Predicting E-Commerce Marketing Analytics Using Machine Learning and Big Data Analytics

Research Paper

Didem B. Aykurt

Colorado State University Global

MIS581: Business Intelligence and Data Analytics

Dr. Steve Chung

January 7, 2023

Table of Contents

ABSTRACT

This comprehensive analysis delves into diverse facets of online sales dynamics by meticulously exploring datasets encompassing pricing, customer engagement, tenure patterns, discount distributions, product performance, financial trends, marketing dynamics, and predictive modeling. Visualizations, statistical analyses, and predictive modeling techniques unveil valuable insights crucial for strategic decision-making. The study begins by illuminating the distribution of average prices, shedding light on frequency distributions and patterns. It further explores customer engagement with promotional coupons, emphasizing the need to optimize conversions from interest to transactions. Customer tenure analysis provides pivotal insights into temporal dynamics, enabling tailored retention strategies. Examining discount percentages, top-selling products, and revenue trends unveils actionable insights for resource allocation, inventory management, and strategic marketing efforts.

Introduction

E-commerce is the process of buying and selling goods and services over the Internet using various platforms and devices. E-commerce has become a vital part of the global economy, as it offers many benefits to businesses and customers, such as convenience, variety, and lower costs. According to a report by eMarketer, global e-commerce sales reached $4.28 trillion in 2020, an increase of 27.6% from the previous year, and are expected to grow to $6.38 trillion by 2024 by Aslmadi, Shuhaiber, Al-Okail, Gasaymeh, and Alrawashdeh (2023).

However, e-commerce faces many challenges and uncertainties, such as increasing competition, changing customer preferences, security and privacy issues, and technological disruptions. Marketing is the process of creating, delivering, and communicating value to customers and stakeholders and involves various activities, such as market research, product

development, pricing, promotion, and distribution (Akter & Wamba, 2016). Marketing analytics measures and analyzes marketing performance and outcomes using various metrics and tools, such as customer segmentation, customer lifetime value, conversion rate, return on investment, etc. Aim to investigate how machine learning and big data analytics can help e-commerce businesses predict and optimize their marketing performance, such as customer segmentation, lifetime value, conversion rate, return on investment, etc. In this digital transformation, marketing analytics has emerged as a pivotal tool for businesses seeking to thrive in the competitive e-commerce sphere. Understanding and harnessing data-driven insights play a critical role in navigating the intricacies of consumer preferences and market trends, as postulated by Fantini and Narayandas (2023).

Potential Of the Data Within Your Dataset and What It Means for The Organization

The dataset provided holds immense potential for uncovering customer behavior, refining marketing strategies, and improving overall business performance. It would be helpful for **Coupon Engagement and Utilization** (Mela, 2018). This is because analysis of coupon engagement and actual utilization can unveil insights into customer preferences, tendencies, and promotion responses. This understanding helps tailor coupon strategies to optimize conversions. It also offers an opportunity for **Demographic and Tenure Analysis**. Analyzing how demographics (like gender and location) correlate with tenure provides insights into customer loyalty and preferences. Understanding tenure patterns aids customer segmentation and targeted marketing (Mela, 2018). The dataset is opportunistic for deciphering **Product Preferences and Sales Trends**. Analysis of product categories, quantities purchased, and average prices offers insights into popular products, customer preferences, and buying behaviors. This knowledge helps optimize inventory and marketing efforts (Mela, 2018).

The dataset also offers the potential for Optimizing Marketing Strategies through Marketing Spend Analysis. Correlating marketing spending (both offline and online) with sales

quantity allows for the identification of effective marketing channels. It guides resource allocation towards more impactful campaigns (Mela, 2018). Datasets can also be utilized **to understand coupon distribution trends, and their effectiveness helps in refining coupon strategies, targeting potential customers, and improving** conversion rates. It would be helpful for merchants to **identify top-performing products. The analysis** of top-selling products provides critical information for inventory management, focusing marketing efforts, and identifying opportunities for cross-selling or bundling (Mela, 2018).

Therefore, the dataset also has the potential to enhance Business Performance **through Revenue Trend Analysis**. Assessing monthly revenue trends identifies peak sales periods, facilitating better inventory management and resource allocation for marketing during high-demand periods.

The business can undertake customer Segmentation and Retention. Customer segmentation based on tenure or behavior enables personalized marketing strategies, enhancing customer retention and overall satisfaction (Mela, 2018). Therefore, through the insights from this dataset, data-driven decisions can be made, marketing expenditures can be optimized, promotional activities can be tailored, customer engagement can be enhanced, and long-term customer relationships can be fostered. Collectively, these analyses contribute to improved business performance, better resource allocation, and strategic decision-making, ultimately leading to sustainable growth and success in the competitive e-commerce landscape.

Future Plans to Analyze the Data Within the Dataset

The first potential analysis revolves around Customer Segmentation. This would be done to group customers based on behavior, preferences, and tenure, allowing for targeted and personalized marketing strategies. The approach would utilize clustering algorithms (such as K-

Means or hierarchical clustering) on variables like tenure, quantity purchased, average price, and demographic information (Li, 2020). We shall seek to segment customers into distinct groups based on similarities in behavior, enabling tailored marketing campaigns for each segment. This would help with an enhanced understanding of customer groups, allowing for personalized marketing strategies, product recommendations, and improved customer retention.

Another analysis would be undertaken for **Predictive Modeling for Sales Forecasting and Customer Value Prediction. This would be done to** forecast future sales trends and identify potential high-value customers for targeted marketing, as Fantini and Narayandas (2023) stipulated. This would employ machine learning algorithms (e.g., regression models, time-series forecasting) using historical sales data, marketing spending, and customer demographics. The analysis would predict future sales volumes and identify customers likely to make high-value purchases based on historical behavior and demographics. Accurate sales forecasts aid in inventory planning, resource allocation, and targeted marketing strategies to maximize revenue. This would help identify high-value customers for personalized marketing and customer relationship management.

There would also be a marketing Spend Analysis for ROI Assessment to evaluate the effectiveness of marketing spending by correlating it with sales performance. The approach would aim to analyze the relationship between offline/online marketing spending and corresponding sales quantities using correlation analysis, as Fantini and Narayandas (2023) prescribed. This would be done by computing the Return on Investment (ROI) by comparing marketing spending against generated sales, considering different channels and periods. This would help to shed insights into the effectiveness of different marketing channels and campaigns.

This is crucial for the optimization of marketing budgets by reallocating resources to high-performing channels and improving overall ROI.

The Data Preparation would involve cleansing and preprocessing data for modeling, ensuring accuracy and consistency. **Model evaluation and refinement would be done using appropriate evaluation metrics and iterating** for refinement to test predictive models on a separate validation dataset to assess real-world performance. This would help to translate findings into actionable strategies and recommendations for marketing teams and decision-makers.

These future analysis plans aim to drive actionable insights, optimize marketing strategies, and improve decision-making by leveraging advanced analytics techniques on the dataset, ultimately enhancing business performance and competitiveness in the e-commerce sector.

Objectives

The main objective of this thesis is to explore how machine learning and big data analytics can be used to predict e-commerce marketing analytics. The thesis will review the existing literature and practice on this topic and identify the main challenges and opportunities.

The main contribution and significance of this thesis are as follows:

- This thesis project stands as a beacon in marketing analytics for e-commerce, amalgamating cutting-edge data analytics techniques with real-world data to provide actionable insights that can steer businesses toward tremendous success in the digital marketplace.
- This thesis provides a comprehensive and systematic review of the existing literature and practice on machine learning and big data analytics for e-commerce marketing analytics. It identifies the main gaps and research questions.

- This thesis proposes a novel framework and methodology for applying machine learning and big data analytics to predict e-commerce marketing analytics. It provides a detailed step-by-step guide for implementing and evaluating the proposed framework and methodology.
- This thesis demonstrates the feasibility and effectiveness of the proposed framework and methodology through a case study of a real-world e-commerce business problem. It shows how the proposed framework and methodology can enhance marketing performance and outcomes.
- This thesis discusses the implications and limitations of the proposed approach and proposes some future research directions.

Overview of Study

This thesis revolves around leveraging marketing analytics to glean invaluable insights within the e-commerce sector. The dataset, spanning from January 2019 to December 2019, encapsulates a wealth of transactional, demographic, and marketing information. It provides a unique opportunity to conduct comprehensive analyses, as supported by Li (2020):

- Understanding customer behaviors, preferences, and purchasing patterns.
- Evaluating the impact of marketing strategies and discount offers on sales.
- Unveiling correlations between marketing spending and revenue generation.
- Assessing the seasonality, trends, and patterns within different product categories and customer segments.

The dataset serves as a rich source of information, enabling advanced analytics approaches such as customer segmentation, predictive modeling, cohort analysis, and market basket analysis

(Mela, 2018). Its comprehensive nature allows for a deep dive into the various facets of e-commerce operations, facilitating informed decision-making and strategy formulation within the digital marketplace (Mela, 2018).

