Project Title: Analyzing trends for Sole Haven Store Shoe company using PySpark.

Team Members: 1) Shamit Kotak
2) Divya Jain
3) Ishita Joshi

ORIGINAL WORK STATEMENT: We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
1	Shamit Kotak	Shamit Kotak
2	Divya Jain	Divya Jain
3	Ishita Joshi	Ishita Joshi

Executive Summary:

In this project we have tried to analyze a comprehensive dataset encompassing shoe products, aiming to glean insights into consumer behavior and optimize profitability strategies for “Sole Haven”, which is a shoe retailer company. Initially, the data preprocessing revealed the presence of null values in some of the potential variables, prompting their removal to ensure the accuracy of subsequent analyses. Further, by creating a new metric, “units sold”, we accurately gauged product popularity and customer preferences. This newfound understanding of customer demands facilitated tailored product offerings and marketing tactics which can be implemented by Sole Haven to enhance its competitive edge.

A pivotal discovery emerged through the synthesis of price and units data sold, unveiling the revenue potential of individual products. Leveraging this insight, we prioritized the high performing items and refined profitability strategies. Furthermore, performing advanced analytics techniques such as k-means clustering provided deeper segmentation insights, enabling the identification of distinct customer segments and their preferences. The integration of graph analytics facilitated comprehensive data exploration and visualization, with actionable insights for informed decision making. Overall, our project’s approach to data-driven decision-making holds promise for enhancing retail profitability and customer satisfaction, underscoring the significance of leveraging data analytics in business strategies.

Data Description

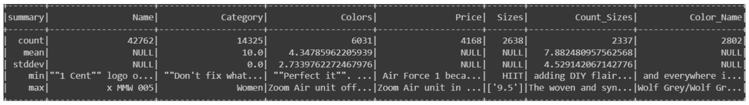
Data source: Kaggle, Self Generated for Project Purpose

Description:

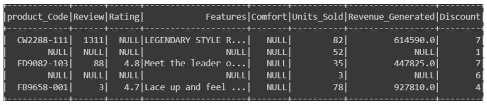
Feature Name	Feature Detail	Feature Type
Name	shoe model name	Categorical
Category	shoe category	Categorical
Colors	total number of color options	Numerical
Price	price of the shoe (in rupees)	Numerical
Sizes	size options	Object
Count_Sizes	total number of size options	Numerical
Color_Name	color options	Categorical
product_code	code of the shoe	Categorical
Review	number of reviews	Numerical
Rating	rating of the shoe	Numerical
Features	unique features of the shoe	Categorical
Comfort	comfort description	Categorical
Brand	brand name	Categorical
Units_sold	total units sold of each model	Numerical
Revenue_Generated	total revenue generated (in rupees)	Numerical

Size: 44476 observations and 13 features

Sample:



```
+-------+--------------------+--------------------+--------------------+--------------------+-------+--------------------+--------------------+
|summary|                Name|            Category|              Colors|               Price|  Sizes|         Count_Sizes|          Color_Name|
+-------+--------------------+--------------------+--------------------+--------------------+-------+--------------------+--------------------+
|  count|               42762|               14325|                6031|                4168|   2638|                2337|                2802|
|   mean|                NULL|                10.0|    4.34785962205939|                NULL|   NULL|   7.882480957562568|                NULL|
| stddev|                NULL|                 0.0|  2.7339762272467976|                NULL|   NULL|   4.529142067142776|                NULL|
|    min|""1 Cent"" logo o...| ""Don't fix what...| ""Perfect it"". ...| Air Force 1 beca...|   HIIT| adding DIY flair...| and everywhere i...|
|    max|           x MMW 005|               Women|Zoom Air unit off...|Zoom Air unit in ...|['9.5']|The woven and syn...|Wolf Grey/Wolf Gr...|
+-------+--------------------+--------------------+--------------------+--------------------+-------+--------------------+--------------------+
```



```
+------------+------+------+--------------------+-------+----------+-----------------+--------+
|product_Code|Review|Rating|            Features|Comfort|Units_Sold|Revenue_Generated|Discount|
+------------+------+------+--------------------+-------+----------+-----------------+--------+
|   CW2288-111|  1311|  NULL|LEGENDARY STYLE R...|   NULL|        82|         614590.0|       7|
|        NULL|  NULL|  NULL|                NULL|   NULL|        52|             NULL|       1|
|  FD9082-103|    88|   4.8|Meet the leader o...|   NULL|        35|         447825.0|       7|
|        NULL|  NULL|  NULL|                NULL|   NULL|         3|             NULL|       6|
|  FB9658-001|     3|   4.7|Lace up and feel ...|   NULL|        78|         927810.0|       4|
+------------+------+------+--------------------+-------+----------+-----------------+--------+
```

Point of Interest:

Gain insights into consumer behavior and optimize profitability strategies for “Sole Haven”.

