DATA DESIGN AND REPRESENTATION FINAL PROJECT

Nike Unisex Footwear Trends: A Web Scraping Data Analysis

Submitted by :

Rachita Harit
Blessy Chinthapalli
Yongija 'Will' Heng

EXECUTIVE SUMMARY

In an advanced analysis initiative, our team utilized sophisticated web scraping techniques to collect and categorize data on Nike's unisex footwear into sale and non-sale segments. Utilizing a structured MySQL database, we applied meticulous data manipulation, wrangling and modeling to craft a high-quality dataset. This dataset then served as the foundation for a Machine Learning (ML) project aimed at uncovering common themes within Nike product descriptions through an unsupervised ML model that emphasized frequency over sequence. Our analysis revealed key themes that are crucial for enhancing Nike's marketing and product development strategies. This data-driven approach not only provides valuable insights into Nike's product positioning but also pioneers the use of analytics in retail strategy optimization.

BACKGROUND

In the fast-paced and highly competitive athletic footwear market, understanding customer preferences and market trends is paramount for a leading brand like Nike. The brand is renowned for its vast range of unisex shoes, catering to a wide spectrum of sports and lifestyle choices. Keeping abreast of market demands and ensuring the alignment of product offerings with consumer expectations is a critical business imperative . That's the reason why we choose to select an e-commerce website to delve deeper and understand the trends

CONTEXT

To stay at the forefront of the industry, our team has implemented a data-driven approach to understand Nike's market positioning better. We systematically collected detailed product data from Nike's online catalog, specifically focusing on the unisex footwear category. This data spans both sale and non-sale items, capturing nuances in pricing, product descriptions, reviews, ratings, and more. The extracted data was then meticulously processed and organized within a MySQL database, forming the basis for in-depth analysis.

DOMAIN KNOWLEDGE

The athletic footwear industry is characterized by fierce competition and swift changes in consumer trends. Products are not only designed for performance but also as fashion statements, requiring a fragile balance between functionality and aesthetic appeal. The branding and marketing of such products must resonate with diverse consumer bases, making the understanding of underlying themes in product descriptions vital for crafting compelling marketing narratives.

Nike, being a household name, holds an extensive collection that goes beyond mere footwear, encompassing a lifestyle. With product lines named after athletes, technologies developed for performance enhancement, and sustainability becoming increasingly important, the company's portfolio reflects a commitment to innovation and social responsibility.

INTRODUCTION TO DATA

Data Sources:

Nike's official online retail store - Nike's official site

The data source is Nike's sales and non-sale webpage, showcasing unisex shoes with and without discounts. It includes a variety of athletic and lifestyle footwear, displaying sale prices, discounts, and product images. This page is ideal for scraping sales trends and pricing data, allows for filtering by categories such as gender and sport, offering insights into consumer preferences and pricing strategies.

WEB SCRAPING ROUTINE(S)

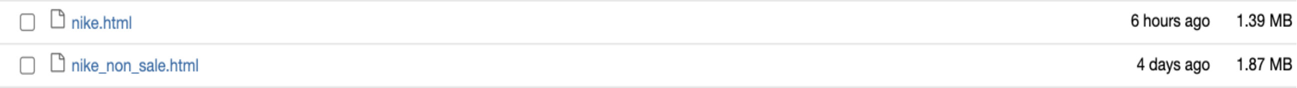
Our web scraping routine targets Nike’s unisex shoe sales, structured into distinct stages: initialization, data extraction, content saving, and subsequent parsing. Initiated with Selenium, the process navigates Nike’s website, interacts with specific web elements to filter and sort products, and employs infinite scrolling to ensure full page content is loaded. The page source is then saved as a local HTML file. This file is parsed using BeautifulSoup to extract detailed product data. The extracted information is ready for further analysis or database storage**.**

I. Scripting and Automation:

The automation process involves a scripted sequence executed through the Selenium WebDriver. Initially, required Selenium modules are imported to facilitate browser interaction. Then, a Chrome WebDriver instance is initiated, which waits up to 10 seconds for page elements to load. The automated browser navigates to Nike's website, where it sequentially clicks through the site’s categories to reach the unisex shoes on sale. To accommodate web pages with lazy loading, an infinite scroll is implemented, ensuring all products are fully loaded for data capture. The same is repeated for the shoes not on sale under the unisex category again. This process sets up the stage for a comprehensive scrape of product listings.

