BAX 422 Data Design Final Project

E-commerce SSENSE

Master of Science in Business Analytics
University of California - Davis

Huixin ‘Blessia’ Li
Hanish Singla
Jialu ‘Carol’ Wang

Table of Contents

Executive Summary

Our project involved scraping data from SSENSE and storing the data in a MongoDB database. By extracting ten specific fields from each product page, we created a comprehensive and structured view of the high-end women's wear items on sale. This database offers great business value, as it enables efficient analysis of branding, consumer behavior, pricing strategies, inventory management, fashion trends, and product offerings. The dataset also highlights the importance of data-driven decision making in the high-end fashion market, and our database provides a valuable tool for gaining these insights and can be used to address a wide range of business problems and scenarios.

Background & Domain Knowledge

The data used in this report was obtained by scraping information from SSENSE.com. SSENSE is a Montreal-based e-commerce platform that specializes in high-end fashion and luxury retail. The company was founded in 2003 by Rami Atallah, and has since grown to become one of the most popular destinations for online shopping in the fashion industry. As a leading e-commerce platform, SSENSE offers a curated selection of designer clothing, shoes, and accessories from over 500 different brands.

The fashion and luxury retail industry is highly competitive, with numerous companies vying for market share. Some of SSENSE's competitors include other high-end e-commerce platforms such as Net-a-Porter, Farfetch, and Matches Fashion, as well as traditional brick-and-mortar retailers like Neiman Marcus and Saks Fifth Avenue. In order to remain competitive, SSENSE

has focused on leveraging technology to provide customers with a unique and engaging shopping experience. This includes the use of artificial intelligence and machine learning to recommend products to customers based on their browsing and purchase history, as well as the incorporation of augmented reality and other immersive technologies to allow customers to visualize products in new and innovative ways. Our project aims to provide an organized and meaningful database that can be used to better understand consumers' behaviors, adjust pricing strategies, and manage inventory in different business scenarios.

Data Design

Data Sources

The data we used originated from SSENSE. Specifically, the data was collected by scraping information on women's wear items that were on sale at the time of data collection. The decision to focus on women's wear was based on the desire to gain better insights into this specific segment of the high-end fashion market, which represents a significant portion of the online retailing market share. According to the Global Fashion Industry Statistics: "revenue generated from the women's apparel market is much higher than that of men's or children's apparel. In 2018, womenswear made up more than half (53%) of global fashion retail spending. Spending on menswear accounted for 31% and childrenswear for 16%." So we believe examining this segment is crucial for understanding consumer behavior and trends in this market. The purpose of collecting this data was to gain insights into the pricing and product offerings of high-end fashion brands on the SSENSE platform and to identify trends and patterns that could be useful in making informed business decisions.

Web-Scraping Routines

The web scraping process involved accessing the HTML code of the SSENSE website and extracting relevant information. By breaking down the scraping processes, detailed steps are as following:

1. Searched for “womenswear” and “on sale” items on SSENSE.com
2. Saved the first two pages from searching results as HTML files, with each page containing 120 items, 240 items in total
3. Extracted the links to each individual product page from step 2’s saved HTML files by using CSS selectors with Beautiful Soup, so 240 URLs in total
4. Went through every URL and saved all 240 items pages as HTML files
5. Again using CSS selectors with Beautiful Soup, we extracted following specific elements from the HTML file for each women's wear item on sale:
    a. Item Brand: The brand of the product
    b. Item Image: The URL of the product image on the SSENSE website
    c. Item Title: The name of the product
    d. Original Price: The original price of the product before any discounts were applied
    e. Discount Price: The price of the product after any discounts were applied
    f. Discounts: The discounts percentage off
    g. Item Color: The color of the product
    h. Item Material: The material or fabric used to make the product
    i. Item Origin: The country where the product was made

   j. Item SKU: Stock Keeping Unit (SKU) number, which is a unique identifier assigned to each product to help manage inventory and track sales

6. Finally, we inserted all 240 items information into MongoDB's collection named "ssense_database"

The dataset was organized with ten features, each of which corresponds to a different aspect of the product information. The final output in MongoDB is a ready-to-use dataset for further business analysis.

