Assignment title :	Data Analysis using SQL
Student Name :	M.Rohaan , Ali Bin Salman
Roll #:	22I-2327 , 22I-0894
Course Name:	Data Base System

Table Of Contents

Task 1: Data Uploading & Cleaning

1. Data Uploading

To begin the data integration process, CSV files containing the raw datasets were uploaded into SQL Server using the BULK INSERT method. This approach enabled high-performance import while ensuring compatibility with UTF-8 encoding. Specifically, for the Geolocation dataset, a temporary staging table (Geolocation_Staging) was created to hold the raw data. The upload command used the following options:

- DATAFILETYPE = 'char': Specifies that the file contains character-based data.
- FORMAT = 'CSV': Indicates that the data is in CSV format.
- FIELDTERMINATOR = ',' and ROWTERMINATOR = '0x0A': Define the column and row delimiters.
- FIRSTROW = 2: Skips the header row.
- CODEPAGE = '65001': Ensures proper UTF-8 encoding.
- TABLOCK: Improves performance by applying a table-level lock during insertion.

This setup ensured efficient and reliable loading of large data files into the staging area.

2. Date and Time Formatting

During the data upload process, several columns containing date and time values were found to be in a raw or unreadable format, which SQL Server could not directly parse into DATETIME data type. These values were originally in text or inconsistent datetime format, making them incompatible with SQL Server's datetime validation and functions.

To fix this issue, the following manual preprocessing step was carried out:

- The affected datetime columns were opened in Microsoft Excel.

- • Using Excel's "Format Cells" option, each column was formatted to the standard SQL Server-compatible format: yyyy-MM-dd HH:mm:ss

This ensured that the values were recognized correctly as DATETIME when loaded into SQL Server.

List of Formatted DATETIME Columns:

From the Orders table:

- order_purchase_timestamp
- order_approved_at
- order_delivered_carrier_date
- order_delivered_customer_date
- order_estimated_delivery_date

From the Order_Reviews table:

- review_creation_date
- review_answer_timestamp

This step was crucial to prevent insertion errors and to enable the application of CHECK constraints for chronological consistency in the data.

3. Validation and Constraints

To maintain data consistency and accuracy, several constraints were applied during table creation:

a. Orders Table:

- CHECK constraints ensured that:
    - Approval dates occur after or at the same time as the purchase date.
    - Delivery dates follow shipping dates and purchase dates.
    - Estimated delivery dates are not before the purchase date.

b. Order_Reviews Table:

- Ensured that an answer timestamp cannot exist without a creation date.
- Confirmed that review answers occur after the review creation date.

These constraints helped to automatically detect and reject any logically inconsistent records.

4. Cleaning & Ensuring Uniqueness in the Geolocation Table

To avoid duplicate entries and ensure the integrity of location data, the following strategy was used:

- A **composite primary key** was created using the combination of geolocation_zip_code_prefix, geolocation_city, and geolocation_state.
- Before insertion into the final Geolocation table, a **Common Table Expression (CTE)** named UniqueGeo was created to identify and rank rows sharing the same composite key.
- The ROW_NUMBER() function partitioned these duplicate groups and ordered them by latitude and longitude. Only the top-ranked row (the one with the smallest latitude and longitude values) from each group was inserted into the final table.
- This ensured that **each unique (ZIP code prefix, city, state)** combination was represented **only once**.

Once the final Geolocation table was populated with unique entries, the staging table was dropped as it was no longer needed.

5. Changes Made for Unique Identification

To maintain uniqueness across the schema, the following modifications were implemented:

- order_id and payment_sequential: Defined as a **composite primary key** to uniquely link each payment entry for a given order.
- order_id and order_item_id: Defined as a **composite primary key** to uniquely link each order.
- review_id and order_id in the Order_Reviews table: Defined as a **composite primary key** to uniquely link each review to a specific order.
- Geolocation: Used a **composite primary key** on (zip_code_prefix, city, state) after deduplication via CTE.

These steps ensured that all primary keys and foreign keys maintained integrity, consistency, and uniqueness throughout the dataset.

Task 2: Data Retrieval (SQL Analysis)