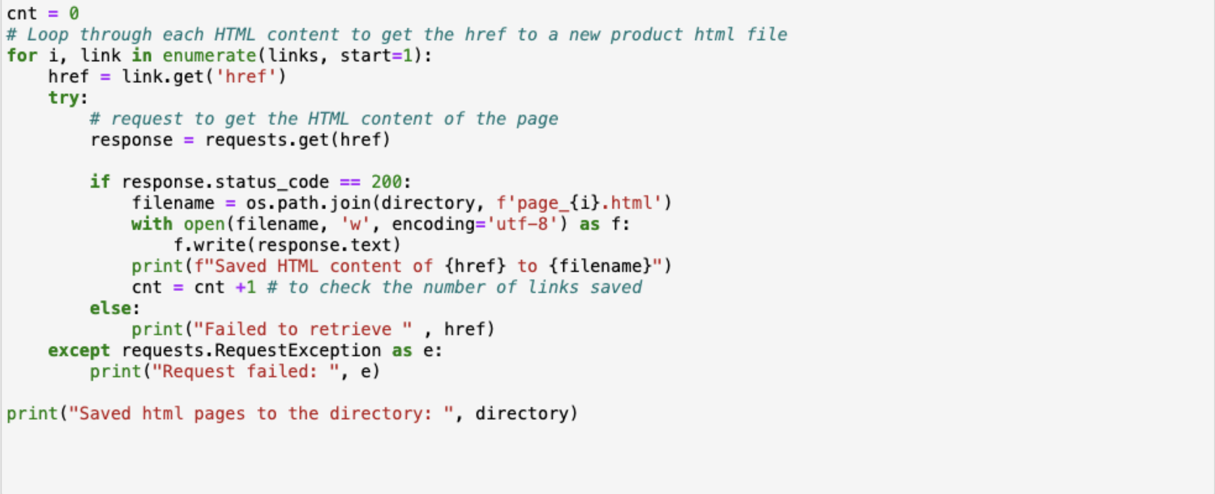
II. Data Extraction and HTML Content Saving:

Step 1 : First, the webpage's HTML source code is saved to a local file named "nike.html" and “nike_non_sale.html” using Python's file handling with UTF-8 encoding. After a short pause, the web browser is closed

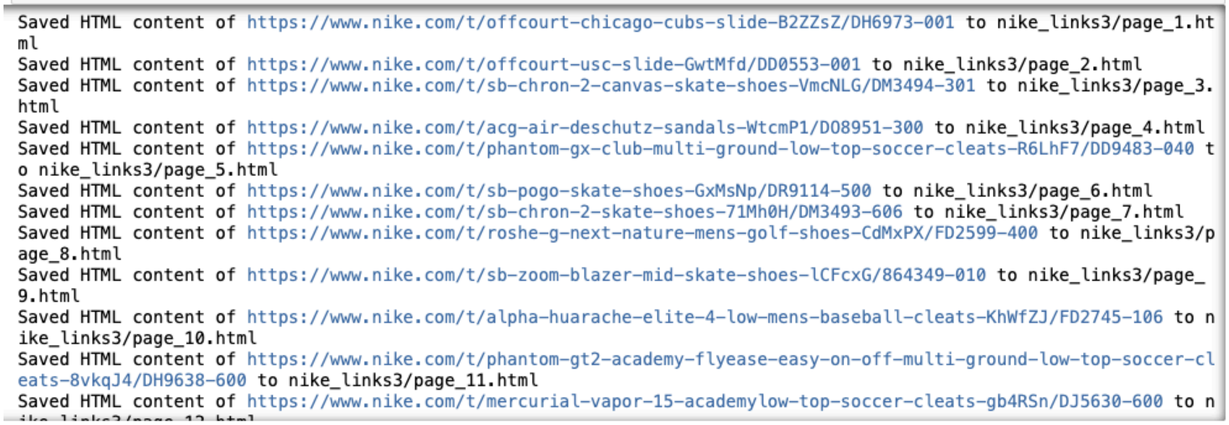


□	nike.html	6 hours ago	1.39 MB
□	nike_non_sale.html	4 days ago	1.87 MB

Step 2 : The second step involves parsing this saved HTML file with BeautifulSoup to extract relevant data, such as product links, for further analysis.