Database Design and Dataset Explanation

The dataset for this report was designed to provide a comprehensive overview of women's wear items on sale at SSENSE.com. The dataset includes information on a wide range of brands and products, such as clothing, shoes, and accessories. To ensure the accuracy and consistency of the data, several procedures were taken during the data scraping and designing process:

1. All price related information was converted into numeric format. Specifically, three features "Original Price", "Discount Price", and "Discounts" were converted to float from string for easier calculating and analyzing purposes. "Discounts" were further divided by 100 for better understanding.
2. Besides the "_id" automatically generated by MongoDB, we added another index on "item SKU." Each product on the SSENSE website has a unique SKU number that identifies it within the company's inventory management system. And the SKU number is a combination of letters and numbers. By adding an index on the "item SKU" field, one can similarly search for documents by their SKU, and MongoDB can quickly find the

document matching the search criteria. With both fields indexed, one can also perform more complex queries that involve both the _id and item_sku fields.

3. All other features were stored in string format after removing meaningless and unnecessary parts, such as “Made in,” “USD,” “$,” and “%” etc. for more efficient use.

We chose to extract these ten features because they provided relevant details of the women's wear items on sale, and could be used to gain insights into the pricing and product information from different potential aspects.

We chose to save information as HTML files because it allowed us to easily view and scrape the data in a human-readable format. By saving each item as an individual HTML file, we were able to view the data in a web browser and manually confirm that the data had been scraped correctly. We used Beautiful Soup as our web scraping tool because it is a powerful and flexible library that allows for easy parsing, which makes it easy to prepare the data for storage in MongoDB. We chose to use MongoDB as our database management system for storing the scraped data due to its flexibility, scalability, and ability to handle unstructured data. MongoDB is a NoSQL database that is designed to store and manage large amounts of unstructured or semi-structured data, making it a good choice for storing the scraped data from SSENSE.

Database Implementation

The dataset we have collected contains detailed and holistic information for each product, which can help address various business questions and provide recommendations for multiple business scenarios. It can be leveraged by employees, managerial levels, and public users to gain insights

from different perspectives by using data-driven approaches. Here are some potential business problems that the dataset can help to address:

1. Brand related:
    a. Which brands frequently have products on sale?
    b. Which brands seldom participate in discounting activities?
    c. Which brands would be good to partner with?
2. Sales volume related:
    a. Does reducing the price of an item promote sales?
    b. How effective are marketing or promotional campaigns when items are on sale?
    c. What are the sales volumes and revenue generated before and after on-sale promotions?
3. Pricing related:
    a. How does pricing change over time?
    b. How does seasonality affect pricing?
    c. What are the pricing strategies for specific brands or categories?
4. Fashion Trend related:
    a. What are the most popular designs or styles nowadays?
    b. Are there any designs or styles that customers do not like?
5. Inventory management related:
    a. Which products should be stocked, and which brands or categories should have priority?
    b. How quickly should new products be introduced, and how quickly should they be put on sale?

c. Which products should be on final sale (no return)?

By analyzing the data and answering these business questions, our dataset can provide recommendations for decision-making in various business scenarios. For example, it can be used to optimize inventory management by identifying top-performing products and brands, and to adjust pricing strategies based on seasonal trends and customer preferences. It can also be used to inform marketing and promotional campaigns by identifying the most effective strategies for promoting on-sale items. The dataset can also be used to identify trends and patterns in consumer behavior and preferences, which can inform marketing campaigns and product development.

