Customer Behavior Analysis for H&M Fashion Retail

The fashion retail industry is characterized by its dynamic nature, heavily influenced by customer preferences, trends, and demographics. To stay competitive and meet customer demands, understanding customer behavior is crucial. This report presents an in-depth analysis of customer behavior based on three primary datasets: "Customers," "Articles," and "Transactions." The analysis encompasses customer demographics, color preferences, purchase patterns, and the impact of club membership on customer lifetime value (CLV).

Data Sets:

- Customers Dataset (~207 MB)
  The "Customers" dataset serves as the foundation for understanding customer characteristics. It includes information on customer demographics, club membership status, fashion news engagement, and more. The dataset has 1371980 rows and 7 columns.

- Articles Dataset (~36 MB)
  The "Articles" dataset plays a pivotal role in discerning customer preferences and purchase patterns. It contains product details such as product types, colors, departments, and other attributes. The dataset has 105542 rows and 25 columns.

- Transactions Dataset (~3.5 GB)
  The "Transactions" dataset includes transactional data, providing insights into sales channels, transaction values, and more. The dataset has 31788324 rows and 5 columns.

Data Pre-processing and Segmentation:

The data was acquired from Kaggle in structured format and stored in pandas data frames. Before embarking on the analysis, we ensured that the raw data is well-structured, and in usable format. We took care of data type conversions for memory optimization by converting numerical attributes to the appropriate numeric data type and categorical attributes to strings or categories.

Data Segmentation: The segmentation of customers into "Low-Value," "Medium-Value," and "High-Value" segments was determined based on the distribution of the transaction price data. "Low-Value Customers" are those whose average transaction price is below the 25th percentile, indicating that they tend to make smaller purchases. "Medium-Value Customers" are customers whose average transaction price falls between the 25th and 75th percentiles, signifying moderate purchase amounts. "High-Value Customers" are those whose average transaction price exceeds the 75th percentile, implying that they make larger and more significant purchases.

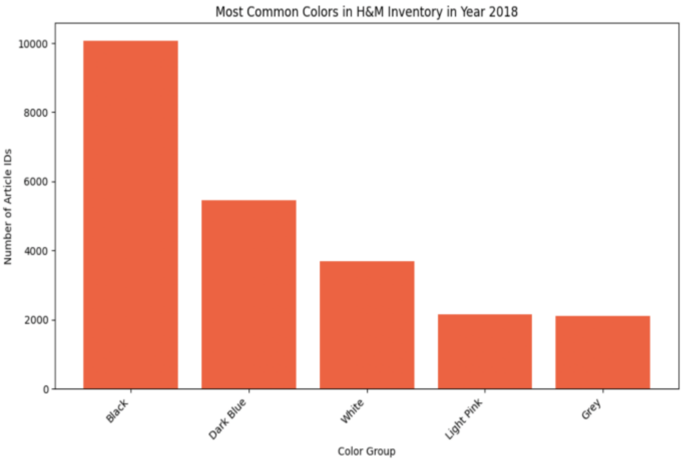
Data Integration: To perform in-depth analysis and uncover meaningful patterns, data from the "Customers," "Articles," and "Transactions" datasets was integrated into a comprehensive data set. Data integration involved combining datasets based on common identifiers (customer_id and article_id) while ensuring data consistency and accuracy.

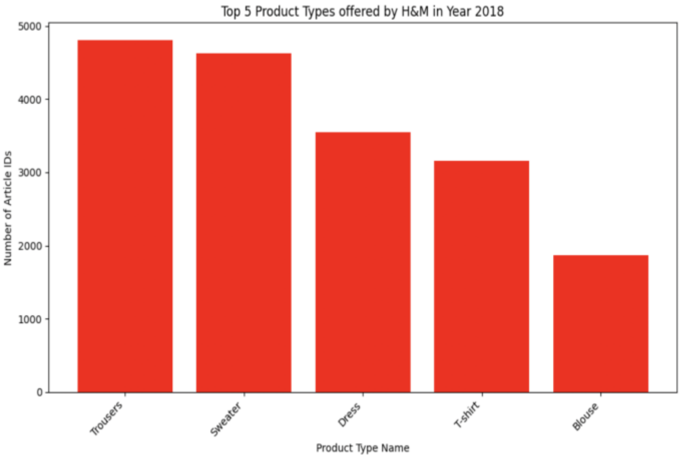
Data aggregation: We performed essential aggregations and summarizations to uncover key insights. These aggregations paved the way for robust and accurate analyses, and enabled us to uncover actionable insights to help drive the informed decision making.

