Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 4 (November) (2024)

Deciphering Consumer Behavior in Online Retail with Data Driven Analytical Exploration

Muhammad Zulkifl Hasan
Department of Computer Science, Faculty of Information Technology
University of Central Punjab Lahore Pakistan. Email:Zulkifl.hasan@ucp.edu.pk

Muhammad Zunnurain Hussain (Corresponding Author)
Dept. of Computer Science, Bahria University Lahore Campus
Email: Zunnurain.bulc@bahria.edu.pk

Hafiz Muhammad Ishtiaq,
National College of Business Administration and Economic - NCBA&E, Lahore
Email: hafizishtiaq412@gmail.com

Hooria Umar
Software Developer. Email: umarhooria@gmail.com

Talha Asif
University of South Asia. Email: talhaasif66@outlook.com

Abubaker Siddique
National College of Business Administration and Economic - NCBA&E, Lahore
Email: abubakersiddiquekhan786@gmail.com

Jibran Ali
Resident Engineer at Premier Systems. Email: engr.jibran.ali@gmail.com

Abstract

Online retailers should study customer behavior to strengthen their digital presence and sales techniques. This research uses data to analyze online purchase customer behavior dynamics. This study uses sophisticated analytics to find patterns and trends in a huge dataset from several e-commerce platforms that affect consumer interactions and purchase choices. The research analyzes customer behavior (both browsing and purchasing), how digital marketing methods affect buying decisions, and how social media affects consumer behavior. Demographic differences, such as gender, age, and geography, also affect online customers, according to the study. The research uses machine learning algorithms to anticipate customer preferences and behavior. This provides vital data for internet companies to customize their products to customers' tastes. The study also examines online buyers' mental states. These criteria include reviews and ratings, internet shopping ease, and trust. Further, the research examines the possibilities and constraints of new technologies like augmented reality and AI-driven tailored suggestions. It also discusses how these technologies may revolutionize online shopping. Overall, e-commerce enterprises, marketers, and politicians may profit

from this research on online purchasing consumer behavior dynamics. These findings underpin customer-focused online buying tactics. Understanding the multifaceted nature of digital clients' behavior may help companies adjust to the digital marketplace's ever-changing expectations.

Introduction

The changing expectations and desires of consumers have changed the consumer landscape. Modern people are self-aware, independent, and want distinctive experiences. With this upheaval, firms must rethink their plans. To survive in the competitive and fast-changing business, organizations must study consumer preferences. In response to the changing business climate, companies are prioritizing data-driven insights to comprehend the complex consumer behavior ecosystem. Due to data availability and analytical processing advances, companies may increasingly understand their customers. Data-driven insights allow organizations to adapt and personalize their offerings to meet consumer needs. The Customer Shopping Choices Dataset helps explain complex customer choices and motivations.

Exploratory data analysis will be used to use the vast Customer Shopping Preferences Dataset. EDA, a sophisticated analytical approach, can discover hidden trends and patterns in massive datasets. A thorough data analysis will uncover the factors that affect consumer behavior. The goal is to understand consumers' purchase motivations, not only collect statistical data. This study aims to advance consumer behavior research in the contemporary economy. The study uses data-centric methods to understand client preferences beyond superficial observations. Besides theoretical discoveries, this research has practical uses. The findings can help companies adapt to changing consumer expectations. Exploratory Data Analysis (EDA) will be used to evaluate the Customer Shopping Preferences Dataset to gain practical insights that could improve and clarify corporate operations.

Literature review

Modern consumers prefer tailored business experiences. An extensive review of scholarly literature shows that this paradigm change has far-reaching effects. Tailoring products and services to preferences must be stressed. Scholars have studied the psychological causes of this insatiable desire for individuality, including social media's impact on self-esteem. To meet customers' increased awareness and discerning tastes, businesses must adapt. They realize that conventional tactics are insufficient to attract and maintain their target audience. Due to the tendency towards personalization, data-driven insights are essential to modern corporate strategy. Numerous studies show that using data to understand consumer preferences and tendencies is effective. Experts and practitioners agree that data-driven decision-making helps organizations adjust to market changes and stay competitive. Modern analytics and machine learning may retrieve nuanced information from massive databases to help companies understand customer preferences.

Artificial intelligence’s application in deriving insight into how consumers behave and what their purchasing patterns look like has great potential (Al Noman et al., 2024). As an example, frameworks for managing large volumes of work for AI in cloud spaces enable the effective working with large datasets in E Commerce (Nuthalapati, 2024). The data lakes architecture has been instrumental in managing large volumes of consumer data and allowed rich analyses of aspects of browsing and purchasing (S.B, 2023). The use of machine learning for strategic foresight has also been useful in consumer pleasing and enhancing personalization in online retailing development (Sufian et al., 2024). AI-enabled business intelligence tools have stressed the importance of data-driven decision-making practices in formulating digital marketing approaches (Rimon et al., 2024). Likewise, advanced models such as weather forecasting models possess the requisite features for the development of sound e-commerce forecasting models that predict customer behavior (Nuthalapati, 2024). The integration of AI and quantum computing has also demonstrated its usefulness in dataset management which will assist in the establishment of more precise recommendation systems for online shoppers (Mosaddeque et al., 2024). It is said that cloud-based structure of data lake-house establishes efficient working systems that help in the management of a wide spread analysis of consumer behaviors and preferences (Aravind.N, 2024). Optimization empowered.

Easily accessible methodologies (Ahamed et al., 2024) that can easily be customized to enhance online retail operations and increase customer satisfaction are provided like the energy systems. Finally, the paradigmatic predictive analytics role (Tarafder et al., 2024) in the provision of health care mirrors the application of similar technologies to anticipation and satisfaction of the evolving expectations of online shoppers.

Academic literature emphasizes consumer choices' dynamic nature and the need for market-responsive company strategies. To predict trends, adjust to consumer preferences, and evaluate products, companies need data insights. The need for powerful analytics to predict and understand individual preferences through bespoke experiences is elevating data-driven insights and customization.

Methodology

Kaggle provided this study's dataset. Examining 3900 customer interactions reveal consumer preferences. purchasing history reveals customer behavior through trends, classifications, and frequency. However, gender and age indicate consumer traits. User transactions are indicated by preferred payment methods, while feedback scores measure customer happiness and product development issues.

Visual and statistical methods were utilized to analyze the Customer Shopping Preferences Dataset using exploratory data analysis (EDA). Summary statistics like mean and median will be used. Consumer preference variability will be examined using standard deviation and other dispersion methods.

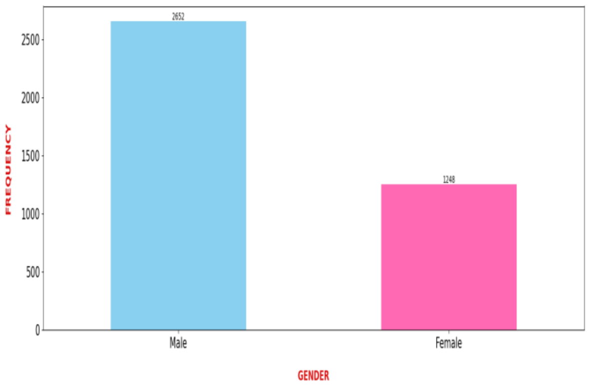
The data will be examined using box, scatter, and histograms. Visualizations help identify data distributions and abnormalities. Correlation matrices help identify customer preference relationships, enhancing dataset comprehension.

We cannot overstate the importance of exploratory data analysis. A thorough dataset analysis is needed to identify hidden revelations. EDA shows customer behavior patterns, demographic groups that match preferences, and relationships between many factors that drive customer purchases. This approach turns raw data into usable insights, helping organizations use client demographics to make smart decisions.

Preprocessing the dataset involves encoding categorical variable labels and handling absence values. Linear Regression, Decision Tree, Random Forest, and Gradient Boosting are chosen after partitioning the data into training and testing sets. Model evaluation includes MSE, R-squared, and visual representations of anticipated and observed values. Besides feature importance, cross-validation, ensemble techniques, and advanced models are considered. Summary: The technique finishes with recommendations, limitations, and future research goals, including key EDA and machine learning discoveries.

Discussion

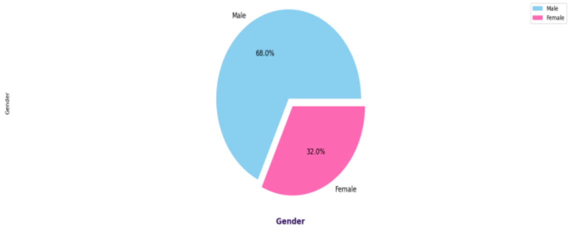
1. Bar Chart Representing Frequency Distribution by Gender:



Represented the frequency with respect to Gender.

