**Exploratory Data Analysis on E-Commerce Data**

**1. Introduction**

**1.1 Purpose of the Project**

The e-commerce sector relies heavily on large amounts of data to understand customer behavior, increase operational efficiency, and gain a competitive advantage. The purpose of this report is to provide insights that support business strategies by analyzing various datasets obtained from an e-commerce platform. The analyses cover topics such as customer demographics, geographical distributions, order details, payment methods, and product and seller performances. This information will help businesses make strategic decisions to improve customer satisfaction, optimize sales, and enhance overall operational performance.

**1.2 About the Dataset**

The datasets used in this project include various operational and customer data from an e-commerce platform. Each dataset represents a specific data category, allowing meaningful relationships to be established between these categories. Here are the datasets and their contents:

* **Customers Table**: This table contains customer information, including customer IDs, unique customer IDs, postal code prefixes, city, and state information. It serves as a primary data source for analyzing the demographic distribution and geographical locations of customers.
* **Geolocation Table**: This table contains geographical information, including postal code prefixes, latitude and longitude coordinates, city, and state information. It is used to analyze the geographical distribution of orders and delivery processes.
* **Order Items Table**: This table contains information about order items, including order IDs, order item IDs, product and seller IDs, shipping limit dates, prices, and shipping values. This information is critical for analyzing sales performance and shipping costs.
* **Order Payments Table**: This table contains payment information, including order IDs, payment sequence, payment types, installment counts, and payment values. It is used to analyze the distribution of payment methods and payment behaviors.
* **Order Reviews Table**: This table contains information about order reviews, including review IDs, order IDs, review scores, review creation dates, and response timestamps. It serves as a primary data source for analyzing customer satisfaction and feedback processes.
* **Orders Table**: This table contains order information, including order IDs, customer IDs, order status, order creation dates, approval dates, delivery dates, and estimated delivery dates. It is used to analyze order processes and delivery performance.
* **Products Table**: This table contains product information, including product IDs, product category names, product name lengths, product description lengths, number of product photos, product weight, and dimensions. It is used to analyze product performance and differences between categories.
* **Sellers Table**: This table contains seller information, including seller IDs, postal code prefixes, city, and state information. It serves as a primary data source for analyzing seller performance and geographical distributions.
* **Product Category Translation Table**: This table contains product category translations, including product category names and their English translations. It is used to match product categories in different languages.

**1.3 Data Analysis Stages**

The data analysis process was conducted by following specific steps and stages. Each stage was carefully prepared and implemented to obtain meaningful insights from the dataset. Here are the steps followed in this process:

**1.3.1 Data Collection and Merging**

The first step in the data analysis process is collecting and merging datasets from different sources. The datasets were provided as CSV files and loaded using Python's pandas library. The merging of datasets was done using common keys, creating a single unified data frame for analysis. This step is crucial for ensuring consistency across the datasets and for conducting accurate analyses.

**1.3.2 Data Cleaning and Preparation**

After merging the datasets, data cleaning operations such as managing missing data, identifying and addressing outliers were performed. This stage is extremely important for ensuring the accuracy and reliability of the analyses. Proper handling of missing data and identification and correction of outliers play a critical role in producing healthy analytical results.

**1.3.3 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is a stage where statistical summaries and visualizations are used to understand the general structure and distribution of the dataset. At this stage, key trends, distributions, and relationships in the dataset were examined. Exploratory analyses help in understanding the general characteristics of the dataset and provide necessary preliminary information for in-depth analyses.

**1.3.4 Specific Analyses and Gaining Insights**

At this stage, in-depth analyses were conducted to address specific business questions and objectives. The analyses covered various topics such as customer segmentation, geographical distributions, order and payment details, and product and seller performances. Each analysis aims to provide meaningful insights that support business strategies and contribute to operational decision-making processes.

* **Customer Analyses**: Analysis of the number of customers by cities and states, segmentation of customers, and examination of customer demographics.
* **Geographical Analyses**: Identification of order density, the relationship between geographical location and delivery times, and determination of the busiest cities.
* **Order and Payment Analyses**: Distribution of payment types, average payment value, the impact of installment count on payment value, and analysis of payment methods.
* **Review Analyses**: Summary statistics of review scores, distribution of review response times, and analysis of customer feedback.
* **Seller and Product Analyses**: Total sales by seller, top-selling products, geographical distribution of sellers, and performance differences between product categories.

**1.3.5 Visualization of Results**

The analysis results were presented using various charts and visuals. These visuals are used to make the insights and findings obtained clearer and more understandable. Visualizations facilitate the interpretation of data and support decision-making processes related to business strategies.

**1.4 Results and Recommendations**

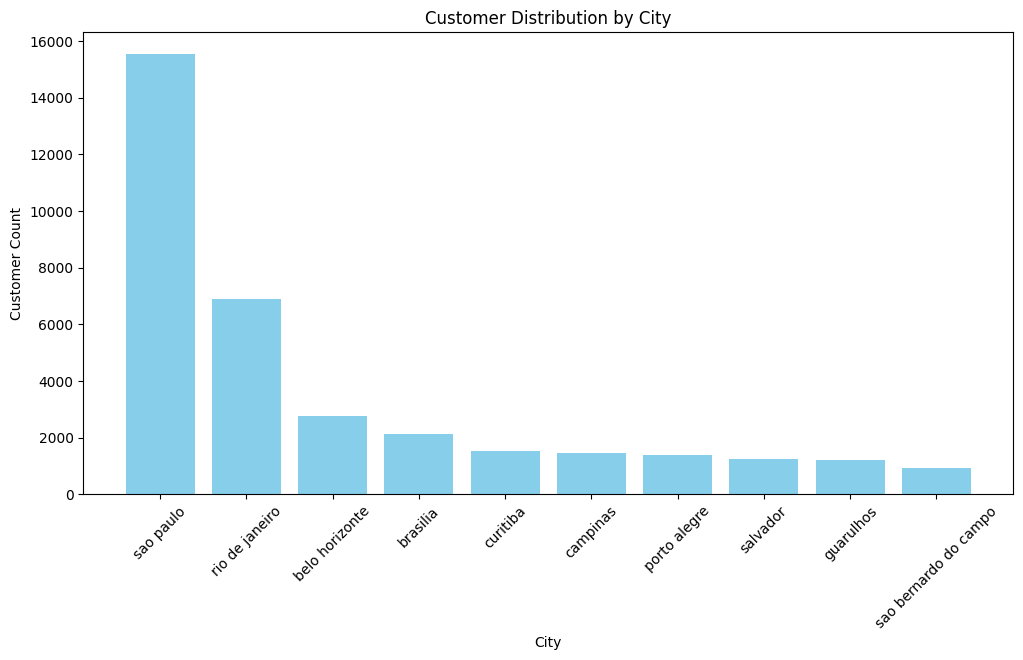
This report comprehensively presents the different stages of data analysis and the findings obtained. As a result of the analyses, meaningful insights have been obtained to improve and enhance business strategies. These findings can be used to make strategic decisions on topics such as customer segmentation, product performance, seller activities, and geographical distributions. For example, identifying the best-selling product categories can help optimize inventory management and marketing strategies. Similarly, customer satisfaction analyses can contribute to the development of customer service and loyalty programs.

**2. Analyses and Findings**

**2.1 Customers Table Analyses**

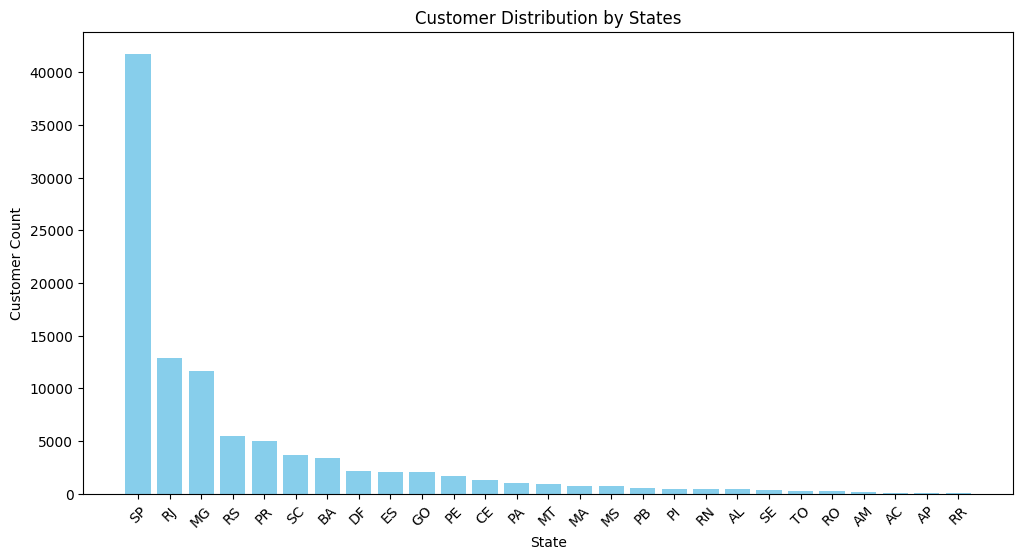
**2.1.1 Distribution of Customer Count by Cities**

Analysis of the distribution of customer count by cities was conducted to determine which cities have the highest customer traffic.



