Web Scraping and Data Analysis

Introduction: Web scraping, is a process of importing data from websites into files or spreadsheets. It is used to extract data from the web, either for personal use by the scraping operator, or to reuse the data on other websites. There are numerous software applications for automating data scraping.

Web scraping is commonly used to:

- Collect business intelligence to inform web content
- Determine prices for travel booking or comparison sites
- Find sales leads or conduct market research via public data sources
- Send product data from eCommerce sites to online shopping platforms like Google Shopping

Tools and Technologies Used: To complete the web scraping successfully I have used **'Python'** as a programming language. Python has extensive support for third-party libraries. The language's ability to handle dynamic websites makes it an ideal choice for scraping large amounts of data from multiple sources.
For the data analysis purpose I have used the **'POWER BI',** which gave good responsive dashboards for the further analysis on data.

Web Scraping: Web scraping is the process of extracting data from websites. It involves fetching the web page content and then parsing it to extract the desired information. This can be useful for various purposes such as data analysis, research, price monitoring, and more.
Here are the basic steps involved in web scraping:

Identify the Target Website: Determine the website from which you want to scrape data.
Understand the Structure of the Website: Understand the structure of the website's HTML code. You'll need to identify the elements that contain the data you want to scrape.

Choose a Web Scraping Tool or Library: There are various tools and libraries available for web scraping in different programming languages. Popular ones include BeautifulSoup.
Fetch the Web Page: Use the web scraping tool or library to fetch the HTML content of the web page.
Parse the HTML: Once you have the HTML content, parse it to extract the relevant data. This involves selecting specific HTML elements, such as <div>, <p>, <table>, etc., based on their class, id, tag names, or other attributes.
Extract the Data: Extract the desired data from the selected HTML elements. This could be text, links, images, or any other information.
Store or Process the Data: After extracting the data, you can choose to store it in a database, CSV file, or process it further according to your requirements.

For the scraping purpose I have chosen ‘https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_States_by_revenue’. A website which contains information about **List of largest companies in the United States by revenue.**

Data Analysis: Power BI is a powerful business intelligence tool developed by Microsoft for data analysis and visualization. It allows users to connect to various data sources, transform data, create interactive reports and dashboards, and share insights across an organization.
The basic guidelines of analyzing the data using POWER BI are:

Import Data: start by importing your data into Power BI. You can connect to a wide range of data sources including databases, Excel files, CSV files, cloud services like Azure and Salesforce, and more.

Data Transformation: Once your data is imported, you may need to clean and transform it to make it suitable for analysis. Power BI provides a range of transformation options such as filtering, splitting columns, merging queries, adding custom columns, and more. Use these tools to prepare your data for analysis.

Data Modeling: Power BI uses a data modeling approach based on relationships between tables. Create relationships between tables if your data is spread across multiple tables.

Data Visualization: Use Power BI's drag-and-drop interface to create visualizations such as charts, graphs, maps, and tables to represent your data. Choose the appropriate visualization type based on the nature of your data and the insights you want to convey. You can customize the appearance of your visualizations and add interactive elements like slicers and filters to allow users to explore the data dynamically.

Create Reports: Combine your visualizations into interactive reports that tell a story or convey specific insights. Arrange your visualizations on a canvas and add text boxes, images, and shapes to provide context and explanation. Power BI allows you to create multiple pages within a report to organize your content effectively.