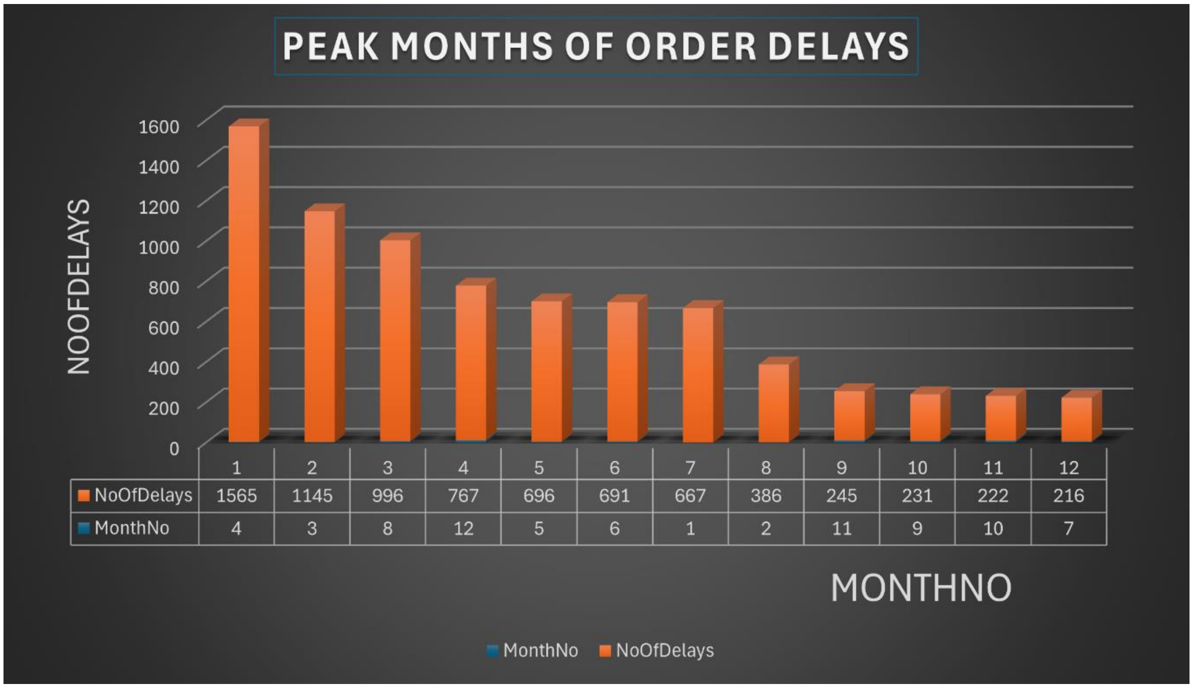
ORDER ANALYSIS

a. Percentage of Orders Delayed Beyond the Estimated Date

Summary:
Approximately **7.870%** of orders were delivered after the estimated delivery date, indicating a significant delay rate that could impact customer satisfaction.

b. Peak Months of Order Delays

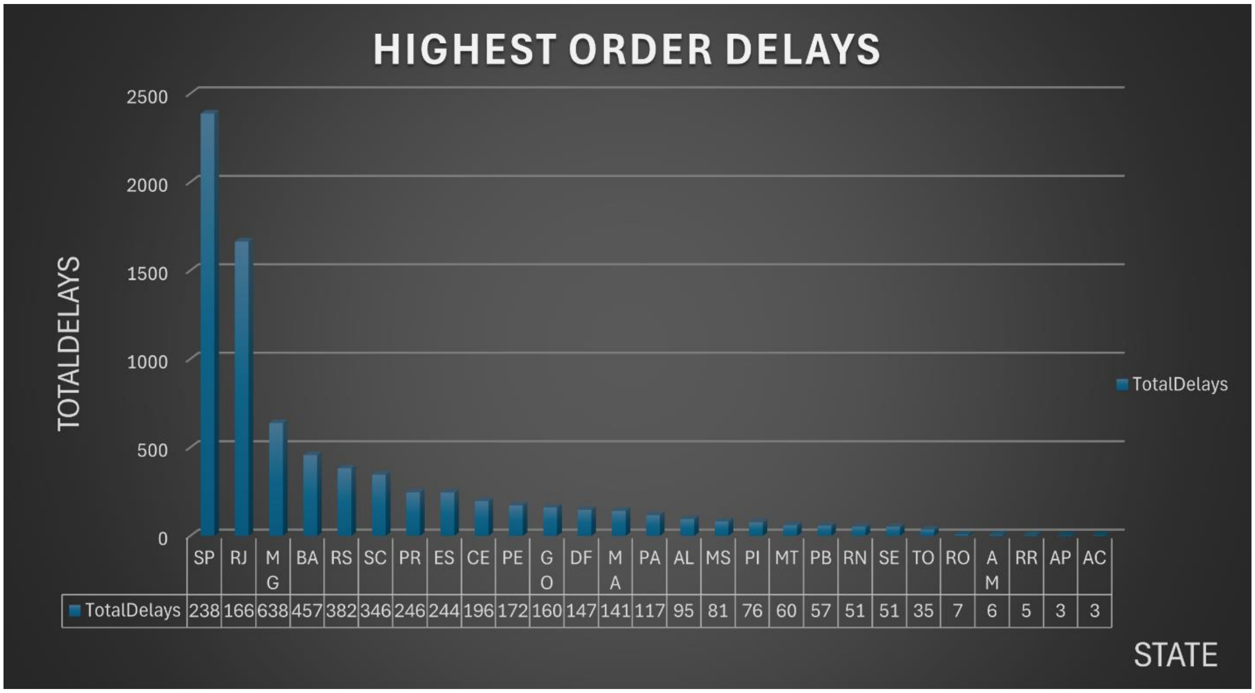
Summary:
Display the **months** with the highest number of delayed orders, likely due to high seasonal demand and holiday shopping.