**2.1.2 Distribution of Customer Count by States**

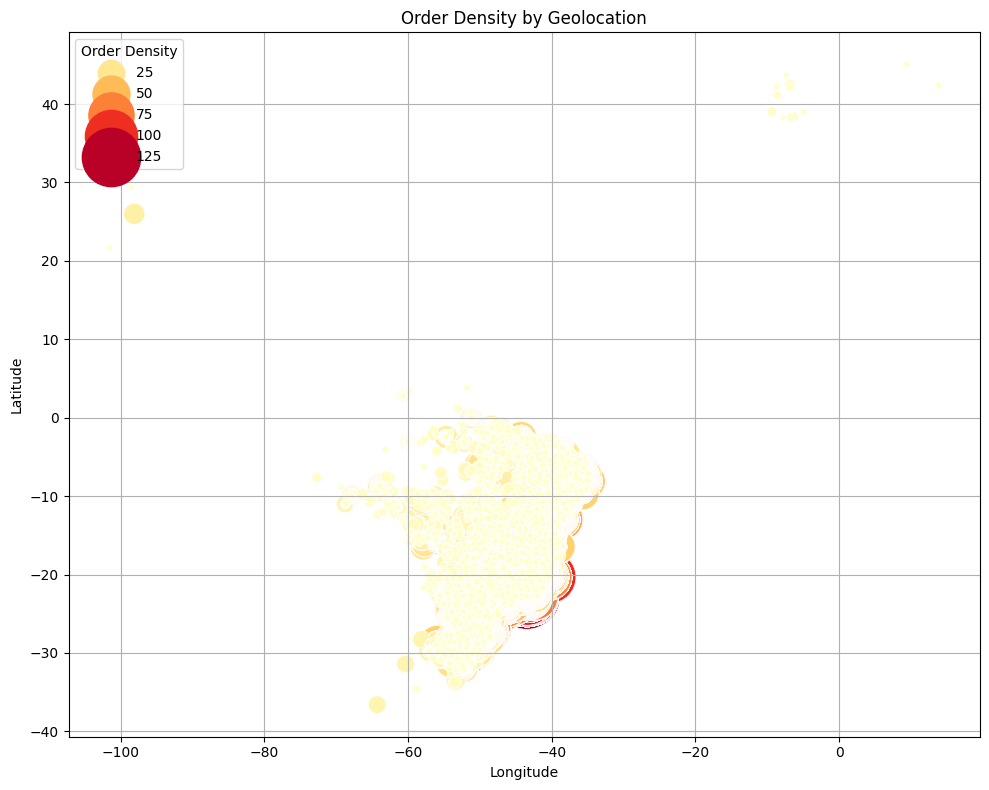
An analysis of the distribution of customer count by states was conducted to identify which states have a higher number of customers.



**2.2 Geolocation Table Analyses**

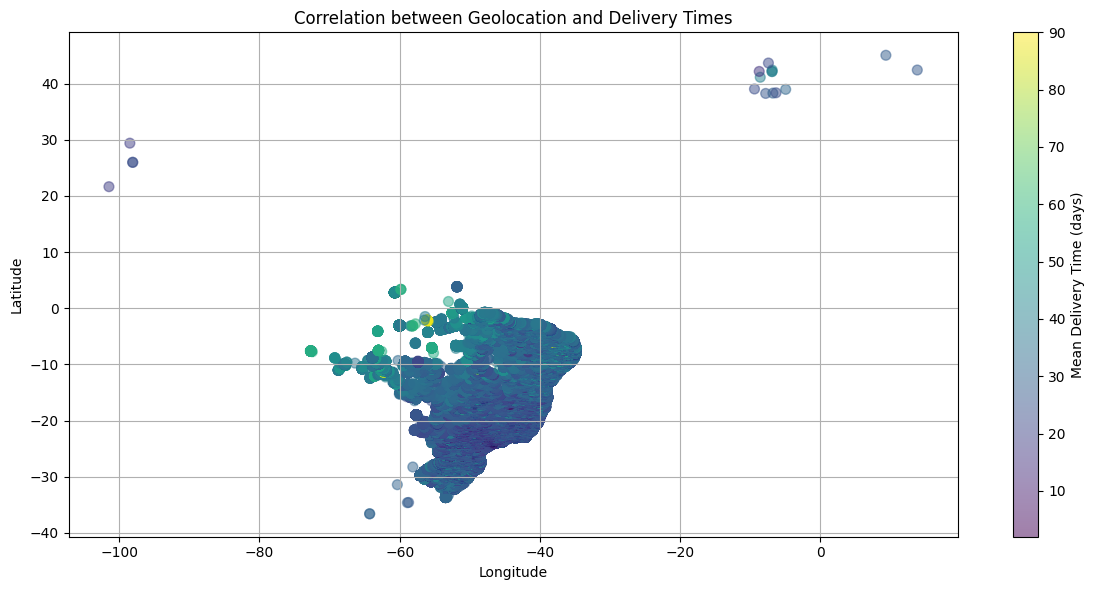
**2.2.1 Order Density by Geolocation**

Order density was examined by geographical locations to determine the regions with the highest order volume.



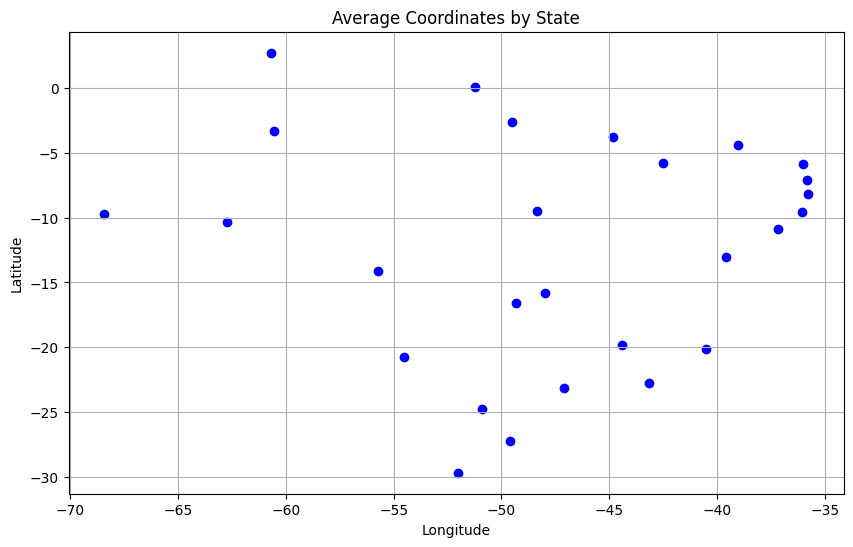
**2.2.2 Correlation between Geolocation and Delivery Times**

The correlation between geographical location and delivery times was analyzed, evaluating the impact of orders from distant regions on delivery times.



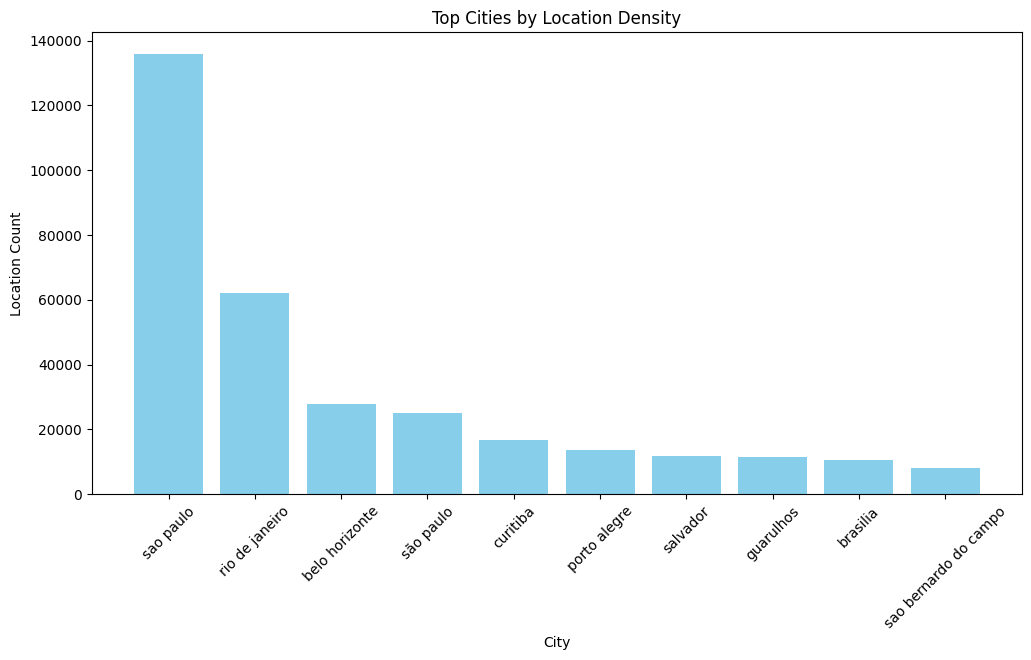
**2.2.3 Average Coordinates by State**

Average coordinates were calculated by state, determining the geographical centers of states.