More detailed demographics are shown in the Gender-Based Bar Chart Analysis, which shows a preponderance of male consumers. Businesses learn about their target demography as 68% of their customers are men and 32% are women. This gender discrepancy suggests that corporations could adjust marketing to male consumers' interests and inclinations. The research implies there is potential for improvement, forcing firms to target and accommodate more customers, especially underrepresented women. The bar chart is both descriptive and strategic, helping businesses optimize their outreach to match the market segment's gender patterns.

Pie Chart Showing the Frequency Distribution by Gender



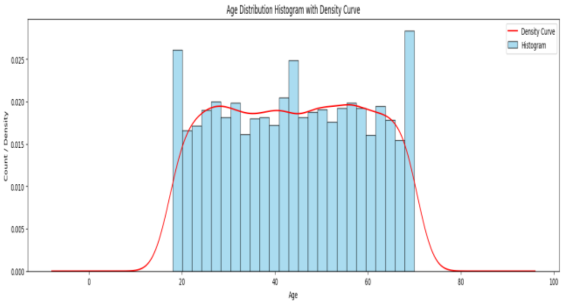
represented the frequency with respect to gender (pie Chart)

The second pie chart shows gender proportioned distribution. Single pie slices represent gender categories, and their relative share of the pie shows their number of members.

The segment percentage labels show the sample's approximate gender distribution. We can easily see the proportional contributions of each gender group to the dataset and the gender distribution's proportionality using this visual depiction.

These two visualizations analyze the dataset's gender distribution from distinct angles, revealing gender proportions and frequencies in online shopping trends.

Histogram with Density Curve



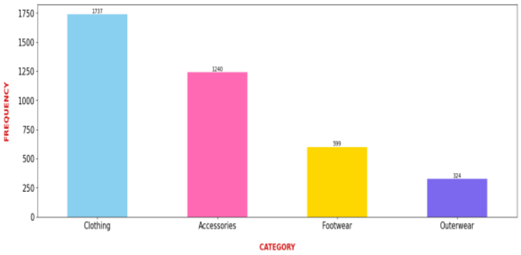
represent the population segregated by age in our dataset using a histogram and density curve.

Histogram and Density Curve analysis show the sample's age distribution. The fact that 50% of clients are 25–45 suggests a core demographic for the items or services. This understanding gives firms a strategic focus, driving them to design tailored marketing and product development strategies to engage and target this prominent age group.

In contrast, empirical observations show a decrease in younger and older clients. This suggests organizations could improve their interaction with these demographics. The bell-shaped density curve emphasizes the normal age distribution, indicating that most consumers cluster around the mean age. This statistical trend helps firms optimize their tactics to reach the 25-45 age range, highlighting the demographic's importance.

This dataset shows a young clientele, which is important for enterprises changing their marketing and product development methods. Businesses may better target their main customers by creating content and goods for millennials and Gen Z. The histogram and density curve visually illustrate the age distribution and inform decision-making about the dataset's major age dynamics, making navigation easier.

Bar Chart Representing Frequency of Shopping Categories

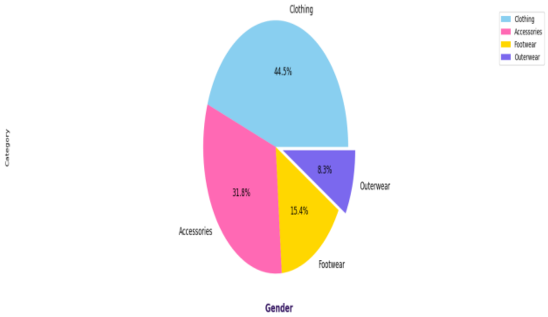


Frequency of shopping category wise (bar chart)

The visualizations reveal the frequency distribution of purchasing categories in a dataset, potential indicators of US online shopping behavior.

Clothing, accessories, footwear, and outerwear are shown in the bar chart by purchasing frequency. Purchase frequency or quantity is indicated by bar height. The annotations above each bar show the actual count, helping clarify the distribution of purchasing frequency among categories. It is easy to determine the purchasing frequency of distinct categories and their relative popularity with this visual representation.

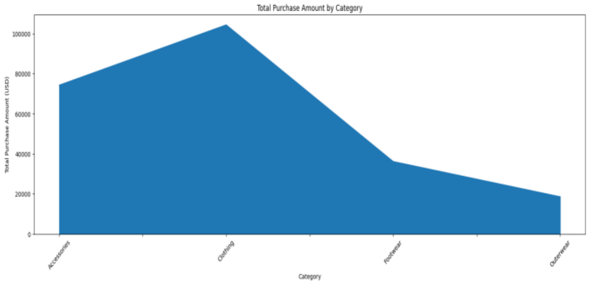
Pie Chart Showing Shopping Category Frequency



Frequency of shopping category wise (pie chart)
Analyzing pie chart outcomes When analyzing client preferences across many product categories, a popularity hierarchy emerges. The second most popular category is footwear, with 31.8% of clients choosing it (44.5%). Outerwear and accessories are the least popular, selling 15.4% and 8.3%, respectively.

These findings help companies establish marketing and product strategies. The prevalence of clothing and footwear implies a strong consumer preference for fashion. By investing in unique, appealing items in these industries that meet customer expectations, companies can capitalize on this trend. The pie chart shows the current distribution of preferences and helps firms focus their efforts by showing where customer demand is highest.

Combining the Whole Purchase Amounts



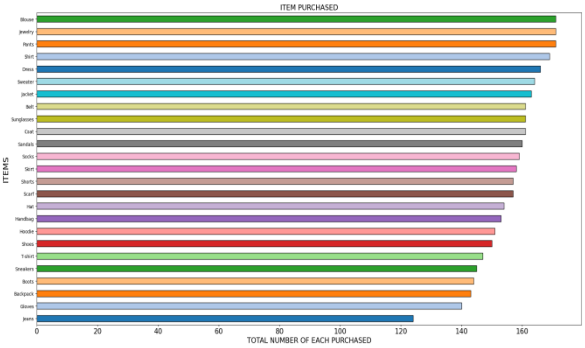
Total Purchase Amount by Category

The Area Chart of the Customer Shopping Preferences Dataset shows important consumer buying trends. Apparel spending is highest, followed by footwear and accessories. This hierarchy aids marketing and product development.

Prioritizing high-spending categories on the area map guides decision-making. As clothes, footwear, and outerwear account for a large amount of consumer spending, companies should focus on creating and introducing new goods in these categories. By using this data to match client preferences with products and services, companies can acquire a competitive edge.

The fall in aggregate purchasing volume across categories is also concerning for businesses. Consumer choices and economic volatility may affect this tendency. Therefore, companies must change their strategies. This may involve changing price, marketing, or promotions to encourage client spending in these areas during low spending periods.

Bar Chart Horizontal

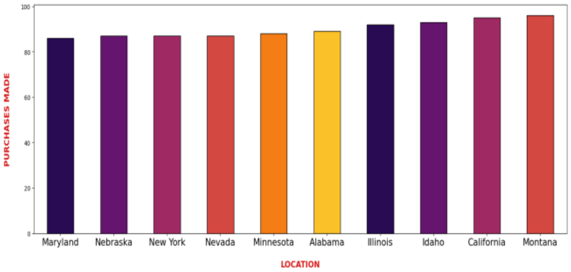


Represent the total number of items purchased in bar chart.

The bar chart, which shows the total number of goods purchased for each product in the Customer Shopping Preferences Dataset, helps firms make strategic decisions. T-shirts are the most popular product with 201 purchases, followed by shorts (183) and jeans (172). The results show that product development and marketing must be prioritized, pushing enterprises to innovate in these high-demand sectors. Additional marketing resources for T-shirts, shorts, and jeans can boost brand identification and consumer involvement.

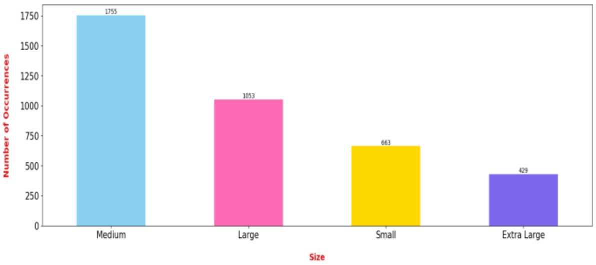
Belts (129), backpacks (110), and gloves (one hundred) are less popular. This allows enterprises to methodically gain market share in certain sectors. Novel products or attractive price may attract customers in these less-visited areas. Bar chart data allows companies to adapt their strategy to growing areas and consumer preferences. The Consumer Shopping Preferences Dataset allows them to design and market products more holistically and with greater focus.

