Customer Preference Analytics

Anant Patel - 0866771

The global retail industry has undergone significant transformations in the past decade, particularly in how consumers engage with fashion products. A key factor in this shift has been the growing dominance of online sales, which now account for 20% of total global retail sales. According to Statista (2023), online sales in the fashion market were valued at $759.5 billion in 2022 and are expected to reach $1.2 trillion by 2027. This surge reflects a broader shift in consumer behavior, as more shoppers turn to digital platforms for their purchases. Fashion, a key component of this transformation, has seen rapid changes in how trends are set and how consumers interact with these trends. The fashion industry, traditionally dominated by physical stores and in-person shopping experiences, is increasingly shaped by online influencers, celebrities, and social media platforms that contribute to fast-moving trends.

New clothing trends are constantly emerging and are widely disseminated through digital platforms. Social media, in particular, has played a pivotal role in reshaping the fashion industry. Influencers and celebrities often lead the way in setting new styles that are quickly adopted by the broader public. Social media platforms such as Instagram, TikTok, and YouTube provide users with immediate exposure to the latest trends, and advertisements targeted at specific consumer segments further amplify this effect (Giri, Thomassey, and Zeng, 2019). This transformation has made fashion more accessible to a global audience, allowing consumers to explore and purchase clothing directly from digital platforms. Fashion giants like H&M and Zara have capitalized on this trend, successfully launching new fashion lines that have gained widespread popularity due to their strong online presence and their ability to influence consumer behavior (@cleofasbrand). These shifts have not only changed how fashion is consumed but have also prompted new ways for customers to engage with the purchase of clothing and accessories.

In addition to the influence of online platforms and digital influencers, the global COVID-19 pandemic has had a significant impact on consumer shopping behavior. As physical stores were forced to close during lockdowns, many consumers turned to e-commerce platforms for their shopping needs. This shift to online shopping was particularly noticeable in the fashion industry, which was one of the hardest-hit sectors during the pandemic (Cleofas et al., n.d.). According to research by Wibowo (2024), the Santoon fashion brand in Indonesia, despite being a market leader, faced challenges with e-commerce adoption. This example underscores the difficulties traditional retailers faced in adapting to the rapid shift toward online shopping. The pandemic highlighted the need for businesses to rethink their strategies and invest more in their online operations.

Understanding these shifts in consumer behavior is crucial for businesses, especially new online fashion retailers looking to succeed in an increasingly competitive market. By analyzing historical data, businesses can gain insights into customer preferences, helping them design strategies that align with current trends and consumer needs. Data-driven strategies are particularly important in the fashion industry, as customer preferences can change rapidly. As Giri (2019) notes, personalized services driven by data analysis are essential for retaining customers and expanding a company's customer base. By leveraging data analytics, businesses can offer tailored recommendations, promotions, and products that meet the specific needs of their customers.

Factors influencing customer preferences in the fashion industry vary between offline and online retail. A key difference is the level of product information available to consumers. In traditional brick-and-mortar stores, customers can physically inspect products, feel the fabric, and try on clothing before making a purchase. However, in the online shopping environment, detailed product information is essential for helping customers make informed purchasing decisions. Wibowo (2024) found that product quality and size are two of the most important factors influencing customer preferences for fashion products. In e-commerce, the availability of detailed product descriptions, high-quality images, and size guides can significantly affect a customer's likelihood of making a purchase. This finding highlights the importance of providing comprehensive product information on online platforms to meet customer expectations.

Customer reviews also play a crucial role in shaping online purchasing decisions. In the absence of the tactile experience of shopping in physical stores, many consumers rely on reviews from other buyers to assess the quality of products. Wibowo (2024) emphasized the importance of customer reviews in enhancing the online shopping experience. Positive reviews can build trust in a product, while negative reviews can prompt consumers to reconsider their purchase. Moreover, online-exclusive offers such as free shipping, limited-time discounts, and special promotions are known to influence customer preferences. These incentives help create a sense of urgency and encourage customers to complete their purchases. Secure payment options, flexible return policies, and efficient delivery services are also factors that can enhance the overall customer experience and drive online purchases.

