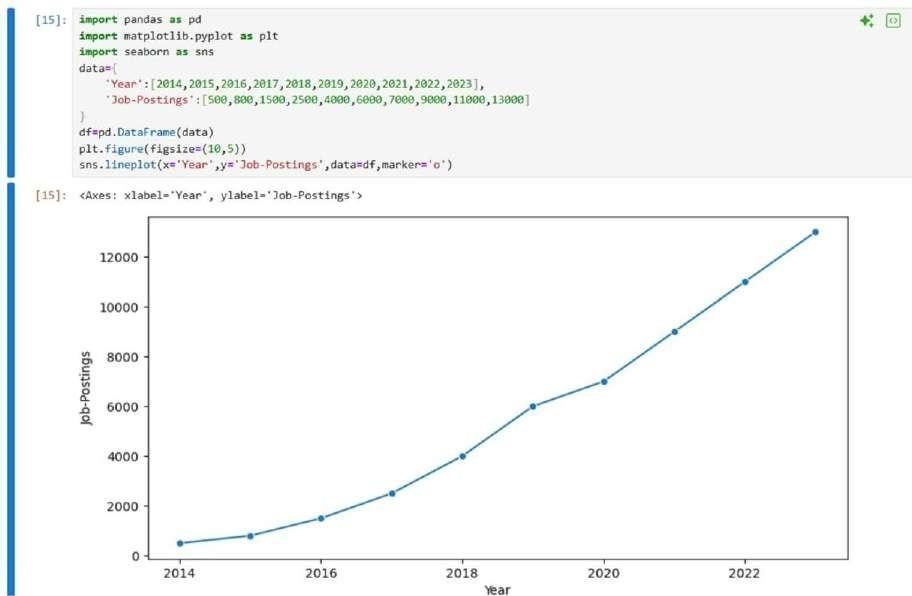
EXPERIMENT NO: 1A

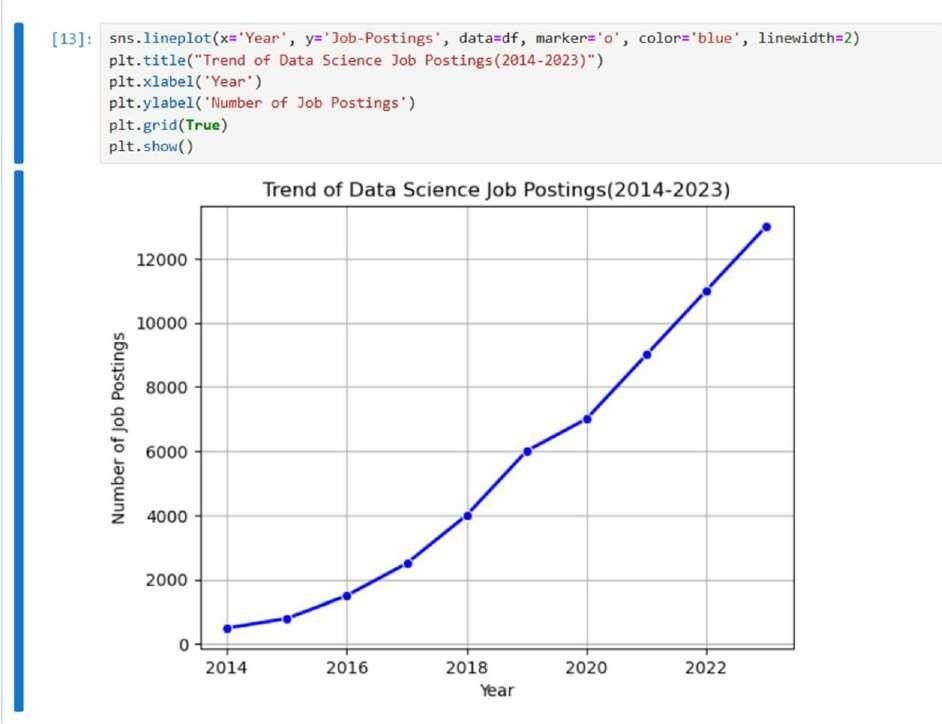
**Analyze the trend of data science job postings over the last decade**

Aim:

To visualize the **trend of Data Science job postings from 2014 to 2023** using a line graph in Python with the help of **Pandas**, **Matplotlib**, and **Seaborn** libraries.

Program:





Result:

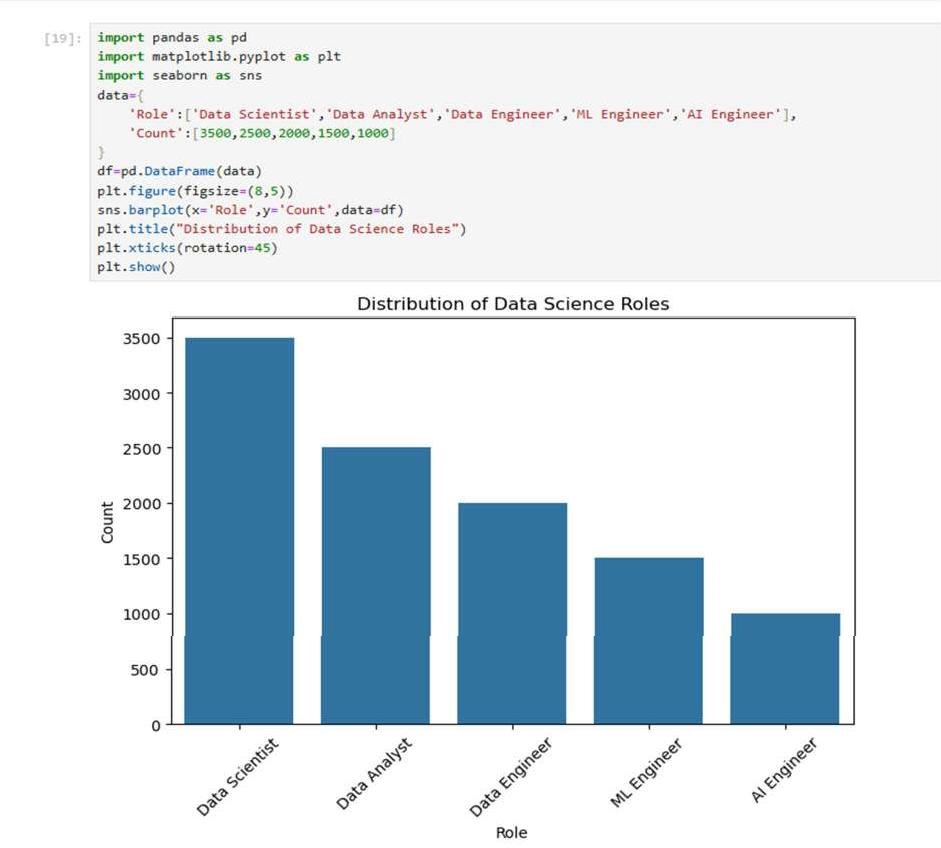
The program successfully displays a **line graph** showing a **steady increase in Data Science job postings** from **2014 (500)** to **2023 (13,000)**.

EXPERIMENT NO: 1B

**Analyze and visualize the distribution of various data science roles** Aim:

To visualize the distribution of different **Data Science roles** and their **counts** using bar and pie charts with Python libraries — **Pandas**, **Matplotlib**, and **Seaborn**.

Program



Result:

The program successfully displays a bar chart and a pie chart showing that **Data Scientist** has the highest count and **AI Engineer** has the lowest count among the data science roles.

EXPERIMENT NO: 1C

**Conduct an experiment to differentiate Structured, Un-structured and Semi structured data**

Aim:

To demonstrate the three types of data — **structured, semi-structured, and unstructured** — and display them in Python.

Program:



Result:

Data can be **structured**, **semi-structured**, or **unstructured**. Structured data is organized in tables, semi-structured data has some organization like JSON, and unstructured data is plain text without a fixed format.

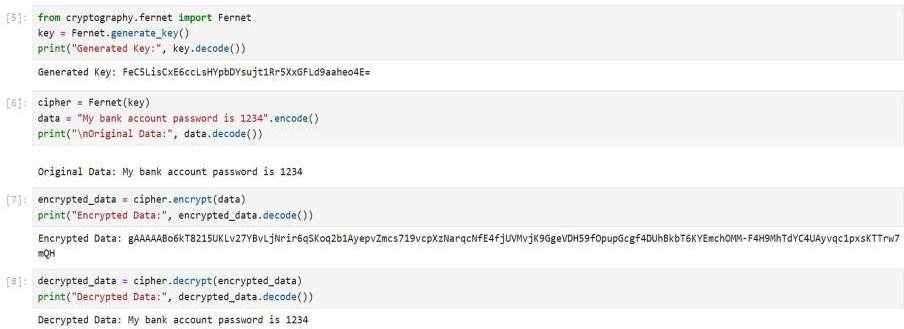
EXPERIMENT NO: 1D

**Conduct an experiment to encrypt and decrypt given sensitive data.**

Aim:

To securely **encrypt and decrypt a sensitive message** (like a password) using symmetric encryption with Fernet from the cryptography library.

Program:



Result:

The program aims to securely encrypt and decrypt a message using Fernet symmetric encryption. It generates a key, converts the message into bytes, encrypts it into unreadable data, and then decrypts it back to the original message. The result shows that the original message is safely restored after encryption.

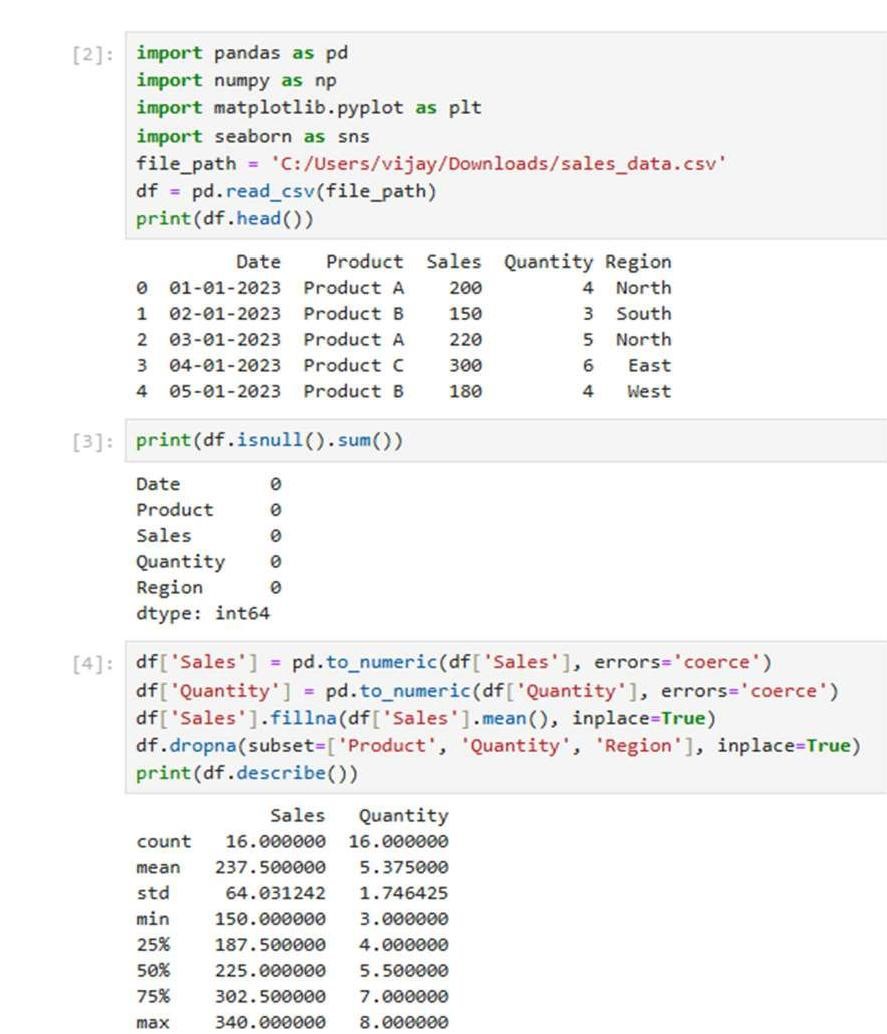
EXPERIMENT NO: 2

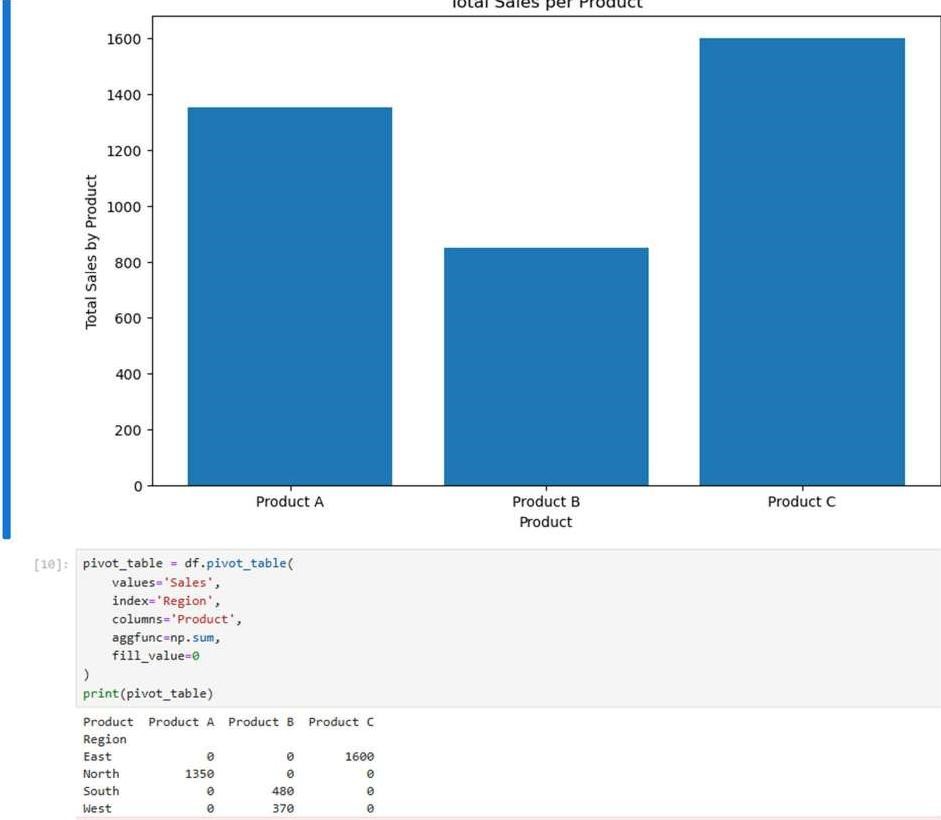
**Upload and Analyze the data set given in csv format and perform data preprocessing and visualization**

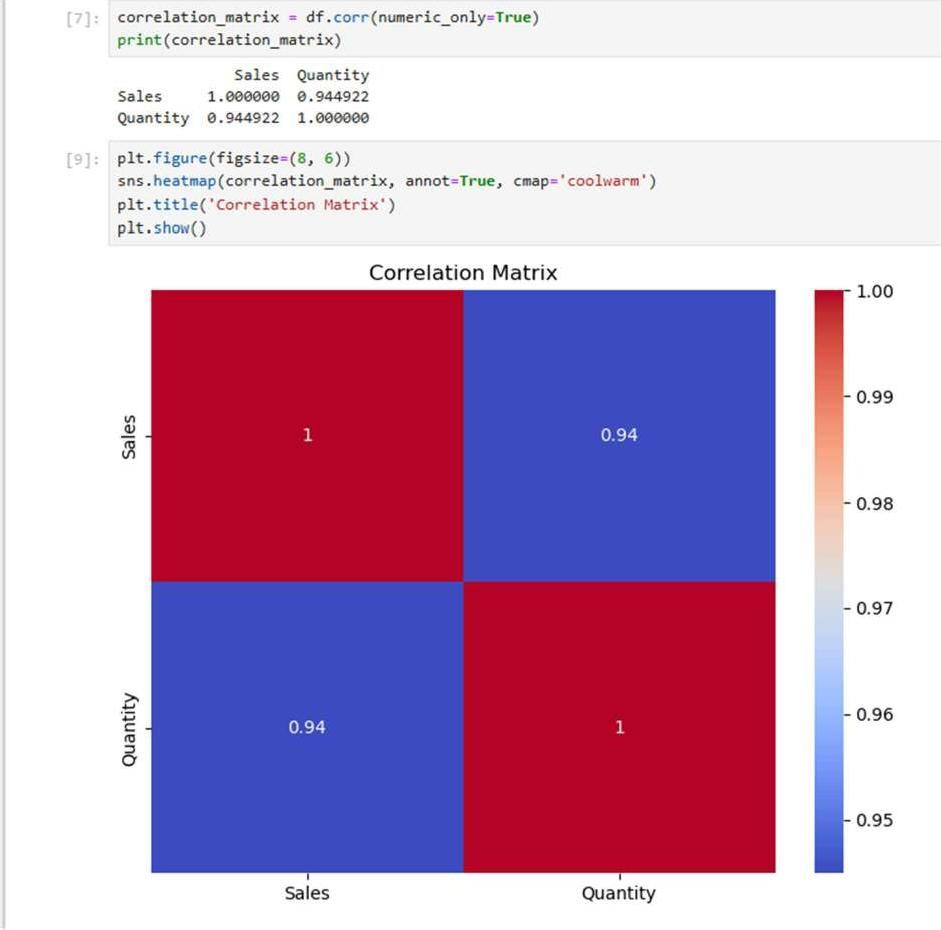
Aim:

To analyze and visualize sales data, clean missing values, summarize total sales and quantities per product, and examine correlations between numeric variables.