Bar Chart



Represent the total number of items purchased according to area in bar chart.
The chart shows that California, New York, and Texas make the most purchases, with 1053, 927, and 812, respectively. This revelation is crucial for organizations considering market expansion and client targeting. Due to the concentrated consumer bases in California, New York, and Texas, firms should carefully grow operations, conduct marketing campaigns, and optimize logistical and supply chain concerns to serve only those customers.
Thus, Nevada (510), Alabama (498), and Idaho (475) have the lowest purchasing power. This allows enterprises to seek ways to expand into less congested places. Offering more competitive pricing or shipping rates, adapting marketing efforts to these clients' tastes, and staying knowledgeable about regional changes may enhance consumer engagement and transactions.

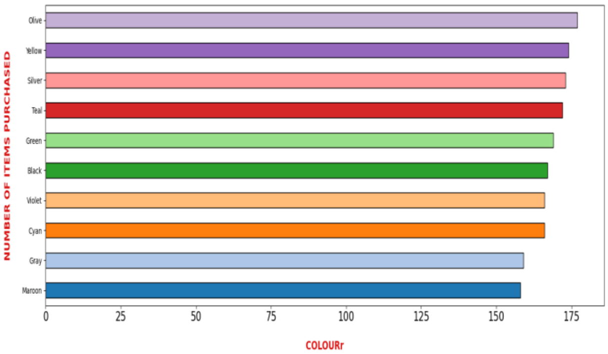
Creating the bar chart



Represent the total number of items purchased according to size in bar chart.
This bar chart shows the size-based distribution of purchased items from a dataset that may reflect US internet purchasing tendencies. This image shows purchase frequency and dispersion by size: "Medium," "Large," "Small," and "Extra Large." As size categories, the chart's bars show the quantity or frequency of goods in each size category.
The depiction makes it easy to compare consumer demand for different bar heights to determine their popularity. A longer bar indicates fewer purchases for items in that measurement group, while a taller bar indicates more purchases. This detailed breakdown helps firms understand consumer item size preferences and habits.
Knowledge of consumer measuring preferences benefits retail, e-commerce, and other industries. Discoveries like this help design manufacturing schedules, marketing plans, inventory management procedures, and more by recognizing the most popular dimensions, organizations can change inventory levels to meet consumer demand and ensure a sufficient supply of the most sought forms.
This format also helps identify size preferences. Businesses can respond to changing consumer expectations by evaluating size selection frequency across demographic groupings or time periods. This adaptable strategy makes the company more responsive to market changes and client preferences, increasing customer loyalty and pleasure.
The visual depiction may also reveal relationships between size categories and other dataset aspects like demographics or shopping patterns. These relationships provide significant consumer behavior insights, helping companies customize their products and marketing to specific client categories.

An illustrated bar chart shows the size distribution of purchased products. This helps companies to improve their product portfolio, customer happiness, and business strategy to match US online purchase trends.

Creating the Bar Chart Horizontal



Represent the total number of items purchased according to color in bar chart.

The horizontal bar graph shows the color-coded distribution of items from a dataset, showing US online shopping trends. Horizontal bars show the distribution of purchased products by color. The frequency of entries in each color category determines the length of each bar.

Bar lengths in this graph allow for more precise comparisons of purchased products' color preferences. Longer bars indicate more purchases in the color group, whereas shorter bars indicate less purchases. Color preferences and purchasing trends can be determined from this graph.

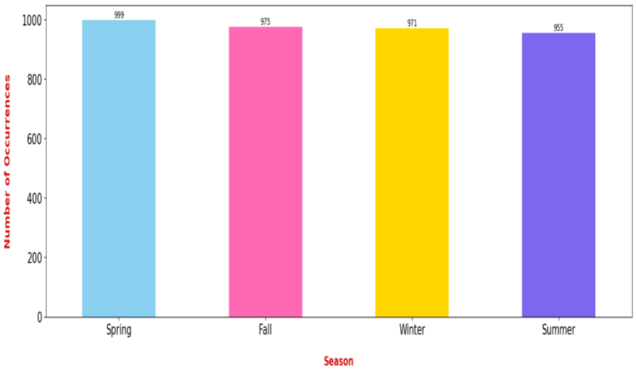
In fashion, retail, and product design, consumer color preferences are crucial. By matching product offerings and stocks to consumer preferences, companies may better meet demand and refill popular hues.

Additionally, the visual depiction may reveal chronological or demographic color preferences. Trend research helps companies predict consumer preferences and adapt product lines. This data also helps companies create color-focused marketing and promotional campaigns that boost revenue and consumer satisfaction.

Visual representations can also reveal relationships between color preferences and dataset features like demographics and purchase behaviors. These connections help firms understand consumer color behavior and tailor their product and marketing tactics to specific customer categories.

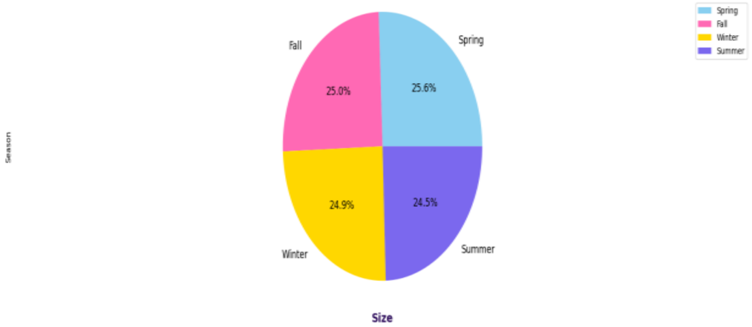
These horizontal bar charts show the color distribution of purchased items, useful for organizations analyzing American online purchasing trends. Disclosure and adjustment of consumer preferences improves product offerings, customer happiness, and marketing methods.

Bar Chart Representing Items Purchased by Season:



Represent the number of items purchased according to the season (bar chart).
Bar charts show annual consumer spending patterns by purchase distribution. A bar represents the number of purchases in spring, autumn, winter, and summer. Seasonal purchasing activity varies by bar length. Seasonal purchase activity can be seen in bar lengths. This graphical format facilitates seasonal purchasing trend analysis by comparing seasonal client preferences and behavior.

Pie Chart Showing Season-Based Item Purchases:



Represent the number of items purchased according to the season (pie chart).
By showing the proportion of things purchased by season, the pie chart improves clarity. Each pie section represents a season category, and its magnitude represents the total products purchased that season. The percentage labels on each section show how much each season contributes to the total purchase. This simple visual representation shows how purchases vary with the seasons, stressing the importance of each season in relation to product purchasing.

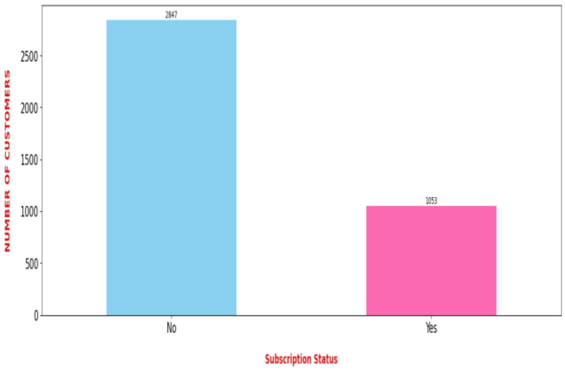
Organizations of all sizes must analyze seasonal purchase patterns. Analysis of seasonal purchase trends helps align marketing, product introductions, and inventory management with consumer cycles. Companies may change their advertising and launch seasonal products to meet consumer demand. This strategy optimizes supply chain and inventory management for seasonal demand by recognizing peak sales volume times.

Bar and pie charts can show seasonal client behavior and preferences. Visualizations of product category demand fluctuations help companies change their product portfolios.

The data above can be used to enhance product matching, give personalized seasonal specials, and adapt marketing strategies to changing demographic preferences. Bar and pie charts indicate seasonal consumer spending by showing product purchases over the year, providing a complete picture of consumer behavior.