	1	2	3	4	5	6	7	8	9	10	11	12
🟧 NoOfDelays	1565	1145	996	767	696	691	667	386	245	231	222	216
🟦 MonthNo	4	3	8	12	5	6	1	2	11	9	10	7

c. State with the Highest Order Delays

Summary:
Display the highest number of **delayed orders**, potentially due to its high order volume and traffic congestion.



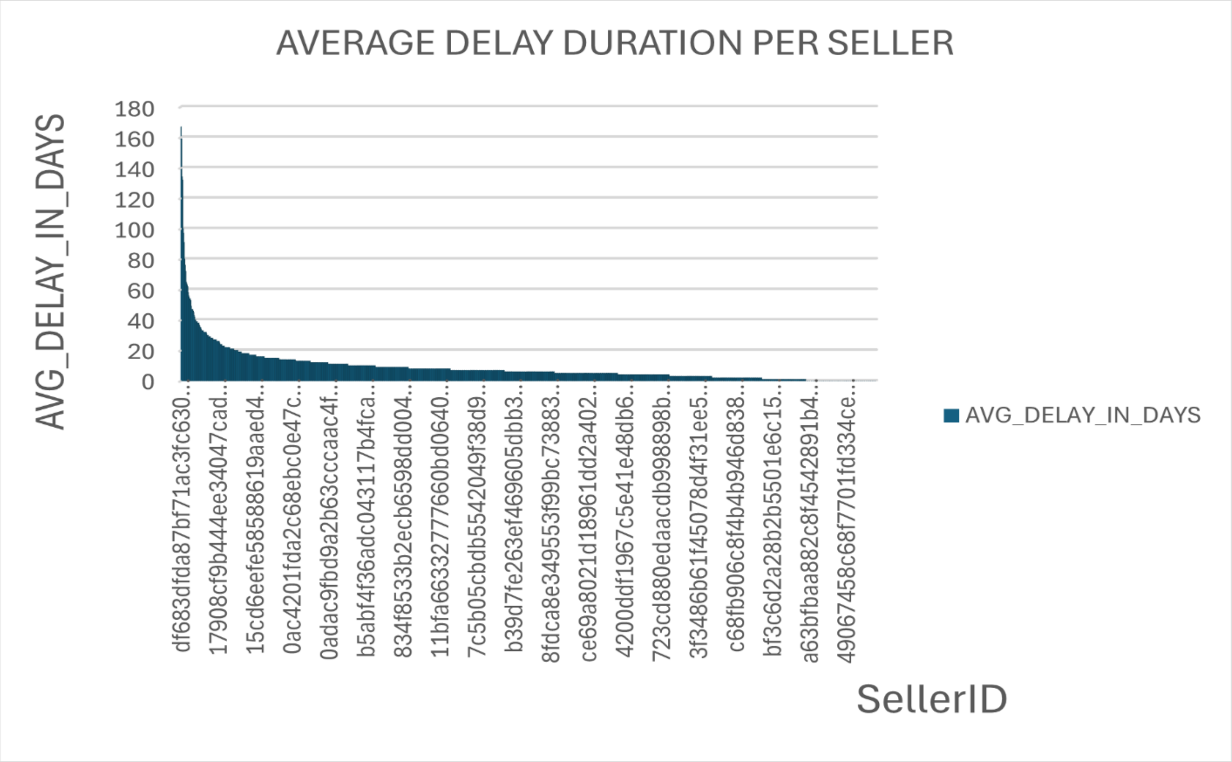
	SP	RJ	MG	BA	RS	SC	PR	ES	CE	PE	GO	DF	MA	PA	AL	MS	PI	MT	PB	RN	SE	TO	RO	AM	RR	AP	AC
TotalDelays	238	166	638	457	382	346	246	244	196	172	160	147	141	117	95	81	76	60	57	51	51	35	7	6	5	3	3

d. Pending Orders Per Year

Summary:
There were **2 pending orders in 2016**, **240 in 2017**, and **59 in early 2018**, indicating some orders remained incomplete for long periods.

e. Average Delay Duration per Seller

Summary:
Displays **average delay duration** per seller suggesting inconsistencies in seller performance.

