深圳大学考试答题纸

(以论文、报告等形式考核专用)

二〇 二四 ～二〇 二五 学年度第 二 学期

课程编号	3601000008	课序号	01	课程名称	数据可视化	主讲教师	陈诗沁	评分	
学　　号	2022363006	姓名	孔梓鲲	专业年级	2022 金融科技（中外合作办学）01				

教师评语：

题目： **Group 2: Visualizing the American Digital Consumer: A Multi-faceted Analysis of the Open E-commerce 1.0 Dataset (2018-2022)**

Abstract

In an era where large-scale transactional data is supplanting traditional survey methods, understanding consumer behavior through rich, accessible datasets is of paramount importance. This report presents a multi-faceted visual analysis of the "Open E-commerce 1.0" dataset, which contains 1.8 million Amazon purchase histories from over 5,000 U.S. consumers from 2018 to 2022, uniquely linked to their demographic profiles. Our methodology encompasses comprehensive data preprocessing, demographic profiling, time-series decomposition to isolate trend and seasonality, and the application of unsupervised machine learning, including K-Means clustering on an RFM (Recency, Frequency, Monetary) model, to identify user archetypes. The analysis reveals several key findings: (1) The cohort's spending patterns strongly validate against market-level performance, confirming the dataset's reliability. (2) The COVID-19 pandemic acted as a significant accelerant, structurally shifting overall spending upwards and increasing behavioral heterogeneity. (3) A clear distinction exists between high-activity demographic segments and high-value-per-transaction cohorts. (4) Behavioral segmentation successfully identified distinct user archetypes, such as 'High-Value Champions' and 'At-Risk' customers, whose value drivers extend beyond simple demographics to include the diversity of their purchases. (5) The dataset demonstrates latent potential for non-commercial research, such as inferring domestic migration patterns from shipping data. Ultimately, this study demonstrates the power of visual analytics to transform raw transactional data into strategic commercial insights and valuable socioeconomic signals, underscoring the potential of democratized data in both business and academic research.

Keywords: *Data Visualization, Consumer Behavior, E-commerce Analytics, Customer Segmentation, RFM Analysis, Machine Learning, Big Data, Open E-commerce*

Task Division

This project was a comprehensive and collaborative endeavor, with tasks strategically divided among team members to leverage individual strengths and ensure a high-quality outcome. The responsibilities were broadly partitioned into two core components: a) research, analysis, and report generation, and b) presentation development and delivery.

The specific roles and contributions of each team member are as follows:

- **Tianyi LYU (Project Lead, Analysis & Reporting):** Served as the project lead, driving the overall direction and progress. Responsible for the primary development of the visualization code and the initial drafting of this comprehensive academic report, ensuring the core analytical narrative was established.
- **Zikun KONG (Data & Technical Optimization):** Spearheaded the data acquisition process, including sourcing the dataset and performing all necessary data processing and cleaning. Additionally responsible for optimizing the visualization code and refining the final report to enhance its analytical depth and clarity.
- **Zinan YE (Presentation Lead & Content Strategy):** Led the development of the final presentation. Tasked with strategically translating the detailed findings from the academic report into a clear, compelling, and visually engaging presentation format.
- **Jinrui ZHAO (Presentation Co-Development):** Collaborated closely on the production of the presentation materials, contributing to the design, content creation, and overall polish of the final presentation to ensure a professional and effective delivery.

The team maintained a seamless workflow through regular and effective communication. The clear division of labor between the report-focused and presentation-focused sub-teams allowed for parallel progress and synergistic integration of all components, culminating in a final project that reflects a truly collaborative and well-coordinated effort.

Content

1. Introduction

1.1 Background: The New Frontier of Consumer Insights

In the contemporary digital era, a significant paradigm shift is underway in the field of economic and social analytics. Traditionally, insights into consumer behavior have been derived from government-led surveys, which produce foundational public datasets and statistics. However, these conventional methods face growing challenges, most notably a sharp decline in response rates in recent years. In parallel, the proliferation of digital platforms has enabled corporations to collect vast quantities of high-frequency transactional data at a scale that far surpasses public agencies. This creates a critical informational asymmetry: while companies possess powerful, granular data, these resources are generally held privately, leaving their full potential for public-interest research largely unrealized.

The democratization of such rich data sources is paramount. Public bodies, like the U.S. Bureau of Labor Statistics, have actively encouraged the sharing of corporate data to improve the accuracy of vital economic indicators such as the Consumer Price Index (CPI). Furthermore, extensive research has demonstrated how large-scale behavioral data can be used to model complex societal phenomena, from human mobility and urban planning to economic well-being and the spread of disease. This report operates at the intersection of this challenge and opportunity, aiming to unlock insights from a novel, publicly accessible dataset of consumer purchase histories.

1.2 Data Source: The Open E-commerce 1.0 Dataset

To address the gap between data potential and accessibility, this report leverages the "Open E-commerce 1.0" dataset, a first-of-its-kind public resource. The dataset was crowdsourced through an online survey with the informed consent of participants and contains 1.8 million Amazon.com purchases from 5,027 U.S. consumers, spanning a five-year period from 2018 through 2022. The transactional records include essential fields such as order date, product title, price, quantity, and category.

The unique power of this dataset lies in its direct linkage of granular purchase histories to comprehensive user surveys. Each user's transactional data is connected to their self-reported information on demographics (e.g., age, gender, income), lifestyle (e.g., household size), and health status. This fusion of behavioral and self-reported data provides an unprecedented opportunity to move beyond aggregate trend analysis and explore the nuanced drivers of consumer behavior across different segments of the population.

1.3 Research Objectives and Structure

The primary objective of this report is to conduct a multi-faceted visual analysis of the Open E-commerce 1.0 dataset to uncover nuanced patterns in American digital consumer behavior. To achieve this, we pose the following key research questions:

1. What are the core demographic and socioeconomic characteristics of the consumer cohort represented in this dataset?
2. What macro-level temporal dynamics and seasonal patterns characterize e-commerce consumption over the five-year period, and how do these align with established market indicators?

3. How do spending behaviors and product category preferences differ across key demographic segments, such as age, income, and gender?
4. What was the discernible impact of major external events, specifically the COVID-19 pandemic, on online purchasing patterns?

To address these questions, this report is structured as follows. Section 2.0 details the data preparation process and presents a demographic profile of the consumer cohort. Section 3.0 examines macro-level e-commerce trends and validates the dataset against public market data. Section 4.0 provides a deep dive into cross-demographic spending behaviors. Section 5.0 presents event-driven case studies on the COVID-19 pandemic and seasonal cycles. Section 6.0 introduces advanced behavioral segmentation models to identify distinct user archetypes. Finally, Section 8.0 offers a concluding summary, discusses the study's limitations, and suggests directions for future research.

2 Data Preparation and Cohort Profile

A rigorous analysis is contingent upon a well-prepared and thoroughly understood dataset. This section outlines the essential preprocessing steps undertaken to ensure data integrity and details the demographic composition of the consumer cohort. This foundational work is critical for establishing the validity of subsequent analyses and for contextualizing the behavioral patterns observed.

2.1 Data Preprocessing

The raw dataset, sourced from the "Open E-commerce 1.0" repository, required several preprocessing steps to prepare it for analysis. The primary objective of this phase was to clean the data, ensure consistency, and refine the dataset to focus on relevant consumer transactions.

The workflow was executed as follows:

- **Data Cleansing and Type Conversion:** Key fields such as Order Date, Purchase Price Per Unit, and Quantity were converted to their appropriate data types (datetime and numeric, respectively). Records containing null values in these critical fields were removed to ensure the reliability of all subsequent calculations.
- **Exclusion of Non-Product Transactions:** An important methodological decision was made to exclude transactions identified as gift cards. These purchases, often representing financial transfers rather than the consumption of goods or services, could otherwise skew spending pattern analysis. A filter was applied to remove any records where the product title contained terms such as "gift card" or "giftcode".
- **Outlier Mitigation:** To mitigate the potential influence of extreme outliers or data entry errors on aggregate statistics, a conservative filter was applied. All purchase records where the calculated Total Amount exceeded the 99.5th percentile of the distribution were trimmed from the dataset.
- **Dataset Integration:** Finally, the cleaned transactional data was merged with the corresponding survey data using an inner join on the unique Survey ResponseID. This procedure resulted in a final, robust dataset where each valid purchase record is enriched with the detailed demographic and lifestyle profile of the consumer.

2.2 Demographic Landscape of the Digital Shopper

Understanding the characteristics of the consumer cohort is fundamental to interpreting their purchasing behavior. The following analysis presents a multi-faceted profile of the unique users in our dataset, leveraging a series of visualizations to illustrate their core demographic and geographic attributes.

The gender distribution of the cohort, presented in Figure 2.1, indicates a majority of female participants (58.6%) compared to male participants (39.5%). This composition, potentially reflecting differing online survey participation rates or platform usage between genders, is a key factor to consider in later analyses where spending patterns may exhibit gender-specific tendencies.