Research Questions and Hypotheses

Research Question

Customer Segments and Sales Differences:

- Research Question (RQ 1): Is there a difference in the quantity of items purchased across different customer segments?

Marketing Effectiveness and Sales Differences:

- Research Question (RQ 2): Is there a difference between marketing spending in online sales versus offline sales?

GST Impact on Product Categories:

- Research Question (RQ 3): How does the GST percentage vary across different product categories, and what is the potential impact on sales?

Predictive Model:

- Research Question (RQ 4): How can the insights from the model be used to develop and implement more effective customer retention strategies?

Overview of the Research Hypotheses

An overview of the hypotheses that could be formulated to address each research question is presented here below:

Customer Segments and Sales Differences:

- Null Hypothesis (H10): There is no significant difference in the mean of items purchased across different customer segments in the population of interest.
- Alternative Hypothesis (H1a): There is a significant difference in the mean of items purchased among distinct customer segments in the population of interest.

Marketing Effectiveness and Sales Differences:

- Null Hypothesis (H20): There is no significant difference between mean offline and online marketing spending of customers in the population of interest.
- Alternative Hypothesis (H2a): There is a significant difference between customers' mean offline and online marketing spend in the population of interest.

GST Impact on Product Categories:

- Null Hypothesis (H30): There is no significant difference in GST percentage variations across different product categories on sales in the population of interest.
- Alternative Hypothesis (H3a): There is a significant difference in GST percentage variations across different product categories on sales in the population of interest.

LITERATURE REVIEW

In modern commerce, the landscape of consumer behavior has evolved exponentially, particularly within the expansive domain of e-commerce. Understanding and dissecting the intricate mechanisms driving customer spending has become imperative for businesses aiming to thrive in the digital marketplace.

Customer Engagement and Coupon Utilization

Scholarly research on customer engagement and coupon utilization unveils multifaceted insights into consumer behavior in leveraging coupons within online settings. Studies such as Mills and Zamudio (2018) presented findings from exploring online coupon usage and shed light on the psychological underpinnings driving consumer engagement. Their findings underscore the multifaceted nature of motivations, revealing that psychological factors like perceived value, scarcity, and social influence significantly impact coupon engagement beyond monetary incentives.

Moreover, Stocchi et al. (2018) emphasized the influence of convenience and perceived effortlessness in online coupon usage. Their study emphasizes that user-friendly interfaces and simplified redemption processes positively correlate with heightened coupon utilization. Correlations between engagement and actual coupon usage have been established, and the study established a direct link between active engagement (such as sharing, bookmarking, or reviewing coupons) and subsequent utilization rates.

Customer Demographics and Tenure Analysis

Bonacchi and Perego (2018) have presented the usefulness of marketing analytics in investigating the correlation between customer demographics and tenure, providing crucial insights into loyalty, retention, and tenure variations across diverse demographics and geographical contexts. For instance, Krautz and Hoffmann (2017) revealed intriguing nuances in the influence of demographics on tenure. Their research highlights that while gender may not directly impact tenure, it interacts with other factors like product preferences or shopping behaviors, indirectly influencing loyalty and retention rates. Kumar and Reinartz (2018) explored the significance of customer satisfaction in different segments and its impact on firm performance. It emphasizes the

importance of understanding diverse customer profiles for effective marketing strategies. Bonacchi and Perego (2018) provided an in-depth analysis of different models and applications of Customer Lifetime Value. They discussed the significance of LTV in marketing strategies and its impact on long-term customer relationships.

Similarly, findings from Wu and Li (2018) demonstrate the significance of geographical contexts in shaping customer tenure. Their study identifies cultural settings and regional preferences as influential factors affecting customer loyalty. They emphasize that localized marketing strategies tailored to regional preferences are pivotal in enhancing tenure among diverse demographic groups.

Marketing Effectiveness and Sales Correlation

Scholarly investigations examining the correlation between marketing spending and sales quantity underscore the complex relationship between marketing strategies and sales performance. For instance, studies by Homburg et al. (2020) delve into the comprehensive impact of marketing spending on sales. Their research analyzes the direct correlation between increased marketing investments and sales volume and dissects the differential effects across various marketing channels. They emphasize the necessity of evaluating ROI to measure the effectiveness of different channels, revealing that while some channels might yield immediate sales, others contribute to long-term brand equity and sustained sales growth.

Furthermore, Halan and Singh's (2023) research focuses on the effectiveness of online marketing vis-à-vis offline strategies. Their findings highlight the shifting landscape and the rising influence of digital marketing on sales. They emphasize the need for a balanced approach, combining online and offline strategies to maximize sales performance. Understanding the interplay between diverse marketing channels and their impact on sales volume aids in crafting

holistic marketing strategies, optimizing budget allocation, and enhancing overall sales effectiveness.

GST Impact on Product Categories

Scholarly investigations into the impact of GST on product categories illuminate the intricate relationship between taxation, consumer behavior, and sales trends. Research by Yeo (2023) provided insights into how varying GST rates influence consumer preferences across product categories. The study emphasizes that differential tax rates affect price sensitivity and consumer choices, leading to fluctuations in demand within specific product categories. For instance, they highlight that essential commodities with lower tax rates exhibit more stable demand patterns than luxury items subject to higher taxes, impacting consumer decisions and sales trends.

The literature review on customer spending and e-commerce analytics underscores the intricate interplay between consumer behavior, marketing strategies, and external factors like taxation. Across diverse studies, a coherent narrative emerges, highlighting the multifaceted nature of consumer engagement with coupons in online settings. Understanding diverse customer profiles becomes pivotal for effective marketing strategies. In evaluating marketing effectiveness and sales correlation, research by Homburg et al. and Halan and Singh emphasizes the multifaceted impact of marketing spending, ROI evaluation, and the significance of a balanced online-offline approach for optimal sales performance. This review of studies underscores the necessity for nuanced approaches in e-commerce analytics. Understanding consumer behavior, leveraging diverse marketing channels effectively, considering demographic nuances, and adapting to external factors like taxation is pivotal for optimizing customer spending and enhancing e-commerce success.

RESEARCH DESIGN

METHODOLOGY

Research Methods

The research methodology employed in this study predominantly falls under quantitative data analysis. Quantitative analysis involves using statistical methods, numerical data, and computational techniques to understand patterns, relationships, and trends within datasets.

The following Statistical Techniques will be utilized in the analysis:

- **Histograms and Kernel Density Estimation:** These methods visually represent the distribution of numerical data, such as 'Avg_Price' and discount percentages, providing insights into their frequency distribution and central tendencies.
- **Descriptive Statistics:** Summary statistics, like mean, median, and interquartile range, quantify aspects of the data distribution, such as tenure duration, helping understand variability and typical values within the dataset.
- **Correlation Analysis:** Examining the relationship between variables, such as online spending and item quantity, using correlation coefficients to measure the strength and direction of these relationships.
- **ANOVA (Analysis of Variance):** Assessing differences among group means, for instance, in the quantity of items purchased across different customer segments, to determine if these differences are statistically significant.

Visualizations such as bar charts and plots will be done. These graphical representations, such as bar charts depicting top-selling products or GST percentages by product category, provide clear visual insights into trends, comparisons, and distributions within the data (Fantini & Narayandas, 2023).

Modeling Techniques will be used, such as K-Means Clustering. It is a method used to identify distinct groups or clusters based on similarities, applied here for customer segmentation based on tenure. Logistic Regression will also be used as a predictive modeling technique to forecast customer retention based on transaction patterns. The null hypotheses in hypothesis testing set up the baseline assumption that no significant difference or relationship exists between specific variables (Fantini & Narayandas, 2023). These hypotheses are then tested statistically to either reject or fail to reject them, providing evidence for or against certain conclusions. The methodology's objective revolves around deriving insights, identifying patterns, and making data-driven decisions, as supported by Li (2020). For instance, they understand customer behavior, optimize marketing strategies, or forecast customer retention.

Each method aims to extract meaningful insights and test specific hypotheses regarding customer behavior, sales trends, and marketing dynamics. The null hypotheses provide a baseline for comparison to derive statistically significant findings and make informed business decisions.

Ethical Considerations for the Data

Ethical considerations are of utmost importance when handling data, mainly quantitative analysis, as they ensure that responsible and ethical research practices are adopted. Several vital ethical considerations in this study are integral to handling diverse data types. The first consideration is **privacy and confidentiality** breaches (Li, 2020)**. Anonymization** and **safeguarding** customer information are imperative, especially in segmentation and retention analysis. It is vital to avoid using personally identifiable information without explicit consent. Careful handling of financial data is necessary to prevent association with individual customer identities, thereby averting potential financial privacy breaches (Li, 2020).

The second consideration is that of **informed consent and data collection.** When data involves customer behavior, such as coupon usage or purchasing patterns, ensuring customer consent for data collection is vital. Clear communication about the purpose and utilization of collected data is necessary.

The third consideration is Data Security and Storage. Robust data security measures must be in place to prevent unauthorized access, breaches, or misuse of sensitive data. Transparency about data usage, analysis methods, and the objectives behind data utilization are fundamental.

