

Topic : Customer Segmentation and Content-Based Image Retrieval for Fashion Products

Team Members

- BP Shashidhar Reddy
- Vishakha Maruti Sonmore
- Aastha Dinesh Singh Kshatriya
- Hari Haran Dhulipala

Course Name: IS 675: Deep Learning for Business

Instructor: Basil Latif

Date: 12/05/2024

Introduction

The e-commerce industry is undergoing a period of remarkable growth and transformation, driven by rapid technological advancements and evolving consumer preferences. With the global market projected to reach \$7.4 trillion by 2025, e-commerce is becoming an integral part of the global economy. This growth is accompanied by increased competition, compelling businesses to adopt innovative strategies to stand out in a crowded market.

To remain competitive, e-commerce companies must harness the vast amounts of data generated daily. This includes structured data, such as transaction records, customer demographics, and product information, as well as unstructured data, including product images, customer reviews, and browsing patterns. Effectively utilizing both types of data is essential for understanding customer behaviour, predicting future trends, and delivering personalized experiences that cater to individual preferences.

Personalization has emerged as a cornerstone of modern e-commerce, enabling businesses to tailor their offerings to the unique needs and desires of each customer. Studies have shown that personalization can significantly enhance customer satisfaction, loyalty, and engagement, directly impacting revenue growth. However, achieving this level of personalization requires advanced analytics and data processing techniques capable of handling diverse and complex datasets.

In response to these challenges, this project leverages predictive Modeling and content-based image retrieval using deep learning. Predictive Modeling is used to anticipate customer needs by analysing historical data, enabling businesses to optimize inventory management, marketing strategies, and pricing decisions. Meanwhile, content-based image retrieval employs advanced computer vision techniques to facilitate visual searches, allowing customers to discover visually similar products effortlessly. This feature aligns with the growing demand for intuitive and engaging shopping experiences, particularly among younger, tech-savvy demographics.

By combining machine learning and deep learning approaches, this project aims to address critical challenges in e-commerce, providing actionable insights and innovative solutions. The integration of predictive analytics and visual search capabilities offers a comprehensive strategy for enhancing user engagement, boosting operational efficiency, and ultimately driving business success in the competitive e-commerce landscape.

Problem Statement

E-commerce platforms face a significant challenge in managing and extracting actionable insights from the vast and complex datasets they generate daily. These datasets often include structured data, such as transaction histories and customer profiles, alongside unstructured data, including product images and textual reviews. While these datasets hold immense potential for driving business growth, the sheer volume, diversity, and complexity of the data make it difficult for businesses to harness their full value effectively.

One critical issue is the inability to derive meaningful patterns from this data to anticipate customer preferences and behaviour. Without advanced analytical capabilities, businesses struggle to optimize critical functions such as inventory management, pricing strategies, and personalized marketing. The lack of effective utilization of available data can result in missed opportunities, reduced customer satisfaction, and diminished competitive advantage.

Adding to this challenge is the growing demand for intuitive and visually engaging shopping experiences. Millennials and Gen Z consumers, who now represent a significant portion of the global e-commerce market, increasingly prefer visual search functionality over traditional text-based searches. Visual search allows users to upload an image or select an existing product and receive recommendations for similar items. However, implementing this feature requires sophisticated content-based image retrieval systems that can analyse, compare, and recommend visually similar products in real time.

Predictive Modeling and deep learning offer promising solutions to these challenges. Predictive Modeling enables e-commerce platforms to forecast customer behaviour and market trends based on historical data, providing actionable insights for personalization and operational optimization. Simultaneously, deep learning techniques, such as convolutional neural networks (CNNs), empower businesses to implement advanced visual search functionalities, enhancing the shopping experience by meeting modern consumer expectations.

By addressing these challenges, this project aims to bridge the gap between the potential and practical use of data in e-commerce. It leverages predictive analytics to improve decision-making and deep learning to create engaging, visually-driven interactions, ultimately enabling businesses to deliver personalized, data-driven experiences that resonate with their customer base.

Project Approach

1. Dataset Selection

Tabular Dataset

The project uses an **e-commerce dataset** to explore customer behaviour and develop predictive and visual Modeling solutions. Below is a summary of the dataset:

1.1 Characteristics

- **Number of Rows:** 51,290 (substantial for statistical and machine learning analyses).
- **Number of Columns:** 21 (includes a mix of categorical, numerical, and date fields).
- **Data Types:** Includes structured tabular data, with potential challenges from mixed data types in some columns.

1.2. Key Features

- **Order Details:** Includes Order ID, Order Date, and Ship Date, providing insights into transaction timing and delivery performance.
- **Product Information:** Contains Product Category and Product fields, categorizing items sold.
- **Sales Metrics:** Features like Sales, Profit, and Discount help in analyzing financial performance.
- **Customer Segmentation:** Columns such as Customer ID, Customer Name, Segment, and Region provide details for clustering and segmentation.
- **Geographic Information:** Includes City, State, Country, and Region, enabling geographic trend analysis.

1.3. Sample Data

A snippet of the dataset is provided below to illustrate its structure:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Order ID	Order Date	Ship Date	Aging	Ship Mode	Product C	Product	Sales	Quantity	Discount	Profit	Shipping C	Order Pric	Customer	Customer Segment	City	State	Country	R	
2	AU-2015-1	11/9/2015	11/17/2015	8	First Class	Auto & Ac	Car Media	\$140.00	2	0.05	\$46.00	\$4.60	Medium	LS-001	Lane Dani	Consumer	Brisbane	Queenslar	Australia	C
3	AU-2015-2	6/30/2015	7/2/2015	2	First Class	Auto & Ac	Car Speak	\$211.00	3	0.03	\$112.00	\$11.20	Medium	IZ-002	Alvarado I	Home Offi	Berlin	Berlin	Germany	C
4	AU-2015-3	12/5/2015	12/13/2015	8	First Class	Auto & Ac	Car Body C	\$117.00	5	0.01	\$31.20	\$3.10	Critical	EN-003	Moon We	Consumer	Porirua	Wellington	New Zeala	C
5	AU-2015-4	5/9/2015	5/16/2015	7	First Class	Auto & Ac	Car & Bike	\$118.00	2	0.05	\$26.20	\$2.60	High	AN-004	Sanchez B	Corporate	Kabul	Kabul	Afghanist	C
6	AU-2015-5	7/9/2015	7/18/2015	9	First Class	Auto & Ac	Tyre	\$250.00	1	0.04	\$160.00	\$16.00	Critical	ON-005	Rowe Jack	Corporate	Townsville	Queenslar	Australia	C
7	AU-2015-6	2/25/2015	3/5/2015	8	First Class	Auto & Ac	Bike Tyres	\$72.00	3	0.04	\$24.00	\$2.40	Critical	TO-006	Carter Bar	Corporate	Bytom	Silesia	Poland	E
8	AU-2015-7	4/9/2015	4/10/2015	1	First Class	Auto & Ac	Car Mat	\$54.00	1	0.05	\$54.00	\$5.40	High	OM-007	Mconnell	Consumer	Chicago	Illinois	United Sta	C
9	AU-2015-8	3/30/2015	4/6/2015	7	First Class	Auto & Ac	Car Seat C	\$114.00	5	0.02	\$22.60	\$2.30	Critical	AN-008	Dennis Ho	Corporate	Suzhou	Anhui	China	N
10	AU-2015-9	2/9/2015	2/16/2015	7	First Class	Auto & Ac	Car Pillow	\$231.00	5	0.03	\$116.40	\$11.60	Critical	EN-009	Wall Olser	Consumer	Juárez	Chihuahua	Mexico	N
11	AU-2015-1	4/21/2015	5/1/2015	10	First Class	Auto & Ac	Car Media	\$140.00	2	0.02	\$54.40	\$5.40	Critical	TT-0010	Shepard V	Consumer	Soyapang	San Salvac	El Salvado	C
12	AU-2015-1	11/16/2015	11/26/2015	10	First Class	Auto & Ac	Car Speak	\$211.00	4	0.01	\$122.60	\$12.30	Critical	ED-0011	Johns Ree	Corporate	Taipei	Taipei City	Taiwan	N
13	AU-2015-1	9/1/2015	9/2/2015	1	First Class	Auto & Ac	Car Body C	\$117.00	4	0.04	\$18.30	\$1.80	High	ON-0012	Doyle Knu	Home Offi	Los Angele	California	United Sta	V
14	AU-2015-1	7/9/2015	7/16/2015	7	First Class	Auto & Ac	Car & Bike	\$118.00	1	0.02	\$35.60	\$3.60	Critical	WN-0013	Butler Bro	Corporate	Saint-Briei	Brittany	France	C
15	AU-2015-1	7/22/2015	7/27/2015	5	First Class	Auto & Ac	Tyre	\$250.00	3	0.04	\$140.00	\$14.00	High	AN-0014	Johnson A	Corporate	Kamina	Katanga	Democrat	A
16	AU-2015-1	10/12/2015	10/21/2015	9	First Class	Auto & Ac	Bike Tyres	\$72.00	4	0.01	\$18.00	\$1.80	Medium	EY-0015	Greene De	Consumer	Brisbane	Queenslar	Australia	C
17	AU-2015-1	2/23/2015	3/5/2015	10	First Class	Auto & Ac	Car Mat	\$54.00	2	0.01	\$27.00	\$2.70	Critical	RN-0016	Bentley Zy	Consumer	Berlin	Berlin	Germany	C
18	AU-2015-1	5/4/2015	5/8/2015	4	First Class	Auto & Ac	Car Seat C	\$114.00	2	0.05	\$22.60	\$2.30	High	CK-0017	Rivera Bla	Consumer	Shouguan	Shandong	China	N
19	AU-2015-1	6/12/2015	6/19/2015	7	First Class	Auto & Ac	Car Pillow	\$231.00	5	0.05	\$93.30	\$9.30	High	RE-0018	Wong Mai	Consumer	New York	New York	United Sta	E
20	AU-2015-1	5/13/2015	5/20/2015	7	First Class	Auto & Ac	Car Media	\$140.00	2	0.05	\$46.00	\$4.60	Critical	ON-0019	Hendricks	Consumer	Behshahr	Mazandar	Iran	E

1.4. Business Relevance

- **Customer Behaviour:** Helps predict future purchases and identify patterns in customer segments.
- **Financial Insights:** Allows for sales forecasting and profit optimization.
- **Geographic Trends:** Supports regional targeting and market analysis.
- **Personalization:** Insights from this data can drive product recommendations and marketing campaigns.