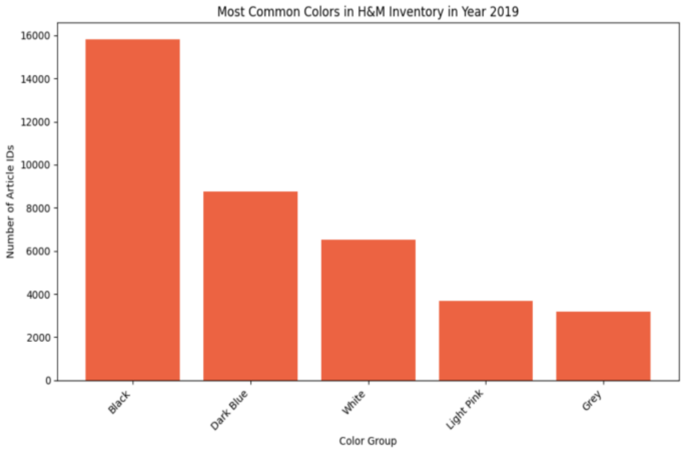
We commenced our analysis by examining the prevalent colors within H&M's inventory and determining the primary product categories that constitute a significant portion of H&M's offerings.

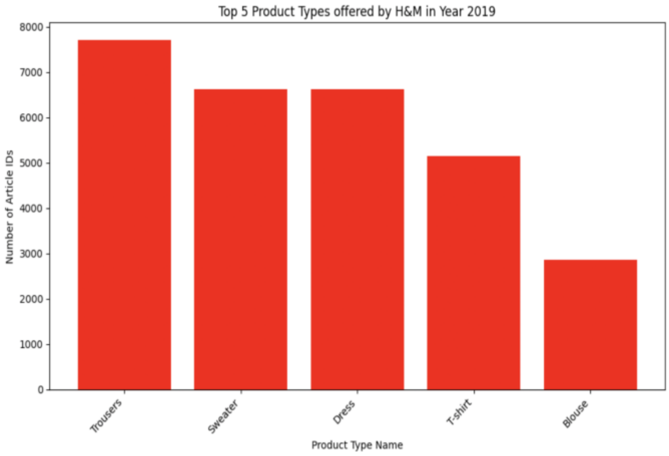
Most common colors in H&M inventory:
Black, Dark Blue, White, Light Pink, Grey
Light Beige

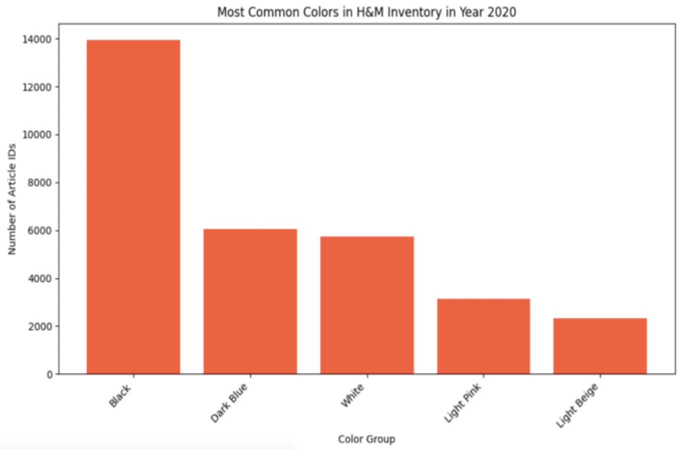
Top 5 Product Types offered by H&M:
Trousers, Sweater, Dress, T-shirt, Blouse,
Shorts

