Project 1
Exploratory Data Analysis

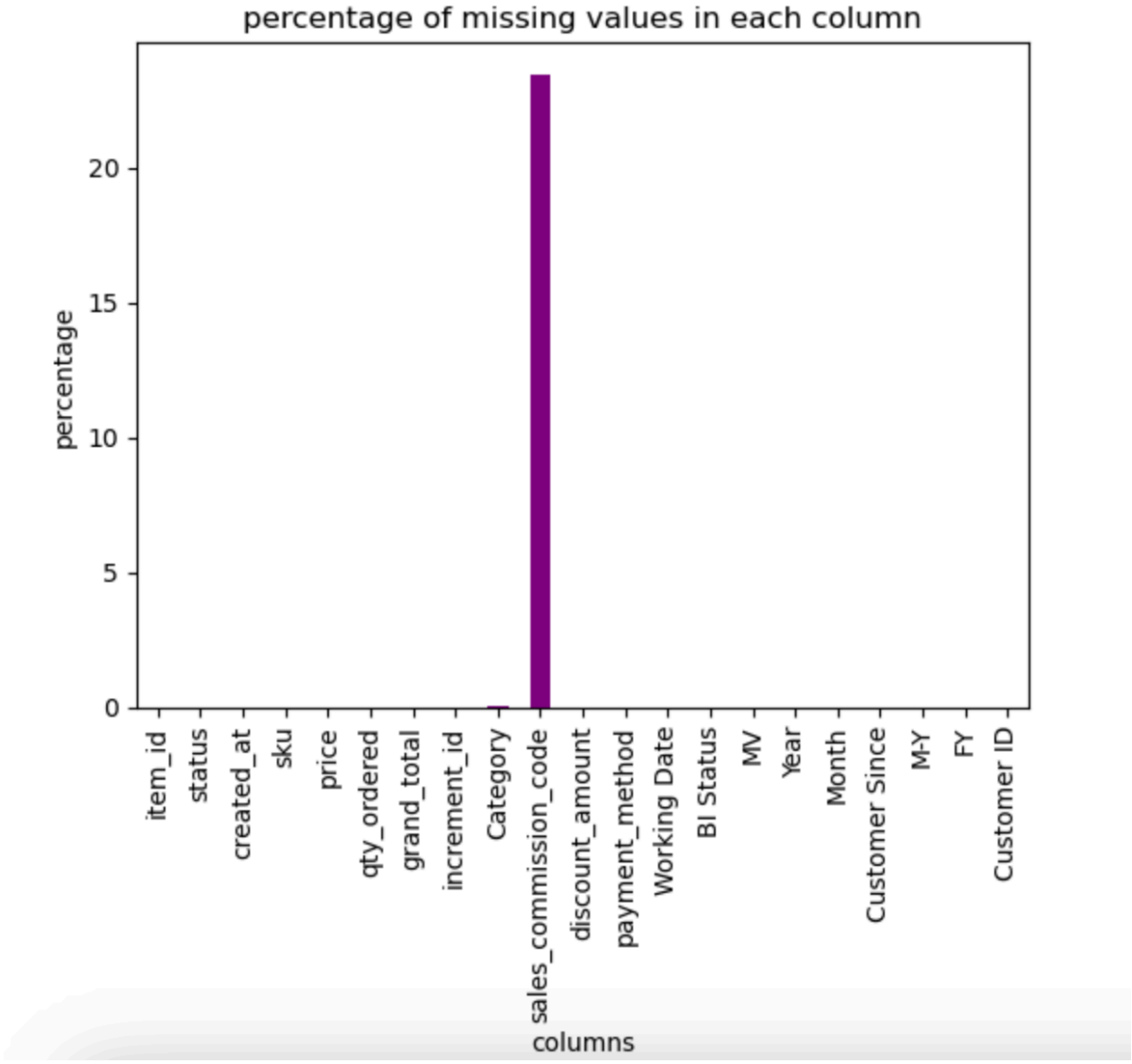
Ecommerce Dataset

Steffi Dorothy
September 18, 2024

Task 2
Statistical Summaries

Visualizing Missing Data

Handling missing values is critical in preparing an e-commerce dataset for analysis. Properly addressing missing data ensures the accuracy and reliability of insights, which is essential for making informed business decisions.



sales_commission_code column has a lot of missing values. I dropped this column from the data as this is not required for analysis

Check for duplicates

Ensuring the dataset is free from duplicate records is an essential step in data preprocessing, as duplicates can skew analysis and lead to inaccurate insights. No duplicates were found.

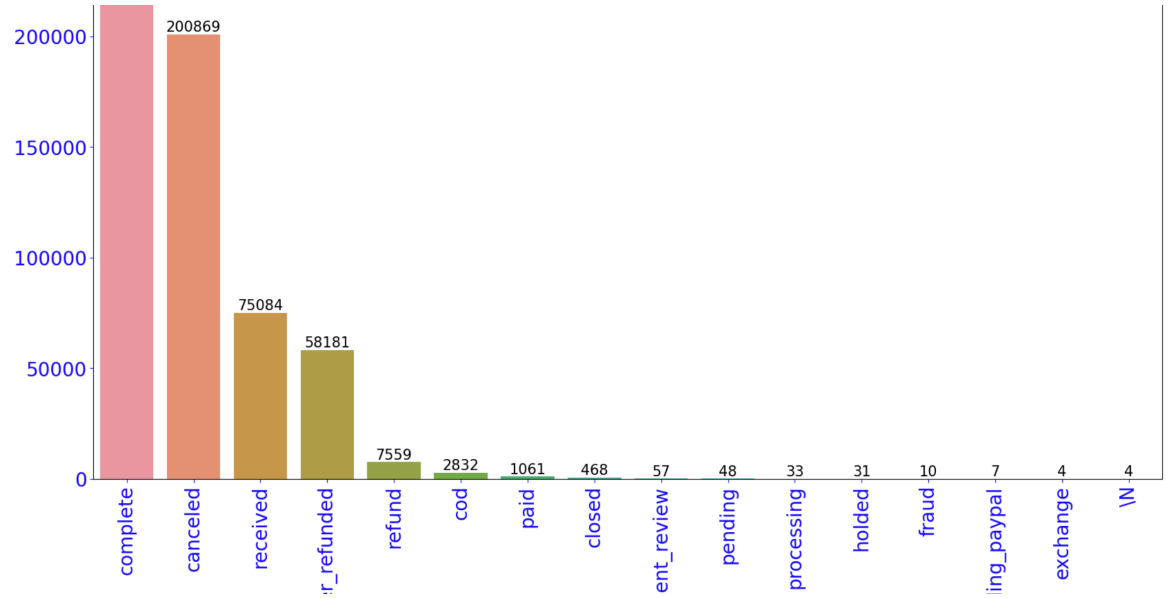
Additional Data Quality Checks

Validated Data Types, Ensured Consistency in Categorical Variables, and Analyzed Date and Time Features.