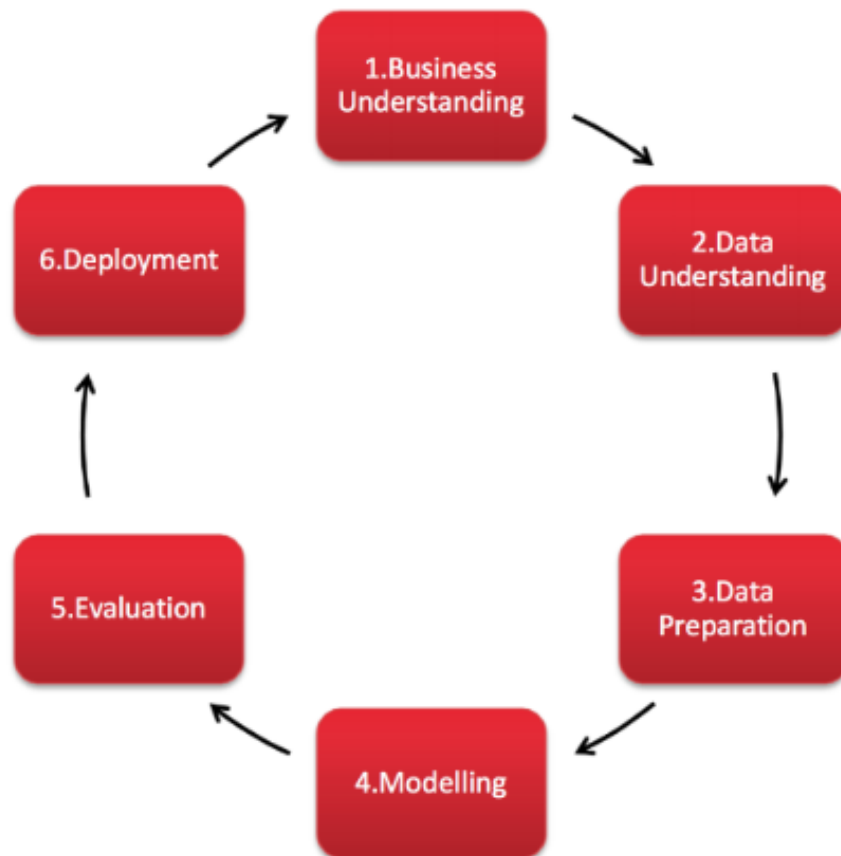
1.5. Unstructured Dataset:

A dataset with 44,441 images of e-commerce products (clothes, beauty products, shoes, etc.) will support a content-based image retrieval system. This aligns with the business goal of enhancing customer experience with visual search.

2. CRISP-DM Methodology

- **Business Understanding:** Focus on enhancing personalization and user experience in e-commerce.
- **Data Understanding:** Analyse the tabular dataset for trends in sales, profits, and regions. For images, explore the distribution and diversity of product categories.
- **Data Preparation:**
 - Tabular Data: Handle missing values, normalize numerical features, encode categorical variables, and remove outliers.
 - Image Data: Resize and preprocess images for input to CNN models.

- **Modeling:** Apply machine learning (ML) and neural network (NN) architectures to both datasets.
- **Evaluation:** Assess model accuracy, runtime, and other metrics for both ML and NN approaches.
- **Deployment:** Propose potential deployment pipelines for integrating these models into business applications.



3. Visualizations


- **Tabular Data:** Use visualizations like histograms, scatterplots, boxplots, and correlation matrices to understand data distribution and relationships.
- **Model Results:** Include visualizations of confusion matrices, accuracy/loss curves, feature importance (for ML models), and embedding projections (e.g., t-SNE for image embeddings).

4. Neural Network Design

- **Tabular Dataset:** Implement feedforward neural networks (FNNs) for predictive modeling.
- **Image Dataset:** Use a pre-trained CNN (e.g., ResNet, VGG) for feature extraction and train the classifier with embeddings. Include models like CNN+KNN for content-based image retrieval

5. Machine Learning Models

- Tabular Dataset: Train Decision Tree, Random Forest, and Gradient Boosted Trees for predictive Modeling.




Predictive Modeling Using Decision Tree Classifier

WHY DECISION TREE ?

Predict the most likely product category a customer will purchase next based on their past behavior. This helps businesses anticipate needs, personalize marketing, and improve product recommendations.

Derived Insights and Business Applications

- Key Drivers of Customer Preferences
- Improved Inventory Management
- Effective Pricing & Discount Strategies



Sales Forecasting Using Random Forest Regressor


WHY RANDOM FOREST?

The goal of this analysis is to forecast future sales volume for each product category using historical sales data. Accurate sales forecasts are crucial for inventory management, optimizing marketing strategies, and planning promotional campaigns.


Derived Insights and Business Applications

- Accurate Sales Forecasts
- Understanding Key Drivers
- Actionable Recommendations

- Unstructured Dataset: Use KNN on CNN-generated embeddings for image classification.




Fashion Visual Search Model using pre-trained CNN and KNN



Feature Extraction Using CNN

- We employed a pre-trained CNN as the backbone for feature extraction.
- The CNN is used to generate feature embeddings.
- VGG16.



Similarity Search Using KNN

The embeddings generated by the CNN are used as input for a KNN classifier.

6. Model Improvement

- Perform hyperparameter tuning for ML and NN models:
 - **ML Models:** Adjust parameters like tree depth, number of estimators, learning rate, etc.
 - **NN Models:** Experiment with optimizers (SGD, Adam), learning rates, batch sizes, dropout rates, and architectures.
- Use libraries like GridSearchCV or Optuna for systematic tuning.

7. Performance Comparison

- Compare ML models (DT, RF, GBT) and NN architectures based on metrics such as:
 - **Accuracy:** Overall correctness of predictions.
 - **Runtime:** Time taken for training and inference.
 - **Precision/Recall:** Performance on imbalanced datasets.
- Evaluate the NN's ability to handle unstructured data, demonstrating its advantage over traditional ML models.

8. Results and Discussion

- **Tabular Data:**
 - Discuss the predictive power of the Decision Tree, Random Forest, and Neural Networks.
 - Present insights like key drivers of sales and category-level preferences.
- **Unstructured Data:**
 - Evaluate CNN's ability to generate meaningful embeddings for similarity search.
 - Discuss challenges like imbalanced product categories. or overfitting.
- **Challenges:** Include preprocessing issues, model convergence difficulties, or hyperparameter tuning roadblocks.

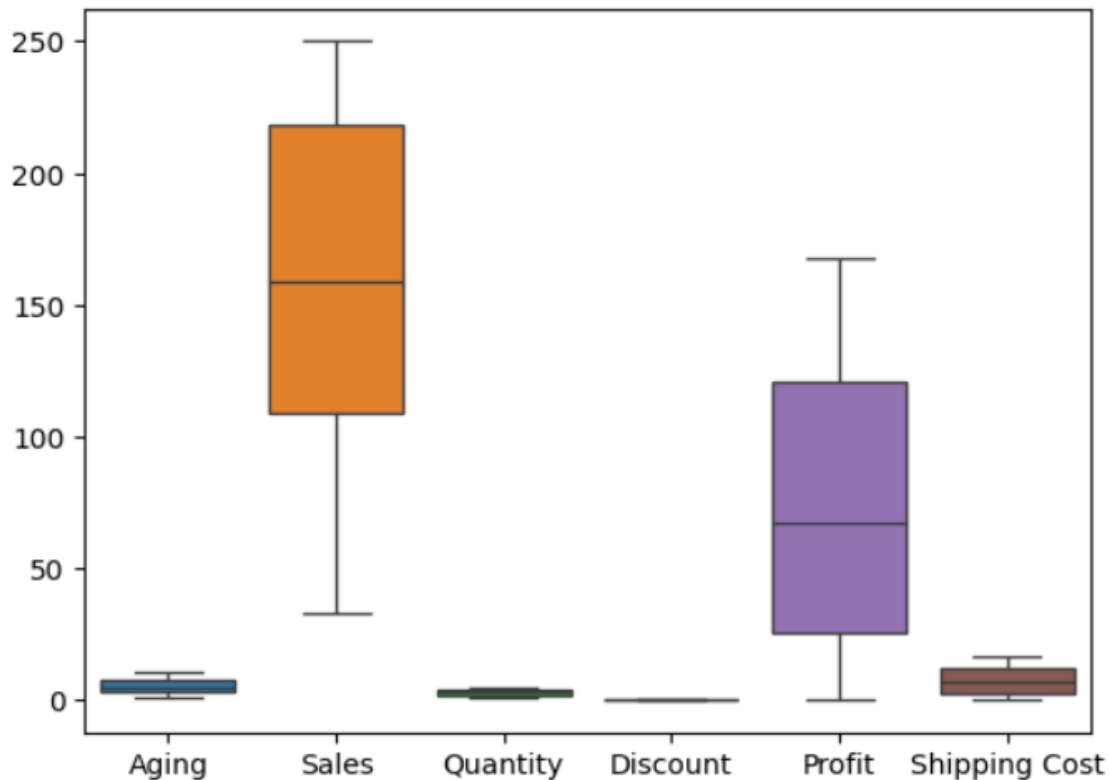
Before and after converting Order Date and Ship Date from Object datatype to Datetime64[ns] and sales, quantity, discount, profit, Shipping cost from object to float64 datatype.