Research Questions:

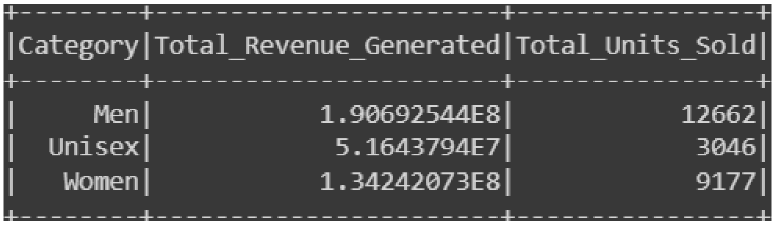
1) What insights can be leveraged to optimize marketing strategies and drive sales revenue across different target segments?
2) How can we enhance customer segmentation, optimize marketing strategies, and improve inventory management in the shoe store, ultimately driving competitive advantage and customer loyalty?
3) How does implementing streaming analytics in our shoe store affect profitability and fame? Can real-time insights optimize inventory, pricing, and marketing to drive profitability and enhance customer satisfaction, thereby boosting the store's reputation?
4) How are different shoe models related to each other based on factors like color availability and category similarity, and how can these relationships inform personalized product recommendations and marketing strategies?

Methodology:

1) Data Preprocessing: We have processed the data by removing the null prices to ensure the data integrity, consistency, and accuracy in subsequent analyses. This step was crucial before performing any analysis to maintain the quality of the dataset and to facilitate meaningful insights. We further added multiple columns in our data set using the "withColumn" method and further using the "rand()" function to generate random numbers. This helped us to prioritize important products and trends in our dataset.
2) Descriptive Statistics: After data preparation, we performed some descriptive statistics such as count, min, max to summarize the key attributes of the dataset, including the prizing, ratings, and sizes. This technique provided an overview of the data distribution and allowed us for initial exploration.
3) Feature Engineering: We added new features, such as "Units_Sold" and "Revenue_Generated", by combining the existing attributes like prices and sales volume. These features helped us to understand the product performance and relevant potential for profitability optimization.
4) Clustering Analysis (K-means): We segmented the customers based on their purchasing behavior and preferences. This ML technique helped us to identify different customer segments, facilitating targeted marketing strategies and personalized product recommendations.
5) Spark Streaming: In order to leverage real-time data for tailored marketing campaigns and enhanced customer experiences, showcasing techniques like personalized recommendations and timely promotions, we have used Spark Streaming techniques. This technique helped us to focus on store optimization by ensuring that popular products are readily available while minimizing excess stock.
6) PySpark GraphFrames: By adapting graph frames technology, the sole haven store can offer personalized recommendations to customers and inturn increasing the likelihood of sales and improving customer satisfaction.

Results and Findings

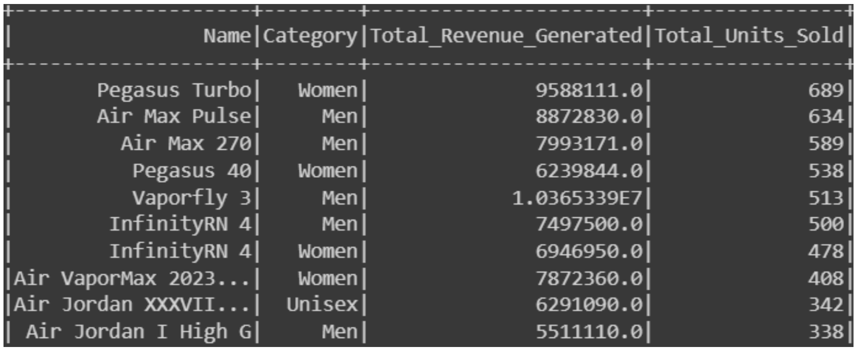
1) Sales revenue across different target segments



```
+--------+-----------------------+----------------+
|Category|Total_Revenue_Generated|Total_Units_Sold|
+--------+-----------------------+----------------+
|     Men|           1.90692544E8|           12662|
|  Unisex|            5.1643794E7|            3046|
|   Women|           1.34242073E8|            9177|
+--------+-----------------------+----------------+
```

Men's shoes drive the majority of total revenue, showcasing high demand. Despite lower unit sales, women's shoes contribute significantly, and unisex shoes, though fewer in units, have a notable impact. This highlights the potential for focused marketing and product development to boost men's shoes' performance and tap into growth opportunities in women's and unisex segments.