Brand recognition is another key factor influencing customer preferences in the fashion industry. According to Cleofas et al. (n.d.), brand identity is the most important factor for customers, contributing 43% to their purchasing decisions. Other factors, such as the functionality of the clothing and the place of purchase, are also important but rank lower in importance compared to brand recognition. This finding suggests that consumers are often willing to pay a premium for products from well-established and trusted brands. Brand loyalty is especially strong in the fashion industry, where consumers often make repeat purchases from brands they trust. Kod (n.d.) also found that factors such as brand name, price, and the convenience of shopping from home were strong motivators for online shoppers.

Age also plays a role in shaping consumer purchasing decisions, although research on this topic has yielded mixed results. In Poland, for example, Kod (n.d.) found no significant differences in purchasing decisions across age groups, even though the Mann-Whitney U test showed notable results for different age groups. In contrast, Glińska and Tomaszewska (2017) found that age does affect online shopping behavior. Younger consumers (under 25) are more likely to view online shopping as a form of entertainment, enjoying the experience of browsing and participating in loyalty programs. Older consumers (over 25), on the other hand, tend to value the convenience of 24/7 availability, the ability to find detailed product information, and the opportunity to purchase rare or unique items. Delimarta and Rahadi (2021) also highlighted that factors such as brand image, price, quality, and design are positively related to customers' willingness to purchase fashion products. This suggests that businesses need to tailor their marketing strategies to the specific needs and preferences of different age groups to maximize their appeal.

The role of data analytics in understanding and responding to consumer preferences is becoming increasingly important for businesses in the fashion industry. Data analysts play a crucial role in uncovering hidden patterns and trends that can guide business decisions. Customer analytics, as defined by Giri, Thomassey, and Zeng (2019), involves analyzing purchase data to uncover insights that can be used to improve customer service and product offerings. Key data sources include website traffic, purchase history, and customer feedback. Website traffic data, such as page views, click-through rates, and time spent on the site, can provide valuable insights into customer behavior and preferences. Similarly, purchase history can reveal patterns related to average order value, preferred brands, and product categories. Customer feedback, gathered through surveys, reviews, and ratings, can help businesses assess product quality, customer service, and overall satisfaction.

Clustering techniques such as K-means clustering can be used to segment customers based on shared characteristics. Sulianta, Ulfah, and Amalia (2024) used this technique to group customers into four clusters based on price sensitivity, brand loyalty, and product preferences. By analyzing these clusters, businesses can develop targeted marketing strategies and personalized offers that cater to the specific needs of each customer segment. Okofu et al. (2024) employed sentiment analysis to assess customer behavior and derive insights related to factors such as product usefulness, ease of use, and purchase intention. This approach allows businesses to identify customer pain points and improve the overall shopping experience.

The benefits of data analytics for fashion retailers are numerous. By understanding age-related differences in consumer behavior, businesses can tailor their marketing strategies to specific customer segments. For example, younger consumers may be more responsive to promotions, loyalty programs, and influencer marketing, while older consumers may prioritize product information, quality, and convenience. Glińska and Tomaszewska (2017) concluded that age-dependent differences in consumer behavior could be used to target different age groups with personalized offers, content, and discounts. In addition, data analysis can help businesses improve product assortment planning, inventory management, and customer service, leading to higher levels of customer satisfaction and retention.

Several methods are commonly used in customer preference analysis. These methods include:

1. \*\*Thematic Analysis\*\*: This method is used to identify recurring patterns and generate themes within data (Wibowo, 2024). It helps researchers understand customer preferences by examining the frequency and context of certain behaviors or attitudes.

2. \*\*Triangulation\*\*: This method involves validating research findings by comparing data from multiple sources. For instance, Wibowo (2024) used triangulation to compare data from customer interviews, company observations, and existing research on consumer decision-making in the fashion industry.

3. \*\*Conjoint Analysis\*\*: This technique helps businesses identify the relative importance of different product attributes in customer purchasing decisions. Cleofas et al. (n.d.) and Suzianti, Faradilla, and Anjani (2015) used conjoint analysis to understand customer preferences for various fashion attributes.