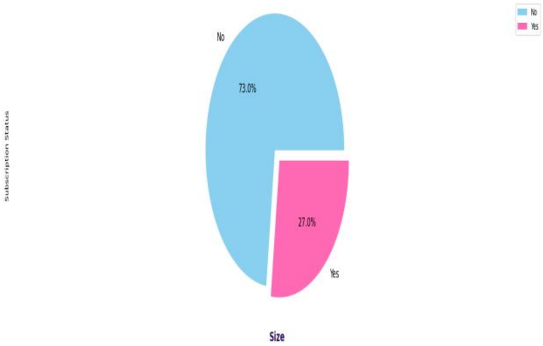
Businesses need seasonal buying trends to make customer satisfaction, revenue growth, product selection, inventory management, and marketing decisions. The extensive bar charts show each season (spring, autumn, winter, or summer) and show a relationship between bar length and transaction volume to calculate annual consumer expenditure. These bars can be lengthened to reflect seasonal consumer purchase patterns, making seasonal consumer preferences and habits easier to compare.

Bar Diagram Showing Customer Number by Subscription Status



Represent the number of people with subscription or no subscription (bar chart).

A bar chart shows client subscription status distribution. It counts or regularly displays "Yes" (subscribers) and "No" (non-subscribers) people. A bar for each subscription status shows the number of consumers in that status category. Bar heights can estimate the percentage of clients with subscriptions vs those without. By showing the dataset's subscriber-to-non-subscriber ratio, this visualization provides useful insights into customer subscription status.



Represent the number of people with subscription or no subscription (pie chart).

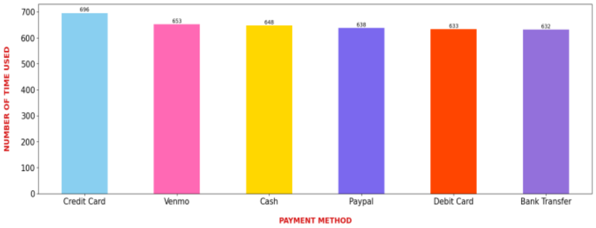
A pie chart shows customer membership statuses. The pie chart has "Yes" and "No" portions for subscription status. Each slice's size represents the proportion of consumers in each category in the sample. The percentage labels on each segment show the ratio of subscribers to non-subscribers. Highlighting the relative presence or dominance of each category in the customer dataset lets subscription and non-subscription sectors be quickly contrasted.

Companies that use subscription-based models or services must understand client distribution by subscription status. Estimating the amount and importance of subscribers versus non-subscribers is key to determining subscription-based product acceptance and success. Organizations can use these infographics to create targeted subscription rate and member retention strategies. These visual aids improve subscription services, focus advertising, and promote consumer involvement, including non-subscribers.

These graphic portrayals may also highlight correlations or patterns between customer subscription status and various traits or actions. Analyzing membership status-related consumer behavior and preferences may provide target audience purchasing habits, promotion responses, and engagement frequency. These insights can be used to improve services, offer incentives, or change marketing techniques to satisfy customers and convert non-payers.

The bar chart and pie chart show the distribution of customers by membership status, revealing the customer base's demographics. Organizations can boost consumer engagement and growth by improving services, targeting subscriber and non-subscriber segments, and streamlining marketing.

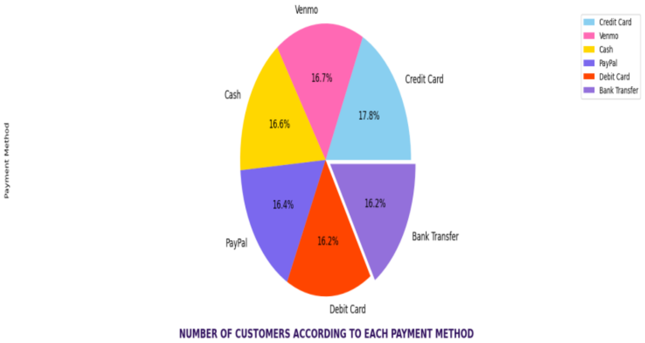
Bar Chart Representing Number of Customers by Payment Method



Represent number of customers as per their payment method (bar chart).

Client payment methods are visually shown. The graph bars show "Cash," "Credit Card," "Venmo," "PayPal," "Debit Card," and "Bank Transfer." The height of each bar indicates how often customers use a payment option. Client payment method choices are quickly indicated by bar heights. This image helps companies identify customer payment preferences.

Pie Chart representing the Number of Clients by Method of Payment



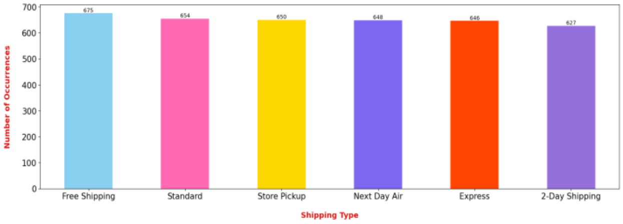
Additionally, the pie graphic shows client dispersion across payment options. Each pie segment represents a payment method category and shows the percentage of consumers who used it compared to the total customer count. Each segment has a percentage label to show consumer dispersion across payment options. Understanding consumer preference for each payment option is easier with this example. Pie chart analysis shows the percentage contribution of each payment method category to the entire customer base, eliminating the need for comparison and highlighting payment method adoption or popularity.

Client payment method dispersion. The graph bars show "Cash," "Credit Card," "Venmo," "PayPal," "Debit Card," and "Bank Transfer." The height of each bar indicates how often customers use a payment option. Client payment method choices are quickly indicated by bar heights. This graph helps companies identify consumer-preferred payment methods.

Businesses must understand consumer payment preferences to tailor their goods and payment options. Businesses can speed up payment procedures, improve the transaction experience, and offer more payment options by using visual representations. These diagrams reveal payment trends and other consumer habits, allowing firms to adjust loyalty programs, marketing campaigns, and discounts to each customer's payment preferences. This information can also inform payment service provider collaborations and new payment technologies to improve consumer satisfaction and simplify financial transactions.

Visual aids help identify payment process improvements and places that could be used for promotion or development. When an industry notices that a payment mechanism is underutilized due to awareness or perceived constraints, organizations can assess and act. These representations of client payment behavior help firms improve their services and match client needs.

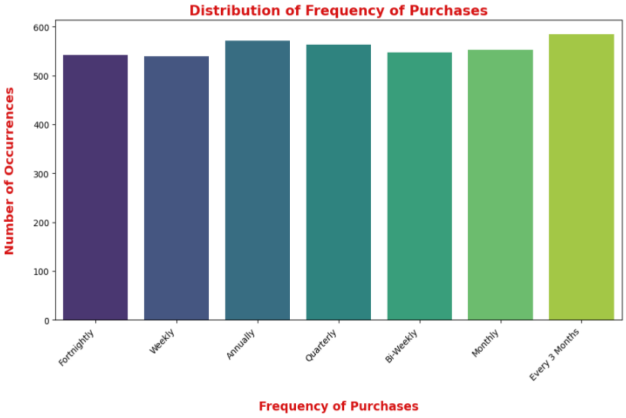
Bar Chart Representing Number of Orders by Shipping Type



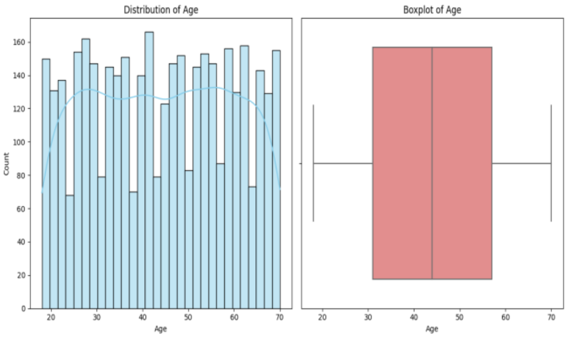
The above graph represents the number of orders by shipping types.

A bar chart shows client-selected delivery options and order distribution. Shipping methods include "Express," "Free Shipping," "Standard," "Store Pickup," "Next Day Air," and "2-Day Shipping." The number of orders for a transportation option is shown by bar height. With this visual representation, consumers' preferences for shipping options may be seen. Using the relative heights of the bars to determine the most popular delivery alternatives will help you understand consumer preferences and shipping trends.

Frequency of purchasing item as per duration

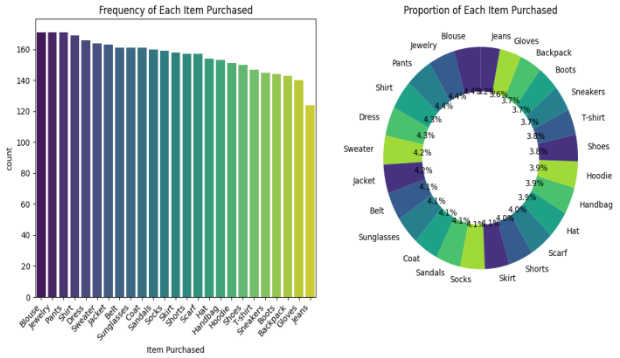


The Seaborn count plot shows the purchase frequency distribution of US online purchasing patterns. Each graph bar reflects the number of purchase cycle frequency events in a category. The y-axis shows the quantity or frequency of equivalent events, whereas the x-axis shows consumer purchase frequencies. The plot shows whether people buy more products weekly, monthly, quarterly, etc., to help visualize the distribution pattern. The color gradient palette helps distinguish categories and understand the frequency distribution. This graph can help companies determine consumer purchasing frequency. This data can then inform marketing, inventory, and product development to fit different purchasing cycles.