Program:







Result:

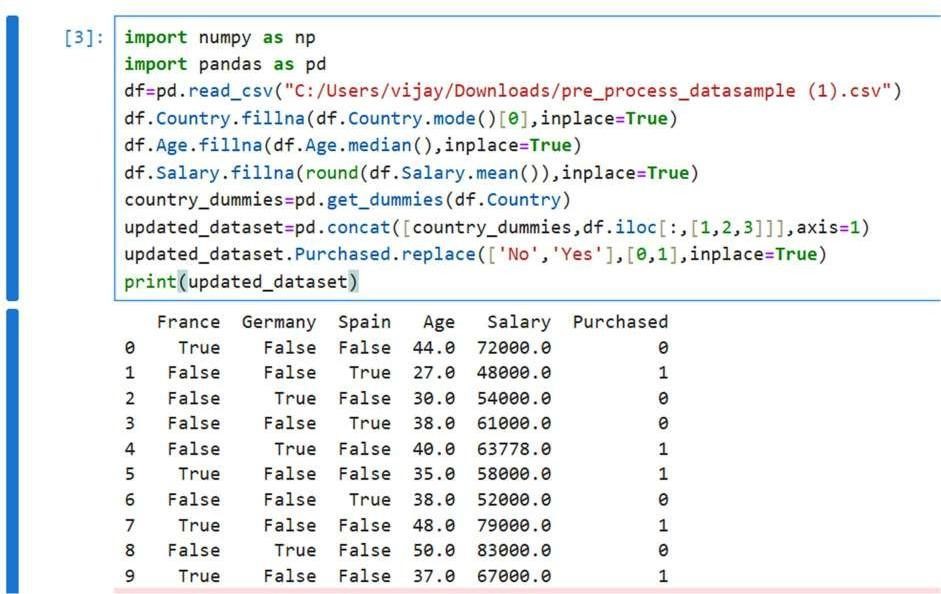
The dataset was cleaned, total sales and quantities per product were calculated, the bar chart highlighted top-selling products, the pivot table showed regional sales distribution, and the correlation matrix revealed a strong positive relationship between quantity and sales.

EXPERIMENT NO: 3A

**Data Preprocessing and Encoding for Machine Learning** Aim:

To preprocess a dataset by handling missing values, encoding categorical variables, and preparing it for analysis or machine learning.

Program:



Result:

A cleaned and transformed dataset where all missing values are replaced, categorical variables are converted into numeric form, and the Purchased column is ready for modeling.

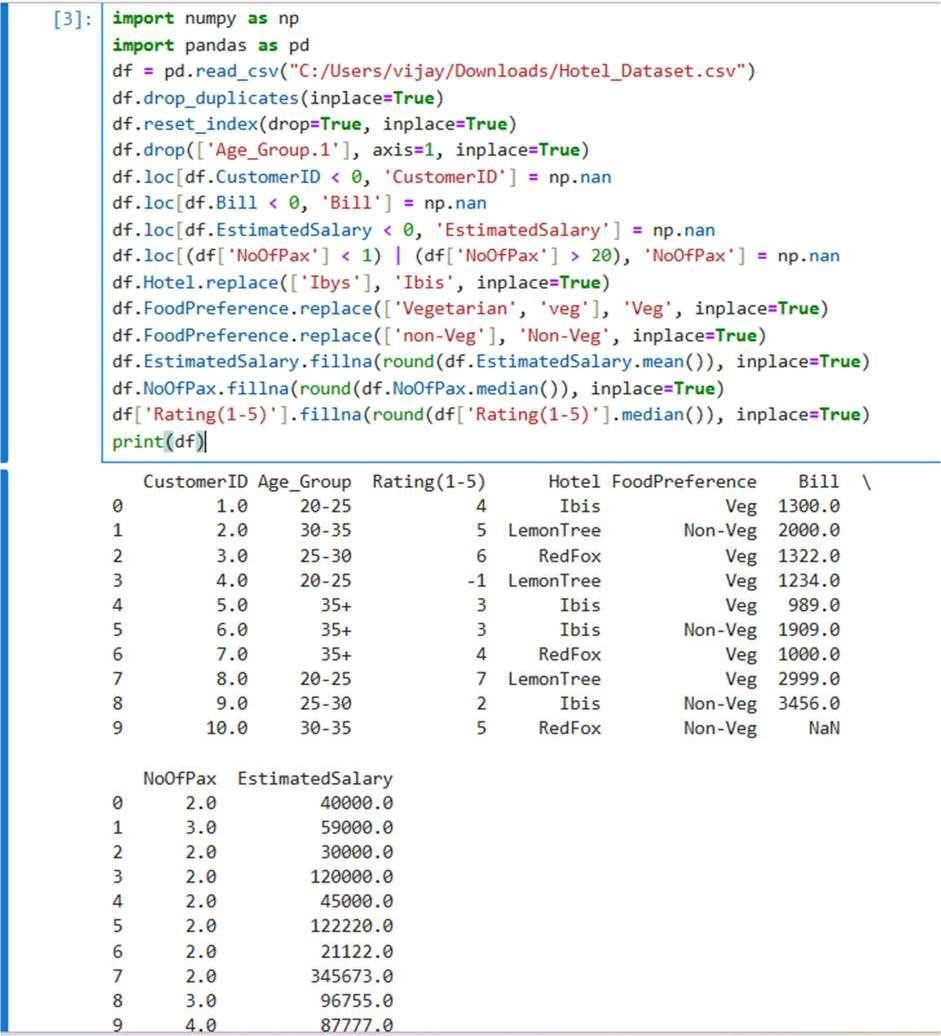
Each row now has numeric values only, suitable for machine learning or statistical analysis.

EXPERIMENT NO: 3B

**Hotel Dataset Cleaning and Preprocessing** Aim:

To clean and preprocess the hotel dataset by removing duplicates, correcting invalid and inconsistent values, filling missing data, and standardizing text entries, making it ready for analysis or machine learning.

Program:



Result:

A cleaned and consistent hotel dataset with no duplicates, corrected text values, invalid numerical entries replaced, and missing values filled, ready for analysis or modeling.

EXPERIMENT NO: 3C

**Airline Flight Dataset Generation and CSV Export** Aim:

To generate a synthetic dataset of airline flights with random flight details (IDs, airlines, source and destination cities, departure/arrival times, durations, and fares) and export it as a CSV file for analysis or testing purposes.

Program:



Result:

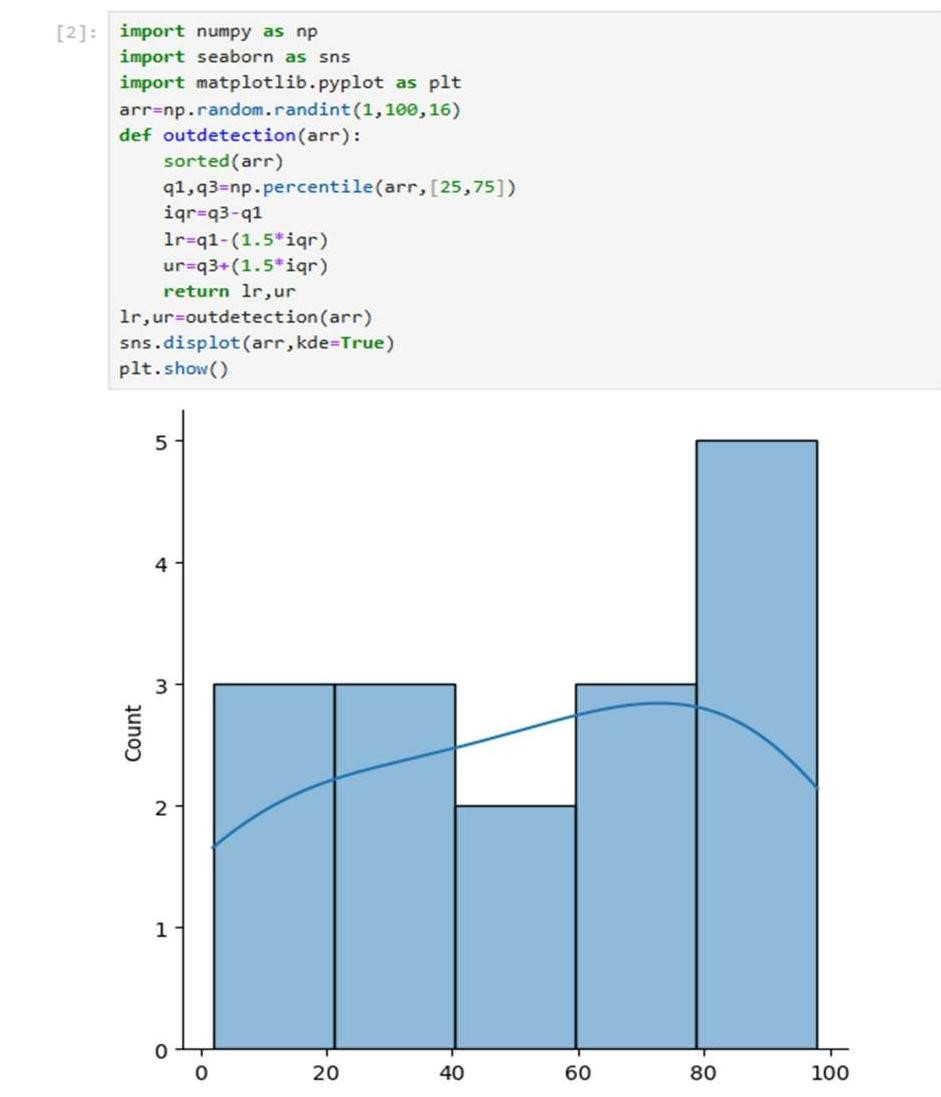
A CSV file "airline.flights" containing 15 randomly generated flight records with columns: Flight\_ID, AirLine, Source, Destination, Departure\_time, Arrival\_time, Duration, and Fare. The dataset can be used for analysis, testing, or simulations.

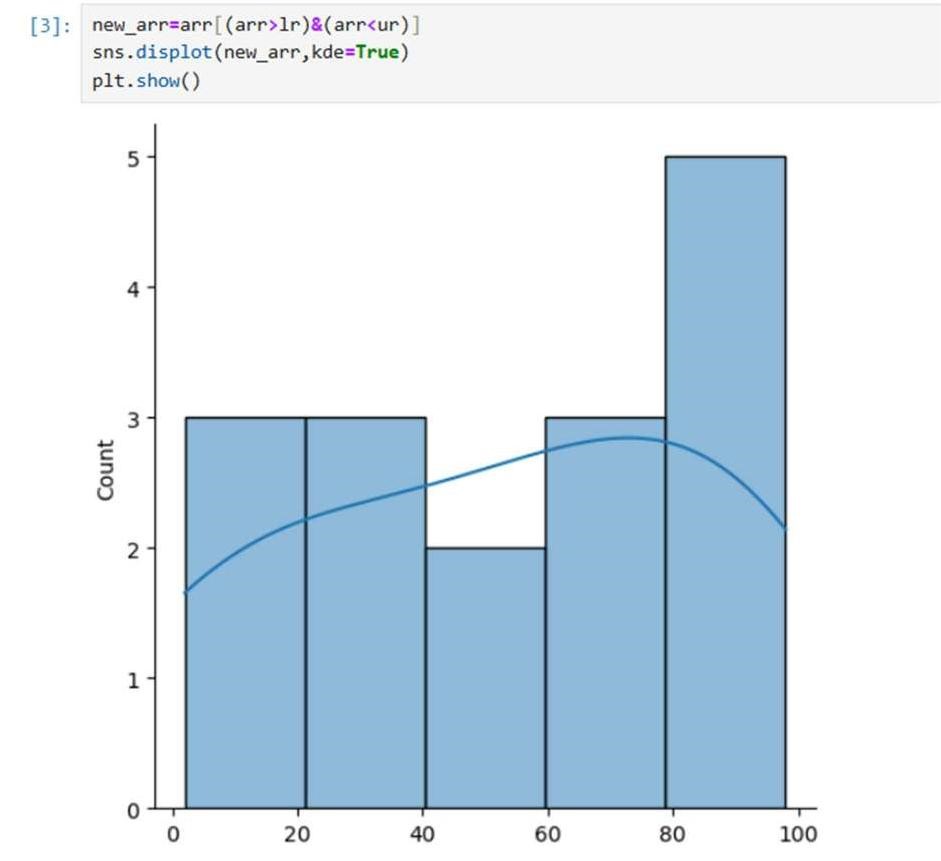
EXPERIMENT NO: 4

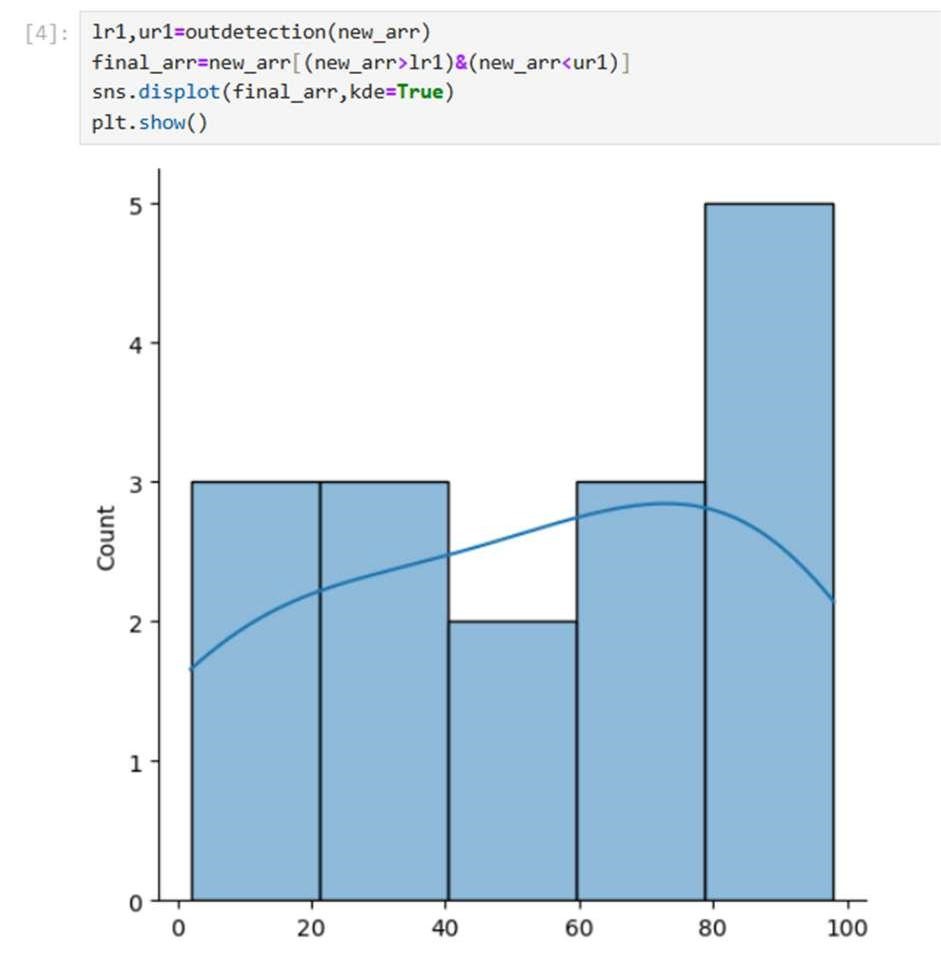
**Outlier Detection and Visualization using IQR** Aim:

To detect and remove outliers from a dataset using the Interquartile Range (IQR) method and visualize the data distribution before and after outlier removal.

Program:







Result:

The code removes extreme values from a random dataset using IQR, showing three plots: original data, after first outlier removal, and after final outlier removal, resulting in a cleaner, more representative dataset.

