Assignment No. 10

Aim:

Data Visualization III

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a box plot for each feature in the dataset.
4. Compare distributions and identify outliers.

Prerequisites:

1. Prior knowledge of Python programming.
2. Google Colab / Python IDE
3. Jupyter Notebook

Objectives: Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a box plot for each feature in the dataset.
4. Compare distributions and identify outliers.

Theory:

1. Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt # Importing the required libraries
import seaborn as sns
%matplotlib inline
```

2. Data Visualization:

The process of finding trends and correlations in our data by representing it pictorially is called Data Visualization. To perform data visualization in python, we can use various python data visualization modules such as Matplotlib, Seaborn, Plotly, etc.

3. What is Data Visualization?

Data visualization is a field in data analysis that deals with visual representation of data. It graphically plots data and is an effective way to communicate inferences from data. Using data visualization, we can get a visual summary of our data. With pictures, maps and graphs, the human mind has an easier time processing and understanding any given data. Data visualization plays a significant role in the representation of both small and large data sets, but it is especially useful when we have large data sets, in which it is impossible to see all of our data, let alone process and understand it manually.

4. Data Visualization in Python

Python offers several plotting libraries, namely Matplotlib, Seaborn and many other such data visualization packages with different features for creating informative, customized, and appealing plots to present data in the most simple and effective way.

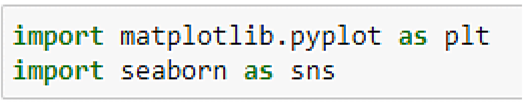
5. Matplotlib and Seaborn

Matplotlib and Seaborn are python libraries that are used for data visualization. They have inbuilt modules for plotting different graphs. While Matplotlib is used to embed graphs into applications, Seaborn is primarily used for statistical graphs.

But when should we use either of the two? Let’s understand this with the help of a comparative analysis. The table below provides comparison between Python's two well-known visualization packages Matplotlib and Seaborn.

Matplotlib	Seaborn
It is used for basic graph plotting like line charts, bar graphs, etc.	It is mainly used for statistics visualization and can perform complex visualizations with fewer commands.
It mainly works with datasets and arrays.	It works with entire datasets.
Seaborn is considerably more organized and functional than Matplotlib and treats the entire dataset as a solitary unit.	Matplotlib acts productively with data arrays and frames. It regards the aces and figures as objects.
Seaborn has more inbuilt themes and is mainly used for statistical analysis.	Matplotlib is more customizable and pairs well with Pandas and Numpy for Exploratory Data Analysis.

Let's consider the apple yield (tons per hectare) in Kanto. Let's plot a line graph using this data and see how the yield of apples changes over time. We start by importing Matplotlib and Seaborn.

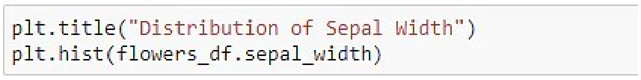


```
import matplotlib.pyplot as plt
import seaborn as sns
```

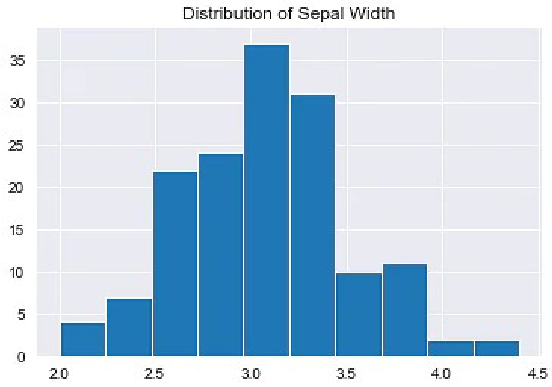
6. Histograms

A Histogram is a bar representation of data that varies over a range. It plots the height of the data belonging to a range along the y-axis and the range along the x-axis. Histograms are used to plot data over a range of values. They use a bar representation to show the data belonging to each range. Let's again use the 'Iris' data which contains information about flowers to plot histograms.

Now, let's plot a histogram using the hist() function.



```
plt.title("Distribution of Sepal Width")
plt.hist(flowers_df.sepal_width)
```



Conclusion:

Thus we have studied various data visualization techniques using maxplotlib, seaborn, histogram and boxplot libraries.