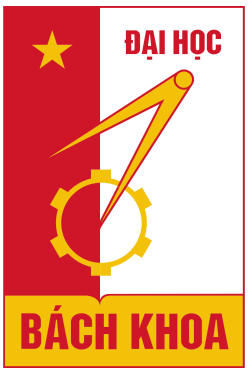
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
Symposium On Information and Communication Technology



Clothes Marketing Prediction

Course: Introduction to Data Science - IT4142E

Class Code: 152473

Supervisor: Associate Prof Than Quang Khoat

Member of our group:

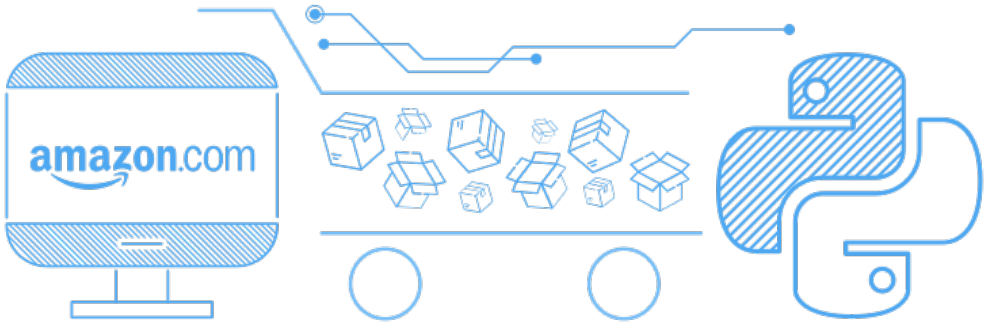
Name	Student ID	Email
Trinh Duy Phong	20220065	phong.td220065@sis.hust.edu.vn
Nguyen Viet Anh	20225434	anh.nv225434@sis.hust.edu.vn
Luu Hoang Phan	20225516	phan.lh225516@sis.hust.edu.vn
Hoang Trung Khai	20225502	khai.ht225502@sis.hust.edu.vn
Nguyen Thanh Minh	20225450	minh.nt225450@sis.hust.edu.vn

Contents

1 Introduction

In the dynamic and competitive clothing industry, understanding the factors that influence pricing is crucial for both consumers and businesses. This project focuses on developing a model to predict clothing prices based on key attributes such as brand, category, material, and other relevant product features. Through in-depth data exploration and analysis, we identify the primary factors that impact pricing. The model development phase utilizes machine learning algorithms to capture the complex relationships within the market. The evaluation and optimization processes ensure the model's accuracy and reliability. The successful implementation of this predictive model offers valuable insights for retailers, manufacturers, and consumers, helping to optimize pricing strategies and decision-making in the clothing market. This report provides a detailed overview of the data exploration, model development, and evaluation processes, offering a deeper understanding of the pricing dynamics in the clothing industry.

2 Data Collection



In this project, one of the most popular e-commerce sites,Amazon, will be mined for information about clothes.

Figure 1: HTTPX toolkit

Figure 2: Asyncio library

Figure 3: Parsel library

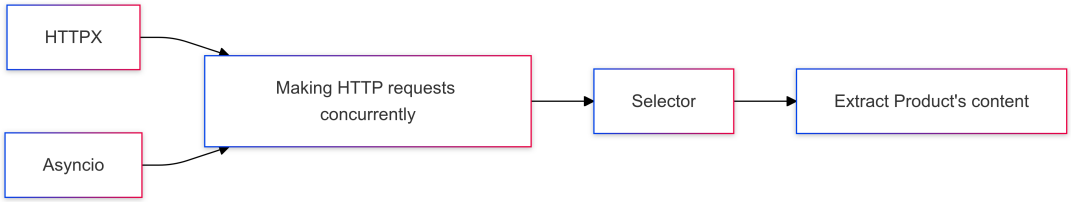


Figure 4: Library flow

In order to scrape Amazon reliably at large scale, we use:

1. HTTPX and Asyncio to make HTTP requests concurrently
2. Selector to extract products' content

2.1 Scraping Flow

We follow this workflow to scrape types of clothes and their variants (clothes in different colors and sizes)

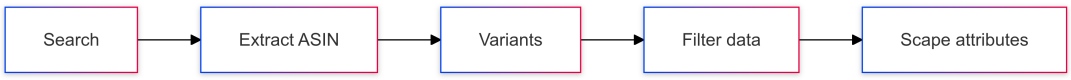


Figure 5: Scrape Flow

1. Search: Search the clothes according to categories (dress, shirt,...)
2. Extract ASIN: Extract ASIN from the links of products
3. Variants: Fetch all variants' ASIN of a product
4. Filter data: Only keep variants with color and size in the list of colors and sizes (further elaboration on Data Preprocessing section) or partly resemble to the list

5. Scrape attributes: Scrape all attributes of a product (list of attributes are on Attributes Scraped section)

2.2 Attributes Scraped

Attribute	Data Type	Description
Brand	Categorical	Brand name of the clothes.
Color	Categorical	Color of the clothes.
Size	Categorical	Size of the clothes.
Price	Categorical	Price of the clothes.
Fabric	Categorical	Fabric of the clothes.
Care	Categorical	How to clean the clothes.
Department	Categorical	Gender which the clothes are for.
Origin	Categorical	Origin of the clothes.
Closure	Categorical	Closure of the clothes.
First date	Categorical	Time the clothes are released to market.
Rating	Categorical	Number of customers rating the clothes.
Star	Categorical	Rating of the clothes.

Table 1: Attributes collected from Amazon.

3 Data Preprocessing

Data preprocessing is an essential step in preparing the dataset for model development. In this section, we detail the steps taken to clean the data, handle missing values, and standardize categorical attributes.

	name	asin	brand	color	size	price	fabric	care	department	origin	closure	first_date	rating	star	url
0	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CT3G46LT	Brand: PARISDIARY	3 Packs-black/Dark Gray/Navy	Small	$44.99	87% Polyester, 13% Spandex	Machine Wash	NaN	Imported	Zipper	NaN	909 ratings	4.5	https://www.amazon.com/dp/B0CT3G46LT
1	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CYGX1VRB	Brand: PARISDIARY	3 Packs-black/Dark Gray/Camouflage Gray	Large	$45.99	87% Polyester, 13% Spandex	Machine Wash	NaN	Imported	Zipper	NaN	909 ratings	4.5	https://www.amazon.com/dp/B0CYGX1VRB
2	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CWLJN23H	Brand: PARISDIARY	3 Packs-black/Black/Navy	X-Large	$44.99	87% Polyester, 13% Spandex	Machine Wash	NaN	Imported	Zipper	NaN	909 ratings	4.5	https://www.amazon.com/dp/B0CWLJN23H
3	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CT3JXKKT	Brand: PARISDIARY	3 Packs-black/Dark Gray/Blue	Medium	$44.99	87% Polyester, 13% Spandex	Machine Wash	NaN	Imported	Zipper	NaN	909 ratings	4.5	https://www.amazon.com/dp/B0CT3JXKKT

Figure 6: Raw data

3.1 Categorical Data Encoding

In our project, the dataset comprises entirely categorical attributes represented as strings. To make the data suitable for model training, we converted these categorical variables into numerical representations using encoding techniques. This transformation

ensures compatibility with machine learning algorithms while preserving the underlying information embedded in the original categories.

3.2 Removing the `first_date` Column

"The `first_date` column contains information about the date when the item was first introduced. However, the dataset lacks diversity in terms of seasons and years, making this column less meaningful for the price prediction task. Due to its limited variability and relevance, we remove the `first_date` column to avoid introducing unnecessary noise and ensure a more focused analysis."

3.3 Standardizing Colors and Sizes

To ensure consistency in the dataset, values in the `color` and `size` columns are standardized. Any variations or partial matches in color and size are mapped to a predefined list of valid categories. The predefined filters are as follows:

- **Color Filters:** `["black", "white", "gray", "beige", "red", "blue", "green", "brown", "pink", "purple", "yellow"]`
- **Size Filters:** `["small", "medium", "large", "x-large", "xx-large"]`

Since the clothing sizes have an inherent order or ranking (small < medium < large < x-large < xx-large), ordinal categorical variables are used. Ordinal encoding assigns each category a numerical value based on its rank or order.
Ordinal Encoding for Clothing Sizes:

- `"small"` $\rightarrow$ 0.0
- `"medium"` $\rightarrow$ 1.0
- `"large"` $\rightarrow$ 2.0
- `"x-large"` $\rightarrow$ 3.0
- `"xx-large"` $\rightarrow$ 4.0

3.4 Handling Missing Values in the `color` Column

Missing values in the `color` column are handled by filling them in a manner that maintains the distribution of the existing color values. We calculate the probability distribution of the available color values and use this distribution to randomly assign missing values, ensuring the overall color distribution is preserved.

```
# Handle missing value for color
import numpy as np
color_probabilities = df['color'].value_counts(normalize=True, dropna=True)
nan_indices = df['color'][df['color'].isna()].index
df.loc[nan_indices, 'color'] = np.random.choice(color_probabilities.index, size=len(n
```

These preprocessing steps ensure that the dataset is clean, consistent, and ready for analysis and model training. Missing values are addressed appropriately, and categorical variables are standardized to improve the accuracy of the model.

	name	asin	brand	color	size	price	department	origin	rating	star	...	Not Bleach	Tumble Dry	Pull on	Tie	Zipper	Button	No closure	Elastic	Lace Up	Drawstring
0	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CT3G46LT	Parisdiary	Navy	0.0	44.99	Men	Imported	909.0	4.5	...	0	0	0	0	1	0	0	0	0	0
1	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CYGX1VRB	Parisdiary	None	2.0	45.99	Men	Imported	909.0	4.5	...	0	0	0	0	1	0	0	0	0	0
2	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CWLJN23H	Parisdiary	Black	3.0	44.99	Men	Imported	909.0	4.5	...	0	0	0	0	1	0	0	0	0	0
3	3 Pack Mens Athletic Sweatpants with Zipper Po...	B0CT3JXKKT	Parisdiary	Blue	1.0	44.99	Men	Imported	909.0	4.5	...	0	0	0	0	1	0	0	0	0	0

Figure 7: Preprocess data

There are 19768 instances of processed data and 35267 instances of raw data.

4 Exploratory Data Analysis

4.1 Outlier Detection

Box plots and scatter plots were used to identify outliers. Rows with extreme values were removed based on thresholds such as:

- CPU speed $>$ 40 GHz.
- RAM $>$ 80 GB.
- PPI $>$ 350.
- Weight $>$ 15 kg.

4.2 Data Visualization

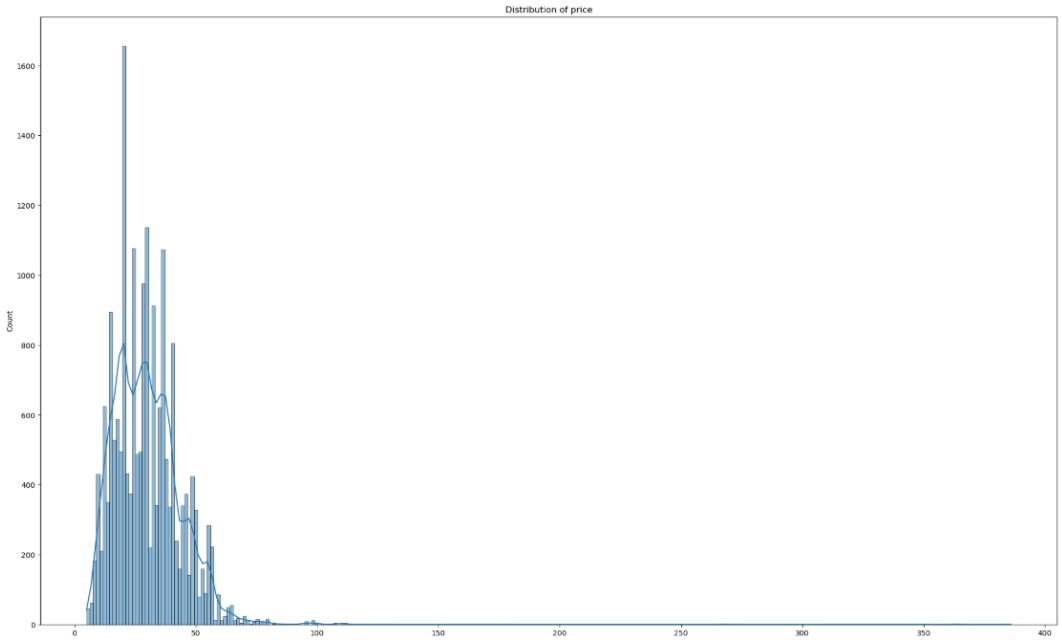


Figure 8: Distribution of Price

The attribute **Price** for clothes exhibits a significant skew, with many outliers beyond the 75% quantile. Another observation is that the distribution of prices is not centered but

is right-skewed, indicating that most prices are concentrated at the lower end of the scale. The above figures show that the most common price range lies between approximately 0–50 dollars. The lowest-priced item in the dataset starts at just a few dollars, while a few premium items extend well beyond 200 dollars, highlighting a wide range of affordability and luxury options in the dataset.