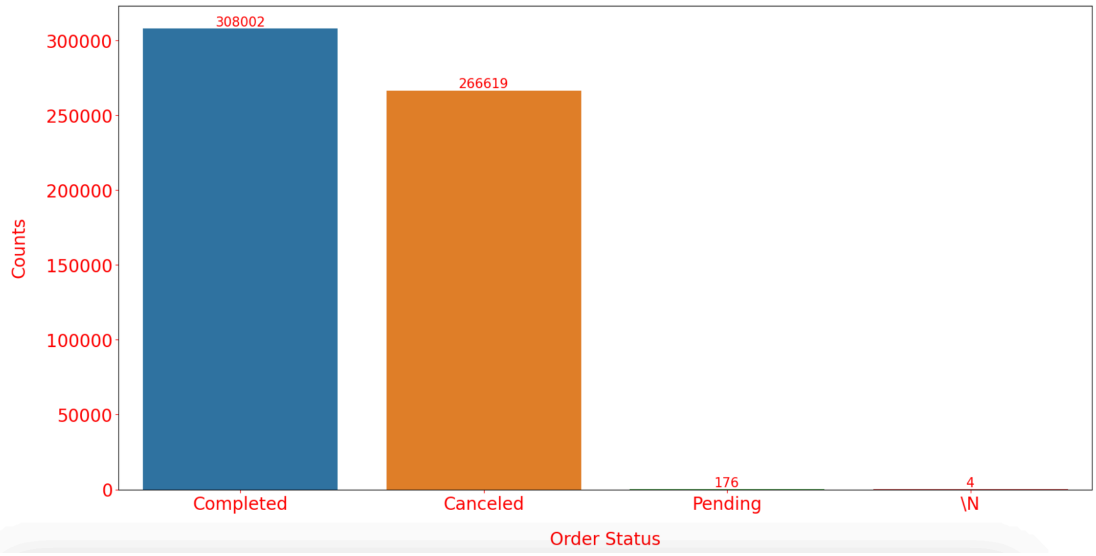
Task 3

Data Analysis

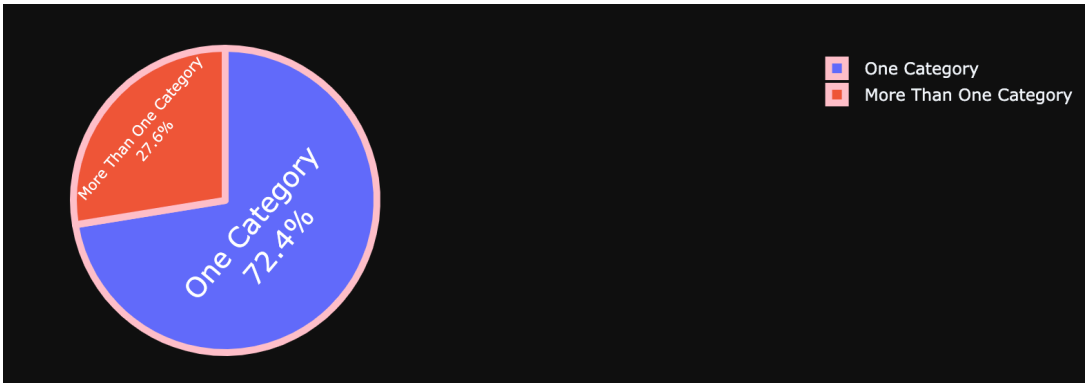
Order Status



Some of these words share similar meanings, so I categorized them together.



Orders Per Category

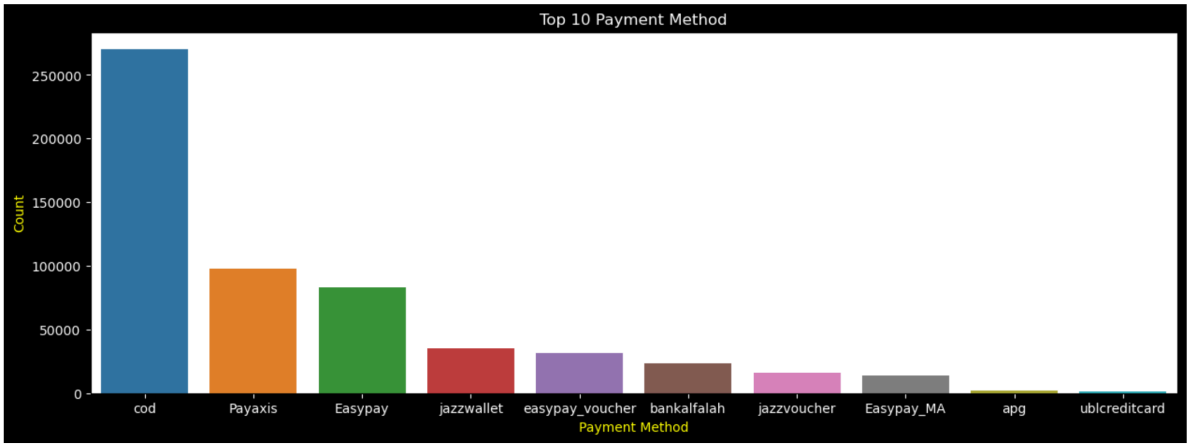


Approximately 72.4% of customers made purchases exclusively from one category, while 27.6% bought products from multiple categories.