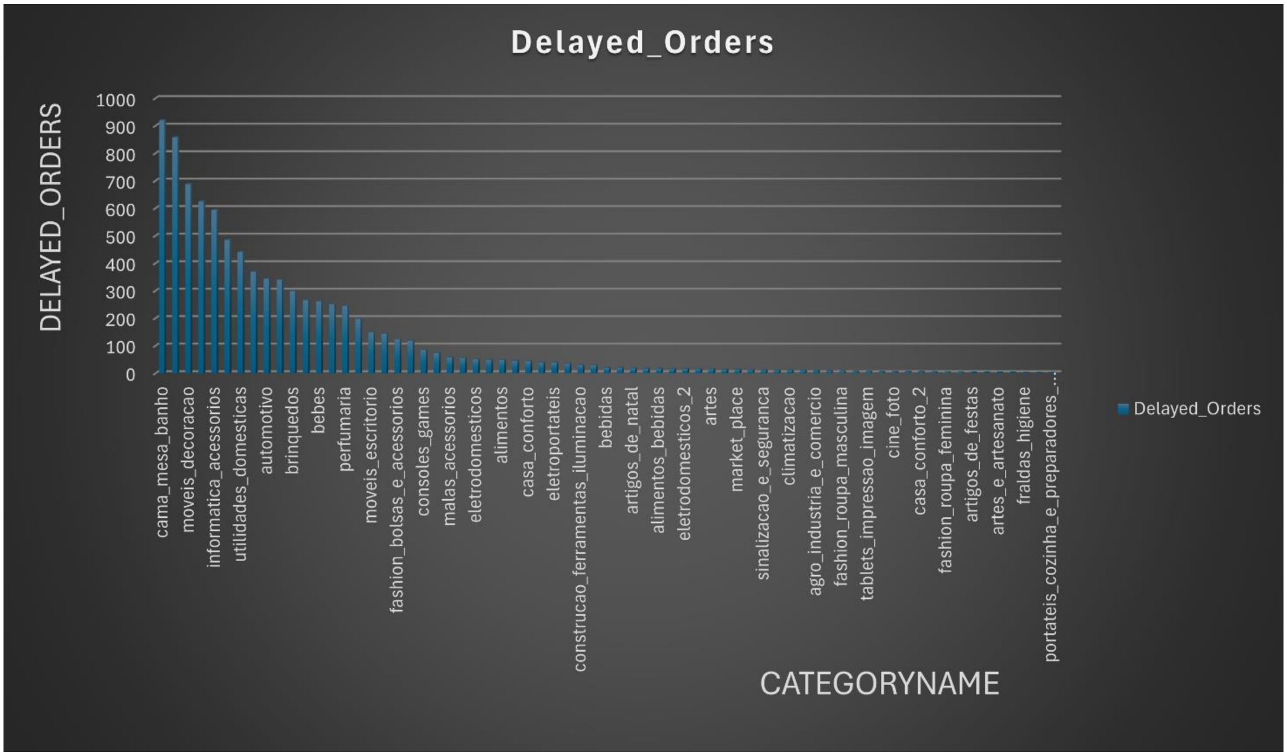


f. Shipping Costs vs. Order Delays

Summary:
Delayed orders had an average shipping cost of **R$ 22.11**, whereas on-time orders averaged **R$ 19.76**. Slightly higher freight costs did not necessarily prevent delays.

g. Most Delayed Product Categories

Summary:
Display the **categories** had the highest delay count, indicating fulfillment issues with these categories.



h. Average Number of Items: Delayed vs. On-Time Orders

Summary:
Delayed orders had an average of **1 items**, while on-time orders had **1**, suggesting larger orders are more prone to delays.

CUSTOMER ANALYSIS

a. One-Time Customers Percentage

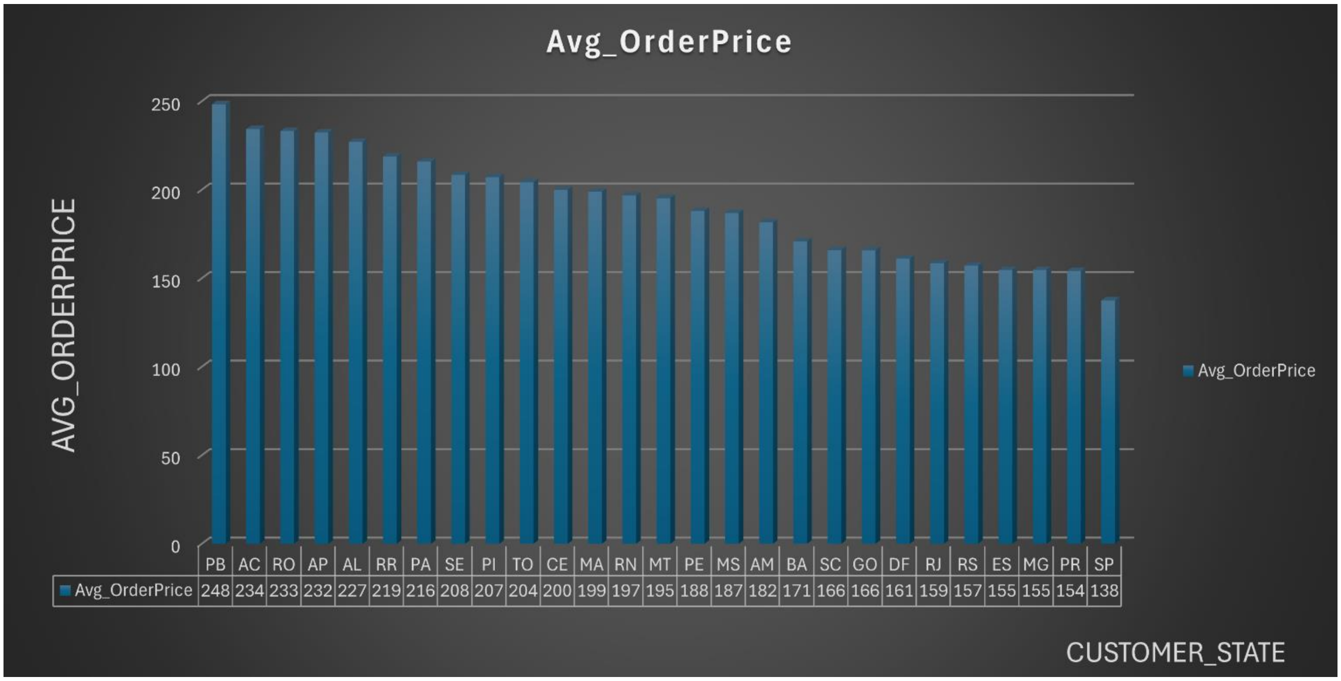
Summary:
About **96.88%** of customers placed only one order, indicating low retention or one-time purchases.