EXPERIMENT NO: 5

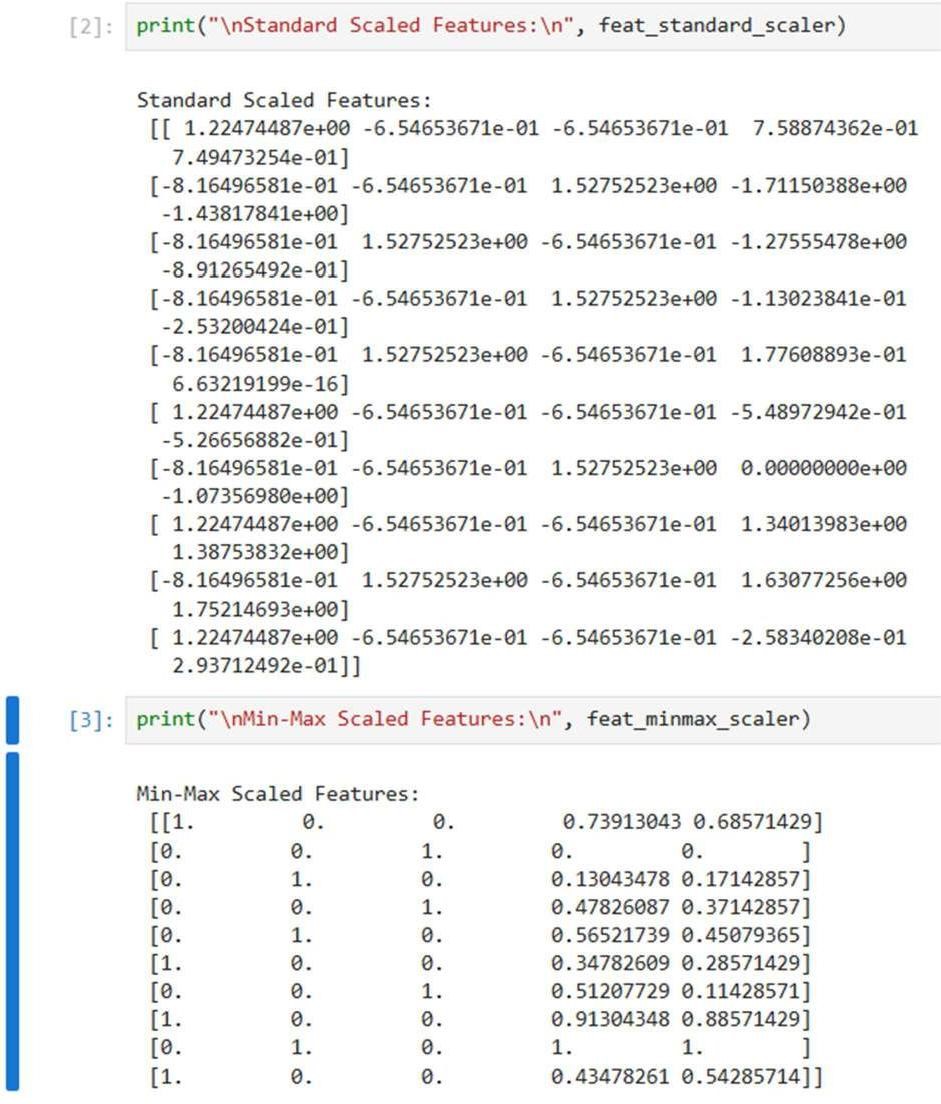
**Experiment to understand feature scaling.**

Aim:

To preprocess a dataset by handling missing values, encoding categorical variables, and scaling features using StandardScaler and MinMaxScaler.

Program:





Result:

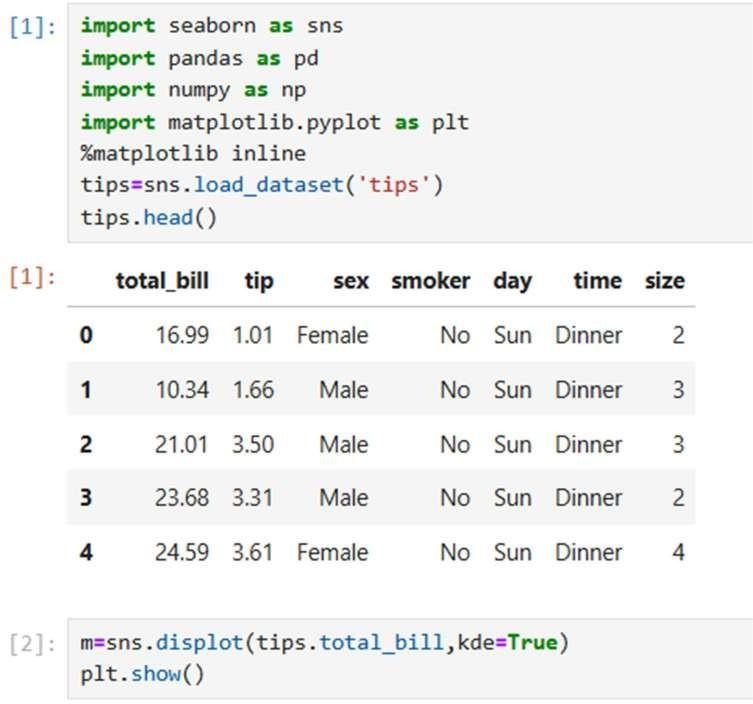
The dataset is cleaned and numeric features are combined with one-hot encoded Country. Standard scaling centers the data, while Min-Max scaling normalizes it between 0 and 1, making it ready for machine learning.

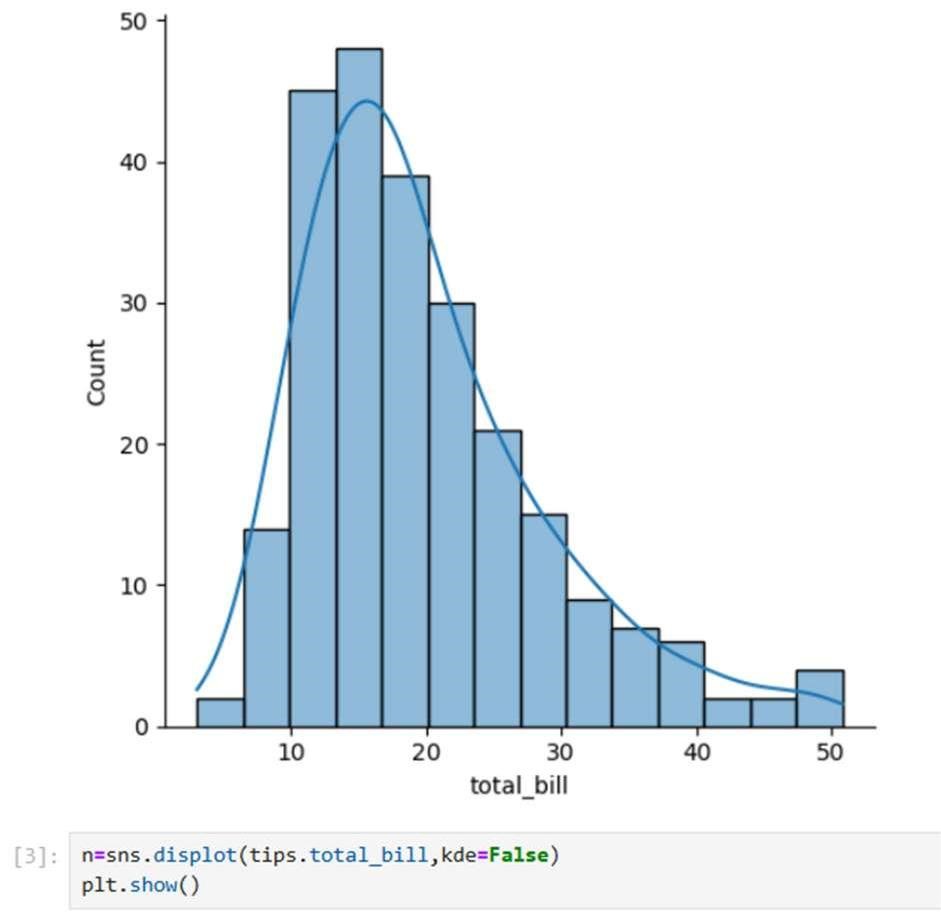
EXPERIMENT NO: 5

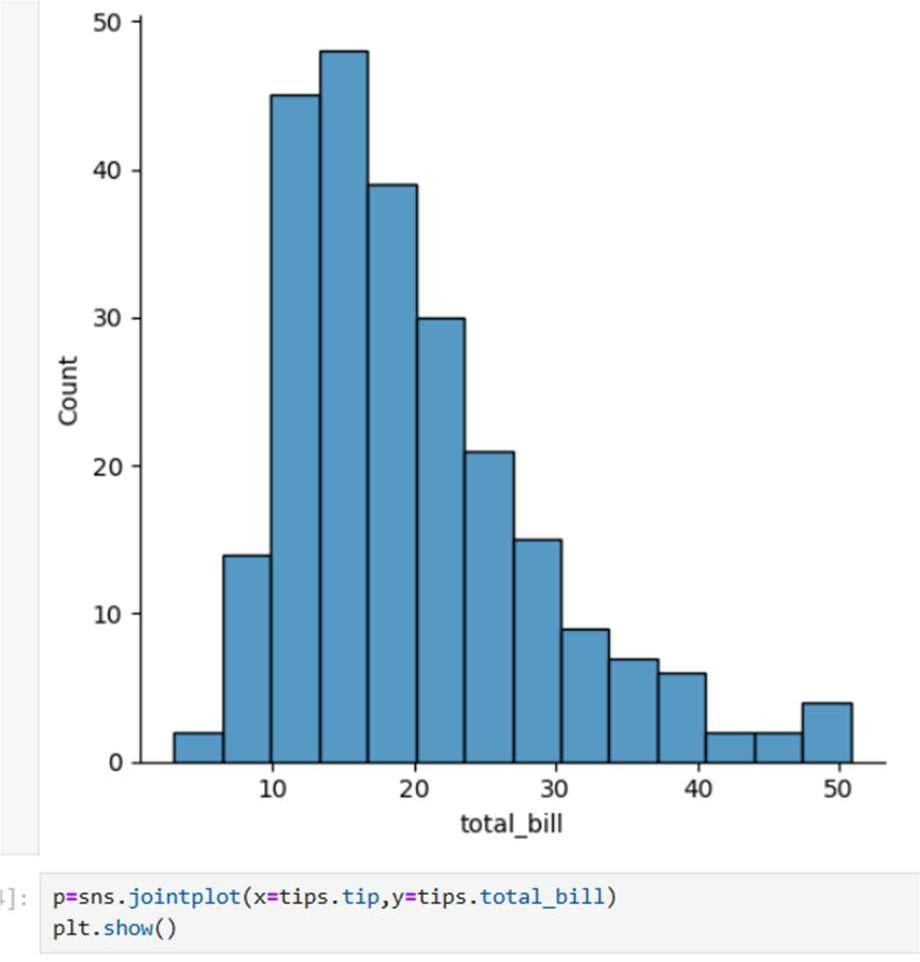
**Data Visualization using Seaborn** Aim:

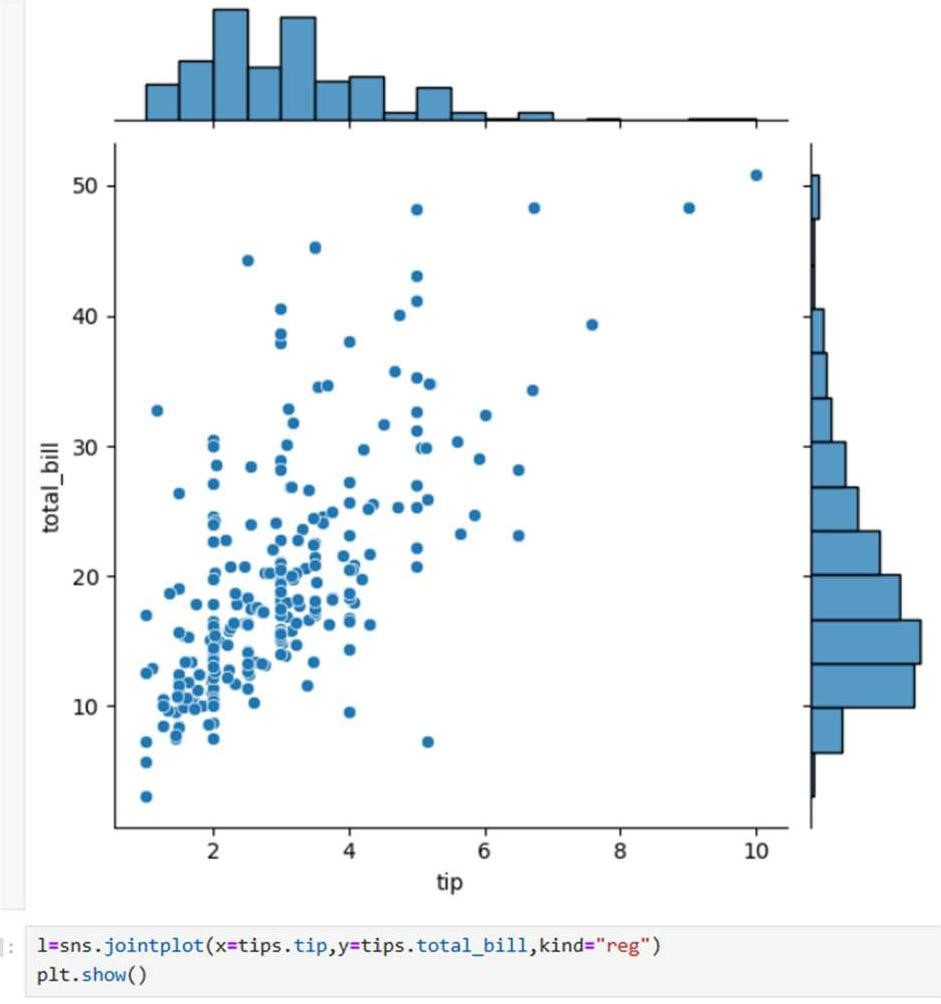
To explore and visualize the tips dataset using various plots to understand distributions, relationships, correlations, and categorical counts.

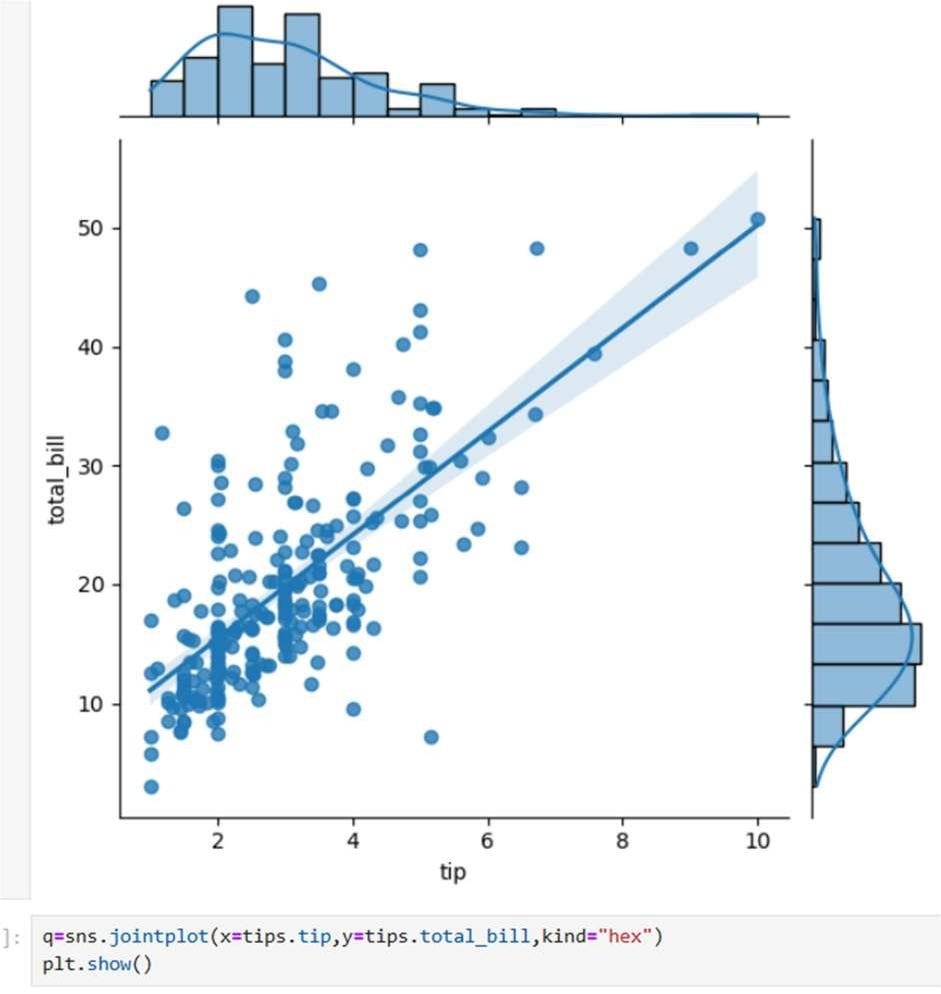
Program:

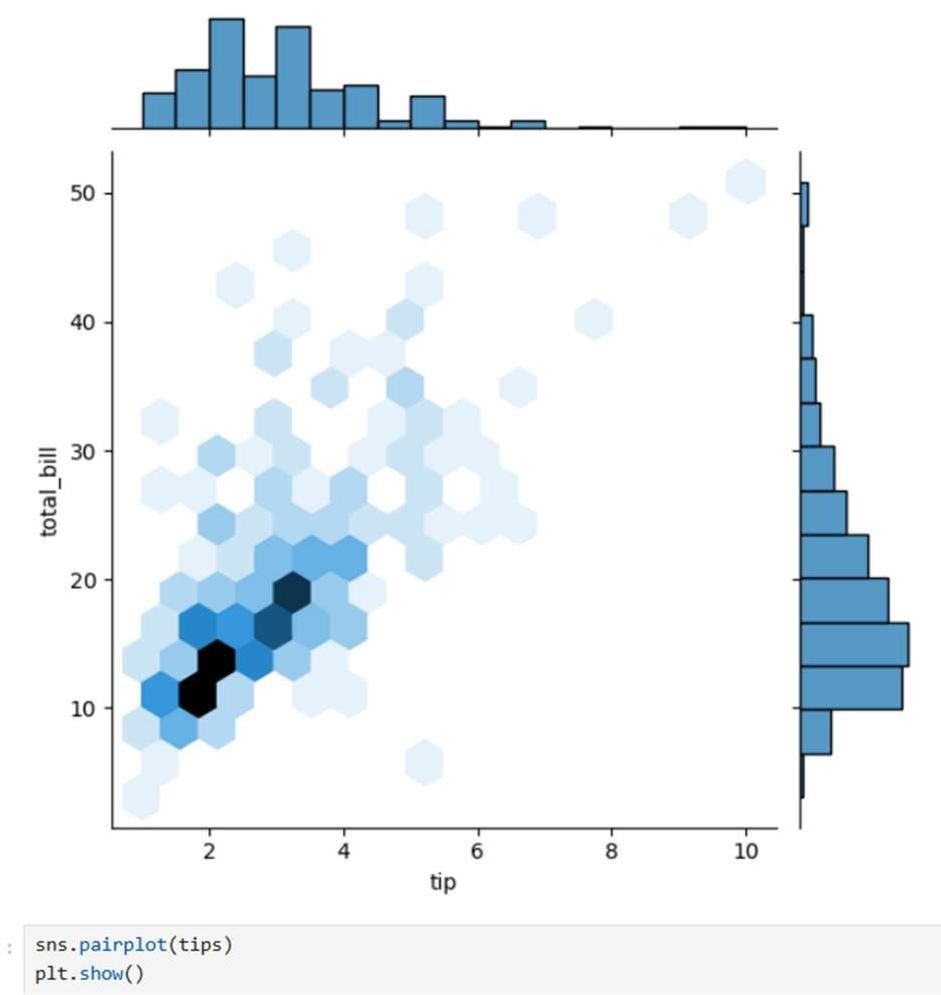


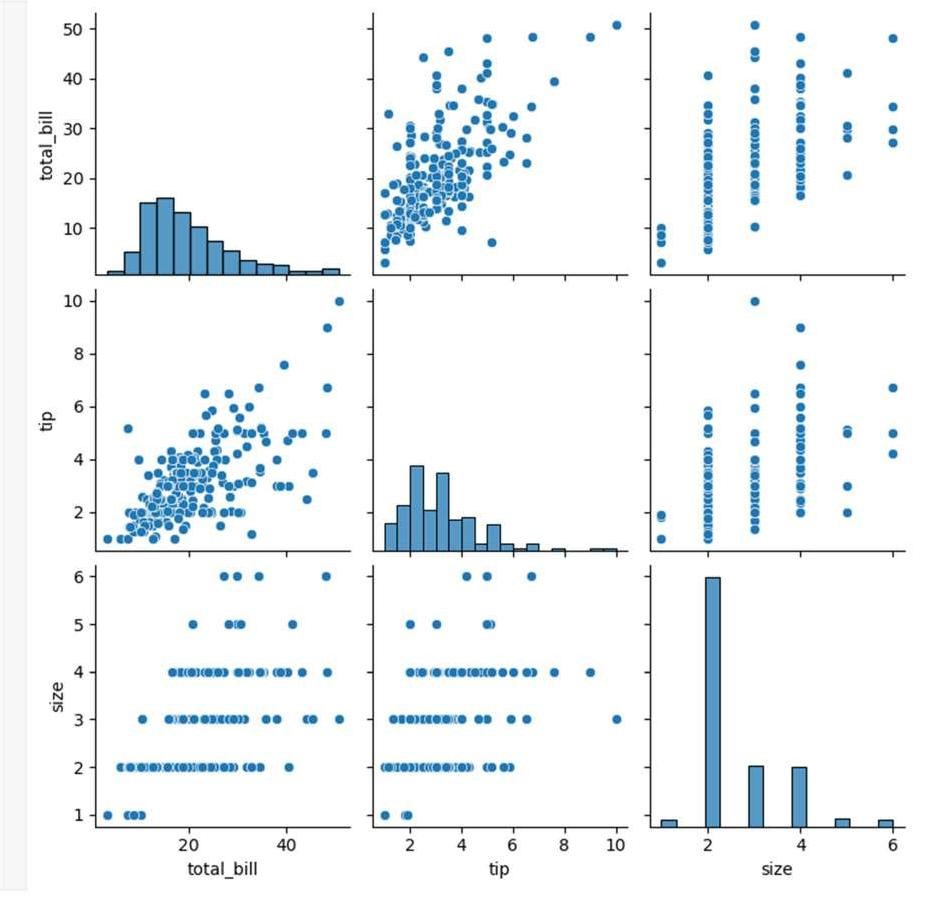
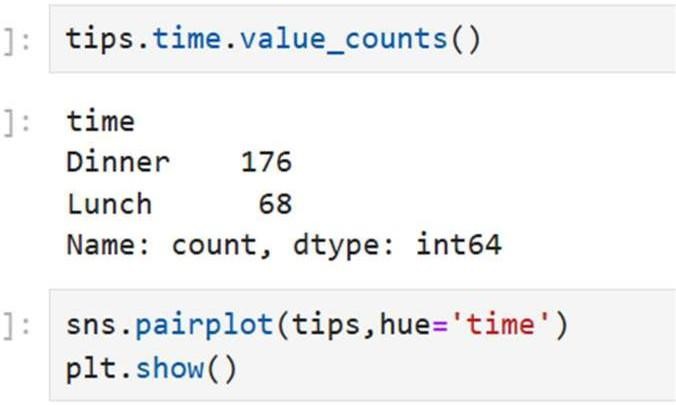