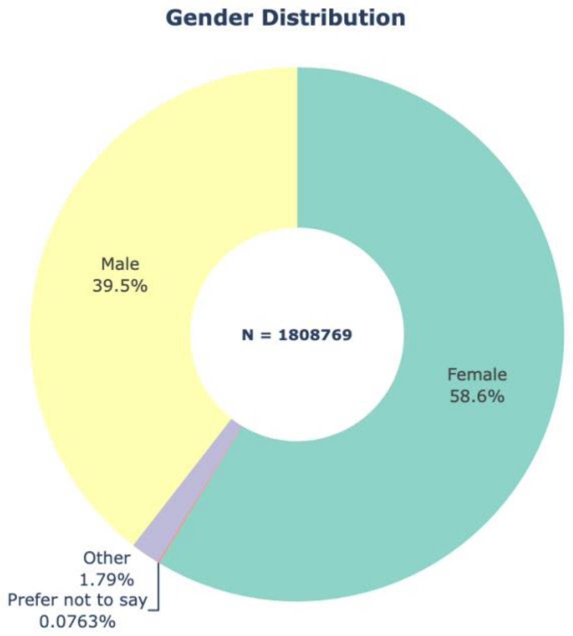

Figure 2.1: Gender Distribution

An examination of the age structure, detailed in Figure 2.2, reveals that the cohort is predominantly composed of young to middle-aged adults. The 25-34 and 35-44 year-old brackets are the largest segments, collectively accounting for over 60% of the participants. This finding suggests the dataset offers a robust view into the consumption habits of the primary digital-native and economically active workforce. Conversely, individuals aged 65 and older are notably underrepresented, a common bias in digital surveys that must be acknowledged.

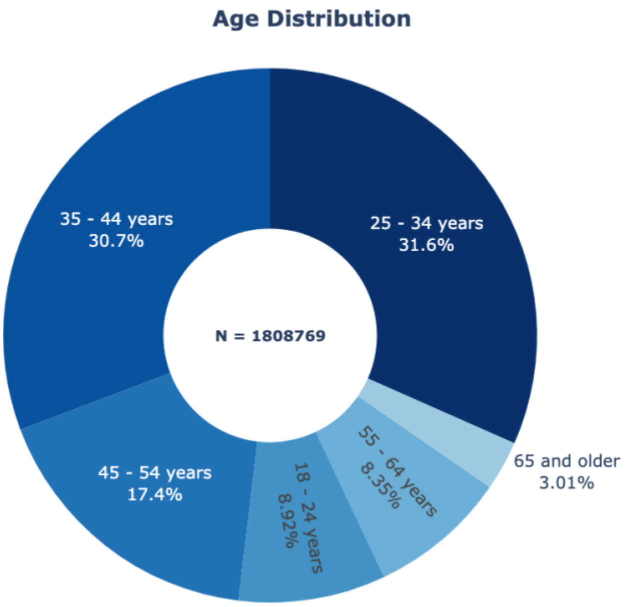

Figure 2.2: Age Group Distribution

The socioeconomic profile of the cohort is illustrated by household income levels in Figure 2.3. The data shows a broad representation across different brackets, with a significant concentration in the middle- to upper-middle-class ranges; the \$50,000-\$74,999 and \$100,000-\$149,999 brackets are the most prominent. This suggests our analysis is particularly relevant for understanding mainstream consumer behavior rather than that at the economic extremes.

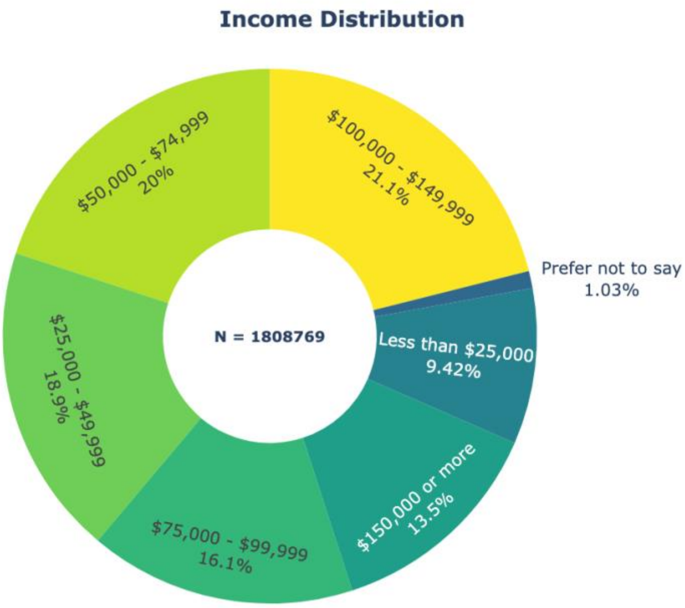

Figure 2.3: Income Level Distribution

Further context is provided by the household structure, as shown in Figure 2.4. The cohort is varied, with two-person households forming the largest group (32.3%), followed closely by households with four or more members (29.5%). This diversity indicates that the purchasing data reflects a mix of consumer units, from couples and individuals to small and large families, each with potentially distinct needs and consumption patterns.

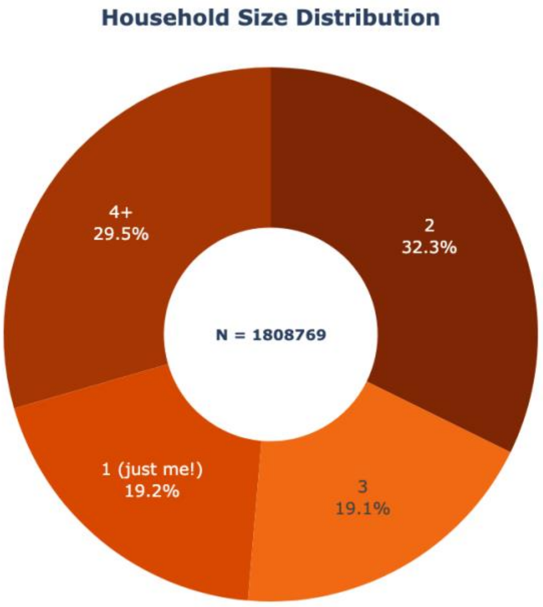

Figure 2.4: Household Size Distribution

Finally, to understand the geographic footprint of the user base, we calculated a user penetration rate—the number of unique users per 100,000 state residents—to normalize for population differences. The resulting choropleth map, Figure 2.5 visualizes this distribution across the United States. While users are present nationwide, the map reveals notable concentrations in states like New Hampshire and Oregon. This visualization moves beyond simple counts to highlight regions with disproportionately high representation in the dataset, which could signal regional variations in platform adoption or survey participation behavior.

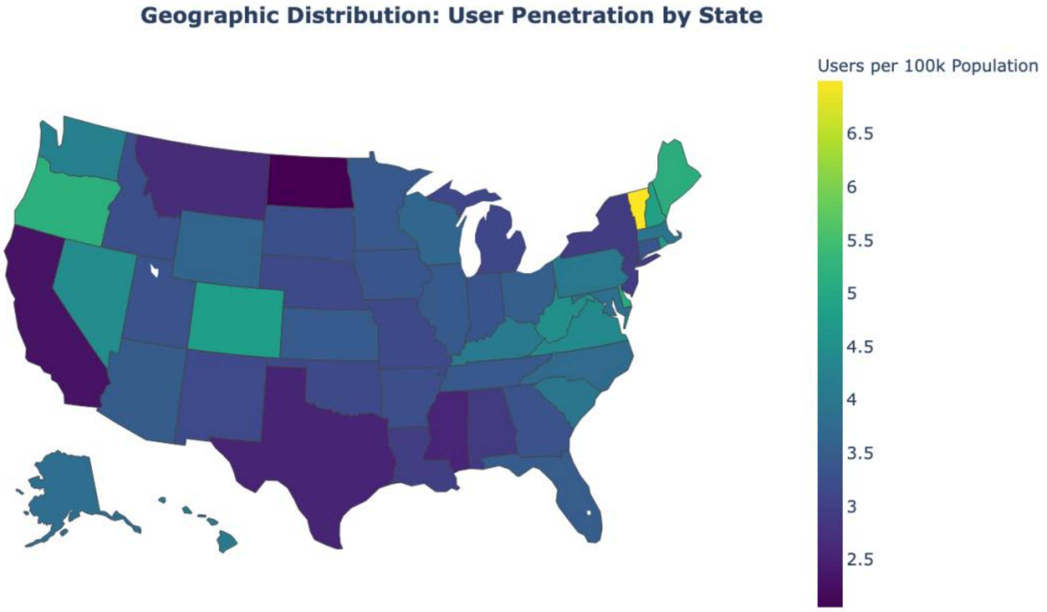
Geographic Distribution: User Penetration by State
Users per 100k Population
6.5
6
5.5
5
4.5
4
3.5
3
2.5

Figure 2.5: Geographic Distribution by State Penetration Rate

3. Macro-Level E-commerce Dynamics and Validation

Having established the profile of our consumer cohort, the next critical step is to understand their collective behavior over time. This section examines the macro-level dynamics of e-commerce consumption from 2018 to 2022. We begin by validating our sample's spending patterns against established market benchmarks to confirm its representativeness. Subsequently, we delve into a granular analysis of temporal trends and seasonal cycles to deconstruct the rhythm of digital commerce.