Order ID	object
Order Date	object
Ship Date	object
Aging	float64
Ship Mode	object
Product Category	object
Product	object
Sales	object
Quantity	object
Discount	object
Profit	object
Shipping Cost	object
Order Priority	object
Customer ID	object
Customer Name	object
Segment	object
City	object
State	object
Country	object
Region	object
Months	object

Order ID	object
Order Date	datetime64[ns]
Ship Date	datetime64[ns]
Aging	float64
Ship Mode	object
Product Category	object
Product	object
Sales	float64
Quantity	float64
Discount	float64
Profit	float64
Shipping Cost	float64
Order Priority	object
Customer ID	object
Customer Name	object
Segment	object
City	object
State	object
Country	object
Region	object
Months	object

Data Visualisations

```
sns.boxplot(data=df1);
```



Visualization 1: Boxplot of Key Features

Code: `sns.boxplot(data=df1)`

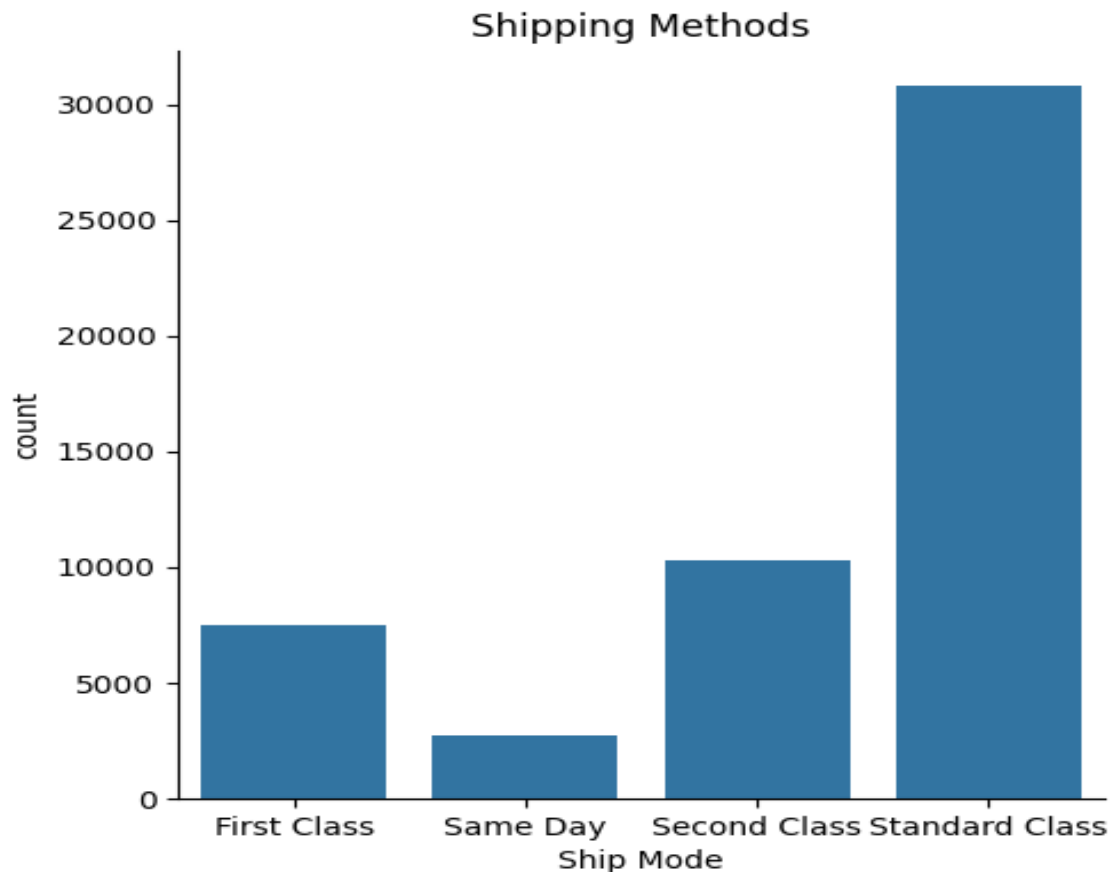
Explanation:

This boxplot visualizes the distribution of key numerical features such as **Aging**, **Sales**, **Quantity**, **Discount**, **Profit**, and **Shipping Cost**:

- **Sales and Profit:** Both show significant variance, with Sales being the most dispersed. This indicates diverse product pricing.
- **Shipping Cost:** Has minimal variability, reflecting standard rates.
- **Aging, Quantity, and Discount:** Show lower distributions and limited outliers, signifying constrained ranges in these metrics.
- Outliers are observed, particularly in **Profit** and **Sales**, which could represent high-value transactions or exceptional discounts.

```
#We note that the most used method is standard shipping
sns.catplot(x='Ship Mode',kind='count', data=df1 )
plt.title('Shipping Methods')
plt.figure(figsize=(50,10))
```

<Figure size 5000x1000 with 0 Axes>



Visualization 2: Count Plot of Shipping Methods

Code:

```
sns.catplot(x='Ship Mode', kind='count', data=df1)
```

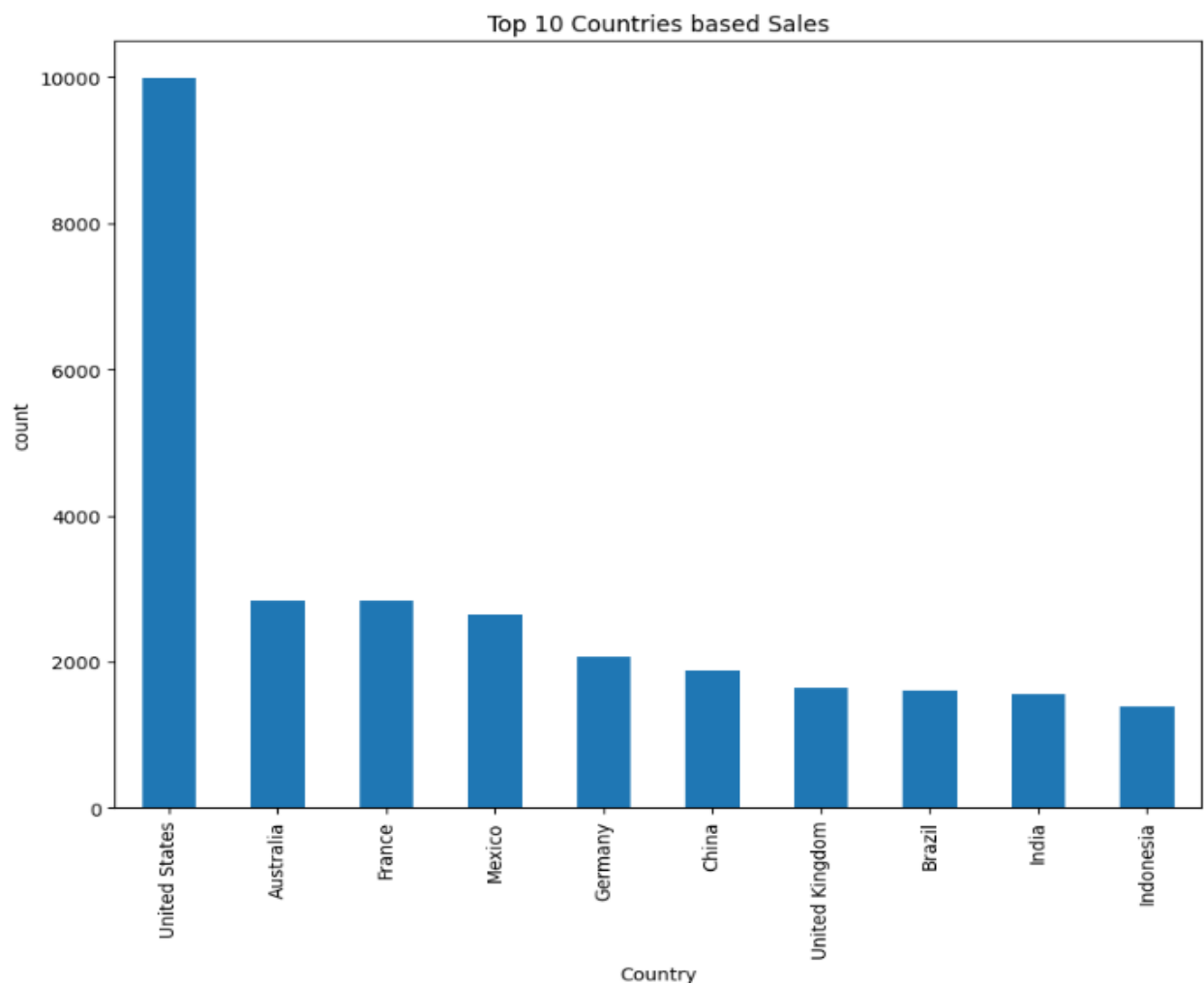
```
plt.title('Shipping Methods')
```

Explanation:

This count plot depicts the usage frequency of different shipping modes:

- **Standard Class** is the most commonly used shipping method, indicating a preference for cost-effective solutions over expedited delivery.
- Other modes, such as **First Class** and **Second Class**, are moderately used, while **Same Day** shipping sees minimal usage, likely due to its premium pricing.

- This insight is crucial for optimizing logistics and understanding customer preferences.