**2.2.4 Determination of the Most Dense Cities**

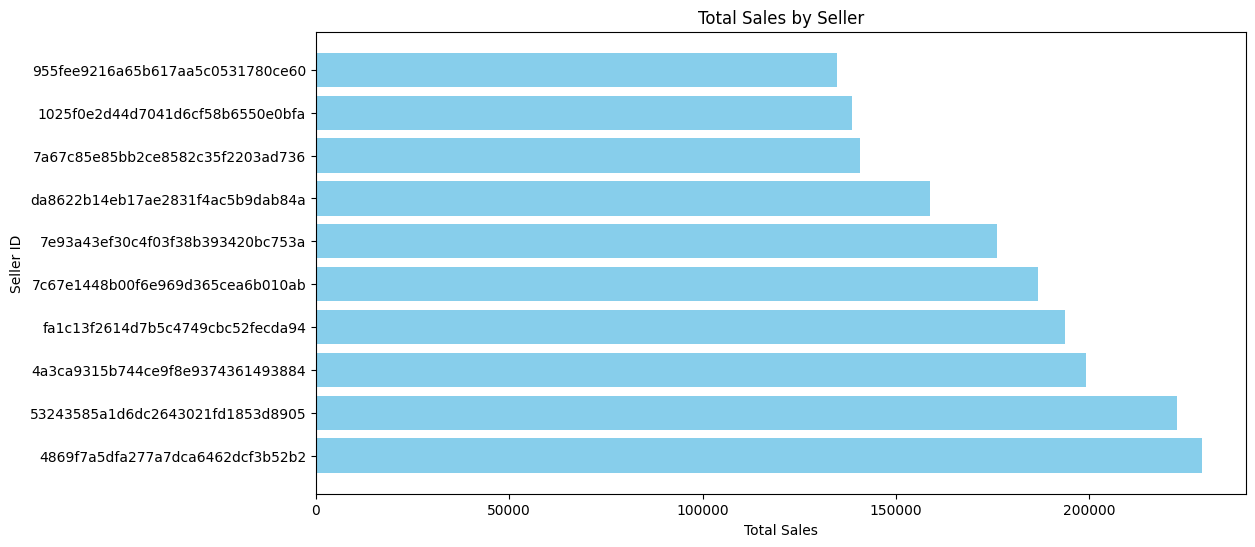
The most densely populated cities were identified, and the order density in these cities was examined.



**2.3 Order Items Table**

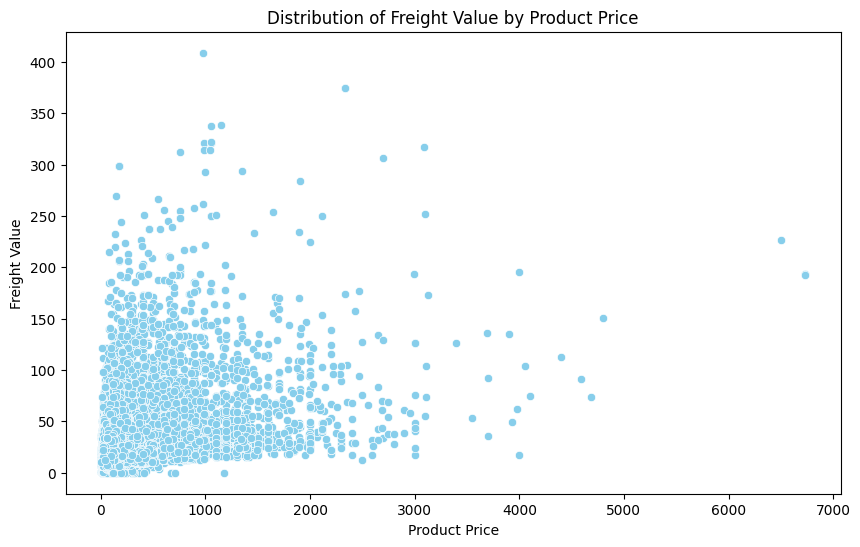
**2.3.1 Total Sales by Seller**

Total sales were analyzed by sellers, and the most successful sellers were identified.



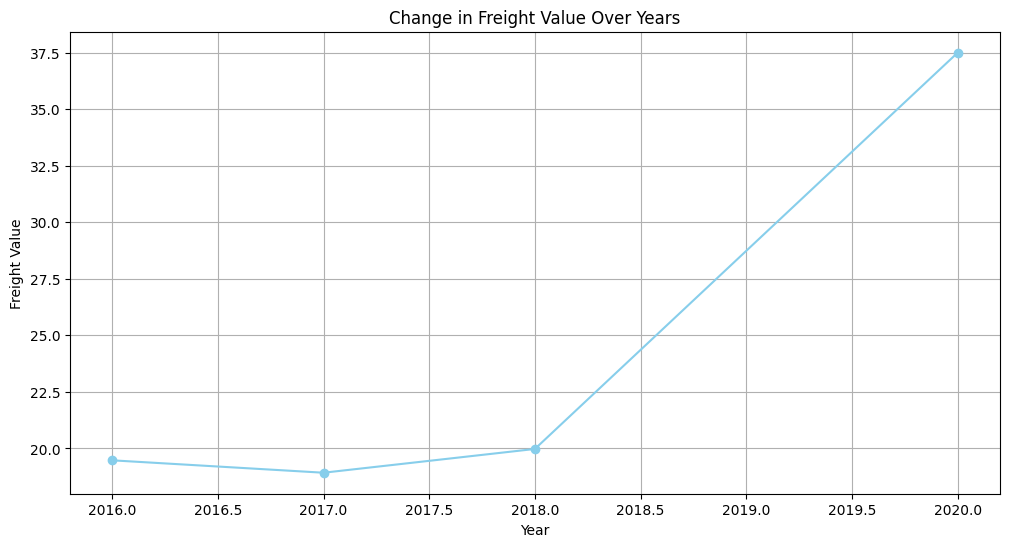
**2.3.2 Distribution of Freight Value by Product Price**

This analysis, the distribution of shipping costs by product prices was examined.



**2.3.3 Change in Freight Value Over Time**

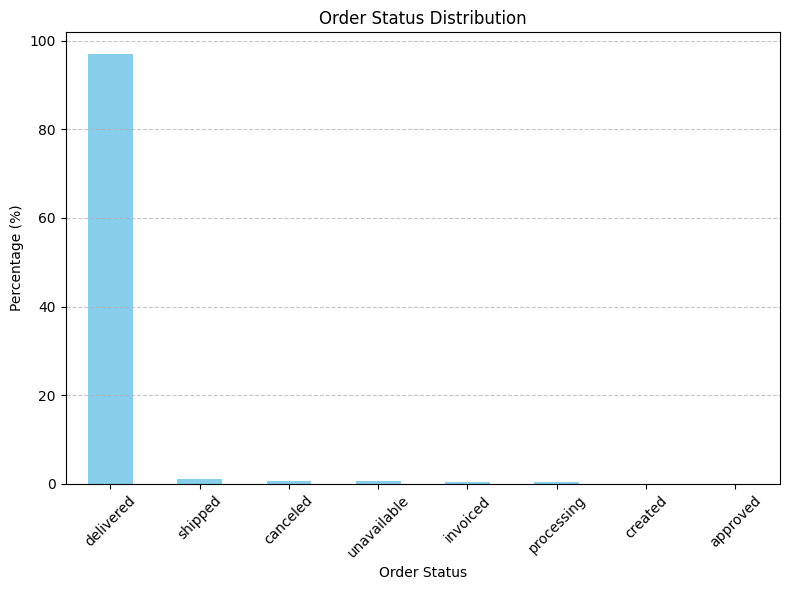
The change in freight value over time was analyzed, determining how shipping costs fluctuated over time.



**2.4 Order Payments Table Analyses**

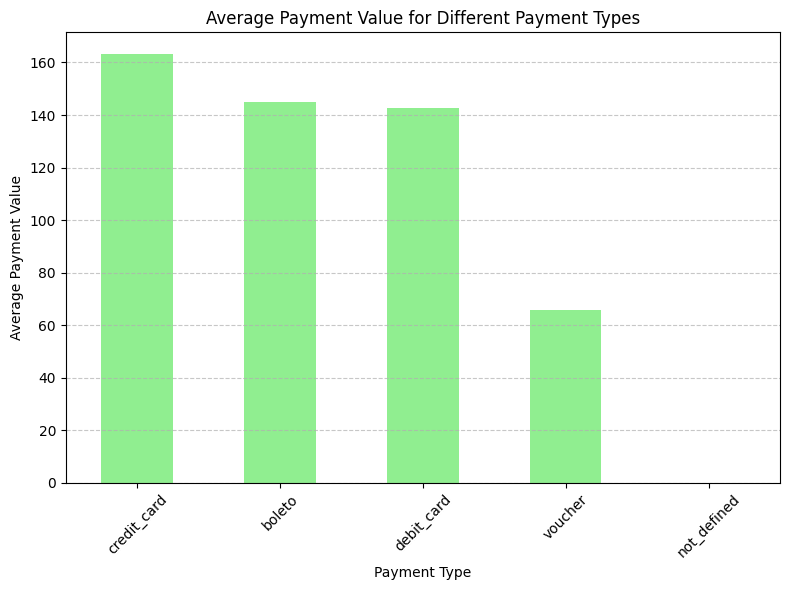
**2.4.1 Distribution of Payment Types**

The distribution of payment types was examined, identifying which payment methods are more common.



