stats exp2

Experiment Number: 2

AIM: Locate an open-source dataset from the web. Provide a clear description of the data and its source.

Software Used: Python with NumPy, SciPy, and possibly Pandas (for data manipulation) in a Jupyter Notebook or similar environment.

Theory:

This experiment focuses on obtaining and understanding data, a crucial first step in any data analysis or machine learning project. It also highlights the importance of Python libraries like NumPy and SciPy in scientific computing and data manipulation.

- **Open-Source Data:** Data that is freely available for anyone to use, modify, and distribute. Many repositories and platforms provide such data for research, education, and analysis.
- **NumPy:** A fundamental library for numerical computing in Python, providing powerful array objects and tools for linear algebra, Fourier transforms, and more.
- **SciPy:** Built on top of NumPy, SciPy offers a collection of algorithms and functions for scientific computing, including optimization, statistics, signal processing, and more.

Dataset Information:

To complete this experiment, you need to choose an open-source dataset. Here are a few suggestions and sources:

- **Kaggle:** A popular platform with numerous datasets for various domains (https://www.kaggle.com/datasets)
- **UCI Machine Learning Repository:** A well-known repository with a diverse collection of datasets (https://archive.ics.uci.edu/ml/index.php)
- **Google Dataset Search:** A search engine specifically for finding datasets (https://datasetsearch.research.google.com/)

Example Dataset (You should choose your own):

- **Dataset Name:** Iris Dataset
- **Source:** UCI Machine Learning Repository
- **Description:** This classic dataset contains measurements (sepal length, sepal width, petal length, petal width) of 150 iris flowers from three different species (setosa, versicolor, virginica). It is often used for classification and clustering tasks.

Code (Illustrative - adapt to your chosen dataset):

Python

```
import numpy as np
import pandas as pd
from scipy import stats

# Load the dataset (assuming it's in a CSV file)
data = pd.read_csv('iris.csv')

# Display basic information about the data
print(data.head())  # Show the first few rows
print(data.info())  # Show column names and data types
print(data.describe())  # Show summary statistics

# Example calculations using NumPy and SciPy
sepal_length_mean = np.mean(data['sepal_length'])
petal_width_std = stats.tstd(data['petal_width'])

print(f"Mean sepal length: {sepal_length_mean}")
print(f"Standard deviation of petal width: {petal_width_std}")
```

Output (Will vary based on your dataset):

(Output will show the first few rows of the data, column information, summary statistics, and the calculated mean and standard deviation.)

Result:

The code successfully loaded the chosen dataset and displayed its basic information and summary statistics. It also demonstrated the use of NumPy and SciPy functions to calculate simple descriptive statistics.

Learning Outcomes:

- Learned how to find and access open-source datasets from the web.
- Understood the importance of data exploration and examination before analysis.
- Gained familiarity with loading and manipulating data using Pandas.
- Applied NumPy and SciPy functions for basic statistical calculations.
- Developed skills in working with real-world datasets for data analysis tasks.