Visualization 3: Top 10 Countries by Sales

Code:

```
plt.bar(top_10_countries['Country'], top_10_countries['Count'])
```

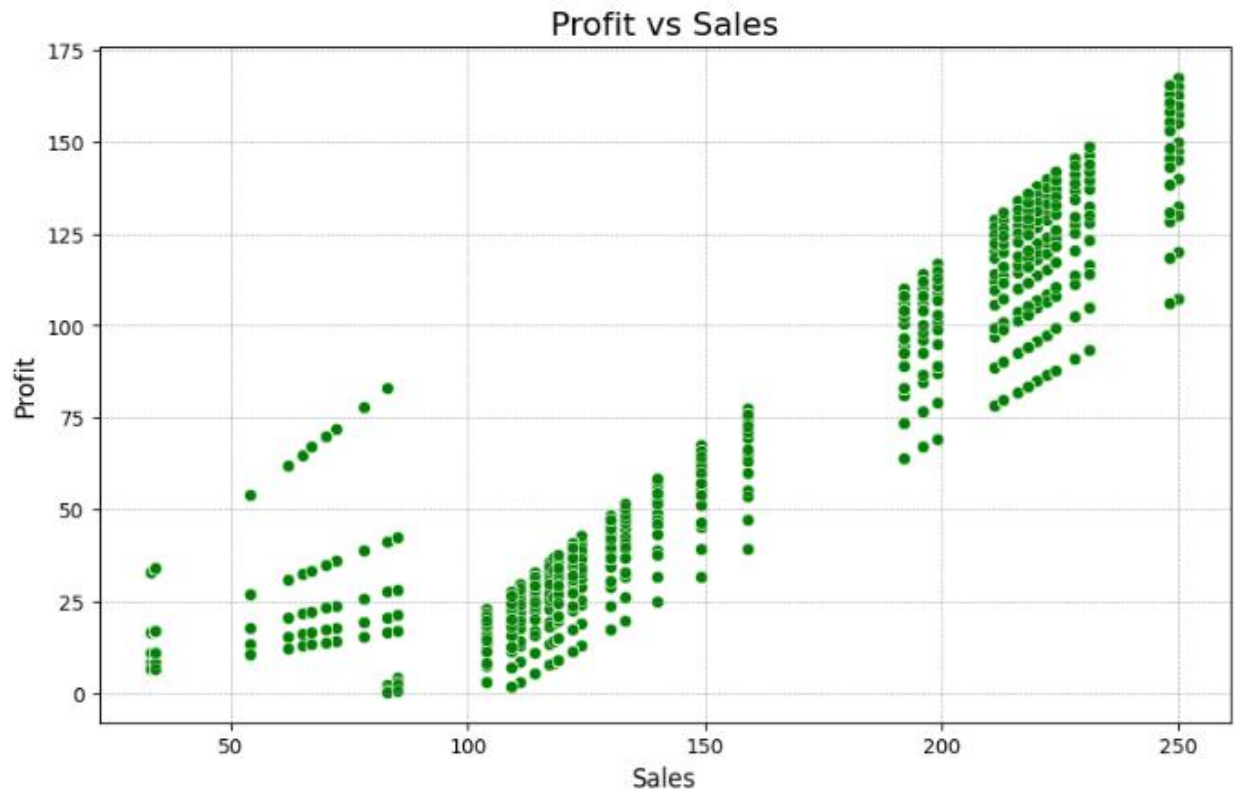
```
plt.title('Top 10 Countries Based on Sales')
```

Explanation:

This bar chart shows the top 10 countries contributing to sales:

- The **United States** dominates, accounting for the highest sales, followed by **Australia, France, and Mexico**.
- Emerging markets like **India and Indonesia** show significant but lesser contributions, indicating potential growth areas.

- These insights are critical for regional marketing strategies and inventory allocation.

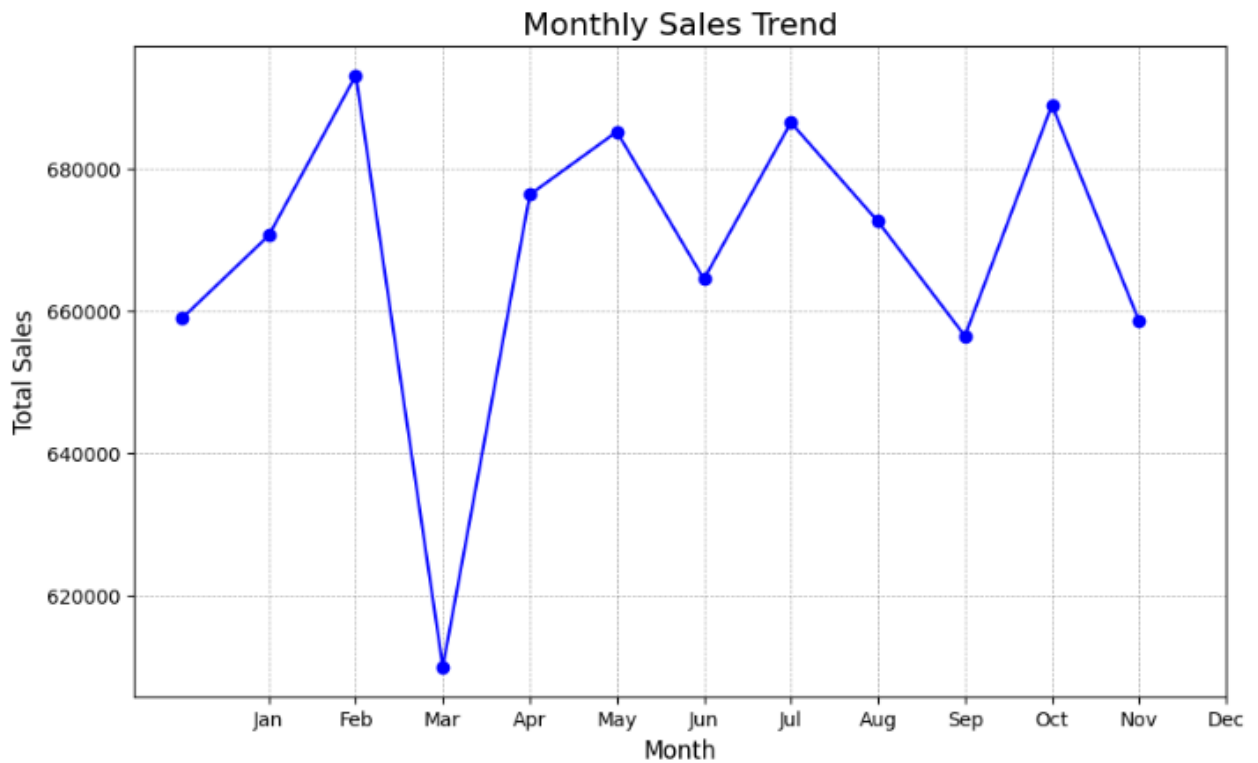


Visualization 4: Profit vs Sales

This scatterplot highlights the relationship between sales and profit for an e-commerce dataset. Each point in the chart represents a specific transaction, with **Sales** plotted on the x-axis and **Profit** on the y-axis. The following insights can be drawn from the chart:

- The chart reveals a positive correlation between sales and profit, indicating that higher sales generally lead to higher profits.
- The data points are clustered into distinct groups, possibly corresponding to different product categories, customer segments, or regions.
- A significant number of transactions with low sales have profits close to zero, suggesting either low-margin items or discounts impacting profitability.

This visualization supports the project's goal of analyzing financial performance and identifying patterns in sales and profitability across segments or regions.

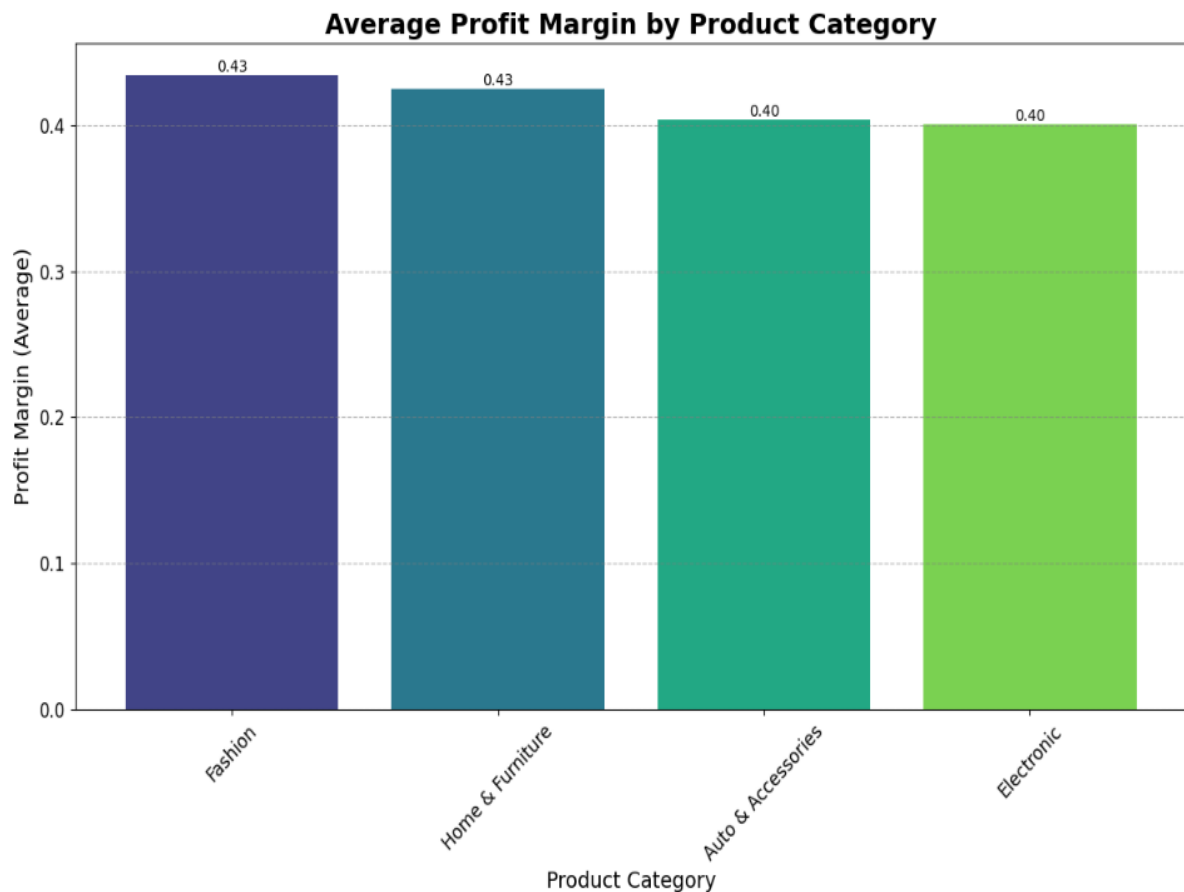


Visualization 5: Monthly Sales Trend

This line chart presents the **Monthly Sales Trend** over a year, with **Month** on the x-axis and **Total Sales** on the y-axis. Key observations include:

- Sales exhibit a fluctuating trend, with peaks in March, July, and October, and a significant dip in April.
- The cyclical nature of sales could correspond to seasonal trends, promotional events, or other external factors influencing customer purchasing behavior.
- Understanding these patterns can help businesses optimize inventory, plan marketing campaigns, and align their strategies with customer demand cycles.