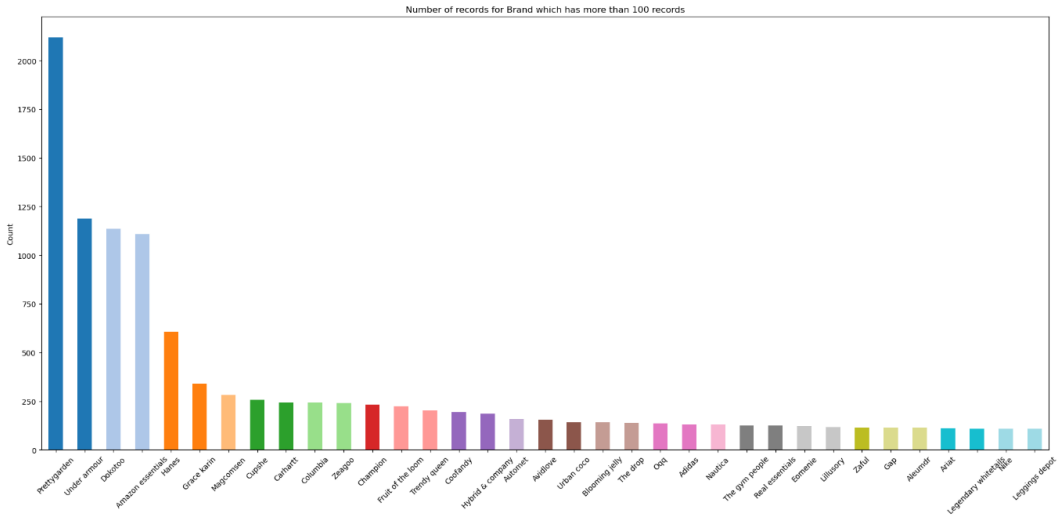


Figure 9: Number of records for Brand which has more than 100 records

The **Brand** attribute exhibits a highly uneven distribution, with only a few brands dominating the dataset. The bar chart highlights that ”Prettgarden” has the highest number of records, surpassing 2000, followed by ”Under Armour,” ”Dokotoo,” and ”Amazon Essentials,” each with over 1000 records. The remaining brands have significantly fewer records, many of them under 500. This indicates a concentration of data within a limited number of brands, which may influence the modeling process and necessitate careful handling to avoid biased predictions toward the dominant brands.

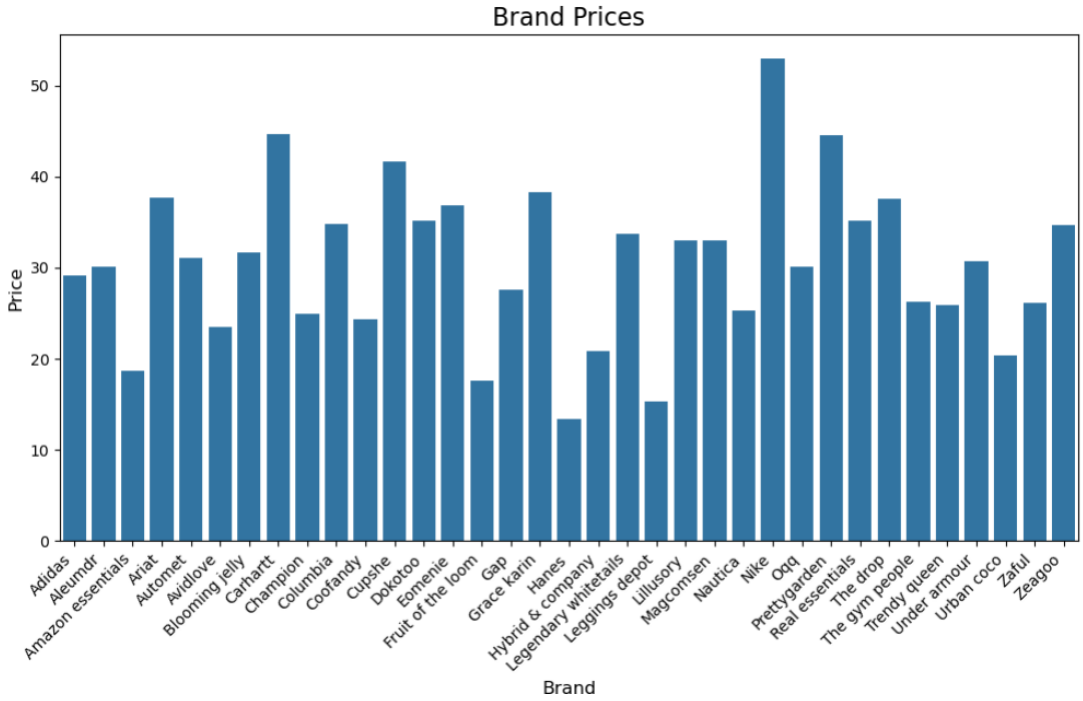


Figure 10: Brand Price

In analyzing the **Price** attribute for clothes, the data revealed a notable skew, with a significant number of outliers exceeding the 75% quantile. Interestingly, the price distribution is far from symmetrical; it leans heavily towards the lower end, creating a pronounced right skew. The figures illustrate that the majority of items are priced between 0 and 50 dollars, representing the most common price range. While the dataset includes budget-friendly options starting at just a few dollars, it also captures a small selection of premium items priced well over 200 dollars. This wide spectrum of pricing highlights both affordability and luxury within the product range.

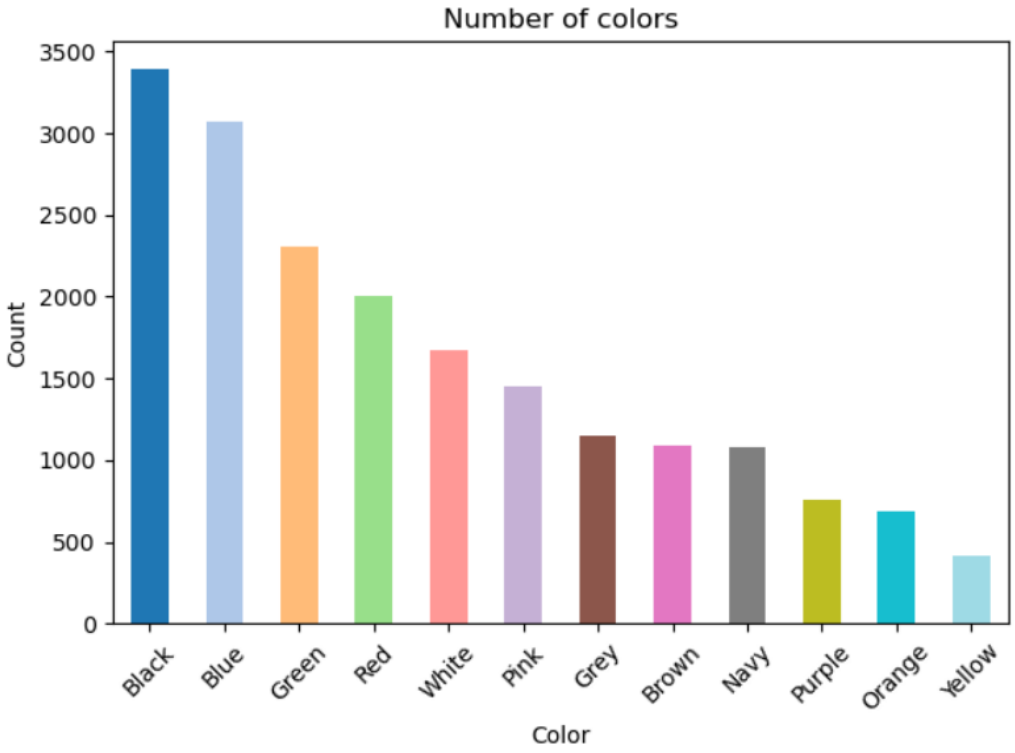


Figure 11: Number of colors

Black is the most popular color, with over 3500 entries, followed by **Blue**, while colors like **Yellow** and **Orange** are much less common, with fewer than 1000 entries each. Overall, the chart highlights a strong preference for neutral and versatile colors, such as Black, Blue, and Grey. This suggests that customers tend to favor classic, easy-to-match colors, providing useful insights for brands when planning inventory or marketing campaigns.

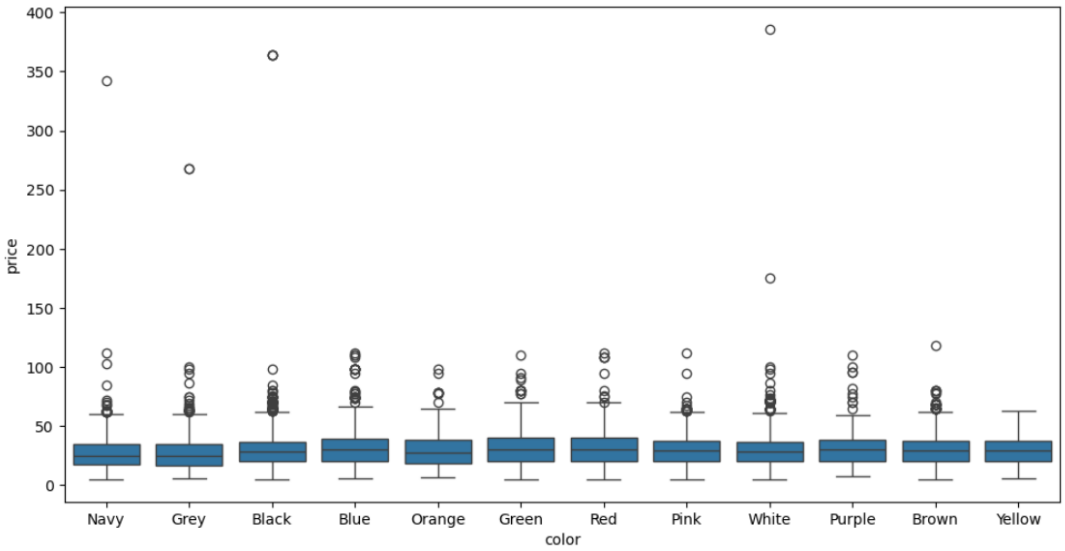


Figure 12: Color Boxplots

The price distribution are compared across different clothing colors. A key observation

is that most colors have a similar median price, clustered around the lower end, with relatively small interquartile ranges. However, outliers are evident in every category, with some prices exceeding 300 dollars, particularly in colors like **Black** and **White**. Overall, this consistent pricing across colors indicates that color may not be a primary determinant of price, but the presence of outliers suggests that certain premium products might skew the data for specific colors. This analysis provides valuable insights for understanding pricing strategies and customer preferences.

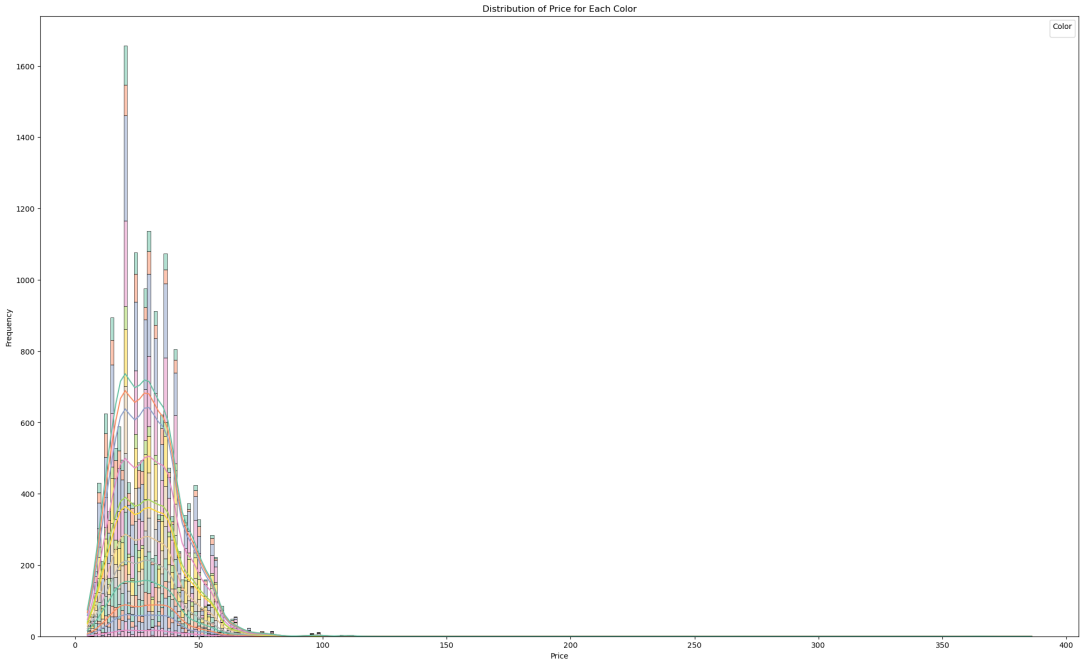


Figure 13: Distribution of Price by Color

The graph presents the distribution of prices across various colors in the clothing dataset. A prominent observation is that the majority of items, irrespective of their color, are priced between 0 and 50 dollars, with a sharp decline in frequency for higher price ranges. Each color displays a similar right-skewed pattern, indicating that lower-priced items dominate for all colors, while premium-priced items are relatively rare. This visualization highlights that pricing trends are consistent across colors, suggesting that price is more influenced by other factors, such as product type or brand, rather than color alone.

