Data Analysis and Mining of User Behavior on E-Commerce Platforms

Shuhan Chen

Shuhan.Chen21@student.xjtlu.edu.cn

(*School of AI and Advanced Computing，Xi’an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215018, China*)

***Abstract:*** *The burgeoning e-commerce sector demands profound insights into user behavior to refine marketing strategies and customer service. This study investigates user behavior on e-commerce platforms through extensive data analysis, employing statistical modeling, cluster analysis, and artificial neural networks to interpret the vast data from user interactions and transactions. Advanced data preprocessing techniques are utilized to ensure data integrity, revealing temporal and geographical consumer behavior patterns that underscore regional variations in purchasing power and preferences. This research enhances the understanding of e-commerce user behavior and lays the groundwork for predictive models capable of forecasting future consumer actions, allowing businesses to tailor their marketing approaches and augment the user experience.*

***Keywords:*** *Data analysis; E-Commerce; Statistical modelling;* *Cluster analysis*

1. Introduction

E-commerce constitutes one of the most significant transactional markets for users today, and the industry has seen a gradual maturation[1]. The platform encompasses a comprehensive range of goods available in the market[2]. Influenced by the Internet and the exigencies of the pandemic, online consumption behaviors have come to satisfy users across all age demographics[3]. Amid the continuous proliferation of e-commerce transactions, Internet data is adept at catering to diverse customer groups, including enterprises, individuals, and smaller units. Each customer segment maintains its distinct user base, with varying habits in e-commerce utilization[4].

For e-commerce platforms, analyzing user behavior—through metrics such as transactional actions, volumes, frequency, and timing—is pivotal for enabling strategies like re-marketing, targeted marketing, and promotional campaigns[5]. Concurrently, machine learning has experienced expeditious advancement within the realm of artificial intelligence. By employing machine learning techniques to dissect customer behavior data, one can swiftly extract personalized features from vast datasets and predict future customer actions[6].

Machine learning's utility is profound, with implications for both academic research and practical application, particularly concerning the processing of large-scale Internet data[7]. Its application is extensive, ranging from uncovering hidden patterns in customer purchasing behavior to enhancing the user experience, and simultaneously assisting platforms in identifying potential commercial insights.

1. Application of the Principle of Artificial Neural Network

With the ubiquitous adoption of the internet and the evolution of e-commerce, a burgeoning array of methods for analyzing and predicting customer behavior has emerged, demonstrating notable adaptability and learning capabilities. Tabianan has proposed a method for discerning customer behavior preferences by employing neural networks, genetic algorithms, and the K-means clustering algorithm to analyze customers' browsing paths and extract characteristics like shopping preferences[8]. Similarly, Bucklin and colleagues have delved into clickstream data from enterprise server logs to study customer decisions to either continue browsing or exit, as well as the duration of their visits to web pages[9]. Nasir et al. have developed a proactive customer behavior prediction model using weighted Markov chains that estimates behaviors by analyzing actions, attributes, and the optimal classification of states[10]. They also improved an attribute reduction algorithm based on the difference matrix, applying it to e-commerce consumer behavior prediction to refine the condition attribute set and distill the rules, leading to a new prediction method grounded in rough set theory. Dhandayudam and colleagues have furthered this work by extracting association rules from Web logs using the rough set method of the difference matrix, using these rules to predict customer behavior[11].

In a different vein, Haughton et al. have compared the C&RT and CHAID algorithms for their efficacy in mining potential customer predictions, with results indicating similar performance by both[12]. Their findings revealed that these contemporary methods are more effective than traditional ones, with neural networks providing greater accuracy than multinomial logistic regression. Eunju and colleagues have proposed a multi-classifier method using genetic algorithms for predicting e-commerce customers' purchasing intentions, and Zhu et al. have utilized an enhanced decision tree algorithm to construct a potential customer mining model based on customer behavior data[13].

1. Data Preprocessing and Analysis

# *Data reprocessing*

The dataset used here has a total of 20,183 samples, each containing 9 features and 1 label, as described in Table 1.

***Table 1: Specific Meanings of Each Data Segment***

|  |  |
| --- | --- |
| Field | Comment |
| Order ID | Unique identification for each order |
| Merchant ID | Unique identification for each merchant |
| User ID | Unique identification for each user |
| Payment Date | Payment date of the order, from January 2022 to June 2022 |
| Actual Payment Amount | The actual amount paid for each order, in Yuan (RMB) |
| Shipping Fee | Courier postage per order |
| Province | The province of the user of the current order |
| City | The city of the user of the current order |
| Quantity | Quantity of goods purchased |
| Usage Status | User Coupon Usage: 1 for Used, 0 for Received but Not Used, None for Not Received |