3.1 Market Validation: Aligning with Amazon's Performance

Before drawing conclusions from a sample-based dataset, it is imperative to validate its trends against broader market indicators. This ensures that the behavioral patterns observed within our cohort are not anomalous, but rather a reliable proxy for the larger population of digital consumers. Following the validation methodology of the original study by Berke et al. (2024), we compared the total quarterly spending of our cohort against Amazon's publicly reported performance.

Figure 3.1 plots the aggregate spending of the sample for each quarter within the study's primary timeframe. The trajectory reveals a distinct and consistent pattern: a significant surge in consumer spending in the fourth quarter (Q4) of each year, corresponding to the holiday shopping season, followed by a moderation in the subsequent first quarter (Q1). This cyclical behavior, coupled with a general upward trend from 2018 through 2022, strongly mirrors the performance of the e-commerce market at large. The pronounced drop-off observed in 2023 is an artifact of the data collection period concluding in March of that year. The high degree of alignment between our sample's spending and market-level data provides strong evidence of our dataset's validity, lending credibility to the more detailed analyses that follow.

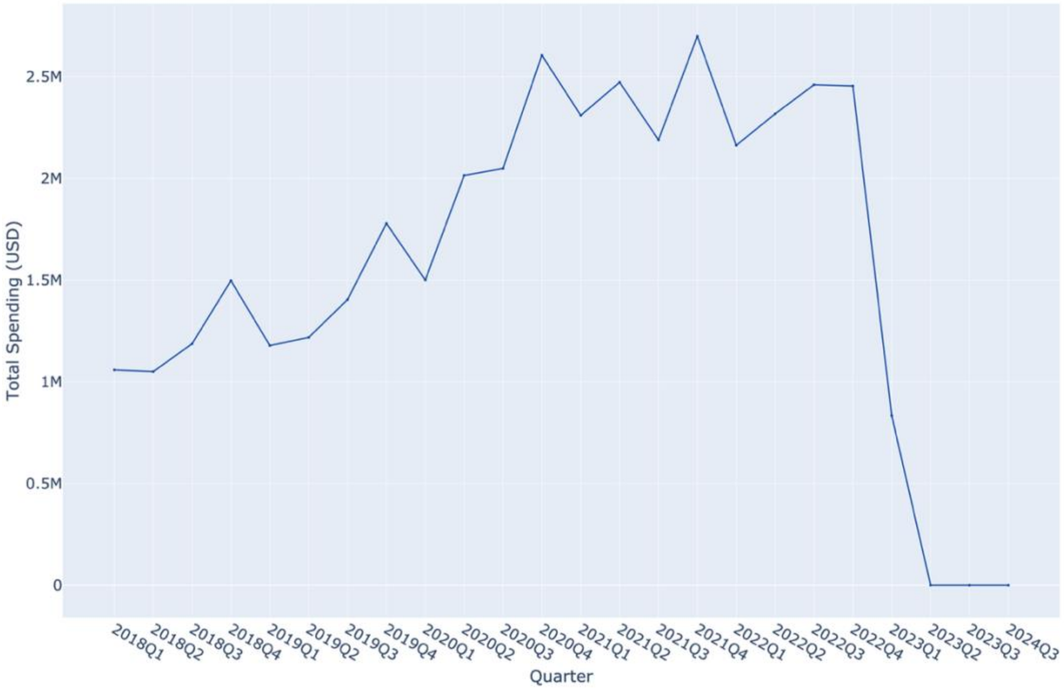

Figure 3.1: Quarterly Spending Evolution

3.2 Temporal Patterns: Uncovering Trends and Seasonality

With the dataset's credibility established, we can now dissect its temporal patterns with greater detail. Figure 3.2 provides a monthly view of consumer behavior, plotting both total spending and the total number of purchases. The visualization clearly shows that these two metrics are tightly coupled; periods of increased spending are consistently accompanied by a higher frequency of purchases. This indicates that growth in consumption is driven by a combination of more active purchasing and potentially higher average order values. The annual rhythm is even more pronounced at this monthly resolution, with sharp peaks consistently appearing in the final months of each year.

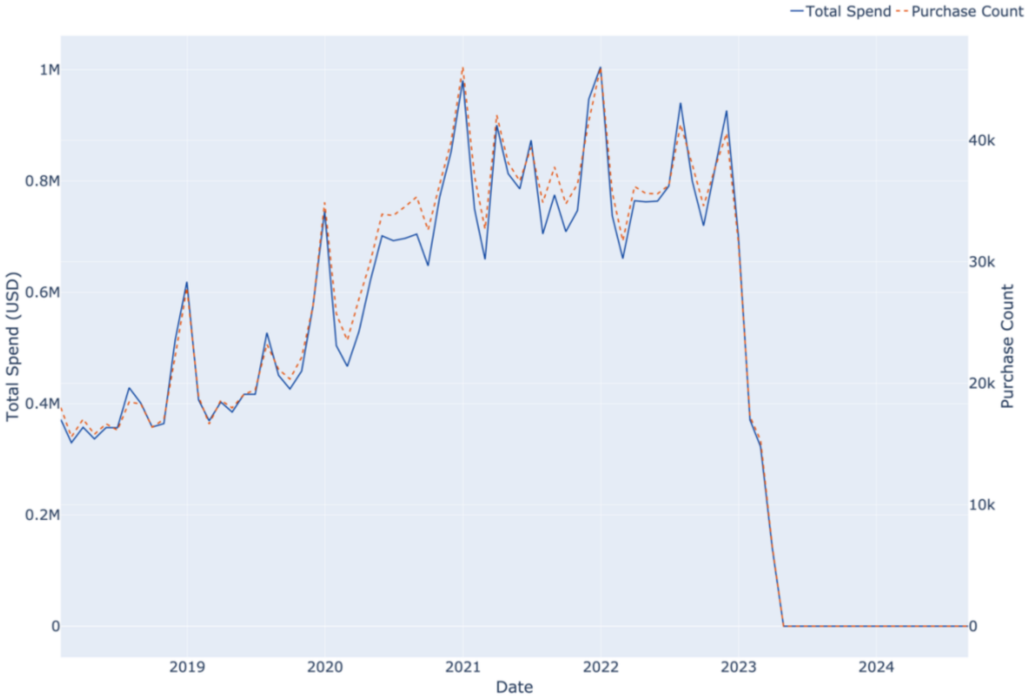

Figure 3.2: Monthly Consumer Activity: Spending vs. Frequency

To deconstruct these dynamics with greater statistical rigor, a seasonal decomposition analysis was performed on the daily spending data, as presented in Figure 3.3. This technique separates the time series into three distinct components:

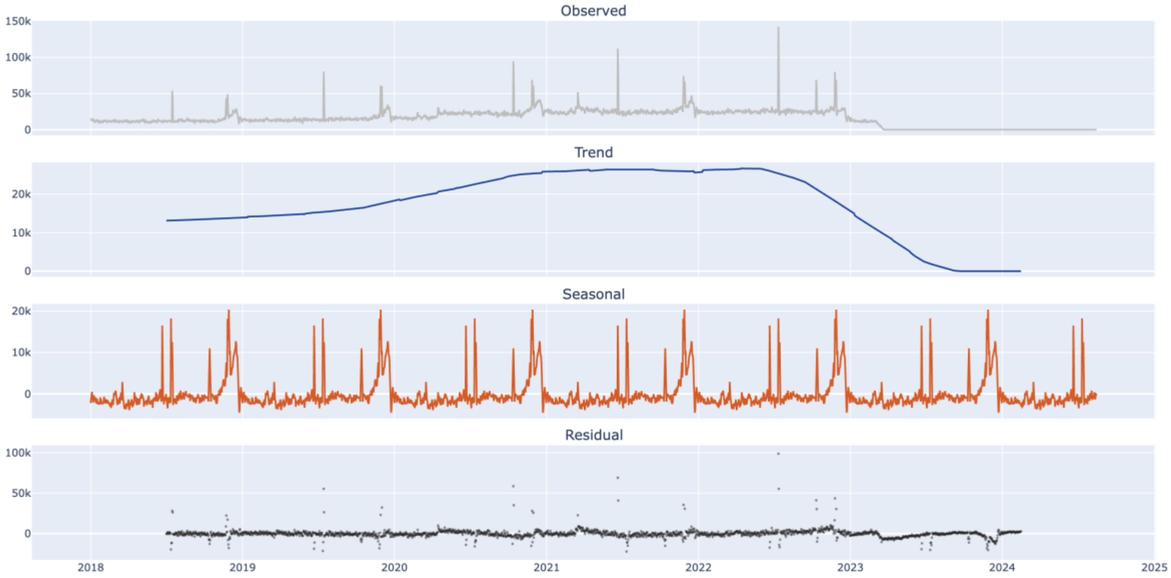

Figure 3.3: Time Series Decomposition of Daily Spending

- **Trend:** The trend component isolates the long-term trajectory of consumer spending, abstracting away from short-term fluctuations. It reveals a clear and sustained growth path from 2018 until reaching a plateau in the 2021-2022 period, reflecting the maturation of the cohort's purchasing habits or broader economic factors.