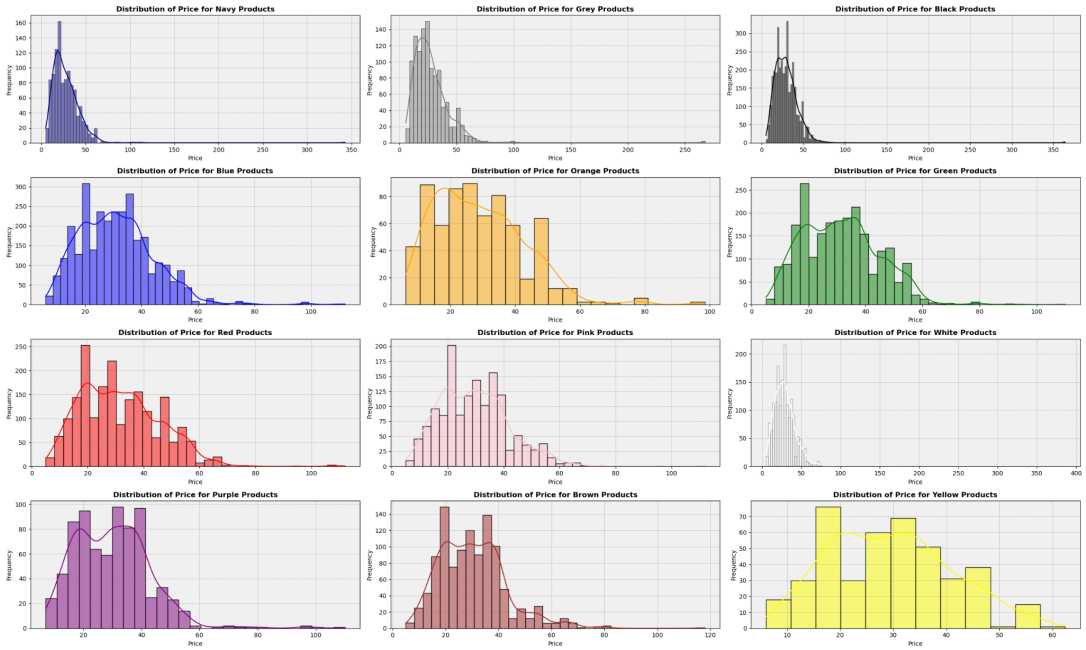


Figure 14: Distribution of Price for Each Color

The set of histograms illustrates the price distributions for products grouped by color. A key feature across all charts is the right-skewed distribution, with most items clustered at lower price ranges and a gradual decrease in frequency as prices rise. While the general trend is consistent, some colors, such as **Black** and **Grey**, exhibit higher maximum prices and more pronounced outliers compared to others like **Yellow** or **Pink**. These insights suggest that color does not drastically impact price but might influence outlier occurrences, which could be linked to specific product types or premium offerings in certain color categories. This analysis aids in identifying pricing trends and variations across different color segments.

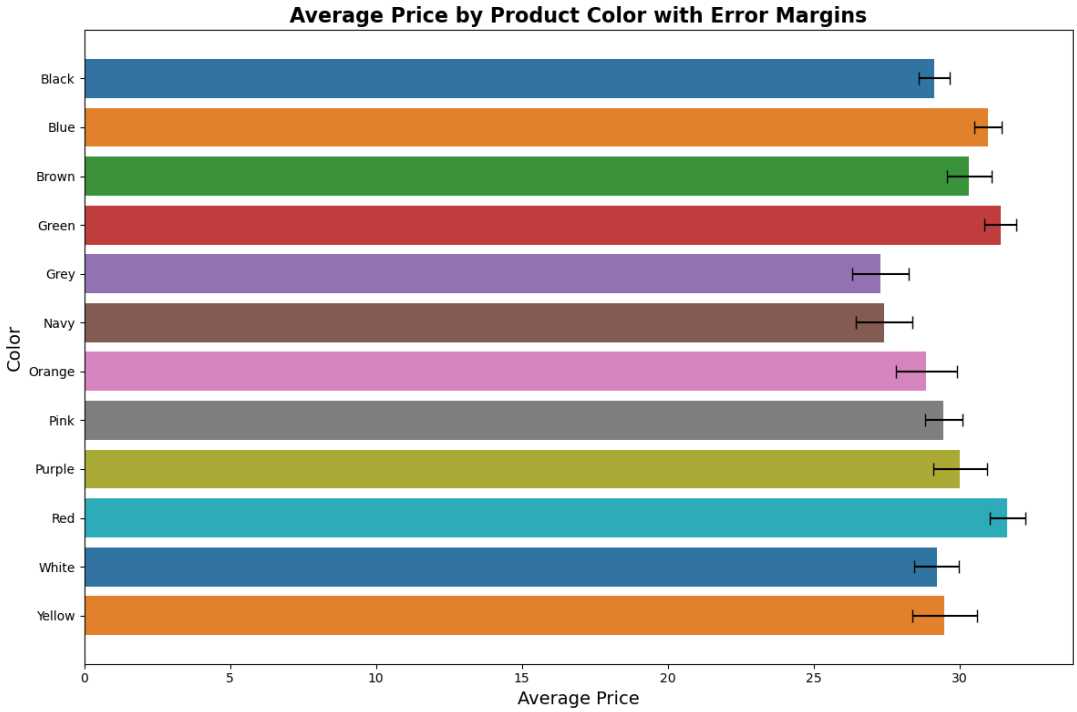


Figure 15: Average Price by Product Color with Error Margins

The bar chart shows the average price of products for each color, with error bars representing the variation in prices within each group. Overall, the average prices are fairly consistent across colors, typically falling between 25 and 30 dollars. While most colors have small error margins, a few, like **Green** and **Brown**, show slightly more variation, suggesting a wider range of prices for those categories. This indicates that, while there is some variability, color doesn't appear to play a major role in determining the average price, pointing to a fairly uniform pricing strategy across the dataset.

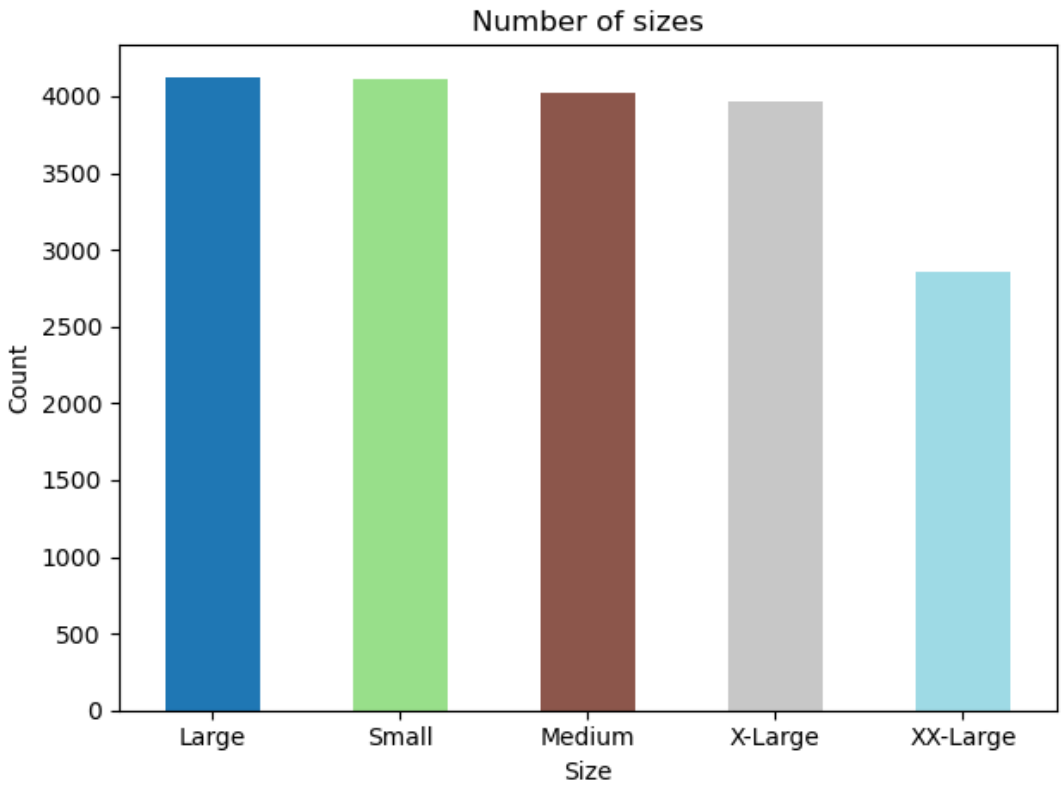


Figure 16: Number of sizes

The distribution of product sizes reveals that **Large**, **Small**, and **Medium** sizes dominate the dataset, each accounting for over 4000 products. In contrast, **X-Large** is moderately represented, and **XX-Large** has the fewest entries. This trend suggests a clear focus on standard sizing, which appears to cater to the majority of consumers. The relatively lower availability of extended sizes like **XX-Large** highlights a potential gap in the market for products targeting customers outside the standard size range.

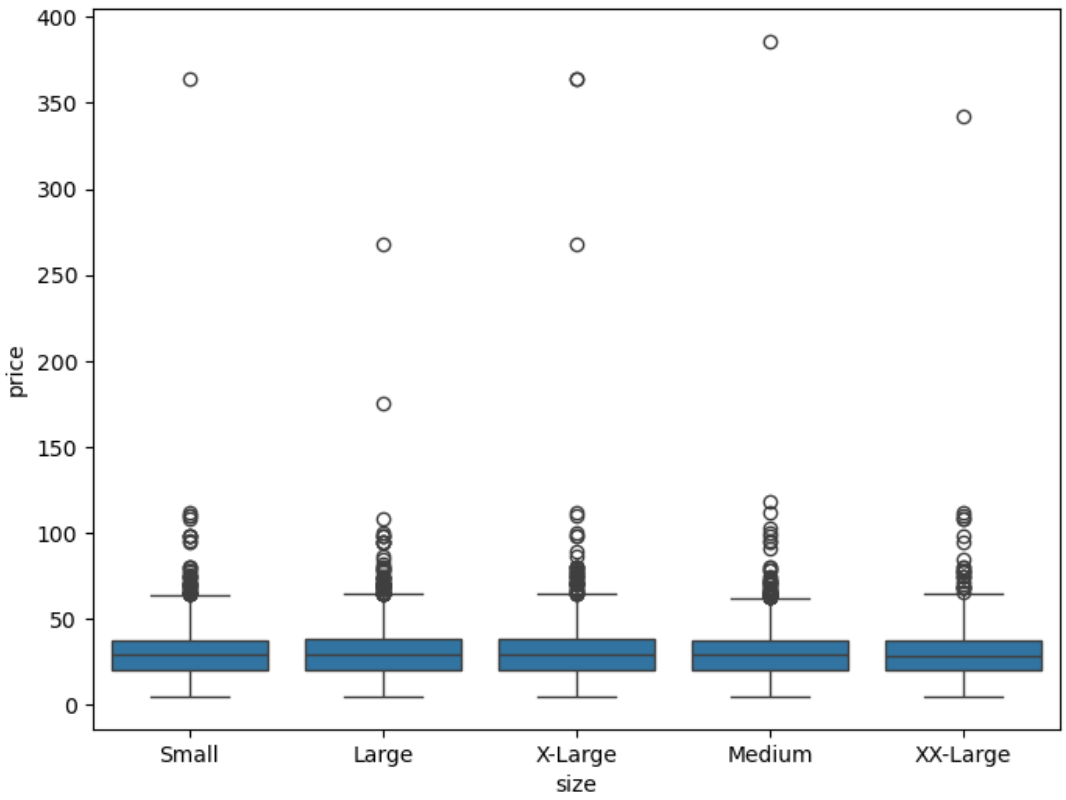


Figure 17: Sizes Boxplots

Across all size categories, the price distribution exhibits a consistent pattern, with most items falling below 50 dollars, as shown by the uniform medians. While the interquartile ranges are similar, indicating that size has little effect on price variation, a notable feature is the presence of outliers. Sizes such as **Large** and **Medium** show more extreme prices, with some items exceeding 350 dollars. This suggests that while most prices remain consistent, high-priced outliers likely represent premium or specialized products, irrespective of size.