Limitations of Analysis

While this analysis provides valuable insights, it is essential to acknowledge its limitations: The analysis's effectiveness heavily relies on the quality and quantity of available data. Incomplete, inaccurate, or insufficient data could skew conclusions or limit the depth of insights. For instance, if specific customer segments or product categories are underrepresented in the dataset, the analysis might not accurately capture their actual behavior or significance. The dataset might not represent the entire customer base or market accurately. It could be not very objective towards specific demographics, behaviors, or purchasing patterns, leading to conclusions that might not be universally applicable. Hypothesis testing assumes certain conditions (like normality, equal variance, etc.), and violations of these assumptions can affect the accuracy of results. Moreover, statistical significance does not always equate to practical significance. Though providing insights, the predictive regression model for customer retention has its limitations. Its accuracy might be moderate, but it might not capture all factors influencing customer retention. Factors beyond the dataset, like external market changes or customer preferences, could significantly impact retention. Addressing these limitations could involve refining data collection methods, incorporating

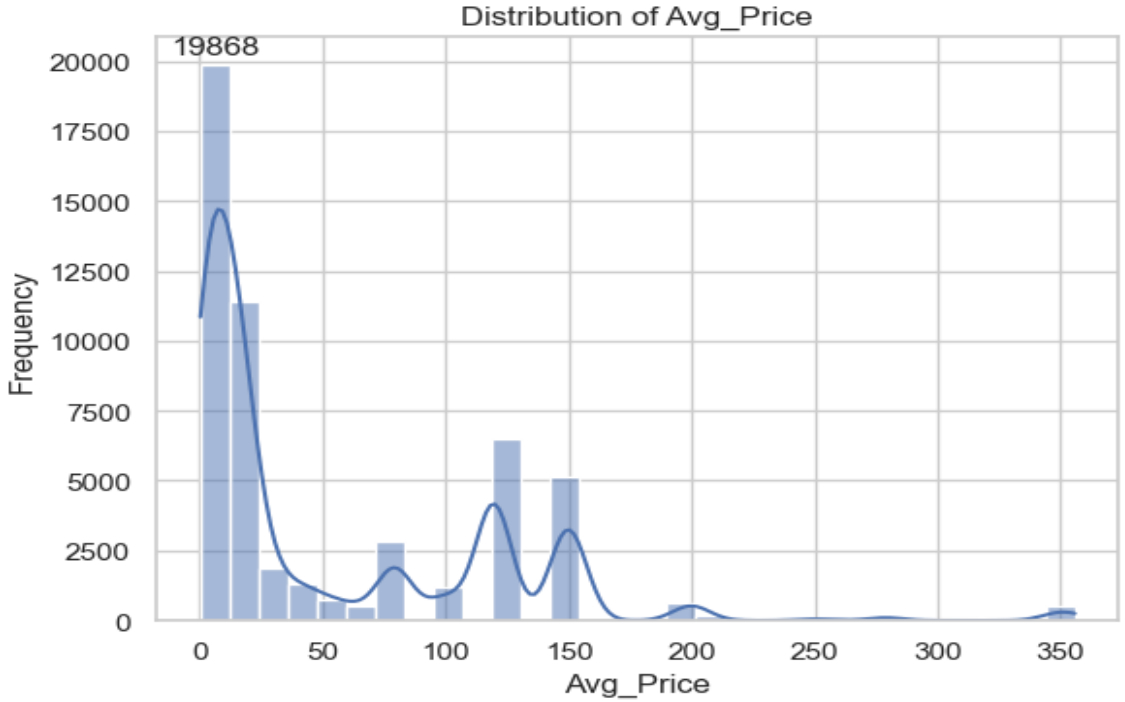
external data sources, conducting more granular analyses, validating models with real-time data, and considering a broader range of market dynamics factors.

Findings

Price Distribution

The distribution of the 'Avg_Price' variable in the Online Sales dataset was effectively visualized using histograms and kernel density estimation. The histograms provide a detailed overview of the frequency distribution of 'Avg_Price' values, with each bar annotated to display the precise count. This visual representation allows for identifying central tendencies and patterns in the data and highlights the frequency of occurrence for different price ranges. Kernel density estimation further enhances the understanding of the distribution shape. The accompanying printed count values offer a comprehensive list of the occurrences of each unique 'Avg_Price' value.

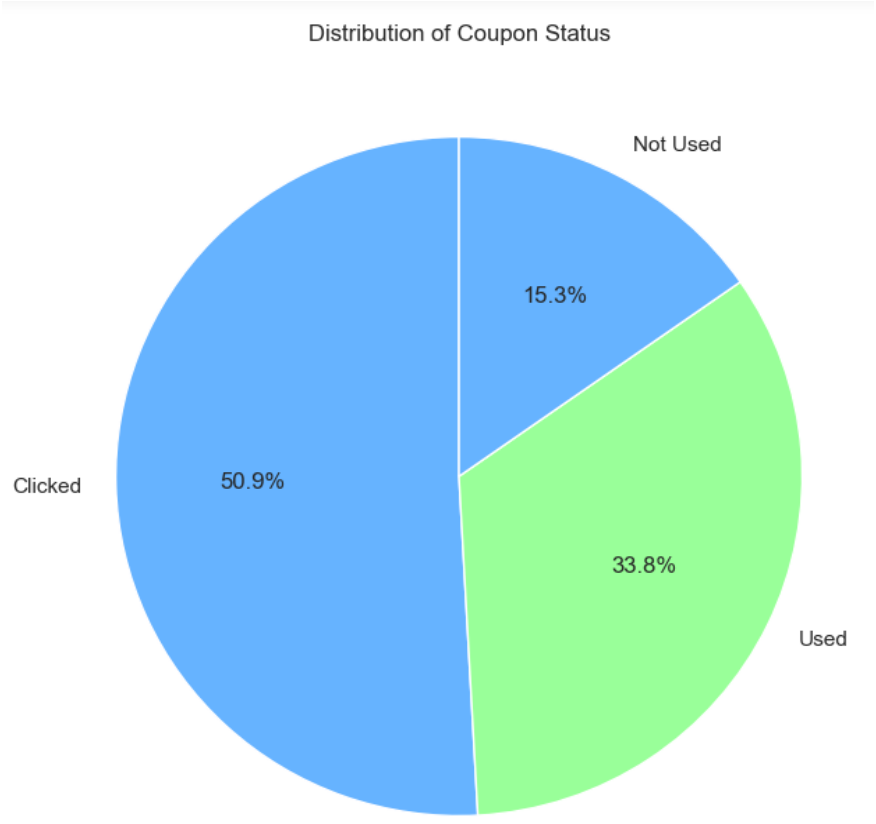
Figure 1: The graph shows the 'Avg_Price' frequency in the Sales dataset.



Exploring Coupon Status Distribution in Online Sales

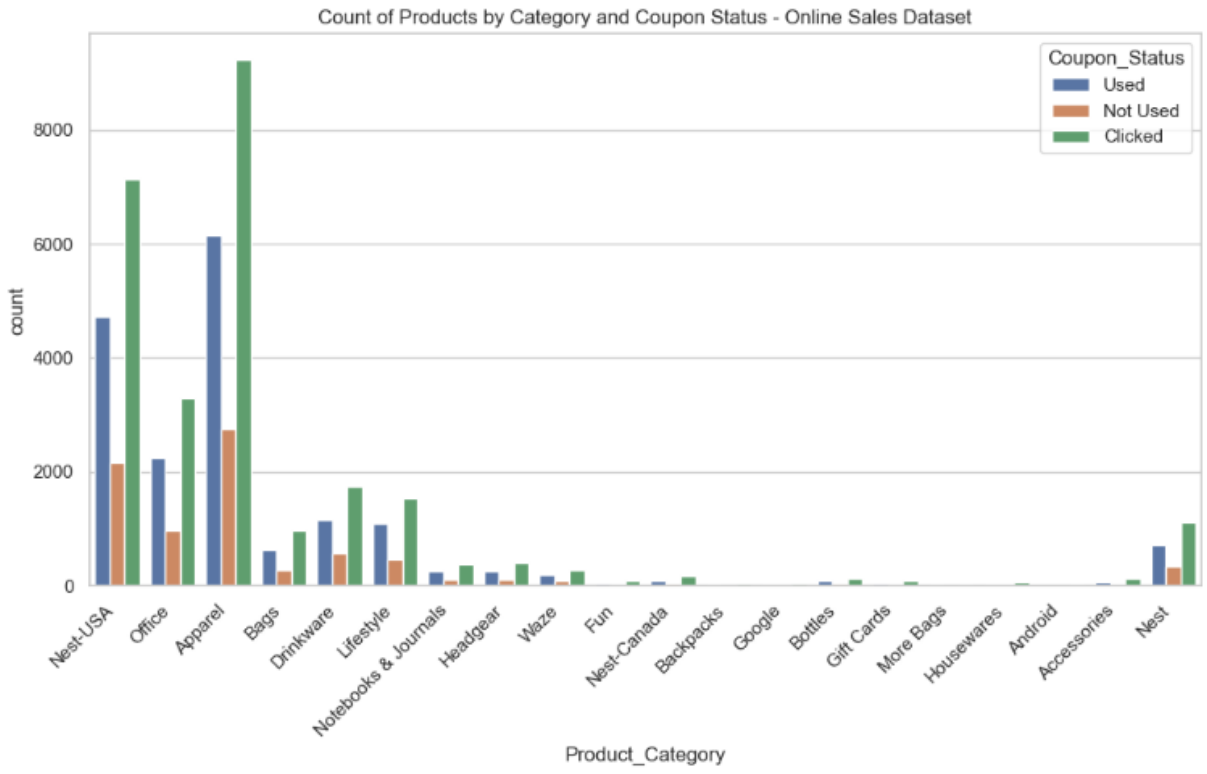
Analyzing the 'Coupon_Status' variable in the Online Sales dataset has unveiled essential patterns in customer engagement with promotional coupons. The prevalence of 'Clicked' coupons, followed by actual coupon utilization ('Used'), suggests a robust interest in discounts among customers. However, the presence of a significant number of 'Not Used' coupons implies untapped opportunities for conversion. These insights can guide strategic marketing efforts, allowing the organization to tailor promotional campaigns to customer preferences. The findings highlight the need to focus on attracting clicks and optimizing the conversion of interest into actual transactions. By leveraging these insights, the organization can refine its marketing strategies, enhance customer engagement, and capitalize on the latent potential within coupon interactions, contributing to the overall success of online sales.