- **Seasonal:** This component powerfully visualizes the dataset's underlying annual rhythm. It showcases a highly regular and predictable cycle, dominated by the major spending spikes of the Q4 holiday season. Smaller, yet consistent, cyclical patterns are also visible throughout the year.
- **Residual:** The residual component captures the irregular, random noise in the data that cannot be explained by the trend or seasonal cycles. Its relative stability suggests that our model has successfully captured the primary systematic patterns in consumer behavior.

In summary, this multi-layered temporal analysis confirms that the consumption patterns within the Open E-commerce 1.0 dataset are not random. Instead, they are a composite of a robust, multi-year growth trend overlaid with a powerful and predictable annual seasonal cycle, both of which are foundational to understanding the behavior of the digital consumer.

4. Deep Dive: Cross-Demographic Spending Behaviors

To explore how spending power varies across the intersection of different demographic groups, we employ a parallel categories plot. This visualization technique is highly effective for revealing relationships and flows between multiple categorical variables, with the color intensity representing a quantitative measure—in this case, the average spending per order.

4.1 Spending Power: Disparities Across Demographics

Figure 4.1 illustrates the spending flows between gender and age cohorts. Each band represents a specific demographic combination (e.g., females aged 25-34), its thickness indicates the relative size of the cohort within the dataset, and its color signifies the average spending value per transaction. A key insight revealed is that spending power is not dictated by a single factor but by their interaction. The highest average spending (indicated by the bright yellow bands) is concentrated among older male cohorts, particularly those in the 55-64 and 65 and older age brackets. Conversely, the youngest cohorts (18-24 years), regardless of gender, consistently exhibit the lowest average spending, as shown by the dark purple bands. Furthermore, the visualization highlights a crucial distinction between activity volume and transactional value: while female users aged 25-44 form the thickest, most active bands, their average spending per order is moderate. This suggests that high user activity does not necessarily equate to the highest per-transaction value, a vital insight for targeted marketing and customer valuation.

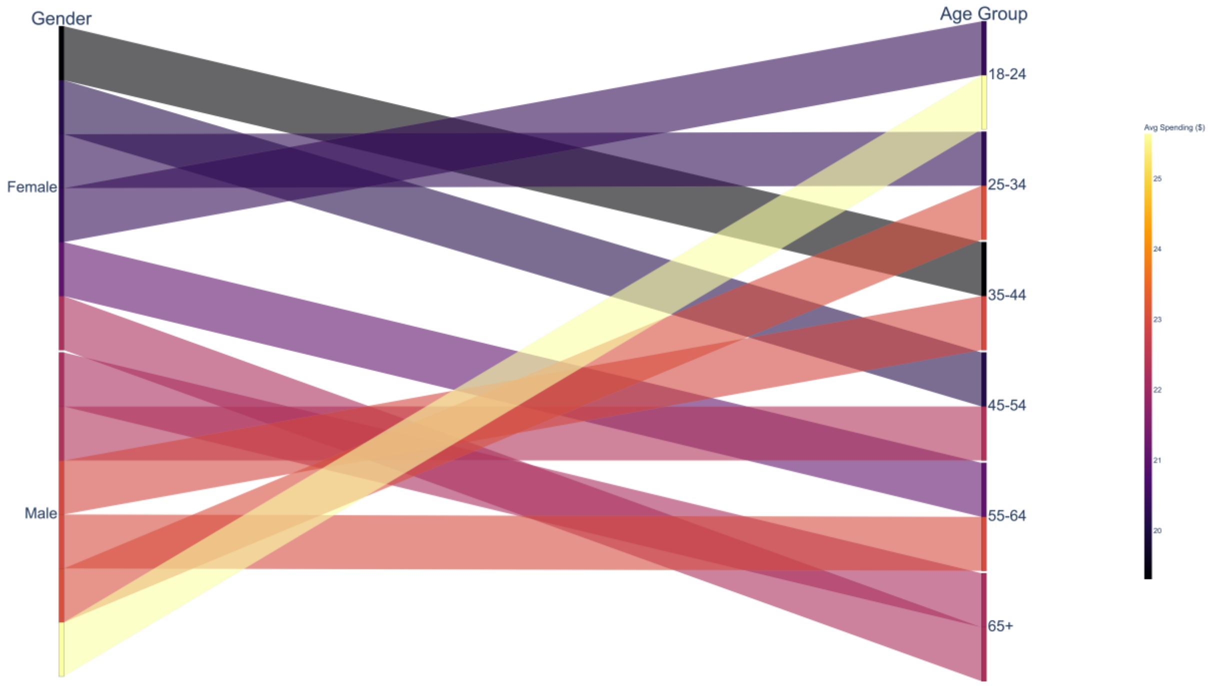

Figure 4.1: Aggregated Spending Patterns by Demographic

4.2 Category Preferences: What Do People Buy?

Understanding what consumers purchase is as important as understanding who is purchasing. This subsection dissects the product category landscape to identify dominant market segments and uncover the distinct characteristics of each.

A high-level overview is provided in Figure 4.2. In this visualization, the area of each rectangle is proportional to the total spending in that category, while the color corresponds to the average spend per order. Immediately, "Abis Book" emerges as the category with the highest aggregate spending. However, the color-coding reveals a more complex story: categories like "Computer Drive or Storage" and "Headphones," while smaller in total volume, exhibit a much darker green hue, indicating a significantly higher average value per order. This highlights a fundamental distinction between high-volume, lower-cost categories ("volume drivers") and high-value, lower-frequency ones ("value drivers").

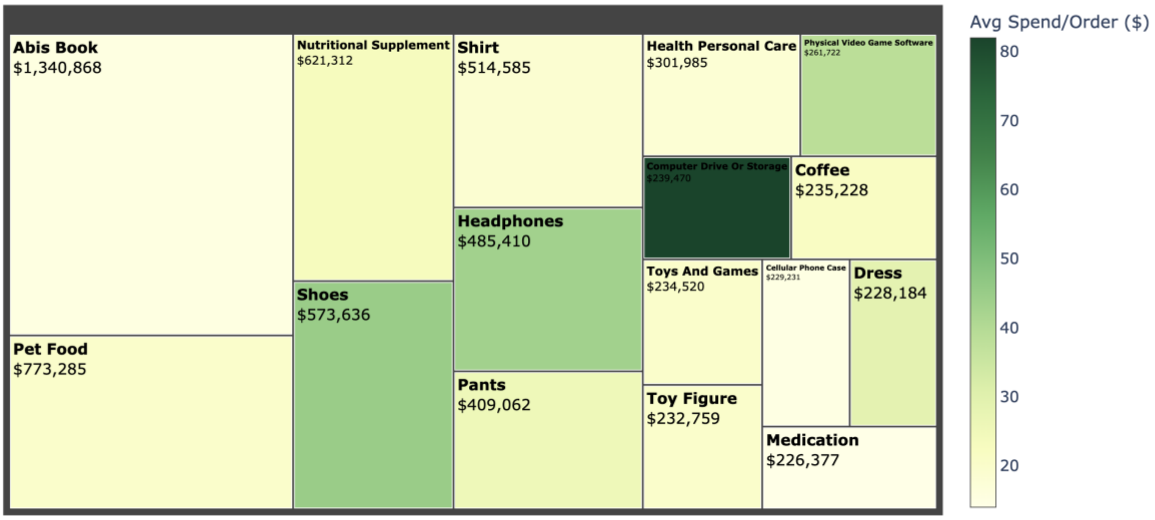

Figure 4.2: Treemap of Product Category Spending

To further dissect these top categories, Figure 4.3 presents a granular breakdown for each, utilizing a "small multiples" approach. This allows for a direct comparison of four key metrics: total spending, total quantity sold, number of unique users, and average spend per order. A comparison between "Abis Book" and "Headphones" is particularly insightful. While books attract the largest user base (4,229 users) and sell in the highest quantity, their average spend per order is relatively low. In contrast, headphones are purchased by fewer users (3,252) but command a much higher average order value, resulting in substantial total spending. This suggests different purchasing logics: books represent a broad-based, frequent, lower-cost category, while headphones signify a more specialized, high-investment purchase. Similarly, categories like "Health Personal Care" and "Shirt" follow the high-user-base, lower-average-value pattern. By combining the macro overview from the treemap with the micro-level details of the small multiples, we can effectively profile and segment product categories, providing a strategic map of the e-commerce landscape.

