**Problem Statement:**

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.

**Objective:**

The describe() function is used **to generate descriptive statistics that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values**.

**Theory:**

**1) df.describe ():**

***-*** The pandas.describe function is used to get a descriptive statistics summary of a given dataframe. This includes mean, count, std deviation, percentiles, and min-max values of all the features.

***-*** The Describe function returns the statistical summary of the dataframe or series. This includes count, mean, median (or 50th percentile) standard variation, min-max, and percentile values of columns. To perform this function, chain .describe() to the dataframe or series.

**2) df.info () :**

The info() method prints information about the DataFrame.

The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non- null values).

**PROGRAM:**

import pandas as pd

import numpy as np

df=pd.read\_csv("User\_Data.csv")

df.describe()

df.info()

**Output:**

**#describe()**

|  | User ID | Age | EstimatedSalary | Purchased |
| --- | --- | --- | --- | --- |
| count | 4.000000e+02 | 400.000000 | 400.000000 | 400.000000 |
| mean | 1.569154e+07 | 37.655000 | 69742.500000 | 0.357500 |
| std | 7.165832e+04 | 10.482877 | 34096.960282 | 0.479864 |
| min | 1.556669e+07 | 18.000000 | 15000.000000 | 0.000000 |
| 25% | 1.562676e+07 | 29.750000 | 43000.000000 | 0.000000 |
| 50% | 1.569434e+07 | 37.000000 | 70000.000000 | 0.000000 |
| 75% | 1.575036e+07 | 46.000000 | 88000.000000 | 1.000000 |
| max | 1.581524e+07 | 60.000000 | 150000.000000 | 1.000000 |

**#info()**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 400 entries, 0 to 399

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 User ID 400 non-null int64

1 Gender 400 non-null object

2 Age 400 non-null int64

3 EstimatedSalary 400 non-null int64

4 Purchased 400 non-null int64

dtypes: int64(4), object(1)

memory usage: 15.8+ KB

**RESULT:**

Thus the data is visualized using plotty framework.