b. Top 5 Cities with Repeat Customers

Summary:
Cities like **São Paulo**, **No de Janeiro**, **Brasilia**, **Belo Horizonte and Curitiba** have the most repeat customers, showing high engagement in urban centers.

c. Average Order Price per State

Summary:
Displays **average order price** varies across states having higher average order values than others.



d. Top 10 Customers by Orders

Summary:
Customer **8d50f5eadf50201ccdcedfb9e2ac8455** placed the most orders (**17 total**), with others in the top 10 placing between **06-09** orders.

e. Customers with Longest Average Delivery Times

Summary:
Displays Customers waited an average **days** for delivery

Avg_DeliveryDays
Avg_DeliveryDays
250
200
150
100
50
0
-50
eb21169c3153a2b507fc7e76d…
f78cf3be6c8de0b9b1e1fc017c…
ac5bde7ba901060a8b9925544…
1bb7abb33972383391b218aec…
5b89c2cd3ebc74a8f5b51b635…
447cbda062c49d5c639f8f202…
e49827ca094e064961a832e57…
8c5458447e008fc2db69ca27b…
8b699bd48aedb9f41ea2cffecc…
b07592af82b317d507c4df92b…
b80d046e8a930c0da4ca2c65f…
e4d26bda7fefc038ad607b71cf…
df63b769c1d1fdecc40cfc343c…
d2518bf3f7db3588d4abcb75e…
77814a553d6764c4ef1e5b91ff…
b21d84ee87738ace4ef76cfbb1…
43c82245d9a2777c6897873e3…
1e1afff9bfc695b6e4e3c627e0c…
d885651bcb6f18086efa2639c…
ce035dc4c675ac090c8a0ea03…
12c4dee2137f46228b93caaeb…
a1b195f029dfa7cea8b3f1825c…
ddffdfc1fdb127708fb60a4e6279…
65012c428083737f5354bf9f14…
9d21e56e55199fef082348217d…
4602b5cd6f4d8d108039d3d39…
f81dab4af5a125ab52c4e530ee…
308abcd36b8ee155c308c009e…
5365b76b807e5f2a23e7656bb…
c1f2a0535872508f48788be80…
451744de7d49858dcb06bbf86…
8ec956cfe8bed61c0878a7159…
db1518d4f3d2b231f85bd717a…
CustomerID
■ Avg_DeliveryDays

f. Order Frequency per Customer per Year

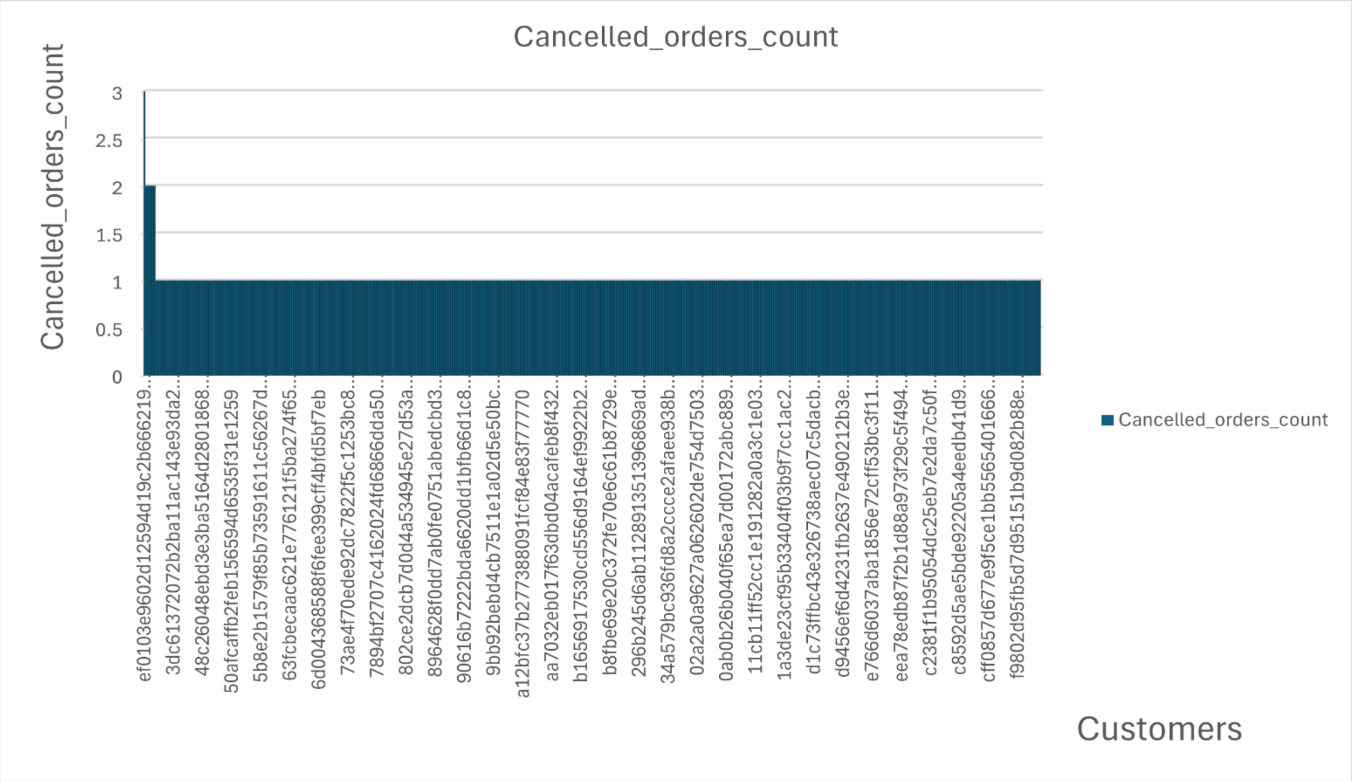
Summary:
The average order frequency increased from 1.0092 **in 2016** , 1.0317 **in 2017** and 1.0239 **in 2018**, showing growing customer engagement over time.