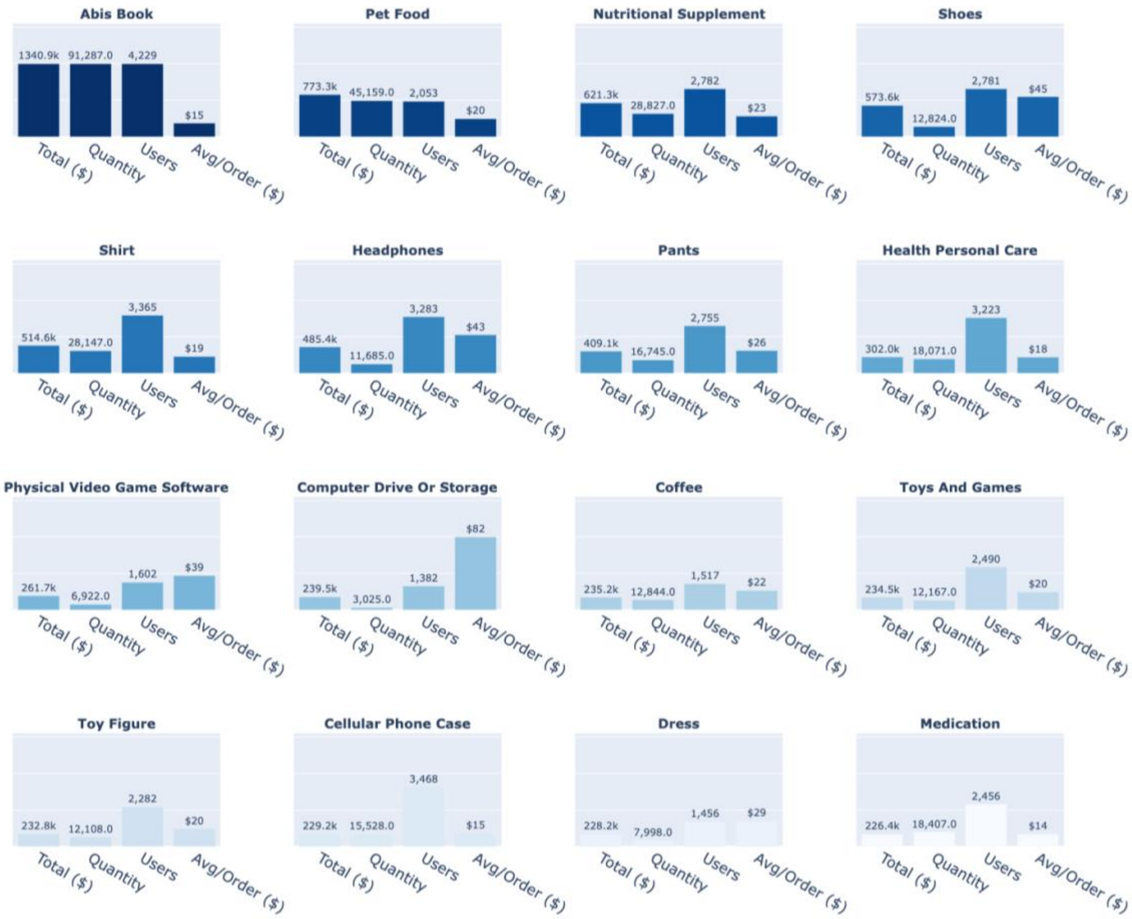

Figure 4.3: Detailed Metrics for Top Product Categories

5 Event-Driven Analysis: Consumption Under External Shocks

While the preceding sections established broad trends and demographic patterns, consumer behavior is often acutely influenced by specific, powerful drivers. This section presents two distinct case studies to explore these event-driven dynamics. First, we examine the impact of an unprecedented external shock—the COVID-19 pandemic—to understand how consumption patterns adapt in times of crisis. Second, we analyze the predictable, cyclical rhythms of seasonality and cultural calendars to uncover the underlying temporal structure of digital commerce.

5.1 Case Study 1: The COVID-19 Pandemic

The COVID-19 pandemic represents one of the most significant societal disruptions in modern history, profoundly altering daily life and, consequently, consumer priorities. The Open E-commerce 1.0 dataset provides a unique, real-time lens through which to observe these shifts. To measure the direct consumer response to the public health crisis, we first analyzed spending on Personal Protective Equipment (PPE), identified through product titles containing keywords such as 'mask' or 'N95'.

As Figure 5.1 illustrates, spending in this category was negligible prior to 2020. The declaration of the global pandemic in March 2020, marked by the vertical line, triggered an immediate and dramatic surge in purchasing that peaked in the subsequent months. This visualization vividly captures the initial wave of consumer adaptation to the new reality. However, the pandemic's impact was not confined to specific product categories. To assess its effect on overall consumer spending behavior, Figure 5.2 presents a box plot comparing the distribution of total spending per user before and after March 2020. The results are striking. The visualization reveals a statistically significant upward shift in the median spending per user during the COVID-19 period. Furthermore, the interquartile range and the spread of outliers are visibly larger, indicating that not only did consumers spend more on average, but their spending behavior also became more heterogeneous. This suggests an acceleration of the general population's shift towards e-commerce for a wider array of goods and services, fundamentally altering the baseline level of digital consumption.