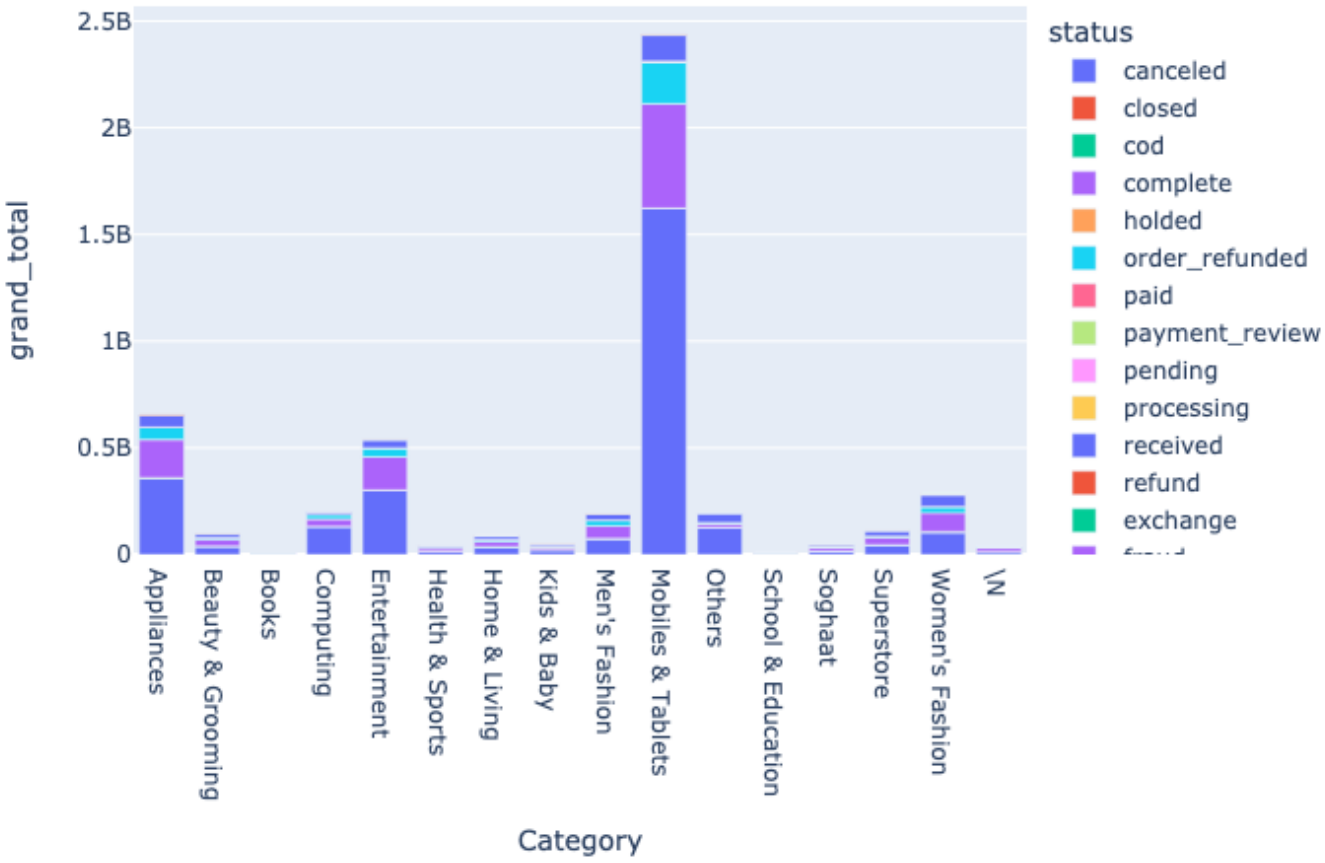
Orders Per Customer

56% of customers have placed more than one order while 44% of customers have placed one order.

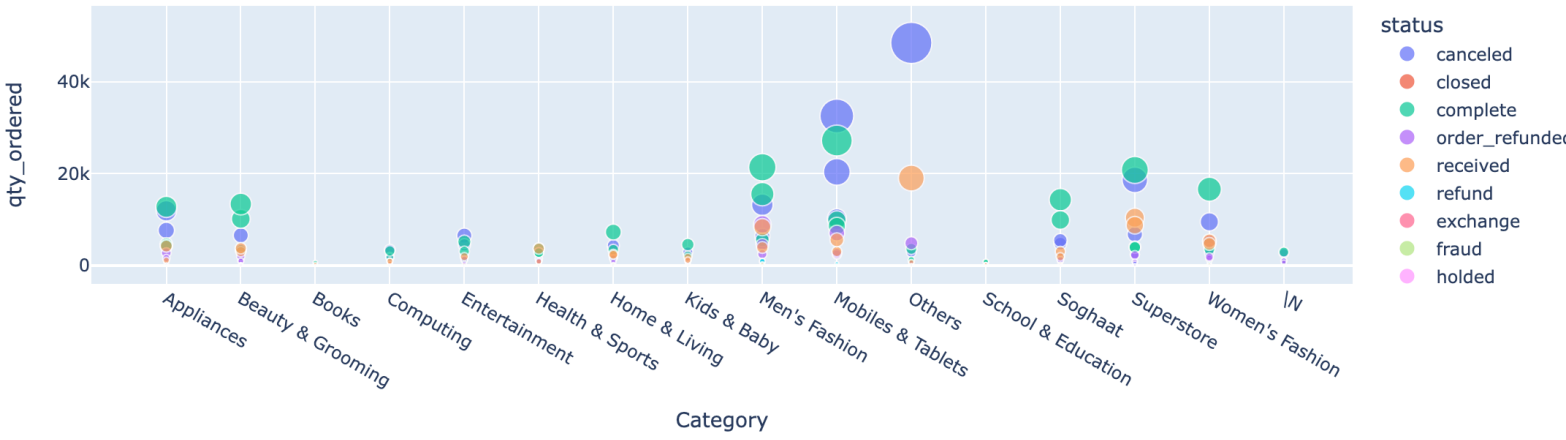
Examination of the Top Ten Payment Methods



Analysis of the total number of items in a specific category in a long format



Valuable insights through a comprehensive analysis of Quantity Ordered in a Long Format utilizing a Scatter Graph