2) Best selling products across different categories

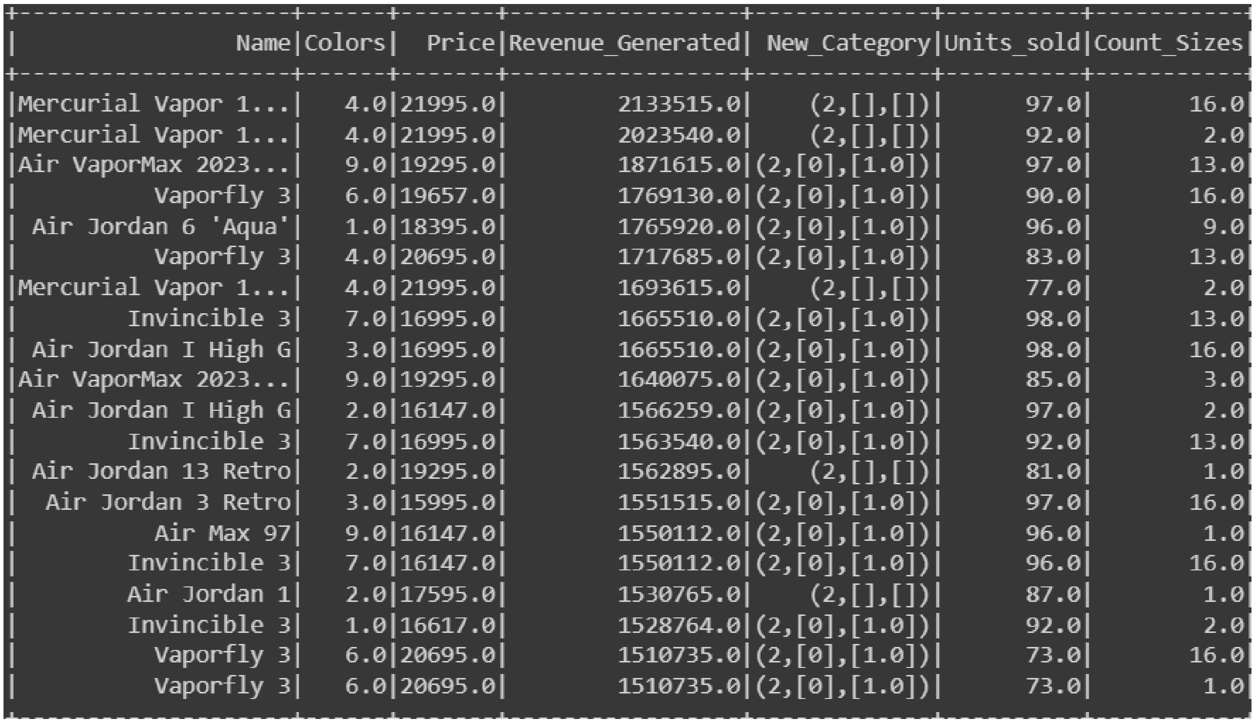


```
+--------------------+--------+-----------------------+----------------+
|                Name|Category|Total_Revenue_Generated|Total_Units_Sold|
+--------------------+--------+-----------------------+----------------+
|        Pegasus Turbo|   Women|              9588111.0|             689|
|       Air Max Pulse|     Men|              8872830.0|             634|
|         Air Max 270|     Men|              7993171.0|             589|
|          Pegasus 40|   Women|              6239844.0|             538|
|          Vaporfly 3|     Men|            1.0365339E7|             513|
|        InfinityRN 4|     Men|              7497500.0|             500|
|        InfinityRN 4|   Women|              6946950.0|             478|
|Air VaporMax 2023...|   Women|              7872360.0|             408|
|Air Jordan XXXVII...|  Unisex|              6291090.0|             342|
| Air Jordan I High G|     Men|              5511110.0|             338|
```

Top-selling products differ by category, with "Pegasus Turbo" and "Pegasus 40" leading in women's shoes, and "Air Max Pulse" and "Air Max 270" dominating in men's shoes. The inclusion of "Air Jordan XXXVII" in the unisex category indicates its broad appeal. This analysis informs retailers on understanding consumer preferences, optimizing product assortment, and tailoring marketing campaigns for maximum sales and revenue across diverse target segments.