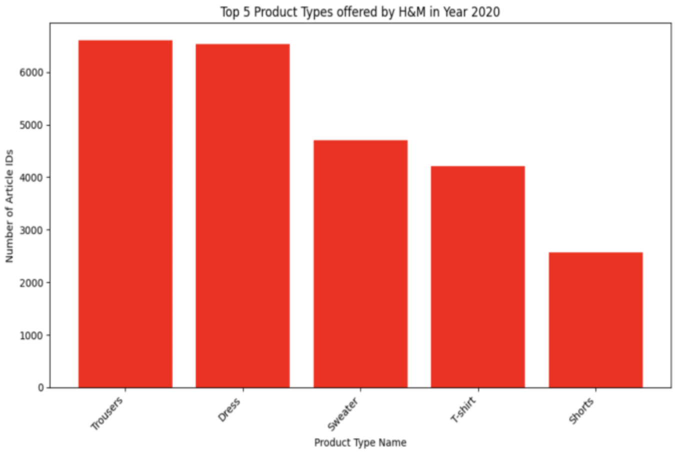




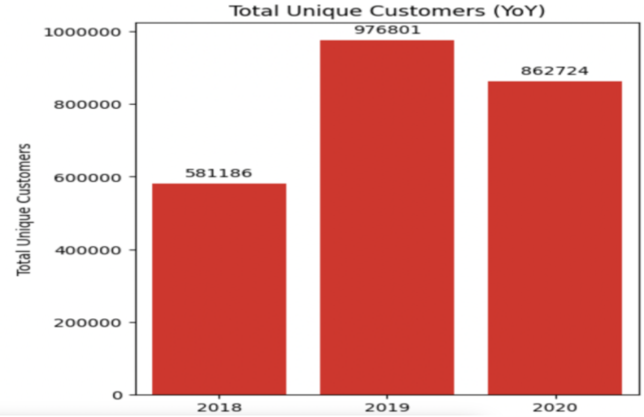


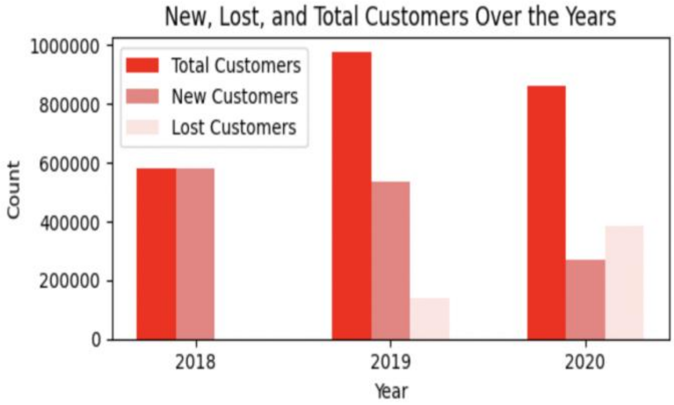




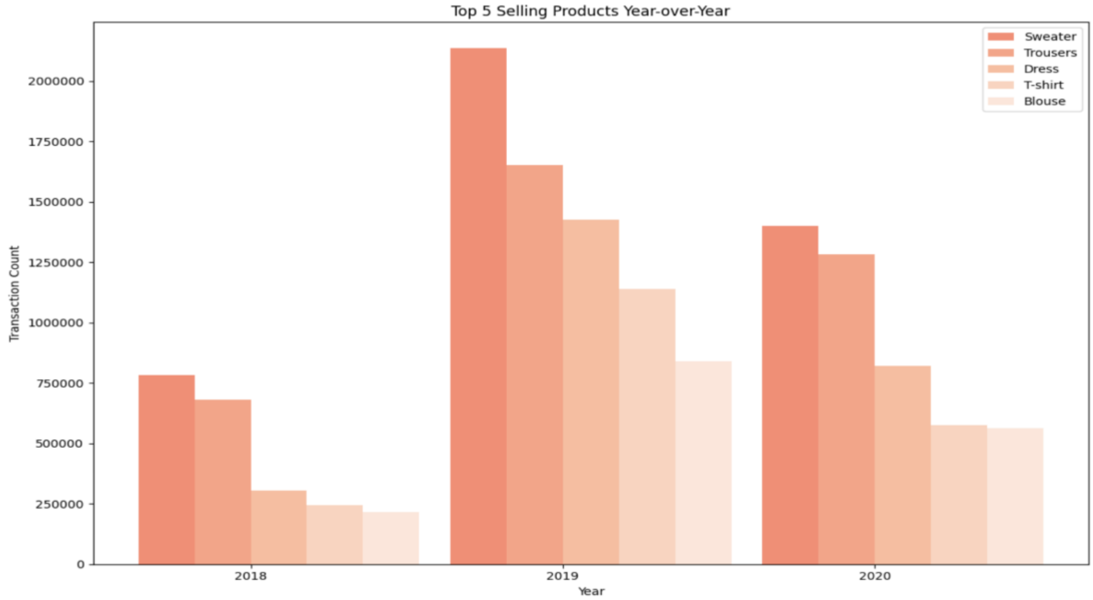


Subsequently, we examined the cumulative count of unique customers over a three-year period and conducted an analysis of customer retention, which involved identifying the total number of customers, new customer acquisitions, and the retention of existing customers. In 2018, the company had 581,186 customers. The following year, in 2019, customer numbers surged by 68.07% to 976,801, signifying significant growth. However, by 2020, there was a notable decrease by 11.68% to 862,724 customers.



