Data Quality and Web Scraping Report for Graham Watches

Introduction

This report provides an overview of the web scraping project conducted on the Graham Watches website. The primary goal was to collect detailed information about the watch models offered by Graham, including but not limited to reference numbers, prices, descriptions, and technical specifications. This data is critical for understanding the brand's offerings, pricing strategies, and product features.

Web Scraping Methodology

URL Identification: The initial step involved identifying the URLs of product listings and individual watch detail pages.

Data Extraction: Using Python with libraries such as BeautifulSoup and Requests, the script extracted relevant data from the HTML content of each page. This included model names, prices, descriptions, specifications, and images.

Data Storage: Extracted data was stored in a structured format using pandas DataFrames, which were then exported to a CSV file for further analysis.

Data Quality Assessment

Accuracy: The data appears to reflect the website's content accurately, with model details correctly matched to their respective attributes. Manual spot checks against the website confirmed the reliability of the scraping script.

Completeness: All targeted data fields were successfully extracted for the majority of watch models. However, some fields were occasionally missing due to variations in the website's HTML structure across different pages.

Challenges and Limitations

Inconsistent HTML Structures: The website's varying HTML layouts for different watch models presented a significant challenge. This variability led to difficulties in uniformly extracting data across all product pages, impacting the dataset's completeness and consistency.

PDF Technical Sheets Extraction: Extracting detailed specifications from PDF technical sheets proved challenging due to their diverse formats. The reliance on regex for text extraction was not always effective, especially with PDFs that contained complex layouts or graphical elements. This limitation affected the accuracy and consistency of extracting key specifications like movement details and power reserve.

Conclusion

The web scraping project successfully collected a comprehensive dataset from the Graham Watches website, offering valuable insights into the brand's product lineup. While the dataset exhibits high quality in terms of accuracy and consistency, ongoing efforts in data cleaning, validation, and periodic updates are essential to address completeness and timeliness.

Future work will focus on refining the scraping process to handle dynamic content more efficiently, automating data validation checks, and establishing a routine for updating the dataset to reflect the latest website content.