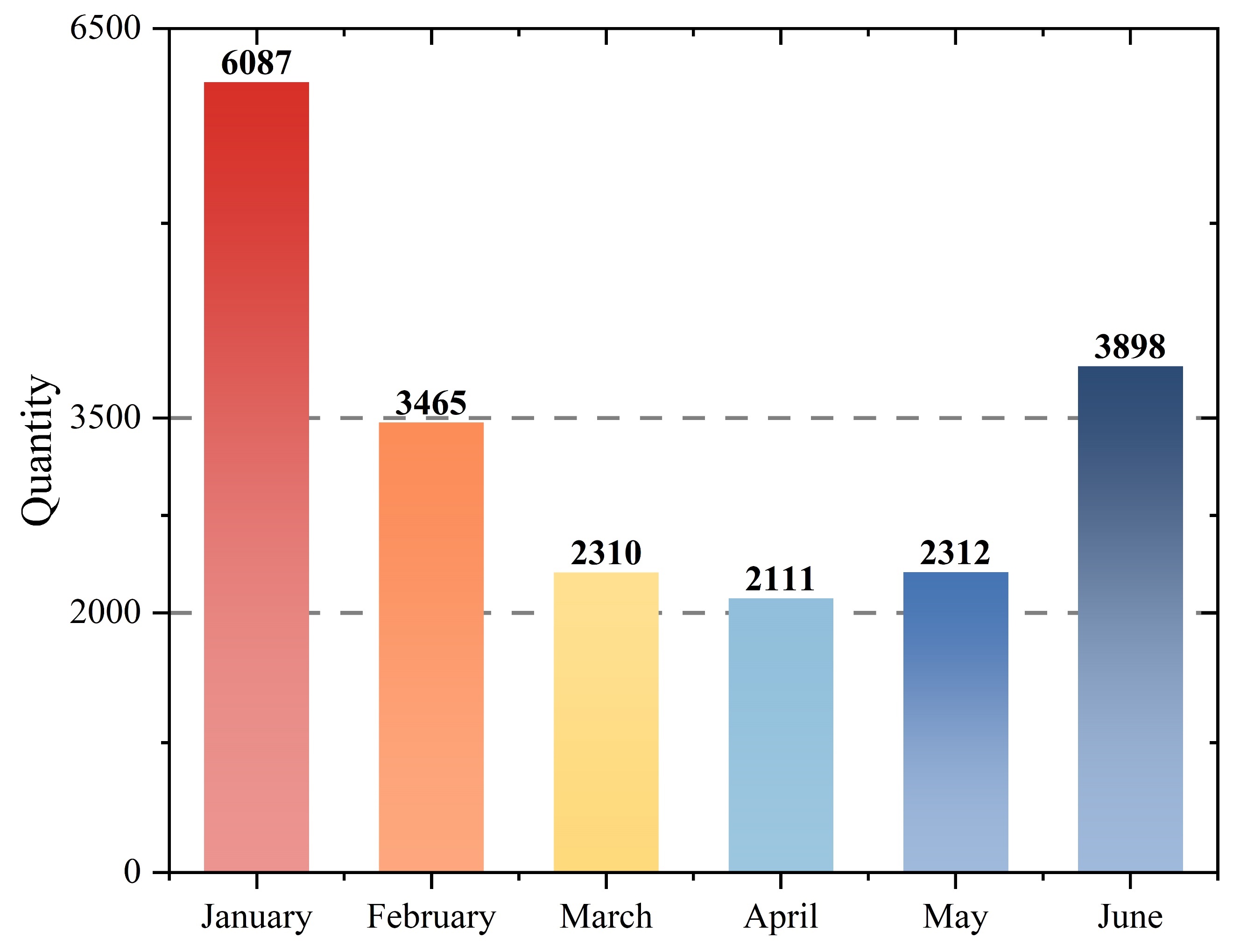
Initially, the three attributes “order ID,” “merchant ID,” and “user ID” function as identifiers and are not indicative of user consumption behavior; consequently, they have been excluded from the analysis. Subsequently, the two attributes “province” and “city” represent categorical variables and have been transformed into quantitative variables to facilitate subsequent statistical analysis. Furthermore, regarding the attribute “Usage Status,” the designation “None” signifies the non-receipt of a coupon. For analytical consistency, “None” has been coded as “2”.

Moreover, a comprehensive review of the dataset for any duplicates or missing entries yielded no such instances, confirming the integrity of the data. With these steps, the data preprocessing phase has been concluded.

# *Data Analysis*

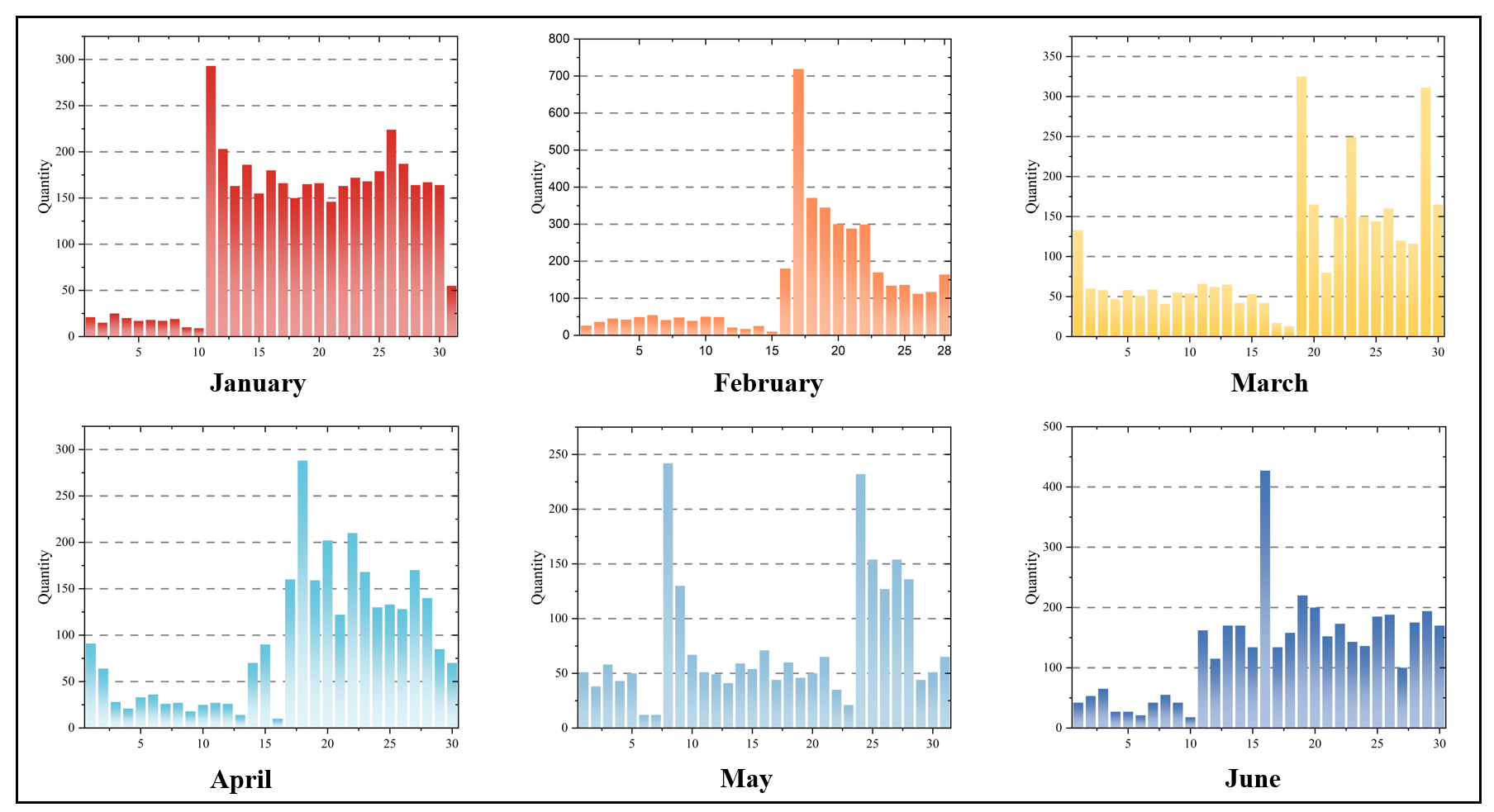
For the "user consumption behavior," a preliminary statistical analysis can be conducted across three dimensions: time, place, and amount.