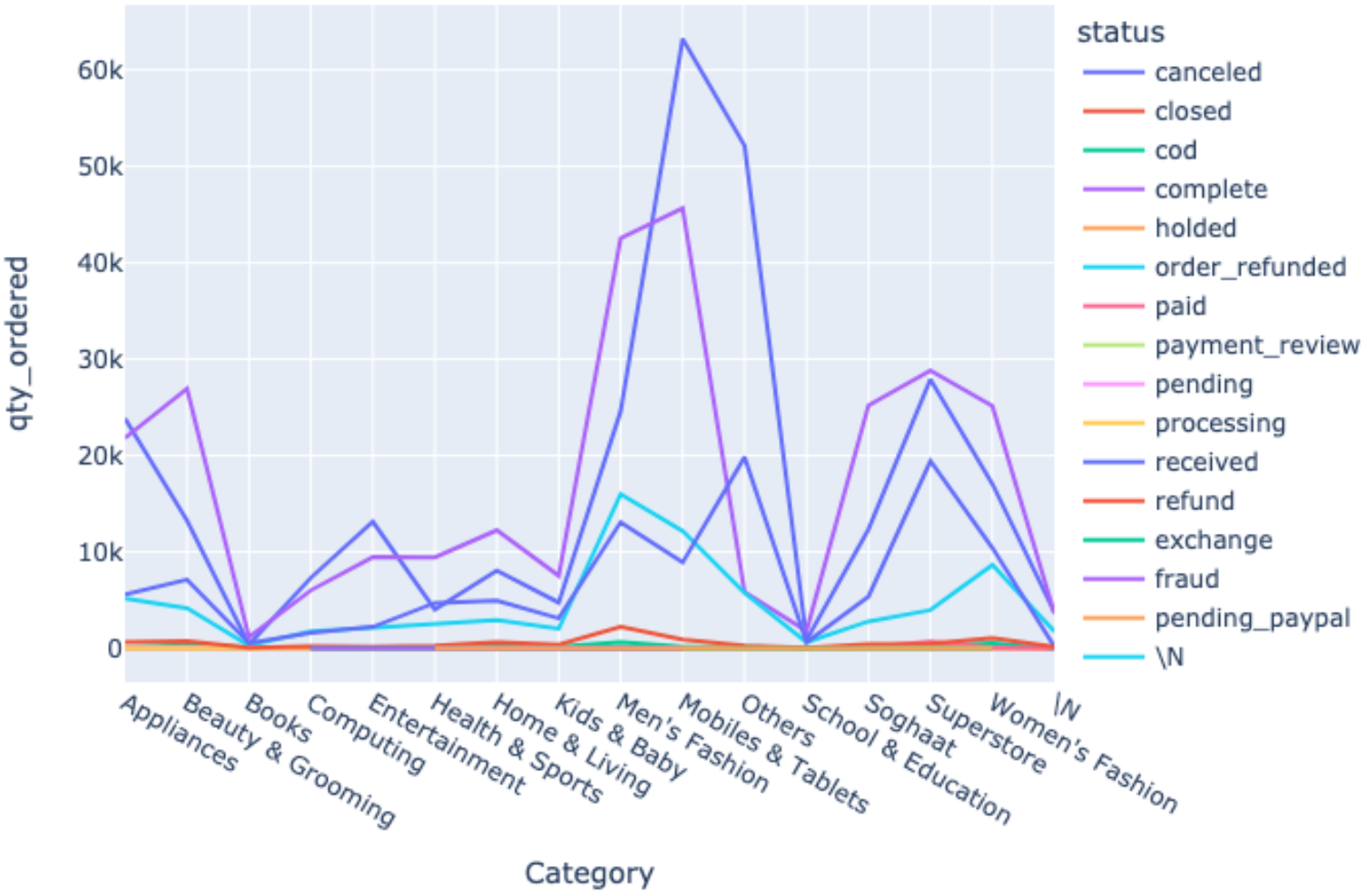


Analyzing the Number of Orders for Items Categorized by Category

The graph shows the **quantity ordered** across various **product categories** by **order status**. A few key observations:

- **Mobiles & Tablets** dominate the number of orders, especially in statuses like **complete** and **canceled**.
- Categories such as **Appliances** and **Men's Fashion** also show significant order volumes, with varying statuses.
- There is a clear discrepancy in the number of orders across different categories, indicating that popular categories like **Mobiles & Tablets** and **Men's Fashion** may skew the dataset, introducing **category bias**.

This suggests a potential over-representation of certain products.

