E-Commerce Data Science Project

Joshua Schaake
Wentworth School of Computing & Data Science

Abstract

My project explores customer behavior within an online retail dataset to understand how demographic and transactional variables relate to satisfaction and spending. I used techniques such as clustering and visualization to identify distinct customer segments and examine key features such as membership, spending habits, and discounts to see their influence on satisfaction levels. I also incorporated customer-level insights in a national e-commerce trend. **Keywords: E-commerce, customer behavior, satisfaction, clustering, and segmentation.**

Introduction

In a world that continues to grow with technology, E-commerce has grown rapidly along with how consumers are interacting with businesses. Understanding customer behavior is very important for business aiming to personalize services as well as changing marketing to become more adaptable to the digital world. In my project, I m analyzing datasets of individual customer transactions to identify what main factors influence spending and satisfaction levels. I combined micro-level consumer data with national statistics to create a full picture of the current online retail environment. In my exploration I had a few specific questions I wanted to gain insight on. They were as follows: How do spending habits and discount usage relate to customers satisfaction? How can customer segmentation be used to personalize pricing strategies? How does customer-level data relate to the overall industry trends?

Data Sets

A. **Dataset 1** which is titled, "E-commerce Customer Behaviors – Sheet1.csv," is a simulated dataset created for educational use in this course. This dataset was designed to show real-world online customer behavior and includes transactional and demographic information. **Dataset 2** which is titled "ECOMNSA.xlsx," and I obtained this from the US Census Bureau's official website. This set covers quarterly E-commerce sales which is published by the US Department of Commerce. The dataset provides time-series data on the national US E-commerce sales from the years 1999 to 2024, although it is not seasonally adjusted. **Dataset 3**, titled "24q4supptables.xlsx," which also came from the US Census Bureau. The set includes quarter 4 of 2024 data that breaks down total and E-commerce retail sales which is broken down by business sector. Originally the data was intended for public use and included government-published economic indicators.

B. **Dataset 1** contains 350 rows and 11 columns. Some of the variables include Gender, City, Membership Type, Satisfaction Level, Discount, Items Purchased, Total Spent, Age, Days Since Last Purchase, and Average Rating. Some data cleaning, I did was label-encode the categorical variables, and got rid of

E-Commerce Data Science Project

Joshua Schaake
Wentworth School of Computing & Data Science

different missing values. This data set for me served as the foundation for all customer-level analysis. **Dataset 2** was fully cleaned to begin with, which made it easy to implement. It was formatted as a time-series table with quarterly observations, which helped me see the growth of E-commerce as a share of total retail sales. I used this to visualize larger scale trends as well as the growth of E-commerce over time. **Dataset 3** consisted of total and E-commerce sales for different retail sectors in the 4th quarter of 2024 alone. Some data cleaning, I did was removing formatting characteristics and non-numeric entries, which helped me see the different sectors side by side. I used both raw dollar figures and comparisons to enhance my visualizations.

Methodology

Following my data cleaning, I was left with data I could work with to create meaningful models and insights. The goal of my **first modeling/visualization** was to explore several key patterns and correlations between the different variables. I used bar plots, box plots, and a heatmap to show the relationships between variables including Total Spent, Items Purchased, Satisfaction Levels, and other variables. For my **second modeling/visualization**, I used K-Means clustering to segment customers into similar behavior groups. I chose variables including Total Spent, Items Purchased, Days Since Last Purchase, Discounts, and Average Rating to create my clustering model. I chose these because I thought these would reflect very well on the transactional behavior and engagement of customers. I generated a 2D scatterplot using Principal Component Analysis, along with 3 clusters, to group customers on the behavioral factors that I chose above. This model allowed me to gain insight into the way customers can be grouped together to show an overall trend in E-commerce. For my **third model/visualization,** I chose to show the larger scale trends of E-commerce in two different visualizations. This will help provide extra information into the specific models I chose for 1 and 2, as well as show the overall growth of E-commerce. The first visual shows a time-series plot from 1999 to 2024 showing the national E-commerce sales numbers. The other plot I created was a bar chart comparing Q4 2024 E-commerce sales to total sales per sector, which helps describe our customer-level data.