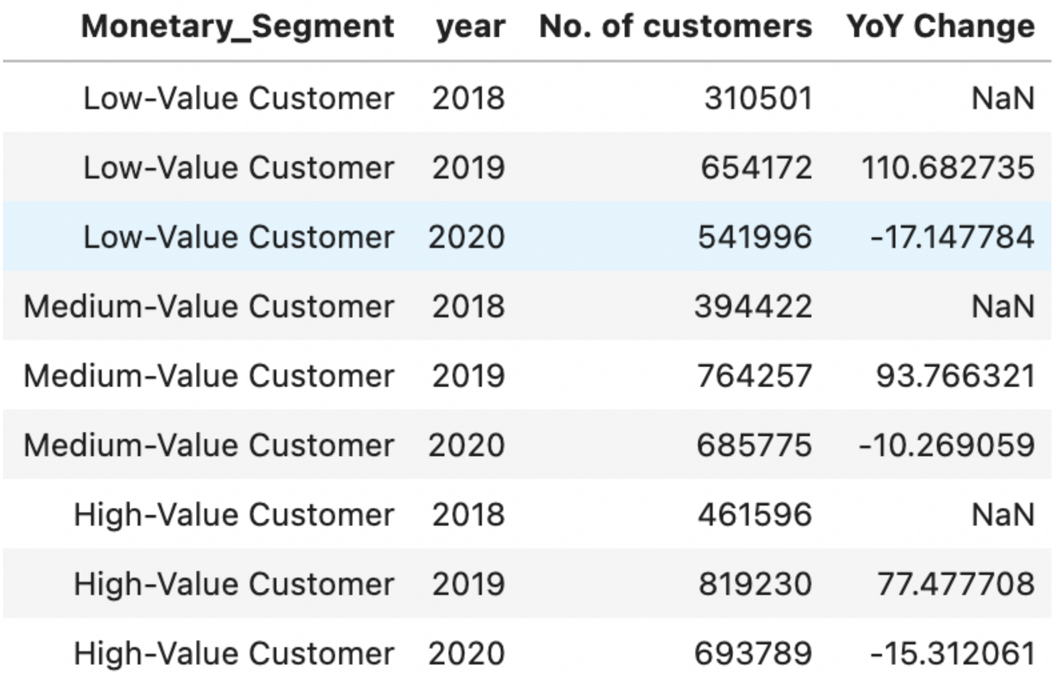


Our exploration of customer behavior began with an in-depth analysis of purchase preferences, where we identified the top 5 products that achieved the highest sales for each of the three individual years. These products reflect the items that attracted the greatest interest from customers, leading to strong sales performance for H&M each year. Notably, Sweater, Trousers, Dress, T-shirt, and Blouse consistently held their positions as the top 5 products with the highest customer interest, maintaining their popularity across the years of analysis.



Analyzing customer retention trends across different monetary segments, it becomes evident that Low-Value Customers exhibited substantial growth from 2018 to 2019, indicating an influx of new customers. However, a notable decline in their numbers in 2020 suggests that many may have not continued their association with the brand. In contrast, Medium-Value Customers displayed a similar pattern of significant growth in 2019 but managed to retain a larger portion of their customer base in 2020, showcasing relatively better customer loyalty. High-Value Customers, while experiencing robust growth in 2019, saw a decline in 2020, although they still maintained a considerable customer base. These observations underscore the varying degrees of customer retention across different monetary segments, highlighting the need for tailored strategies to foster long-term customer relationships, particularly among Low-Value Customers.

