TASK 1: WEB SCRAPING

Introduction:
Web Scraping is a technique used to extract large amounts of data from websites automatically. This data is usually unstructured and is converted into structured form for analysis.

Objectives of Web Scraping:
• Collect real-time data from public websites
• Create custom datasets for analysis
• Reduce manual data collection effort

Tools Used:
• Python (Requests, BeautifulSoup, Scrapy)
• No-code tools: Octoparse, ParseHub

Process of Web Scraping:
1. Identify the target website and dataset
2. Study the HTML structure using browser developer tools
3. Send HTTP request to fetch web page
4. Parse HTML and extract required elements
5. Store the extracted data in CSV/Excel/JSON format

Python Code for Web Scraping:

```
import requests
from bs4 import BeautifulSoup
import csv

url = "https://example.com"
response = requests.get(url)

soup = BeautifulSoup(response.text, "html.parser")

titles = soup.find_all("h2")

with open("output.csv", "w", newline="") as file:
    writer = csv.writer(file)
    writer.writerow(["Title"])

    for title in titles:
        writer.writerow([title.text.strip()])

print("Data scraped successfully")
```

TASK 2: EXPLORATORY DATA ANALYSIS (EDA)

Introduction:
Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics using statistical methods and visualizations.

Objectives of EDA:
• Understand data structure and variables
• Identify trends and patterns
• Detect outliers and missing values
• Validate assumptions before modeling

Steps in EDA:
1. Load the dataset
2. Understand rows, columns, and data types
3. Perform summary statistics
4. Visualize data distributions
5. Detect anomalies and data issues

Python Code for EDA:

```
import pandas as pd
import matplotlib.pyplot as plt

# Load dataset
data = pd.read_csv("output.csv")

# Basic information
print(data.info())
print(data.describe())

# Check missing values
print(data.isnull().sum())

# Visualization
data['Title'].value_counts().head(10).plot(kind='bar')
plt.show()
```

Outcome of EDA:
• Clear understanding of dataset quality
• Identification of key variables
• Dataset ready for further analysis or machine learning

Conclusion:
Web Scraping helps in collecting data from the web, while EDA helps in understanding and preparing the data. Both are fundamental steps in the data science lifecycle.