Figure 2: The pie chart shows the 'Coupon_Status' percentage for each group in the Sales dataset.



The inspection of the count of products by category and coupon status from the online sales dataset, which categories have the highest or lowest counts of products, and how they vary across coupon statuses. The “Apparel” category has the highest count for used and not used coupons. The used coupon status has a higher count than the non-used one for most categories, except for “Office” and “Nest’s USA.” This means that customers are more likely to buy products from those categories if they have a coupon. Also, see that the clicked coupon status has a very high count for all categories, which means that customers are very interested in clicking on the coupons.

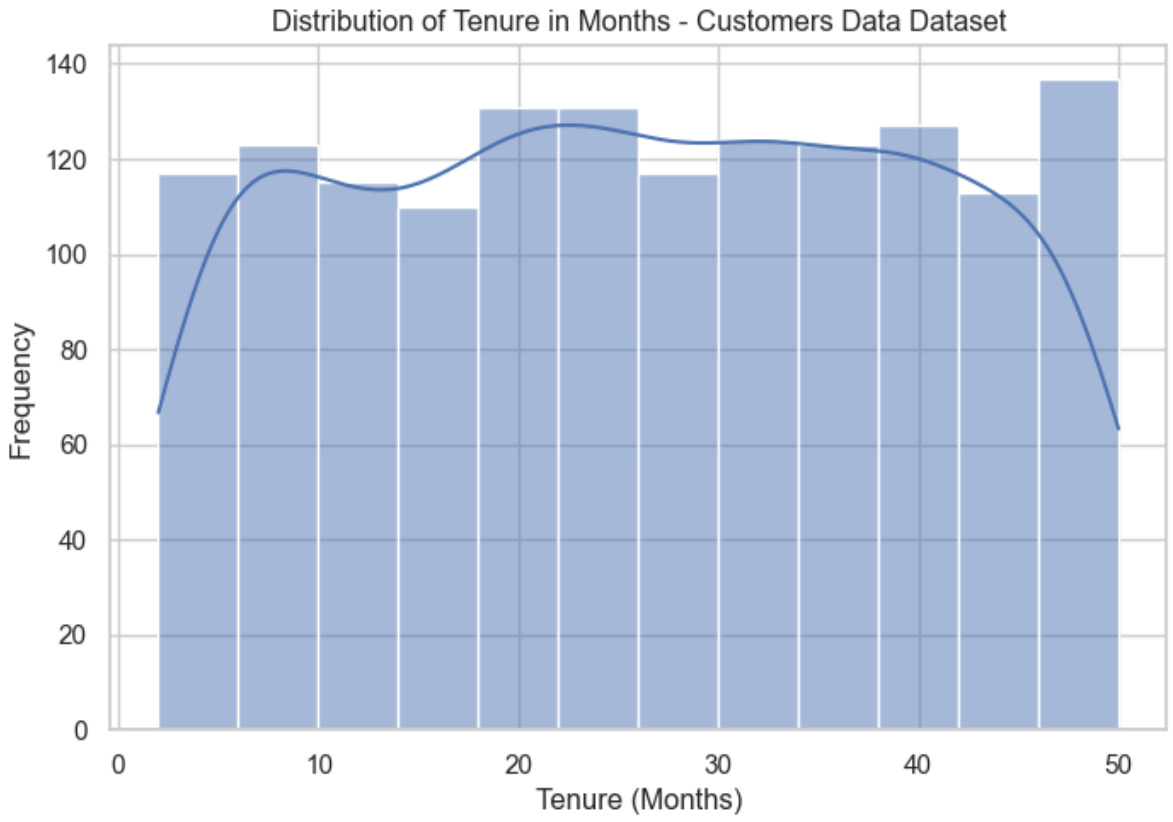
Figure 3: The bar graph shows the product count by category and coupon status in the Sales dataset.



Exploring Customer Tenure in the Customer Data Dataset

The exploration of customer tenure patterns within the Customers Data dataset has illuminated crucial insights into the temporal dynamics of customer relationships. The histogram with kernel density estimation, presented in the visual representation 'Distribution of Tenure in Months,' effectively captures the concentration of customers across varying tenure ranges. The quantitative summary statistics further reinforce this understanding, indicating a mean tenure of approximately 25.91 months and a diverse range from 2 to 50 months. The pronounced interquartile range from 14 to 38 months highlights the substantial duration of customer engagements. This comprehensive analysis provides a clear snapshot of tenure distribution and sets the stage for strategic decision-making. Recognizing the longevity and variability of customer relationships equips the organization with valuable insights to tailor retention strategies, enhance customer satisfaction, and foster long-term loyalty, contributing to the business's overall success.

Figure 4: Summary of the 'Tenure_Months' variable in the Customer dataset



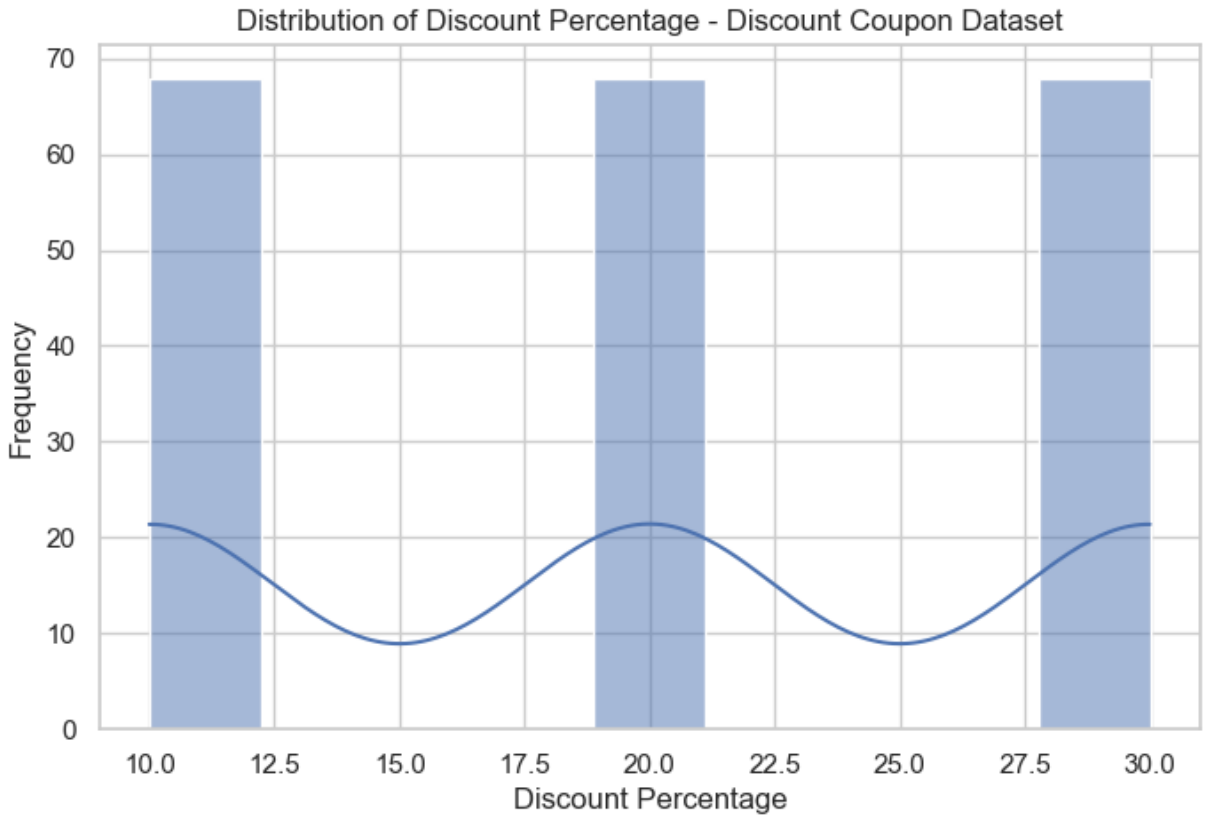

```
Summary Statistics of Tenure_Months
count    1468.000000
mean       25.912125
std        13.959667
min         2.000000
25%        14.000000
50%        26.000000
75%        38.000000
max        50.000000
```

Analyzing Coupon Distribution: Unveiling Discount Percentages

The distribution of discount percentages in the Discount Coupon dataset is unveiled through a comprehensive histogram and kernel density estimation. The visualization, centered around the theme 'Distribution of Discount Percentage,' delineates the spread of discount values on the x-axis against their respective frequencies on the y-axis. A deeper analysis employs a group-by approach, highlighting the count of coupons associated with distinct discount percentages. Specifically, 10%, 20%, and 30% discounts are each represented by 68 coupons. This detailed breakdown augments

the visual representation, offering a nuanced perspective on the prevalence of various discount ranges within the dataset.