Additionally, the chosen database implementation MongoDB surpasses traditional SQL by several advantages. Firstly, MongoDB's document-based model allows for greater flexibility in data storage, as it does not require a predefined schema like SQL databases, allowing for storing unstructured data in a structured format and making it easy to query and analyze the data. Secondly, MongoDB's ability to handle a wide range of data types without the need for complex data normalization or indexing. Thirdly, MongoDB's built-in tools and APIs make it easy to perform complex data manipulations and aggregations on the data, which can be more challenging in SQL databases. Finally, MongoDB's ability to scale horizontally makes it a more scalable option than SQL databases. MongoDB can handle the addition of more data or more fields, without sacrificing performance or requiring significant changes to the database architecture.

Conclusion

By scraping data from SSENSE, a leading e-commerce in the fashion and luxury retail industry, and implementing data into MongoDB, our project provides a comprehensive and reliable overview of women's wear items on sale at SSENSE. With a structured format and database design choices, we were able to create an informative dataset that can be leveraged to gain insights into the high-end fashion market and to make data-driven decisions across various aspects of the business, from branding to sales to inventory management. The dataset can also be used to identify trends and patterns in consumer behavior and preferences, which can inform marketing campaigns and product development.

References

"Global Fashion Industry Statistics." FashionUnited,

https://fashionunited.com/global-fashion-industry-statistics.

Appendix: Database Overview

	brand	original price	discount price	discounts	item title	item color	item material	item origin	item image	item sku
9471	Jacquemus	370.0	104.0	0.72	Brown 'Le Body Yauco' Bod	Brown	85% viscose, 15	Portugal	https://img.ssen	222
9471	Versace Underw	395.0	162.0	0.59	Black Printed Bikini Top	Black/White	Body: 78% polye	Italy	https://img.ssen	222
9471	Jacquemus	140.0	69.0	0.51	Navy 'Le Bonnet' Beanie	Navy	80% wool, 20%	Italy	https://img.ssen	222
9471	Ann Demeuleme	590.0	319.0	0.46	Black Elisabeth Shirt	Black	100% cotton.	Italy	https://img.ssen	222
9471	Ann Demeuleme	590.0	260.0	0.56	White Elizabeth Shirt	White	100% cotton.	Italy	https://img.ssen	222
9471	Versace Underw	350.0	144.0	0.59	Black Printed Bikini Bottom	Black/White	Body: 78% polye	Italy	https://img.ssen	222
9471	Totême	600.0	378.0	0.37	Gray Wrap Midi Dress	Grey melange	100% wool.	China	https://img.ssen	222
9471	Chloé	550.0	347.0	0.37	Green Betty Boots	Dark emerald	Upper: rubber. S	Romania	https://img.ssen	222
9471	Totême	420.0	281.0	0.33	Off-White Mini Cable Pullov	Off-white	100% merino wo	China	https://img.ssen	221
9471	We11done	420.0	151.0	0.64	Gray Basic Washed Sweats	Charcoal	100% cotton.	South Korea	https://img.ssen	222
9471	We11done	295.0	115.0	0.61	Gray Old School Campus T-	Grey	100% cotton.	South Korea	https://img.ssen	222
9471	We11done	455.0	132.0	0.71	Gray Printed Sweatshirt	Charcoal	100% cotton.	South Korea	https://img.ssen	222
9471	We11done	455.0	137.0	0.7	Black Printed Sweatshirt	Black	100% cotton.	South Korea	https://img.ssen	222
9471	We11done	320.0	102.0	0.68	White Cursive Sweatshirt	White	100% cotton.	South Korea	https://img.ssen	222
9471	1XBLUE	240.0	106.0	0.56	SSENSE Exclusive Black Te	Black	Body: 100% woc	China	https://img.ssen	222
9471	Eckhaus Latta	190.0	139.0	0.27	Green Lapped Baby Turtlen	Fishbowl	100% cotton.	Peru	https://img.ssen	222
9471	AMI Alexandre M	200.0	130.0	0.35	Black Ami de Cœur Long Sl	Black	Body: 100% orga	Portugal	https://img.ssen	221
9471	AMI Alexandre M	200.0	122.0	0.39	White Ami de Cœur Long S	White	Body: 100% orga	Portugal	https://img.ssen	221

Please refer to our “final_project.py” file for detailed code information.