Initially, we examine the temporal characteristics by segmenting the data into units of "month," "day," and "hour." The findings are presented in Figure 1.



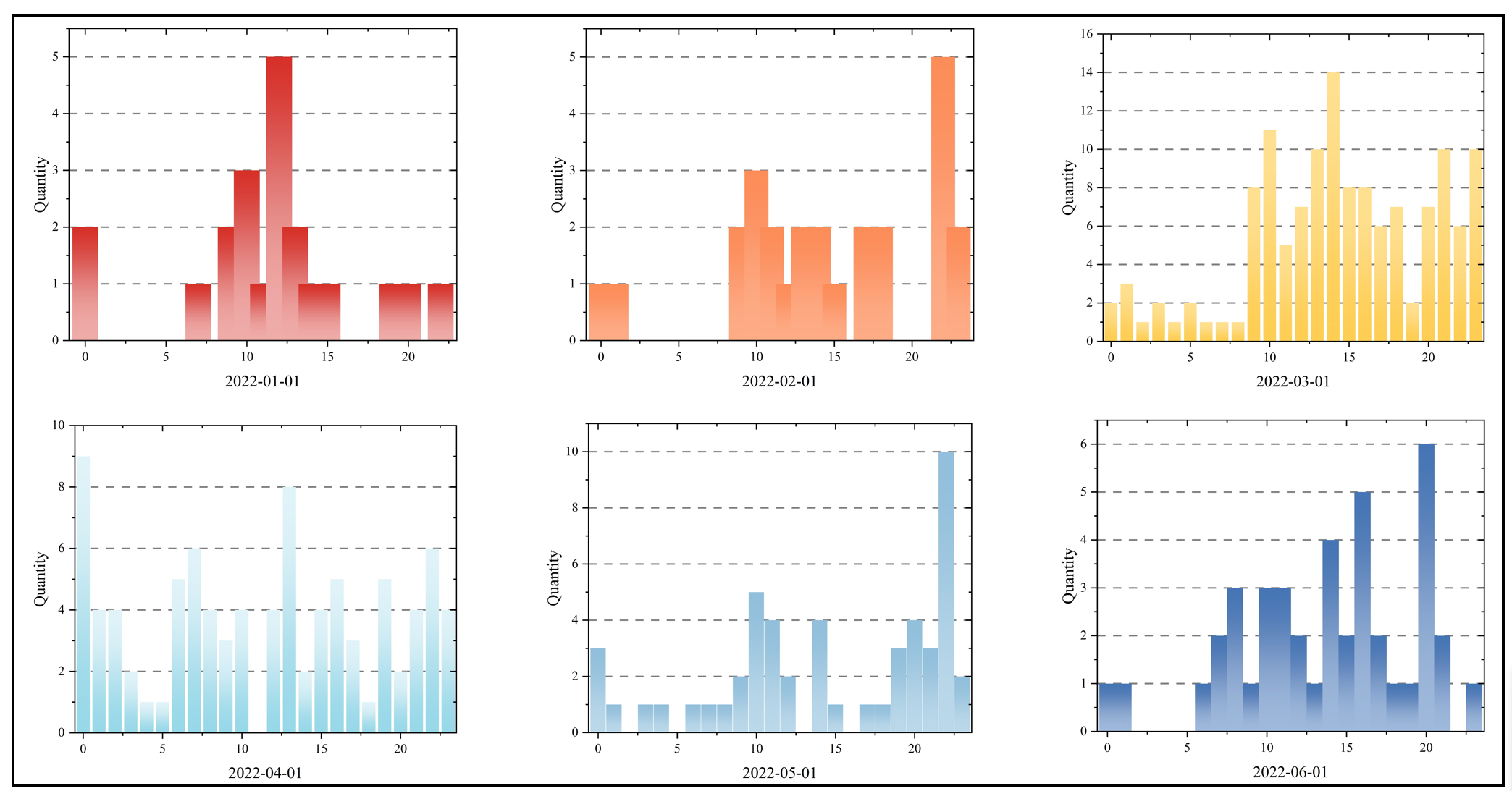
*Figure 1. Number of Orders by Month*

Figure 1 reveals that the product's sales volume exhibits a pattern of initial decline followed by an increase over the months. Notably, sales peak in January and June, while reaching their nadir in April. Preliminary analysis suggests that these cyclical variations may be linked to seasonal events, such as holidays and school vacations. For example, increased sales in January could be attributed to New Year promotions, whereas the surge in June might result from summer discounts or end-of-season clearance events. Additionally, consumer behavior could be influenced by anticipation of these promotional periods, possibly deferring purchases in expectation of more favorable prices, which can lead to exceptionally high sales volumes in subsequent months. It can be hypothesized that there is a discernible monthly pattern in sales volume changes. Further investigation is warranted to ascertain if there are consistent intra-monthly sales patterns over time.