4. \*\*CRISP-DM (Cross-Industry Standard Process for Data Mining)\*\*: This methodology provides a structured approach to data mining, involving stages such as data collection, cleaning, analysis, and interpretation. These phases produce specific outputs that guide decision-making (Sulianta, Ulfah, and Amalia, 2024).

5. \*\*K-means Clustering\*\*: This unsupervised machine learning technique groups data points into clusters based on their similarities. It is widely used in customer segmentation to identify groups of customers with shared characteristics, such as price sensitivity or brand preference.

These are the common methods used by the researchers to conduct Customer Preference Analysis.

# Data description

These aspects or variables must be present in the dataset because the study's goal is to determine client preferences based on behavioral and demographic characteristics including age, gender, recency, preferred delivery mode, and transaction amount. The dataset that will be used in this study may be found on Bhadramohit (2024). This dataset shows US consumer patterns and was gathered from a number of retail and e-commerce platforms. Numerous analytics, including customer segmentation and demographic analysis, can be performed using this dataset. In addition, a range of visualizations can be derived from this dataset because of its diversity. The dataset also makes use of a number of GitHub repositories that gathered demographic information (season, age, gender, and purchase amount). The dataset contains a total of 3900 rows and 22 columns. The description of the Columns is listed in the below given table:

Table for Column Description

| Feature Name | Description |
| --- | --- |
| Customer.ID | Unique identifier for each customer. |
| ———————————– | ———————————————————————— |
| Age | Age of the customer. |
| ———————————– | ———————————————————————— |
| Gender | Gender of the customer. |
| ———————————– | ———————————————————————— |
| Item.Purchased | Item that was purchased by the customer. |
| ———————————– | ———————————————————————— |
| Category | Category of the purchased item. |
| ———————————– | ———————————————————————— |
| Purchase.Amount..USD. | Purchase amount in USD. |
| ———————————– | ———————————————————————— |
| Location | Geographical location of the customer. |
| ———————————– | ———————————————————————— |
| Size | Size of the purchased item. |
| ———————————– | ———————————————————————— |
| Color | Color of the purchased item. |
| ———————————– | ———————————————————————— |
| Season | Season during which the item was purchased. |
| ———————————– | ———————————————————————— |
| Review.Rating | Customer’s rating for the item. |
| ———————————– | ———————————————————————— |
| Subscription.Status | Whether the customer has a subscription. |
| ———————————– | ———————————————————————— |
| Payment.Method | Payment method used for the purchase. |

| Shipping.Type | Type of shipping used for the purchase. |
| --- | --- |
| Discount.Applied | Whether a discount was applied during the purchase. |
| ———————————– | ———————————————————————— |
| Promo.Code.Used | Whether a promo code was used for the purchase. |
| ———————————– | ———————————————————————— |
| Previous.Purchases | Number of previous purchases made by the customer. |
| ———————————– | ———————————————————————— |
| Preferred.Payment.Method | Customer’s preferred method of payment. |
| ———————————– | ———————————————————————— |
| Frequency.of.Purchases | Frequency with which the customer makes purchases. |

* The categorical variables are Gender ( Male, Female), Item.Purchased (Dress, Backpack,…), Category (Clothing, Footwear, Outerwear, Accessories), Location (New Jersey, Ohio,… ), Size (S, M, L, XL), Color (Yellow, White, …), Season (Winter, Spring, Summer, Fall), Payment.Method (Credit, Debit, Venmo, Bank Transfer, Cash, Paypal), Preferred.Payment.Method (Credit, Debit, Venmo, Bank Transfer, Cash, Paypal), Shipping.Type ( Express, Free shipping, Next Day Air, Standard, 2-Day Shipping, Store Pickup)
* Subscription.Status, Discount Applied, and Promo.Code.Used are binary categorical variable with Yes or No as possible values.
* Age, Purchase.Amount..USD., Review.Rating, Previous.Purchases, Frequency.of.Purchases are a continuous variable.