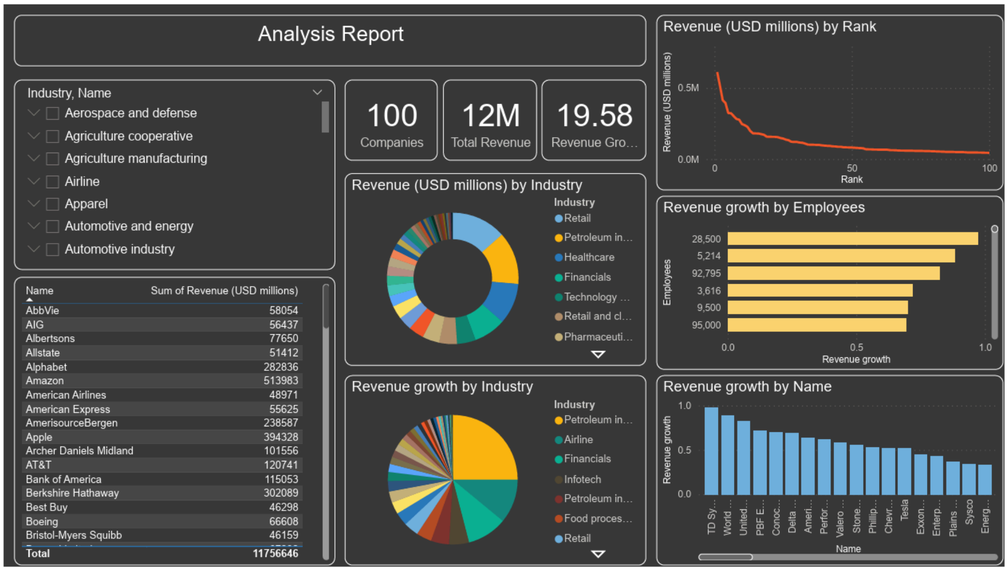
Create Dashboards: Dashboards in Power BI provide a high-level overview of key metrics and insights. Pin visualizations from your reports to a dashboard to create a customized dashboard that highlights the most important information. You can also add tiles such as text boxes, images, and web content to provide additional context.

Analyze Data: Once you have created your reports and dashboards, use Power BI's interactive features to analyze your data. Drill down into specific details, apply filters and slicers to focus on subsets of data, and perform ad-hoc analysis using natural language queries with Power BI's Q&A feature.

Share and Collaborate: Share your reports and dashboards with others in your organization using Power BI Service. Publish your content to the Power BI Service where users can view and interact with it using a web browser or mobile app. You can also create content packs and apps to distribute your insights to specific user groups.

Monitor Performance: Use Power BI's built-in monitoring and analytics capabilities to track the performance of your reports and dashboards. Monitor usage metrics, identify popular content, and gather feedback from users to continuously improve your data analysis and visualization efforts.

Result of POWER BI on analyzing the web scraping:



Analysis Report

Industry, Name
- □ Aerospace and defense
- □ Agriculture cooperative
- □ Agriculture manufacturing
- □ Airline
- □ Apparel
- □ Automotive and energy
- □ Automotive industry

Name	Sum of Revenue (USD millions)
AbbVie	58054
AIG	56437
Albertsons	77650
Allstate	51412
Alphabet	282836
Amazon	513983
American Airlines	48971
American Express	55625
AmerisourceBergen	238587
Apple	394328
Archer Daniels Midland	101556
AT&T	120741
Bank of America	115053
Berkshire Hathaway	302089
Best Buy	46298
Boeing	66608
Bristol-Myers Squibb	46159
Total	**11756646**

100 Companies

12M Total Revenue

19.58 Revenue Gro…

Conclusion: In conclusion, leveraging Python for web scraping and Power BI for data analysis presents a robust and efficient approach to extracting valuable insights from web-based sources. Through Python's versatile libraries such as BeautifulSoup and Scrapy, web scraping becomes streamlined, allowing for the seamless extraction of data from diverse websites. Once the data is collected, Power BI offers an intuitive platform for transforming raw data into actionable insights, with its rich visualization capabilities and interactive features. Together, this combination empowers users to uncover trends, patterns, and correlations within the scraped data, facilitating informed decision-making processes. By seamlessly integrating web scraping with Python and data analysis with Power BI, organizations can harness the power of web-based information to drive strategic initiatives, enhance operational efficiency, and gain a competitive edge in today's data-driven landscape.

Reported by,

Noone Harshan