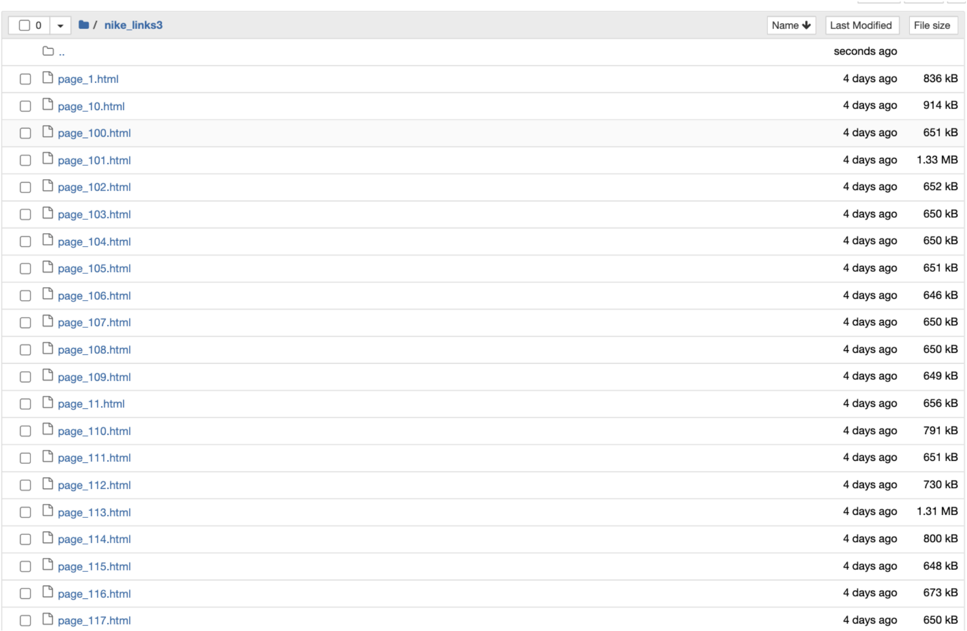
```
cnt = 0
# Loop through each HTML content to get the href to a new product html file
for i, link in enumerate(links, start=1):
    href = link.get('href')
    try:
        # request to get the HTML content of the page
        response = requests.get(href)

        if response.status_code == 200:
            filename = os.path.join(directory, f'page_{i}.html')
            with open(filename, 'w', encoding='utf-8') as f:
                f.write(response.text)
            print(f"Saved HTML content of {href} to {filename}")
            cnt = cnt +1 # to check the number of links saved
        else:
            print("Failed to retrieve " , href)
    except requests.RequestException as e:
        print("Request failed: ", e)

print("Saved html pages to the directory: ", directory)
```



```
Saved HTML content of https://www.nike.com/t/offcourt-chicago-cubs-slide-B2ZZsZ/DH6973-001 to nike_links3/page_1.ht
ml
Saved HTML content of https://www.nike.com/t/offcourt-usc-slide-GwtMfd/DD0553-001 to nike_links3/page_2.html
Saved HTML content of https://www.nike.com/t/sb-chron-2-canvas-skate-shoes-VmcNLG/DM3494-301 to nike_links3/page_3.
html
Saved HTML content of https://www.nike.com/t/acg-air-deschutz-sandals-WtcmP1/DO8951-300 to nike_links3/page_4.html
Saved HTML content of https://www.nike.com/t/phantom-gx-club-multi-ground-low-top-soccer-cleats-R6LhF7/DD9483-040 t
o nike_links3/page_5.html
Saved HTML content of https://www.nike.com/t/sb-pogo-skate-shoes-GxMsNp/DR9114-500 to nike_links3/page_6.html
Saved HTML content of https://www.nike.com/t/sb-chron-2-skate-shoes-71Mh0H/DM3493-606 to nike_links3/page_7.html
Saved HTML content of https://www.nike.com/t/roshe-g-next-nature-mens-golf-shoes-CdMxPX/FD2599-400 to nike_links3/p
age_8.html
Saved HTML content of https://www.nike.com/t/sb-zoom-blazer-mid-skate-shoes-lCFcxG/864349-010 to nike_links3/page_
9.html
Saved HTML content of https://www.nike.com/t/alpha-huarache-elite-4-low-mens-baseball-cleats-KhWfZJ/FD2745-106 to n
ike_links3/page_10.html
Saved HTML content of https://www.nike.com/t/phantom-gt2-academy-flyease-easy-on-off-multi-ground-low-top-soccer-cl
eats-8vkqJ4/DH9638-600 to nike_links3/page_11.html
Saved HTML content of https://www.nike.com/t/mercurial-vapor-15-academylow-top-soccer-cleats-gb4RSn/DJ5630-600 to n
```