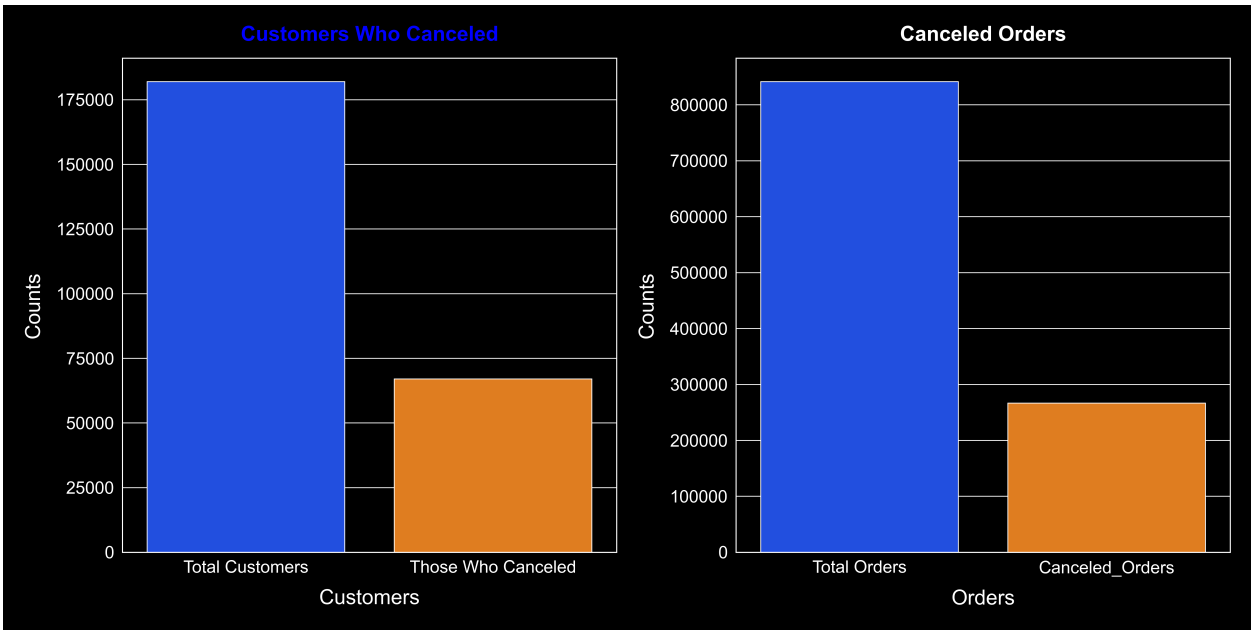


Task 4

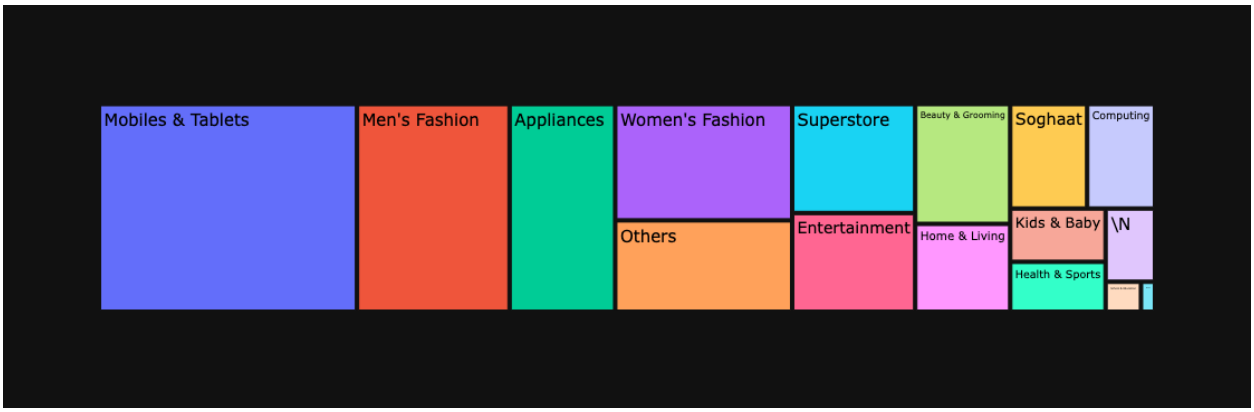
Questions

What are the cancellation and return rates for different categories? Do certain categories experience higher return rates?

Canceled Orders



Famous Category: Canceled Orders

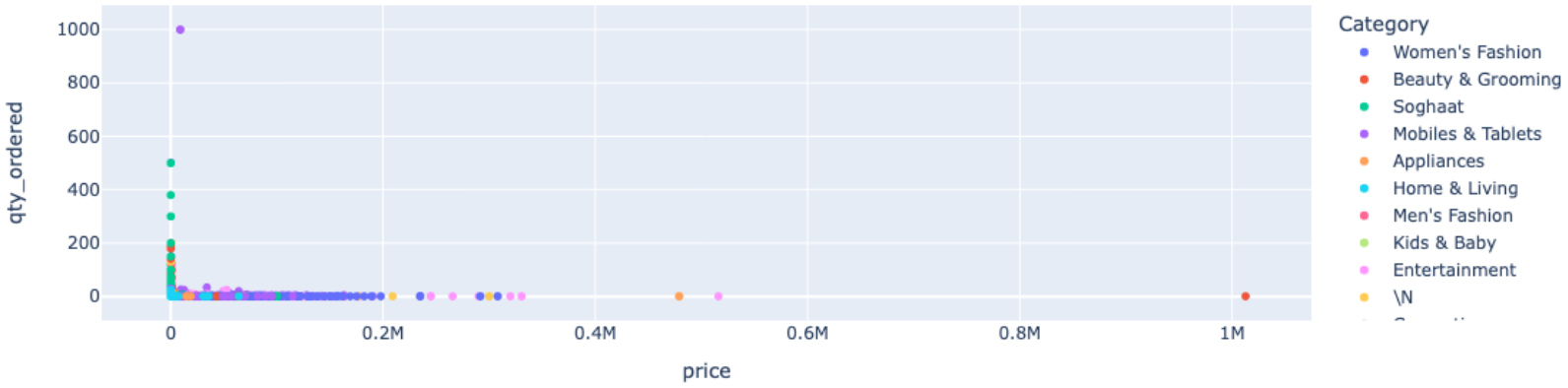


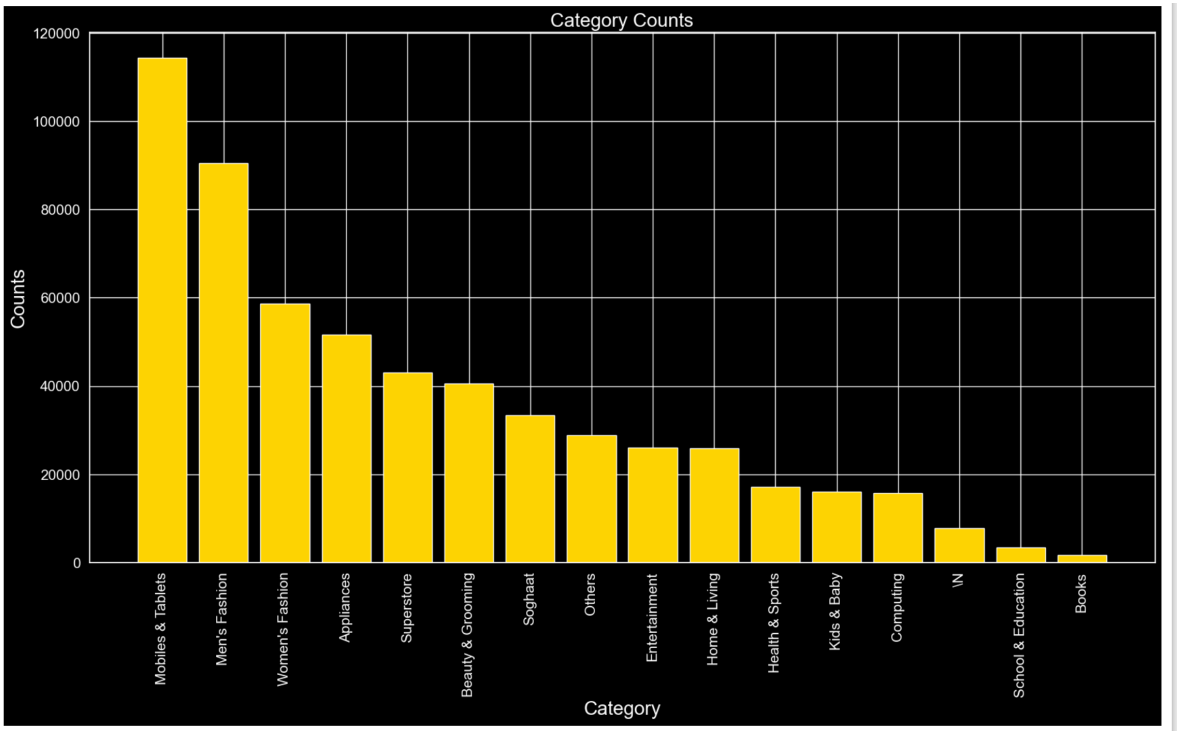
We have a total of 266619 canceled orders. Some categories have higher cancellation rates than others, which suggests that certain products are more likely to be returned or result in buyer's remorse. Identifying these categories could help us manage inventory levels more effectively. The majority of the orders were canceled for Mobile & tablets, Men's Fashion, and Appliances.