*Figure 2. Number of Orders by Day*

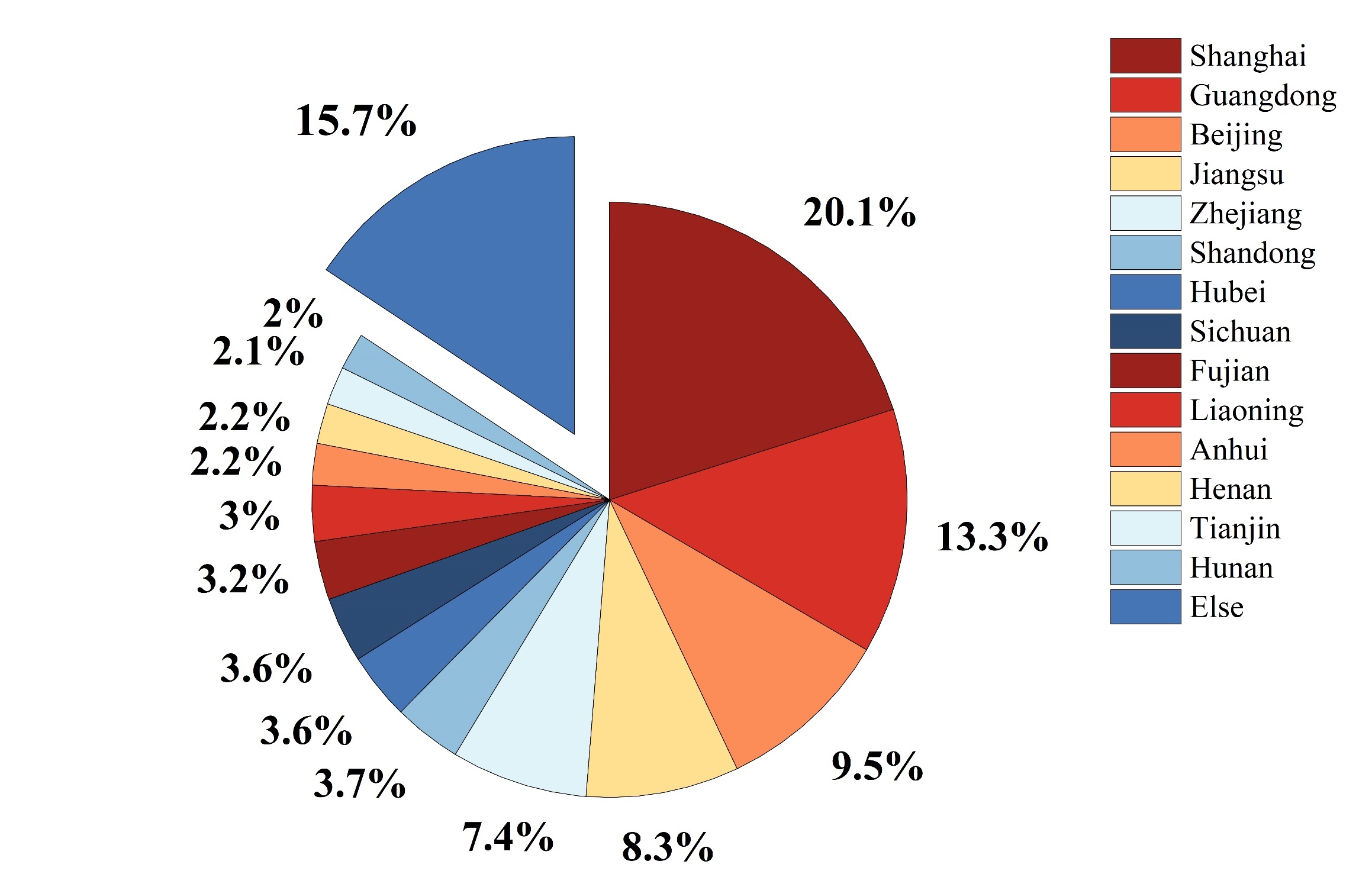
Figure 2 demonstrates that, within each month, sales volumes in the second half consistently exceed those in the first half, with this trend being most pronounced from January to March when the total sales for the first half may not equal the sales of a single day in the latter half. The analysis indicates that the timing of salary payments, typically at the end of the month or the beginning of the next, may lead consumers to defer significant purchases until they have received their income. The distribution of tax returns, bonuses, and other seasonal incomes at the start of the year may further elucidate the consumption patterns observed from January to March. Additionally, the lower spending in early January could be a result of financial recovery from holiday expenses, whereas the latter half of the month may experience a boost from New Year promotions. Moreover, the sales peak around Valentine's Day on February 14th could contribute to the increased sales in the first half of February. To investigate daily sales variations, further analysis has been undertaken, focusing on the first day of each month for brevity. The findings are detailed in Figure 3.



*Figure 3. Number of Orders by Hour*

Figure 3 shows that within each day, sales volumes peak from 10 AM to 12 PM and from 10 PM to 12 AM. The increase during 10 AM to 12 PM may correspond to common break times in many work environments, a period when individuals couldengage in online shopping. Similarly, the spike from 10 PM to 12 AM possibly aligns with the typical post-work relaxation period, which is also a high time for online shopping activity. Furthermore, the rise in evening shopping could be linked to decision fatigue at the end of the day, potentially leading to more impulsive purchasing behavior.

After analyzing how sales volume varies over time, the analysis now shifts to examine the relationship between geographical location and sales volume. The objective of this subsequent analysis is to determine the influence of different regions or commercial centers on the product's sales volume. Figure 4 illustrates the statistical results of this investigation.