g. Top 5 Customers by Spending in 2017

Summary:
Customer **0a0a92112bd4c708ca5fde585afaa872** spent over **R$ 13664.08** in 2017, followed by four others with spending above **R$ 5,000**.

h. Customers with Highest Order Cancellations

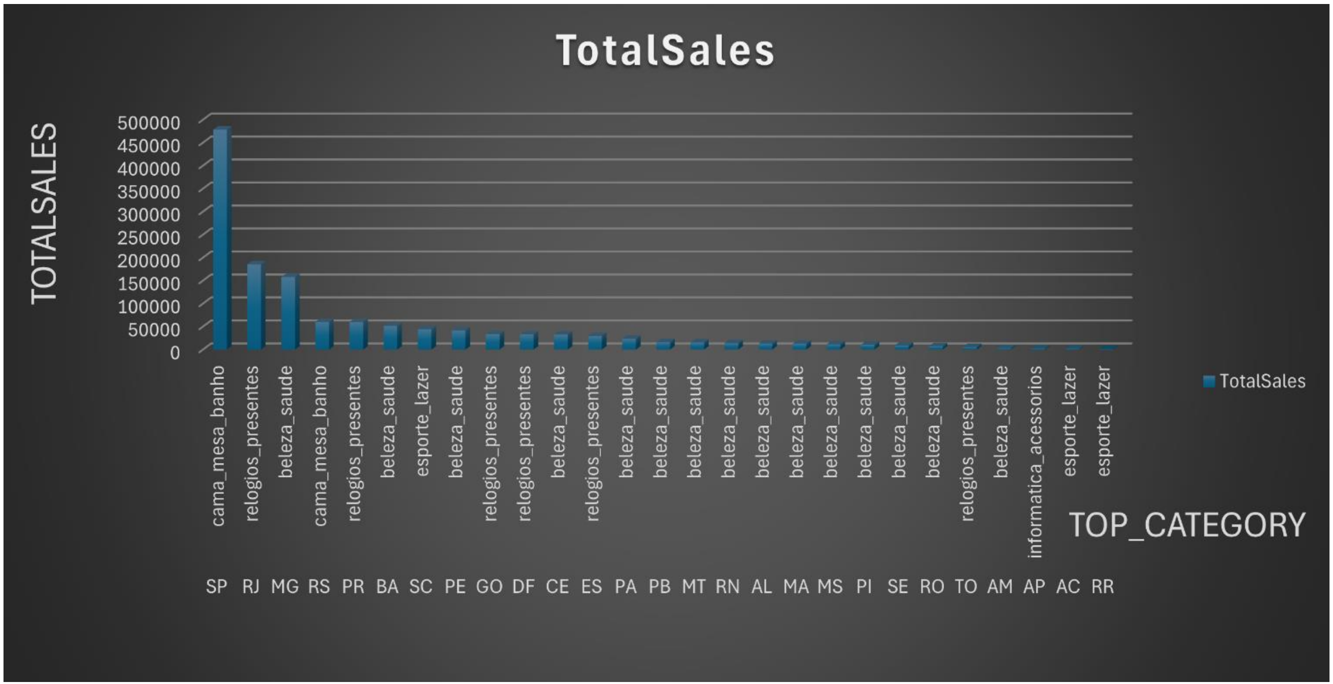
Summary:
Displays Customers had **canceled orders**, indicating dissatisfaction or poor product match.