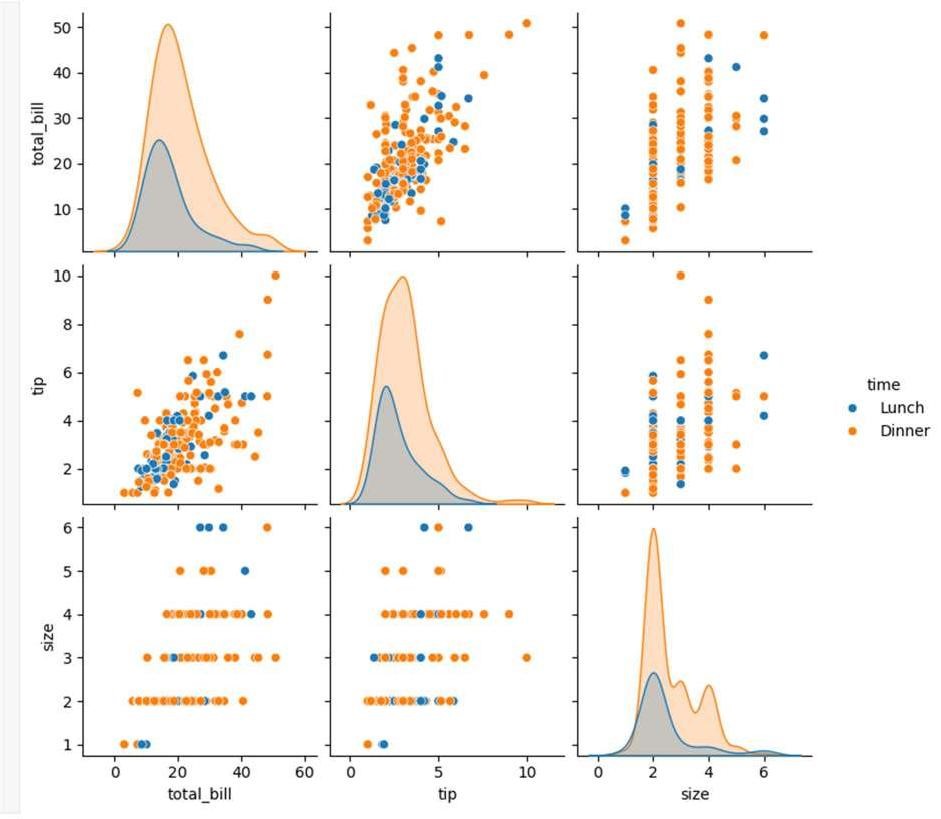


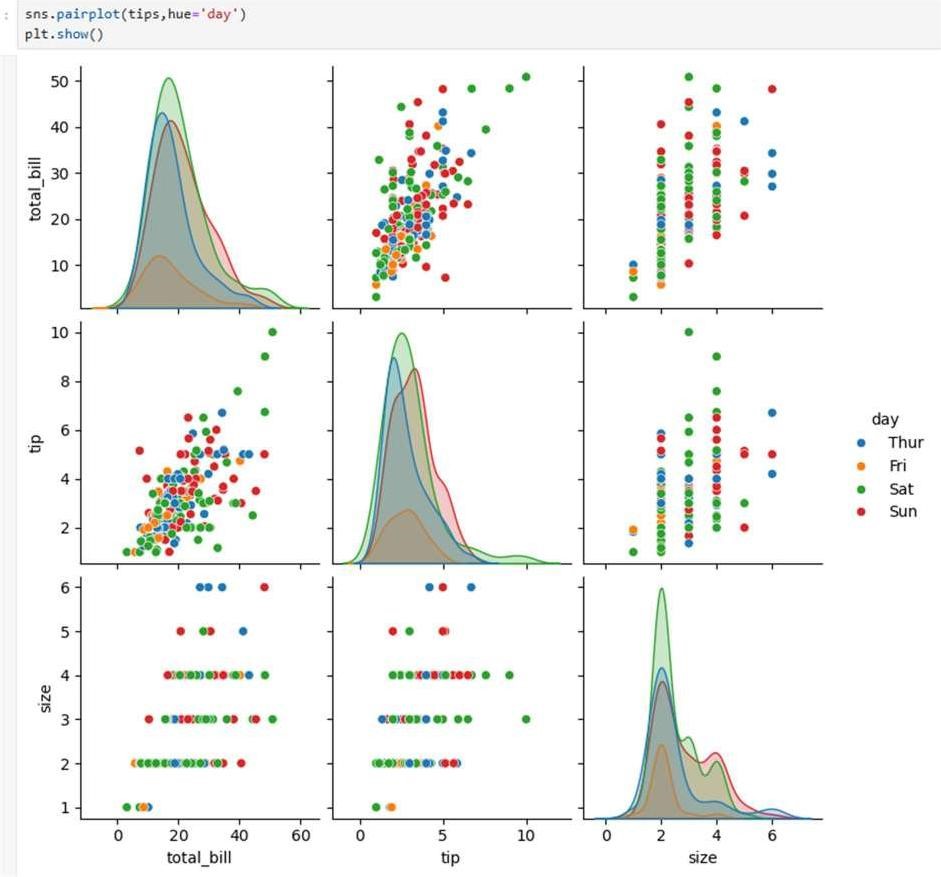


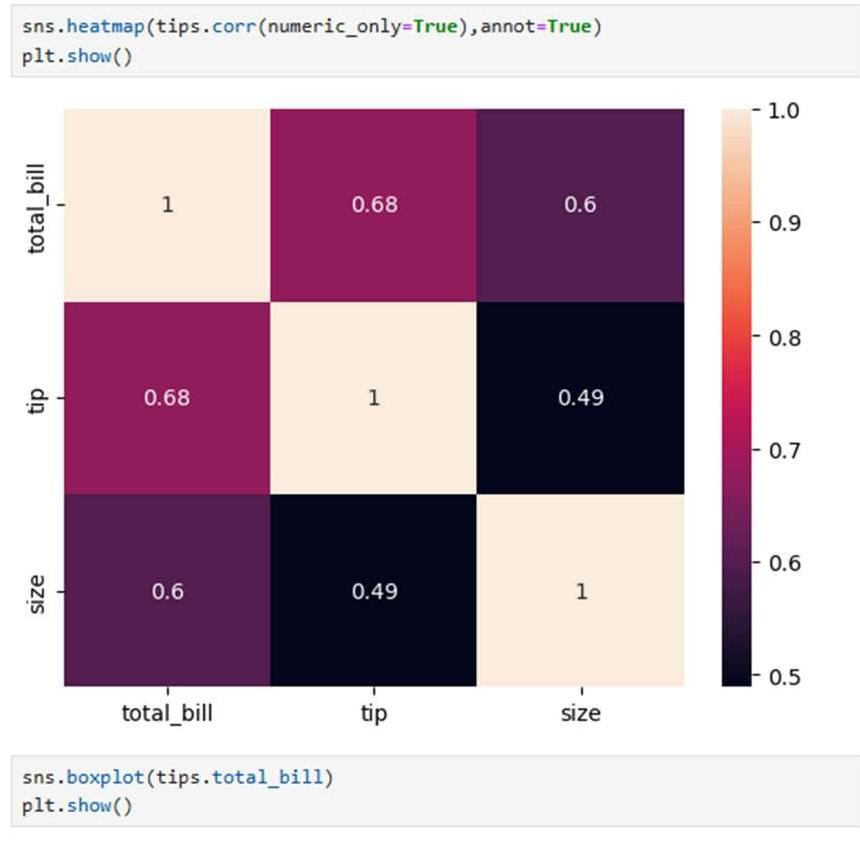


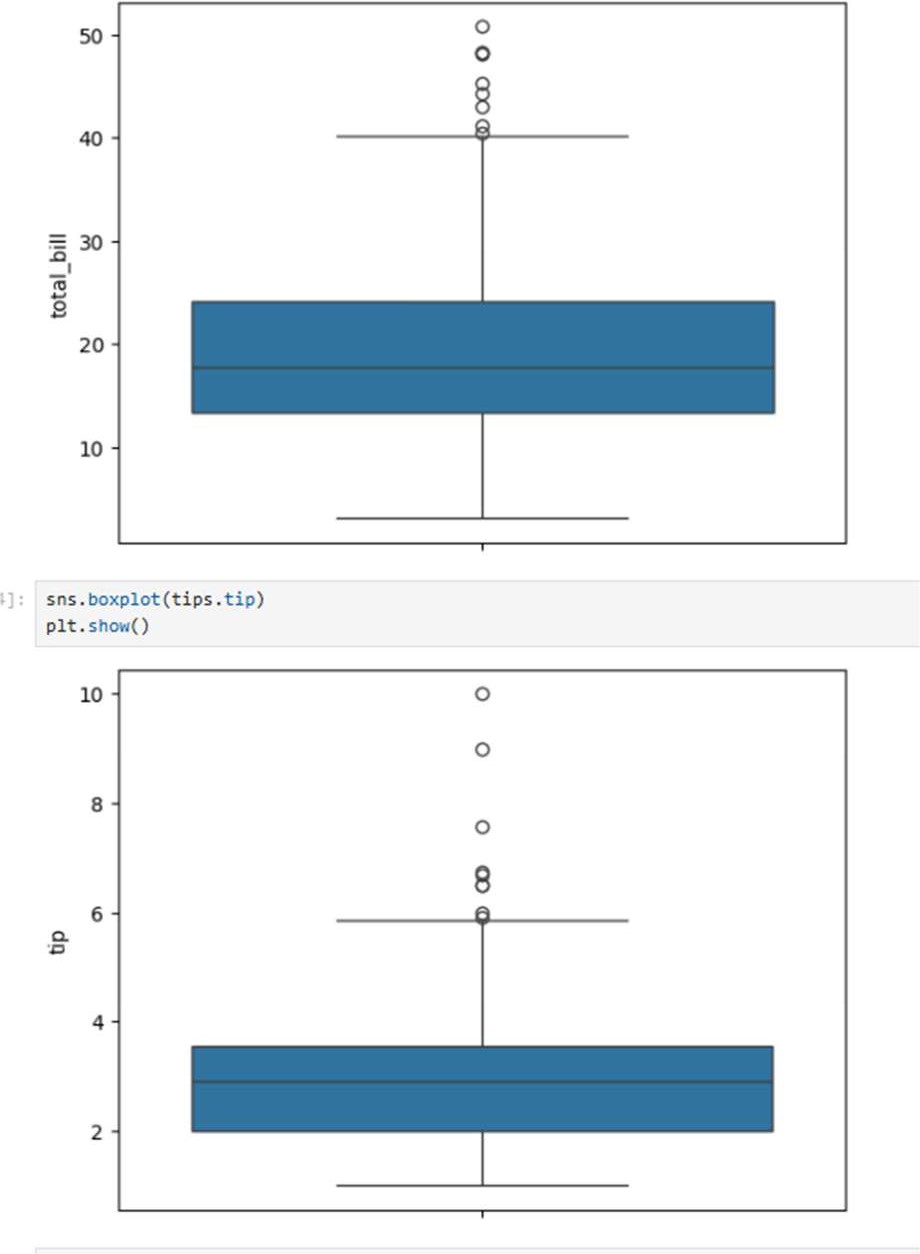


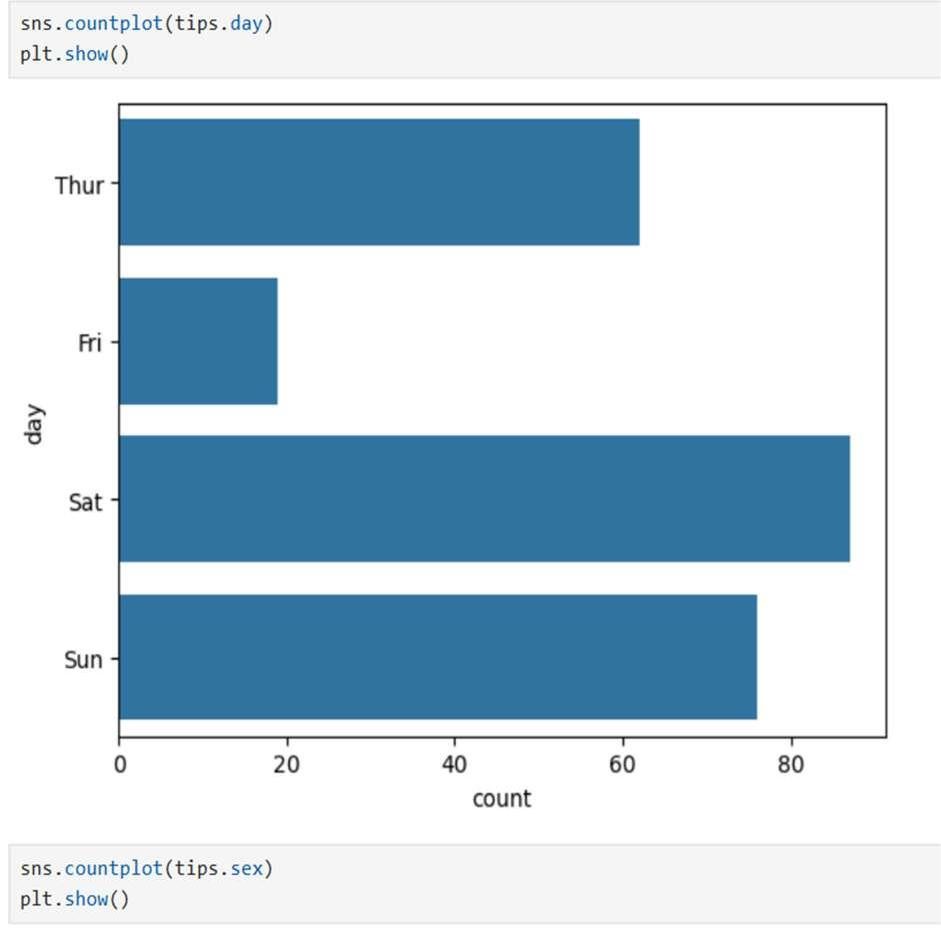


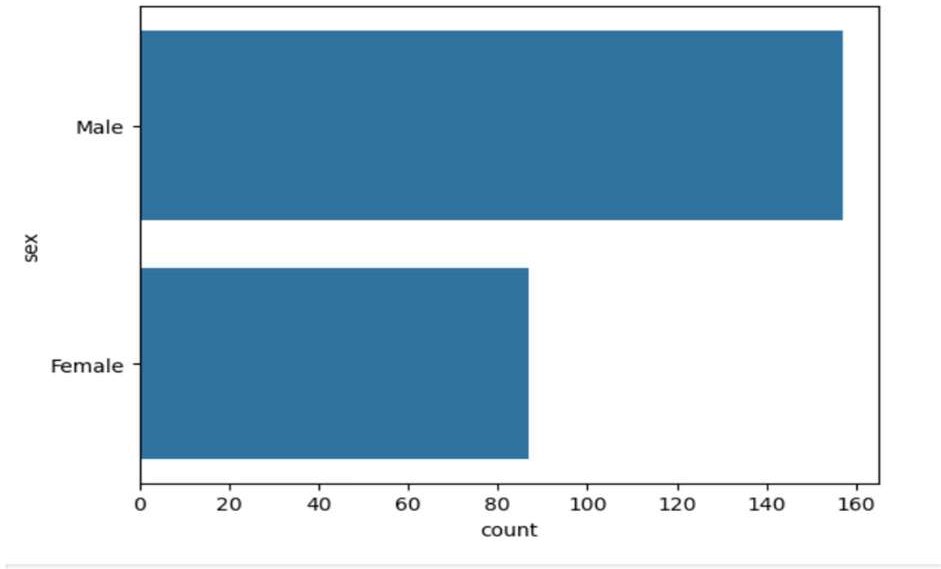












Result:

The visualizations show distributions of bills and tips, relationships between tip and total bill, correlations among numeric features, outliers, and counts by day and gender, providing a clear overview of the dataset.