Figure 5: Frequency of 'Discount_pct' from Discount Coupon dataset.

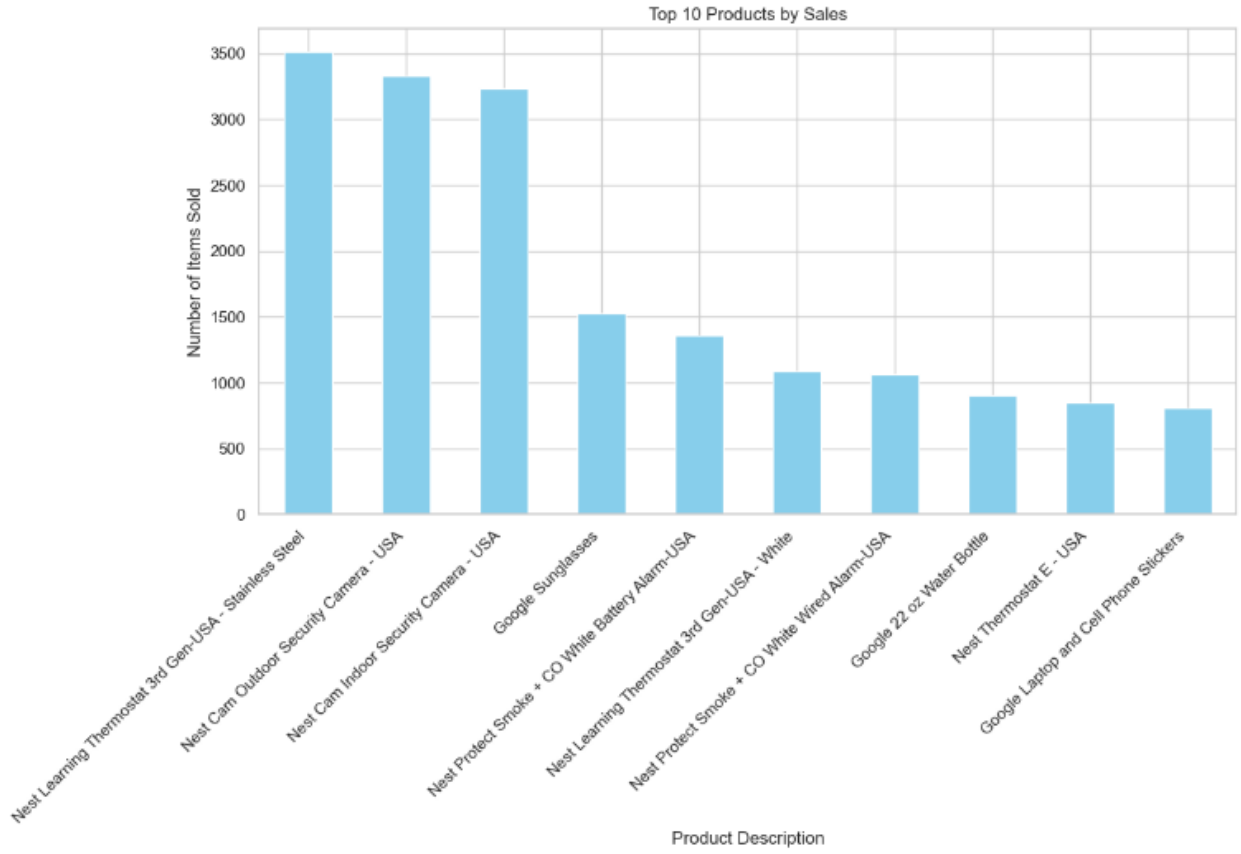


Unveiling Sales Trends: Top 10 Products Analysis

Examining the Online Sales dataset unveiled compelling insights into product performance, prominently featured in a bar chart showcasing the top 10 products by sales. The Nest Learning Thermostat 3rd Gen-USA in Stainless Steel is leading the chart, boasting an impressive 3,511 units sold. Noteworthy contenders include the Nest Cam Outdoor Security Camera and Nest Cam Indoor Security Camera. This comprehensive analysis is a pivotal tool for strategic decision-making within the organization. Understanding these top products' popularity and sales dynamics enables precise resource allocation, targeted marketing efforts, and streamlined inventory management.

The visual representation is a quick reference for stakeholders, empowering them to make informed decisions that drive sales growth and elevate customer satisfaction.

Figure 6: The bar chart shows the top 10 by counting occurrences from the Sales dataset.

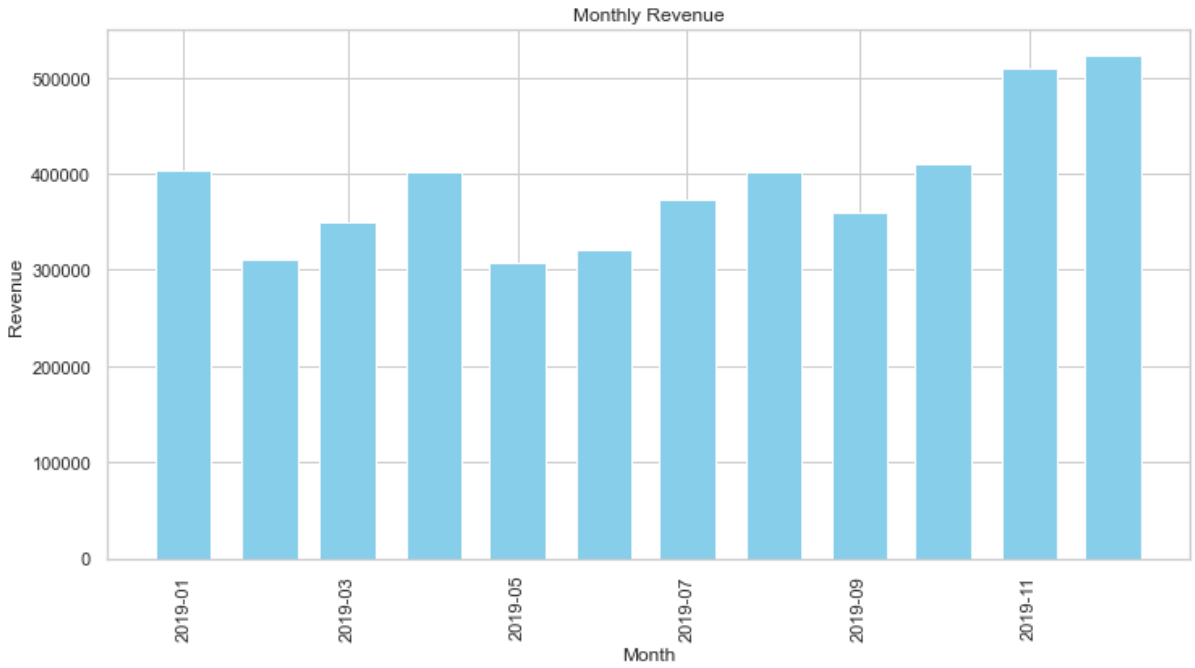


Unveiling Financial Trends: Monthly Revenue Analysis

The meticulous monthly revenue analysis in the Online Sales dataset illuminates crucial financial insights throughout the year. A comprehensive overview of monthly financial performance emerges by transforming transaction dates and calculating revenue through the product of quantity and average price. The resulting analysis depicts discernible trends, with peak revenue in November at $508,942.62 and December at $523,258.19, indicative of heightened sales during the holiday season. The visual representation, encapsulated in the' Monthly Revenue bar chart,' serves as a dynamic tool for decision-makers. It offers a concise yet impactful snapshot of revenue

fluctuations, enabling strategic resource allocation and targeted marketing efforts. This detailed financial analysis empowers the organization to make informed decisions, capitalize on peak sales periods, and foster sustained revenue growth.