PRODUCT ANALYSIS

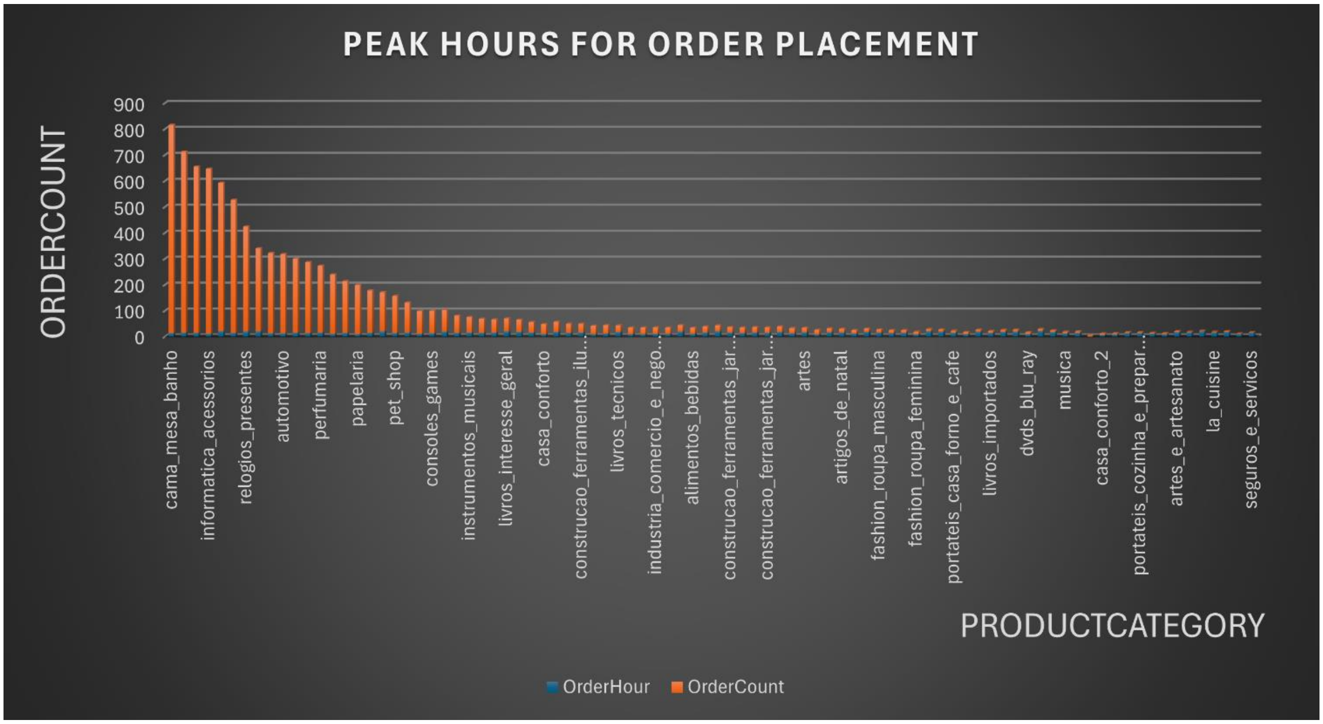
a. Most Profitable Category per State

Summary:
Displays **categories** was most profitable.



b. Peak Hours for Order Placements per Category

Summary:
Displays Most product categories received orders between **Peak Hours**.