Bar Plot - Which Category has the highest number of orders?

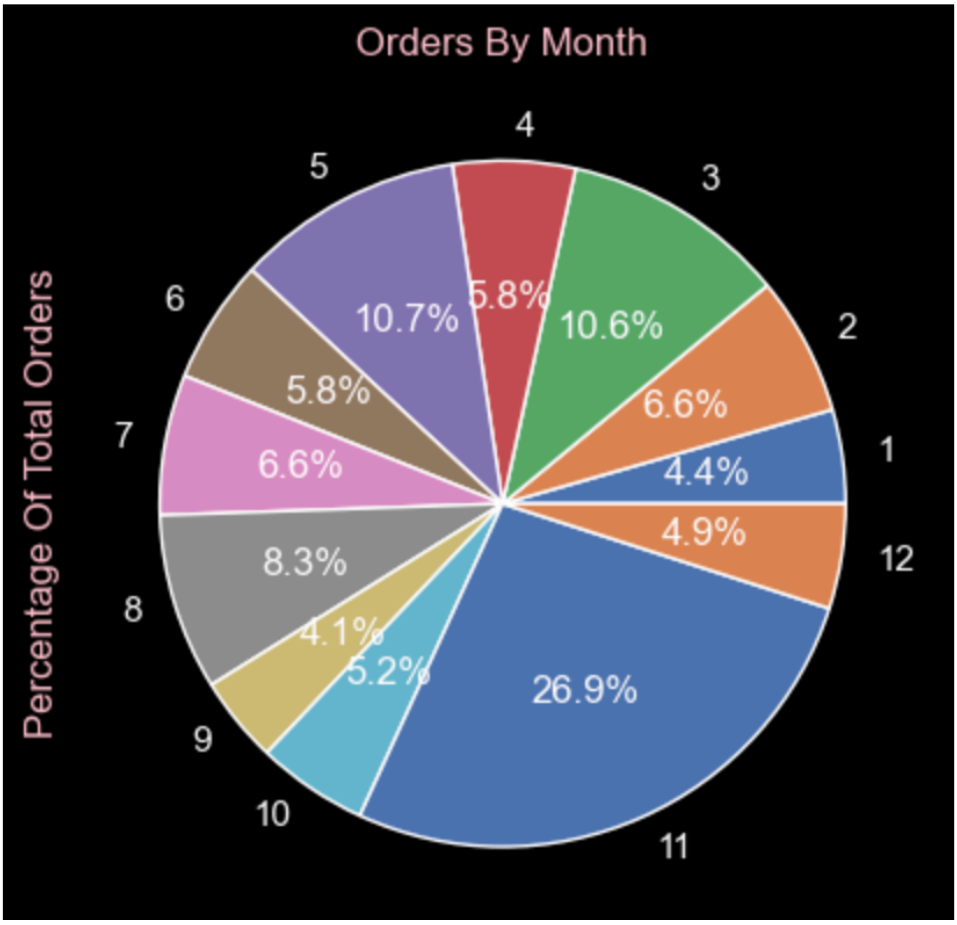
- **Mobiles & Tablets**:
  - Highest number of orders: **Up to 1000**.
  - Maximum price: **163k**, with only one order.
- **Beauty & Grooming**:
  - Orders: **200** for products priced at **400**.
  - Highest-priced product: **1M**, with only one order.
- **Women's Fashion**:
  - Highest number of orders: **150** for products priced at **300**.
  - Maximum price: **65k**, with only one order.

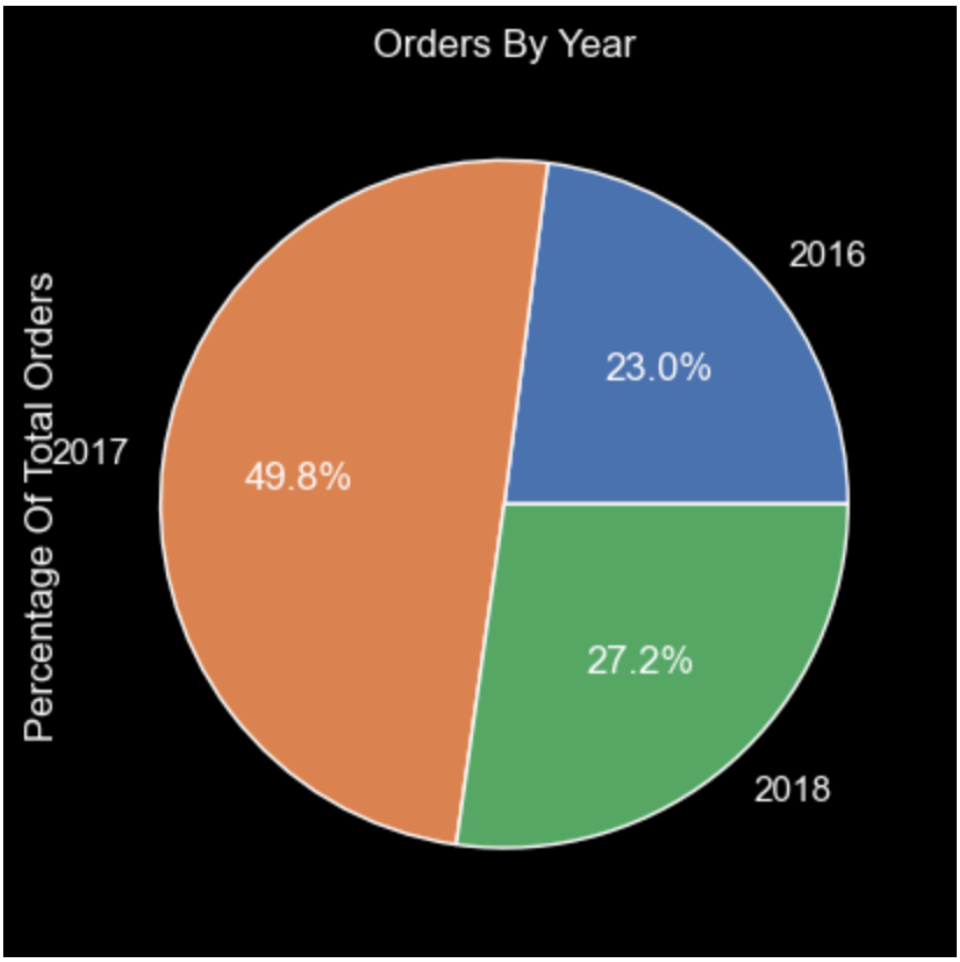
These insights help identify purchasing patterns across different product categories.





