Assignment No 10:

Title: **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.

Theory:

Features of a Dataset

The features of a dataset may allude to the columns available in the dataset. The features of a dataset are the most critical aspect of the dataset, as based on the features of each available data point, will there be any possibility of deploying models to find the output to predict the features of any new data point that may be added to the dataset.

Some possible features of a dataset are:

Numerical Features: These may include numerical values such as height, weight, and so on. These may be continuous over an interval, or discrete variables.

Categorical Features: These include multiple classes/ categories, such as gender, colour, and so on.

Metadata: Includes a general description of a dataset. Generally in very large datasets, having an idea/ description of the dataset when it’s transferred to a new developer will save a lot of time and improve efficiency.

Size of the Data: It refers to the number of entries and features it contains in the file containing the Dataset.

Formatting of Data: The datasets available online are available in several formats. Some of them are JSON (JavaScript Object Notation), CSV (Comma Separated Value), XML (eXtensible Markup Language), DataFrame, and Excel Files (xlsx or xlsm). For particularly large datasets, especially involving images for disease detection, while downloading the files from the internet, it comes in zip files which will be needed to extract in the system to individual components.

Algorithm:

Step 1: Download the data set of Iris

Step 2: Importing Libraries

```
import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

df1=pd.read_csv("Iris_data_sample.csv")

df1.head()
```

Step 3: Identify Features & data types

```
df1.shape

df1.info()

df1["Species"].unique()

df1.groupby("Species").size()

corr=df1.corr()

plt.subplots(figsize=(6,6))

sns.heatmap(corr, annot=True)
```

Step 4: Draw Box Plot

```
def graph(y):

sns.boxplot(x="Species", y=y, data=df1, color="purple")

plt.figure(figsize=(20,20))
```

Step 5: Adding the subplot at the specified grid position

```
plt.subplot(221)

graph('SepalLengthCm')

plt.subplot(222)

graph('SepalWidthCm')

plt.subplot(223)

graph('PetalLengthCm')

plt.show()
```

Questions: Identify the outliers from the boxplot drawn for iris dataset.

Conclusion: Implemented successfully data visualization on dataset features using Python on Iris dataset.