This chart demonstrates the utility of temporal analysis for uncovering trends that can inform inventory and marketing decisions, aligning with the project's objectives in predictive modeling.

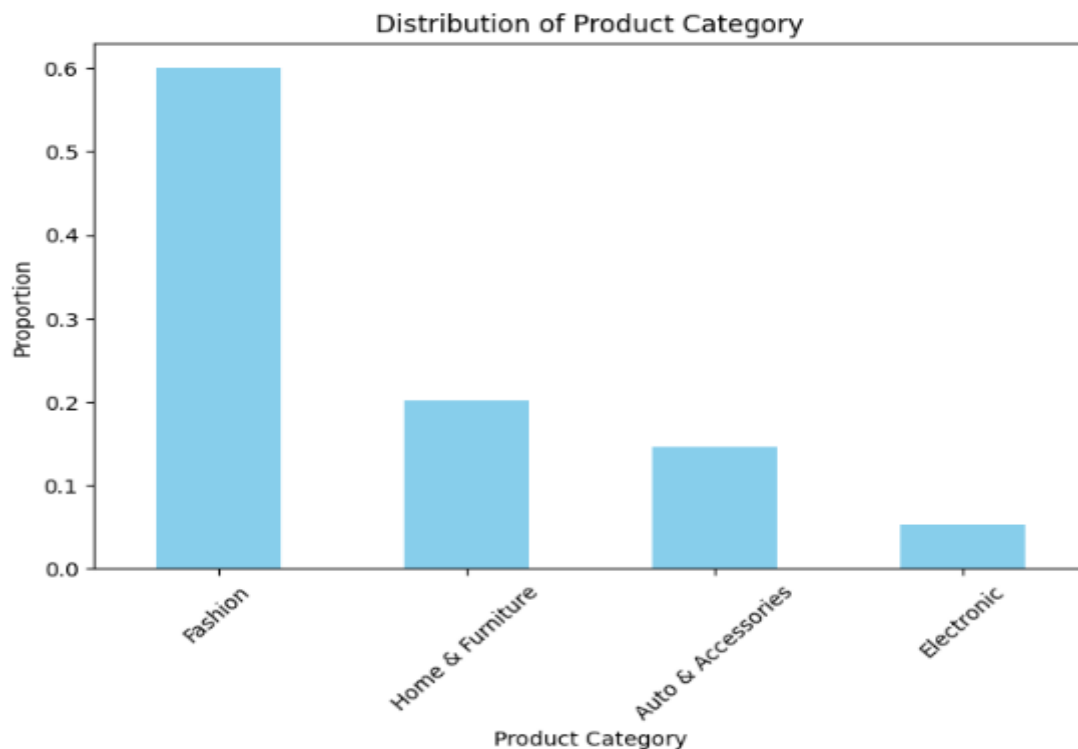


Visualization 6:

The bar chart compares the average profit margins across four product categories: Fashion, Home & Furniture, Auto & Accessories, and Electronics. Fashion and Home & Furniture lead with the highest average profit margin of 43%, indicating that these categories are slightly more profitable compared to the others. Auto & Accessories and Electronics follow closely with a profit margin of 40%, showing only a marginal difference in profitability among the categories.

This visualization highlights the consistency in profit margins across all categories, with a small range of variation. Such uniformity could suggest that these industries operate under similar pricing strategies or cost structures. Businesses in these sectors might benefit from focusing on other differentiating factors, such as volume, customer retention, or operational efficiency, to enhance profitability further.

```
Target Variable Distribution (Percentage):  
Product Category  
Fashion          0.600148  
Home & Furniture  0.201037  
Auto & Accessories 0.146142  
Electronic       0.052673  
Name: proportion, dtype: float64
```



The distribution of product categories, as shown in the data table and bar chart, highlights a significant imbalance in representation. Fashion holds the majority share, accounting for 60% of the dataset. This dominance indicates that Fashion is the most prevalent category, possibly due to a higher demand, production volume, or emphasis in the dataset's source. In contrast, the Home & Furniture category follows at 20%, while Auto & Accessories accounts for 14.6%. Electronics has the smallest share, comprising only 5.2%, making it the least represented category.

Such a skewed distribution suggests that the dataset is heavily centered around Fashion. This can have implications for any analyses derived from this data. For instance, conclusions about trends, profitability, or customer preferences might disproportionately favor the Fashion category due to its overwhelming presence. On the other hand, the relatively small proportion of Electronics could result in underrepresentation of this category's impact or potential in the broader analysis.

The variation in category representation may reflect the source's operational focus or market dynamics. For example, if the dataset originates from a retail platform or business that specializes in Fashion products, it would naturally explain the category's dominance. Similarly, the lower share of Electronics could indicate a niche focus or a smaller product

range in that segment. Understanding these nuances is critical to interpreting the data accurately and deriving actionable insights.

To address this imbalance, analyses may need to include normalization or weighting techniques to ensure that smaller categories like Electronics and Auto & Accessories are not overlooked. Alternatively, if the dataset is intended to prioritize Fashion, this focus should be explicitly acknowledged in any conclusions. By considering the broader context and addressing the distribution disparities, stakeholders can make well-rounded decisions that account for all categories fairly.



The heatmap above represents the correlation matrix between various features such as Aging, Sales, Quantity, Discount, Profit, Shipping Cost, and Profit Margin. Correlation coefficients range from -1 to 1, where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values around 0 indicate little to no correlation.

Sales and Profit show a strong positive correlation (0.92), indicating that higher sales are associated with higher profits. Similarly, Shipping Cost and Profit also have a strong correlation (0.92), suggesting that shipping costs might be directly tied to sales and profit levels. This trend is expected in businesses where increased shipping costs arise from higher sales volumes.

Profit Margin exhibits a moderate positive correlation with Profit (0.74) and Shipping Cost (0.74). However, it has a weaker correlation with Sales (0.47), suggesting that profit margin is influenced not only by sales but also by factors like pricing strategy, discounts, and shipping costs. On the other hand, Quantity has a weak negative correlation with Profit Margin (-0.39), indicating that selling larger quantities may reduce per-unit profit margins, possibly due to volume discounts or reduced pricing.

Interestingly, features like Aging and Discount have almost no correlation with most variables, suggesting they play a minimal role in determining the overall profit or profit margin. This heatmap provides valuable insights into relationships between variables, guiding decisions for optimizing sales, profits, and shipping strategies to enhance overall performance.



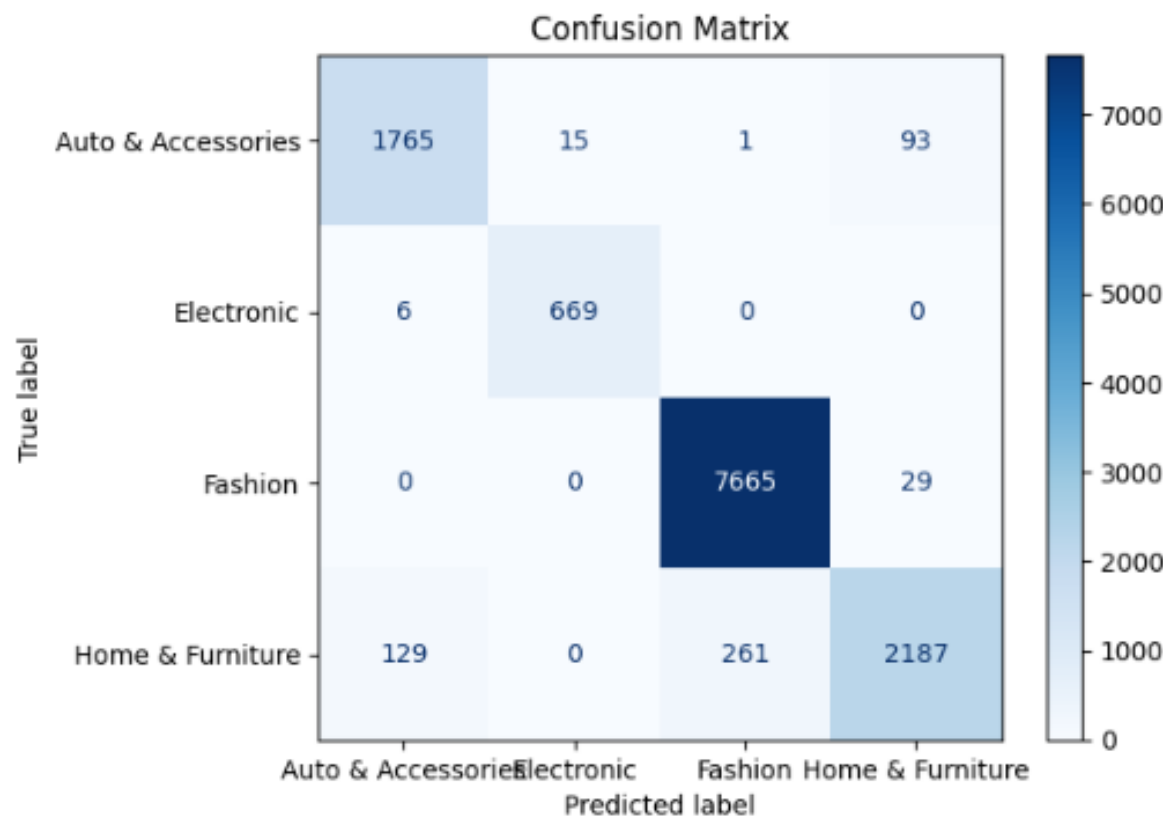
This scatter plot compares the actual sales values to the predicted sales values generated by a Random Forest Regressor. Each blue dot represents a pair of actual and predicted sales values, while the red line represents the perfect prediction line, where predicted values exactly match the actual values.