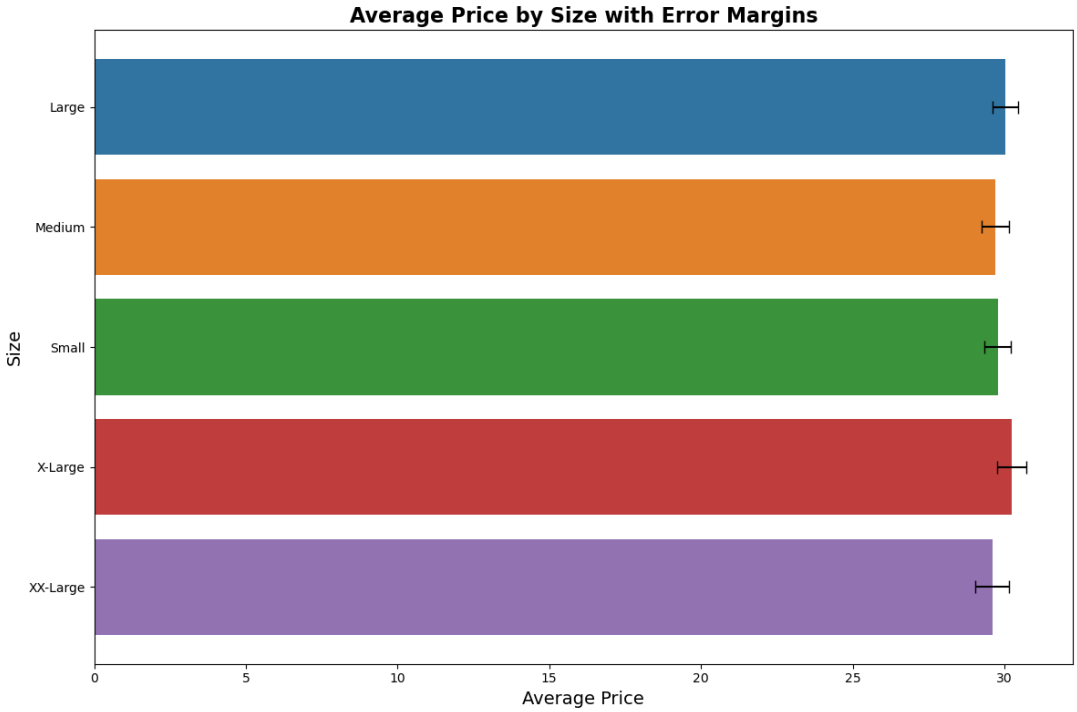


Figure 18: Average Price by Product Color with Error Margins

The average price remains consistent across all size categories, falling within the range of 25 to 30 dollars. The error margins, shown as small bars, reveal minimal variability in prices for each size, suggesting a uniform approach to pricing regardless of size. This uniformity indicates that size has little influence on price differentiation, providing further evidence of a standardized pricing strategy across the dataset.

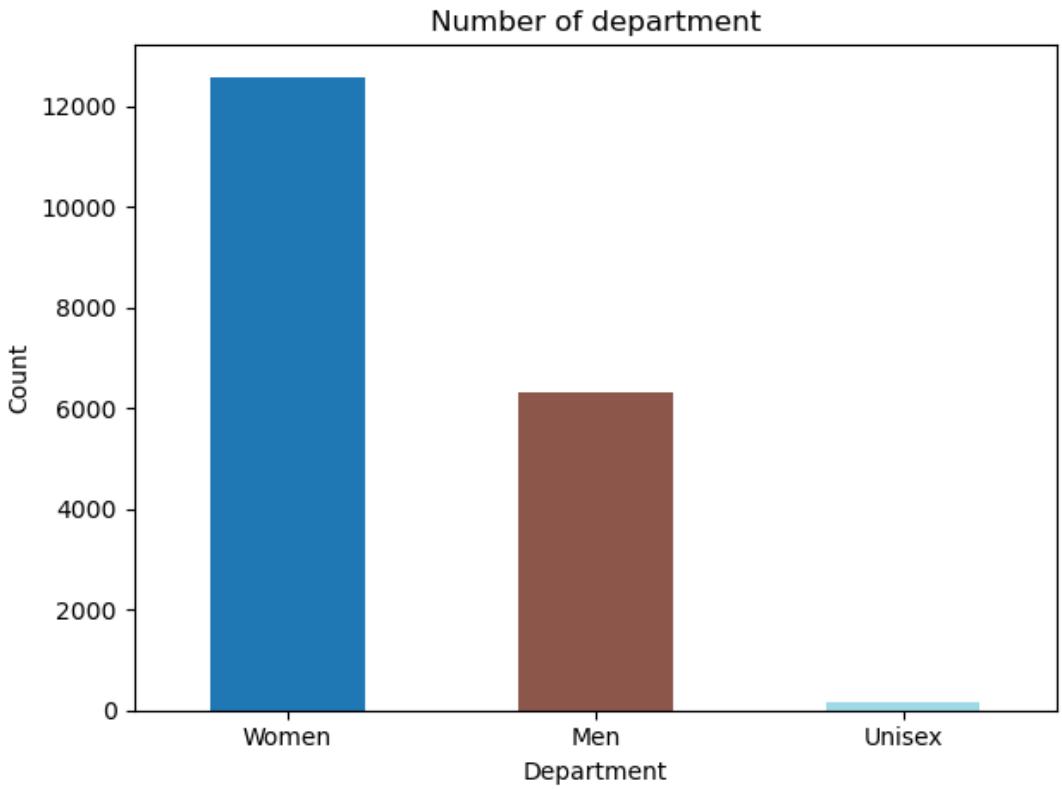


Figure 19: Number of department

The bar chart illustrates the distribution of products across different departments: **Women**, **Men**, and **Unisex**. A significant observation is that the **Women** department dominates, with over 12,000 products, followed by the **Men** department, which has about half as many. The **Unisex** category represents only a small fraction of the total, with minimal entries. This trend suggests a strong market focus on women's clothing, indicating that product offerings are largely tailored to meet the demands of female consumers, with less emphasis on men's or unisex products.

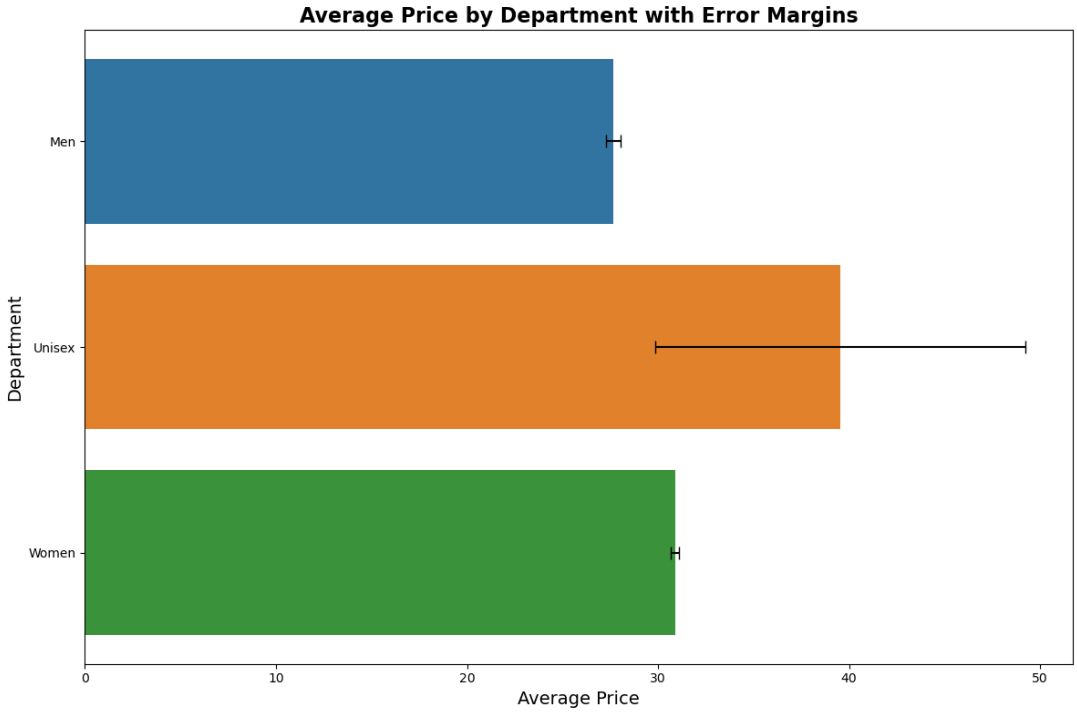


Figure 20: Average Size by Product Color with Error Margins

The **Unisex** department has the highest average price, with a large error margin indicating significant variability in pricing. In contrast, the **Men** and **Women** departments have similar average prices, ranging between 25 and 30 dollars, and their smaller error margins reflect more consistent pricing. This suggests that the Unisex category likely includes a broader mix of products, ranging from standard to premium, while pricing for Men and Women's items is more uniform.

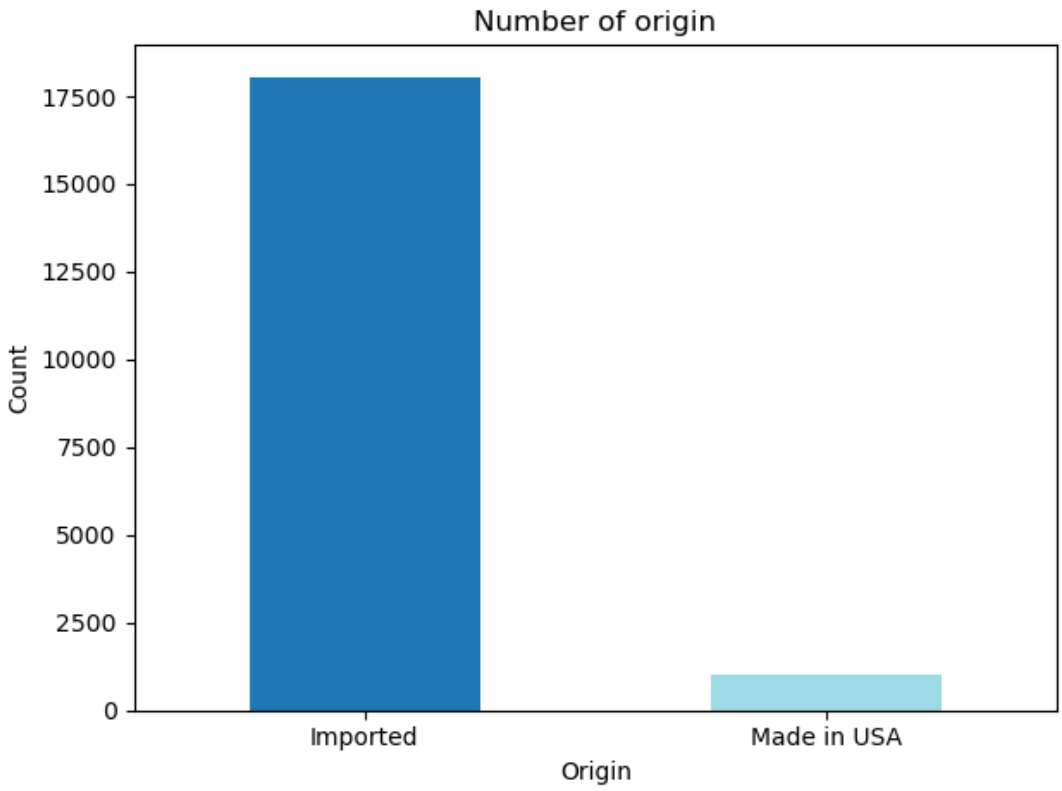


Figure 21: Number of origin

The distribution of products based on their origin highlights a significant disparity, with the vast majority being **Imported**, accounting for over 17,500 entries. In contrast, products labeled as **Made in USA** form a small fraction of the dataset, with only a few thousand entries. This imbalance suggests a strong reliance on imported goods in the dataset, which could reflect broader market trends favoring global sourcing over domestic manufacturing.