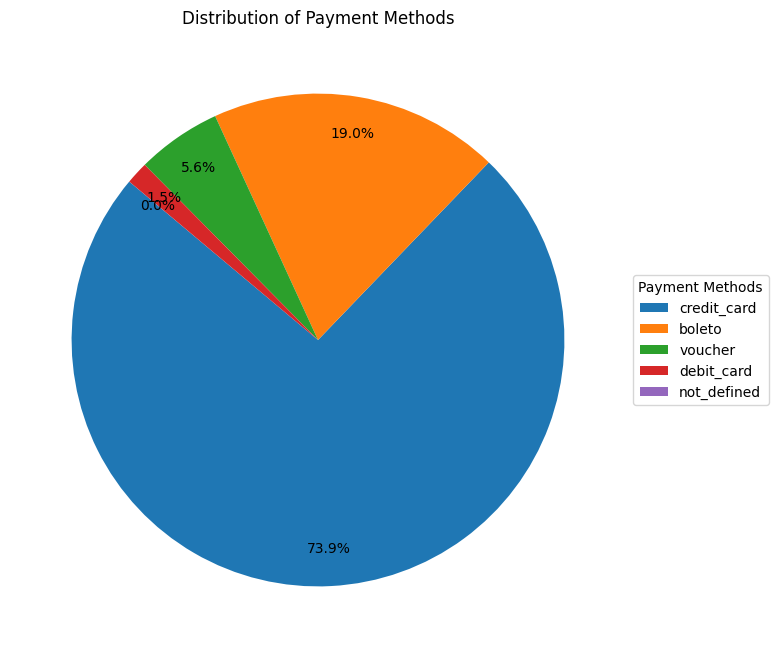
**2.4.2 Average Payment Value for Different Payment Types**

Average payment values were calculated for different payment types, determining which payment types are associated with higher values.



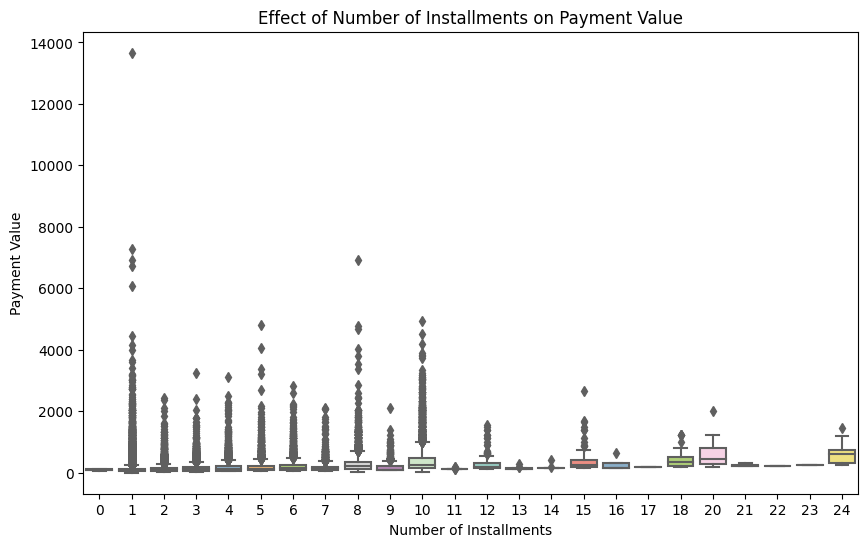
**2.4.3 Distribution of Payment Methods**

The distribution of payment methods was analyzed, identifying which payment methods customers prefer.



**2.4.4 The Effect of Number of Installments on Payment Value**

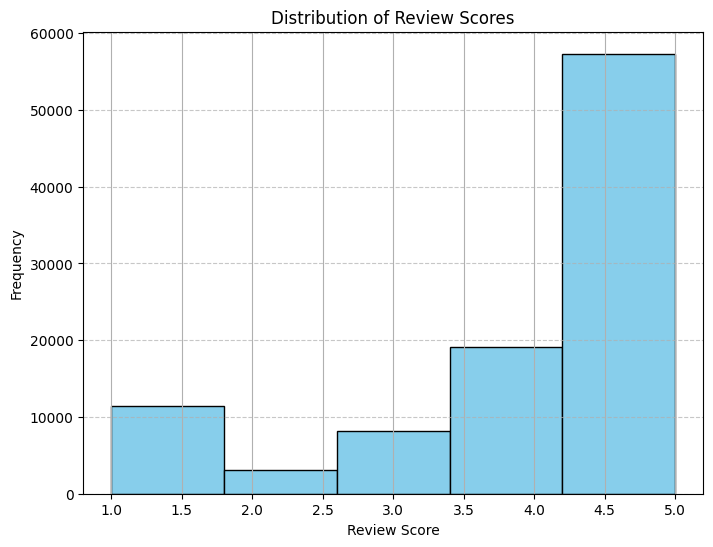
The effect of the number of installments on payment value was examined, evaluating how installment count affects the payment amount.



**2.5 Order Reviews Table Analyses**

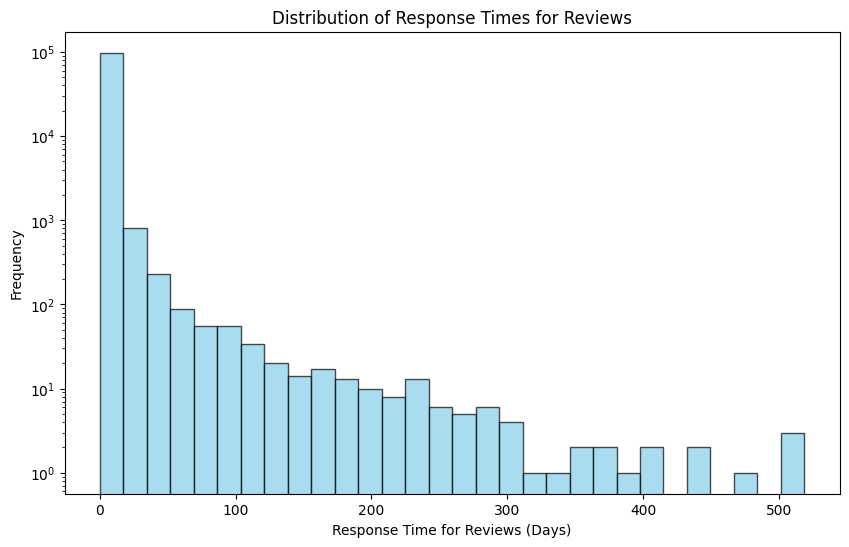
**2.5.1 Summary Statistics of Review Scores**

Summary statistics of review scores were calculated, examining the average scores.



**2.5.2 Distribution of Review Response Times**

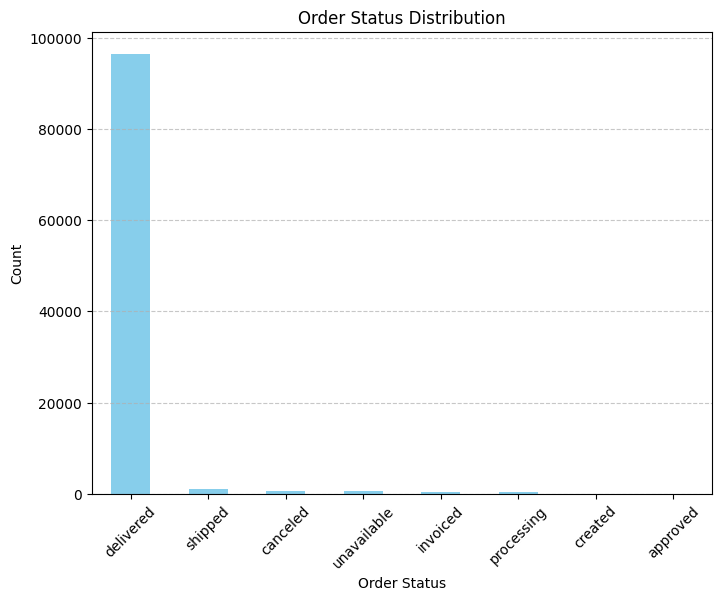
The distribution of review response times was analyzed, determining how quickly responses to customer feedback are provided.