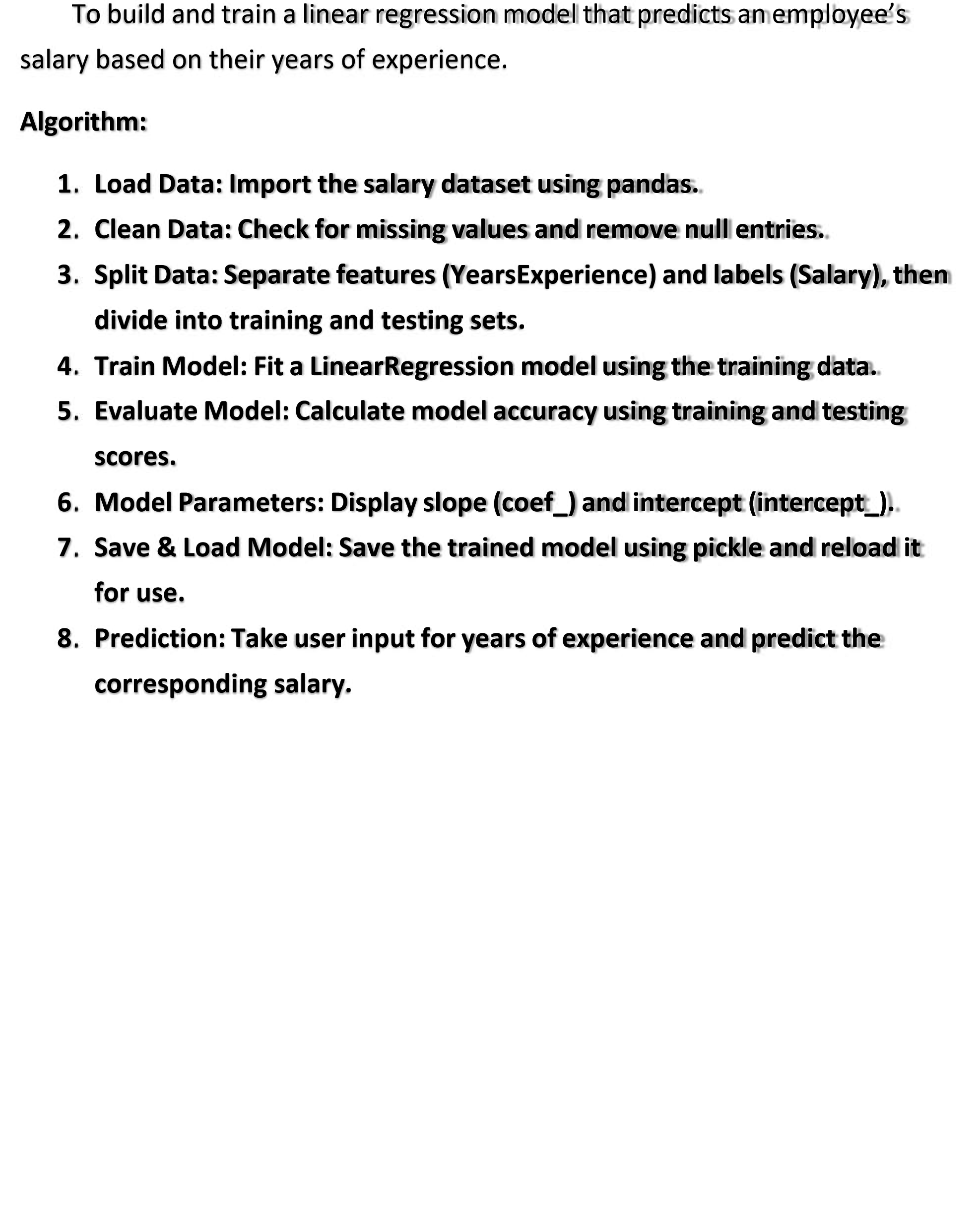
EXPERIMENT NO: 7

# Salary Prediction using Linear Regression



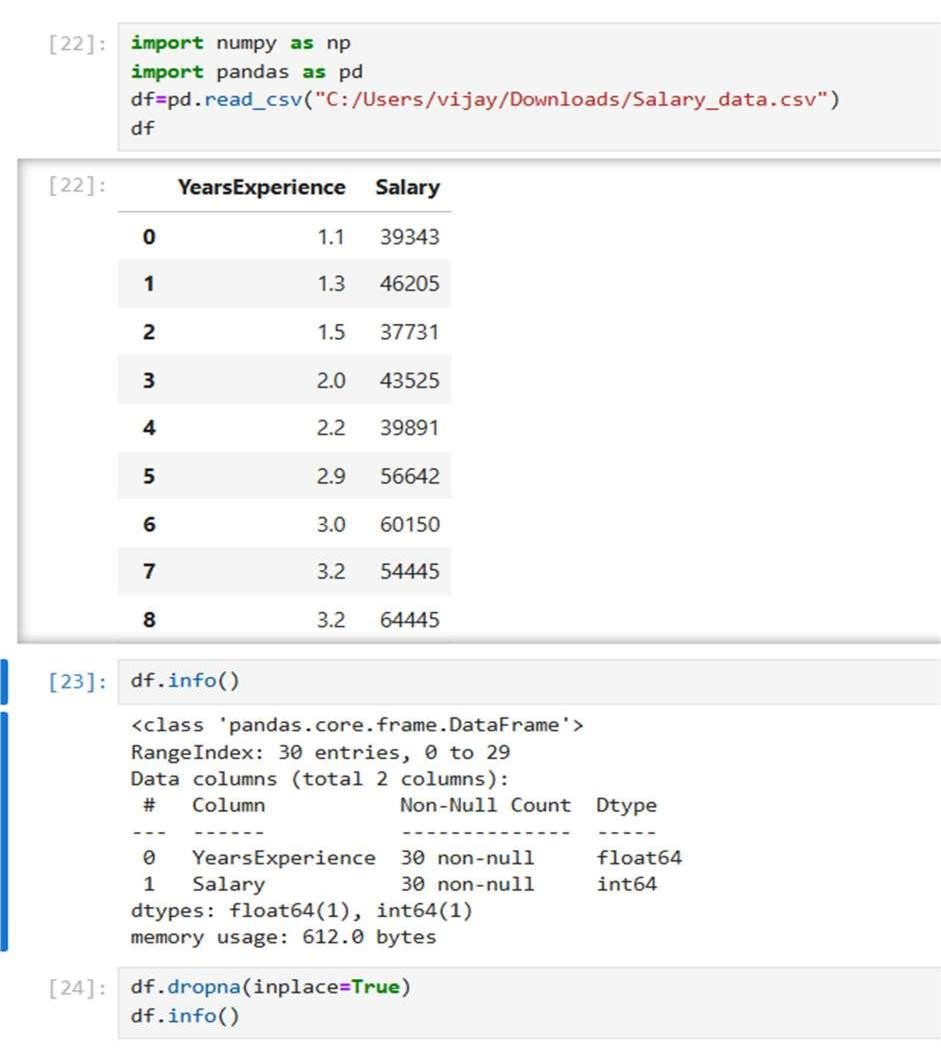
**Aim:**

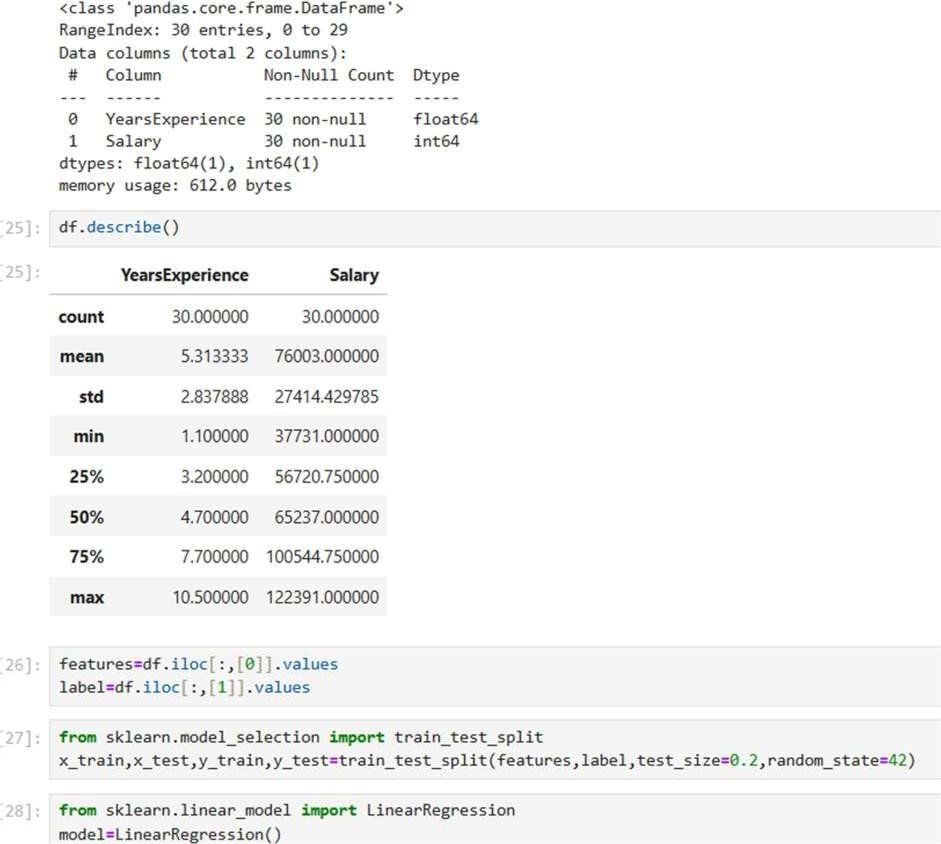


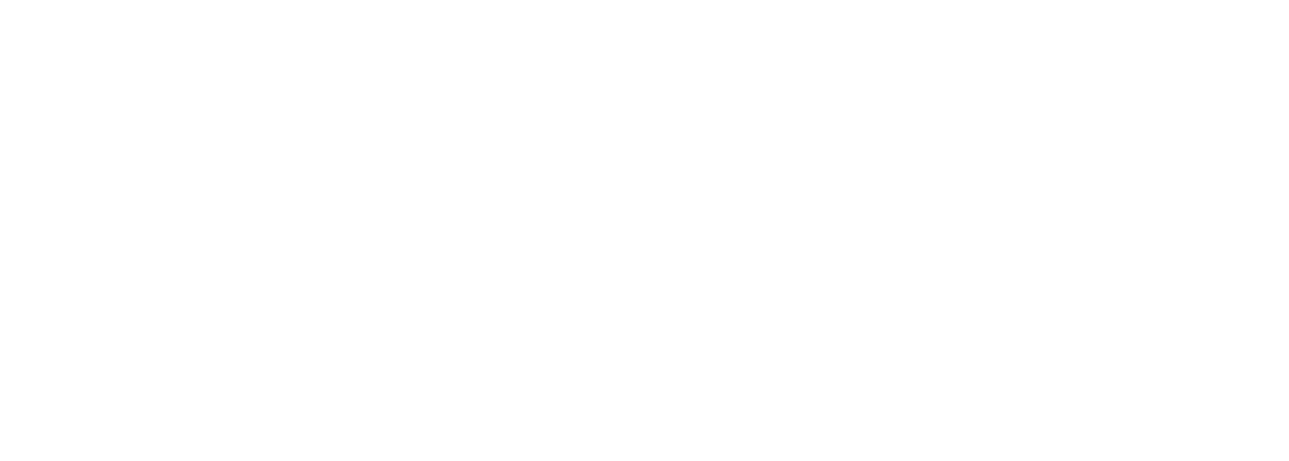
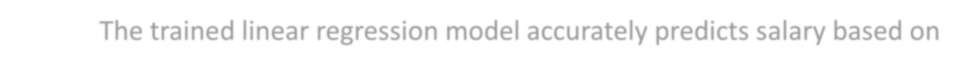
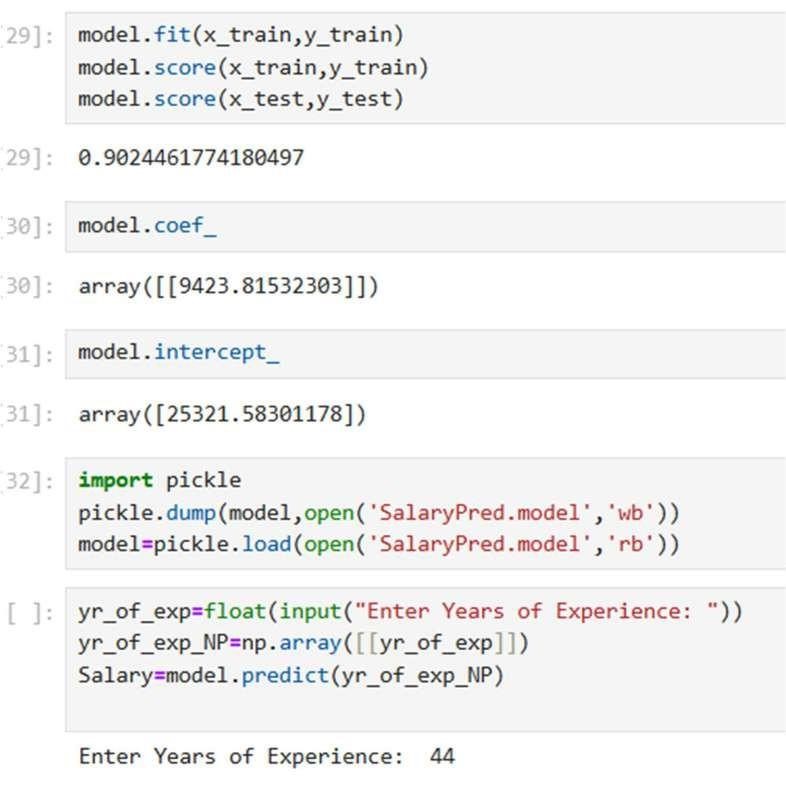




**Program:**







**Result:**

The

trained

linear

regression

model

accurately

predicts

salary

based

on

years of experience. The model is saved for future use, and upon entering a

number of years, it outputs the estimated salary value.

EXPERIMENT NO: 8

# Iris Flower Classification using K-Nearest Neighbors (KNN)



**Aim:**



**.**

**.**

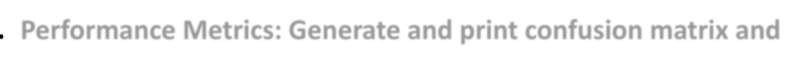
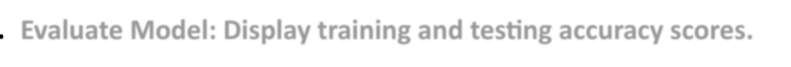
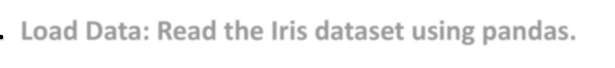
**.**

**.**

**.**

**.**

**.**



To

classify

Iris

flower

species

based

on

their

sepal

and

petal

measurements

using the K

-

Nearest Neighbors algorithm.

**Algorithm:**

**1**

**Load**

**Data:**

**Read**

**the**

**Iris**

**dataset**

**using**

**pandas.**

**2**

**Explore**

**Data:**

**Check**

**data**

**info,**

**count**

**species,**

**and**

**display**

**sample**

**records.**

**3**

**Split**

**Data:**

**Separate**

**features**

**)**

**(**

**measurements**

**and**

**labels**

**species**

**).**

**(**

**4**

**Train**

**-**

**Test**

**Split:**

**Divide**

**data**

**into**

**80**

**%**

**training**

**and**

**20**

**%**

**testing**

**sets.**

**5**

**Model**

**Training:**

**Initialize**

**KNN**

**with**

**5**

**neighbors**

**and**

**fit**

**it**

**on**

**training**

**data.**

**6**

**Evaluate**

**Model:**

**Display**

**training**

**and**

**testing**

**accuracy**

**scores.**

**7**

**Performance**

**Metrics:**

**Generate**

**and**

**print**

**confusion**

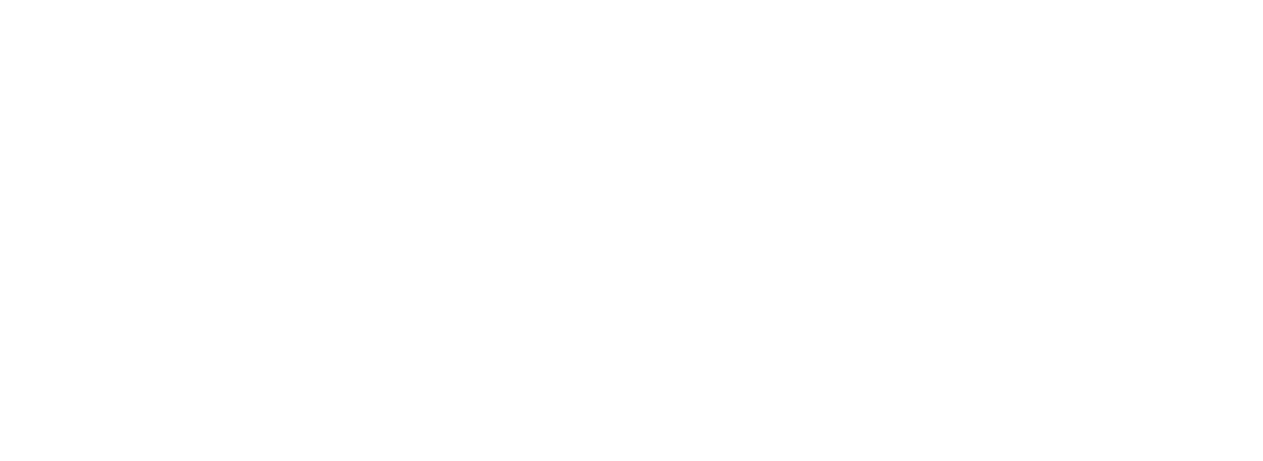
**matrix**

**and**

**classification report.**



**Program:**



**Result:**

The

KNN

model

successfully

classifies

Iris

flowers

into

their

respective

species with high accuracy. The confusion matrix and classification report

confirm that the model performs well with minimal misclassifications.