The alignment of most data points close to the red line indicates that the model performs well, accurately predicting sales in most cases. However, there are a few points slightly deviating from the line, representing instances where the predictions differ from the actual

values. These deviations might result from model limitations, data variability, or noise in the dataset.

The Random Forest Regressor seems to capture the relationship between the input features and sales effectively, as shown by the tight clustering of points near the perfect prediction line. This suggests the model has a strong predictive capability, making it suitable for forecasting sales based on the given dataset.

Overall, this visualization demonstrates the model's accuracy and reliability while highlighting opportunities to further refine the model by analyzing the instances where the predictions deviate from the actual values.



The confusion matrix visualizes the performance of a classification model across four product categories: Auto & Accessories, Electronic, Fashion, and Home & Furniture. Each cell shows the number of instances where the true labels (rows) are classified as the predicted labels (columns). The diagonal cells represent correct predictions, while off-diagonal cells represent misclassifications.

For **Auto & Accessories**, the model correctly classified 1,765 instances, with some misclassifications as Home & Furniture (93 instances) and Electronic (15 instances).

Electronic products show 669 correctly classified instances, with minimal misclassification. **Fashion**, the dominant category, has the highest number of correct predictions (7,665), but 29 instances were misclassified as Home & Furniture. **Home & Furniture** products have 2,187 correct predictions, with some confusion with Auto & Accessories (129 instances) and Fashion (261 instances).

The matrix highlights that the model performs exceptionally well for Fashion, likely due to its dominance in the dataset. However, there is notable confusion between Home & Furniture and other categories, possibly due to overlapping features or similarities between these categories.

This confusion matrix provides insights into the model's strengths and weaknesses. To improve the model, further feature engineering or balancing the dataset could help reduce misclassifications, particularly in underrepresented categories like Electronics and Auto & Accessories.

```
Column: Sales
Lower Bound: -54.5
Upper Bound: 381.5
Outliers Count: 0

Column: Profit
Lower Bound: -116.64999999
Upper Bound: 262.95
Outliers Count: 0

Column: Quantity
Lower Bound: -1.0
Upper Bound: 7.0
Outliers Count: 0

Column: Discount
Lower Bound: -0.0099999999
Upper Bound: 0.07
Outliers Count: 0

Column: Shipping Cost
Lower Bound: -11.65
Upper Bound: 26.35
Outliers Count: 0
```

The table presents an analysis of the numerical columns in a dataset, including **Sales**, **Profit**, **Quantity**, **Discount**, and **Shipping Cost**, to identify outliers and define their lower and upper bounds. For each column, the bounds are determined using statistical methods to highlight the range of typical values. Remarkably, no outliers were detected across all columns, indicating that the data is well-contained within the calculated bounds. This suggests consistency and reliability in the dataset, with values like Sales ranging from -54.5 to 381.5 and Profit from -116.65 to 262.95.

The absence of outliers in columns like Quantity (ranging from -1.0 to 7.0) and Discount (from -0.01 to 0.07) further confirms the uniformity of the data. Similarly, Shipping Costs, bounded between -11.65 and 26.35, exhibit no anomalies. This clean dataset provides a solid foundation for further analysis or model building without the need for additional preprocessing to handle extreme values. The lack of outliers ensures that results derived from this data will be unbiased and robust.

Dataset – 2 : Deep Learning

```
img_df.head()
```

	filename	link
0	15970.jpg	http://assets.myntassets.com/v1/images/style/p...
1	39386.jpg	http://assets.myntassets.com/v1/images/style/p...
2	59263.jpg	http://assets.myntassets.com/v1/images/style/p...
3	21379.jpg	http://assets.myntassets.com/v1/images/style/p...
4	53759.jpg	http://assets.myntassets.com/v1/images/style/p...

This image shows a sample from a DataFrame named `img_df`, which contains two columns: `filename` and `link`. The `filename` column lists the names of image files (e.g., 15970.jpg, 39386.jpg), while the `link` column provides the corresponding URLs for these images, hosted on the `assets.myntassets.com` server.

This dataset appears to be structured for tasks like image retrieval, classification, or analysis. Each row in the DataFrame represents a mapping between an image file and its online location. Such a format is commonly used in projects involving computer vision or e-commerce platforms to link product images to their respective data or features.

```
styles_df.head()
```

	id	gender	masterCategory	subCategory	articleType	baseColour	season	year	usage	productDisplayName
0	15970	Men	Apparel	Topwear	Shirts	Navy Blue	Fall	2011.0	Casual	Turtle Check Men Navy Blue Shirt
1	39386	Men	Apparel	Bottomwear	Jeans	Blue	Summer	2012.0	Casual	Peter England Men Party Blue Jeans
2	59263	Women	Accessories	Watches	Watches	Silver	Winter	2016.0	Casual	Titan Women Silver Watch
3	21379	Men	Apparel	Bottomwear	Track Pants	Black	Fall	2011.0	Casual	Manchester United Men Solid Black Track Pants
4	53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012.0	Casual	Puma Men Grey T-shirt

This image displays a sample of a DataFrame named `styles_df`, which contains detailed metadata about various products. The columns in this DataFrame are as follows:

1. **id**: A unique identifier for each product (e.g., 15970, 39386).
2. **gender**: Indicates the target gender for the product (e.g., Men, Women).
3. **masterCategory**: Broad category of the product (e.g., Apparel, Accessories).
4. **subCategory**: Specific subcategory within the master category (e.g., Topwear, Bottomwear, Watches).
5. **articleType**: The type of the product (e.g., Shirts, Jeans, Tshirts).
6. **baseColour**: The primary color of the product (e.g., Navy Blue, Blue, Silver).
7. **season**: The season associated with the product (e.g., Fall, Summer, Winter).
8. **year**: The year the product is associated with (e.g., 2011, 2016).
9. **usage**: The intended usage of the product (e.g., Casual).
10. **productDisplayName**: The descriptive name of the product (e.g., Turtle Check Men Navy Blue Shirt, Titan Women Silver Watch).

This dataset is used for cataloguing products, enabling tasks like product categorization, recommendation, or e-commerce search optimization. Combining this DataFrame with the `img_df` (image links) shown earlier could facilitate a complete mapping of product metadata to their images, which would be valuable for applications such as machine learning models in fashion or e-commerce platforms.

```
styles_df['filename'] = styles_df['id'].astype(str) + '.jpg'
```

	id	gender	masterCategory	subCategory	articleType	baseColour	season	year	usage	productDisplayName	filename
0	15970	Men	Apparel	Topwear	Shirts	Navy Blue	Fall	2011.0	Casual	Turtle Check Men Navy Blue Shirt	15970.jpg
1	39386	Men	Apparel	Bottomwear	Jeans	Blue	Summer	2012.0	Casual	Peter England Men Party Blue Jeans	39386.jpg
2	59263	Women	Accessories	Watches	Watches	Silver	Winter	2016.0	Casual	Titan Women Silver Watch	59263.jpg
3	21379	Men	Apparel	Bottomwear	Track Pants	Black	Fall	2011.0	Casual	Manchester United Men Solid Black Track Pants	21379.jpg
4	53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012.0	Casual	Puma Men Grey T-shirt	53759.jpg
...
44419	17036	Men	Footwear	Shoes	Casual Shoes	White	Summer	2013.0	Casual	Gas Men Caddy Casual Shoe	17036.jpg
44420	6461	Men	Footwear	Flip Flops	Flip Flops	Red	Summer	2011.0	Casual	Lotto Men's Soccer Track Flip Flop	6461.jpg
44421	18842	Men	Apparel	Topwear	Tshirts	Blue	Fall	2011.0	Casual	Puma Men Graphic Stellar Blue Tshirt	18842.jpg
44422	46694	Women	Personal Care	Fragrance	Perfume and Body Mist	Blue	Spring	2017.0	Casual	Rasasi Women Blue Lady Perfume	46694.jpg
44423	51623	Women	Accessories	Watches	Watches	Pink	Winter	2016.0	Casual	Fossil Women Pink Dial Chronograph Watch ES3050	51623.jpg

This table displays the `styles_df` DataFrame after adding a new column named `filename`, which is created by appending the `.jpg` extension to the `id` column. The updated DataFrame

now provides a direct mapping of each product's unique identifier to its associated image file name.

Key columns include:

1. **id**: The unique identifier for each product.
2. **gender, masterCategory, subCategory, articleType, baseColour, season, year, usage**: Provide detailed metadata about the product, such as its category, color, target gender, usage type, and seasonal association.
3. **productDisplayName**: A descriptive name for each product, which often includes brand and product type information.
4. **filename**: The new column added, which represents the image file name associated with each product.

By incorporating the filename column, this dataset can now be easily merged with the `img_df` DataFrame (containing image URLs). This integration would allow for a comprehensive linkage of product metadata to their respective images, enabling tasks like building an image-based recommendation system, training machine learning models for product classification, or enhancing search and filtering functionality in e-commerce platforms.