How does the number of orders vary over time?


Orders By Month
Percentage Of Total Orders
4
3
2
1
12
11
10
9
8
7
6
5
5.8%
10.6%
6.6%
4.4%
4.9%
26.9%
5.2%
4.1%
8.3%
6.6%
5.8%
10.7%


Orders By Year
Percentage Of Total Orders
2016
2017
2018
23.0%
49.8%
27.2%

- **Seasonal Trends**:
    - **Highest order volumes** were observed in **November**, reflecting typical holiday shopping behavior.
    - **Lowest order volumes** occurred in **September**, possibly due to a lull before the holiday shopping season.

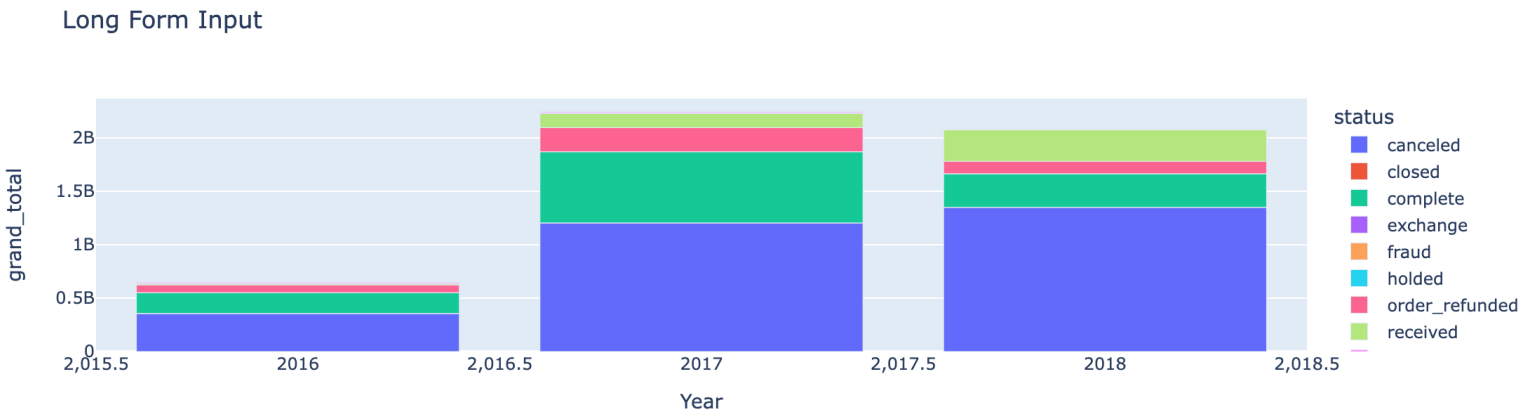
- **Year-over-Year Trends**:
    - Order volumes **peaked in 2017**, indicating that broader **economic conditions** and **industry trends** influenced demand during this time.

These insights demonstrate the influence of both seasonal factors and macroeconomic trends on consumer purchasing behavior.

How many customers return? Over time, what percentage of orders are from returning versus new customers?

- **Customer Retention**:
  - **55.99%** of all orders were placed by **return customers**, underscoring the importance of retaining existing customers.
  - While attracting new customers is essential for business growth, focusing on **customer retention** can be a more effective way to drive **long-term profitability**.

To maximize retention, the business might consider strategies like loyalty programs, personalized offers, and excellent customer service.



In each year order cancellation is high.