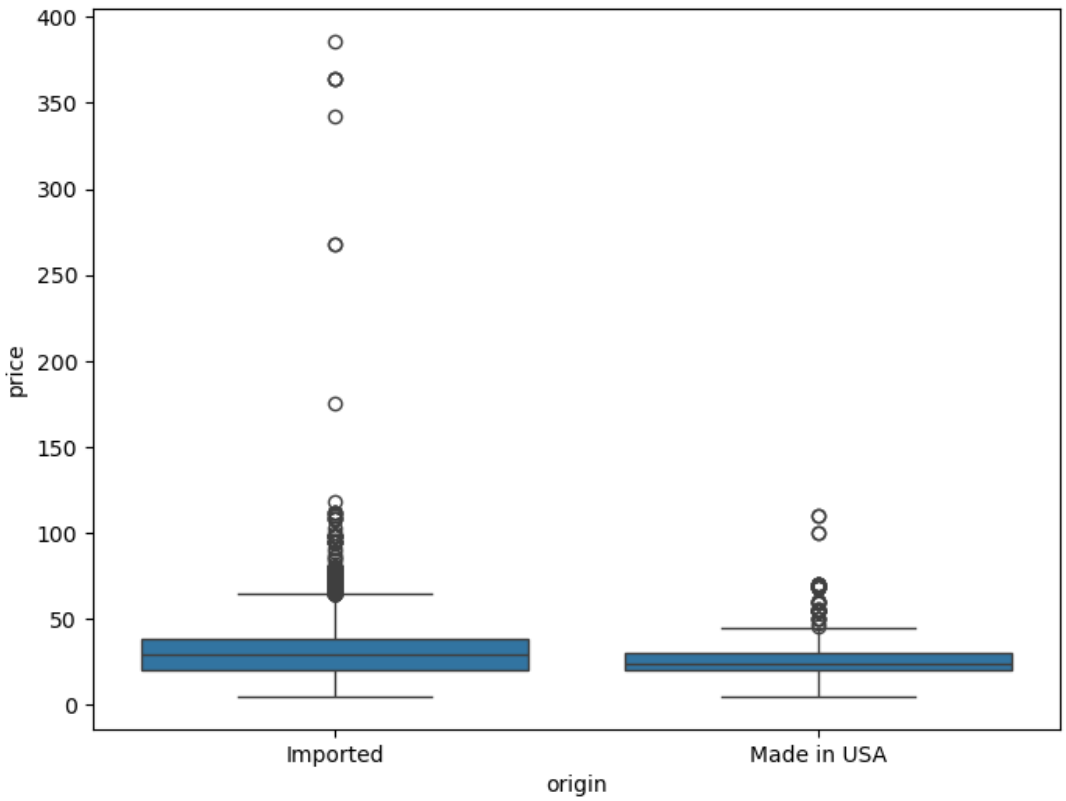


Figure 22: Origin Boxplots

Products based on their origin, **Imported** and **Made in USA**, demonstrate similar median prices, typically below 50 dollars. A notable difference is the variability, as **Imported** products have a wider price range and more outliers, with some prices exceeding 350 dollars. On the other hand, **Made in USA** products exhibit a more compact distribution with fewer extreme values. This suggests that imported items cater to a broader market, including premium offerings, while domestic products are priced more consistently.

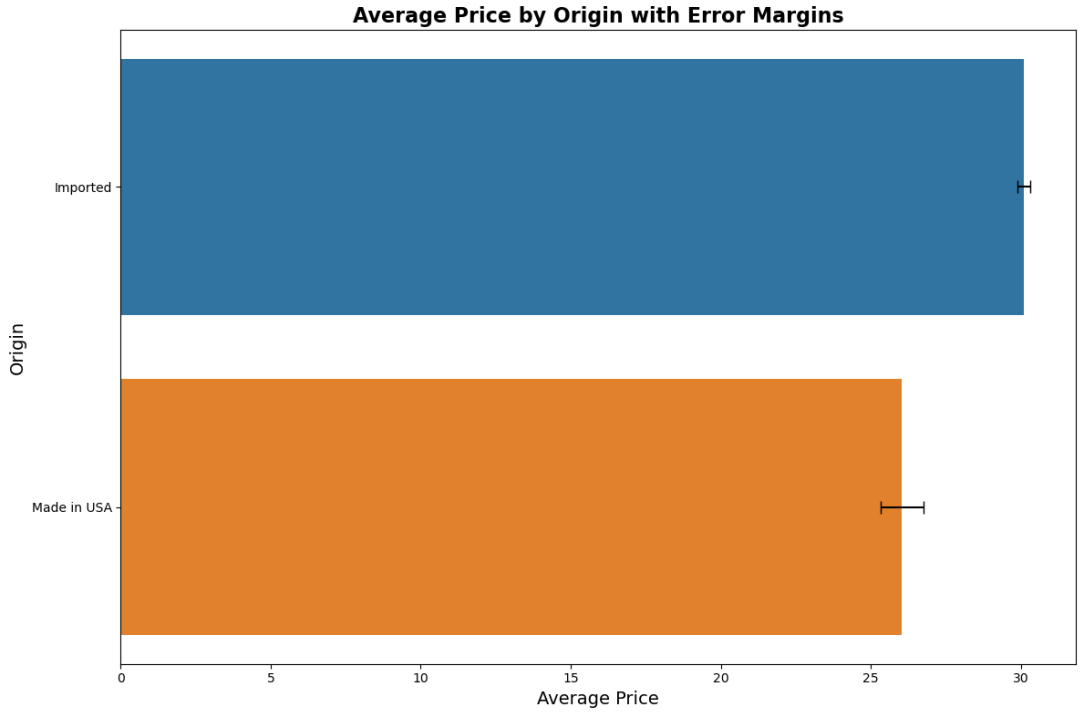


Figure 23: Average Origin by Product Color with Error Margins

Products of **Imported** origin have a slightly higher average price compared to those **Made in USA**, with the averages being close to 30 dollars and 25 dollars, respectively. The error margins for both categories are narrow, indicating minimal variability in average pricing. This consistency suggests that while imported goods are marginally more expensive on average, pricing strategies for both origins remain fairly stable and uniform.

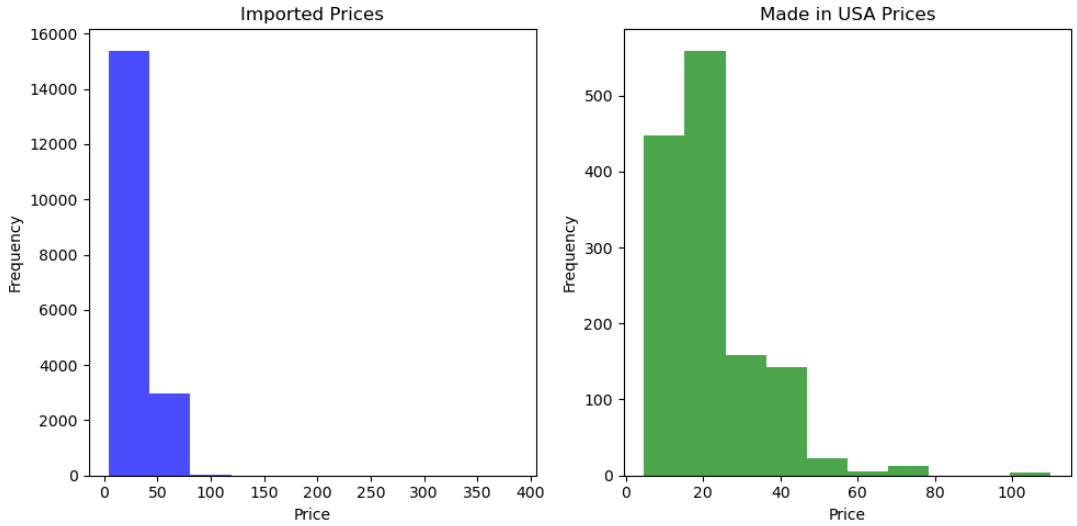


Figure 24: Distribution of Price for Each Origin

The price distributions for **Imported** and **Made in USA** products reveal distinct patterns. **Imported** products are heavily concentrated below 50 dollars, with an overwhelming majority priced in this range, and very few entries exceeding 100 dollars. In contrast, **Made in USA** products display a wider spread, with a noticeable number of

items priced between 20 and 60 dollars, and a smaller frequency of high-priced items reaching up to 100 dollars. These distributions suggest that imported products cater predominantly to lower price ranges, while domestic products exhibit greater variability and include a broader range of mid-priced offerings.

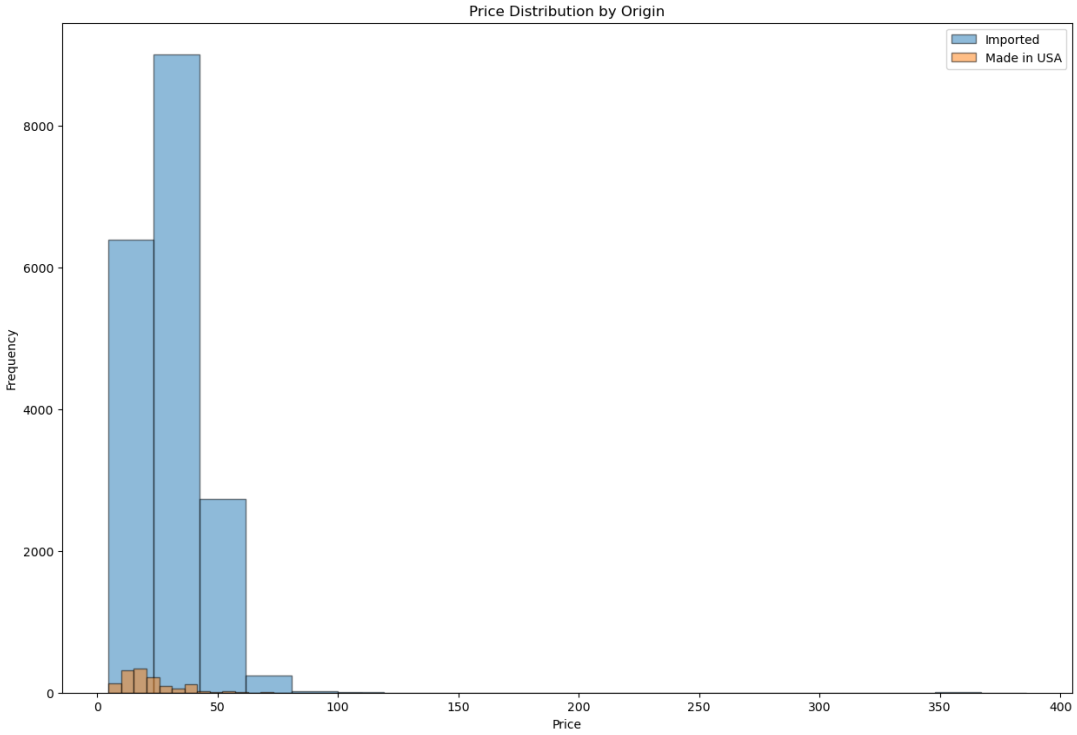


Figure 25: Distribution of Price by Origin

The price distribution for **Imported** and **Made in USA** products highlights a striking difference in frequency. **Imported** items dominate the dataset, with the majority priced below 50 dollars and only a few exceeding this range. In comparison, **Made in USA** products are far fewer in number but show a broader spread within the lower price range. This pattern suggests a heavy reliance on imported goods for low-cost offerings, while domestically produced items cater to a more limited but potentially varied market.

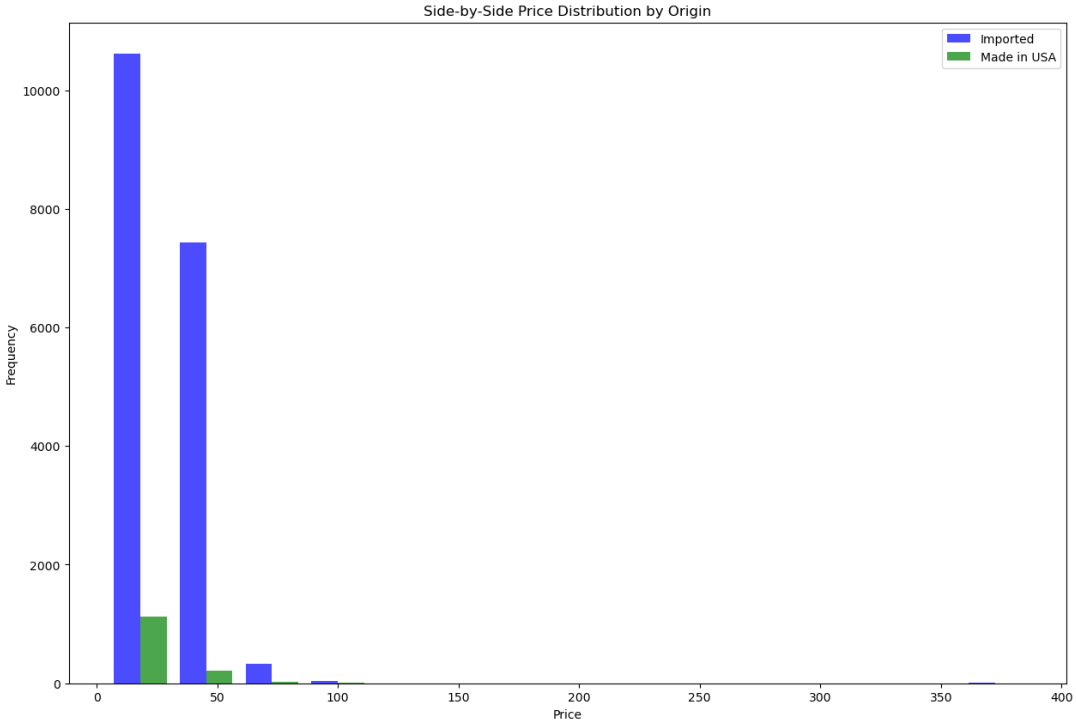


Figure 26: Side-by-Side Price Distribution by Origin

A clear distinction exists between **Imported** and **Made in USA** products in terms of frequency and pricing patterns. **Imported** items are significantly more frequent, with the majority priced below 50 dollars, and only a small fraction exceeding this range. On the other hand, **Made in USA** products, though fewer in number, exhibit a slightly wider spread in the lower price range. This disparity reflects a heavy reliance on imported goods for low-cost offerings, while domestic products seem to target a more niche market.