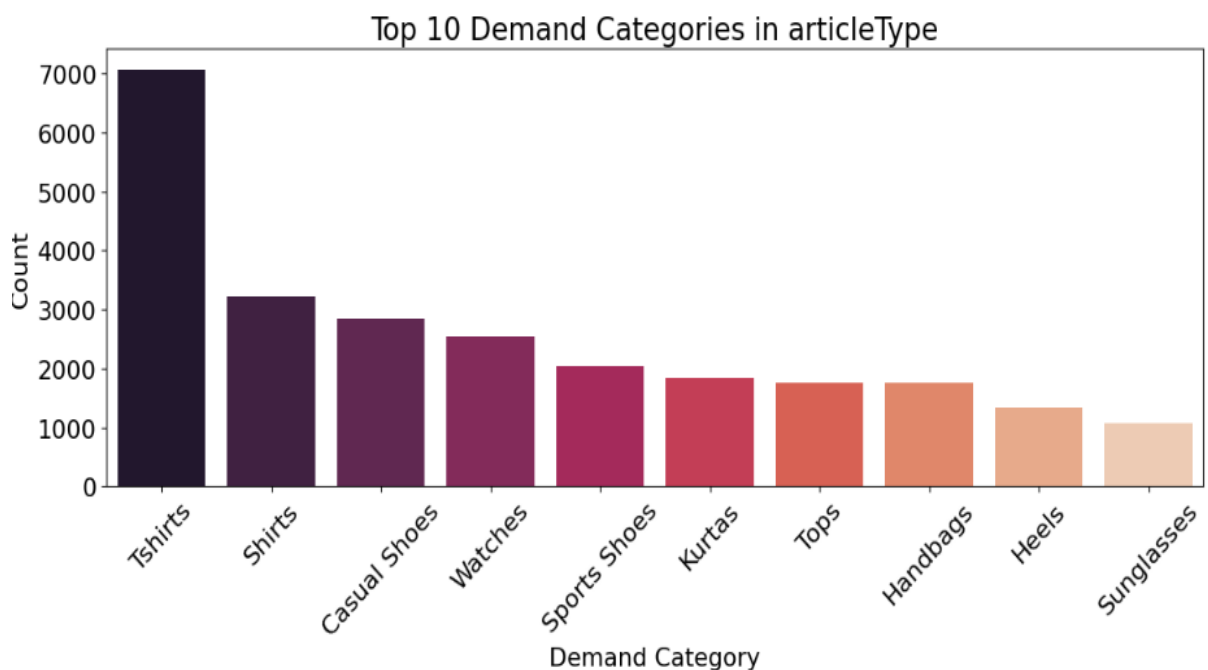
```
Most frequent categories:
masterCategory
Apparel          21397
Accessories      11274
Footwear         9219
Personal Care    2403
Free Items       105
Sporting Goods   25
Home             1
Name: count, dtype: int64
```

This table shows the count of items across different master Category values, indicating the distribution of products in the dataset. The most frequent categories are:

1. **Apparel**: The largest category with 21,397 items, indicating a significant focus on clothing-related products in the dataset.

2. **Accessories:** The second most frequent category with 11,274 items, representing a diverse range of non-clothing products such as watches, bags, and jewelry.
3. **Footwear:** Includes 9,219 items, highlighting its importance but with a smaller share compared to Apparel and Accessories.
4. **Personal Care:** Contains 2,403 items, likely comprising cosmetics, skincare, and related products.
5. **Free Items:** A minor category with 105 items, possibly promotional or complimentary items.
6. **Sporting Goods:** An even smaller category with just 25 items, indicating a minimal presence in the dataset.
7. **Home:** The least represented category with only 1 item.

This distribution reveals a strong emphasis on Apparel, Accessories, and Footwear, which dominate the dataset. Categories like Sporting Goods and Home are underrepresented, possibly due to limited offerings or a narrower scope in the dataset source. This imbalance could influence analyses or recommendations, necessitating consideration of the dataset's focus areas.



This bar chart displays the top 10 most in-demand categories in the articleType column based on their counts in the dataset. The demand is measured by the number of items available in each category, indicating their popularity or importance.

1. **Tshirts** dominate the chart, with the highest count of over 7,000 items, highlighting their widespread demand and significance in the dataset.
2. **Shirts** and **Casual Shoes** follow with relatively similar counts, reflecting their popularity among consumers.
3. Other significant categories include **Watches**, **Sports Shoes**, **Kurtas**, and **Tops**, all of which show substantial demand.
4. **Handbags**, **Heels**, and **Sunglasses** complete the list, with slightly lower counts but still within the top 10.

This chart emphasizes the focus on casual and versatile products like Tshirts, Shirts, and Shoes, which are essential wardrobe staples. It also highlights the importance of accessories such as Watches, Handbags, and Sunglasses, which cater to style and functionality. Businesses can use this data to prioritize these high-demand categories when making inventory or marketing decisions.

Most frequent subcategories:

subCategory	
Topwear	15402
Shoes	7343
Bags	3055
Bottomwear	2694
Watches	2542
Innerwear	1808
Jewellery	1079
Eyewear	1073
Fragrance	1011
Sandal	963

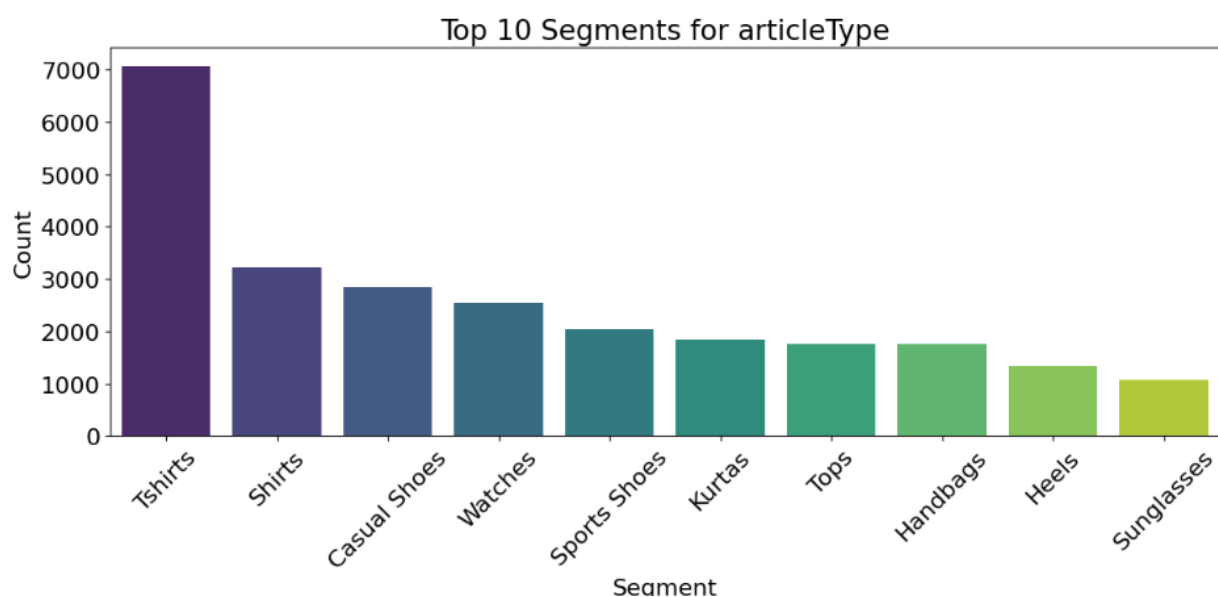
Name: count, dtype: int64

This table lists the most frequent subcategories in the dataset along with their counts, indicating the distribution of products across different subcategories:

1. **Topwear** is the most frequent subcategory with 15,402 items, showcasing its dominance, likely due to the broad variety and essential nature of topwear products.
2. **Shoes** follow with 7,343 items, highlighting their importance as a staple product category.
3. **Bags**, **Bottomwear**, and **Watches** appear next, with counts of 3,055, 2,694, and 2,542, respectively, reflecting their moderate yet significant presence in the dataset.

4. **Innerwear** and **Jewellery** have counts of 1,808 and 1,079, indicating a niche but notable representation.
5. **Eyewear**, **Fragrance**, and **Sandal** complete the list, each having over 1,000 items, showing their presence as specialized product categories.

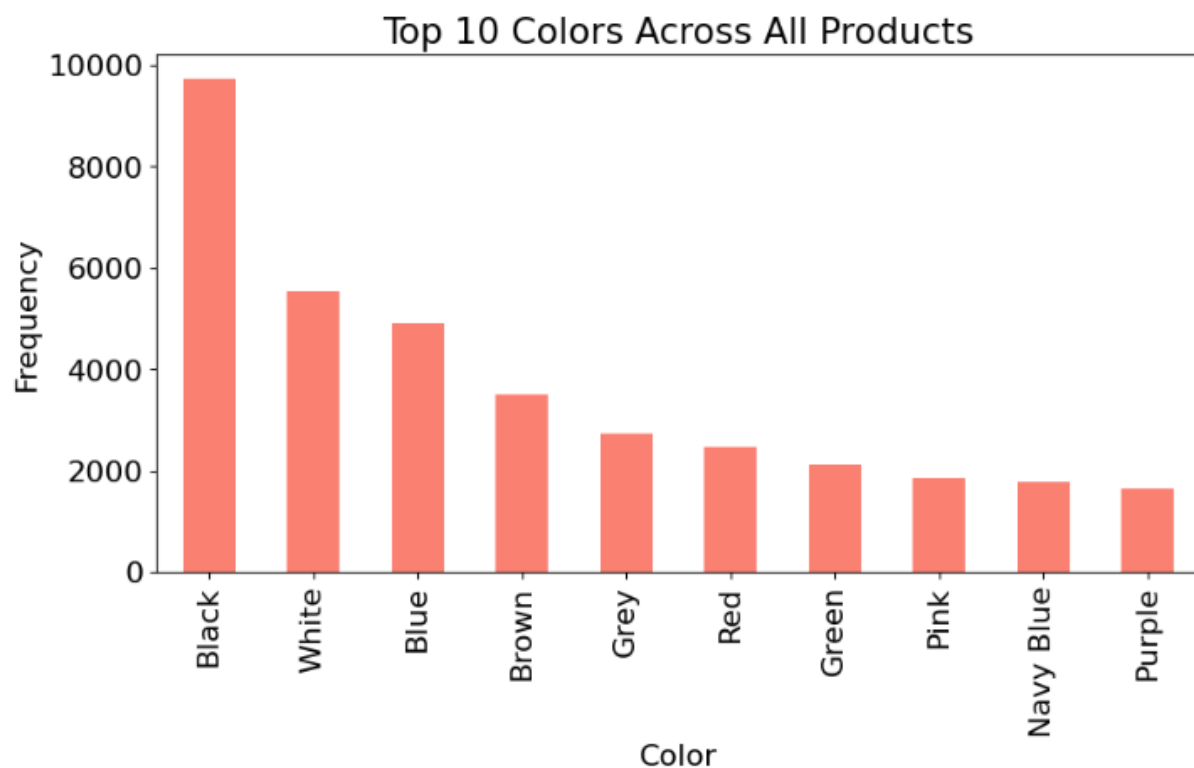
This distribution highlights the dataset's focus on essential and high-demand categories like Topwear and Shoes, while also including niche products like Fragrance and Sandals. These insights could guide targeted marketing strategies, inventory planning, or category-specific promotions.