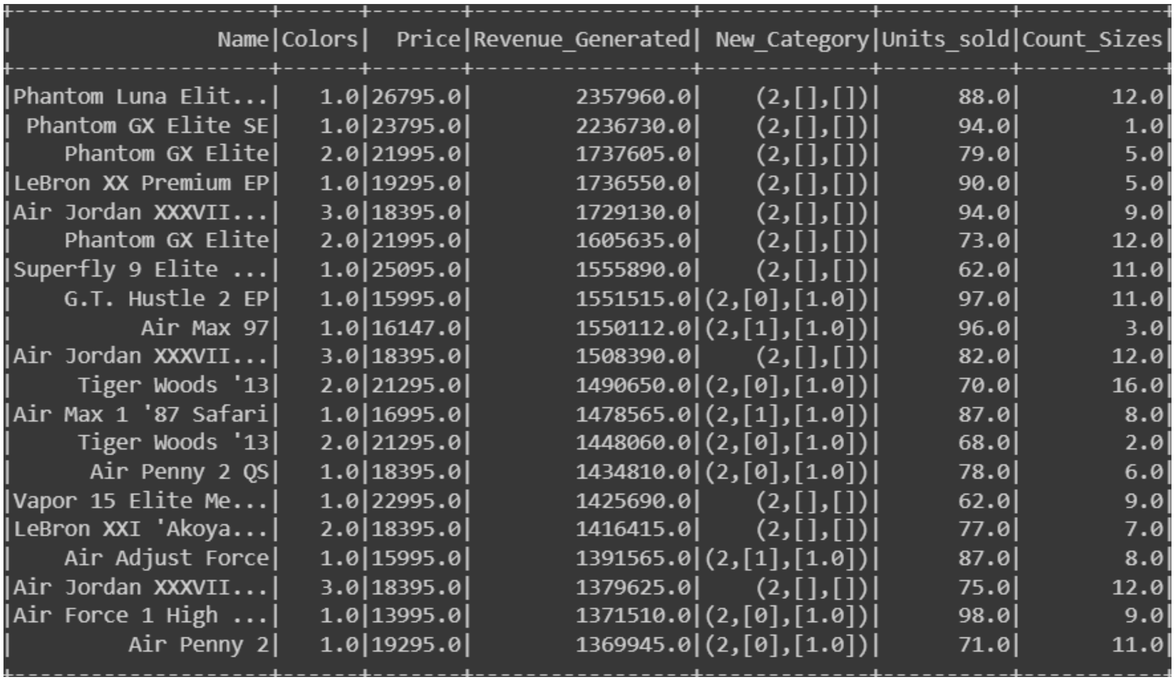
3) Identifying distinct clusters within the shoe data to personalize customer experiences

Cluster 1: Shows higher-priced shoes with moderate to high units sold, indicating demand for premium products.