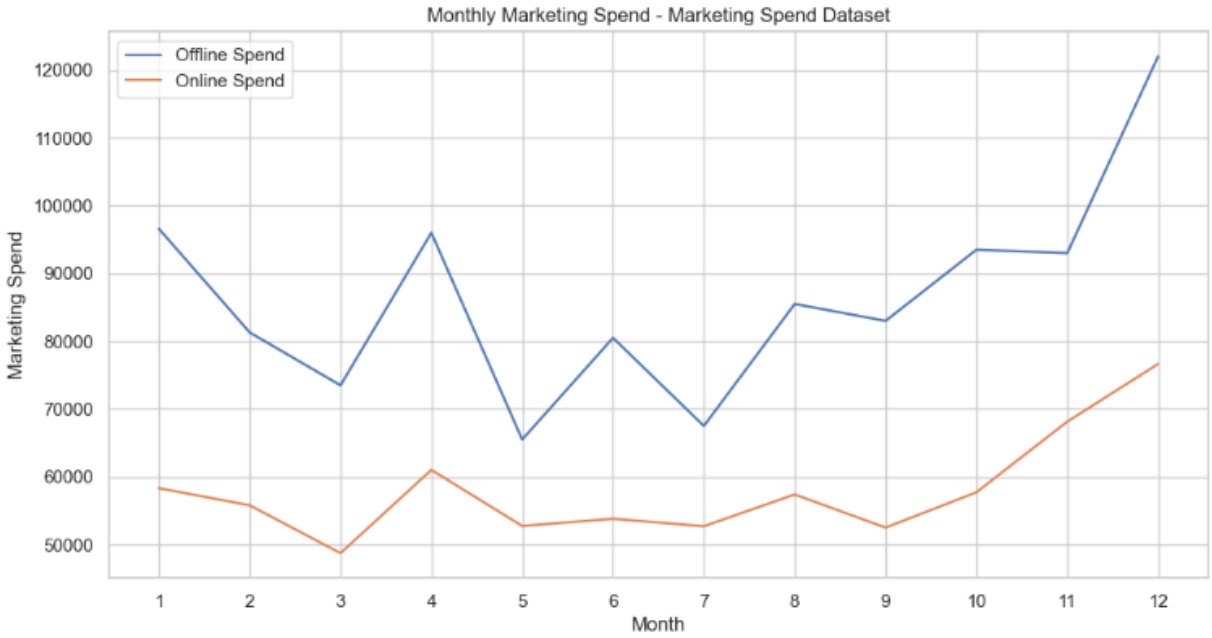
Figure 7: The bar chart shows monthly revenue from the Sales dataset.



Unveiling Financial Spend: Month Marketing Spend

A line graph shows the monthly marketing spend on offline and online channels over a year. The blue line represents offline spending, and the orange line represents online spending. The offline spending fluctuates more than the online spending, and both channels increase their spending towards the end of the year. The offline spending dips in March, May, and August, which could be related to lower demand or budget in those months. Also, offline spending peaks in December, which could be related to higher demand or higher budget during the holiday season. Additionally, offline and online spending move in the same direction for most of the year, except for months 9 to 11, where online spending increases while offline spending decreases. This could indicate that the channels are complementary or substitutable, depending on the situation.

Figure 8: The line graph shows monthly online and offline marketing spending from the Marketing Spend dataset.



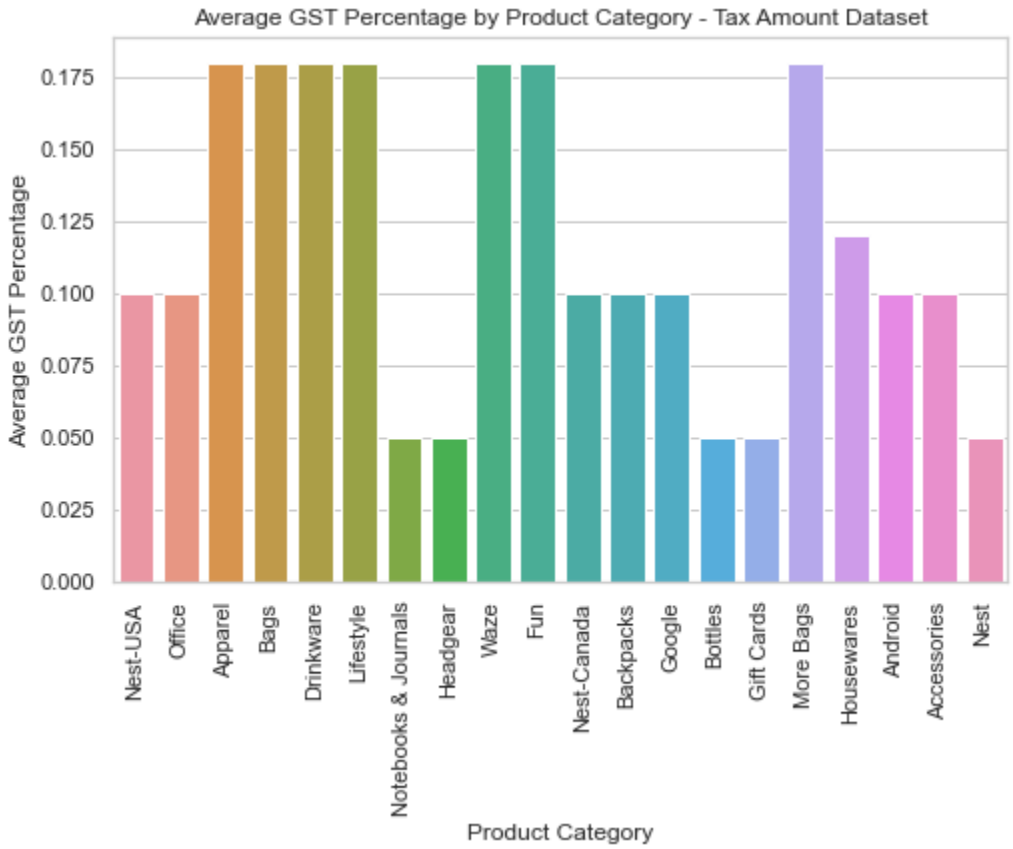
Exploring GST Percentage by Product Category

The bar plot is a visual guide to understanding the distribution of Goods and Services Tax (GST) percentages across diverse product categories within the Tax Amount dataset. The meticulously crafted figure, eight by 5 inches, employs a bar plot with 'Product Category' on the x-axis and 'Average GST Percentage' on the y-axis. Aptly titled 'Average GST Percentage by Product Category - Tax Amount Dataset,' the visualization provides a succinct yet comprehensive overview.

This analysis unveils the variation in GST percentages among different product categories, offering valuable insights for strategic decision-making. By discerning these differences, the organization can refine pricing strategies, enhance financial planning, and ensure compliance with tax regulations tailored to each product category. The rotation of x-axis labels enhances visual

readability, making this analysis a crucial tool for a nuanced understanding of GST dynamics within the diverse landscape of product categories.

Figure 9: The bar chart shows GST percentage distribution from the Tax dataset.



Hypothesis Testing

Investigating Customer Segmentation

Exploring customer segmentation involves a K-Means clustering analysis utilizing the 'Customers_Data.csv' dataset. Key steps include selecting relevant features for clustering, standardizing the features using StandardScaler, and determining the optimal number of clusters using the silhouette score. The silhouette score optimization process resulted in the identification of two clusters as the optimal configuration.

Clustering Process:

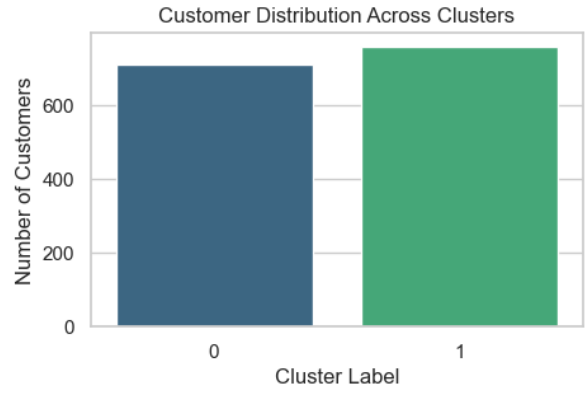
1. **Dataset:** The 'Tenure_Months' feature was selected as a significant factor for clustering.
2. **Standardization:** Features were standardized using StandardScaler to ensure uniform scaling across variables.
3. **Optimal Clusters:** The silhouette score was employed to determine the optimal number of clusters, with two clusters identified as the most suitable configuration.
4. **K-Means Clustering:** The K-Means algorithm was applied with the optimal number of clusters, assigning each customer to a specific cluster.

Insights:

The K-Means clustering revealed two distinct customer segments. The count of customers in each cluster is as follows:

Figure 10: The result of the K-Mean cluster with a count of customers in each cluster from the Customer dataset

```
Count of Customers in Each Cluster:
Cluster_Label
1    758
0    710
Name: count, dtype: int64
```



This segmentation provides a foundation for targeted marketing strategies, personalized customer experiences, and tailored business approaches based on the unique characteristics of each cluster. Understanding customer tenure patterns facilitates strategic decision-making, allowing the organization to optimize engagement strategies and enhance customer satisfaction within each identified segment.