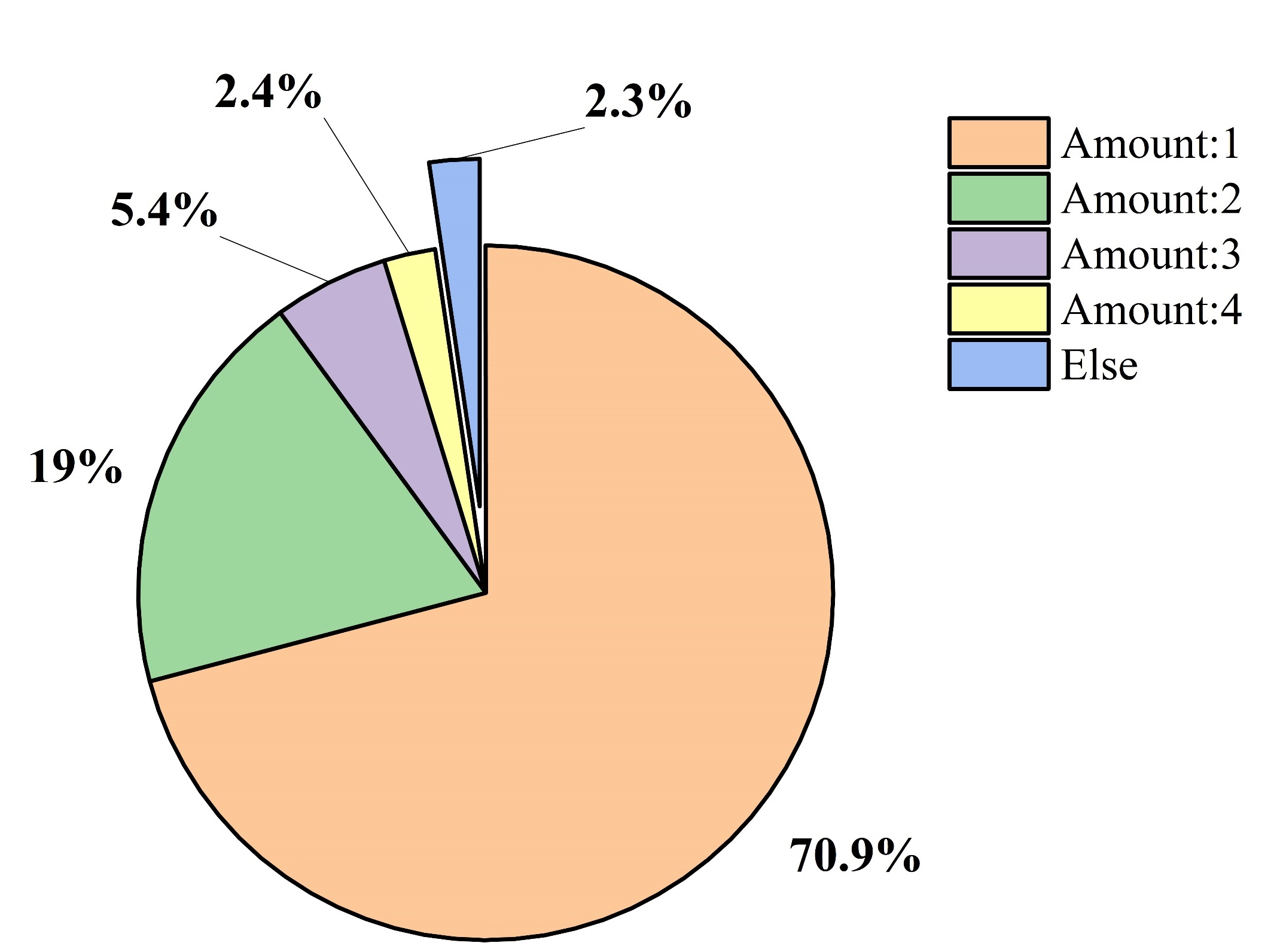


*Figure 4. Sales Volume Proportions by Province*

Figure 4 indicates that the highest sales volumes are recorded in Shanghai, followed by Guangdong Province. Beijing, Jiangsu Province, and Zhejiang Province also demonstrate robust sales. Conversely, Tibet, Qinghai, and Ningxia report the lowest sales volumes. The ensuing analysis explores the financial and urban development factors:

1. **Economic Development Level:** Shanghai, recognized as one of China's most economically prosperous cities with substantial GDP per capita, and Guangdong Province, an economic hub, lead in sales. Beijing, Jiangsu, and Zhejiang follow suit as economically advanced regions. Typically, a higher level of economic development correlates with increased consumer spending capacity and purchasing power. Conversely, Tibet, Qinghai, and Ningxia, being economically less developed, have lower per capita disposable incomes and, consequently, reduced purchasing power.
2. **Population Structure and Density:** The populous and dense regions of Shanghai, Guangdong, Beijing, Jiangsu, and Zhejiang contrast with the sparser populations of Tibet, Qinghai, and Ningxia. A higher population density often equates to a larger market size and, by extension, higher sales volumes.
3. **Consumer Preferences and Lifestyle:** In more developed regions, consumers may exhibit a greater propensity to embrace new products and technologies, contributing to elevated sales volumes. Urbanization and contemporary lifestyles can lead to diversified consumer behaviors, while less developed regions may adhere to more traditional and conservative consumption patterns.
4. **External Investment and Business Activities:** Regions with advanced development tend to attract greater external investment and business establishment, fostering employment and enhancing residents' income levels, which in turn supports increased purchasing power. In contrast, limited investment and business activity can curtail consumption growth in less economically developed regions.

After concluding the analysis of the relationship between location and sales volume, the analysis now turns to a detailed examination of consumer purchase quantities. The objective of this stage is to derive insights into consumer behavior within individual transactions, with a particular emphasis on purchase quantity. By thoroughly analyzing these data, we can uncover patterns in consumer purchasing habits and preferences. This, in turn, provides valuable information for refining product positioning, optimizing inventory management, and developing targeted sales strategies.



*Figure 5. Proportions of Purchase Quantities*

The data from Figure 5 indicates that the majority of consumers typically purchase only a single item, which represents approximately 70% of total purchases. Consumers buying two items constitute the next significant proportion, whereas those purchasing more than two items account for less than 10% of the total. Economic analysis of these trends suggests that for certain goods, frequent multiple purchases may be unnecessary. For instance, durable or high-value items are often bought less frequently. Conversely, products that are daily necessities are likely to be purchased regularly but in quantities that meet immediate needs. In the case of non-essential goods, even smaller quantities might be purchased.