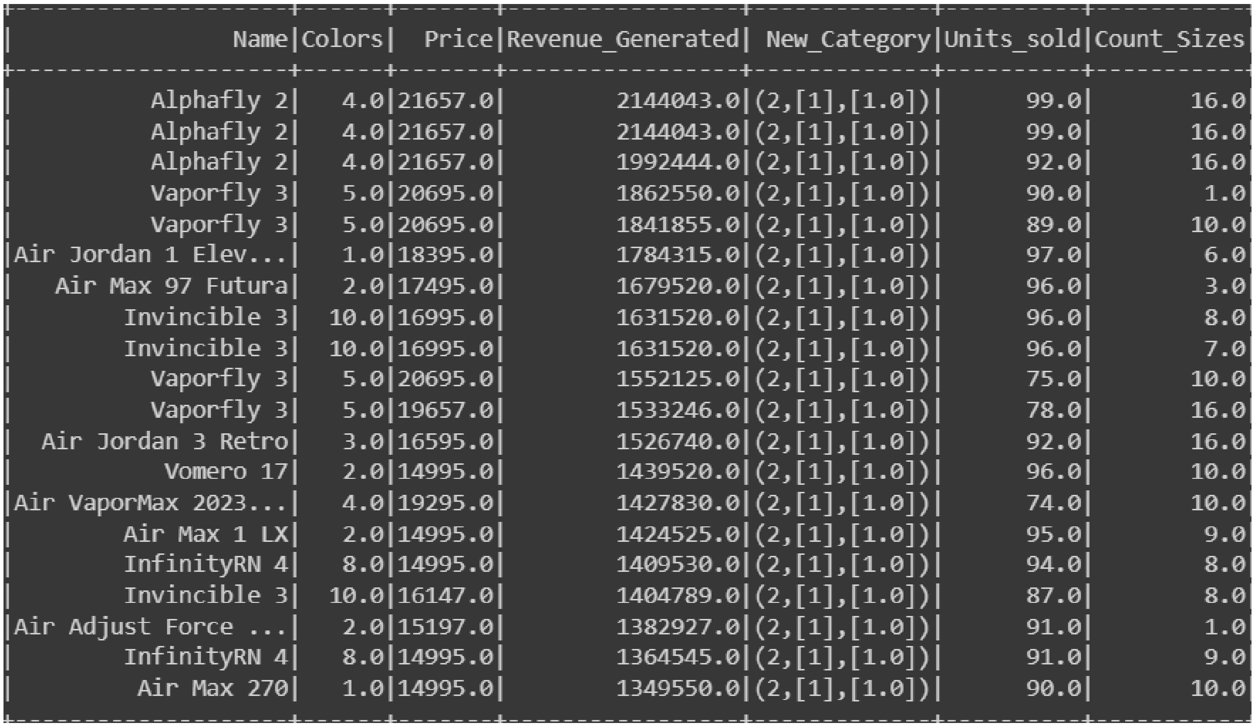
```
+--------------------+------+-------+-----------------+-------------+----------+-----------
|                Name|Colors|  Price|Revenue_Generated| New_Category|Units_sold|Count_Sizes|
+--------------------+------+-------+-----------------+-------------+----------+-----------
|Mercurial Vapor 1...|   4.0|21995.0|        2133515.0|    (2,[],[])|      97.0|       16.0|
|Mercurial Vapor 1...|   4.0|21995.0|        2023540.0|    (2,[],[])|      92.0|        2.0|
|Air VaporMax 2023...|   9.0|19295.0|        1871615.0|(2,[0],[1.0])|      97.0|       13.0|
|          Vaporfly 3|   6.0|19657.0|        1769130.0|(2,[0],[1.0])|      90.0|       16.0|
| Air Jordan 6 'Aqua'|   1.0|18395.0|        1765920.0|(2,[0],[1.0])|      96.0|        9.0|
|          Vaporfly 3|   4.0|20695.0|        1717685.0|(2,[0],[1.0])|      83.0|       13.0|
|Mercurial Vapor 1...|   4.0|21995.0|        1693615.0|    (2,[],[])|      77.0|        2.0|
|        Invincible 3|   7.0|16995.0|        1665510.0|(2,[0],[1.0])|      98.0|       13.0|
| Air Jordan I High G|   3.0|16995.0|        1665510.0|(2,[0],[1.0])|      98.0|       16.0|
|Air VaporMax 2023...|   9.0|19295.0|        1640075.0|(2,[0],[1.0])|      85.0|        3.0|
| Air Jordan I High G|   2.0|16147.0|        1566259.0|(2,[0],[1.0])|      97.0|        2.0|
|        Invincible 3|   7.0|16995.0|        1563540.0|(2,[0],[1.0])|      92.0|       13.0|
| Air Jordan 13 Retro|   2.0|19295.0|        1562895.0|    (2,[],[])|      81.0|        1.0|
|  Air Jordan 3 Retro|   3.0|15995.0|        1551515.0|(2,[0],[1.0])|      97.0|       16.0|
|          Air Max 97|   9.0|16147.0|        1550112.0|(2,[0],[1.0])|      96.0|        1.0|
|        Invincible 3|   7.0|16147.0|        1550112.0|(2,[0],[1.0])|      96.0|       16.0|
|        Air Jordan 1|   2.0|17595.0|        1530765.0|    (2,[],[])|      87.0|        1.0|
|        Invincible 3|   1.0|16617.0|        1528764.0|(2,[0],[1.0])|      92.0|        2.0|
|          Vaporfly 3|   6.0|20695.0|        1510735.0|(2,[0],[1.0])|      73.0|       16.0|
|          Vaporfly 3|   6.0|20695.0|        1510735.0|(2,[0],[1.0])|      73.0|        1.0|
+--------------------+------+-------+-----------------+-------------+----------+-----------
```