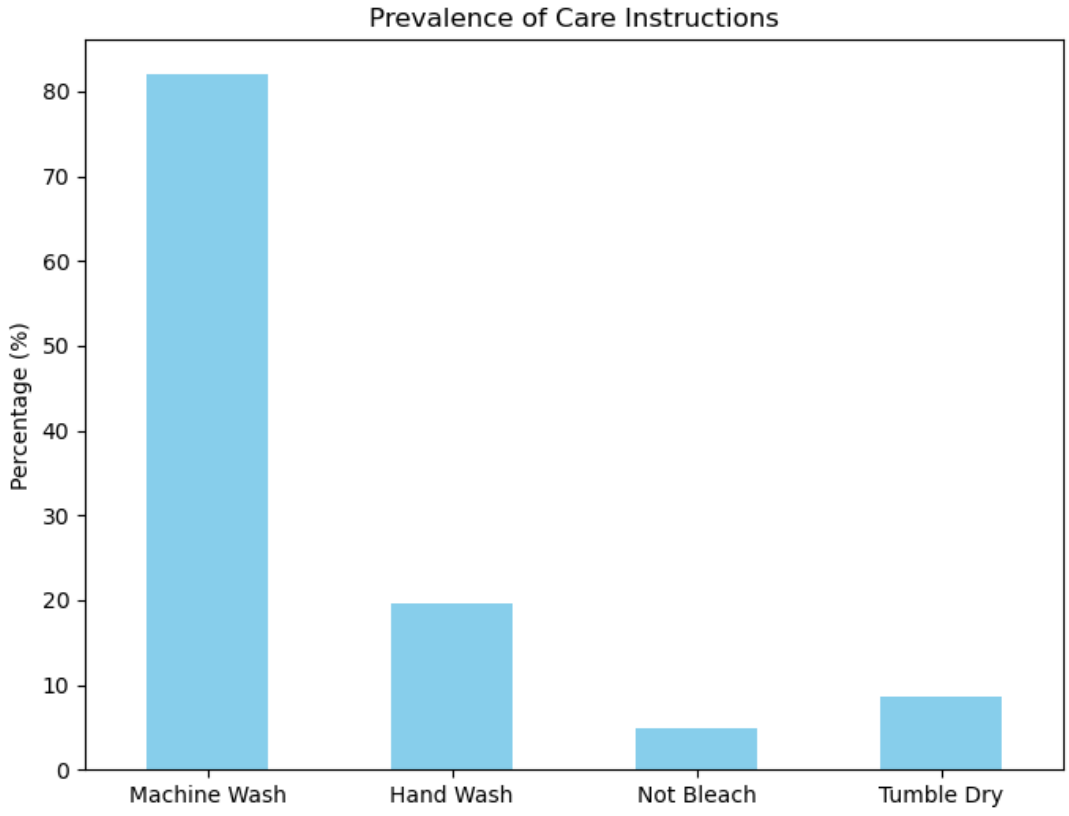


Figure 27: Prevalence of Care Instructions

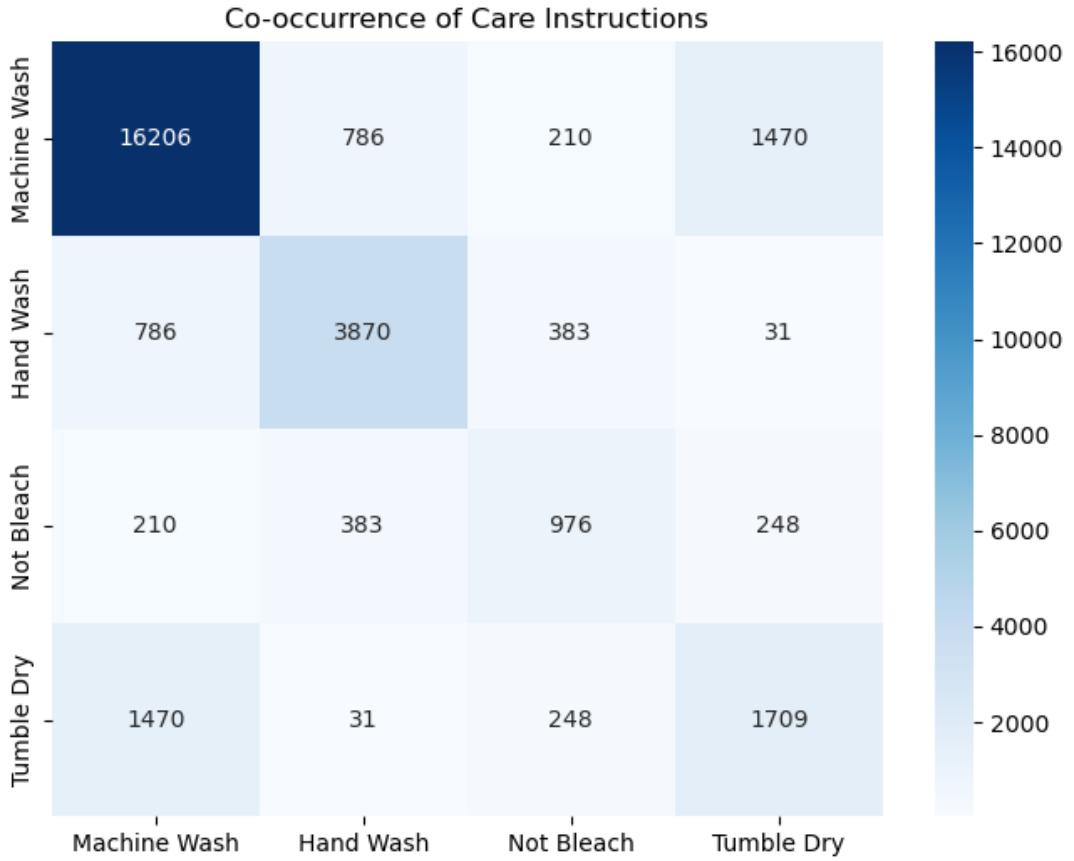


Figure 28: Co-occurrence of Care Instructions

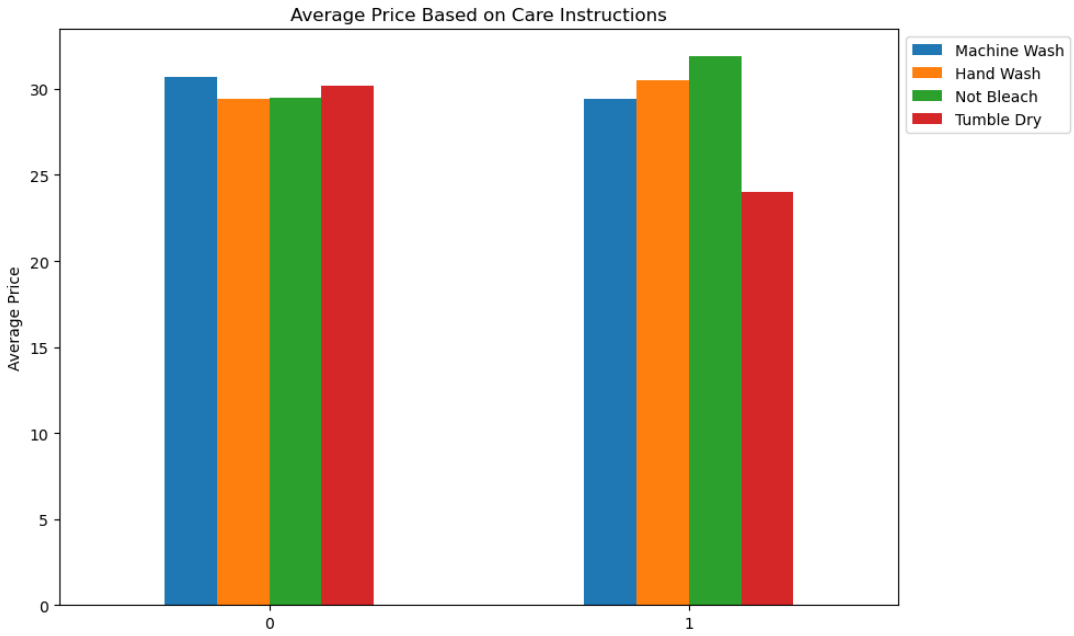


Figure 29: Average Price Based on Care Instructions

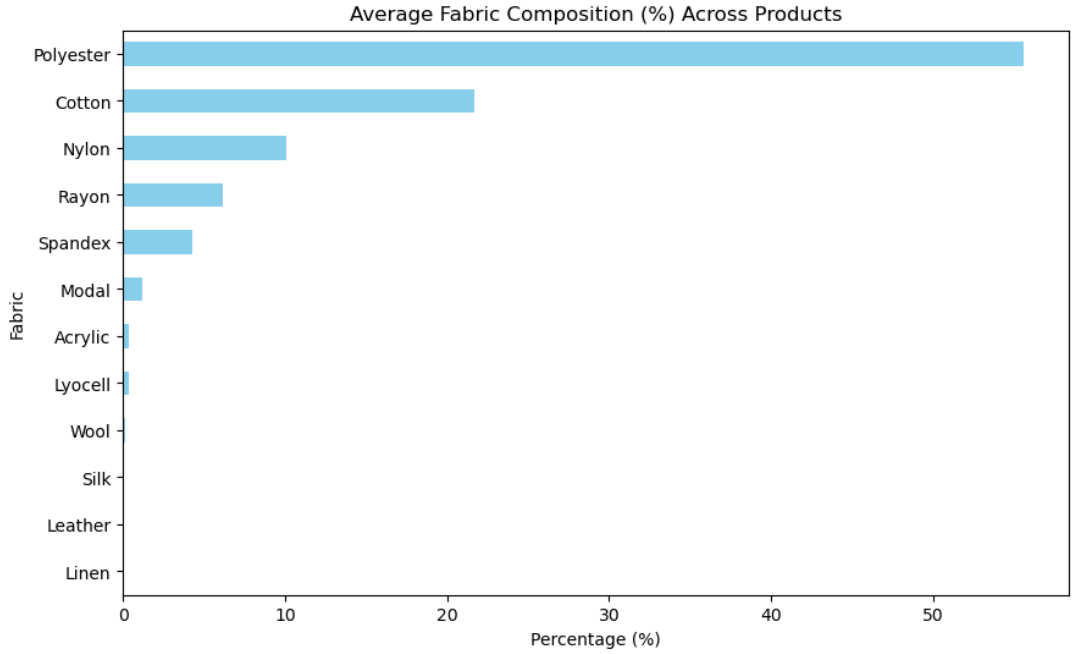


Figure 30: Average Fabric Composition (percent) Across Products

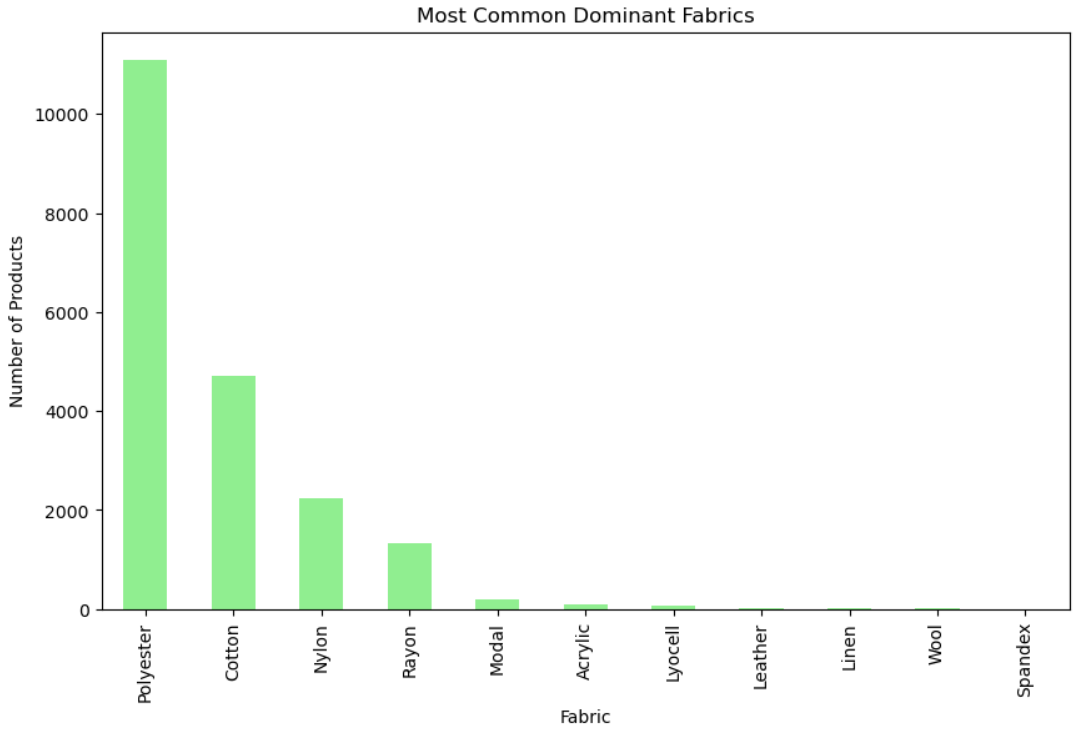


Figure 31: Most Common Dominant Fabrics

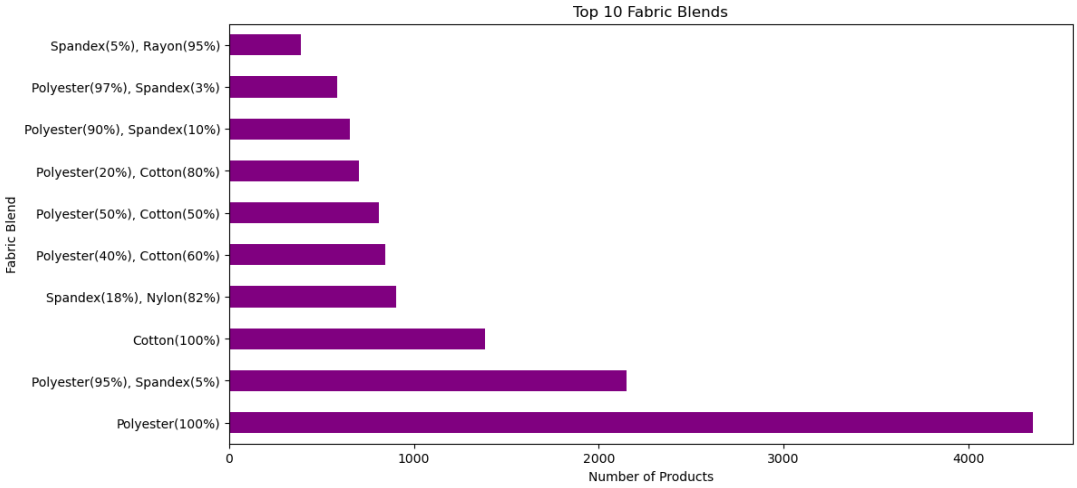


Figure 32: Top 10 Fabric Blends

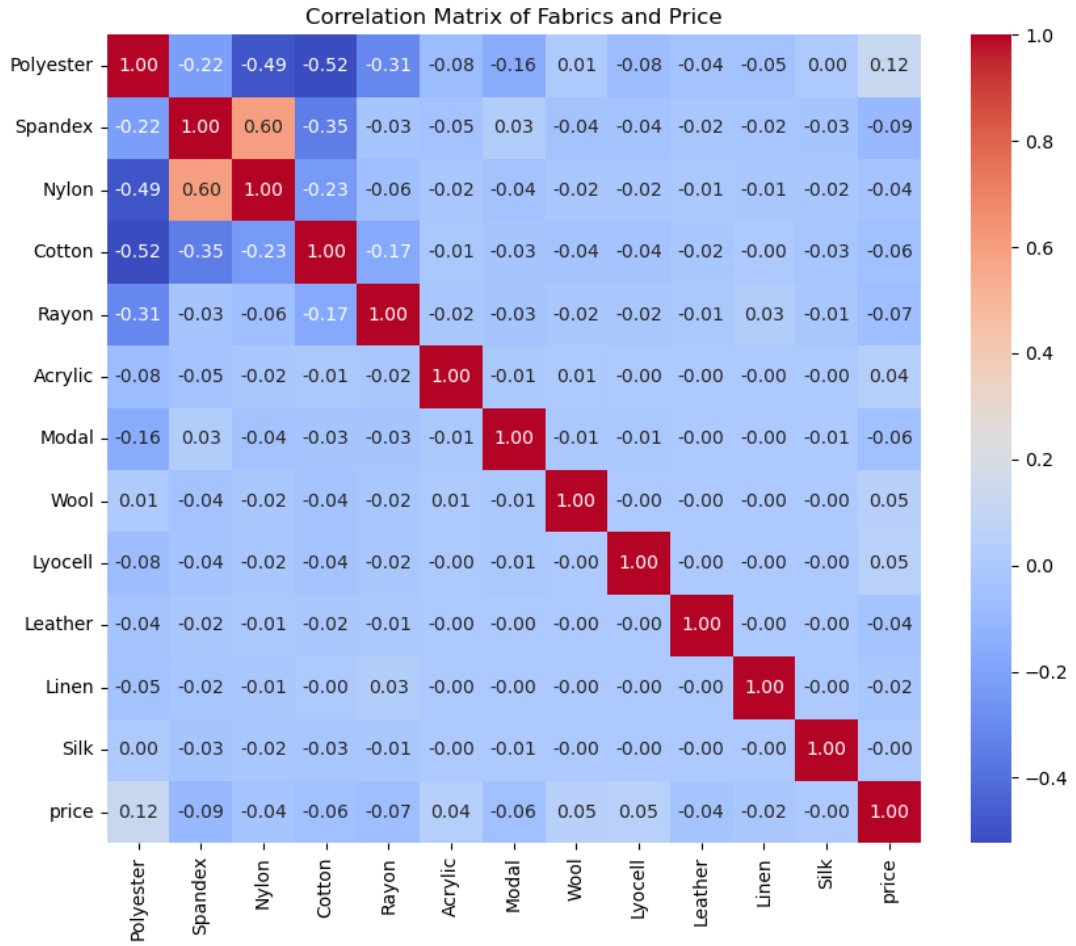


Figure 33: Correlation Matrix of Fabrics and Price

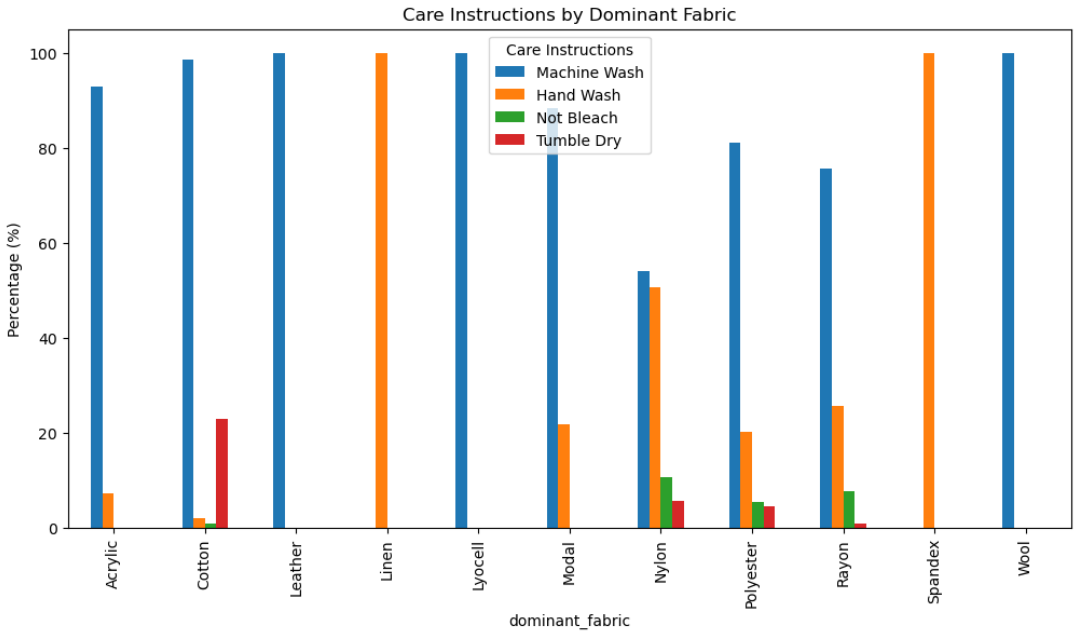


Figure 34: Care Instructions by Dominant Fabric

5 Modeling

5.1 Decision Tree

Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It works by splitting the dataset into subsets based on the feature that best separates the data at each node. The tree structure consists of internal nodes representing feature tests and leaf nodes representing the final prediction. In the context of this project, decision trees can model the relationship between various features (such as brand, material, and size) and the target variable (price) by creating a sequence of decision rules. These models are easy to interpret, but they can be prone to overfitting if not properly tuned or pruned.

5.2 Random Forest

Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to improve accuracy and robustness. Each tree is trained on a random subset of the data, using a technique known as bootstrapping, and at each node, a random subset of features is considered for splitting. This randomization helps to reduce overfitting and increases the model's generalizability. Random Forest can capture complex relationships between features, such as brand, fabric, and size, by averaging the predictions from many individual decision trees. As a result, it tends to outperform a single decision tree, providing more accurate and stable predictions.

5.3 XGBoost

XGBoost is an advanced implementation of gradient boosting, used for both classification and regression tasks. It builds multiple decision trees sequentially, where each tree attempts to correct the errors made by the previous ones. XGBoost includes regularization (L1 and L2) to prevent overfitting and improve model generalization. It uses an efficient algorithm to compute gradients, leading to fast training times and high accuracy. For clothes price prediction, XGBoost can effectively capture non-linear relationships between features such as material, brand, and color, making it a popular choice for high-performance predictive modeling. Its ability to handle missing data and support parallel processing further enhances its suitability for large, complex datasets.

5.4 FCNNs

Fully Connected Neural Networks (FCNNs) are a fundamental type of artificial neural network, commonly used for both classification and regression tasks. They consist of multiple layers of interconnected neurons, where each neuron in a layer is connected to every neuron in the subsequent layer. FCNNs excel at capturing complex relationships in data through their non-linear activation functions, making them highly versatile. Regularization techniques such as dropout and weight decay can be applied to prevent overfitting and improve generalization. For clothes price prediction, FCNNs can model intricate interactions between features like material, brand, and design, providing robust predictive performance. Their scalability and ability to learn from large datasets make them a strong choice for problems requiring high accuracy and adaptability.