**2.6 Orders Table Analyses**

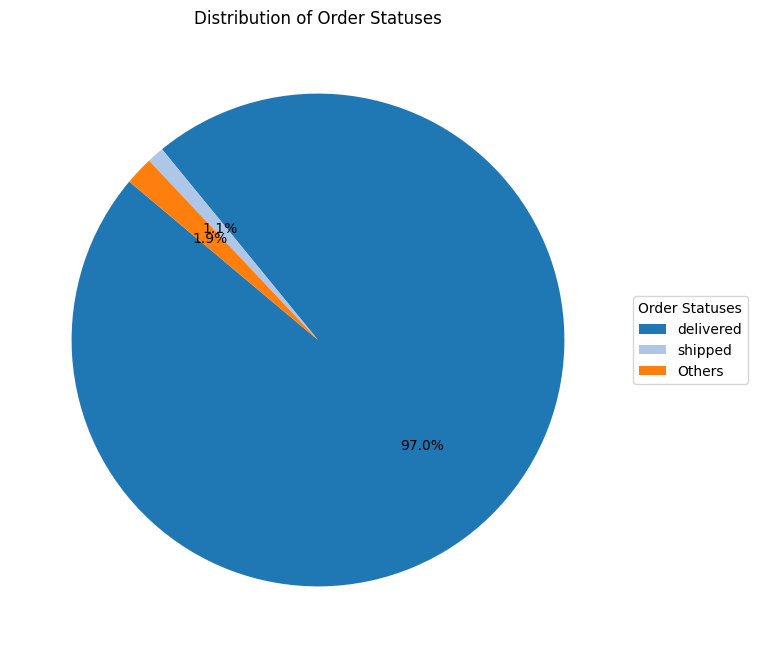
**2.6.1 Order Status Distribution**

The distribution of order statuses was examined, identifying which stages orders concentrate on.



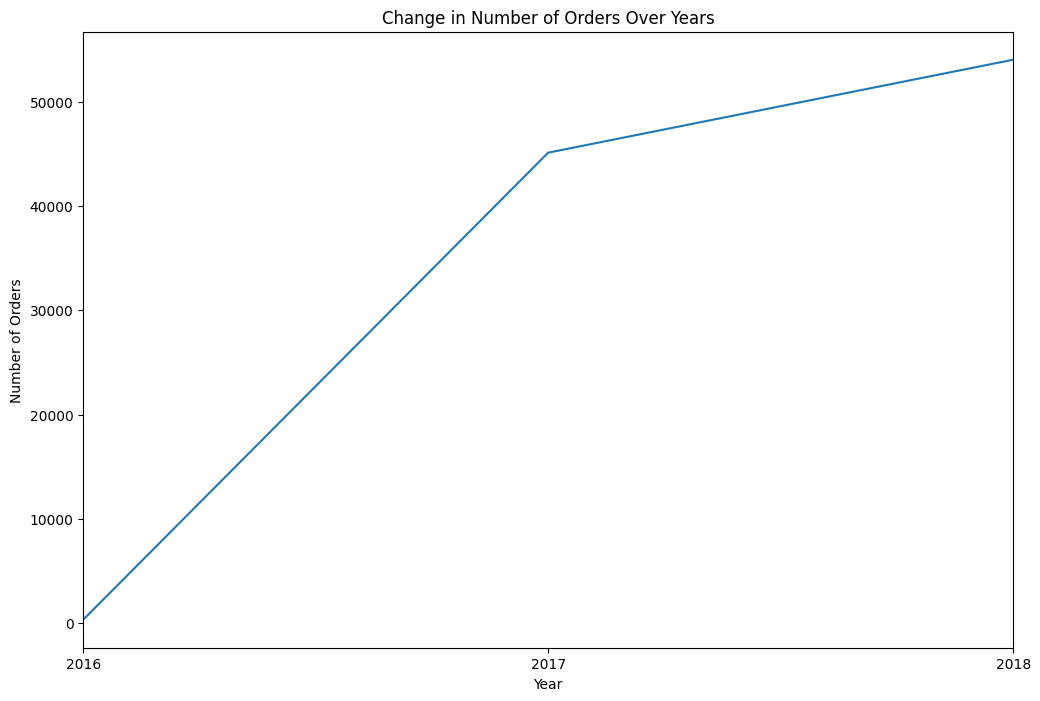
**2.6.2 Distribution of Order Statuses**

The distribution of order statuses was analyzed, identifying which stages orders are at and how many are completed.



**2.6.3 Change in Number of Orders Over Time**

The change in the number of orders over time was examined, determining the time periods when orders are concentrated.



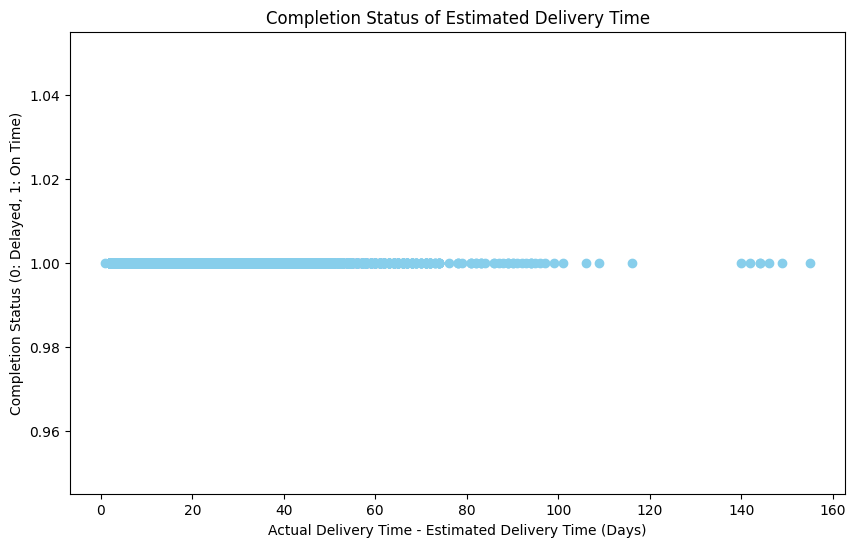
**2.6.4 Distribution of Delivery Times**

The distribution of delivery times was analyzed, determining the average delivery time.



**2.6.5 Completion Status of Estimated Delivery Time**

The completion status of estimated delivery time was examined, evaluating how accurate estimated times are compared to actual deliveries.



**2.7 Sellers Table Analyses**

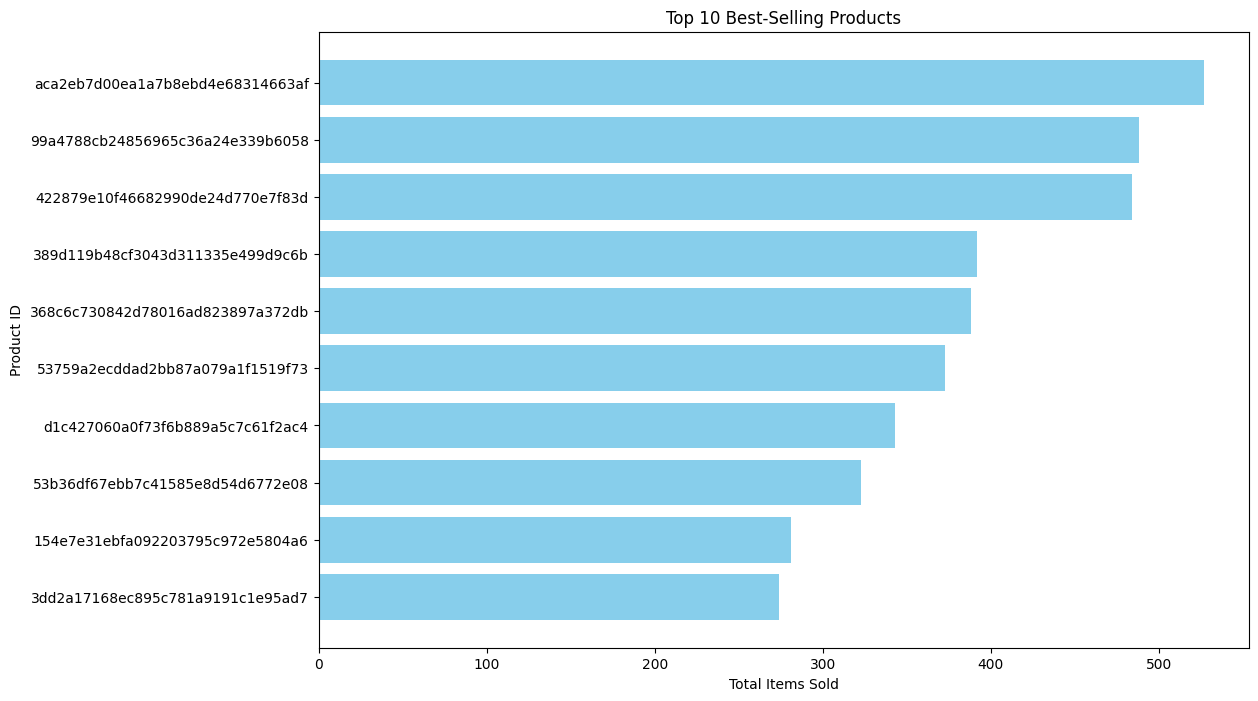
**2.7.1 Total Sales Value for Each Product Category**

Total sales value for each product category was examined, determining which categories have higher sales.