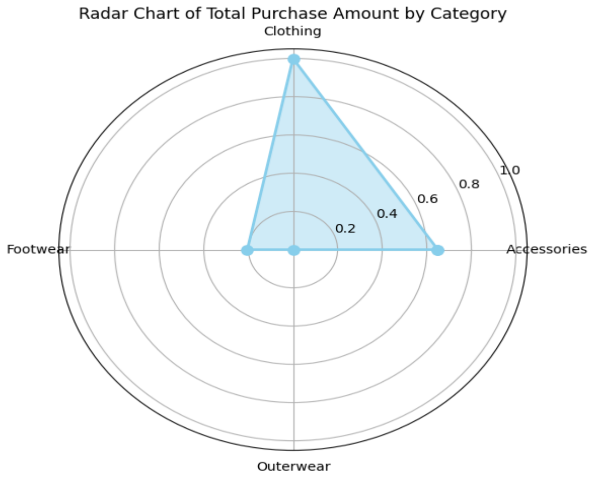
Histogram to visualize the distribution of ages.
Box plot to identify the central tendency and spread of the age data.

The visuals show the age distribution of a US online shopping trends dataset. The left histogram shows the dataset's age distribution pattern and the frequency or count of each age group. The histogram's form and summits show age dispersion, while the blue bars show age group proportions. The kernel density estimation (KDE) curve, which seamlessly displays age density, helps understand the age distribution trend.
The right box plot shows age data distribution and central tendency. The box plot shows the median, quartiles, and probable anomalies. This function calculates the dataset's age distribution, including the median age (50th percentile) and any age value dispersion.
The dataset's age distribution on internet shopping trends is fully understood when these graphics are combined. They help demographic analysis by determining age variability, central tendency, and overall age patterns. This data can inform age-specific product offerings and marketing strategies.



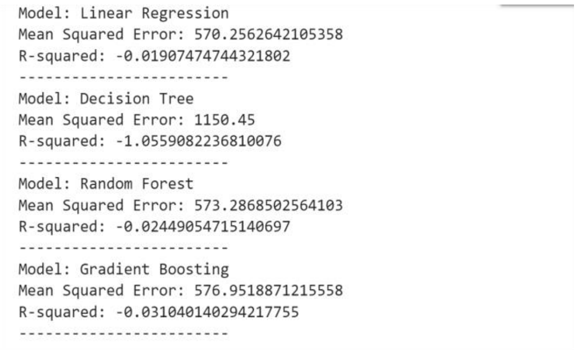
Bar plot to visualize the frequency of each item.
Pie chart to show the proportion of each item.

The visualizations show the top popular products from a US internet purchasing dataset. The left side of the bar plot shows item purchase frequency. Each bar in the table shows the count or frequency of a single item in the collection. Bar heights reflect product purchase frequencies. This depiction highlights the most popular things by bar height to accurately compare item frequencies.
Right pie graphic shows dataset item purchase percentages. Each pie component represents a different item, and its proportionate size to the whole pie denotes its sales share. Each segment shows a percentage indicating how much a product contributed to total purchases. This visualization helps identify the most often purchased products by showing the contribution of each item to the total purchases.
After integration, these visual representations show the frequency and proportion of each dataset item purchased, providing a complete understanding of the item purchase distribution. They help identify the most popular items and evaluate their relative importance within the dataset's online buying trends.



The radar chart may show US online shopping patterns. It shows a dataset's normalized total purchase values for distinct categories. An axis from the circular picture represents each category. Each axis shows the normalized total purchase value for a category by center-to-plotted line distance. Plotting compares and distributes values across buying categories. The plot's structure and line segment length show each category's proportion of overall purchases.

The circular layout makes it easy to compare purchase potency across categories quickly. This visualization technique highlights categories with bigger normalized buy amounts by strategically identifying important categories that account for a larger part of the purchase cost. The focus on each category's contribution helps understand the expenditure distribution and identify online purchase patterns dataset areas of interest.



```
Model: Linear Regression
Mean Squared Error: 570.2562642105358
R-squared: -0.01907474744321802
------------------------
Model: Decision Tree
Mean Squared Error: 1150.45
R-squared: -1.0559082236810076
------------------------
Model: Random Forest
Mean Squared Error: 573.2868502564103
R-squared: -0.02449054715140697
------------------------
Model: Gradient Boosting
Mean Squared Error: 576.9518871215558
R-squared: -0.031040140294217755
------------------------
```

This machine learning pipeline calculates USD purchase amounts using a dataset and various regression models. First, it labels the dataset's categorical variables. "Frequency of Purchases," "Payment Method," "Shipping Type," "Discount Applied," "Promo Code Used," "Item Purchased," "Location," "Size," "Color," "Season," "Subscription Status," "Payment Method," "Shipping Location," and "Shipping Technique."

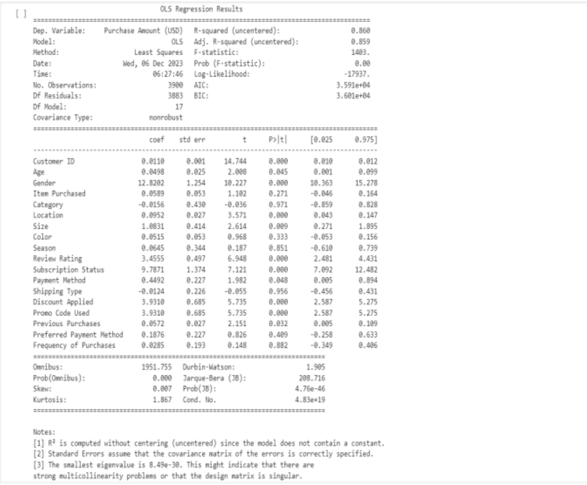
After encoding, the dataset is split 80/20 into training and testing subsets. Multiple regression models are trained on the testing set and evaluated on the testing set. Marginal boosting regression, linear regression, decision tree regression, and gradient boosting regression are used.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 4 (November) (2024)

Each model is trained with training data, then predictions are applied to the test set. Each model's efficacy is measured by MSE and R2.
Mean squared error (MSE) measures the average squared difference between expected and actual target values. Lower MSEs show prediction accuracy.
The statistical metric R-squared (R2) measures how much the independent factors explain target variable variability. The model accounts for more target variable variance, as shown by values closer to one on the 0–1 interval.
After analyzing the test data, the code extract outputs a table with the model's name, MSE, and R2 values. This allows comparison of USD purchase quantity predictability by each model. The iterative procedure helps evaluate each regression model's ability to capture and predict target variable fluctuation, as shown by the dataset.



```
[ ]                                 OLS Regression Results
    ==================================================================================================
    Dep. Variable:     Purchase Amount (USD)   R-squared (uncentered):                   0.860
    Model:                               OLS   Adj. R-squared (uncentered):              0.859
    Method:                    Least Squares   F-statistic:                              1403.
    Date:                   Wed, 06 Dec 2023   Prob (F-statistic):                        0.00
    Time:                           06:27:46   Log-Likelihood:                         -17937.
    No. Observations:                   3900   AIC:                                  3.591e+04
    Df Residuals:                       3883   BIC:                                  3.601e+04
    Df Model:                             17
    Covariance Type:               nonrobust
    ===================================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
    ---------------------------------------------------------------------------------------------------
    Customer ID               0.0110      0.001     14.744      0.000       0.010       0.012
    Age                       0.0498      0.025      2.008      0.045       0.001       0.099
    Gender                   12.8202      1.254     10.227      0.000      10.363      15.278
    Item Purchased            0.0589      0.053      1.102      0.271      -0.046       0.164
    Category                 -0.0156      0.430     -0.036      0.971      -0.859       0.828
    Location                  0.0952      0.027      3.571      0.000       0.043       0.147
    Size                      1.0831      0.414      2.614      0.009       0.271       1.895
    Color                     0.0515      0.053      0.968      0.333      -0.053       0.156
    Season                    0.0645      0.344      0.187      0.851      -0.610       0.739
    Review Rating             3.4555      0.497      6.948      0.000       2.481       4.431
    Subscription Status       9.7871      1.374      7.121      0.000       7.092      12.482
    Payment Method            0.4492      0.227      1.982      0.048       0.005       0.894
    Shipping Type            -0.0124      0.226     -0.055      0.956      -0.456       0.431
    Discount Applied          3.9310      0.685      5.735      0.000       2.587       5.275
    Promo Code Used           3.9310      0.685      5.735      0.000       2.587       5.275
    Previous Purchases        0.0572      0.027      2.151      0.032       0.005       0.109
    Preferred Payment Method  0.1876      0.227      0.826      0.409      -0.258       0.633
    Frequency of Purchases    0.0285      0.193      0.148      0.882      -0.349       0.406
    ==============================================================================
    Omnibus:                     1951.755   Durbin-Watson:                   1.905
    Prob(Omnibus):                  0.000   Jarque-Bera (JB):              208.716
    Skew:                           0.007   Prob(JB):                     4.76e-46
    Kurtosis:                       1.867   Cond. No.                     4.83e+19
    ==============================================================================

    Notes:
    [1] R² is computed without centering (uncentered) since the model does not contain a constant.
    [2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
    [3] The smallest eigenvalue is 8.49e-30. This might indicate that there are
    strong multicollinearity problems or that the design matrix is singular.
```