Hypothesis Testing

Customer Segments and Sales Differences:

The hypothesis aimed to discern potential differences in purchasing behavior among distinct customer segments.

Hypothesis:

- Null Hypothesis (H0): There is no significant difference in the mean of items purchased across different customer segments in the population of interest.
- Alternative Hypothesis (Ha): A significant difference exists in the mean of items purchased among distinct customer segments in the population of interest.

ANOVA Result:

Figure 11: The result of the ANOVA test for customer segmentation from the Sales and customer dataset.

```
ANOVA Result for Customer Segmentation:
F_onewayResult(statistic=7.785165506101219, pvalue=0.005269580727491623)
```

The ANOVA test yielded a statistically significant result with an F-statistic of 7.79 and a p-value of 0.005. This indicates that there is sufficient evidence to reject the null hypothesis, suggesting that there is a significant difference in purchasing behavior across customer segments. The insights derived from this analysis can guide strategic decision-making, allowing the organization to tailor marketing strategies and customer experiences based on the identified differences in purchasing behavior among different customer segments.

Marketing Dynamics

Exploring marketing dynamics involves a dual-analysis approach, encompassing correlation and ANOVA analyses. The datasets 'marketing_spend' and 'online_sales' were merged based on the 'Date' column, creating a comprehensive dataset for analysis. Monthly correlations were calculated to discern potential patterns in the relationship between online spending and the quantity of items sold. Subsequently, an ANOVA analysis was conducted to evaluate the impact of marketing effectiveness on monthly sales.

Hypothesis:

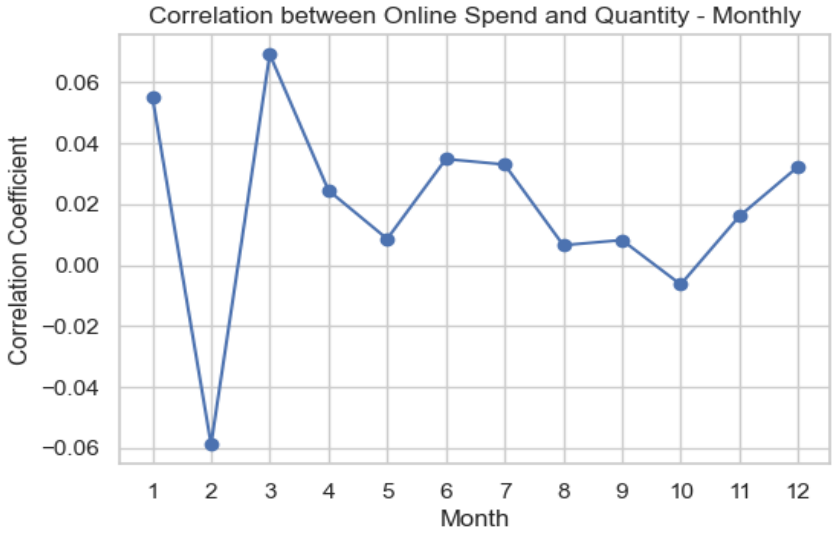
- Null Hypothesis (H0): There is no significant correlation between online spending and item quantity in the population of interest.

- Alternative Hypothesis (Ha): There is a significant correlation between online spending and item quantity in the population of interest.

ANOVA Result:

Figure 12: The correlation between online spending and monthly item quantity.

```
Correlation Results for Each Month:
    Month  Correlation   P-value
0       1     0.055157  0.000436
1       2    -0.058863  0.000738
2       3     0.069173  0.000005
3       4     0.024510  0.114397
4       5     0.008602  0.560930
5       6     0.034807  0.024205
6       7     0.033028  0.016694
7       8     0.006526  0.608867
8       9     0.008214  0.590751
9      10    -0.006354  0.681897
10     11     0.016054  0.312443
11     12     0.032312  0.030161
```



The low p-value suggests a significant difference in monthly sales across different groups, categorized based on marketing effectiveness. Therefore, we have sufficient evidence to reject the null hypothesis (H0). This implies that marketing effectiveness, as measured by online spending, substantially impacts the quantity of items sold every month.

The hypothesis testing for the Monthly Marketing Spend analysis aims to assess whether there is a significant difference between offline and online marketing expenditures. The data was initially processed by converting the 'Date' column to a datetime format and creating a new column for the

month. The marketing spend data was then grouped by month, and each month's offline and online spending was calculated.

The resulting Figure 8 line plot vividly illustrates the monthly offline and online marketing expenditures trends. The null hypothesis (H0) posits no significant difference between offline and online marketing spend, while the alternative hypothesis (Ha) suggests a significant difference.

GST Impact on Product Categories:

Creating a hypothesis for GST's tax amount impact on product categories is important because it allows us to test whether GST and 'Product_Category' have any effect on the sales and profitability of different types of goods and services. A hypothesis is a tentative statement that can be verified or falsified by empirical evidence. The datasets tax_amount and 'online_sales' were merged based on the 'ProductCategory' column, creating a comprehensive dataset for analysis. By creating an ANOVA test, dependent variable sales of 'Quantity' and independent variables 'GST' and 'Product_Category.'

Predictive Model: Logistics Regression

This customer retention prediction analysis implemented a logistic regression model to forecast whether customers would remain engaged with the online sales platform based on their transaction patterns. The dataset underwent preprocessing, including converting transaction dates and creating a 'Cohort_Month' column representing the month of the customer's first purchase. Customers were then labeled as 'Retained' or 'Not Retained' based on the recency of their transactions. Feature engineering involved creating dummy variables for product categories and calculating additional features like 'Total_Spend' and 'Total_Quantity' for each customer. The dataset was split into training and testing sets, and a logistic regression model, incorporated into a pipeline with standard scaling, was trained on the training set. The model demonstrated its predictive capabilities on the

test set, yielding an accuracy score, a confusion matrix, and a classification report to comprehensively evaluate its performance. This analysis provides valuable insights into customer retention dynamics, empowering businesses to address customer engagement and enhance strategic decision-making proactively. The dataset fits the model moderately well, but there is room for improvement for the accuracy of 0.67. We might try to improve the accuracy by adding more predictors, removing irrelevant predictors, or transforming the predictors to improve their relationship with the outcome.

Figure 14: The result of Logistic Regression for customer engagement and strategic decision-making.

```
Accuracy: 0.67
Confusion Matrix:
[[2229 2017]
 [1464 4875]]
Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.52      0.56      4246
           1       0.71      0.77      0.74      6339

    accuracy                           0.67     10585
   macro avg       0.66      0.65      0.65     10585
weighted avg       0.67      0.67      0.67     10585
```

The predictive model for customer retention, employing logistic regression and feature engineering, achieved an accuracy of 0.67 on the test set. The confusion matrix reveals that 2,229 instances were correctly classified as 'Not Retained' (0), while 4,875 instances were correctly classified as 'Retained' (1). However, there were misclassifications, with 2,017 instances of 'Not Retained' being predicted as 'Retained' and 1,464 instances of 'Retained' being predicted as 'Not Retained.' The classification report provides a more detailed assessment, indicating that the model demonstrates higher precision and recall for the 'Retained' class (1) compared to the 'Not Retained'

class (0). With a weighted average F1-score of 0.67, this predictive model showcases moderate effectiveness in identifying and classifying retained and non-retained customers. Further refinement and exploration may enhance the model's predictive power and contribute to strategic decision-making for customer retention efforts.

CONCLUSION

The comprehensive analysis conducted across various facets of the datasets has illuminated critical insights crucial for strategic decision-making in online sales and marketing dynamics. Several key findings have emerged through visualizations, hypothesis testing, and predictive modeling.

Exploring price distributions, coupon engagement, customer tenure, discount percentages, top-selling products, revenue trends, GST variations, and customer segmentation has provided a robust foundation for tailored strategies. These insights enable businesses to refine marketing approaches, optimize customer engagement, and drive revenue growth.

This multifaceted analysis provides a rich understanding of sales dynamics, customer behavior, and marketing effectiveness. Leveraging these insights equips businesses with the tools needed to refine strategies, foster customer loyalty, optimize resources, and ultimately drive sustainable growth in the competitive landscape of online sales. Continued exploration and refinement of these findings can further enhance decision-making processes and contribute to the business's ongoing success.

RECOMMENDATIONS

Based on the extensive analysis conducted across various dimensions of the Online Sales dataset, several actionable recommendations emerge to bolster business strategies:

1. Coupon Optimization:

- **Leverage 'Clicked' Coupons:** Convert clicked coupons ('Not Used') into 'Used' ones by refining the offering or enhancing visibility.
- **Conversion Strategies:** Develop targeted campaigns to bridge the gap between interest ('Clicked') and transactions ('Used').

2. Customer Engagement & Retention:

- **Segmented Strategies:** Tailor marketing efforts based on the identified customer clusters, focusing on different retention strategies for each group.
- **Long-Term Engagement:** Capitalize on insights from customer tenure analysis to foster loyalty and satisfaction over extended periods.