|  |  |
| --- | --- |
| **product\_category\_name** | **price** |
| beleza\_saude | 1252690 |
| relogios\_presentes | 1203007 |
| cama\_mesa\_banho | 1029335 |
| esporte\_lazer | 983274 |
| informatica\_acessorios | 907286 |
| moveis\_decoracao | 724953 |
| cool\_stuff | 632713 |
| utilidades\_domesticas | 628346 |
| automotivo | 590342 |
| ferramentas\_jardim | 482651 |
| brinquedos | 480964 |
| bebes | 409822 |
| perfumaria | 396708 |
| telefonia | 320125 |
| moveis\_escritorio | 272426 |
| papelaria | 229803 |
| pcs | 222917 |
| pet\_shop | 213102 |
| instrumentos\_musicais | 191159 |
| eletroportateis | 190360 |
| eletronicos | 158489 |
| consoles\_games | 156821 |
| fashion\_bolsas\_e\_acessorios | 151655 |
| construcao\_ferramentas\_construcao | 144214 |
| malas\_acessorios | 139608 |
| eletrodomesticos\_2 | 113214 |
| casa\_construcao | 82767 |
| eletrodomesticos | 79847 |
| agro\_industria\_e\_comercio | 72470 |
| moveis\_sala | 68516 |
| telefonia\_fixa | 59422 |
| casa\_conforto | 58315 |
| climatizacao | 54877 |
| audio | 50490 |
| portateis\_casa\_forno\_e\_cafe | 47430 |
| livros\_interesse\_geral | 46540 |
| moveis\_cozinha\_area\_de\_servico\_jantar\_e\_jardim | 46203 |
| construcao\_ferramentas\_iluminacao | 40981 |
| construcao\_ferramentas\_seguranca | 40453 |
| industria\_comercio\_e\_negocios | 39617 |
| alimentos | 29049 |
| market\_place | 28168 |
| construcao\_ferramentas\_jardim | 25571 |
| artes | 24059 |
| fashion\_calcados | 23343 |
| bebidas | 22266 |
| sinalizacao\_e\_seguranca | 21421 |
| moveis\_quarto | 19997 |
| livros\_tecnicos | 18960 |
| construcao\_ferramentas\_ferramentas | 15852 |
| alimentos\_bebidas | 15034 |
| fashion\_roupa\_masculina | 10699 |
| fashion\_underwear\_e\_moda\_praia | 9458 |
| artigos\_de\_natal | 8699 |
| tablets\_impressao\_imagem | 7499 |
| cine\_foto | 6890 |
| musica | 6014 |
| dvds\_blu\_ray | 5948 |
| livros\_importados | 4613 |
| artigos\_de\_festas | 4460 |
| moveis\_colchao\_e\_estofado | 4360 |
| portateis\_cozinha\_e\_preparadores\_de\_alimentos | 3961 |
| fashion\_roupa\_feminina | 2766 |
| fashion\_esporte | 2093 |
| la\_cuisine | 2054 |
| artes\_e\_artesanato | 1799 |
| fraldas\_higiene | 1565 |
| pc\_gamer | 1541 |
| flores | 1088 |
| casa\_conforto\_2 | 733 |
| cds\_dvds\_musicais | 730 |
| fashion\_roupa\_infanto\_juvenil | 564 |
| seguros\_e\_servicos | 283 |

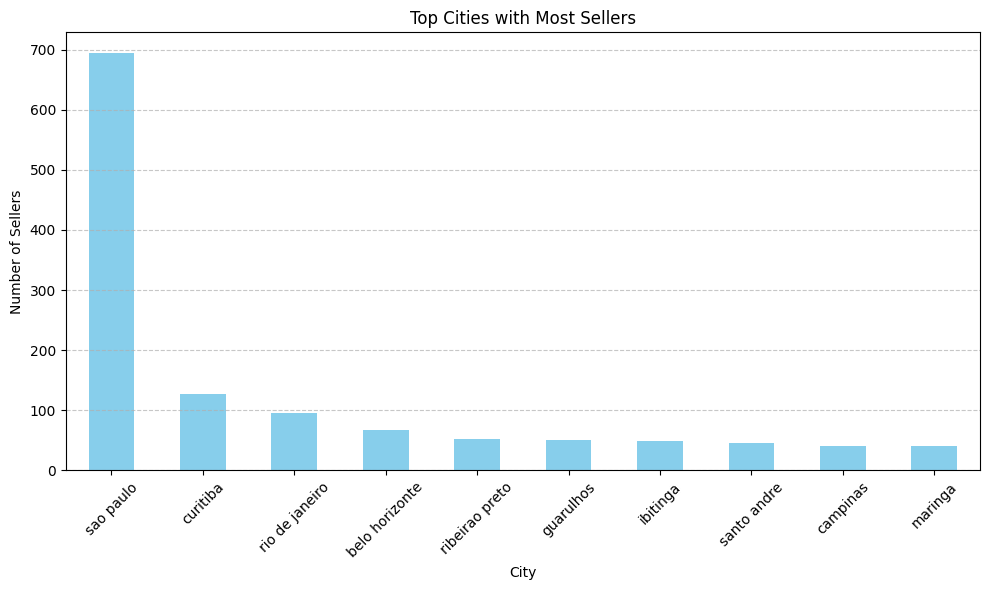
**2.7.2 Top 10 Best-Selling Products**

The top 10 best-selling products were identified, and their performance was analyzed.



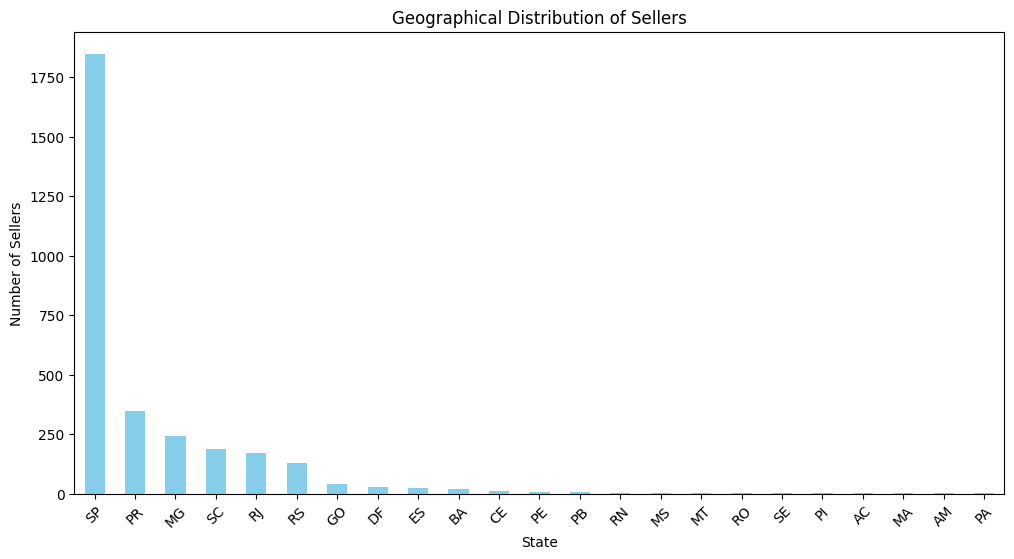
**2.7.3 Top Cities with Most Sellers**

The cities with the most sellers were identified, and the geographical distribution of sellers was examined.



**2.7.4 Geographical Distribution of Sellers**

The geographical distribution of sellers was analyzed, and the number of sellers in different regions was examined.



**2.8 Other Analyses**

**2.8.1 Average Order Value**

Average order value was calculated, and customers' average spending was examined.

Average Order Value: 139.56

**2.8.2 Total Sales**

Total sales quantity was calculated, and the overall performance of the business was evaluated.

Total Sales: 15721763

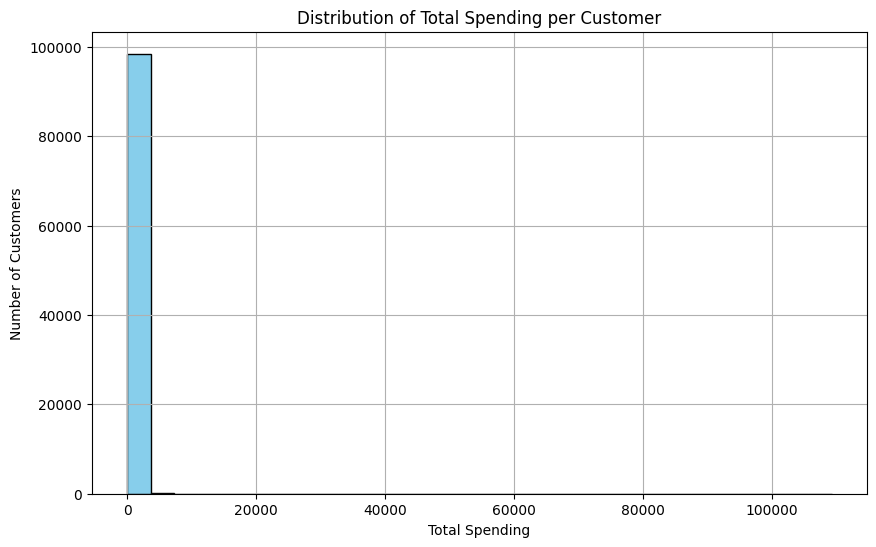
**2.8.3 Number of Orders per Customer Segment**

The number of orders per customer segment was examined, determining the purchasing habits of different customer groups.