Cluster 2: Exhibits a mix of mid to high-priced shoes with varied colors and sizes, suggesting diversity in consumer preferences.



```
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|                Name|Colors|  Price|Revenue_Generated| New_Category|Units_sold|Count_Sizes|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|Phantom Luna Elit...|   1.0|26795.0|        2357960.0|    (2,[],[])|      88.0|       12.0|
| Phantom GX Elite SE|   1.0|23795.0|        2236730.0|    (2,[],[])|      94.0|        1.0|
|    Phantom GX Elite|   2.0|21995.0|        1737605.0|    (2,[],[])|      79.0|        5.0|
|LeBron XX Premium EP|   1.0|19295.0|        1736550.0|    (2,[],[])|      90.0|        5.0|
|Air Jordan XXXVII...|   3.0|18395.0|        1729130.0|    (2,[],[])|      94.0|        9.0|
|    Phantom GX Elite|   2.0|21995.0|        1605635.0|    (2,[],[])|      73.0|       12.0|
|Superfly 9 Elite ...|   1.0|25095.0|        1555890.0|    (2,[],[])|      62.0|       11.0|
|    G.T. Hustle 2 EP|   1.0|15995.0|        1551515.0|(2,[0],[1.0])|      97.0|       11.0|
|          Air Max 97|   1.0|16147.0|        1550112.0|(2,[1],[1.0])|      96.0|        3.0|
|Air Jordan XXXVII...|   3.0|18395.0|        1508390.0|    (2,[],[])|      82.0|       12.0|
|     Tiger Woods '13|   2.0|21295.0|        1490650.0|(2,[0],[1.0])|      70.0|       16.0|
|Air Max 1 '87 Safari|   1.0|16995.0|        1478565.0|(2,[1],[1.0])|      87.0|        8.0|
|     Tiger Woods '13|   2.0|21295.0|        1448060.0|(2,[0],[1.0])|      68.0|        2.0|
|      Air Penny 2 QS|   1.0|18395.0|        1434810.0|(2,[0],[1.0])|      78.0|        6.0|
|Vapor 15 Elite Me...|   1.0|22995.0|        1425690.0|    (2,[],[])|      62.0|        9.0|
|LeBron XXI 'Akoya...|   2.0|18395.0|        1416415.0|    (2,[],[])|      77.0|        7.0|
|    Air Adjust Force|   1.0|15995.0|        1391565.0|(2,[1],[1.0])|      87.0|        8.0|
|Air Jordan XXXVII...|   3.0|18395.0|        1379625.0|    (2,[],[])|      75.0|       12.0|
|Air Force 1 High ...|   1.0|13995.0|        1371510.0|(2,[0],[1.0])|      98.0|        9.0|
|         Air Penny 2|   1.0|19295.0|        1369945.0|(2,[0],[1.0])|      71.0|       11.0|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
```

