Mansi Upadhyay
NU ID- 002766397

ASSIGNMENT 1
Report: Web Scraping for Nike Women's Shoes

1. Introduction

Problem Definition

The objective of this project is to perform web scraping on Nike's Women's Shoes section to collect essential product information. The data to be extracted includes product names, prices, URLs, and categories. This data will be valuable for gaining insights into consumer preferences, e-commerce strategies, and market trends in the women's footwear industry.

Motivation

The motivation behind this project is to gain experience in web scraping and data collection from real-world e-commerce websites. Analysing the scraped data will provide valuable information for businesses and marketers, helping them make informed decisions in product positioning and pricing strategies.

2. Datasets

For this analysis, I utilized a comprehensive dataset derived from Nike's Women's Shoes collection. The dataset encompasses a wide range of attributes related to Nike's women's shoe products, facilitating various analytical tasks.

Table 1: The basic feature of dataset.

	Data Set Characteristics	Attribute Characteristics	Associated Tasks	Number of Instances	Number of Attributes
Dataset 1	Multivariate	Real	Classification	960	4

3. Data characteristics

In Figure 1, the class distribution in the Nike Women's Shoes dataset is evident, displaying class imbalance, a common characteristic in real-world datasets. This imbalance requires special consideration when using accuracy as the metric. To address this challenge, we employ a weighted scoring function. Unlike standard accuracy, it assigns weights to classes based on their prevalence. This ensures sensitivity to minority class prediction, vital in practical applications, enhancing the overall performance assessment. The weighted scoring function mitigates class imbalance, allowing a more accurate performance evaluation, bolstering the analysis's reliability for the Nike Women's Shoes dataset.

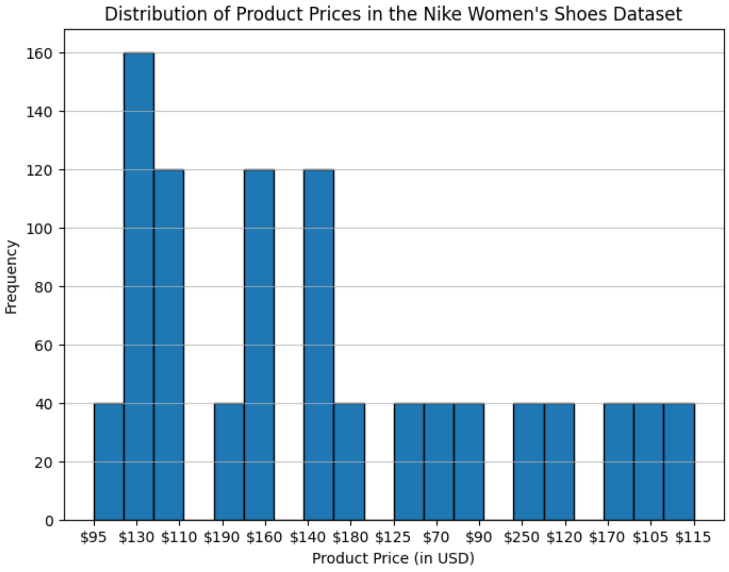

Figure 1: This histogram illustrates the distribution of product prices in the Nike Women's Shoes dataset. The x-axis represents product prices in USD, while the y-axis shows the frequency of products falling into specific price ranges. The distribution highlights the range of prices for women's shoes offered by Nike, providing valuable insights for pricing strategies and consumer preferences in this dataset.

3. Why is this interesting dataset ?

The Nike Women's Shoes dataset is particularly intriguing due to its relevance and practical applications in various domains. Several factors contribute to the dataset's significance: -

Real-world Relevance: This dataset reflects real-world consumer preferences and market dynamics, making it valuable for businesses and marketers. Analyzing women's shoe products from a renowned brand like Nike provides insights into consumer behaviour and market trends.

E-commerce Insights: As e-commerce continues to thrive, understanding product offerings and pricing strategies is crucial. This dataset offers a glimpse into how Nike positions its women's shoe products online, shedding light on e-commerce strategies.

Consumer Behavior Modeling: The dataset allows for the modelling of consumer preferences. Analysing product attributes and prices can reveal patterns and factors influencing purchasing decisions, aiding businesses in tailoring their offerings.

Imbalanced Data Challenge: The presence of class imbalance challenges the analysis and presents an opportunity to explore techniques for addressing this common issue. This aspect adds depth to the dataset's appeal, as handling imbalanced data is a practical concern in machine learning and data science.

Scalability: With a substantial number of instances and attributes, the dataset offers scalability for various analytical tasks, including classification, clustering, and regression, making it versatile for different research areas.

In summary, the Nike Women's Shoes dataset is captivating due to its applicability in understanding consumer behavior, e-commerce dynamics, and imbalanced data challenges. Its real-world relevance and scalability make it an engaging resource for researchers and practitioners alike.

4. Methodology

Environment Setup

- **Code Editor**: Google colab
- **Programming Language**: Python

Tools Used

- **Requests**: Used for sending HTTP GET requests to the Nike website.
- **Beautiful Soup**: Employed for parsing and navigating the HTML content of web pages.
- **Pandas**: Used for structuring and storing the scraped data in a tabular format.

5. Solution

The solution involves the following steps:

3.1 Sending HTTP Requests

We initiate the data collection process by sending HTTP GET requests to the Nike Women's Shoes section on the website. This step is essential to access the web pages and retrieve the HTML content for further analysis.

3.2 HTML Parsing

Once the HTML content is fetched, we utilize the **Beautiful Soup** library to parse and extract relevant data from the web pages. Beautiful Soup simplifies the process of navigating the HTML structure and locating specific elements.

3.3 Data Extraction

The key information, including product names, prices, URLs, and categories, is extracted from the HTML elements. We identify and extract these data points from the parsed HTML content.

3.4 Data Structuring

The extracted data is structured into a pandas DataFrame for easy manipulation and analysis. A DataFrame is a tabular data structure that allows us to store and work with structured data efficiently.

3.5 Data Display and Storage

We display the first 1000 rows of the DataFrame to provide an overview of the collected data. Additionally, we save the structured data as a CSV file named 'nike_womens_shoes.csv' for future reference and analysis.

6. Challenges and Outlook

Challenges

- **Rate Limiting**: There might be rate limiting or IP blocking mechanisms in place on the Nike website to prevent web scraping. Implementing rate limiting logic could be beneficial to avoid issues.
- **HTML Structure Changes**: If the structure of the website's HTML changes, the code may need to be updated to adapt to these changes.
- **Web Driver Location**: The location of the Chrome WebDriver is hardcoded, which may not be the same on all systems, affecting portability.

Outlook

- To make the code more robust, dynamic parsing of the page structure could be implemented, allowing it to adapt to changes in the website's layout.

- Consider using WebDriver manager libraries to handle the WebDriver's location and version, ensuring better portability across different systems.

7. Conclusion

This project successfully scraped valuable data from Nike's Women's Shoes section. The collected data can be further analyzed to gain insights into consumer behavior and e-commerce strategies. While challenges such as rate limiting and HTML structure changes exist, the project provides a foundation for future improvements and enhancements in web scraping techniques.

References

https://www.nike.com/w/womens-shoes-5e1x6zy7ok

https://www.scrapingbee.com/blog/web-scraping-101-with-python/

https://www.blog.datahut.co/post/scrape-indeed-using-selenium-and- beautifulsoup