This bar chart visualizes the top 10 segments in the article Type category based on their counts in the dataset, reflecting their relative popularity and demand:

1. **Tshirts** stand out as the most popular segment, with a count exceeding 7,000, highlighting their high demand and universal appeal.
2. **Shirts** and **Casual Shoes** follow with significant counts, emphasizing their importance as essential fashion and footwear items.
3. **Watches**, **Sports Shoes**, and **Kurtas** are mid-ranked segments, showcasing a balanced demand for both accessories and traditional wear.
4. **Tops**, **Handbags**, **Heels**, and **Sunglasses** round out the top 10, with relatively lower but still notable counts, indicating their value as complementary or fashion-driven items.

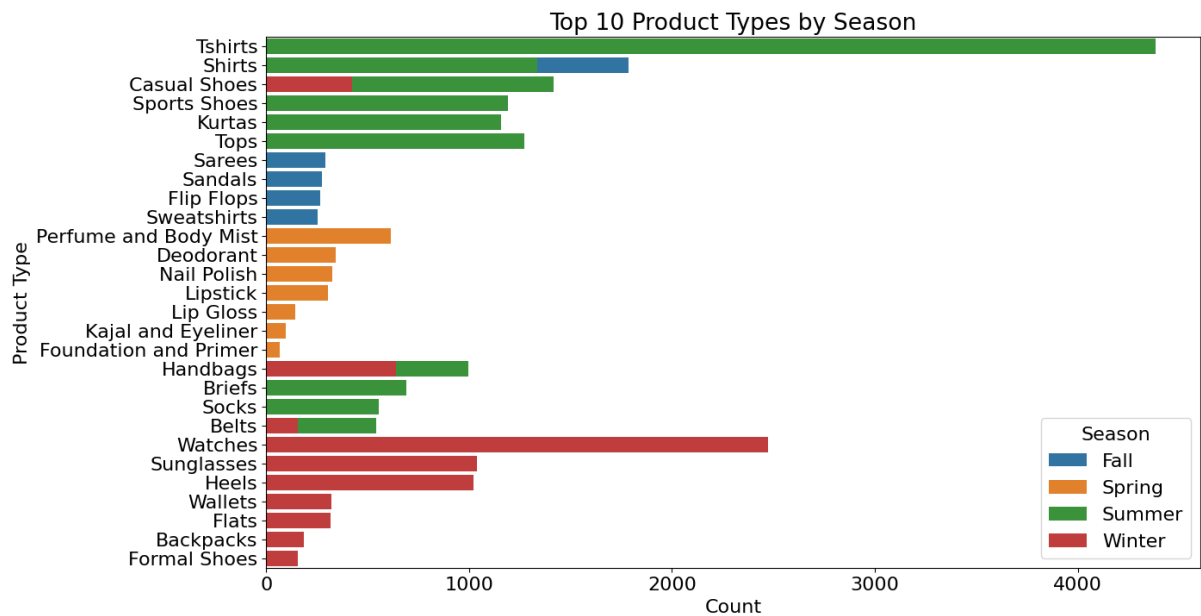
This chart underscores the dataset's focus on versatile and widely-used items like Tshirts and Shoes while maintaining a diverse range of categories such as accessories and traditional apparel. These insights are valuable for businesses in prioritizing inventory, creating marketing campaigns, and addressing customer preferences in high-demand segments.



This bar chart illustrates the top 10 colors used across all products in the dataset, showcasing their frequency of occurrence:

1. **Black** is the most dominant color, with a frequency of nearly 10,000, highlighting its universal appeal and versatility in fashion and product design.
2. **White** and **Blue** follow, with high frequencies, reflecting their popularity as staple colors across various product categories.
3. **Brown**, **Grey**, and **Red** rank in the mid-range, indicating their balanced demand for specific use cases or aesthetics.
4. **Green**, **Pink**, **Navy Blue**, and **Purple** complete the top 10, with relatively lower frequencies, showcasing their use as secondary or niche colors.

This distribution underscores the preference for neutral and versatile colors like Black, White, and Blue, which dominate consumer choices. These insights can help businesses prioritize popular color options while maintaining a diverse range of secondary colors to cater to specific customer preferences.



This bar chart displays the top 10 product types by season, with each bar divided into segments representing Fall, Spring, Summer, and Winter. The chart illustrates the distribution of popular product types and their association with specific seasons:

1. **Tshirts** and **Shirts** are most prevalent in the Summer season, reflecting their demand as lightweight, breathable clothing options during warm weather.
2. **Casual Shoes** and **Sports Shoes** maintain steady demand across multiple seasons, showcasing their versatility as year-round footwear.
3. **Kurtas** and **Tops** are primarily popular in Summer but also see demand during other seasons, especially in warmer climates.
4. **Watches** and **Sunglasses** are most associated with Winter and Summer, respectively, suggesting their practical and seasonal accessory roles.
5. **Sweatshirts** and **Sandals** align with colder and warmer seasons, respectively, underscoring their functional seasonal use.

This chart provides valuable insights into seasonal trends, helping businesses align their inventory and marketing strategies with customer preferences across seasons. For example, they can stock Tshirts and Sunglasses in Summer while focusing on Sweatshirts in Winter.

Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/vgg16/vgg16_weights_tf_dim_ordering_tf_kernels_notop.h5
58889256/58889256 ————— 2s 0us/step
Model: "functional"

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 224, 224, 3)	0
vgg16 (Functional)	(None, 7, 7, 512)	14,714,688
global_average_pooling2d (GlobalAveragePooling2D)	(None, 512)	0

Total params: 14,714,688 (56.13 MB)
Trainable params: 0 (0.00 B)
Non-trainable params: 14,714,688 (56.13 MB)

This table represents the summary of a pre-trained **VGG16** model loaded without its top classification layers. It is structured as a Keras functional model and provides details about the layers, their output shapes, and the number of parameters:

1. Input Layer:

- Type: InputLayer
- Output Shape: (None, 224, 224, 3)
- Represents the input shape for RGB images (224x224 pixels with 3 color channels).

2. VGG16 Layer:

- Type: Functional
- Output Shape: (None, 7, 7, 512)
- Consists of the convolutional base of VGG16, which outputs feature maps of size 7x7 with 512 filters.

3. Global Average Pooling Layer:

- Type: GlobalAveragePooling2D
- Output Shape: (None, 512)
- Reduces the spatial dimensions of the feature maps to a vector of size 512, enabling the output to serve as input for fully connected layers or other downstream tasks.

Parameters:

- Total Parameters: 14,714,688
- Trainable Parameters: 0
- Non-Trainable Parameters: 14,714,688

This indicates that the model is being used in a transfer learning setup, where the pre-trained VGG16 weights are frozen (non-trainable) and likely combined with a custom classifier for a specific task. The use of pre-trained weights helps leverage the powerful feature extraction capabilities of VGG16 without the need for training from scratch, saving time and computational resources.



This image demonstrates a product recommendation system focused on footwear. The system takes an **input image** (a pair of black sandals) and identifies **similar products**, displayed as five visually similar items labelled "Similar Product #1" through "Similar Product #5."

How it Works:

1. **Input Image Processing:** The system analyses the visual characteristics of the input image, such as shape, colour, and texture.
2. **Feature Extraction:** A deep learning model extracts features from the input image to represent it in a feature space.
3. **Similarity Search:** The extracted features are compared against a database of footwear images to find items with high similarity scores.
4. **Output Display:** The most similar items are ranked and presented to the user as recommendations.

Applications:

This system is useful in e-commerce platforms to enhance user experience by offering visually similar alternatives to a selected product. It can aid in better product discovery, increase customer satisfaction, and drive sales by suggesting related items.



This image illustrates a visual search or recommendation system for a fashion product. The system takes an **input image** (e.g., a man wearing a black shirt) and retrieves **similar products** based on visual similarity. The results are displayed as a set of five images labeled as "Similar Product #1" to "Similar Product #5."

Such a system likely uses techniques like:

1. **Feature Extraction:** A pre-trained deep learning model (e.g., VGG16 or ResNet) extracts visual features from the input image.
2. **Similarity Matching:** These features are compared to a database of product images using metrics like cosine similarity or Euclidean distance.
3. **Ranking:** The system ranks products based on their similarity scores, displaying the most similar items.

This approach is commonly used in e-commerce platforms to enhance the customer shopping experience by enabling users to find visually similar products, facilitating product discovery, and increasing engagement.

Conclusion

- This model bridges the gap between what a customer wants and what the catalog offers by relying on visual cues rather than textual descriptions.
- By analyzing the retrieved items for frequently searched products, we can identify top trends and optimize our inventory to meet customer demand.
- Through visual similarity, we can promote items that might have otherwise gone unnoticed, increasing sales and reducing unsold inventory.