Every feature, including age, gender, frequency of purchases, fashion product size, whether or not a promo coupon was used, shipping type, and many more, is required by the study. Additionally, features like payment method, product category, and subscription can be used to further determine the undiscovered relationship. These connections can be used to draw attention to certain novel elements influencing consumer choices. These insights can also be used by businesses to strategically plan their firm for overall development. This data set can be further segmented according to behavioral and demographic characteristics. Based on analysis, demographic characteristics can be utilized to target customers by age and gender, for example, offering different deals for men and women or exclusive offers for kids. Purchase amount, frequency of purchases, and preferred payment methods are behavioral attributes that can be used to group both new and devoted consumers and then make offers based on those clusters. In addition, a Market Basket Analysis can be performed depending on the category and item purchased to identify products that are sold together, like socks and shoes or t-shirts and slacks.

# Methods

## Various researchers have employed different methods for customer preference analysis, including Thematic Analysis (Wibowo, 2024), Triangulation (Wibowo, 2024), Conjoint Analysis (Cleofas et al., n.d.; Suzianti, Faradilla, & Anjani, 2015), CRISP-DM (Sulianta, Ulfah, & Amalia, 2024), K-Means Clustering (Sulianta, Ulfah, & Amalia, 2024), and the Elbow Method (Sulianta, Ulfah, & Amalia, 2024).

## For this study, I will use the CRISP-DM methodology, also known as the Cross-Industry Standard Process for Data Mining, which was applied by Sulianta, Ulfah, and Amalia (2024) in their research on identifying consumer preferences in the fashion industry through K-Means Clustering. The CRISP-DM method consists of four key steps:

## Data Collection: In this phase, data is gathered from various retail and e-commerce platforms in the United States, as well as several GitHub repositories. This data is then compiled and made available by Bhadramohit (2024) on Kaggle.

## Data Cleaning: This step involves preparing the dataset for analysis by removing missing values (NA) and performing basic feature engineering to ensure the data is clean and usable.

## Data Analysis: After cleaning, the preprocessed data is explored and analyzed to extract valuable insights that can inform the study’s objectives.

## Interpretation: The insights gained from the analysis are then interpreted to uncover trends and patterns, helping to answer the research questions effectively.

## 1. Data Collection

The data is collected by bhadramohit (2024) which can be downloaded and loaded into Rstudio using following code.

df <- read.csv("shopping\_trends.csv")  
df\_2 <- df

The original dataframe is copied into another variable, df\_2. This variable will be used in further analysis.

## 2. Data Cleaning

Firstly, the categorical features have to be converted into factors for analysis and better structuring of data. Therefore, looping through each column and comparing their class with character which then is converted into factor column can be done in following way. Moreover, the Size column has a natural order of S < M < L < XL, so Size feature can be factorized using ordered = TRUE.

# Loop to convert character column in factor column  
for (i in names(df\_2)){  
 if(class(df\_2[[i]]) == "character"){  
 df\_2[[i]] <- factor(df\_2[[i]])  
   
 }  
}  
  
# Factorizing Size column for ordering the levels  
df\_2$Size <- factor(df\_2$Size, levels=c("S","M","L","XL"), ordered = TRUE)  
  
str(df\_2)

Secondly, the dataset should be checked for any NA values. This can be done by finding the column sum from each variable after applying is.na().

# Check null values  
colSums(is.na(df\_2))

The age is continuous variable, to check if age have any effect in customer preference it needs to be binned into two categories: Under 25 ( Age <= 25 ), and Above 25 ( Age > 25 ) (Glińska and Tomaszewska (2017)).

# Binning the age in to two categories  
df\_2$Age.bin <- ifelse(df\_2$Age <= 25,"Under 25","Above 25")  
  
# Factorizing the column  
df\_2$Age.bin <- factor(df\_2$Age.bin)  
  
str(df\_2$Age.bin)

## 3. Data Analysis

First of all, the new binned age (Age.bin) needs to be analyzed with other features to find if there is a significant difference in the Age groups.

Frequency of Purchase may vary for individuals under 25 and above 25. There may be a significance that Age has a relationship with Frequency of purchase. To visualize both the groups based on the count of different frequencies of purchase category a horizontal bar chart for each age group can be created.