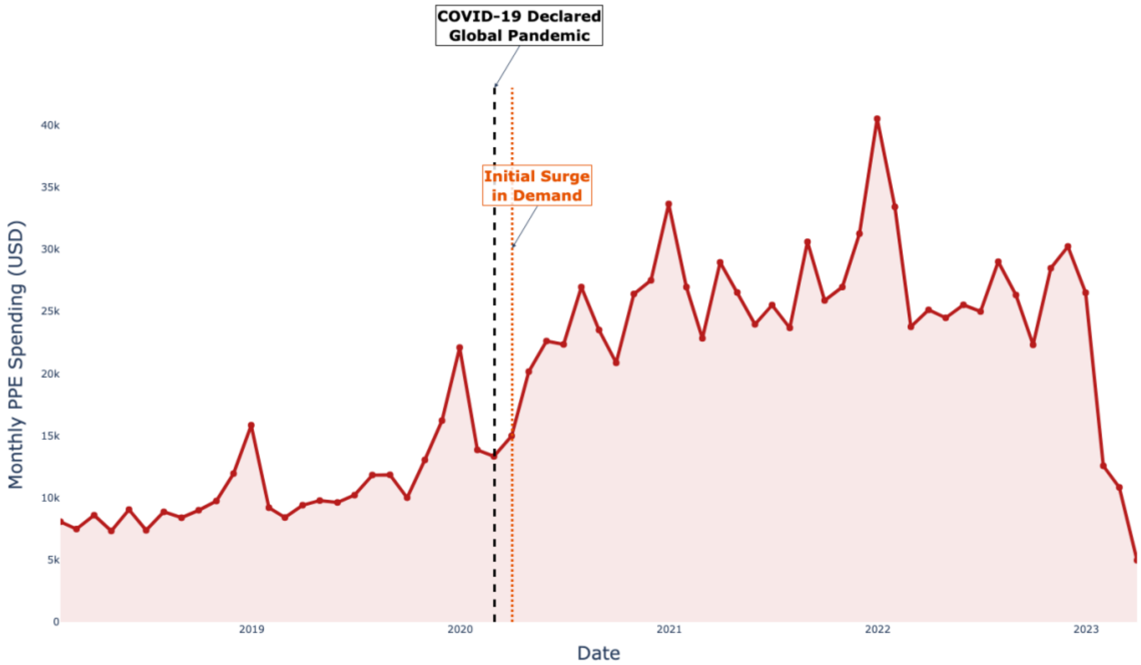

Figure 5.1: Purchasing Surge in Personal Protective Equipment

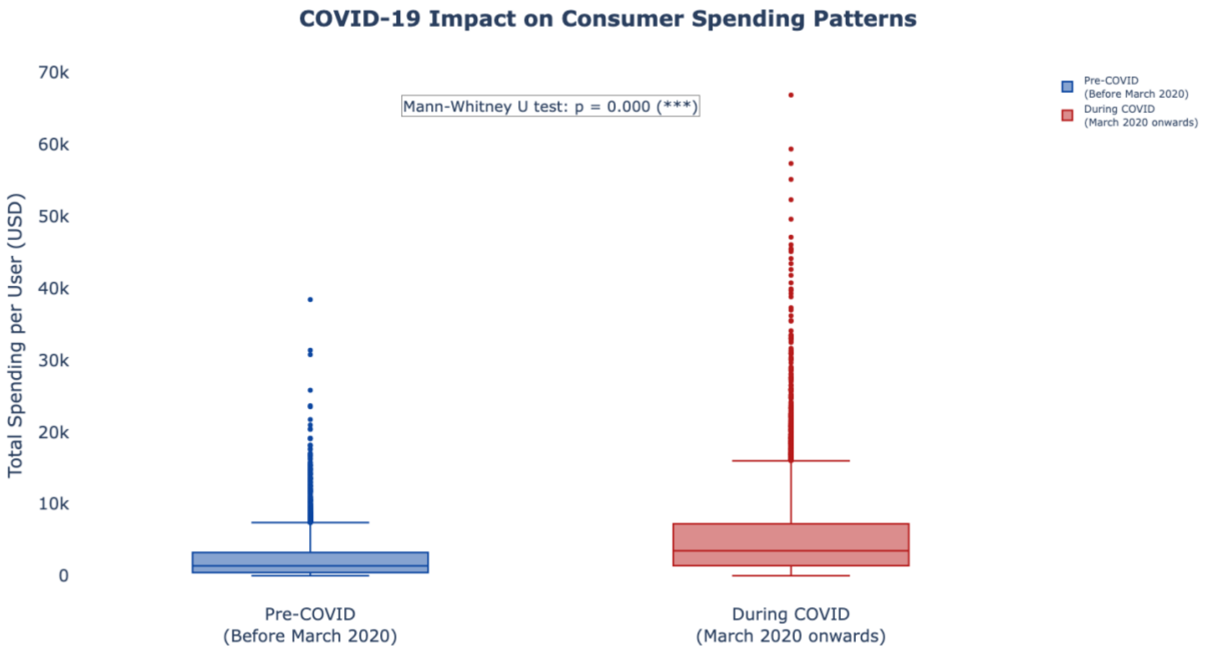

Figure 5.2: Comparative Analysis of User Spending Pre- and During-COVID

5.2 Case Study 2: Seasonality and Daily Rhythms

In contrast to the sudden shock of the pandemic, consumer behavior is also governed by predictable, recurring patterns. A classic illustration of seasonality is found in apparel. Figure 5.3 plots the monthly spending on winter items (e.g., boots) against summer items (e.g., sandals). The two categories exhibit a near-perfect inverse correlation: spending on summer footwear peaks in the warm months of May and June, while spending on winter footwear surges dramatically in the fourth quarter. This clear, opposing seasonality validates the dataset's ability to capture fine-grained, category-specific consumer cycles.

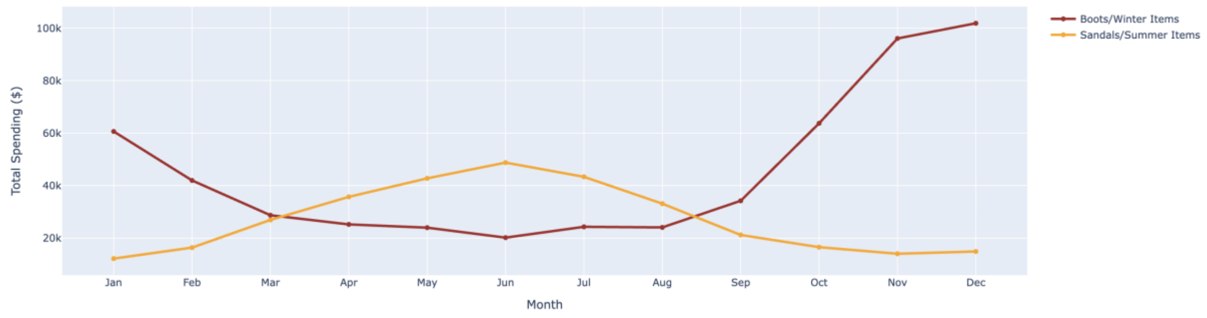

Figure 5.3: Opposing Seasonality in Footwear Purchases

Beyond annual seasons, purchasing behavior also follows micro-rhythms. Figure 5.4 visualizes aggregate spending based on the hour of the day and the day of the week. This reveals that purchasing is not uniformly distributed, but rather concentrated in specific time blocks, such as weekday evenings, potentially reflecting post-work online activity. To pinpoint the impact of specific holidays and modern sales events, Figure 5.5 presents the entire year's daily spending intensity. The darkest bands on the calendar clearly highlight weeks of intense consumer activity. These periods, such as the one around Week 47-48 which includes Black Friday and Cyber Monday, stand in stark contrast to the lower-level, routine spending of ordinary weeks. Together, these visualizations demonstrate that purchasing data contains a multi-layered temporal structure, capturing everything from broad annual seasons to specific, culturally-driven shopping events.

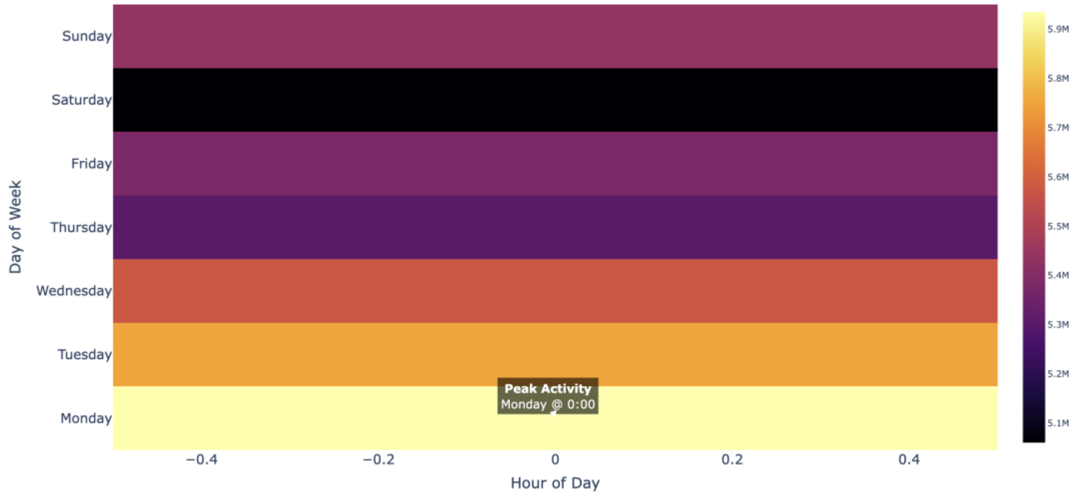

Figure 5.4: Heatmap of Hourly and Daily Spending Patterns

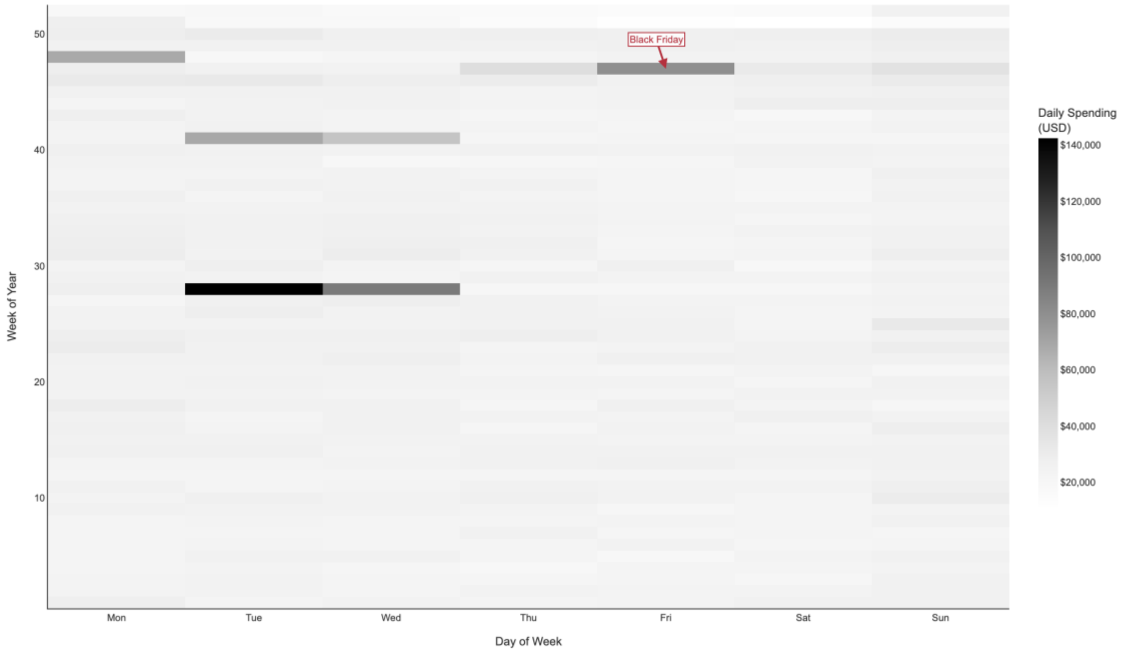

Figure 5.5: 2022 Daily E-commerce Spending Calendar

6 Advanced Behavioral Segmentation and User Archetypes

While previous sections analyzed consumer behavior through the lens of pre-defined demographic segments, this chapter advances the analysis by employing unsupervised machine learning to derive segments directly from behavioral data. This data-driven approach allows us to move beyond simple comparisons and identify naturally occurring clusters of users, or "archetypes," based on their actual purchasing patterns. The objective is to construct a more nuanced and actionable understanding of the consumer base, revealing distinct personas that demographic data alone cannot capture.

6.1 RFM Model for Customer Value Segmentation

To quantify and segment users based on their transactional value, we first implement the widely recognized Recency, Frequency, and Monetary (RFM) model. This framework evaluates each user along three critical dimensions:

- **Recency (R):** How recently a customer has made a purchase.
- **Frequency (F):** How often they make purchases.
- **Monetary (M):** How much money they spend.