|  |  |
| --- | --- |
| **Customer State** | **Order Count** |
| AC | 81 |
| AL | 413 |
| AM | 148 |
| AP | 68 |
| BA | 3380 |
| CE | 1336 |
| DF | 2140 |
| ES | 2033 |
| GO | 2020 |
| MA | 747 |
| MG | 11635 |
| MS | 715 |
| MT | 907 |
| PA | 975 |
| PB | 536 |
| PE | 1652 |
| PI | 495 |
| PR | 5045 |
| RJ | 12852 |
| RN | 485 |
| RO | 253 |
| RR | 46 |
| RS | 5466 |
| SC | 3637 |
| SE | 350 |
| SP | 41746 |
| TO | 280 |

**2.8.4 Distribution of Total Spending per Customer**

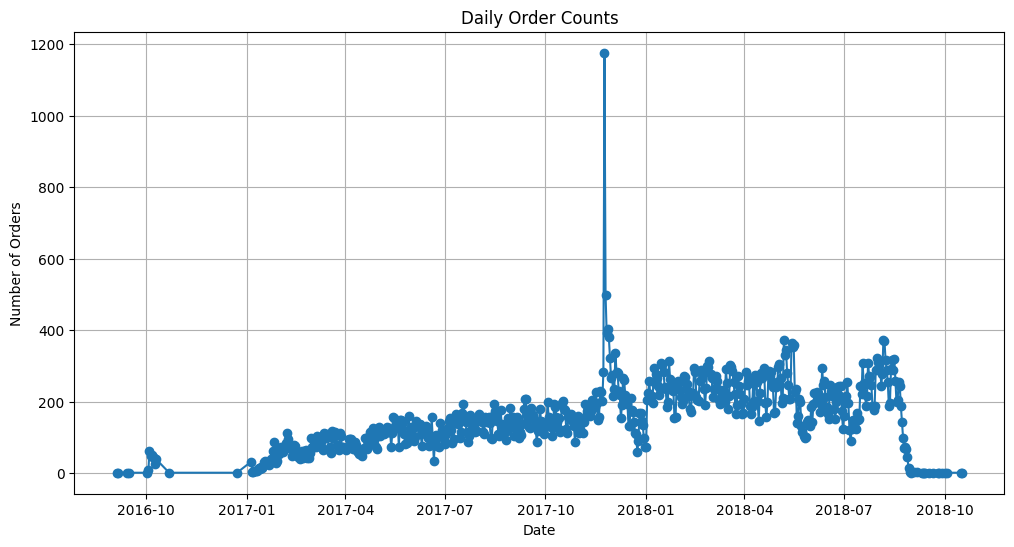
The distribution of total spending per customer was examined, determining which customers spend more.



#### 2.9 Additional Analyses

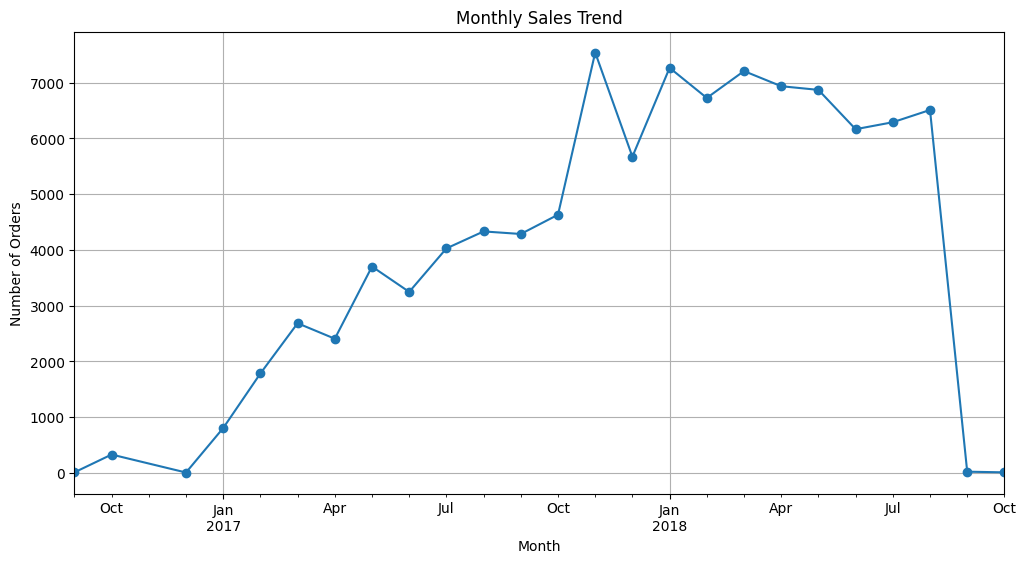
##### 2.9.1 Daily Order Count

The daily order count was examined, determining the daily order volume of the business.



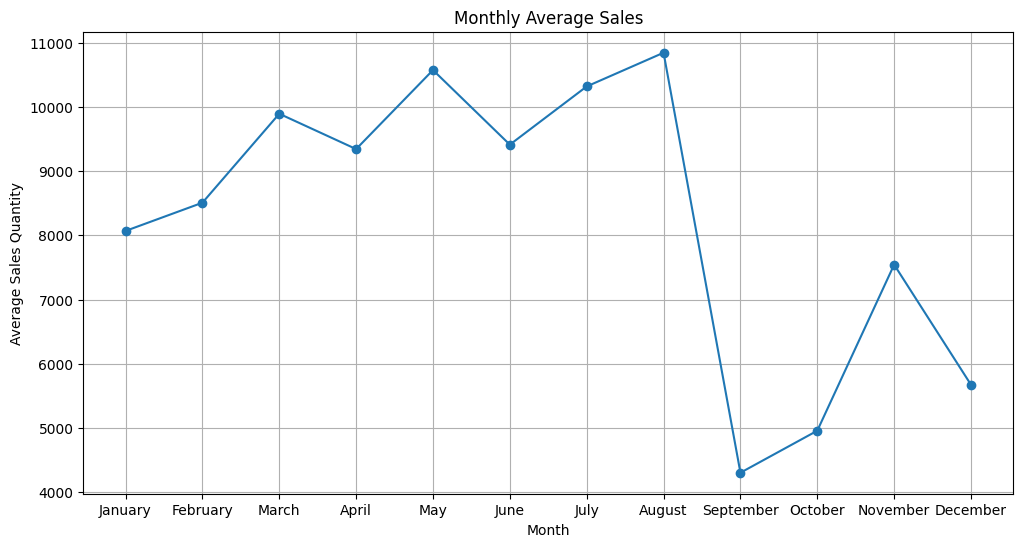
##### 2.9.2 Sales Trends Over Time

Sales trends over time were analyzed, examining changes in the business's performance over different time periods.



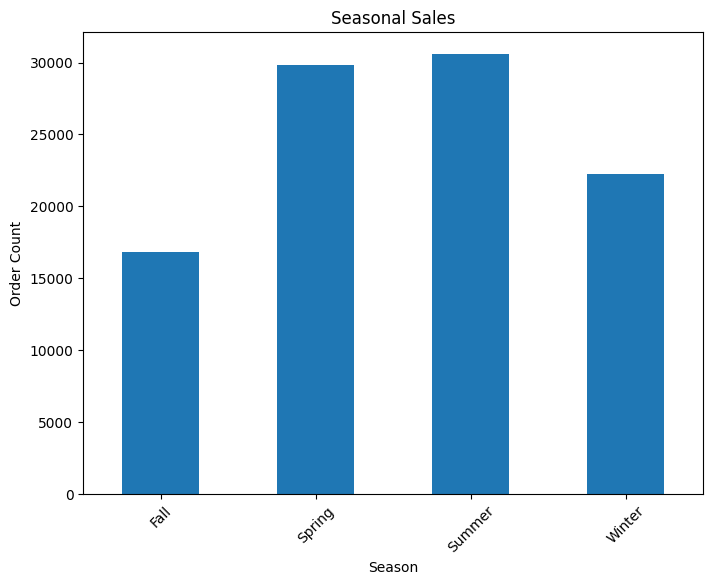
##### 2.9.3 Average Monthly Sales

Average monthly sales were examined, determining which months have higher sales volumes.