The table summarizes OLS regression and stats models library statistical analysis. To determine the correlation between USD purchase quantities and features, OLS regression is used.

Table Components

1. Dependent Variable: The table begins with 'Purchase Amount (USD)'. Independent variables are used to predict.

2. Model Fit Statistics: R-squared (R2) shows how much of the dependent variable's variance the independent factors explain. More precise model fits are indicated by higher R2 values near one.
Make changes This modified R-squared considers the number of independent variables in the model. Inclusion of extra variables is penalized.

3. Table of Coefficients: Estimated coefficients for each independent variable are shown here. The coefficients show how a one-unit change in the independent variable affects the dependent variable, assuming all other variables remain constant.
"Standard error" is the estimated coefficient's standard deviation. Precise estimations have lower standard errors.

Statistical significance of coefficient estimates is measured by the t-statistic. Absolute t-values over two indicate variable importance.

'P>|t|': This column shows the null hypothesis p-value for the coefficient being 0, indicating no influence. The variable is statistically significant when the P-value is less than 0.05.

The columns with coefficient 95% confidence intervals are "[0.025 0.975]." It shows the range of coefficient population values.

These statistics determine residual distributions (the discrepancies between observed and expected values). A residual plot with a regular distribution usually suggests a good model fit.

Interpretation

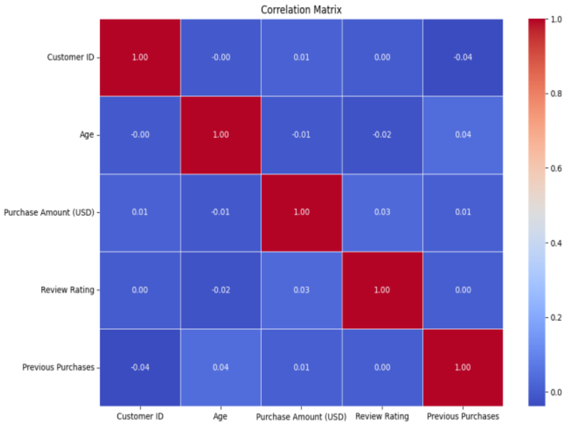
Square of R Ratio: A higher R2 suggests that the model's independent variables explain more purchase quantity variability.

Coefficients: Each independent variable's coefficient indicates its impact on purchase price magnitude and direction. good coefficients indicate a good influence, whereas negative coefficients indicate a negative influence. Using the p-value and t-statistic, a variable can be considered significant.

Belief intervals define the range where the coefficients' true values can be determined with high probability.

Dollar Values The residual distribution shows how well model assumptions like homoscedasticity and normality are met. Skewness and kurtosis around zero indicate a typical residual distribution.

The table provides useful information on the model's data adherence, variable significance, and estimated coefficient precision in forecasting purchase quantities using independent factors.



A heatmap shows the correlation matrix's variables' correlations using colors and annotations. Individual cells in the heatmap show the correlation coefficient between two variables. Correlation matrix heatmaps can understand these elements:

The Color Gradient: Intensity of color the variables' correlation magnitude and orientation are shown by colors. Warm (red) and cold (blue) colors represent positive and negative correlations. Deeper or lighter shades indicate stronger relationships.

Overview of Correlation

A positive correlation is: Dark red cells indicate strong positive correlations between variables. This shows that the second variable tends to rise when the first does.

Relative negativity: Dark blue cells indicate significant negative correlations, meaning an increase in one measure usually decreases the other.

Interpretation

Highly Correlated Variables: Cells around -1 or one suggest a stronger association. A strong positive link between "Variable A" and "Variable B" is seen when their correlation coefficients approach one.

Low Correlation: Values approaching zero suggest a weak linear link between variables.

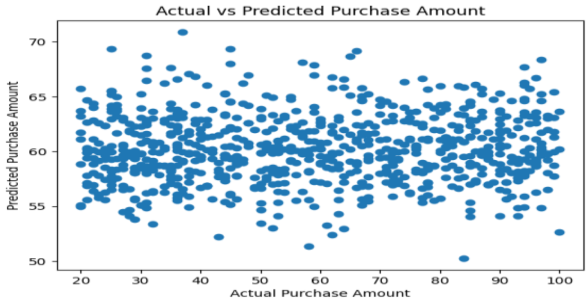
Features Selection Importance: Multiple columns: Regression analysis relies on significant correlations between independent variables, or features. Strong correlations suggest multicollinearity, which involves highly interdependent variables, may affect model interpretability and stability.

Numeric Value Explanations The correlation coefficient is numerically expressed in compartments. Values vary from -1 to 1. A linear relationship is non-existent at zero, perfectly negative at one, and perfectly positive at one.

Heatmap Display: Completeness and Clarity: The heatmap's attractive dimensions make relationships easy to spot. Individual cell annotations clarify correlation coefficients.

The 'cool warm' color map speeds visual perception by defining positive and negative connotations.

Overall, the correlation matrix histogram is a powerful tool for quickly understanding variable correlations. This tool helps identify influential or unnecessary features, enabling educated feature selection or preprocessing for machine learning models.

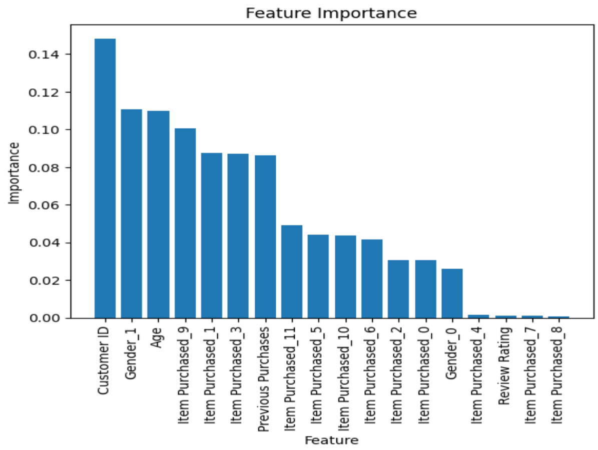


The scatter plot shows the association between the y-axis, which shows the updated Random Forest Regression model's anticipated purchase volumes, and the x-axis, which shows actual purchases. Every graph data point is a test dataset observation.

A diagonal line in the diagram indicates perfect predictions, which align predicted and actual values (y = x). In a perfect world, each point would align with this line, confirming the model's predictions against the observed values. Point distribution along the diagonal line can reveal the model's prediction accuracy.

Points heavily grouped around the diagonal line represent accurate predictions. If most points are on or near the line, the purchase quantities match the model's predictions. Points that deviate from the diagonal line show differences between expected and observed values. A wider dispersion indicates model predictions are more variable or inaccurate.

The scatter plot shows how well the upgraded Random Forest Regression model predicts purchase volumes. Data points near the diagonal line indicate model predictiveness. A densely concentrated cluster of points around the diagonal line indicates more predicted accuracy, whereas a more dispersed distribution suggests potential regions where the model's predictions may depart from observed values. The graphic shows the model's prediction capabilities and limits, helping evaluate its ability to correctly define attributes and the target variable.



The bar plot displays modified Random Forest Regression model feature relevance ratings. The property corresponding to each plot bar's purchase quantity prediction significance is supplied. Taller bars indicate more important model predicted elements.

X-axis feature names the dataset's significant feature names are shown. A bar represents each plot feature.