A K-Means clustering algorithm was applied to the standardized RFM metrics of each user to partition the cohort into distinct segments. To determine the optimal number of clusters, we employed three validation methods: the Elbow Method, Silhouette Analysis, and Gap Statistic (see Figures 6.0a, 6.0b, and 6.0c respectively), all of which converged to support k=4 as the optimal clustering solution. Figure 6.1 provides a three-dimensional scatter plot that visualizes each user's position in the RFM space, with each point colored according to its assigned cluster. The analysis successfully identified four primary archetypes:

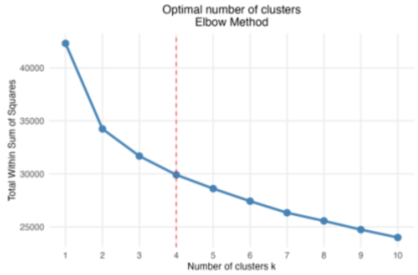

Figure 6.0a

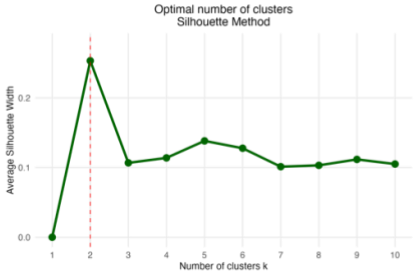

Figure 6.0b

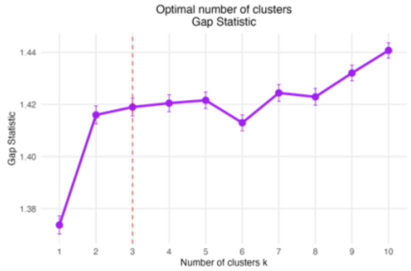

Figure 6.0c

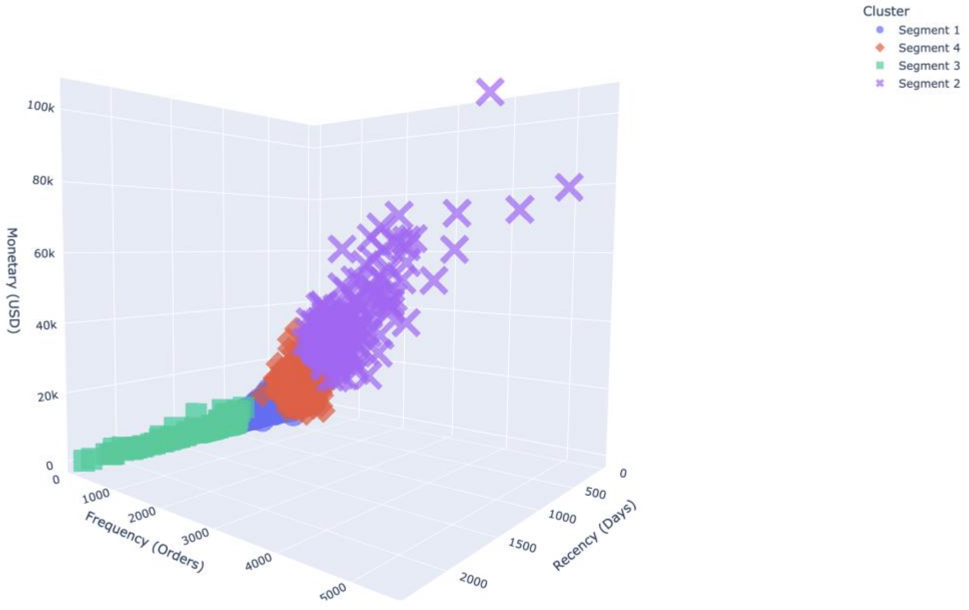

Figure 6.1: 3D Visualization of RFM Segments

- **High-Value Champions (Segment 2, Purple Crosses):** Occupying the high-Frequency and high-Monetary space, this crucial segment represents the most valuable customers. Their consistent and high-volume purchasing behavior forms the commercial backbone of the user base.

- **Loyal Customers (Segment 4, Red Diamonds):** This group is characterized by low Recency (recent purchases) and moderate Frequency and Monetary values. They are actively engaged and represent a stable source of revenue.
- **At-Risk or Lapsed Users (Segment 3, Green Squares):** A significant cluster is found along the high-Recency axis, regardless of their past frequency or monetary value. These users have not made a purchase in a long time and are in danger of churning, making them prime targets for re-engagement campaigns.
- **New or Occasional Shoppers (Segment 1, Blue Circles):** This segment typically clusters at lower Frequency and Monetary values. They represent users with high potential who require nurturing and targeted promotions to increase their loyalty and spending.

6.2 Unsupervised Clustering for Behavioral Personas

To complement the model-driven RFM analysis, we performed an exploratory visualization of the overall behavioral landscape. Figure 6.2 plots each user based on their total number of orders against their total spending. The main scatter plot reveals a strong positive correlation between these two metrics, yet with significant variance. Critically, the marginal histograms along the top and right axes show that both distributions are heavily right-skewed. This confirms a key e-commerce principle: a relatively small number of highly active users are responsible for a disproportionately large share of total orders and revenue.

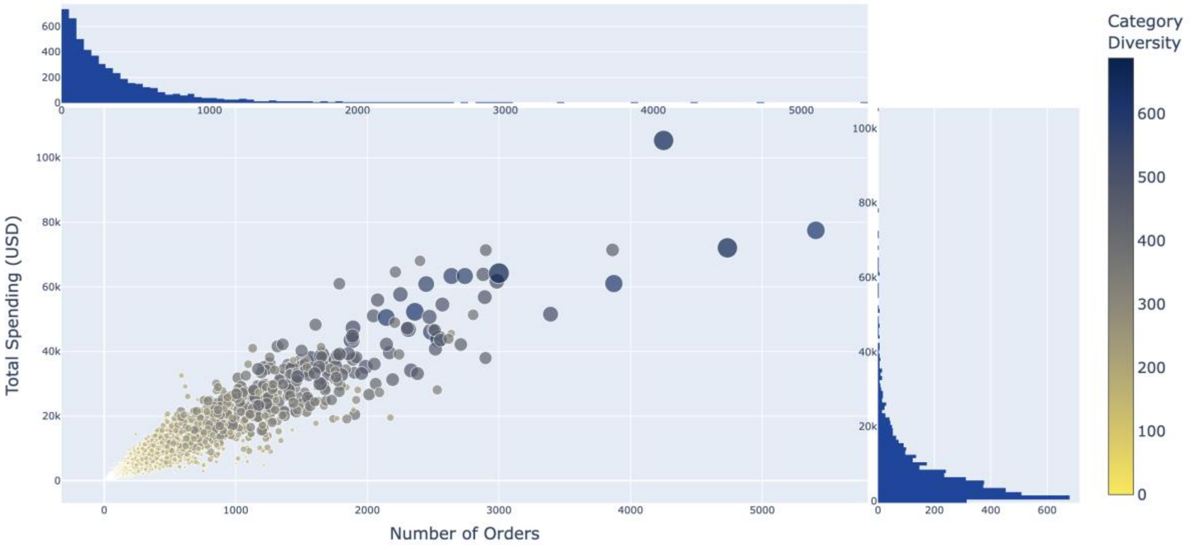

Figure 6.2: Behavioral Distribution with Marginal Histograms

Further depth is added by incorporating a third behavioral dimension: product category diversity. In Figure 6.3, the color of each point corresponds to the number of unique product categories a user has purchased from. It is immediately evident that the highest-spending and most frequent shoppers (located in the upper-right quadrant) are also those who explore the widest range of product categories, as indicated by the bright yellow hues. This powerful insight suggests that the path to creating a high-value customer is not merely about encouraging more purchases, but about deeply integrating the platform into their lives across a diverse array of needs. Together, these visualizations reveal that the consumer base is a heterogeneous ecosystem, composed of distinct archetypes whose value and behavior can be understood and cultivated through data-driven strategies.