# 1.1 Frequency of Purchase  
  
# Plot to display the count of Customer under 25 and their Frequency of Purchase  
data.frame(table(df\_2[df\_2$Age.bin == "Under 25", "Frequency.of.Purchases"])) %>%  
 mutate(name = fct\_reorder(Var1, desc(Freq))) %>%  
 ggplot(aes(x = name, y = Freq)) +   
 geom\_bar(stat = "identity") +   
 labs( x = "Frequency of Purchases", y = "Count",) +  
 coord\_flip()  
  
# Plot to display the count of Customer above 25 and their Frequency of Purchase  
data.frame(table(df\_2[df\_2$Age.bin == "Above 25", "Frequency.of.Purchases"])) %>%  
 mutate(name = fct\_reorder(Var1, desc(Freq))) %>%  
 ggplot(aes(x = name, y = Freq)) +   
 geom\_bar(stat = "identity") +   
 labs(x = "Frequency of Purchases", y = "Count") +  
 coord\_flip()

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | (a) Frequency of purchase for Age <= 25 | | |  | | --- | | (b) Frequency of purchase for Age > 25 | |

Figure 1: Plots showing the frequency purchase in different age group

A chi-squared test of independence is conducted to find if there is any significant effect of age on the frequency of purchase.

# Chi square test of independence  
chisq.test(table(df\_2$Age.bin,df\_2$Frequency.of.Purchases))

As mentioned by Glińska and Tomaszewska (2017) younger generation saw online shopping as part of loyalty programs. Promo codes are one of the aspects of loyalty programs where customer can get promo codes based on their frequency of purchase. Therefore, a relationship between age and promo code used can be analyzed using a heatmap with the count in each category and a chi-squared test to prove the effect.

# 1.2 Promo codes  
age\_promo\_table <- data.frame(table(df\_2$Age.bin,df\_2$Promo.Code.Used))   
  
colnames(age\_promo\_table) <- c("Age.bin", "Promo.Code.Used", "Count")  
  
age\_promo\_table %>%  
 ggplot(aes(x = Age.bin, y = Promo.Code.Used, fill = Count)) +  
 geom\_tile() +  
 scale\_fill\_viridis(discrete = FALSE) +  
 labs(title = "Heatmap for Age group and Promo code used",  
 x = "Age group", y = " Promo Code Used", fill = "Count") +  
 theme\_minimal() +   
 geom\_text(aes(label = Count), color = "white", size = 6)

age\_promo\_code <- table(df\_2$Age.bin, df\_2$Promo.Code.Used)  
chisq.test(age\_promo\_code)

Furthermore, Wibowo (2024) concludes that customer reviews play a crucial role. So, to find which age group gives a higher rating a boxplot of the distribution of ratings and a t-test to confirm the effect of age and reviews rating needs to be computed.

# 1.3 Review Rating  
df\_2 %>%  
 ggplot(aes(x=Age.bin, y = Review.Rating, fill = Age.bin)) +  
 geom\_boxplot()

Before performing the t-test the assumptions need to be checked to verify if a parametric test can be conducted or not. After conducting a t-test and getting a significant p-value < 0.05, a Cohen’s d test will be conducted to find the effect size of the relationship.

# T test  
Age.grouping <- group\_by(df\_2,Age.bin)  
  
# Assumptions checking  
# Outlier checking  
identify\_outliers(Age.grouping, Review.Rating)  
# Normality checking  
shapiro\_test(Age.grouping, Review.Rating)  
# Homogeniety of variance  
levene\_test(df\_2, Review.Rating ~ Age.bin)  
  
t\_test(Review.Rating ~ Age.bin, data = df\_2)  
  
cohens\_d(df\_2,Review.Rating ~ Age.bin, var.equal = FALSE)

The promo codes are very crucial during online purchases therefore, gender may have an influence on promo code usage. This can be conducted using a chi-squared test and post-hoc test to find which gender uses promo-code most.