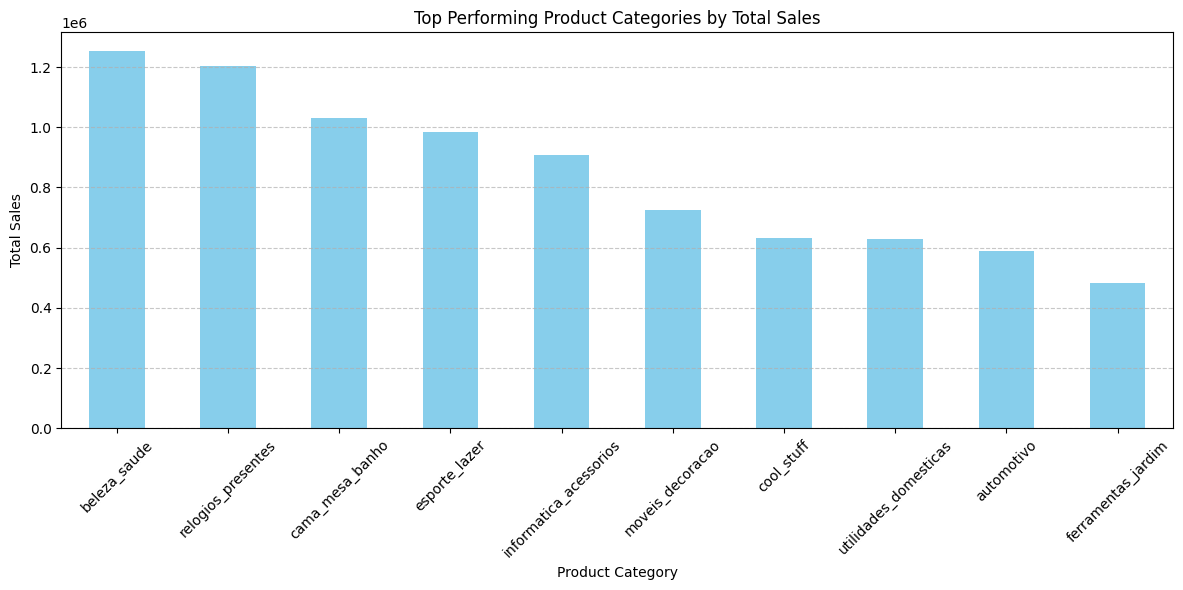
##### 2.9.4 Seasonal Sales

Seasonal sales were analyzed, determining which seasons see higher sales for the business.



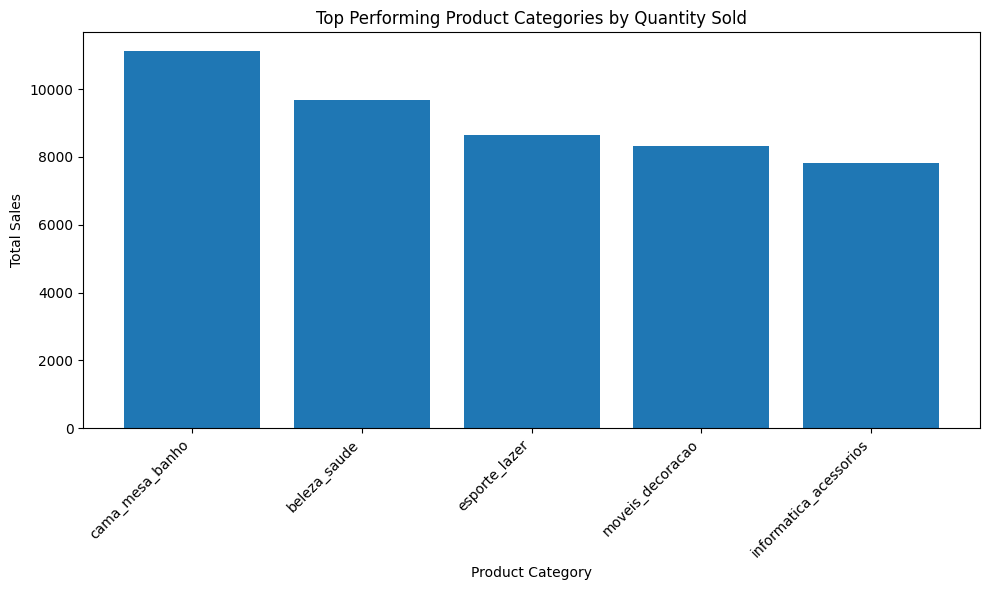
##### 2.9.5 Top Performing Product Categories by Total Sales

The top-performing product categories were identified based on total sales, and their sales were examined.



##### 2.9.6 Top Performing Product Categories by Quantity Sold

The top-performing product categories were identified based on the quantity sold, and their sales were examined.



##### 2.9.7 Contribution of Top 20 Customer Segments to Total Revenue

The contribution of customer segments to total revenue was analyzed, determining which customer segments contribute more to the business's revenue. Specifically, a focus was placed on the top 20 customers in terms of their contribution.

|  |  |  |
| --- | --- | --- |
| **customer\_id** | **price** | **contribution\_percentage** |
| 1617b1357756262bfa56ab541c47bc16 | 13440 | 0,099379 |
| ec5b2ba62e574342386871631fafd3fc | 7160 | 0,052943 |
| c6e2731c5b391845f6800c97401a43a9 | 6735 | 0,049801 |
| f48d464a0baaea338cb25f816991ab1f | 6729 | 0,049756 |
| 3fd6777bbce08a352fddd04e4a7cc8f6 | 6499 | 0,048056 |
| 05455dfa7cd02f13d132aa7a6a9729c6 | 5934 | 0,043878 |
| df55c14d1476a9a3467f131269c2477f | 4799 | 0,035485 |
| 24bbf5fd2f2e1b359ee7de94defc4a15 | 4690 | 0,034679 |
| e0a2412720e9ea4f26c1ac985f6a7358 | 4598 | 0,033999 |
| 3d979689f636322c62418b6346b1c6d2 | 4590 | 0,03394 |
| cc803a2c412833101651d3f90ca7de24 | 4400 | 0,032535 |
| 1afc82cd60e303ef09b4ef9837c9505c | 4399 | 0,032528 |
| 35a413c7ca3c69756cb75867d6311c0d | 4099 | 0,030309 |
| e9b0d0eb3015ef1c9ce6cf5b9dcbee9f | 4059 | 0,030013 |
| 926b6a6fb8b6081e00b335edaf578d35 | 3999 | 0,02957 |
| c6695e3b1e48680db36b487419fb0398 | 3999 | 0,02957 |
| 3be2c536886b2ea4668eced3a80dd0bb | 3980 | 0,029429 |
| 31e83c01fce824d0ff786fcd48dad009 | 3930 | 0,02906 |
| eb7a157e8da9c488cd4ddc48711f1097 | 3899 | 0,02883 |
| 19b32919fa1198aefc0773ee2e46e693 | 3700 | 0,027359 |

**Conclusion**

This analysis of the e-commerce dataset provides valuable insights into the sales and customer behavior of an online retailer in Brazil. Several key findings emerge:

**Sales Performance:**

Consistent Sales Growth: The number of orders has steadily increased over time, suggesting a strong and expanding customer base.

Seasonal Fluctuations: Monthly sales exhibit clear seasonal patterns, peaking during the summer months and experiencing a dip in the winter. This indicates potential opportunities for targeted marketing campaigns during peak seasons.

High Order Volume: The daily order counts suggest a high volume of transactions, indicating a thriving business with a significant customer base.

Top-Performing Product Categories: Certain product categories like 'cama\_mesa\_banho', 'beleza\_saude', 'esporte\_lazer', and 'moveis\_decoracao' dominate both quantity sold and total sales. This highlights popular customer preferences and opportunities for expanding inventory within these categories.

Distribution of Order Statuses: While the majority of orders are successfully delivered, a small percentage remain in other statuses like 'shipped' or 'others'. This suggests an efficient fulfillment process, although potential areas for optimization could be explored to minimize delays.

**Customer Analysis:**

Concentrated Customer Base: The majority of customers are concentrated in the state of São Paulo, highlighting the importance of targeted marketing efforts within this region.

High Average Spending: The distribution of total spending per customer reveals a significant number of customers who make substantial purchases, suggesting potential for customer loyalty programs and personalized offers.

Review Score Distribution: The majority of customer reviews are positive, demonstrating overall customer satisfaction.

Payment Preferences: Credit card emerges as the most popular payment method, indicating the importance of offering secure and convenient payment options.

**Seller Analysis:**

Wide Range of Seller Activity: The distribution of total sales by seller shows a significant number of sellers contributing to the platform's overall success.

Geographic Concentration of Sellers: A majority of sellers operate in São Paulo, highlighting the strong e-commerce ecosystem in this region.

Top Cities with Most Sellers: São Paulo, Curitiba, and Rio de Janeiro have the highest concentration of sellers, suggesting an active and competitive e-commerce landscape in these urban centers.

**Logistics and Delivery:**

Reliable Delivery Time: The distribution of delivery times suggests that a significant portion of orders are delivered within a reasonable timeframe.

Consistent Freight Values: The change in freight values over time reveals a gradual increase, possibly driven by factors such as fuel prices or changing delivery patterns.

Delivery Time Accuracy: The scatter plot of delivery times versus estimated delivery times highlights a high rate of on-time deliveries.

Geolocation Impact on Delivery: The scatter plot shows a correlation between geolocation and delivery times, implying that location can influence delivery efficiency. The heatmap further illustrates the higher density of orders in specific regions, suggesting potential logistical challenges or opportunities.

Overall, the analysis reveals a thriving e-commerce platform with a robust customer base, strong sales performance, and efficient logistics. While the data suggests significant success, potential areas for improvement include:

Expanding into New Markets: While São Paulo is a key market, expanding into other regions could unlock new growth opportunities.

Optimizing Delivery Efficiency: Analyzing potential logistical bottlenecks and exploring innovative solutions could further improve delivery times and customer satisfaction.

Leveraging Customer Data for Personalized Marketing: Utilizing customer data to personalize marketing efforts could enhance customer engagement and drive sales.

Further investigation and analysis of the data can provide more granular insights into specific areas of improvement and inform strategic decision-making for the e-commerce platform.