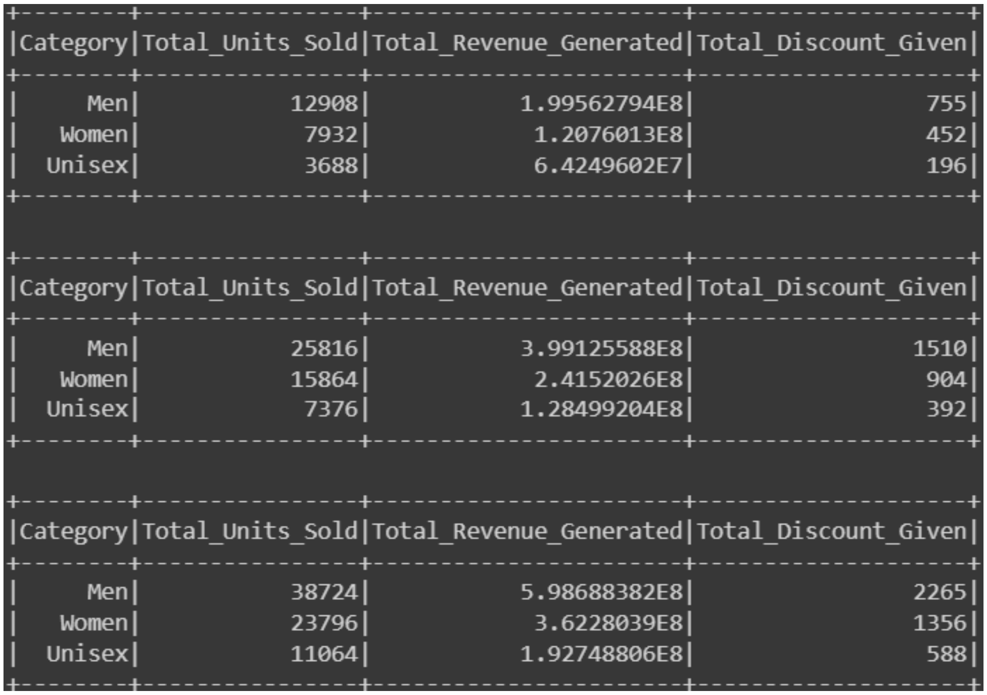
Cluster 3: Displays higher-priced shoes with comparatively lower units sold, indicating niche or specialized products.



```
+--------------------+------+-------+-----------------+-------------+----------+-----------
|                Name|Colors|  Price|Revenue_Generated| New_Category|Units_sold|Count_Sizes
+--------------------+------+-------+-----------------+-------------+----------+-----------
|          Alphafly 2|   4.0|21657.0|        2144043.0|(2,[1],[1.0])|      99.0|       16.0
|          Alphafly 2|   4.0|21657.0|        2144043.0|(2,[1],[1.0])|      99.0|       16.0
|          Alphafly 2|   4.0|21657.0|        1992444.0|(2,[1],[1.0])|      92.0|       16.0
|          Vaporfly 3|   5.0|20695.0|        1862550.0|(2,[1],[1.0])|      90.0|        1.0
|          Vaporfly 3|   5.0|20695.0|        1841855.0|(2,[1],[1.0])|      89.0|       10.0
|Air Jordan 1 Elev...|   1.0|18395.0|        1784315.0|(2,[1],[1.0])|      97.0|        6.0
|   Air Max 97 Futura|   2.0|17495.0|        1679520.0|(2,[1],[1.0])|      96.0|        3.0
|        Invincible 3|  10.0|16995.0|        1631520.0|(2,[1],[1.0])|      96.0|        8.0
|        Invincible 3|  10.0|16995.0|        1631520.0|(2,[1],[1.0])|      96.0|        7.0
|          Vaporfly 3|   5.0|20695.0|        1552125.0|(2,[1],[1.0])|      75.0|       10.0
|          Vaporfly 3|   5.0|19657.0|        1533246.0|(2,[1],[1.0])|      78.0|       16.0
|  Air Jordan 3 Retro|   3.0|16595.0|        1526740.0|(2,[1],[1.0])|      92.0|       16.0
|           Vomero 17|   2.0|14995.0|        1439520.0|(2,[1],[1.0])|      96.0|       10.0
|Air VaporMax 2023...|   4.0|19295.0|        1427830.0|(2,[1],[1.0])|      74.0|       10.0
|        Air Max 1 LX|   2.0|14995.0|        1424525.0|(2,[1],[1.0])|      95.0|        9.0
|        InfinityRN 4|   8.0|14995.0|        1409530.0|(2,[1],[1.0])|      94.0|        8.0
|        Invincible 3|  10.0|16147.0|        1404789.0|(2,[1],[1.0])|      87.0|        8.0
|Air Adjust Force ...|   2.0|15197.0|        1382927.0|(2,[1],[1.0])|      91.0|        1.0
|        InfinityRN 4|   8.0|14995.0|        1364545.0|(2,[1],[1.0])|      91.0|        9.0
|         Air Max 270|   1.0|14995.0|        1349550.0|(2,[1],[1.0])|      90.0|       10.0
+--------------------+------+-------+-----------------+-------------+----------+-----------
```

Analyzing these clusters can help in tailoring marketing strategies, optimizing inventory, and identifying trends. For instance, Cluster 1 may be targeted towards high-end consumers, while Cluster 2 might represent products appealing to a broader audience.