# Dataframe containing Contingency table for gender vs promo-code use  
gender\_promo\_table <- data.frame(table(df\_2$Gender, df\_2$Promo.Code.Used))   
  
# renaming column names of the gender\_promo\_table dataframe  
colnames(gender\_promo\_table) <- c("Gender", "Promo.Code.Used", "Count")  
  
# Heat map to visualize the contingency table  
gender\_promo\_table %>%  
 ggplot(aes(x = Gender, y = Promo.Code.Used, fill = Count)) +  
 geom\_tile() +  
 labs(title = "Heatmap for Gender and Promo Code Used",  
 x = "Gender", y = "Promo Code Used", fill = "Count") +  
 theme\_minimal() +   
 geom\_text(aes(label = Count), color = "white", size = 6)

# Continegency table  
gender\_promo\_code <- table(df\_2$Gender, df\_2$Promo.Code.Used)  
  
# Chi-square test  
chisq.test(gender\_promo\_code)  
  
# Post-hoc test  
chisq.posthoc.test(gender\_promo\_code)

Wibowo (2024) also highlighted that offers like free shipping affects customer preference. Therefore, this difference can be visualized using a horizontal grouped bar chart. In addition, a chi-square test can be performed to identify if there is any significance in relationship between gender and shipping type. Also, a post-hoc test will be performed to find which shipping method is prefered by each gender.

# 2.6 Shipping Type  
gender\_shipping <- data.frame(table(df\_2$Gender, df\_2$Shipping.Type))   
colnames(gender\_shipping) <- c("Gender","Shipping\_Type","Count")  
  
gender\_shipping %>%  
 ggplot(aes(fill = Gender, y = Count, x = Shipping\_Type)) +   
 geom\_bar(position = "dodge", stat = "identity") +  
 coord\_flip()

shipping\_table <- table(df\_2$Gender,df\_2$Shipping.Type)  
chisq.test(shipping\_table)  
chisq.posthoc.test(shipping\_table)  
  
# Free shipping is preferred by Females more than males

# Bibiliography

bhadramohit. 2024. “Customer Shopping Latest Trends Dataset.” <https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset>.

Cleofas, Maria Arielle, Yogi Tri Prasetyo, Ardvin Kester S Ong, and Satria Fadil Persada. n.d. “Brand or Clothing Function? Consumer Preference Analysis on Clothing Apparel Attributes and Design: A Conjoint Analysis Approach.”

Delimarta, Florencia Devina, and Raden Aswin Rahadi. 2021. “Customer Preferences on Sustainable Fashion Purchases: A Conceptual Model.” *International Journal of Entrepreneurship and Management Practices* 4 (13): 78–88.

Giri, Chandadevi, Sebastien Thomassey, and Xianyi Zeng. 2019. “Customer Analytics in Fashion Retail Industry.” In *Functional Textiles and Clothing*, 349–61. Springer.

Glińska, Ewa, and Ewelina Tomaszewska. 2017. “Customer Preferences Related to Shopping Online.” *Annales Universitatis Mariae Curie-Skłodowska, Sectio H–Oeconomia* 51 (2): 87.

Kod, JEL. n.d. “Customer Preferences Related to Shopping Online.” *Chemistry* 1: 56.

Okofu, Sebastina Nkechi, Kizito Eluemunor Anazia, Maureen Ifeanyi Akazue, Margaret Dumebi Okpor, Amanda Enadona Oweimieto, Clive Ebomagune Asuai, Geoffrey Augustine Nwokolo, Arnold Adimabua Ojugo, and Emmanuel Obiajulu Ojei. 2024. “Pilot Study on Consumer Preference, Intentions and Trust on Purchasing-Pattern for Online Virtual Shops.” *Int. J. Adv. Comput. Sci. Appl* 15 (7): 804–11.

Statista. 2023. “Global Online Fashion Market Value from 2017 to 2027.” <https://www.statista.com/statistics/228706/global-online-fashion-market-value/>.

Sulianta, Feri, Khaerani Ulfah, and Endang Amalia. 2024. “Revealing Consumer Preferences in the Fashion Industry Using k-Means Clustering.” *International Journal of Engineering Continuity* 3 (2): 34–53.

Suzianti, Amalia, Nurza Dwi Prisca Faradilla, and Shabila Anjani. 2015. “Customer Preference Analysis on Fashion Online Shops Using the Kano Model and Conjoint Analysis.” *International Journal of Technology* 6 (5): 881–85.

Wibowo, Amira Luthfiyya. 2024. “Exploring Santoon’s Customers’ Preferences That Affect Their Purchase Decision for Buying Fashion Products.” *Journal of Consumer Studies and Applied Marketing* 2 (2): 139–45.