III. Extracting Product Attributes for Database Structuring:

The automation script navigates to a saved directory of Nike product pages, where it reads the HTML content and employs BeautifulSoup to parse it. The script meticulously extracts key product details including **title**, **category**, **pricing**, **discounts**, **descriptions**, **product IDs**, **number of reviews**, **ratings** and **colors**. These details are formatted into a dictionary per product, which is then collated into a list—forming a structured dataset ready for database integration, ensuring a robust foundation for subsequent data analysis.



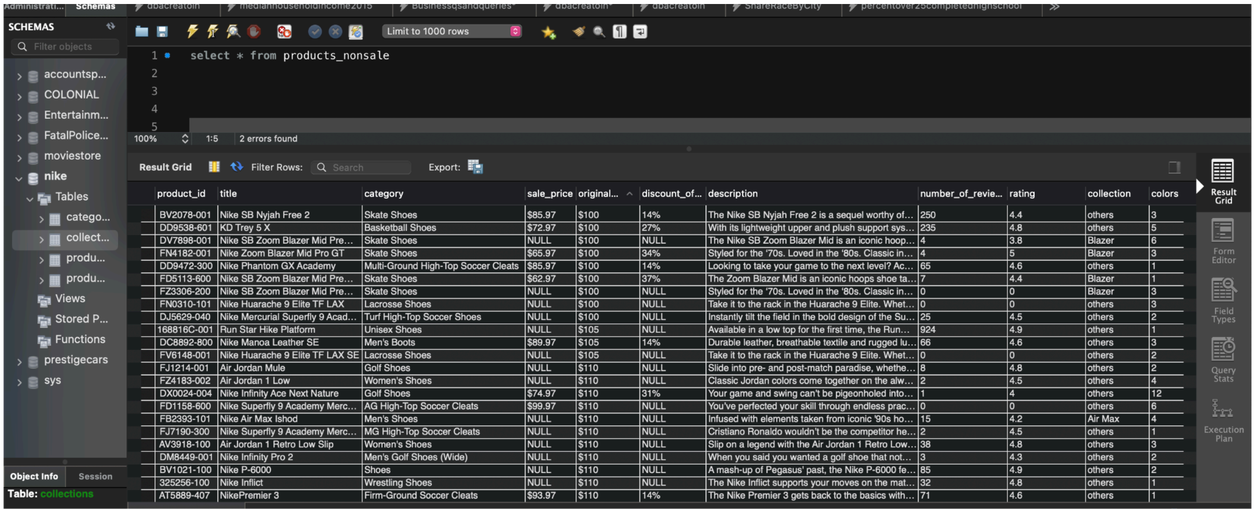
☐ 0 ▾ / nike_links3	Name ↓ Last Modified	File size
..	seconds ago	
☐ page_1.html	4 days ago	836 kB
☐ page_10.html	4 days ago	914 kB
☐ page_100.html	4 days ago	651 kB
☐ page_101.html	4 days ago	1.33 MB
☐ page_102.html	4 days ago	652 kB
☐ page_103.html	4 days ago	650 kB
☐ page_104.html	4 days ago	650 kB
☐ page_105.html	4 days ago	651 kB
☐ page_106.html	4 days ago	646 kB
☐ page_107.html	4 days ago	650 kB
☐ page_108.html	4 days ago	650 kB
☐ page_109.html	4 days ago	649 kB
☐ page_11.html	4 days ago	656 kB
☐ page_110.html	4 days ago	791 kB
☐ page_111.html	4 days ago	651 kB
☐ page_112.html	4 days ago	730 kB
☐ page_113.html	4 days ago	1.31 MB
☐ page_114.html	4 days ago	800 kB
☐ page_115.html	4 days ago	648 kB
☐ page_116.html	4 days ago	673 kB
☐ page_117.html	4 days ago	650 kB

