Assignment No - 10 (Group A)

Problem statement:

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

Pre-requisite

1. Basic of Python Programming
2. Seaborn Library, Concept of Data Visualization.
3. Types of variables

Objective

Students should be able to perform the Statistical operations using Python on any open source dataset.

Software and Hardware requirements:-

1. **Operating system:** Linux- Ubuntu 16.04 to 17.10, or Windows 7 to 10,
2. **RAM-** 2GB RAM (4GB preferable)
3. IDE :-Anaconda Jupiter Notebook / pycharm /Visual Studio

Theory-

Data Visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decision.

Histogram in Data Visualization:

A histogram is a chart that displays numeric data in ranges, where each bar represents how frequently numbers fall into a particular range. Like a bar chart, histograms consist of a series of vertical bars along the x-axis. Histograms are most commonly used to depict what a set of data looks like in aggregate. At a quick glance, histograms tell whether a dataset has values that are clustered around a small number of ranges or are more spread out

Boxplot in Data Visualization:

Box Plot is the visual representation of the depicting groups of numerical data through their quartiles. Boxplot is also used for detect the outlier in data set. It captures the summary of the data efficiently with a simple box and whiskers and allows us to compare easily across groups

. Boxplot summarizes a sample data using 25th, 50th and 75th percentiles.

These percentiles are also known as the lower quartile, median and upper quartile. A box plot consist of 5 things.

- Minimum
- First Quartile or 25%
- Median (Second Quartile) or 50%
- Third Quartile or 75%
- Maximum

Draw the boxplot using seaborn library:

Syntax :

seaborn.boxplot(x=None, y=None, hue=None, data=None, order=None, hue_order=None, orient=None, color=None, palette=None, saturation=0.75, width=0.8, dodge=True, fliersize=5, linewidth=None, whis=1.5, notch=False, ax=None, **kwargs)

Parameters:

x = feature of dataset

y = feature of dataset

hue = feature of dataset

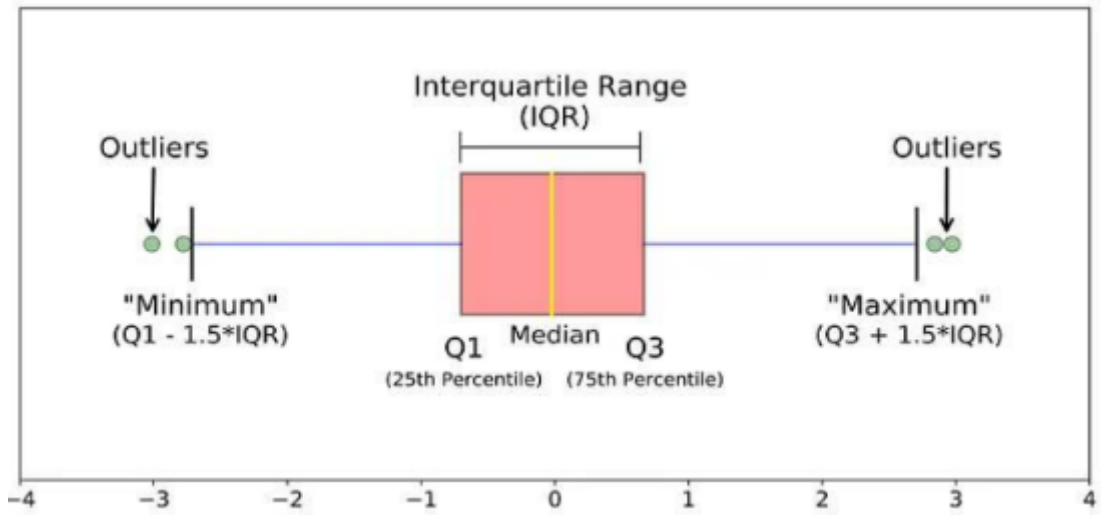
data = dataframe or full dataset

color = color name

Example: # load the dataset tips = sns.load_dataset('tips') tips.head()

Box Plot Visualize Outlier:

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. outlier visualization the box plot is the easiest way to grasp valuable information about your data's outliers. But before visualizing any outliers let's understand what's a box plot and its different components.



As we can see in the image above, a box plot has a lot of components and every one of them helps us to represent and understand the data:

 Q1. 25% of the data is below this data point.

 Median. The central value of the data set. It can also be represented as Q2. 50% of the data is below this data point.

 Q3. 75% of the data is below this data point.

 Minimum. The data point with the smallest value in the data set that isn't an outlier.

 Maximum.

The data point with the biggest value in the data set that isn't an outlier.

 IQR. Represents all the values between Q1 and Q3. Once we understood all the components of a box plot let's visualize it for a given variable in our data set:

```
fig = plt.figure(figsize=(10,5))
sns.boxplot(df.MedInc)
plt.title('Box Plot: Median income for households within a block (tens of thousands of dollars)', fontsize=15)
plt.xlabel('Median Income (tens of thousand of dollars)', fontsize=14)
 plt.show()
```

Conclusion:

In this way we have done how we can draw Boxplot and compare distributions and identify outliers using the Seaborn library. We have seen how to histogram in Seaborn.