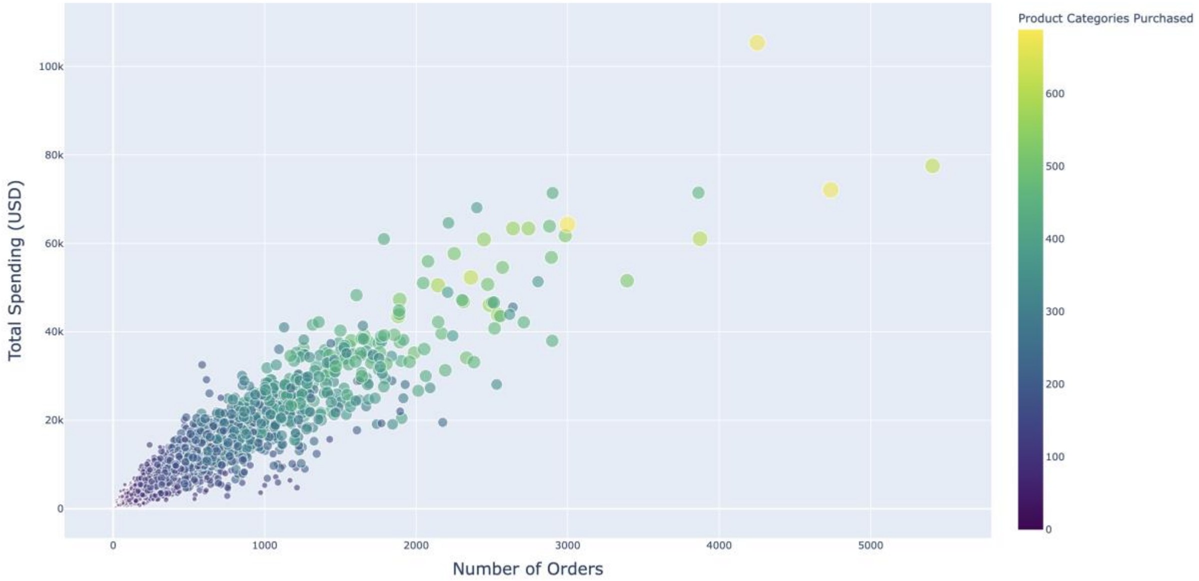

Figure 6.3: Consumer Segmentation by Order Frequency, Spending, and Category Exploration

7 Special Focus: Inferring Domestic Migration Patterns

Beyond its direct application in consumer behavior analysis, the Open E-commerce 1.0 dataset contains latent geographic information that offers a unique proxy for tracking human mobility. This special focus section explores the potential of using multi-year shipping address data to infer patterns of domestic migration within the United States. The objective is to demonstrate how transactional data can serve as a valuable, high-frequency signal for demographic research, complementing traditional census-based methods.

Our methodology involved identifying each user's initial and final state of residence based on their first and last recorded shipping addresses within the dataset's timeframe. These states were then mapped to the four major U.S. census regions (Northeast, Midwest, South, and West) to aggregate state-level moves into broader regional flows.

To visualize these inferred migration pathways, a Sankey diagram was constructed, as presented in Figure 7.1. In this diagram, the nodes represent the four regions, and the connecting flow paths illustrate the movement of users between them. The thickness of each path is directly proportional to the number of individuals inferred to have migrated. The visualization effectively maps the existence and relative scale of these inter-regional movements, providing a compelling proof-of-concept for the viability of this analytical approach.

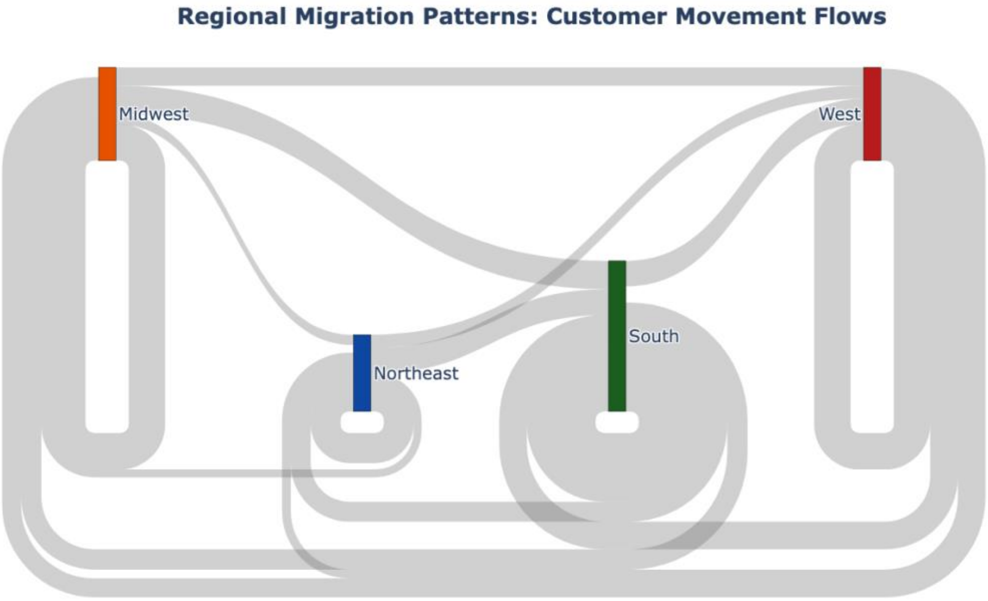

Figure 7.1: Inferred Domestic Migration Flows Between US Regions

The validity of using this dataset as a proxy for migration has been previously established. A comparative analysis with the U.S. Census Bureau's American Community Survey (ACS) found a statistically significant positive correlation, confirming that e-commerce shipping data captures a real demographic signal. It is, however, imperative to acknowledge the limitations of this method. Our approach assumes that a user's first and last shipping addresses accurately represent their primary locations at the beginning and end of the period, which may not capture more complex moving patterns. Furthermore, as noted in the source publication, the absolute number of migrants in a sample of this size is small. Therefore, this dataset is better suited for identifying the *direction* and *relative magnitude* of major migration corridors rather than for precise population counts.

In conclusion, this exploratory analysis demonstrates the remarkable latent value within e-commerce data. It showcases the potential to inform research far beyond its primary purpose, including in fields like demography and urban studies, by using novel proxies to understand complex human behaviors.

8 Conclusion and Future Outlook

This report has undertaken a multi-faceted visual analysis of the Open E-commerce 1.0 dataset, a unique repository of consumer purchase histories enriched with detailed demographic information. By leveraging a suite of visualization techniques, from descriptive statistics to advanced machine learning models, we have moved beyond aggregate trends to dissect the nuanced behaviors of the American digital consumer. This concluding chapter summarizes the principal findings of our investigation, acknowledges its inherent limitations, and outlines promising directions for future research.

8.1 Summary of Key Findings

Our analysis has yielded several key insights into the landscape of modern e-commerce consumption:

1. **Validated Macro-Level Dynamics:** Our cohort's purchasing behavior demonstrated a

strong alignment with broader market trends, including robust long-term growth and powerful, predictable Q4 holiday seasonality. The application of time-series decomposition statistically validated these patterns, separating underlying growth from cyclical effects.

2. **Nuanced Demographic Drivers:** Cross-demographic analysis revealed a critical distinction between high-activity and high-value consumer segments. While younger female cohorts represent the largest volume of purchasing activity, older male cohorts consistently exhibit a higher average spend per transaction, highlighting different forms of value within the user base.
3. **Event-Driven Behavioral Shifts:** The dataset proved highly sensitive to external events. The COVID-19 pandemic acted as a significant catalyst, not only triggering a surge in demand for specific products like PPE but also accelerating a structural shift towards higher overall e-commerce spending. Furthermore, analysis of seasonal and daily data captured the predictable rhythms of consumption, from annual apparel cycles to weekly purchasing cadences.
4. **Data-Driven User Archetypes:** Moving beyond demographics, our application of the RFM model and K-Means clustering successfully identified distinct behavioral archetypes, such as 'High-Value Champions,' 'Loyal Customers,' and 'At-Risk' users. This demonstrates that consumer segmentation can be effectively derived from transactional behavior alone, offering actionable insights for targeted marketing and retention.
5. **Latent Value in Transactional Data:** Our special focus on domestic migration served as a powerful proof-of-concept, demonstrating that shipping data can be used as a novel proxy to infer and visualize macro-level human mobility patterns, showcasing the dataset's value beyond its primary commercial context.

8.2 Limitations

In the spirit of academic rigor, we must acknowledge the limitations of this study:

- **Sample Representativeness:** While macro trends align with market data, the cohort is not a perfectly representative sample of the U.S. population, with known biases such as the underrepresentation of older adults. Conclusions should therefore be interpreted with this context in mind.
- **Methodological Assumptions:** Our analysis relies on specific methodological choices, such as the 99.5th percentile threshold for outlier mitigation and the selection of four clusters in our K-Means model. The inference of migration is also based on a simplified model of first-and-last shipping addresses, which may not capture all real-world complexity.
- **Data Granularity:** The analysis is constrained by the provided product 'Category' labels. A more granular, hierarchical product taxonomy would enable more sophisticated basket analysis and a deeper understanding of product relationships.

8.3 Future Directions

The richness of this dataset opens numerous avenues for future investigation, moving beyond the diagnostic analysis presented herein. An immediate extension lies in the realm of predictive modeling, where the identified user archetypes could serve as a foundation for forecasting customer churn or lifetime value, thereby enabling proactive retention strategies. Furthermore, significant potential resides within the textual data; applying Natural Language Processing (NLP) to the 1.8 million product titles could yield a far more granular product taxonomy and uncover emerging consumer trends that simple category analysis cannot detect. Beyond commercial applications, the dataset's unique linkage of purchasing to self-reported life events and health information invites critical socioeconomic research. Future studies could explore the consumption shocks associated with events like job loss or childbirth, providing valuable data for public health and economic policy. Finally, this dataset provides a rich context for examining the crucial ethical questions of the digital age. Research could be conducted to audit the potential for inferring sensitive attributes from purchasing data alone, contributing to the vital discourse on algorithmic fairness, data privacy, and responsible AI.