Results

Out of my different models, I was able to gain various valuable insights into the world of E-commerce. The correlation heatmap from dataset 1, revealed a few different strong positive relationships and several moderate correlations. Membership Type had a correlation above .9 for Average Rating, Items Purchased, and Total Spent. Additionally, Average Rating showed a moderate correlation with Satisfaction Level, which suggests that service quality may play a factor. Aside from those, we also saw many negative correlations with some

E-Commerce Data Science Project

Joshua Schaake
Wentworth School of Computing & Data Science

of the numeric categories like Age, Days Since Last Purchase, and City. In the visualization of Total Spend by Membership Type, we observed a clear trend: Gold members consistently had the highest average spend, followed by Silver and Bronze. This supports the idea that higher-tier memberships are associated with more frequent or higher-value purchases. The box plot comparing spending by gender indicates that males had slightly more variability in spending, though medians were similar across the 2 genders. The count plot of Satisfaction Level by Discount Applied showed a noticeable increase in satisfied customers among those who received discounts. This makes sense as I would expect that discounts would improve satisfaction levels of an order since there is less money spent. For my K-means clustering, the clusters showed three distinct groups of customers. PCA Component 1, (x-axis), which is dominated by Total Spent and Items purchased, and PCA Component 2, (y-axis), which is dominated by Days Since Last Purchase, are the 2 axes for my clustering model. Cluster 0 included high spenders with recent purchases, while cluster 1 included low spenders who had not purchased recently, and cluster 2 being a midpoint of the two. The clear separation between these clusters confirms that K-Means captured meaningful behavioral differences. National context was provided through Dataset 2, which showed a steady upward trend in U.S. e-commerce sales from 1999 to 2024. This illustrates how the industry continues to grow as consumers shift toward digital channels. Dataset 3 revealed that sectors such as clothing, electronics, and furniture had particularly high E-commerce sales in Q4 2024, highlighting sectors where digital purchasing is most prevalent.

Discussion

The results of my models show that even with a limited dataset, a few different patterns in customer behavior can still be uncovered. My clustering model was particularly useful in differentiating customer types and providing a foundation for marketing strategies for specific groups. Companies can use data like this to change their loyalty programs or discount incentives to try to improve customer engagement and satisfaction. My main limitation was in dataset 1, since it did not include more specific behavioral topics like page views, time on sight, device usage, and seasonality. If even a few of these variables were included I believe it would significantly enhance my models in a real-world outlook. On a different note, my one model that did not work was a logistic regression that I attempted to run on my data. This failed due to the lack of a binary separation, although I believe it would’ve added great insight into my topic. For future work, I would hope to find a way to make logistic regression work with a new data set or further data munging. Additionally, as I mentioned above a few extra variables like page views or time on sight would enhance my models greatly.

Conclusion

My project shows that spending behavior and discount usage are key indicators of customer satisfaction, and that K-Means clustering is an effective method for segmenting customers by engagement and purchase activity. My findings can help E-

commerce businesses target discounts more efficiently, tailor communication, and optimize retention strategies. The insights align with my original research questions and demonstrate how individual behavior reflects broader industry trends in the e-commerce sector. Clustering and visual analysis provide a simple yet powerful toolkit for generating practical business intelligence from transactional data.

References

[1] U.S. Census Bureau. "Quarterly Retail E-Commerce Sales." https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf

[2] U.S. Census Bureau. "Q4 2024 Supplemental Tables from E-Stats." https://www.census.gov/programs-surveys/e-stats.html

[3] Scikit-learn Documentation. https://scikit-learn.org/stable/modules/clustering.html

[4] Seaborn Documentation. https://seaborn.pydata.org

[5] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.