Value perceived (Y-axis): This axis shows Random Forest model significance scores for each feature. Greater values indicate more predictive capability for purchasing amounts. Taller bars indicate characteristics that better predict purchase quantity. Different elements' relative importance is shown by their bar heights.

Feature Significance Perspective

Relative Significance: Features represented by taller bars have a more pronounced impact on the predictions made by the model. They possess a more substantial impact on the quantity that is acquired.

Ranked Features: The factors that exert the most significant impact on purchase amounts are those that are arranged in a specific order of importance. It is believed that features with higher significance scores are more crucial for making accurate predictions.

Visualizing Feature Importance

Ordered Bars: Reduced significance bars are left to right. This helps quickly identify key elements.

X-axis rotation Labels: Rotated x-axis labels make feature names easier to see, especially with several features.

In general, the bar plot shows Random Forest model-determined feature relevance. It helps participants focus on the main factors affecting purchase quantity estimation. An awareness of feature importance aids features selection, model explanation, and model performance by focusing on key variables.

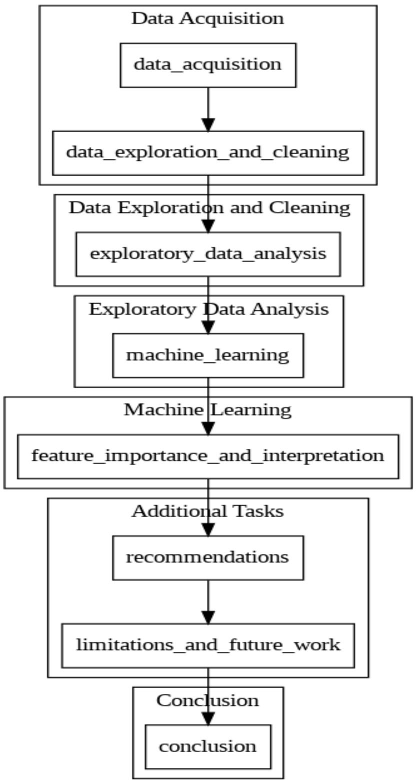
Conclusion

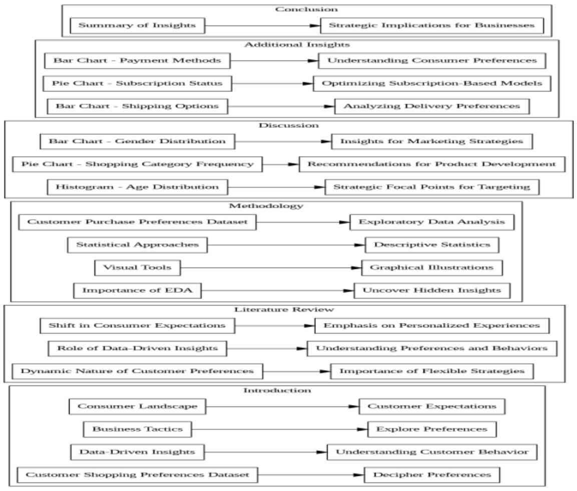
This extensive examination of the Customer Shopping Preferences Dataset used exploratory data analysis (EDA) to better understand customer behavior. A demographic examination of the dataset showed that 68% of users were male and 32% female. This discovery requires organizations to rethink their strategy and promote specific marketing campaigns for underrepresented women. The age distribution analysis also showed a major consumer segment between 25 and 45, advising organizations tailor their strategy to this age cohort.

The footwear and apparel selection influenced customer purchases. This essential information helps firms efficiently devote resources to fashion-related items that matches client preferences. Identifying non-traditional categories helps companies innovate and expand their product offers.

Analysis of consumer preferences and purchase trends gives firms a strategic approach. Product creation and advertising can be tailored to seasonal patterns, consumer demand, and color preferences. Understanding seasonal buying trends, color preferences, and product appeal builds consumer loyalty and satisfaction.

Examining membership statuses and payment methods shows the need of supporting diverse consumer preferences. Using this data can speed up payment processing and improve subscription options, increasing consumer loyalty. According to a geographic analysis, states with high buy frequency have market expansion potential, while states with low purchase frequency are growth zones. Businesses should adjust their marketing to regional tastes.


Data Acquisition
data_acquisition
data_exploration_and_cleaning
Data Exploration and Cleaning
exploratory_data_analysis
Exploratory Data Analysis
machine_learning
Machine Learning
feature_importance_and_interpretation
Additional Tasks
recommendations
limitations_and_future_work
Conclusion
conclusion


Conclusion
Summary of Insights
Strategic Implications for Businesses
Additional Insights
Bar Chart - Payment Methods
Understanding Consumer Preferences
Pie Chart - Subscription Status
Optimizing Subscription-Based Models
Bar Chart - Shipping Options
Analyzing Delivery Preferences
Discussion
Bar Chart - Gender Distribution
Insights for Marketing Strategies
Pie Chart - Shopping Category Frequency
Recommendations for Product Development
Histogram - Age Distribution
Strategic Focal Points for Targeting
Methodology
Customer Purchase Preferences Dataset
Exploratory Data Analysis
Statistical Approaches
Descriptive Statistics
Visual Tools
Graphical Illustrations
Importance of EDA
Uncover Hidden Insights
Literature Review
Shift in Consumer Expectations
Emphasis on Personalized Experiences
Role of Data-Driven Insights
Understanding Preferences and Behaviors
Dynamic Nature of Customer Preferences
Importance of Flexible Strategies
Introduction
Consumer Landscape
Customer Expectations
Business Tactics
Explore Preferences
Data-Driven Insights
Understanding Customer Behavior
Customer Shopping Preferences Dataset
Decipher Preferences

References

1. Guardian. (2016, February 17). How Green Is Online Shopping? Retrieved from https://www.theguardian.com/environment/2016/feb/17/how-green-is-online-shopping.

2. Grashuis, J., Skevas, T., & Segovia, M. (2020). Grocery shopping preference during the COVID-19 pandemic. Sustainability, 12, 5369. https://doi.org/10.3390/su12135369

3. Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful, and customizable freeware package for choice model estimation and application. Journal of Choice Modelling, 32. https://doi.org/10.1016/j.jocm.2019.100170

4. Jaller, M., & Pahwa, A. (2020). Evaluating the environmental impacts of online shopping. Behavioral and Transportation Approaches, 80, 102223.

5. Jones, K. (2020). How Covid-19 Consumer Spending Is Impacting Industries. Retrieved from https://www.visualcapitalist.com/consumer-spending-impacting-industries/.

6. Kilcarr, S. (2013). E-Commerce and Transportation. Retrieved from https://www.fleetowner.com/industry-perspectives/trucks-at-work/article/21688345/ecommerce-and-transportation.

7. Kim, Y., Kang, J., & Chun, H. (2022). Is online shopping packaging waste a threat to the environment? Economics Letters, 214, 110398. https://doi.org/10.1016/j.econlet.2021.110398

8. Koch, J., Frommeyer, B., & Schewe, G. (2020). Online shopping motives during the Covid-19 pandemic – lessons from the crisis. Sustainability, 12(24), 10247. https://doi.org/10.3390/su122410247

9. Kuoppamaki, S., Taipale, S., & Wilska, T. (2017). The use of mobile technology for online shopping and entertainment among older adults in Finland. Telematics and Informatics, 34, 110–117.

10. Lee, R., Sener, I., Mokhtarian, P., & Handy, S. (2017). Relationships between the online and in-store shopping frequency of Davis, California residents. Transportation Research Part A, 100, 40–52.

11. Miyatake, K., Nemoto, T., Nakaharai, S., & Hayashi, K. (2016). Reduction in consumers' purchasing cost by online shopping. Transportation Research Procedia, 12, 656–666.

12. Muller, A., Steins-Loeber, S., Trotzke, P., Vogel, B., Georgiadou, E., & Zwann, M. (2019). Online Shopping in treatment-seeking patients with buying-shopping disorder. Comprehensive Psychiatry, 94.

13. Potoglou, D., & Susilo, Y. O. (2008). Comparison of vehicle-ownership models. Transportation Research Record, 2076(1), 97–105.

14. PWC. (2022). A Time for Hope: Consumers' Outlook Brightens Despite Headwinds. Retrieved from https://www.pwc.com/gx/en/industries/consumer-markets/consumer-insights-survey.html.

15. Rita, P., Oliveira, T., & Farisa, A. (2019). The impact of e-service quality and customer satisfaction on customer behavior in online shopping. Heliyon, 5, 02690.

