Project Report: Web Scraping & Analysis of H&M Products

Introduction

In this project, we aimed to collect and analyze product data from multiple clothing websites, both international and Egyptian. Our goal was to gain insights into various product attributes and their impact on pricing. However, we encountered several challenges during the web scraping phase, which led us to focus on the **H&M website**, as it provided open access to its product data.

Data Collection & Challenges

- Initially, we attempted **web scraping** on multiple fashion websites, but many of them had strict access restrictions, preventing us from extracting data.
- We then shifted our focus to **H&M**, which allowed us to scrape its product listings without major restrictions.
- We successfully collected **4,536 product entries** across all available categories, including attributes such as:
    - **Item Title** (Product Name)
    - **Age Group**
    - **Garment Length**
    - **Presentation Product Type**
    - **Composition**
    - **Category**
    - **Price**

Data Exploration, Cleaning & Processing

After collecting the dataset, we performed an extensive **Exploratory Data Analysis (EDA)** and data cleaning process:

- **Handling Missing Values:** Checked for null values and either filled or removed them accordingly.
- **Removing Duplicates & Outliers:**
    - **Duplicates** were identified and removed to avoid redundant information.
    - **Outliers** in numerical data (e.g., Price, Garment Length) were removed to enhance data consistency.
- **Encoding Categorical Variables:**
    - **Nominal Features** (e.g., Item Title, Category): Applied **One-Hot Encoding** without using pd.get_dummies().
    - **Ordinal Features** (e.g., Age Group): Applied **Label Encoding** to maintain order.
- **Scaling Numerical Variables:** Applied **scaling** to all numerical features except Price, which remained in its original form to retain interpretability.
- **Final Data Size After Cleaning:** The dataset was **reduced to 1,822 entries** after removing duplicates and outliers.

Data Visualization & Key Insights

To better understand the dataset, we visualized key relationships between features using various techniques:

- **Heatmap (Feature Correlation Analysis)**
    - We plotted a heatmap to explore correlations between Price and other features.
    - **Finding:** The heatmap showed **no strong correlation** between Price and any other feature.

- **Price Distribution Analysis**
    - Created histograms and boxplots to examine how prices are distributed.
    - **Finding:** Some categories had extreme outliers, which we removed during cleaning.
- **Garment Length vs. Price**
    - We examined the relationship between **Garment Length** and **Price**.
    - **Finding:** While some variations were observed, Garment Length alone was not a strong determinant of Price.

Challenges & Limitations

1. **Access Restrictions on Fashion Websites**
    - Many popular clothing sites had anti-scraping measures, limiting data access.
    - H&M was chosen because it allowed scraping.
2. **Lack of Strong Correlation in Data**
    - Price did not show a strong correlation with any of the extracted features.
    - More features (e.g., brand reputation, seasonal demand) may be needed for better insights.

Conclusion

- This project successfully **scraped and analyzed 1,822 product entries** from H&M after cleaning, providing insights into pricing and product attributes.
- While we faced **challenges in data access and weak correlations**, we still gained valuable experience in web scraping, data processing, and visualization.