YoY Change in Customer Counts by Monetary Segment:

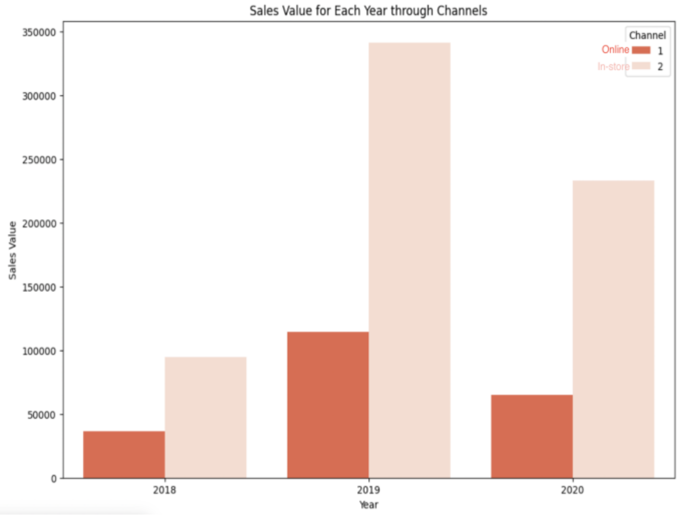


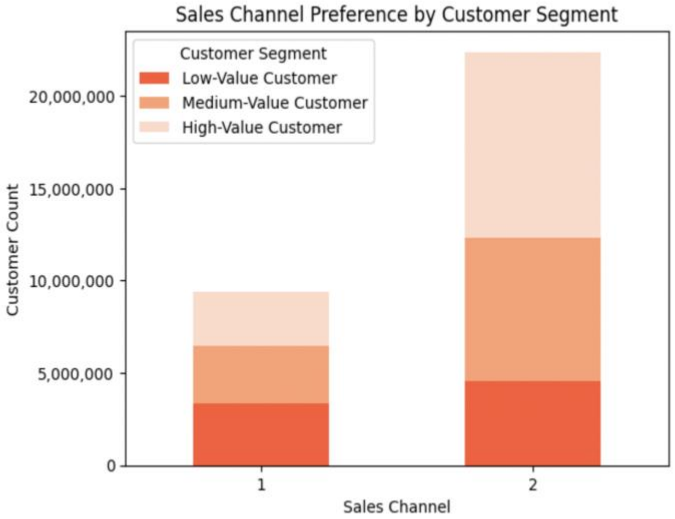
Monetary_Segment	year	No. of customers	YoY Change
Low-Value Customer	2018	310501	NaN
Low-Value Customer	2019	654172	110.682735
Low-Value Customer	2020	541996	-17.147784
Medium-Value Customer	2018	394422	NaN
Medium-Value Customer	2019	764257	93.766321
Medium-Value Customer	2020	685775	-10.269059
High-Value Customer	2018	461596	NaN
High-Value Customer	2019	819230	77.477708
High-Value Customer	2020	693789	-15.312061

In-store purchases(Channel 2) remained the favoured choice for customers in terms of sales across all three years, consistently showing higher sales volumes. The chart clearly indicate the sustained customer preference for channel 2 throughout the three-year period.

- Low-Value Customers: Approximately 3.37 million opt for online(channel 1) purchases, and around 4.58 million choose channel 2, indicating a balanced presence in both channels.
- Medium-Value Customers: Over 3.12 million prefer channel 2, while approximately 312,000 opt for channel 1, highlighting a clear preference for channel 2.
- High-Value Customers: About 2.91 million favor channel 1, while around 10.05 million prefer channel 2, showcasing a strong inclination towards channel 2.

In summary, Medium and High-Value Customers predominantly favor in-store(channel 2) purchases, while Low-Value Customers are more evenly distributed between the two channels.



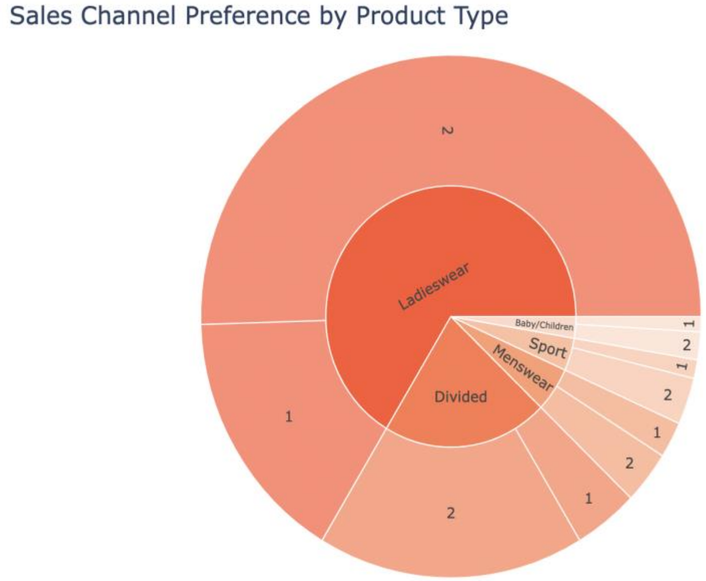


Customers exhibit varying preferences for product categories across sales channels:

- For Baby/Children products, in-store purchases (Channel 2) outperform online platform sales (Channel 1), with approximately $15,224 in sales compared to $8,299.
- In the Divided category, Channel 2 dominates with sales of about $152,471, significantly surpassing the figures from the online platform (Channel 1) at $36,799.

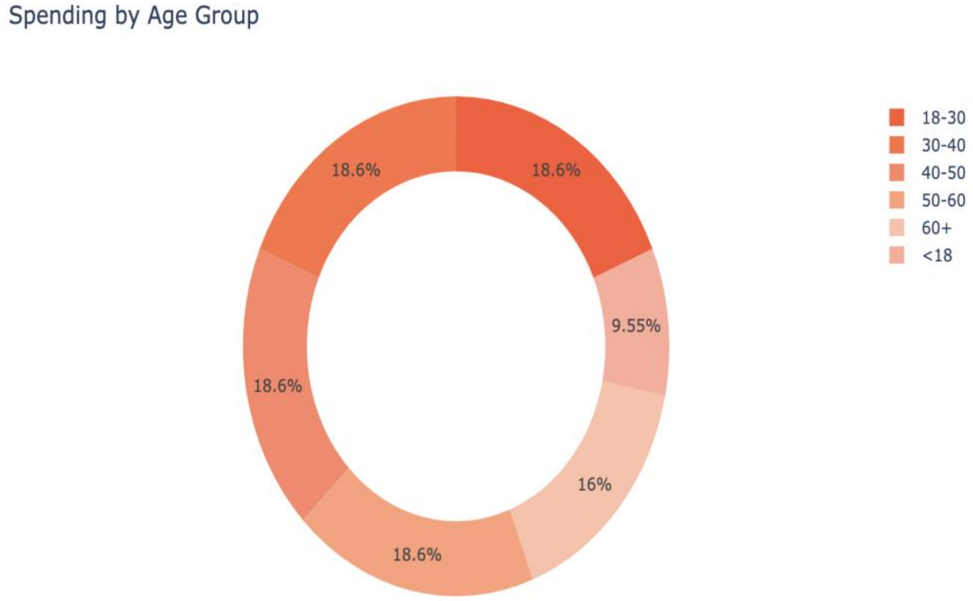
- Ladieswear sales excel in Channel 2, generating approximately $446,612, compared to $140,660 in Channel 1.
- In the Menswear category, Channel 2 prevails with approximately $28,960 in sales, compared to Channel 1's $19,756.
- For Sport products, Channel 2 records approximately $25,696 in sales, while Channel 1 lags behind with around $10,170.

In summary, customers prefer in-store purchases for various product categories, making it crucial for H&M to maintain a robust product offering in this channel to cater to customer preferences effectively.



The analysis of customer behavior reveals valuable insights into their communication frequency, age groups, and spending habits. Firstly, regarding communication frequency, a significant portion of customers in various age groups prefers not to receive fashion news, with proportions ranging from approximately 45% to 59%. Conversely, a substantial proportion subscribes to fashion news, with "Regularly" and "Monthly" frequencies showing higher engagement among customers, particularly in the older age groups.

Interestingly, spending habits among all age groups seem consistent, except for the "60+" group, which records slightly lower spending. This suggests that spending patterns are relatively uniform across age groups.



Hypotheses and Testing:

We formulated and rigorously tested three crucial hypotheses that shed light on various aspects of customer behavior and brand performance. These hypotheses were carefully constructed with the aim of not only understanding customer preferences but also providing actionable insights to enhance brand strategies.

Hypothesis 1: Perceived Color Differences Across Departments

Background and Purpose:

The first hypothesis delved into the world of color psychology and its impact on customer choices. Colors play a significant role in product perception and can influence purchasing decisions. We aimed to assess whether perceived color values significantly differ between different departments of the brand's inventory. This investigation was vital as it could help the brand understand which colors resonate most with customers in specific product categories.