1. Disclosed Artificial Intelligence IoT System Based on Container Virtualization

In the previous section, we conducted a detailed analysis of the changes in order quantities across three key dimensions: time, geographical location, and order size. This involved examining time-series fluctuations, assessing the influence of geographical location on sales volume, and analyzing the distribution of order sizes across different regions. By integrating these dimensions, we identified temporal trends in order quantity, as well as the relationships between location and sales volume, and between sales density and order size in various areas.

Building on the results from this comprehensive analysis, the following section will leverage these findings to construct detailed consumer personas. A consumer persona is a representation of a user, crafted from observed behaviors, preferences, and demographic characteristics, which enables businesses to gain a deeper understanding of their target customer segments.

# *Cluster analysis*

Initially, the dataset undergoes hierarchical clustering, where a hierarchical tree of clusters is constructed by calculating and comparing the distances — typically Euclidean or other measures of similarity — between data points. The history of these distance values, which increase at each level of the clustering hierarchy, can be tracked during the process. The dendrogram illustrates significant changes in the merging of clusters, revealing potential optimal cluster counts at the “elbow point”, where a pronounced increase in distance values suggests a decrease in intra-cluster similarity and an increase in inter-cluster differentiation.

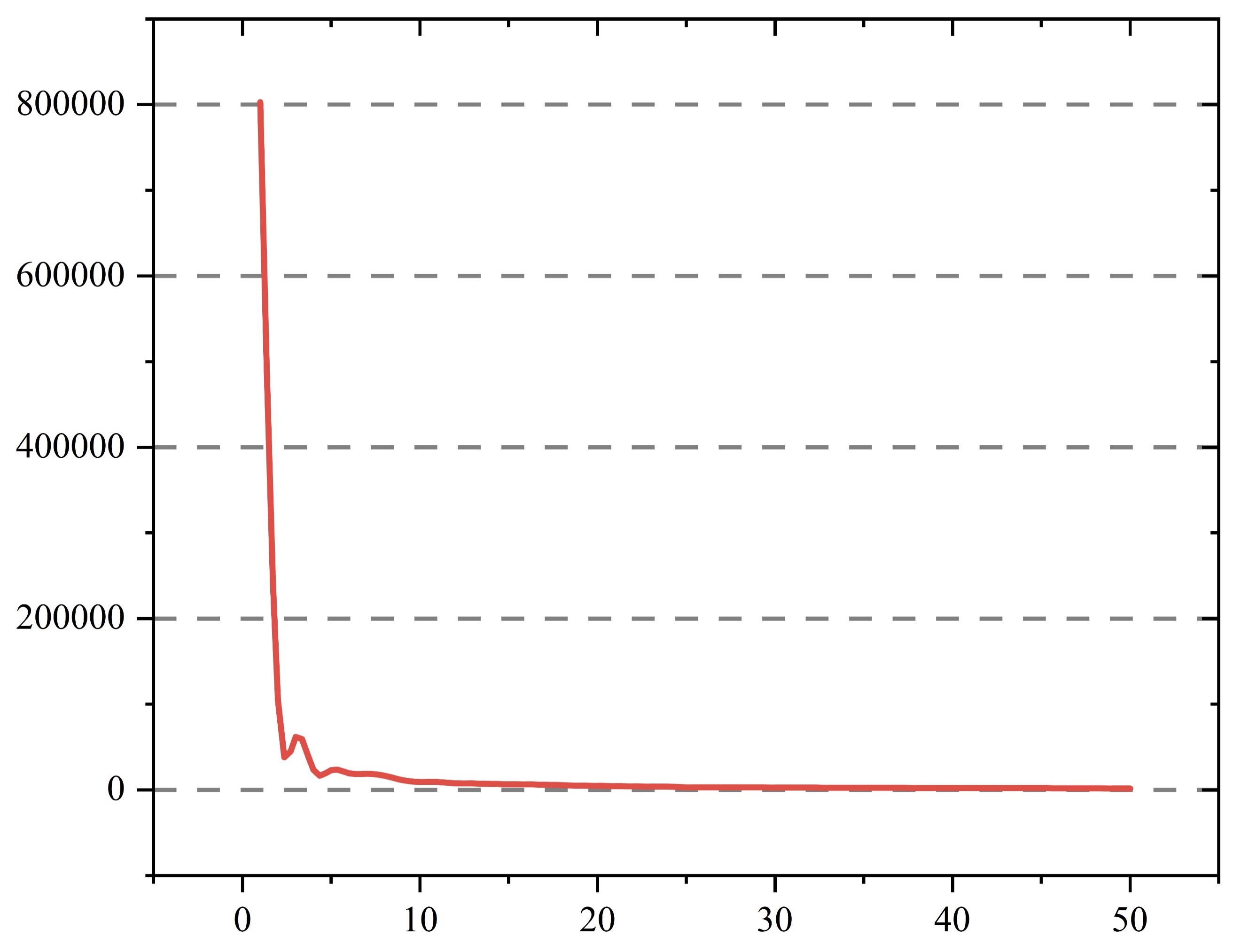
In the formula, and are any two points in an -dimensional space, with their coordinates being and , respectively.

Upon identifying the optimal number of clusters, the dataset is re-clustered using the algorithm. This enhanced version of selects initial cluster centers more precisely, improving both the quality of clustering and the convergence rate. The algorithm iteratively assigns data points to the nearest cluster center and recalculates the centers until they stabilize, indicating convergence to a final cluster division.

These stable clusters, obtained through , are then analyzed to identify common characteristics and behavioral patterns within each group. The insights drawn from these patterns underpin the construction of consumer personas, with each cluster corresponding to a distinct user group characterized by specific purchasing behaviors, preferences, and needs. Utilizing these personas enables businesses to hone their marketing strategies and customize their product offerings to better suit each distinct market segment.