IV. Database Insertion:

This step pertains to database management for storing scraped data. Initially, the script defines the structure of a MySQL database and a table with appropriate data fields like product ID, title, category, prices, discount, number of reviews, and rating. Then, using the `mysql.connector` library, it connects to the MySQL server, creates the database if it doesn't exist, and sets up the table with the defined structure. The script handles potential errors during the process and confirms successful database and table creation with a print statement. This structured database is essential for efficient storage, retrieval, and analysis of the scraped product data.

V. Data Insertion into Tables:

This step involves inserting scraped data into the MySQL database. The script connects to the MySQL server and selects the 'nike' database. It then prepares an SQL insert query to add product data into the 'products and ‘products_nonsale’' tables. Looping through the list of product dictionaries, it inserts each product's data into the table, committing the transaction for each entry. Upon successful insertion, it prints the count of rows added and closes the connection to the database. If an error occurs during insertion, it catches and prints the error message.



```
select * from products_nonsale
```

product_id	title	category	sale_price	original...	discount_of...	description	number_of_revie...	rating	collection	colors
BV2078-001	Nike SB Nyjah Free 2	Skate Shoes	$85.97	$100	14%	The Nike SB Nyjah Free 2 is a sequel worthy of...	250	4.4	others	3
DD9538-601	KD Trey 5 X	Basketball Shoes	$72.97	$100	27%	With its lightweight upper and plush support sys...	235	4.8	others	5
DV7898-001	Nike SB Zoom Blazer Mid Pre...	Skate Shoes	NULL	$100	NULL	The Nike SB Zoom Blazer Mid is an iconic hoop...	4	3.8	Blazer	6
FN4182-001	Nike Zoom Blazer Mid Pro GT	Skate Shoes	$65.97	$100	34%	Styled for the ‘70s. Loved in the ‘80s. Classic in...	4	5	Blazer	3
DD9472-300	Nike Phantom GX Academy	Multi-Ground High-Top Soccer Cleats	$85.97	$100	14%	Looking to take your game to the next level? Ac...	65	4.6	others	1
FD5113-600	Nike SB Zoom Blazer Mid Pre...	Skate Shoes	$62.97	$100	37%	The Zoom Blazer Mid is an iconic hoops shoe ta...	7	4.4	Blazer	1
FZ3306-200	Nike SB Zoom Blazer Mid Pre...	Skate Shoes	NULL	$100	NULL	Styled for the ‘70s. Loved in the ‘80s. Classic in...	0	0	Blazer	3
FN0310-101	Nike Huarache 9 Elite TF LAX	Lacrosse Shoes	NULL	$100	NULL	Take it to the rack in the Huarache 9 Elite. Whet...	0	0	others	3
DJ5629-040	Nike Mercurial Superfly 9 Acad...	Turf High-Top Soccer Shoes	NULL	$100	NULL	Instantly tilt the field in the bold design of the Su...	25	4.5	others	2
168816C-001	Run Star Hike Platform	Unisex Shoes	NULL	$105	NULL	Available in a low top for the first time, the Run...	924	4.9	others	1
DC8892-800	Nike Manoa Leather SE	Men's Boots	$89.97	$105	14%	Durable leather, breathable textile and rugged lu...	66	4.6	others	3
FV6148-001	Nike Huarache 9 Elite TF LAX SE	Lacrosse Shoes	NULL	$105	NULL	Take it to the rack in the Huarache 9 Elite. Whet...	0	0	others	2
FJ1214-001	Air Jordan Mule	Golf Shoes	NULL	$110	NULL	Slide into pre- and post-match paradise, whethe...	8	4.8	others	2
FZ4183-002	Air Jordan 1 Low	Women's Shoes	NULL	$110	NULL	Classic Jordan colors come together on the alw...	2	4.5	others	4
DX0024-004	Nike Infinity Ace Next Nature	Golf Shoes	$74.97	$110	31%	Your game and swing can’t be pigeonholed into...	1	4	others	12
FD1158-600	Nike Superfly 9 Academy Merc...	AG High-Top Soccer Cleats	$99.97	$110	NULL	You’ve perfected your skill through endless prac...	0	0	others	6
FB2393-101	Nike Air Max Ishod	Men's Shoes	NULL	$110	NULL	Infused with elements taken from iconic '90s ho...	15	4.2	Air Max	4
FJ7190-300	Nike Superfly 9 Academy Merc...	MG High-Top Soccer Cleats	NULL	$110	NULL	Cristiano Ronaldo wouldn’t be the competitor he...	2	4.5	others	1
AV3918-100	Air Jordan 1 Retro Low Slip	Women's Shoes	NULL	$110	NULL	Slip on a legend with the Air Jordan 1 Retro Low...	38	4.8	others	3
DM8449-001	Nike Infinity Pro 2	Men's Golf Shoes (Wide)	NULL	$110	NULL	When you said you wanted a golf shoe that not...	3	4.3	others	2
BV1021-100	Nike P-6000	Shoes	NULL	$110	NULL	A mash-up of Pegasus' past, the Nike P-6000 fe...	85	4.9	others	2
325256-100	Nike Inflict	Wrestling Shoes	NULL	$110	NULL	The Nike Inflict supports your moves on the mat...	32	4.8	others	1
AT5889-407	NikePremier 3	Firm-Ground Soccer Cleats	$93.97	$110	14%	The Nike Premier 3 gets back to the basics with...	71	4.6	others	1