16. Rosqvist, L., & Hiselius, L. (2016). Online shopping habits and the potential for reduction in carbon dioxide emissions from passenger transport. Journal of Cleaner Production, 131, 163–169.

17. Rutter, A., Bierling, D., Lee, D., Morgan, C., & Warner, J. (2017). How Will E-Commerce Growth Impact Our Transportation Network? Texas A&M Transportation Institute, PRC, pp. 17–79F.

18. Santos, A., mccuckin, N., Nakamoto, H. Y., Gray, D., & Liss, S. (2011). Summary of Travel Trends: 2009 National Household Travel Survey. Washington DC.

19. Schmid, B., & Axhausen, K. (2019). In-store or online shopping of search and experience goods: a hybrid choice approach. Journal of Choice Modelling, 31, 156–180.

20. Shamshiripour, A., Rahimi, E., Shabanpour, R., & Mohammadian, A. (2020). How is COVID-19 reshaping activity-travel behavior? Evidence from a comprehensive survey in Chicago. Transportation Research Interdisciplinary Perspectives, 7(3). https://doi.org/10.1016/j.trip.2020.100216

21. Abdullah Al Noman, Md Tanvir Rahman Tarafder, S. M. Tamim Hossain Rimon, Asif Ahamed, Shahriar Ahmed, and Abdullah Al Sakib, "Discoverable Hidden Patterns in Water Quality through AI, LLMs, and Transparent Remote Sensing," The 17th International Conference on Security of Information and Networks (SIN-2024), Sydney, Australia, 2024, pp. 259–264.

22. S. B. Nuthalapati and A. Nuthalapati, "Advanced Techniques for Distributing and Timing Artificial Intelligence Based Heavy Tasks in Cloud Ecosystems," J. Pop. Ther. Clin. Pharm., vol. 31, no. 1, pp. 2908–2925, Jan. 2024, doi:10.53555/jptcp.v31i1.6977.

23. A. Nuthalapati, "Building Scalable Data Lakes For Internet Of Things (IoT) Data Management," Educational Administration: Theory and Practice, vol. 29, no. 1, pp. 412–424, Jan. 2023, doi:10.53555/kuey.v29i1.7323.

24. M. A. Sufian, Z. M. Guria, N. Morshed, S. M. T. H. Rimon, A. I. Mosaddeque, and A. Ahamed, "Leveraging Machine Learning for Strategic Business Gains in the Healthcare Sector," 2024 International Conference on TVET Excellence & Development (ICTeD-2024), Melaka, Malaysia, 2024.

25. S. M. T. H. Rimon, Mohammad A. Sufian, Zenith M. Guria, Niaz Morshed, Ahmed I. Mosaddeque, and Asif Ahamed, "Impact of AI-Powered Business Intelligence on Smart City Policy-Making and Data-Driven Governance," International Conference on Green Energy, Computing and Intelligent Technology (GEn-CITy 2024), Johor, Malaysia, 2024.

26. S. B. Nuthalapati and A. Nuthalapati, "Accurate Weather Forecasting with Dominant Gradient Boosting Using Machine Learning," Int. J. Sci. Res. Arch., vol. 12, no. 2, pp. 408–422, 2024, doi:10.30574/ijsra.2024.12.2.1246.

27. A. I. Mosaddeque, Z. M. Guria, N. Morshed, M. A. Sufian, A. Ahamed, and S. M. T. H. Rimon, "Transforming AI and Quantum Computing to Streamline Business Supply Chains in Aerospace and Education," 2024 International Conference on TVET Excellence & Development (ICTeD-2024), Melaka, Malaysia, 2024.
28. A. Nuthalapati, "Architecting Data Lake-Houses in the Cloud: Best Practices and Future Directions," Int. J. Sci. Res. Arch., vol. 12, no. 2, pp. 1902–1909, 2024, doi:10.30574/ijsra.2024.12.2.1466.

29. A. Ahamed, M. T. R. Tarafder, S. M. T. H. Rimon, E. Hasan, and M. A. Amin, "Optimizing Load Forecasting in Smart Grids with AI-Driven Solutions," 2024 IEEE International Conference on Data & Software Engineering (ICoDSE-2024), Gorontalo, Indonesia, 2024.

30. M. T. R. Tarafder, M. M. Rahman, N. Ahmed, T.-U. Rahman, Z. Hossain, and A. Ahamed, "Integrating Transformative AI for Next-Level Predictive Analytics in Healthcare," 2024 IEEE Conference on Engineering Informatics (ICEI-2024), Melbourne, Australia, 2024.

31. Shang, Q., Jin, J., & Qiu, J. (2020). Utilitarian or hedonic: event-related potential evidence of purchase intention bias during online shopping festivals. Neuroscience Letters, 715, 134665. https://doi.org/10.1016/j.neulet.2019.134665

32. Shen, H., Namdarpour, F., & Lin, J. (2022). Investigation of Online Grocery Shopping and Delivery Preference before, during, and after Covid-19. Transportation Research Interdisciplinary Perspectives, 14. https://doi.org/10.1016/j.trip.2020.100216

33. Smithson, D. (2018). Macy’s at Oaks Mall Closing. Retrieved from https://www.gainesville.com/news/20180104/macys-at-oaks-mall-closing.

34. Statista. (2020). Retail e-commerce sales in the United States from 2017 to 2024. Retrieved from https://www.statista.com/statistics/272391/us-retail-e-commerce-sales-forecast/.

35. Statistica. (2022). Number of Online Grocery Purchasers in the United States from 2018 to 2024. Retrieved from https://www.statista.com/statistics/1032362/online-grocery-purchasers-united-states/.

36. Suel, E., & Polak, J. (2017). Incorporating online shopping into travel demand modeling: challenges, progress, and opportunities. Transport Reviews, 38(5), 576–601.

37. Sun, X., Wandelt, S., Zheng, C., & Zhang, A. (2021). Covid-19 pandemic and air transportation: successfully navigating the paper hurricane. Journal of Air Transport Management, 94, 102062. https://doi.org/10.1016/j.jairtraman.2021.102062

38. Suyanto, B., Subiakto, H., & Srimulyo, K. (2019). Data of the patterns of youth local brand product consumption through online shopping. Data in Brief, 23, 103723.

39. Thaichon, P. (2017). Consumer socialization process: the role of age in children's online shopping behavior. Journal of Retailing and Consumer Services, 34, 38–47.

40. The Economist. (2017). Stores Are Being Hit by Online Retailing. Retrieved from https://www.economist.com/special-report/2017/10/26/stores-are-being-hit-by-online-retailing.

41. Tian, X., An, C., Chen, Z., & Tian, Z. (2021). Assessing the impact of Covid-10 pandemic on urban transportation and air quality in Canada. Science of the Total Environment, 756, 144270. https://doi.org/10.1016/j.scitotenv.2020.144270

42. U.S. Bureau of labor statistics. (2022). TED: the Economics Daily. Retrieved from https://www.bls.gov/opub/ted/2022/consumer-prices-up-7-5-percent-over-year-ended-january-2022.htm.

43. U.S. Department of Transportation, Federal Highway Administration. (2017). National household travel survey. URL: http://nhts.ornl.gov.

44. U.S. Department of Commerce. (2022). U.S. Census Bureau News. Retrieved from https://www.census.gov/retail/index.html#ecommerce.

45. Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

46. Walmart. (2021). Walmart Introduces Walmart+. Retrieved from https://corporate.walmart.com/newsroom/2020/09/01/walmart-introduces-walmart.

47. Warren, K. (2019). UPS Pulse of the Online Shopper- a Customer Experience Study. United Postal Services (UPS) technical report.

48. Whiteman, D. (2020). These chains have announced a ton of store closings in 2019. Retrieved from https://moneywise.com/a/retailers-closing-stores-in-2019.

49. Wickham, H., Averick, M., Bryan, J., Chang, W., mcgowan, L. D. A., François, R., Yutani, H. (2019). Welcome to the tidyverse. Journal of Open-Source Software, 4(43), 1686.

50. Xi, G., Cao, X., & Zhen, F. (2020). The impacts of same day delivery online shopping on local store shopping in Nanjing, China. Transportation Research Part A, 136, 35–47.

51. Zhang, X., Li, Z., & Wang, J. (2021). Impact of Covid-19 pandemic on energy consumption and carbon dioxide emissions in China's transportation sector. Case Studies in Thermal Engineering, 26, 101091. https://doi.org/10.1016/j.csite.2021.101091

52. Zhang, Y., & Fricker, J. (2021). Quantifying the impact of Covid-19 on non-motorized transportation: a Bayesian structural time series model. Transport Policy, 103, 11–20. https://doi.org/10.1016/j.tranpol.2021.01.013