Task 1: Exploration of Data Visualization Tools like Tableau, Python libraries, D3.js

• Connecting Dataset

• Preparation of data

1.a)Extent a dataset by list the attributes in any one of the format (CSV,Excel) .

- Import python libraries whatever needed.
- Read and display the details of the dataset
- Show the dimensionality of data, columns, types and missing values
- Compute statistics on numerical features
- Compute shape of the dataset.

Aim:

To explore and analyze a dataset using Python libraries by connecting, preparing, and summarizing the data.

Algorithm:

1. Start the program.
2. Import the necessary Python libraries (pandas, numpy, matplotlib, seaborn).
3. Load the dataset from a CSV file using pandas.read_csv().
4. Display the dataset to check the contents.
5. Find the shape (rows and columns) of the dataset.
6. Display column names, data types, and missing values.
7. Compute statistics (mean, median, std, min, max) on numerical features using describe().
8. Display the results.
9. End the program.

Python Code:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

```
data = pd.read_csv("students.csv")
print("Dataset:\n", data)
rint("\nShape of dataset:", data.shape)
print("\nColumns:", data.columns.tolist())
print("\nData Types:\n", data.dtypes)
print("\nMissing Values:\n", data.isnull().sum())
print("\nStatistical Summary:\n", data.describe())
sns.histplot(data["Marks"], kde=True, color='skyblue')
plt.title("Distribution of Marks")
plt.show()
```

Sample Dataset (students.csv):

ID	Name	Age	Gender	Marks	Department
1	Anu	20	Female	85	CSE
2	Ravi	21	Male	76	ECE
3	Priya	19	Female	90	IT
4	Kiran	22	Male	65	ME
5	Divya	20	Female	88	CSE

Output :

```
Shape of dataset: (5, 6)
Columns: ['ID', 'Name', 'Age', 'Gender', 'Marks', 'Department']
Data Types:
ID            int64
Name          object
Age           int64
Gender        object
Marks         int64
```

Department object

dtype: object

Missing Values:

ID 0

Name 0

Age 0

Gender 0

Marks 0

Department 0

Statistical Summary:

	ID	Age	Marks
count	5.00000	5.000000	5.000000
mean	3.00000	20.400000	80.800000
std	1.58114	1.140175	9.667797
min	1.00000	19.000000	65.000000
max	5.00000	22.000000	90.000000

Result:

The dataset was successfully imported, analyzed, and summarized using Python libraries. Basic statistics and data characteristics (shape, missing values, types) were computed successfully.

1.b) Consider any one of the dataset from kaggle

- Read the dataset and display 5 lines of your dataframe.
- Identify the display the count of null values(missing values) in each columns
- Clean up the blank(null) column and display it
- Identify the duplicate entries from data set
- Remove the duplicate entries from data set

Algorithm:

1. pd.read_csv("/content/student.csv"): This reads the CSV file and returns a DataFrame object. The contents of the DataFrame are printed to the console using print(a).
2. a.shape: This returns a tuple containing the dimensions of the DataFrame. The shape of the DataFrame is (1000, 8).
3. a.info(): This prints a concise summary of the DataFrame, including the data types of each column and the number of non-null values.
4. a.describe(): This returns a statistical summary of the DataFrame, including count, mean, standard deviation, minimum, and maximum values for each column.
5. a.head(): This returns the first 5 rows of the DataFrame.
6. pd.isna(a).sum(): This returns the number of missing values in each column.
7. a.dropna(): This removes all rows with missing values.
8. a.duplicated(): This returns a boolean Series indicating which rows are duplicates.
9. a.drop_duplicates(): This removes all duplicate rows from the DataFrame.

Code:

```
import pandas as pd

import numpy as np
```

```
a=pd.read_csv("/content/student.csv")

print(a)

sh=a.shape

print(sh)

a.info()

a.describe()

print(a.head())

pd.isna(a).sum()

a.dropna()

a.duplicated()

a.drop_duplicates()
```

Output:

	student_id	student_name	age	gender	marks	grade
0	1	abhinash	10	male	85	a
1	2	nithin	20	male	99	s
2	3	tarun	20	male	95	s
3	4	pooja	20	female	98	s
4	5	siri	17	NaN	65	c
5	3	tarun	20	male	95	s

(6, 6)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   student_id    6 non-null      int64 
 1   student_name  6 non-null      object
 2   age           6 non-null      int64 
 3   gender        5 non-null      object
 4   marks         6 non-null      int64 
 5   grade         6 non-null      object
dtypes: int64(3), object(3)
memory usage: 416.0+ bytes
```

	student_id	age	marks
count	6.000000	6.000000	6.000000
mean	3.000000	19.333333	89.500000
std	1.414214	1.211060	12.988456
min	1.000000	17.000000	65.000000
25%	2.250000	19.250000	87.500000
50%	3.000000	20.000000	95.000000
75%	3.750000	20.000000	97.250000
max	5.000000	20.000000	99.000000

Result:

The dataset was successfully imported, analyzed, and summarized using Python libraries. Basic statistics and data characteristics were computed successfully.