4) Streaming analytics in the shoe store



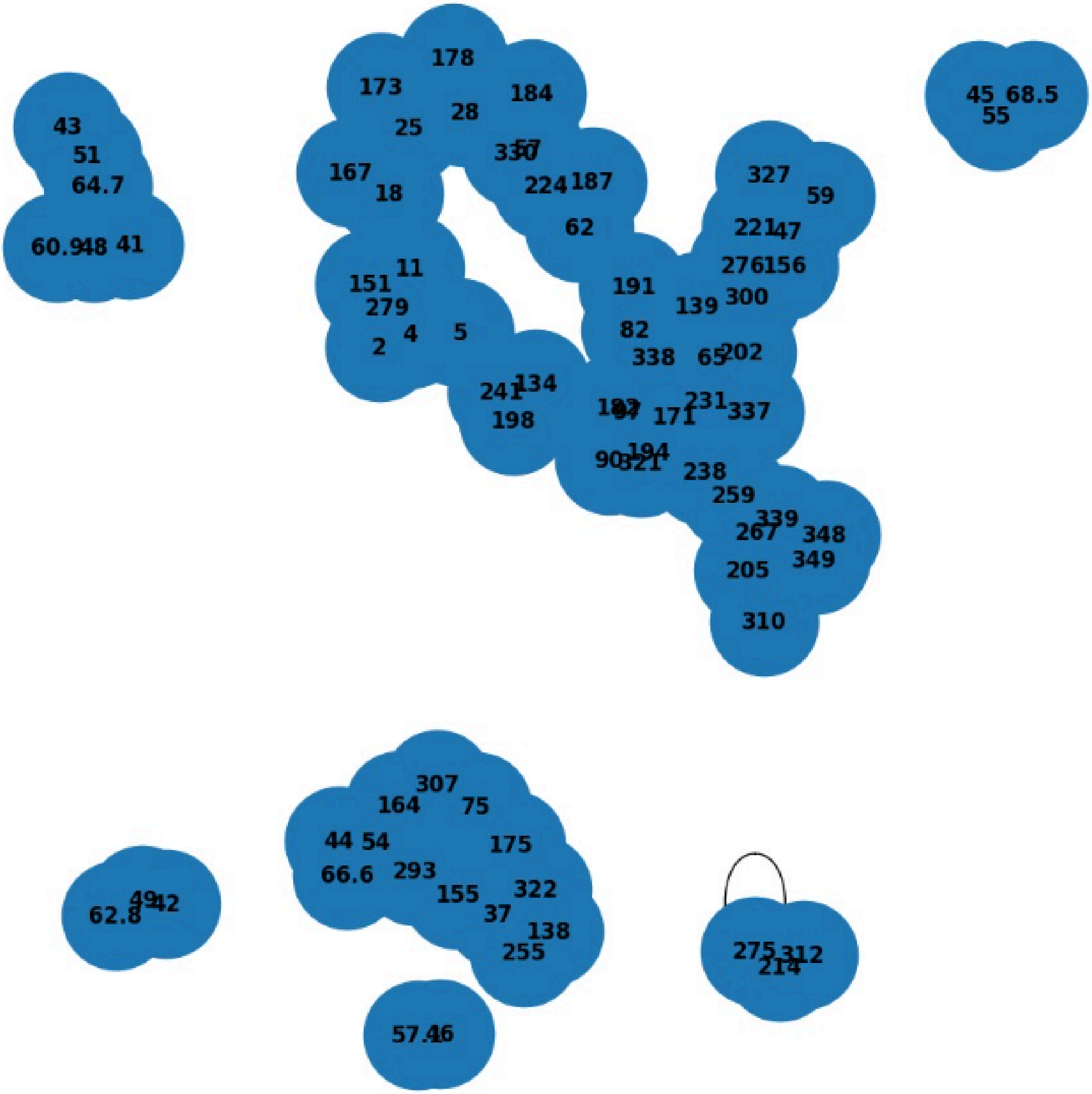
```
+--------+----------------+-----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+-----------------------+--------------------+
|     Men|           12908|           1.99562794E8|                 755|
|   Women|            7932|            1.2076013E8|                 452|
|  Unisex|            3688|            6.4249602E7|                 196|
+--------+----------------+-----------------------+--------------------+


+--------+----------------+-----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+-----------------------+--------------------+
|     Men|           25816|           3.99125588E8|                1510|
|   Women|           15864|            2.4152026E8|                 904|
|  Unisex|            7376|           1.28499204E8|                 392|
+--------+----------------+-----------------------+--------------------+


+--------+----------------+-----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+-----------------------+--------------------+
|     Men|           38724|           5.98688382E8|                2265|
|   Women|           23796|            3.6228039E8|                1356|
|  Unisex|           11064|           1.92748806E8|                 588|
+--------+----------------+-----------------------+--------------------+
```

Implementing data streaming in the shoe store enhances profitability and fame. Real-time insights optimize inventory, reduce costs, and capture maximum value through dynamic pricing. Tailored marketing campaigns based on real-time data elevate customer experiences, increasing satisfaction and loyalty. This approach not only drives efficient operations for profitability but also enhances the store's reputation within the industry, delivering exceptional customer experiences.

5) Identifying relationships between shoes



Understanding the relationships between products can also inform the development of engaging customer experiences, such as curated collections, thematic product displays, and interactive online features. By creating compelling narratives around product relationships, the store can captivate customers and foster brand loyalty.

Conclusion

In conclusion, our analysis of Sole Haven's shoe data offers valuable insights and strategies to overcome challenges and elevate profitability. Through careful preprocessing, we addressed missing values, ensuring reliable data for subsequent analyses. The examination of sales across categories guided targeted marketing and product development efforts, emphasizing the potential in men's, women's, and unisex segments.

The implementation of K-means clustering enhanced customer segmentation for personalized strategies and improved inventory management. Exploring streaming analytics showcased its transformative impact on profitability and fame, optimizing operations and enhancing customer satisfaction in real-time.

Finally, our analysis of shoe relationships allows for personalized product recommendations and innovative marketing strategies. Sole Haven is now equipped with a strategic roadmap to navigate the industry, optimize operations, and enhance customer satisfaction, poised for renewed success and prominence.