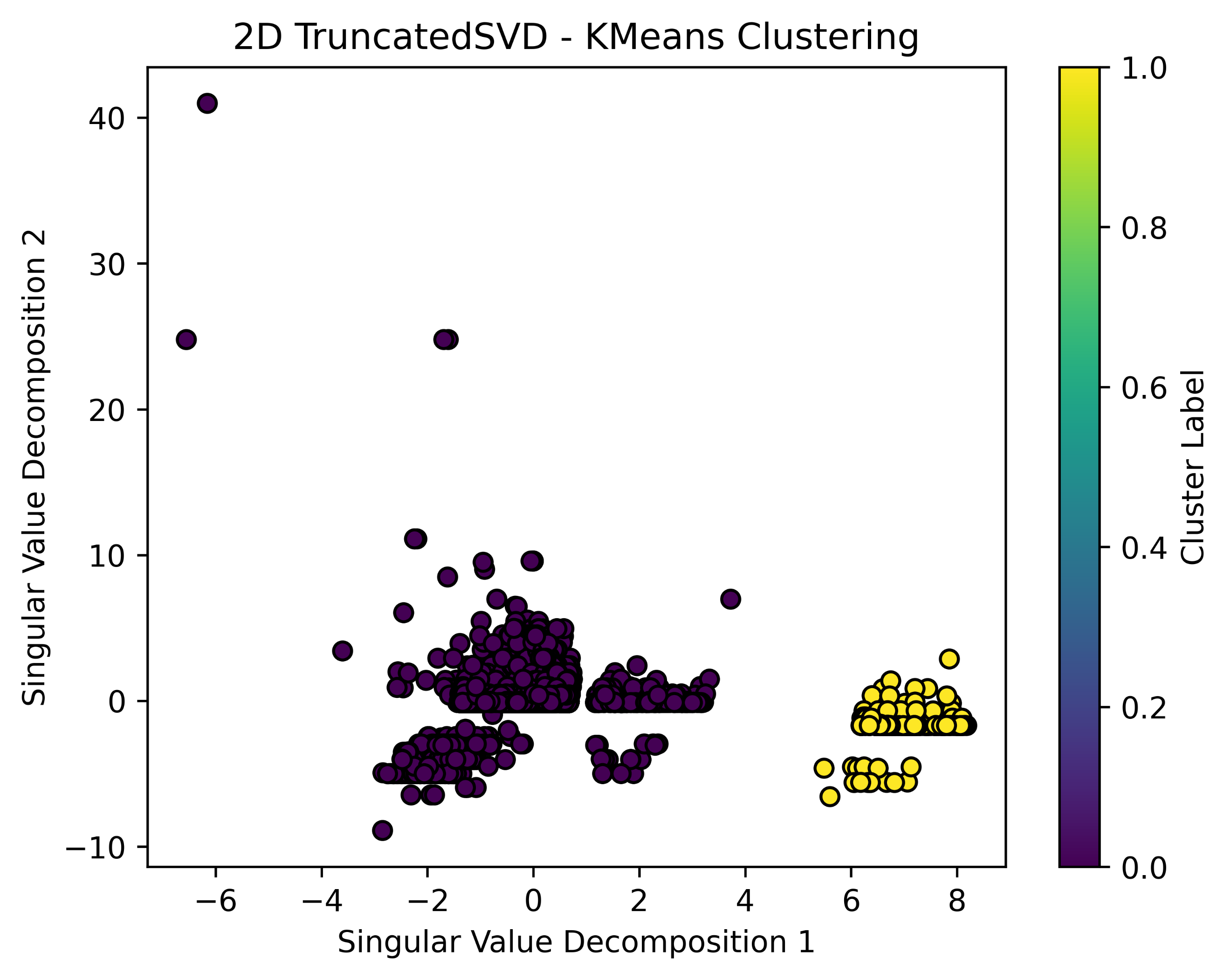
# *Clustering results*

First, the entire dataset is clustered using a hierarchical clustering algorithm, and the resulting trend of the 'distance values,' which reflects the similarity between data points, is analyzed and presented in Figure 6.



*Figure 6. Trends in distance values*

From Figure 6, it is apparent that, based on the selected criteria (such as the elbow method), the optimal clustering effect is achieved with 2 clusters. Therefore, we will proceed by selecting 2 clusters for the algorithm to perform a comprehensive clustering analysis of the entire dataset.