6 Experiments and Results

6.1 Hyperparameter Tuning

6.1.1 Decision Tree Model

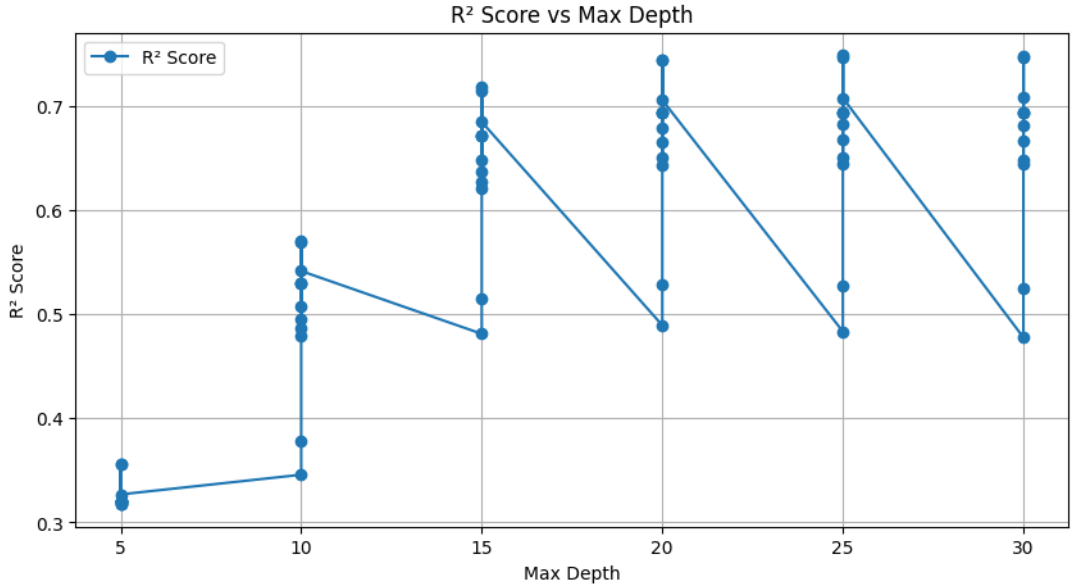


Figure 35: Descision Tree R2 Score

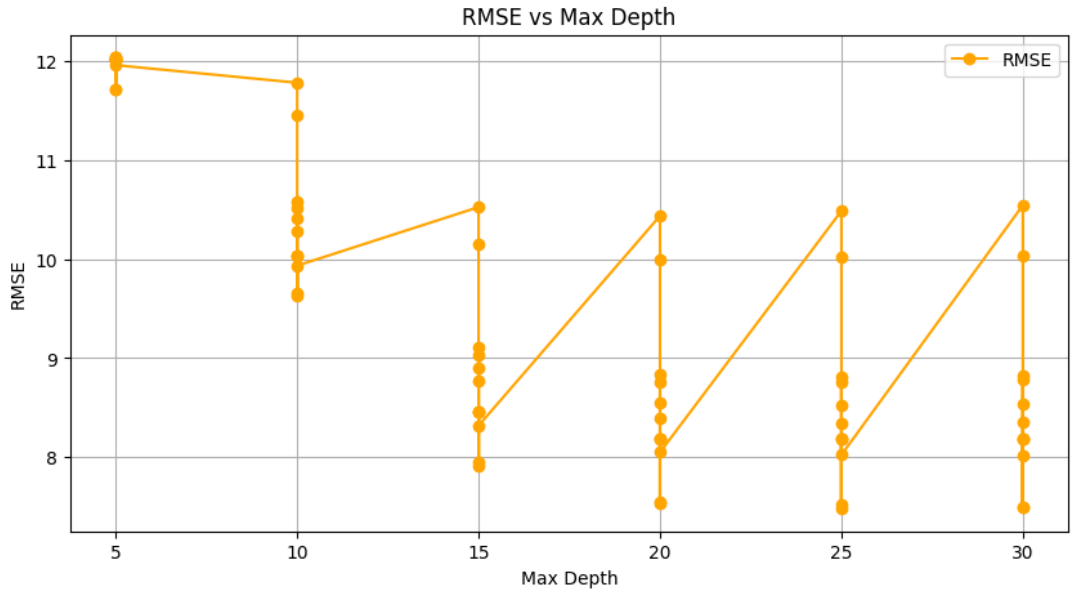


Figure 36: Descision Tree RMSE Score

The analysis revealed that RMSE decreases sharply at smaller depths and stabilizes as the depth increases, while R^2 scores improve significantly up to a peak before becoming inconsistent at higher depths. The optimal `max_depth` was chosen as 15, as it achieves a balance between capturing sufficient data complexity and maintaining low errors.

- `max_depth` = 15

6.1.2 Random Forest Model

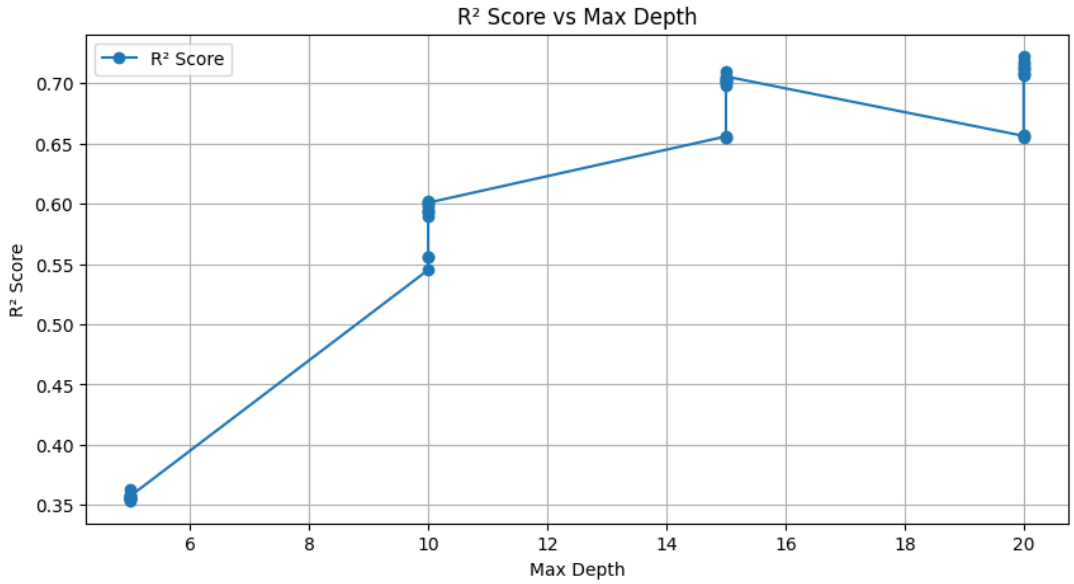


Figure 37: Random Forest R2 Score

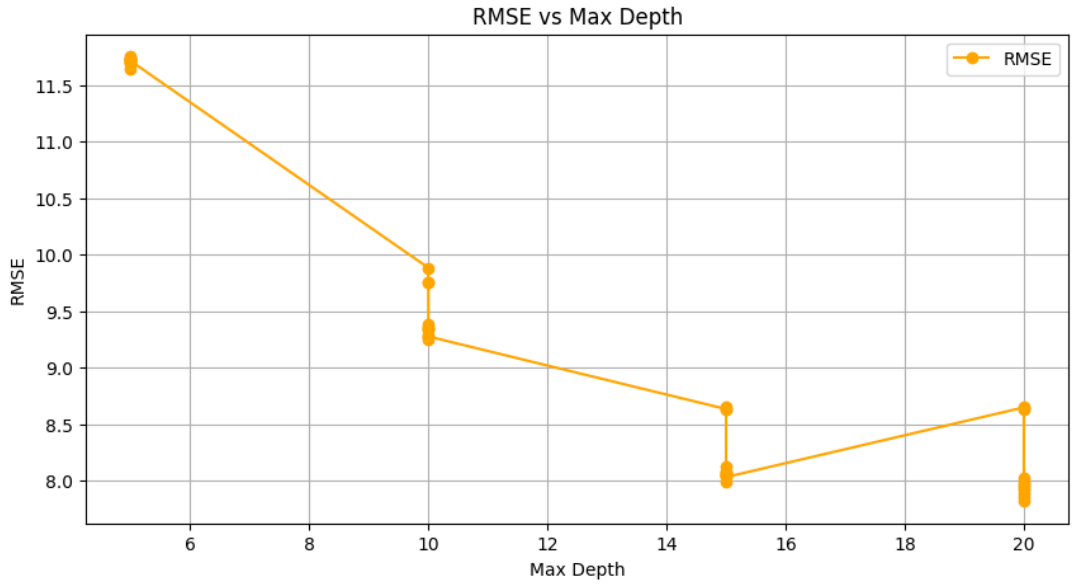


Figure 38: Random Forest RMSE Score

For the Random Forest model used to predict clothes prices, the hyperparameter `max_depth` was evaluated to determine its optimal value. The graphs display the model's performance metrics as `max_depth` increases: the R-squared (R^2) score rises steadily, peaking at 20, while the RMSE decreases initially but stabilizes near its lowest value at a similar depth. These trends highlight the trade-off between model complexity and performance, with deeper trees improving accuracy up to a point. Based on these observations, the optimal `max_depth` was selected as 20, as it allows the model to effectively capture complexity for precise predictions while keeping the error low. So the hyperparameter is:

- `max_depth` = 20

6.1.3 XGBoost

For the XGBoost model, a comprehensive random search was conducted to optimize the hyperparameters. The parameters considered include:

- `colsample_bytree` : (0.6 - 1.0)
- `gamma` : (0 - 0.2)
- `learning_rate` : (0.05 - 0.15)
- `max_depth` : (3 - 7)
- `min_child_weight` : (1 - 4)
- `n_estimators` : (200 - 500)
- `reg_alpha` : (0 - 0.005)
- `reg_lambda` : (0 - 0.005)
- `subsample` : (0.8 - 1.0)

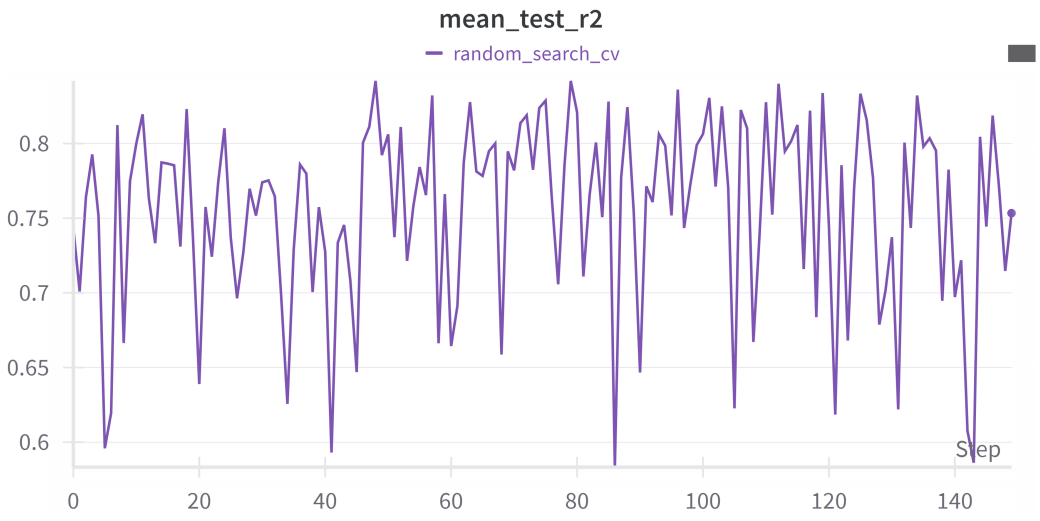


Figure 39: XGBoost R2 Score

There are fluctuations in mean test R-squared scores across various hyperparameter combinations, with a noticeable tendency to cluster between 0.75 and 0.85. This indicates the sensitivity of the model's performance to hyperparameter selection. As a result, the best hyperparameters identified are:

- `colsample_bytree` = 0.829
- `gamma` = 0.046
- `learning_rate` = 0.137

- `max_depth` = 6
- `min_child_weight` = 2
- `n_estimators` = 475
- `reg_alpha` = 0.00098
- `reg_lambda` = 0.00102
- `subsample` = 0.854

which together enable the model to achieve optimal predictive performance.

6.1.4 FCNNs Model

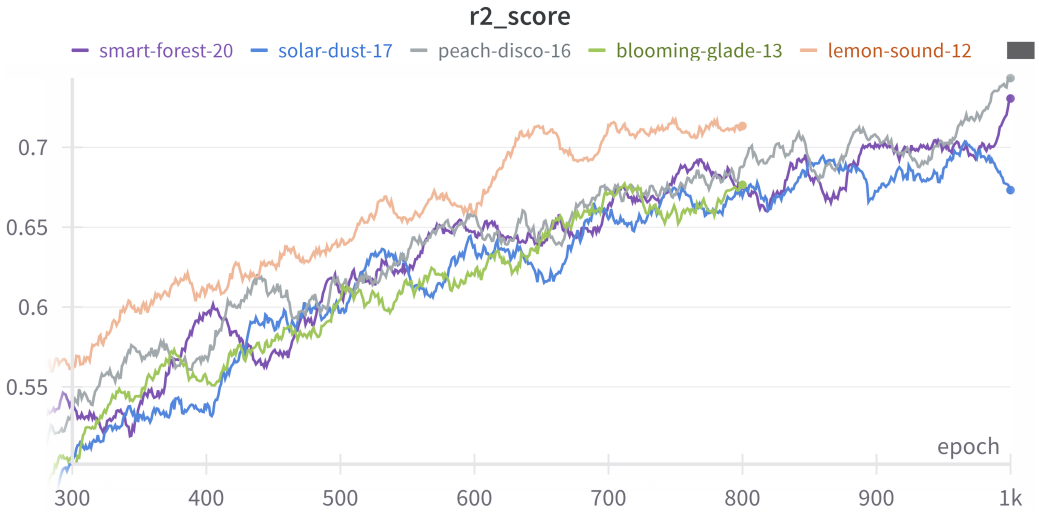


Figure 40: FCNNs R2 Score

The graph depicts the R-squared (R^2) scores of the Fully Connected Neural Network (FCNNs) model over 1000 training epochs across five runs with different initialization seeds. All runs show a steady upward trend, with early variability converging to stable scores around 0.7 or higher, indicating consistent performance improvement. This suggests that the model is robust and benefits from a well-calibrated training process. The chosen hyperparameters are:

- `n_epochs` = 1000
- `batch_size` = 512
- `learning_rate` = 0.0001
- Optimizer: `AdamW`

6.2 Performance Comparison

Model	R^2 **Score**	**MAE**	**RMSE**
Decision Tree	0.590	0.870	9.806
Random Forest	0.600	1.079	9.689
XGBoost	0.817	0.098	0.186
FCNN	0.782	3.891	5.680

Table 2: Performance metrics for models.

7 Conclusion and Future Work

Random Forest demonstrated the best performance with the highest R^2 score. Future work includes testing on external datasets and exploring real-world validation scenarios.

References

1. Hilt, Donald E.; Seegrist, Donald W. (1977). *Ridge, a computer program for calculating ridge regression estimates.*

2. Scikit-learn documentation: `https://scikit-learn.org/`

3. Python Scrapy Playbook: `https://scrapeops.io/python-scrapy-playbook/`