Testing Method:

To address this hypothesis, an analysis of variance (ANOVA) test was conducted. This test allowed us to compare the perceived color values of products across various departments.

Result and Implication:

The ANOVA test revealed a significant difference in perceived color values between departments, as indicated by the ANOVA F-statistic of 6.36 and an exceptionally low p-value of 2.34e-152. These findings offer the brand valuable insights into customer preferences for color across different product categories. Utilizing this information, the brand can optimize its product design and marketing strategies to align with these preferences, potentially increasing customer engagement and sales.

Hypothesis 2: Age Group and Product Type Association

Background and Purpose:

Understanding customer demographics and their preferences is pivotal for tailoring marketing efforts. In this hypothesis, we aimed to explore the association between age groups and product types. This investigation could provide insights into which age groups show a higher interest in specific product categories, enabling the brand to create more targeted marketing campaigns.

Testing Method:

To evaluate this hypothesis, we conducted chi-squared tests, a statistical method for analyzing the relationship between two categorical variables.

Result and Implication:

The Chi-Squared Statistic of 638,533.22 and a p-value of 0.0 indicated a significant association between age groups and product types. This result signifies that different age groups exhibit varying preferences for the brand's product categories. Armed with this knowledge, the brand can refine its product recommendations, promotions, and marketing channels to better cater to each age group, enhancing customer satisfaction and loyalty.

Hypothesis 3: Club Member Lifetime Value Difference

Background and Purpose:

Customer loyalty programs are instrumental in retaining valuable customers. In this hypothesis, we explored whether 'ACTIVE' club members exhibit a higher customer lifetime value (CLV) compared to non-club members. The objective was to assess the effectiveness of the club membership program in increasing customer loyalty.

Testing Method:

To test this hypothesis, we employed a t-test, which is commonly used to compare means between two groups.

Result and Implication:

The t-statistic of 448.63 and a p-value of 0.0 presented strong evidence to reject the null hypothesis. This result implies that 'ACTIVE' club members indeed have a higher CLV than non-club members. Therefore, the club membership program proves to be a valuable asset for the brand, encouraging customer loyalty and potentially leading to higher long-term revenues.

Conclusion

Our journey through customer behavior analysis has unveiled key insights crucial for H&M's continued success in the dynamic fashion retail industry. Notably, color preferences, including black, dark blue, white, light pink, and grey, have played a pivotal role in shaping purchase patterns. Sweaters, Trousers, Dresses, T-shirts, and Blouses have consistently garnered the highest customer interest, underscoring their significance in H&M's inventory.

In the realm of customer retention, we observed nuanced dynamics across monetary segments. Low-Value Customers showed promising growth but required tailored strategies for sustained retention. Medium-Value Customers exhibited strong loyalty, while High-Value Customers maintained a considerable base. These findings emphasize the need for segment-specific retention efforts.

Sales channel analysis highlighted in-store purchases (Channel 2) as the perennial favorite, particularly among Medium and High-Value Customers. Tailoring strategies for diverse customer segments remains pivotal.

Strategically, leveraging color psychology, aligning marketing with age group preferences, and nurturing club memberships can enhance customer satisfaction and loyalty. This data-driven approach equips H&M to thrive in a competitive market, ensuring brand excellence as it evolves with ever-changing trends and customer preferences.

References

We acknowledge the data sources and external tools used during the project.

1. Kaggle (n.d.). Kaggle Competitions Notebooks. https://www.kaggle.com/competitions
2. Stack Overflow (n.d.). https://stackoverflow.com/
3. Kaggle Resources (n.d.). Kaggle Learning Resources. https://www.kaggle.com/learn/overview

These references were consulted for information and guidance during the course of this analysis and report preparation.

Link to Notebook: https://www.kaggle.com/code/rajatchelani/ist652-final-project-code
Link to Datasets: https://www.kaggle.com/c/h-and-m-personalized-fashion-recommendations/data

Note: The implementation and execution of the Jupyter Notebook for this project were conducted on the Kaggle platform due to resource limitations on Google Colab. Kaggle provided the necessary computing resources and libraries required for this analysis. This decision was made to ensure smooth execution and efficient processing of the datasets and analyses presented in this report.

Submitted by:
Rajat Chelani