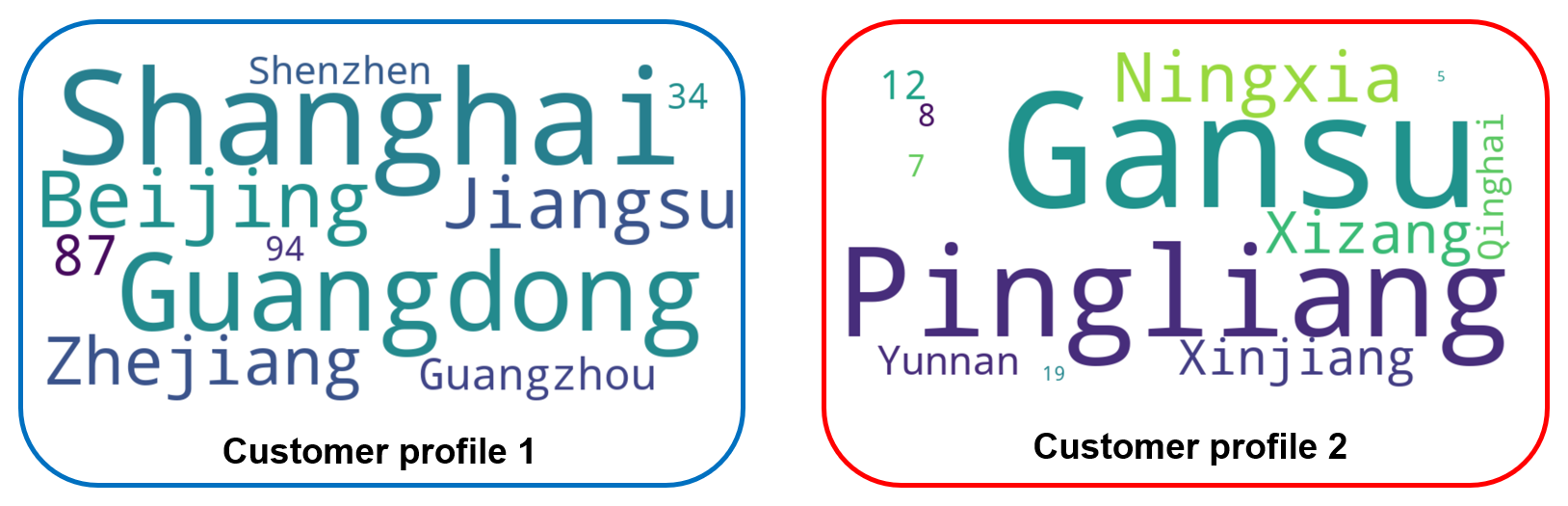


*Figure 7. Trends in distance values*

Figure 7 displays the results after clustering, illustrating that the inter-cluster distance is significant, which suggests well-defined clusters. Subsequently, these clusters will be employed to delineate distinct customer profile.

# *Customer profile*

Based on the results of the cluster analysis, consumer personas representing two distinct consumer behavior patterns have been established. These personas highlight significant differences within the consumer groups and are instrumental in identifying two primary consumer sub-markets. The detailed outcomes, including the key characteristics and behaviors of these sub-markets, are presented in Figure 8.



*Figure 8. Customer profiles*

Analysis of the information presented in Figure 8, when combined with economic knowledge, allows for the inference that Beijing, Shanghai, Zhejiang, and Hangzhou are among the more economically advanced regions in China. These areas feature well-developed industries, high income levels, strong purchasing power, and advanced consumption patterns. Consequently, higher consumer spending amounts can be expected in the user personas from these regions.

In contrast, provinces such as Tibet, Xinjiang, Qinghai, and Gansu, located in western China, generally exhibit lower levels of economic development compared to the eastern coastal areas. Consequently, the income levels and consumption capabilities of the residents in these regions are comparatively lower, leading to smaller consumer spending amounts.

Furthermore, consumer spending power is typically closely correlated with residents' income levels. In regions with higher incomes, consumers tend to purchase higher-priced goods and services, resulting in greater consumer spending amounts. On the other hand, in economically less developed areas, residents have lower disposable incomes, which limits their purchasing power and, accordingly, the consumer spending amounts are relatively smaller.

Additionally, financial services are often more developed and widespread in economically advanced regions, encompassing credit products, investment tools, and payment systems. This development can facilitate residents' consumption and investment activities. In contrast, the absence of such services in less developed areas may constrain consumers' access to credit and payment convenience, thereby influencing consumption levels.

Lastly, the policy framework and level of support in different regions can significantly impact consumer behavior. For example, tax incentives for specific industries, consumer subsidies, and infrastructure development can all catalyze the growth of local consumer markets.

1. Conclusion

In conclusion, our study equips e-commerce platforms with a more nuanced understanding of user behavior and furnishes them with actionable insights to refine marketing initiatives. Future research should persist in examining the evolving landscape of consumer behavior, taking into account nascent trends and technological advancements that may reshape the e-commerce domain. Additionally, incorporating real-time data analysis could advance the accuracy of behavioral forecasts and the efficacy of marketing endeavors. The confluence of sophisticated data analytics and strategic business applications holds the promise of propelling the e-commerce industry to unprecedented levels of achievement and consumer contentment.

References

1. Ponzoa, J.M. and A. Erdmann, *E-Commerce Customer Attraction: Digital Marketing Techniques, Evolution and Dynamics across Firms.* Journal of Promotion Management, 2021, **27**(5): 697-715.

2. Vatsa, P., J. Li, P.Q. Luu, et al., *Internet use and consumption diversity: Evidence from rural China.* Review of Development Economics, 2023, **27**(3): 1287-1308.

3. Wandoko, W. and I.E. Panggati. *Building e-loyalty Toward online Shop in Indonesia Through e-Trust Amid Pandemic COVID-19*. in *2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS)*. 2021.

4. Vamsee, C.S., D. Rakesh, I. Prathyusha, et al. *Demographic and Psychographic Customer Segmentation for Ecommerce Applications*. in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. 2023.

5. Alojail, M. and S. Bhatia, *A Novel Technique for Behavioral Analytics Using Ensemble Learning Algorithms in E-Commerce.* IEEE Access, 2020, **8**: 150072-150080.

6. Rai, S., A. Gupta, A. Anand, et al. *Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods*. in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2019.

7. Zan, C., *Development of e-commerce Big data model based on machine learning and user recommendation algorithm.* International Journal of System Assurance Engineering and Management, 2023.

8. Tabianan, K., S. Velu, and V. Ravi *K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data*. Sustainability, 2022. **14**, DOI: 10.3390/su14127243.

9. Bucklin, R.E. and C. Sismeiro, *Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing.* Journal of Interactive Marketing, 2009, **23**(1): 35-48.

10. Nasir, M. and C.I. Ezeife, *Semantic enhanced Markov model for sequential E-commerce product recommendation.* International Journal of Data Science and Analytics, 2023, **15**(1): 67-91.

11. Dhandayudam, P. and I. Krishnamurthi *Customer Behavior Analysis Using Rough Set Approach*. Journal of Theoretical and Applied Electronic Commerce Research, 2013. **8**, 21-33 DOI: 10.4067/S0718-18762013000200003.

12. Haughton, D. and S. Oulabi, *Direct marketing modeling with CART and CHAID.* Journal of Direct Marketing, 1997, **11**(4): 42-52.

13. Zhu, Y., A. Garai, R. Karmakar, et al., *Customer-centric policies for environmentally sustainable manufacturing of deteriorating items with varying quality control practices under disruptions.* Computers & Industrial Engineering, 2024, **189**: 109895.