Task 5

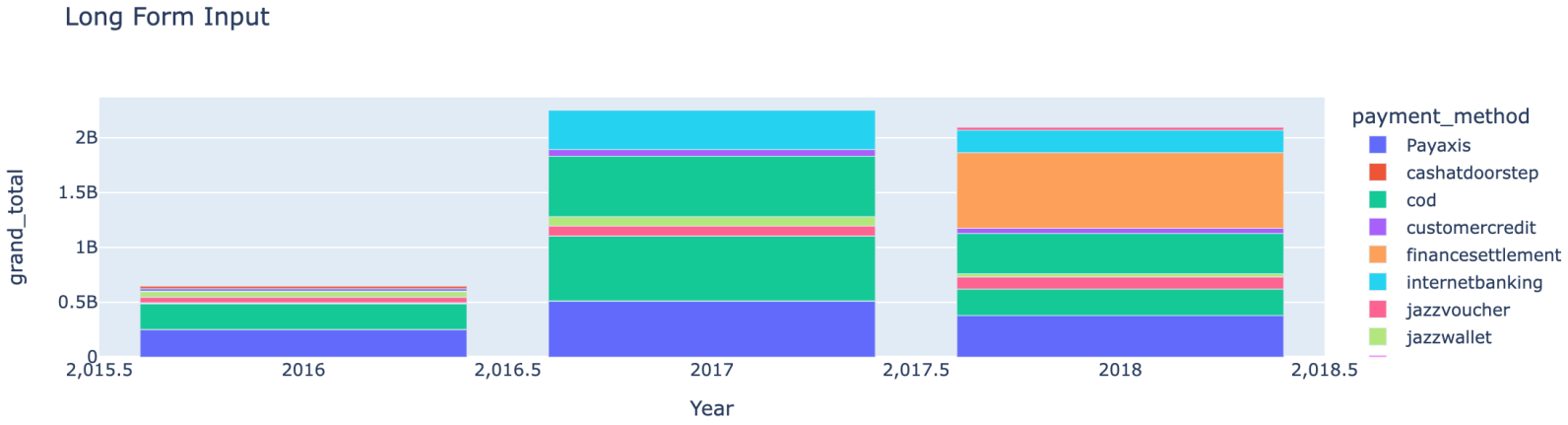
Biases

Distribution of sales by payment method over the years

Selection bias : user preference bias and availability bias

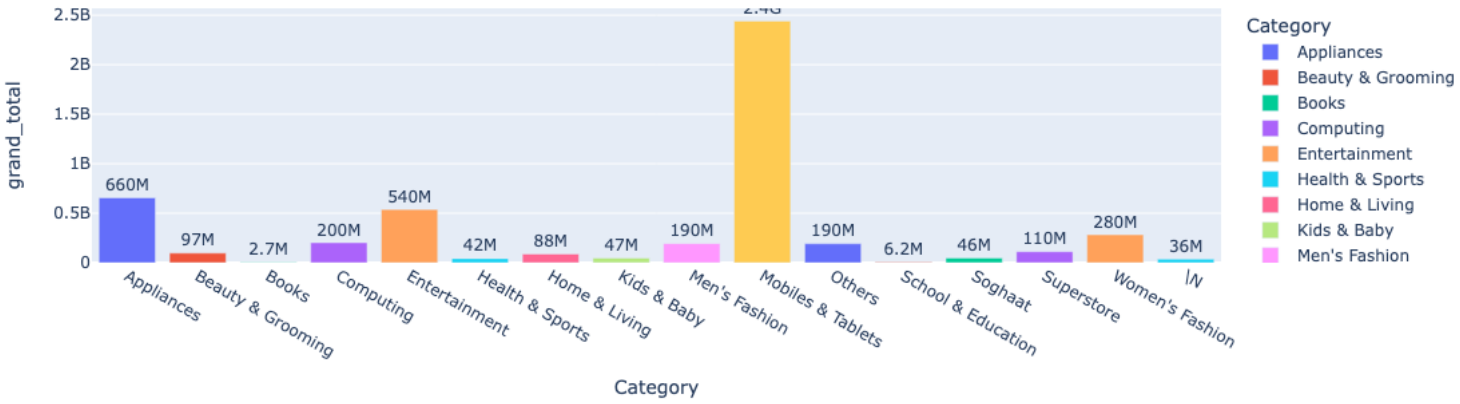
- **User Preference Bias**: Customers may overwhelmingly prefer using **COD (Cash on Delivery)** over other payment methods, possibly due to familiarity or trust issues with online payments.
- **Availability Bias**: Some payment methods (e.g., **Jazzvoucher**, **Jazzwallet**) are underrepresented, possibly due to limited availability or lower promotion, which could skew insights in favor of more common payment methods.

This makes the dataset less reflective of potential customer behaviors if all methods were equally accessible or promoted.



Category-wise, the total of items is analyzed : sales distribution

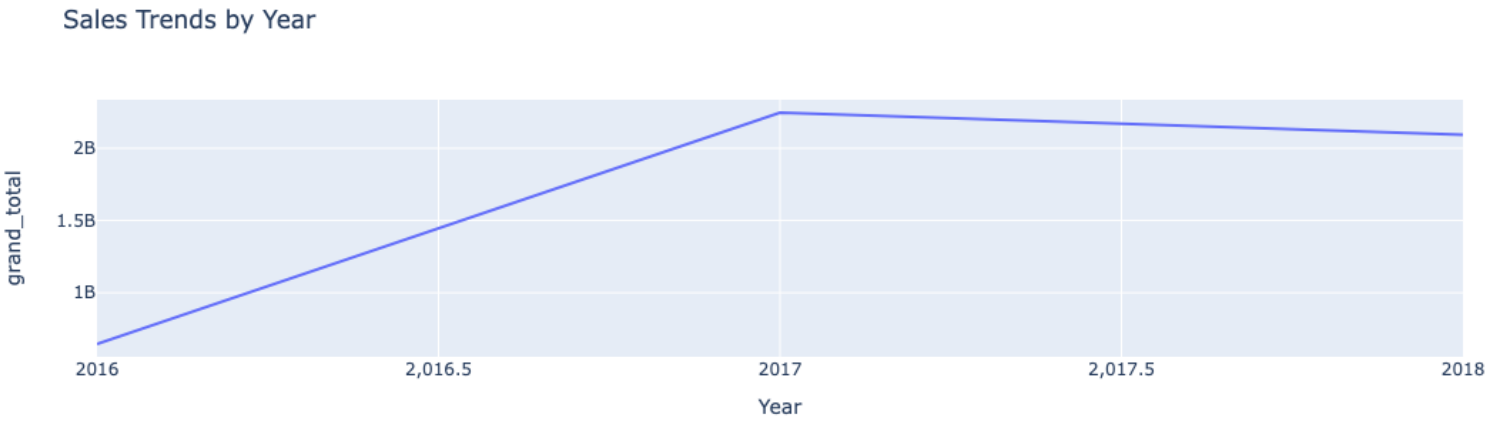
- **Product Category Bias:** Certain categories (e.g., Mobiles & Tablets) may dominate the sales, overshadowing others. This could indicate that the dataset favors popular categories and underrepresents niche products.
  - **Unequal representation**: Some categories are much more represented than others, which could skew business insights.
  - **Market focus**: The dataset might focus more on popular categories, underrepresenting other product types.



Sales trends by year

It appears that sales peaked in 2017 and slightly declined in 2018.

- **Temporal Bias**: The dataset may be limited to a specific time range (2016-2018). If only these years are analyzed, broader long-term trends may be missed.



Customer retention trends monthly

There seems to be a significant fluctuation in retention rates across months, with noticeable peaks and troughs.

- **Seasonal Bias:** Peaks in retention (e.g., month 6 and 10) could reflect seasonal shopping patterns (e.g., mid-year sales or holidays).

External Events: Sudden drops (e.g., month 8) might be due to external factors not captured by the dataset (economic changes, product shortages).

- **Data Imbalance:** Certain months may have more data points (e.g., promotions or campaigns) influencing higher retention. There is a significant peak in **November** (37,171 orders), indicating a likely **seasonal bias**, possibly driven by events like **Black Friday** or other holiday shopping trends. The second-highest volume occurs in **March** (20,926 orders). Months like **January** and **September** have much lower order counts (7,854 and 8,201, respectively), suggesting a potential underrepresentation during off-peak periods, further reflecting seasonal trends in consumer behavior.

```
Month
11    37171
3     20926
5     16070
8     13123
2     12331
7     10998
4      9429
12     9320
10     9057
6      8784
9      8201
1      7854
Name: count, dtype: int64
```