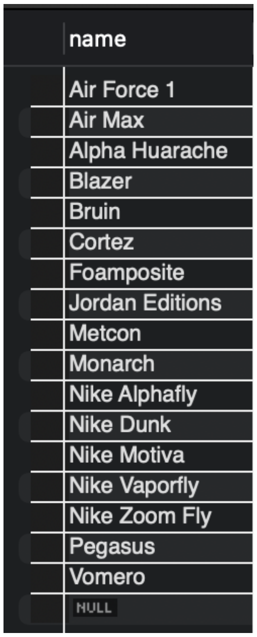
DATA WRANGLING

This step is focused on data wrangling within a MySQL database to make the product data more analytically useful. The process:

- **Database Connection:** The script establishes a connection to the MySQL server and selects the 'nike' database.
- **Column Check/Addition:** IIt checks if a 'collection' column exists in the 'products' table and adds it if it's not present. This step ensures that the database schema includes all necessary fields for analysis.
- **Data Update :** The script retrieves product IDs and titles from the database and iterates through them. It initializes the collection name to 'others' and then updates this value if the product title contains a known collection name.

- **Updating the Table:** An update query is executed for each product to set the collection name accordingly in the database.

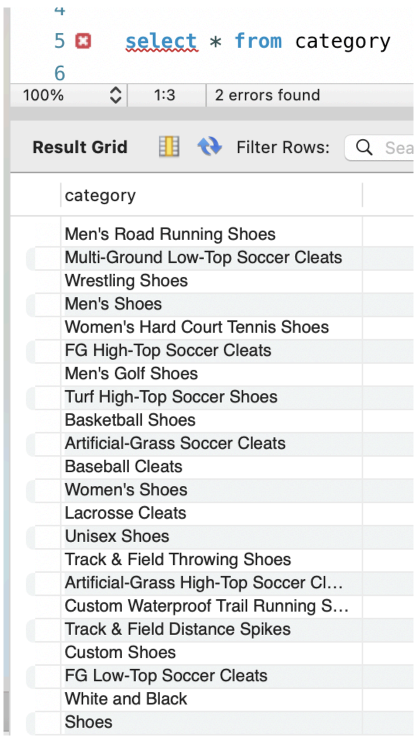
Upon successful completion of updates, the script confirms that the 'collection' column is updated and then closes the MySQL connection. If an error occurs, it catches and prints the error message. This wrangling step is crucial for categorizing products into collections for later analysis.



name
Air Force 1
Air Max
Alpha Huarache
Blazer
Bruin
Cortez
Foamposite
Jordan Editions
Metcon
Monarch
Nike Alphafly
Nike Dunk
Nike Motiva
Nike Vaporfly
Nike Zoom Fly
Pegasus
Vomero
NULL

DATABASE MODELING:

The data modeling in this project involves establishing relational structures within the MySQL database to better organize and connect various product attributes. This includes creating tables with specified cardinality, where the 'product.collection' field is linked to 'collection.name', and 'category.id', with 'style id' serving as primary keys. These relationships are carefully defined to ensure integrity and facilitate complex queries, which will support more nuanced analyses such as identifying product trends and customer preferences.



```
select * from category
```

2 errors found

category
Men's Road Running Shoes
Multi-Ground Low-Top Soccer Cleats
Wrestling Shoes
Men's Shoes
Women's Hard Court Tennis Shoes
FG High-Top Soccer Cleats
Men's Golf Shoes
Turf High-Top Soccer Shoes
Basketball Shoes
Artificial-Grass Soccer Cleats
Baseball Cleats
Women's Shoes
Lacrosse Cleats
Unisex Shoes
Track & Field Throwing Shoes
Artificial-Grass High-Top Soccer Cl…
Custom Waterproof Trail Running S…
Track & Field Distance Spikes
Custom Shoes
FG Low-Top Soccer Cleats
White and Black
Shoes

DATABASE CHOICE

MySQL was chosen for its reliability, strong transactional support, and ease of integration with a wide array of applications, which is highly beneficial for structuring and querying data in web scraping endeavors. Its predefined schema aligns well with the structured nature of product data, such as categories, prices, and ratings, that our project requires. MySQL’s robustness in handling structured data ensures data integrity and facilitates complex queries, which are necessary for the analysis of product trends and consumer preferences. Furthermore, MySQL's widespread adoption and extensive community support provide an abundance of tools and resources for troubleshooting and optimization. These attributes make MySQL a suitable database management system for our project, which demands regular data updates, structured storage, and quick data retrieval for ongoing analysis.

SUMMARY & CONCLUSION

Our project demonstrates the effective use of web scraping to analyze market trends within the athletic footwear industry, focusing on Nike's unisex shoe category. Utilizing Python’s Selenium for automation and BeautifulSoup for parsing, we gathered detailed product data, including prices, descriptions, and customer reviews. This data underwent meticulous wrangling to ensure structure and uniformity suitable for database storage.Throughout the process, MySQL was employed for its robust data management capabilities, hosting the well-structured data essential for insightful analysis. By capturing and organizing data into a relational database, we facilitated complex queries and data integrity essential for our analytical goals.The end result is a rich dataset housed in MySQL, ready for comprehensive analysis to uncover trends, preferences, and opportunities in Nike’s product offerings. Our approach sets the groundwork for strategic decision-making in product marketing and development, leveraging data-driven insights to maintain Nike's competitive edge in the market.

BUSINESS IMPLICATIONS:

- **Targeted Product Strategy:** By analyzing text from product descriptions, Nike gains insights into the themes that resonate with their audience. This can influence design, marketing campaigns, and inventory decisions.
- **Customer-Centric Innovation:** Identifying prevalent themes helps Nike to align its products with consumer trends, potentially enhancing customer satisfaction and loyalty.
- **Gap Identification for Market Advantage:** The project reveals any gaps between customer expectations and current product features, allowing Nike to address these and gain a competitive edge.

BUSINESS QUESTIONS ANSWERED

The data and analysis provide answers to critical business questions:

1. **Which themes in product descriptions are most common and how do they influence sales?**
2. **Are there patterns in customer reviews that can guide product improvements?**
3. **What is the impact of price and discount strategies on consumer perception and sales volume?**

The project demonstrates the strategic role of web scraping and machine learning in product trend analysis and marketing within the retail sector. For Nike, this approach not only augments the understanding of product positioning but also drives data-backed decisions for enhancing market presence and customer engagement. This structured and systematic analysis via web scraping is a strategic tool for ongoing product and marketing optimization.

NEXT STEPS: ML PROJECT

Next, we'll advance to the ML Project phase, where we'll apply an unsupervised machine learning model to analyze text in Nike's product descriptions, aiming to identify prevalent themes. Leveraging LDA (Latent Dirichlet Allocation) and LSTM (Long Short-Term Memory) networks, we'll dissect the frequency of word occurrences, bypassing their order, to gain insights for marketing strategies and feature enhancement.

APPENDIX :

Ways to use scroll using selenium - https://www.browserstack.com/guide/selenium-scroll-tutorial

3 rd party review for Nike - https://www.consumeraffairs.com/sporting_goods/nike.html