3. Product Performance & Inventory Management:

- **Top Products Focus:** Allocate resources and marketing efforts towards top-performing products identified from sales analysis.
- **Inventory Optimization:** Ensure adequate stock of popular items during peak sales periods to meet demand.

4. Financial Planning & Pricing:

- **GST Variation Utilization:** To maintain competitiveness, adjust pricing strategies based on GST variations among product categories.
- **Revenue Insights:** Utilize monthly revenue trends to plan resource allocation, especially during peak sales months.

5. Marketing Strategies:

- **Correlation of Marketing Spend:** Increase focus on online marketing spend as it correlates significantly with item quantity sold.
- **Offline vs. Online Spend:** Evaluate strategies to balance offline and online marketing expenditures, considering their impact on sales.

6. Customer Retention Predictions:

- **Refinement of Predictive Model:** Further refine the logistic regression model for customer retention to enhance its predictive accuracy.
- **Focused Retention Strategies:** Deploy targeted strategies based on the model's predictions to retain customers more effectively.

7. Continuous Improvement:

- **Iterative Analysis:** Regularly revisit and update analyses to adapt to changing market dynamics and customer behaviors.
- **Feedback Utilization:** Integrate customer feedback into strategies for continuous improvement.

8. Holistic Decision-Making:

- **Cross-functional Collaboration:** Foster collaboration among departments to ensure insights translate into actionable organizational strategies.
- **Data-Driven Decisions:** Continue prioritizing data-driven decision-making for sustained growth and adaptability.

Implementing these recommendations based on the comprehensive analysis conducted across various facets of the Online Sales dataset will enable the organization to refine strategies, enhance customer satisfaction, optimize resources, and drive sustained growth in the competitive online sales landscape.

References

Alsmadi, A. A., Shuhaiber, A., Al-Okail, M., Al-Gasaymeh, A. & Alrawashdeh, N. (2023). *Big data analytics and innovation in e-commerce: current insights and future directions.* https://link.springer.com/article/10.1057/s41264-023-00235-7

Akter, S. & Wamba, S. F. (2016). *Big data analytics in E-commerce: a systematic review and agenda for future research*. https://link.springer.com/article/10.1007/s12525-016-0219-0

Bonacchi, M., & Perego, P. (2018). Customer analytics: Definitions, measurement, and Models. *Customer Accounting*, 13–35. https://doi.org/10.1007/978-3-030-01971-6_2

Britt, P. (2023). *3 Ways Al-Powered Predictive Analytics Are Transforming E-commerce.* https://www.cmswire.com/analytics/3-ways-ai-powered-predictive-analytics-are-transforming-ecommerce/

Court, D. (2019). *How US shopping behavior is changing*. Mckinsey & company. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Marketing%20and%20Sales/Our%20Insights/The%20great%20consumer%20shift/ten-charts-show-how-us-shopping-behavior-is-changing.pdf

Fantini, F., & Narayandas, D. (2023). *Analytics for marketers*. Harvard Business Review. https://hbr.org/2023/05/analytics-for-marketers

Halan, D., & Singh, E. P. (2023). Enemies to frenemies: Coopetition between online and offline retailers amidst crises. *International Journal of Retail & Distribution Management*, *51*(4), 425–443. https://doi.org/10.1108/ijrdm-06-2022-0208

Homburg, C., Theel, M., & Hohenberg, S. (2020). Marketing excellence: Nature, measurement, and investor valuations. *Journal of Marketing*, *84*(4), 1–22. https://doi.org/10.1177/0022242920925517

Kashyap, A. (2019). *Predictive marketing analytics using BigQuery ML machine learning templates.* https://cloud.google.com/blog/products/data-analytics/predictive-marketing-analytics-using-bigquery-ml-machine-learning-templates

Krautz, C., & Hoffmann, S. (2017). The tenure-based customer retention model: A Cross-Cultural Validation. *Journal of International Marketing*, *25*(3), 83–106. https://doi.org/10.1509/jim.16.0040

Kumar, V., & Reinartz, W. (2018). Customer analytics part I. *Springer Texts in Business and Economics*, 79–99. https://doi.org/10.1007/978-3-662-55381-7_5

Li, X. (2020). Business analytics in e-commerce: A literature review. *Journal of Industrial Integration and Management*, *06*(01), 31–52. https://doi.org/10.1142/s2424862220500207

Mela, C. F. (2018). *Why Marketing Analytics has not lived up to its promise*. Harvard Business Review. https://hbr.org/2018/05/why-marketing-analytics-hasnt-lived-up-to-its-promise

Mills, P., & Zamudio, C. (2018). Scanning for discounts: Examining the redemption of competing mobile coupons. *Journal of the Academy of Marketing Science*, *46*(5), 964–982. https://doi.org/10.1007/s11747-018-0592-7

Purnomo, Y. J. (2023). Digital marketing strategy to increase sales conversion on e-commerce platforms. *Journal of Contemporary Administration and Management (ADMAN)*, *1*(2), 54–62. https://doi.org/10.61100/adman.v1i2.23

Rosário, A., & Raimundo, R. (2021). Consumer marketing strategy and e-commerce in the last decade: A literature review. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(7), 3003–3024. https://doi.org/10.3390/jtaer16070164

Stocchi, L., Michaelidou, N., Pourazad, N., & Micevski, M. (2018). The rules of engagement: How to motivate consumers to engage with branded mobile apps. *Journal of Marketing Management*, *34*(13–14), 1196–1226. https://doi.org/10.1080/0267257x.2018.1544167

Tolstoy, D., Nordman, E. R., & Vu, U. (2022). The indirect effect of online marketing capabilities on the international performance of e-commerce SMEs. *International Business Review*, *31*(3), 101946. https://doi.org/10.1016/j.ibusrev.2021.101946

Wu, Y.-L., & Li, E. Y. (2018). Marketing mix, customer value, and customer loyalty in Social Commerce. *Internet Research*, *28*(1), 74–104. https://doi.org/10.1108/intr-08-2016-0250

Yeo, J. (2023). An empirical analysis of shopping basket similarities across consumers. *Applied Economics*, *56*(1), 98–116. https://doi.org/10.1080/00036846.2023.2166673

Appendix

Description of Dataset

This dataset compilation comprehensively views customer transactions, demographics, marketing expenditures, discount coupon utilization, and tax implications across various product categories within an e-commerce platform. The dataset provides a rich source for in-depth analysis to extract insights crucial for strategic decision-making in marketing, sales, and customer engagement initiatives. The dataset consists of several components. The first segment is customers' online Sales data, containing transactional data from January 1st, 2019, to December 31st, 2019. It consists of the following variables:

Table 1: Online Sales Data description of listings, variables names, and variables' type.

Variable	**Type of Variable**	**Description**
CustomerID	Nominal	Unique identifier for customers.
Transaction_ID	Numerical (Discrete)	Unique transaction identifier.
Transaction_Date	Numerical (Temporal)	Date of the transaction.
Product_SKU	Nominal	Unique ID for products sold.
Product_Description	Nominal	Description of the product.
Product_Category	Nominal	Categorization of products.
Quantity	Numerical (Discrete)	Number of items purchased in a transaction.
Avg_Price	Numerical (Continuous)	Average price per quantity.
Delivery_Charges	Numerical (Continuous)	The cost associated with delivery.
Coupon_Status	Nominal	Indicates if a discount coupon was applied.

The second segment of data involves **Customers Data**. It contains customer demographics and has the following variables:

Table 2: Customer Data description of listings, variables name, and variables' type.

Variable	Type of Variable	Description
CustomerID	Nominal	Unique customer identifier.
Gender	Nominal	Gender of the customer.
Location	Nominal	Geographic location of customers.
Tenure_Months	Numerical (Discrete)	Duration of the customer's engagement in months.

The following data segment is that of **Discount Coupons,** and it records discount coupons applied to different product categories in various months. It contains the following variables:

Table 3: Discount Coupons Data description of listings, variables name, and variables' type.

Variable	Type of Variable	Description
Month	Nominal	The month when the discount coupon was applied.
Product_Category	Nominal	Category to which the coupon applies.
Coupon_Code	Nominal	Unique code for each coupon.
Discount_pct	Numerical (Continuous)	Percentage of discount offered.

The following data segment consists of **Marketing Spend variables,** which record daily marketing expenditures on offline and online channels. The following variables are part of it:

Table 4: Marketing Spend Data description of listings, variables name, and variables' type.

Variable	Type of Variable	Description
Date	Temporal	Date of marketing spend.
Offline_Spend	Numerical (Continuous)	Expenditure on offline marketing channels.
Online_Spend	Numerical (Continuous)	Expenditure on online marketing channels.

Lastly, there is a segment of **data on tax Amount** which provides GST details for different product categories. It contains the following variables:

Table 5: Tax Amount Data description of listings, variables names, and variables' type.

Variable	Type of Variable	Description
Product_Category	Nominal	Product category.

GST	Numerical (Continuous)	Percentage of Goods and Services Tax applied to each category.