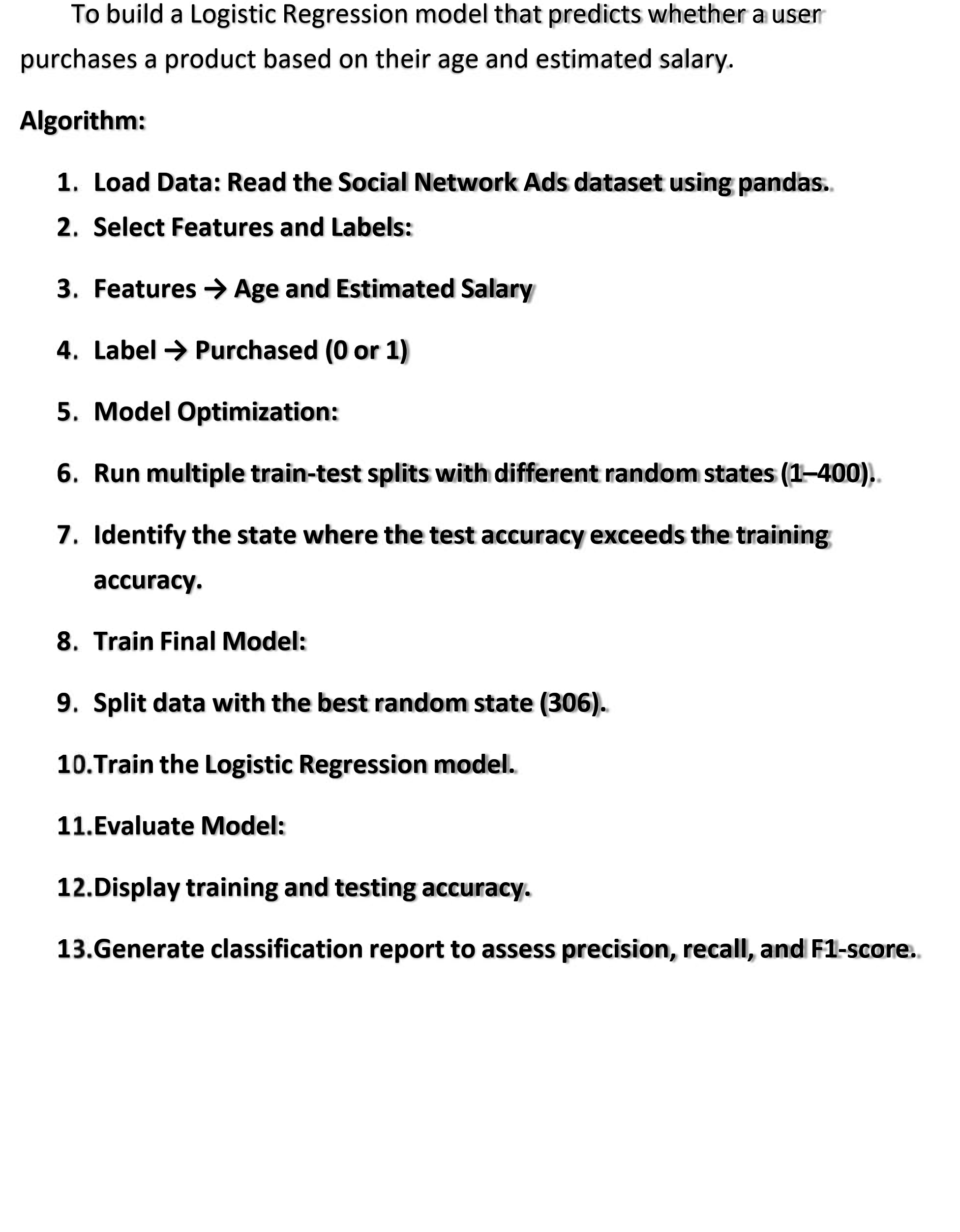
EXPERIMENT NO: 9

# Social Network Ads Classification using Logistic Regression



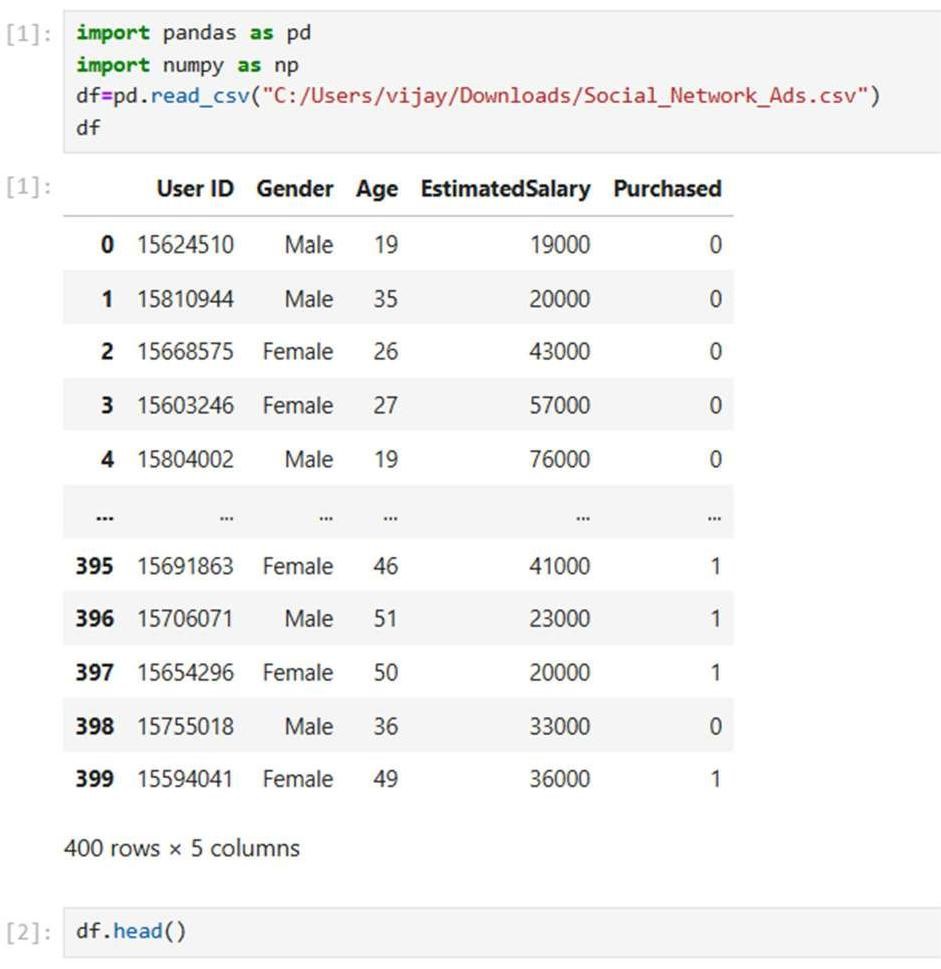
**Aim:**

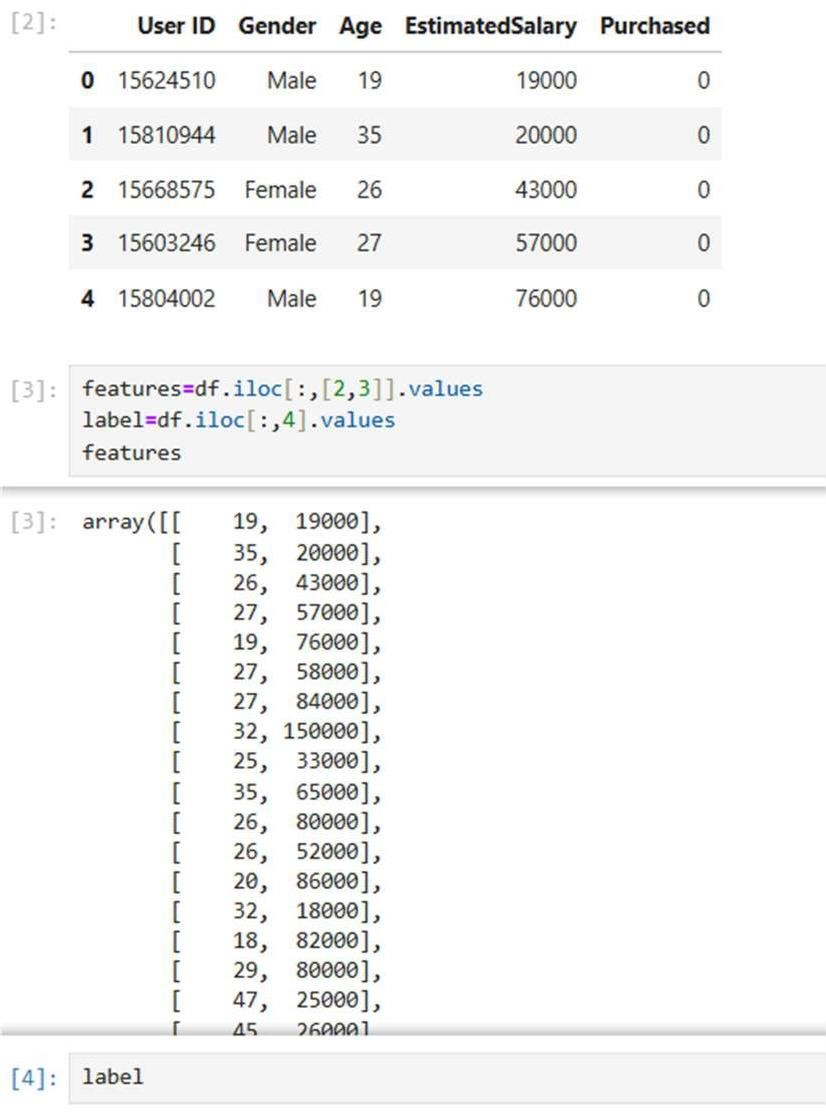




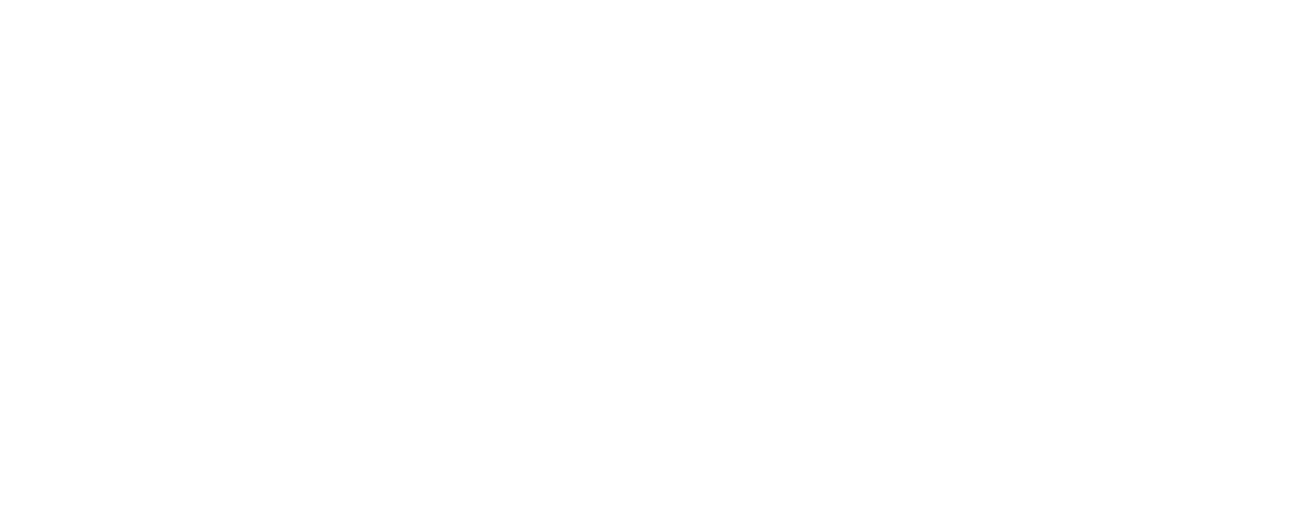
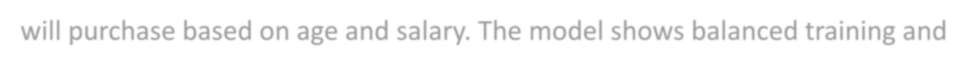
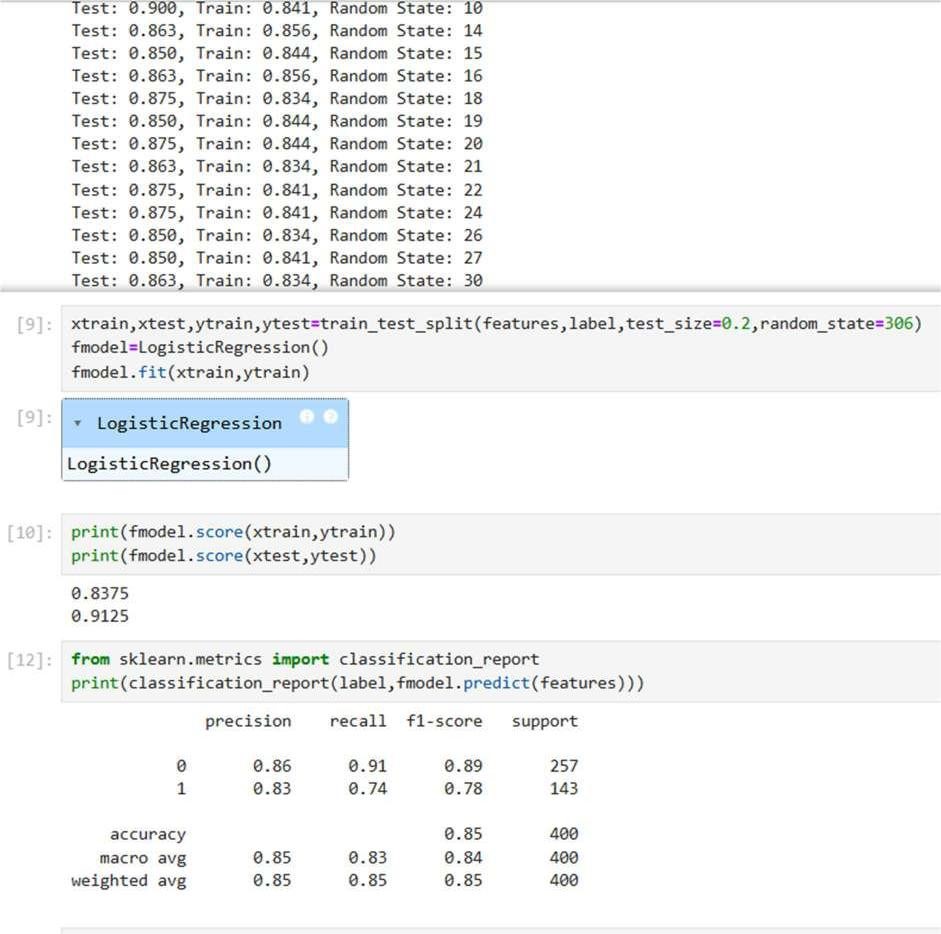


**Program:**









**Result:**

The Logistic Regression model effectively predicts whether a customer

will

purchase

based

on

age

and

salary.

The

model

shows

balanced

training

and

testing accuracy, and the classification report confirms good overall

performance.

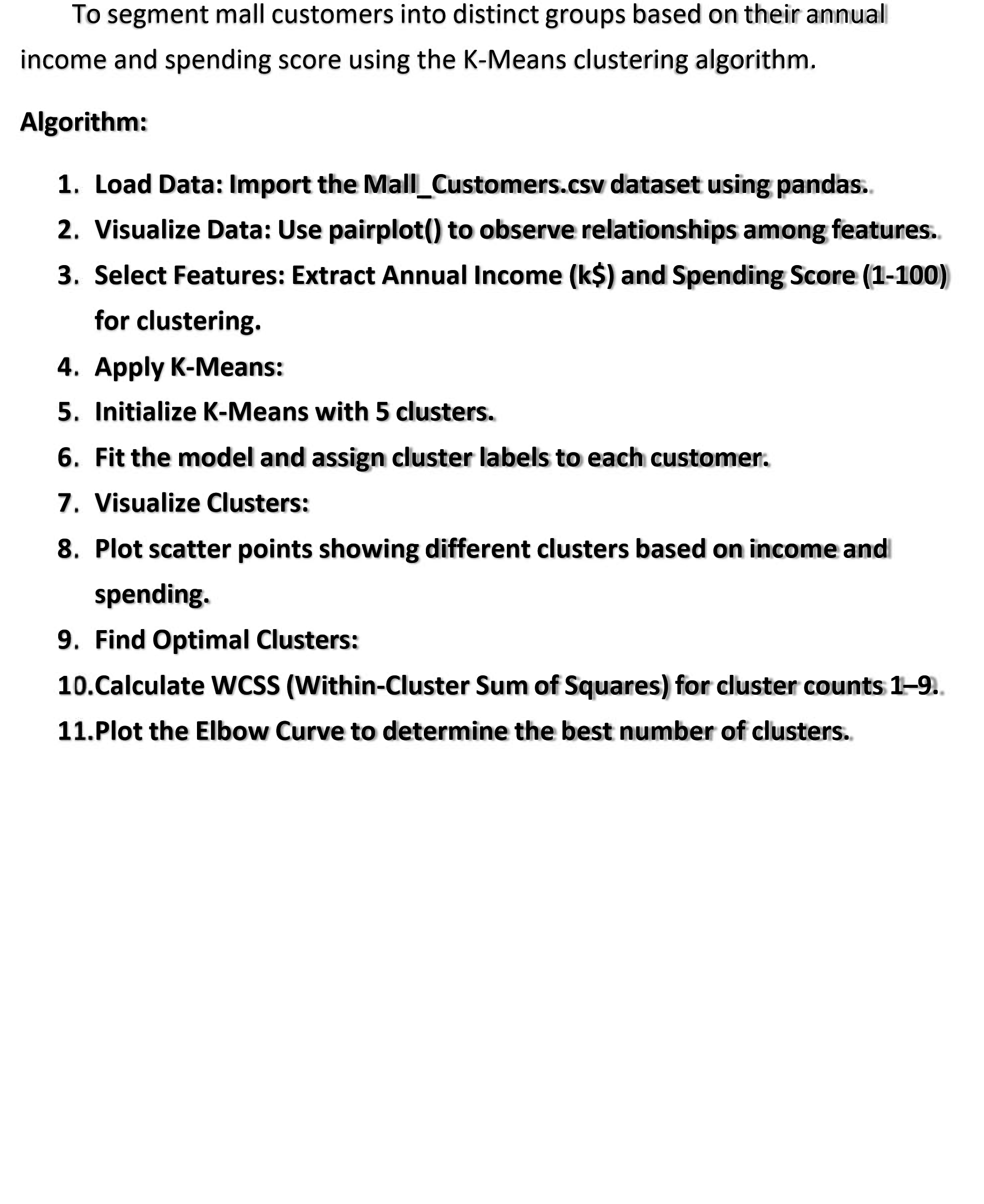
EXPERIMENT NO: 10

# Customer Segmentation using K-Means Clustering



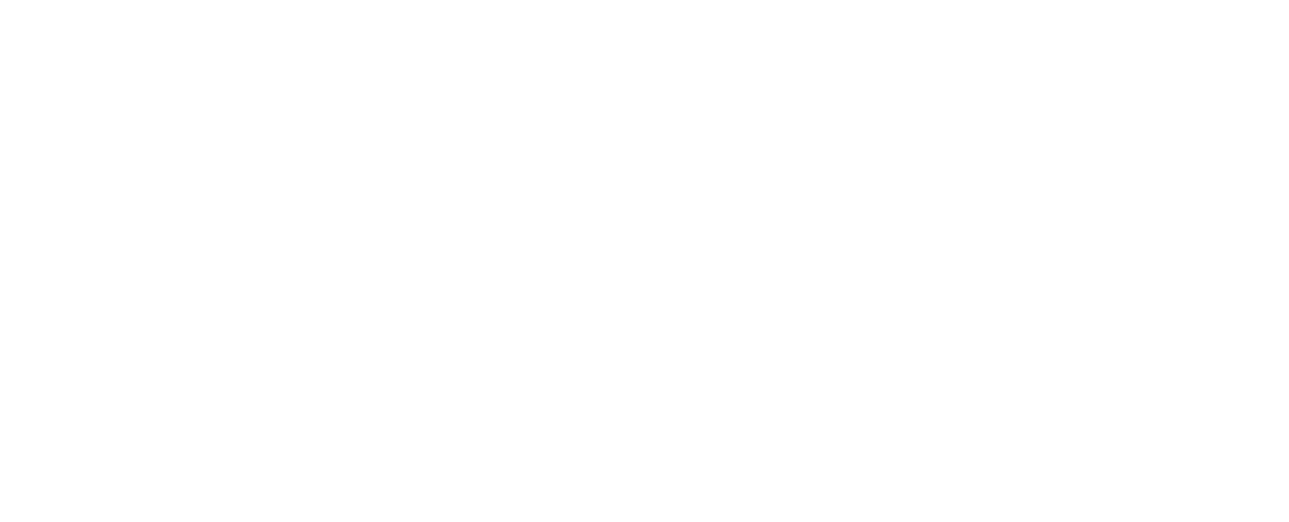
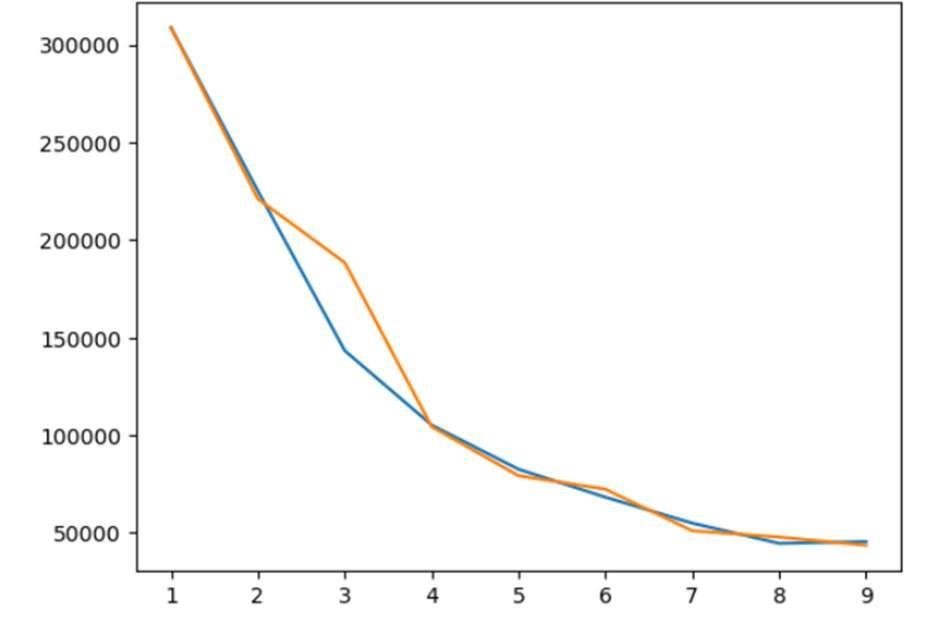
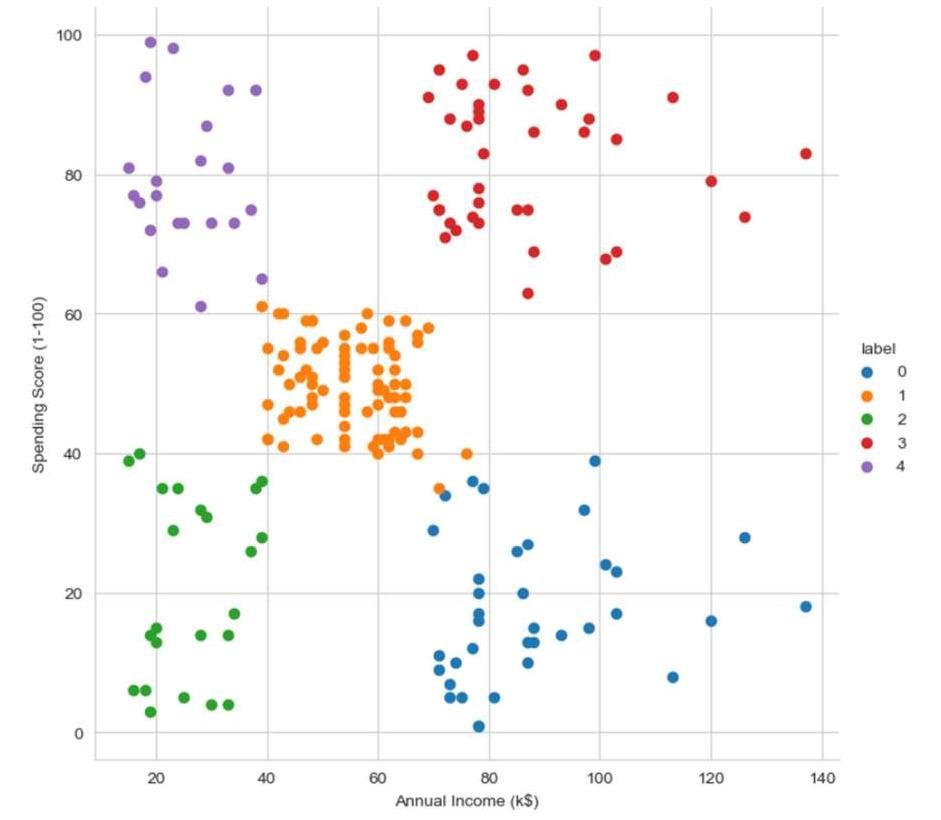
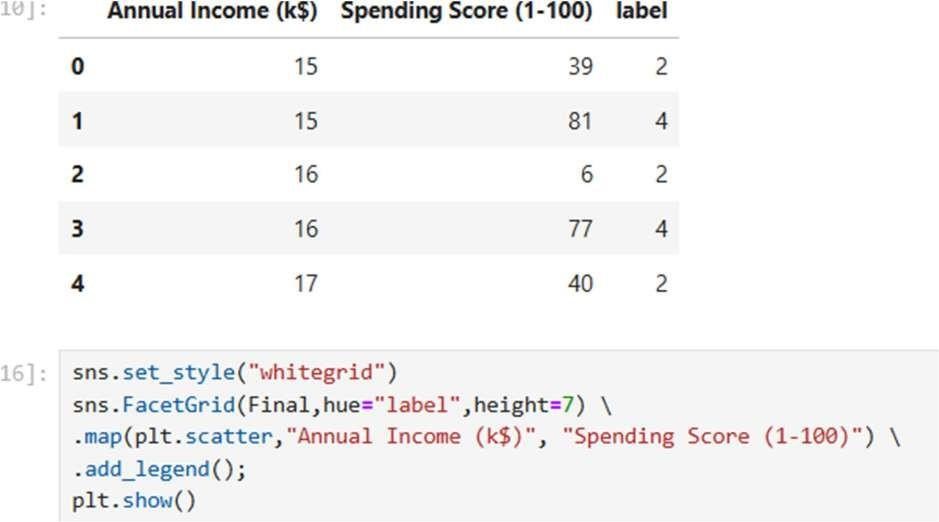
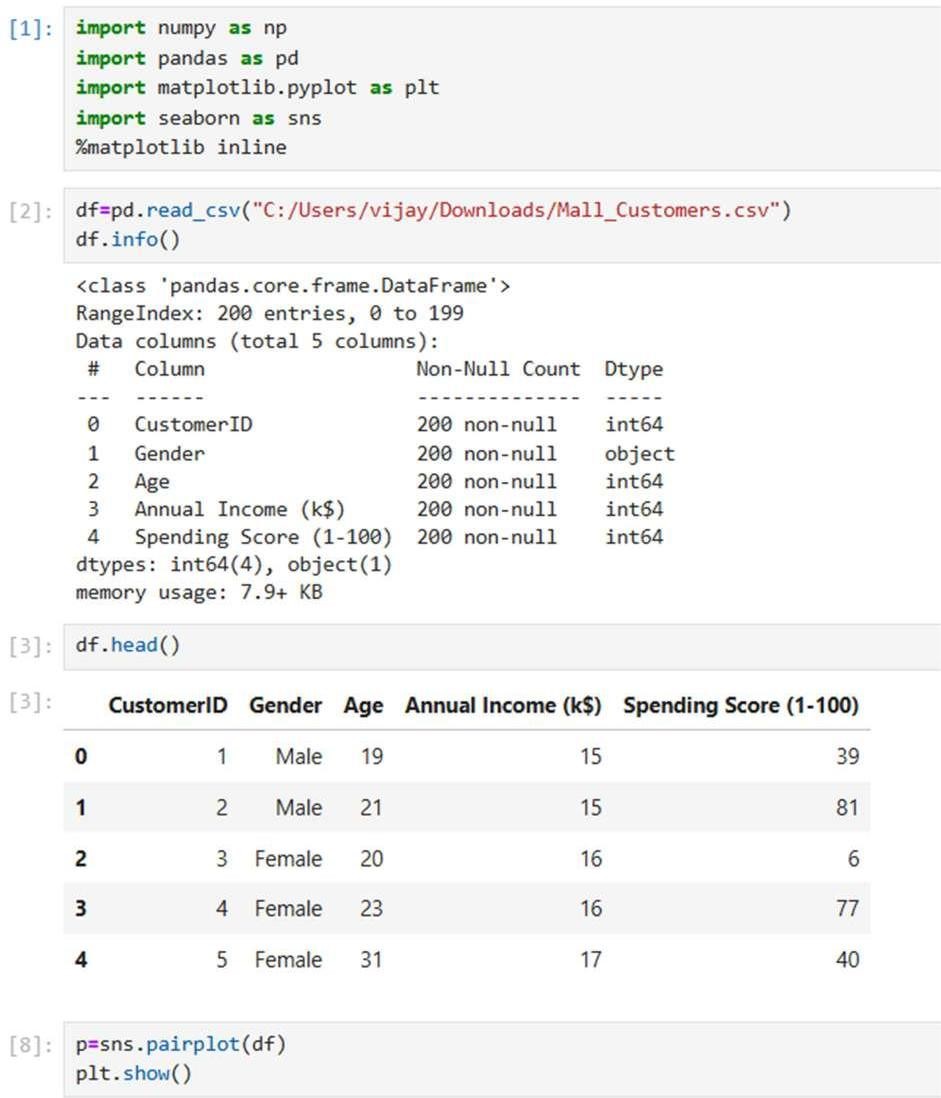
**Aim:**







**Program:**



**Result:**

The

K

-

Means

algorithm

groups

customers

into

5

clusters

based

on

their

spending habits and

income. The Elbow Method

helps confirm that 5 clusters

provide the most balanced segmentation, clearly showing different customer

behavior groups.

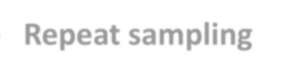
# EXPERIMENT NO: 11



Random Sampling and Sampling Distribution



Aim:



To

explore

random

sampling

from

a

population

and

understand

the concept of sampling distribution using Python in Jupyter

Notebook.

Algorithm:

•

**Initialize**

**population**

**parameters**

•

**Generate**

**the**

**population**

•

**Select**

**sample**

**sizes**

•

**Repeat**

**sampling**

•

**Compute**

**sample**

**means**

•

**Form**

**sampling**

**distributions**

•

**Plot**

**histograms**

•

**Mark**

**population**

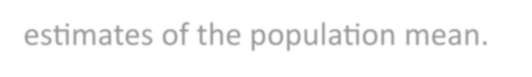
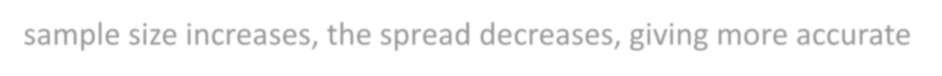
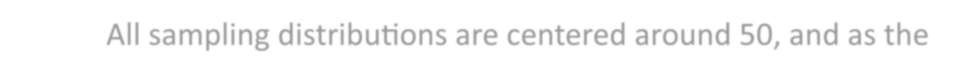
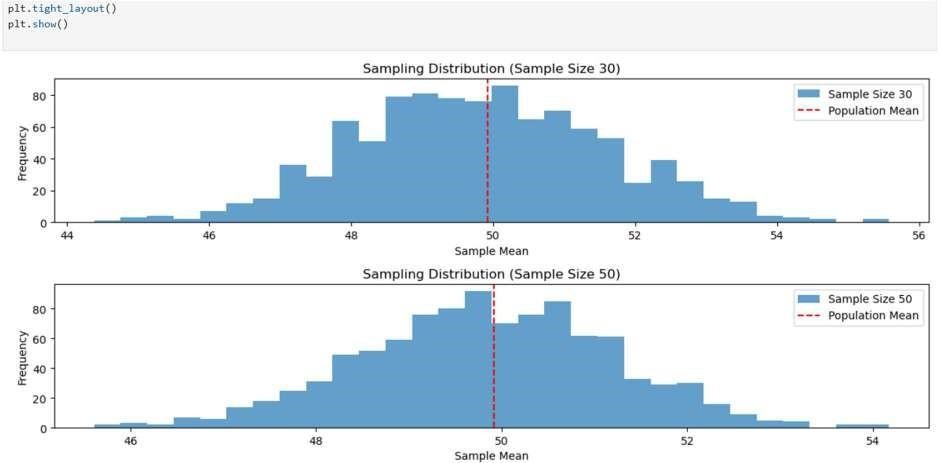
**mean**

•

**Analyze**

**results**

Code:



Result:

All

sampling

distributions

are

centered

around

50

,

and

as

the

sample size increases, the spread decreases, giving more accurate

estimates of the population mean.

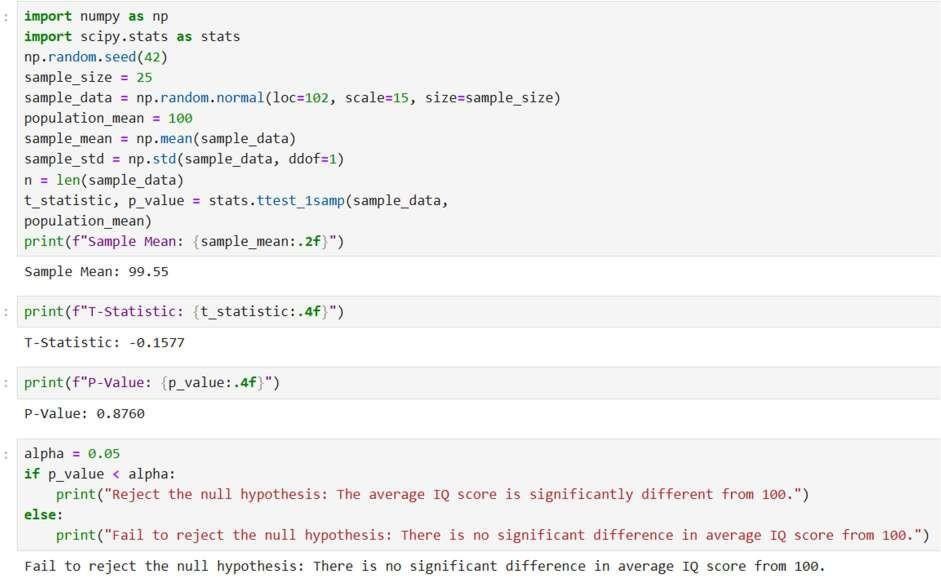
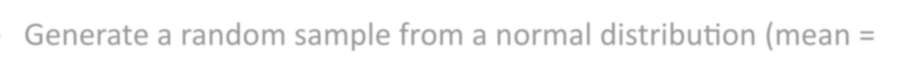
# EXPERIMENT NO: 12



Hypothetical using Z-Test



Aim:



To

test

whether

the

sample

mean

IQ

differs

significantly

from

the

population mean (100) using a one

-

sample t

-

test.

Algorithm:

•

Generate

a

random

sample

from

a

normal

distribution

(

mean

=

102

, std

= 15).

•

Calculate

the

**sample**

**mean**

and

**sample**

**standard**

**deviation**

.

•

Perform

a

**one**

**-**

**sample**

**t**

**-**

**test**

comparing

the

sample

mean

to

the population mean (100).

•

Compare

the

**p**

**-**

**value**

with

the

significance

level

(α

=

0.05).

•

If

*p*

*<*

*α*

,

reject

the

null

hypothesis;

otherwise,

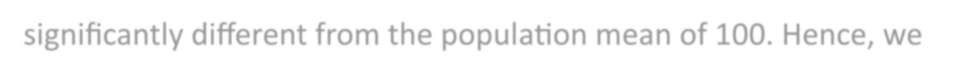
fail

to

reject

it.

Code:



Result:

The test shows that the sample’s average IQ score is not

significantly

different

from

the

population

mean

of

100.

Hence,

we

fail to reject the null hypothesis.

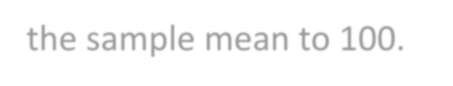
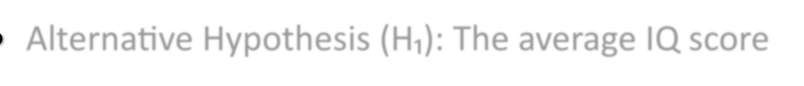
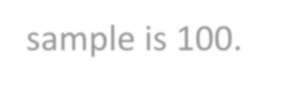
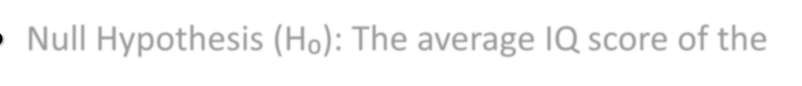
# EXPERIMENT NO : 13 Hypothetical using T-Test



Aim:

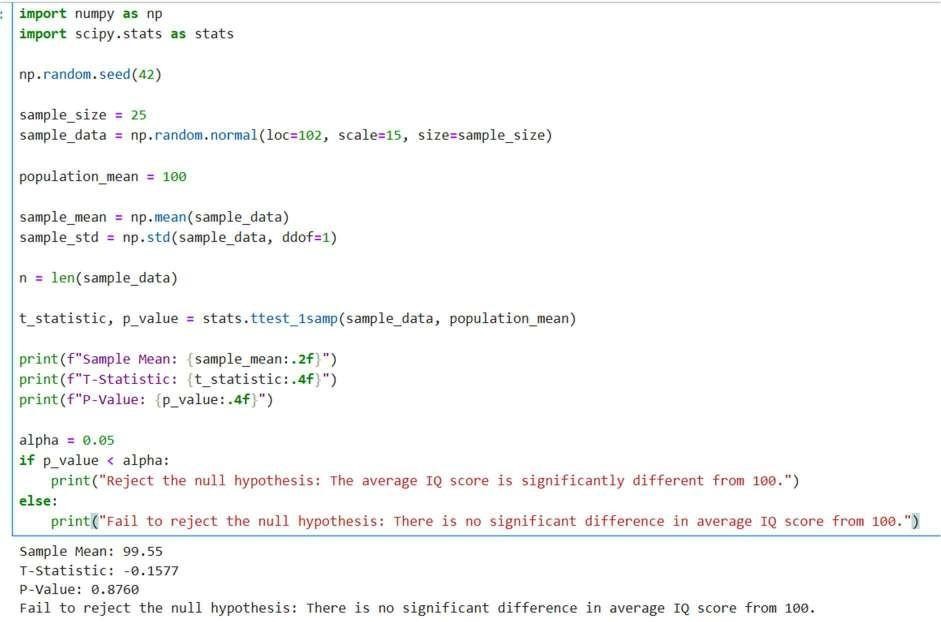


To test whether the average IQ score of a sample of students differs significantly from a population mean IQ score of 100.



Algorithm:

* Null Hypothesis (H₀): The average IQ score of the sample is 100.
* Alternative Hypothesis (H₁): The average IQ score of the sample is not 100.
* Sample: Measure the IQ scores of 25 randomly selected students.
* T-Test: Conduct a one-sample T-test to compare the sample mean to 100.
* Decision Rule: Use a significance level of α = 0.05.

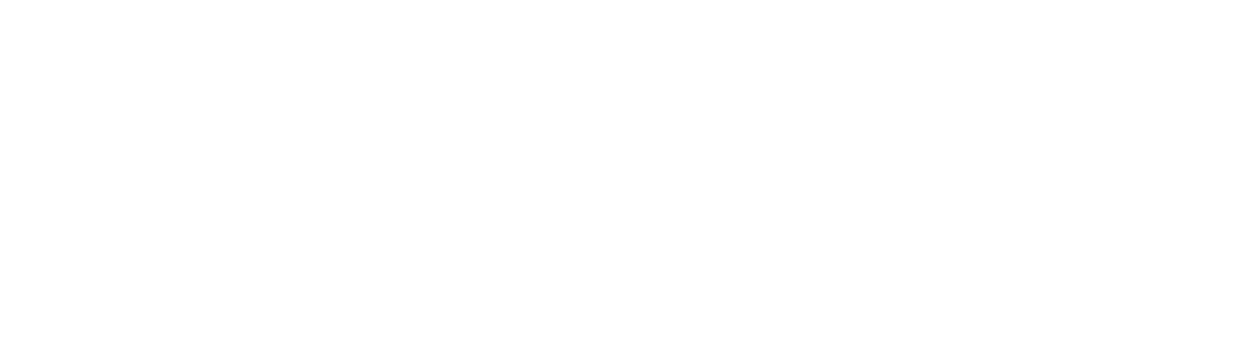


Result:



Code:

The test shows **no significant difference** between the sample mean IQ and the population mean of 100.



Hence, we **fail to reject the null hypothesis**.

# EXPERIMENT NO : 14



Hypothetical using ANOVA-Test



Aim:



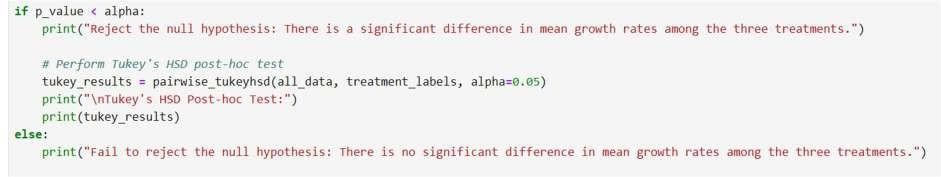
To test whether there is a significant difference in the **mean growth rates** of plants under three different treatments (A, B, and C) using the **One-Way ANOVA** method.



Algorithm:

* Import required libraries: numpy, scipy.stats, and statsmodels.
* Generate three random samples (growth rates) for treatments A, B, and C.
* Perform **One-way ANOVA** using stats.f\_oneway().
* Compare the obtained p-value with significance level α = 0.05.
* If significant, perform **Tukey’s HSD post-hoc test** to find which groups differ.

Code:



Result:

The

p

-

value

(<

0.05)

indicates

that

there

is

a

significant

difference

in

the

mean

growth

rates among treatments A, B, and C.