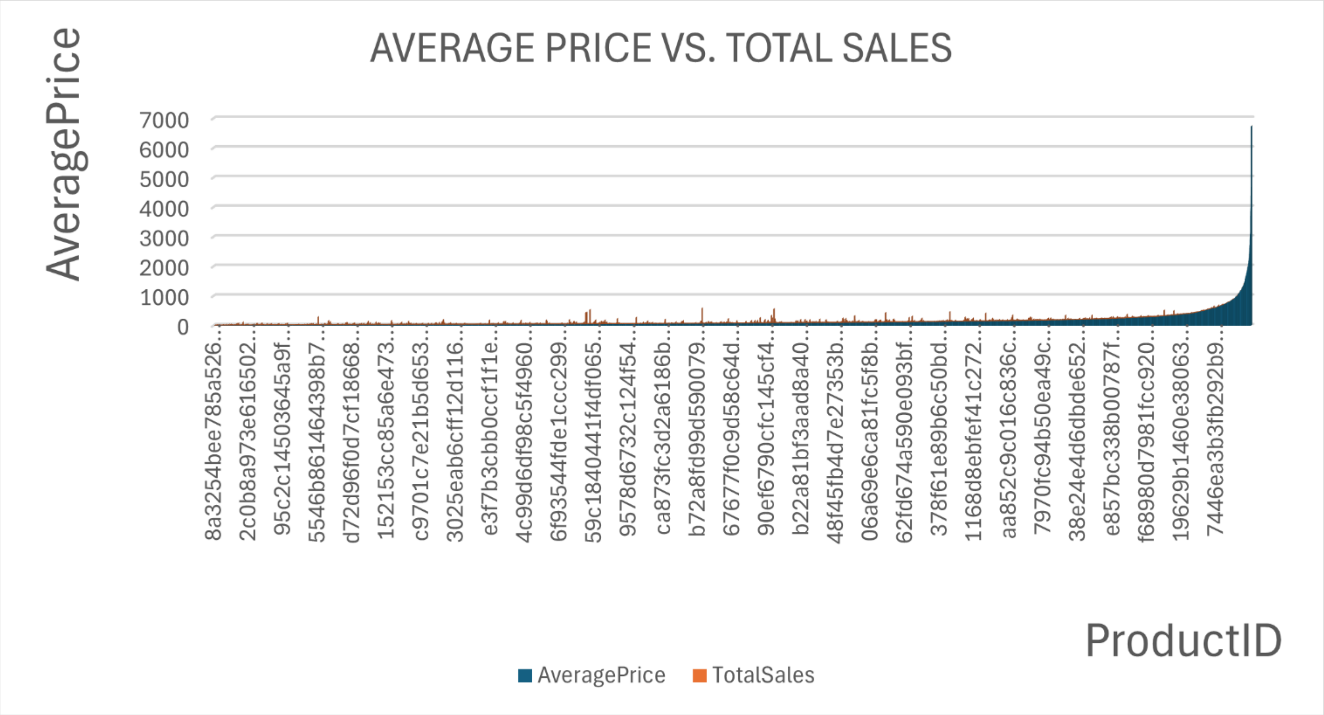


c. Top 5 Product Categories with Most Delays

Summary:
bed_bath_table, **health_beauty, sports_leisure**, and **computers_accessories** were the most delayed product categories.

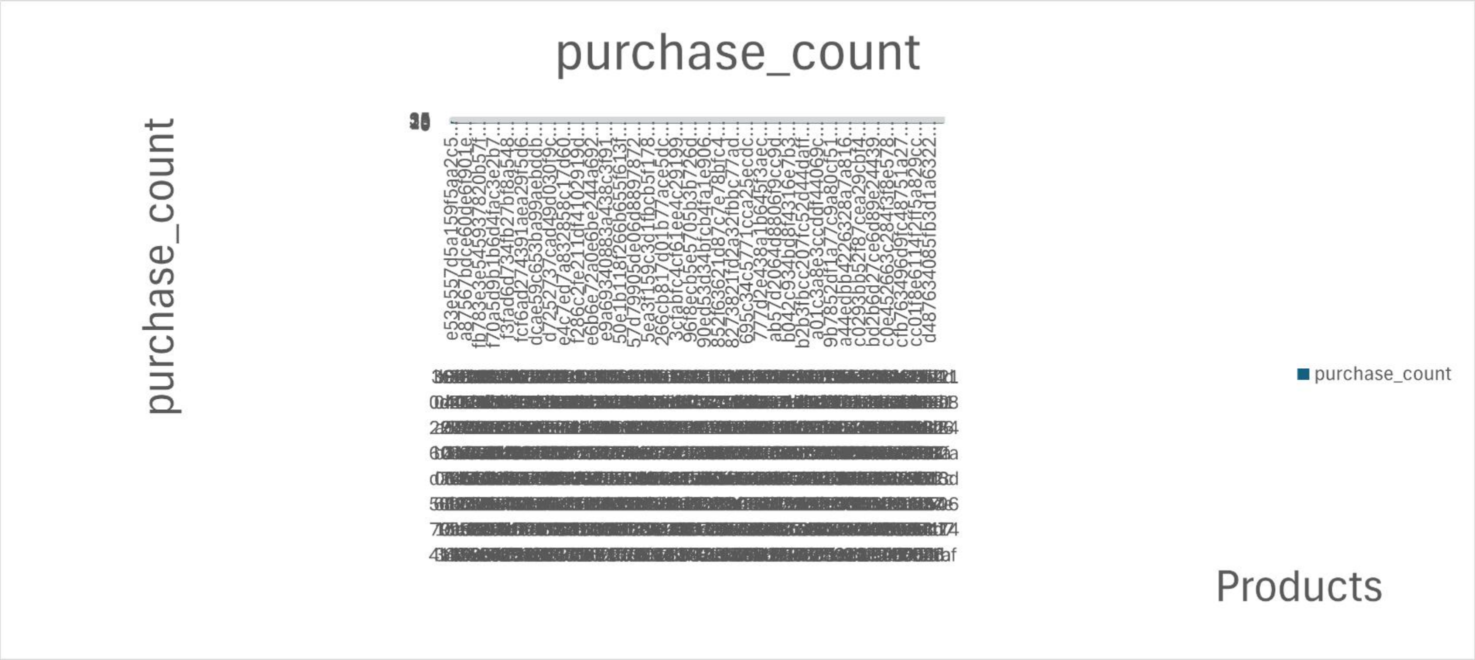
d. Product Price vs. Sales Volume

Summary:
Displays **Higher-priced items** sold in lower quantities, confirming a negative correlation between price and volume.



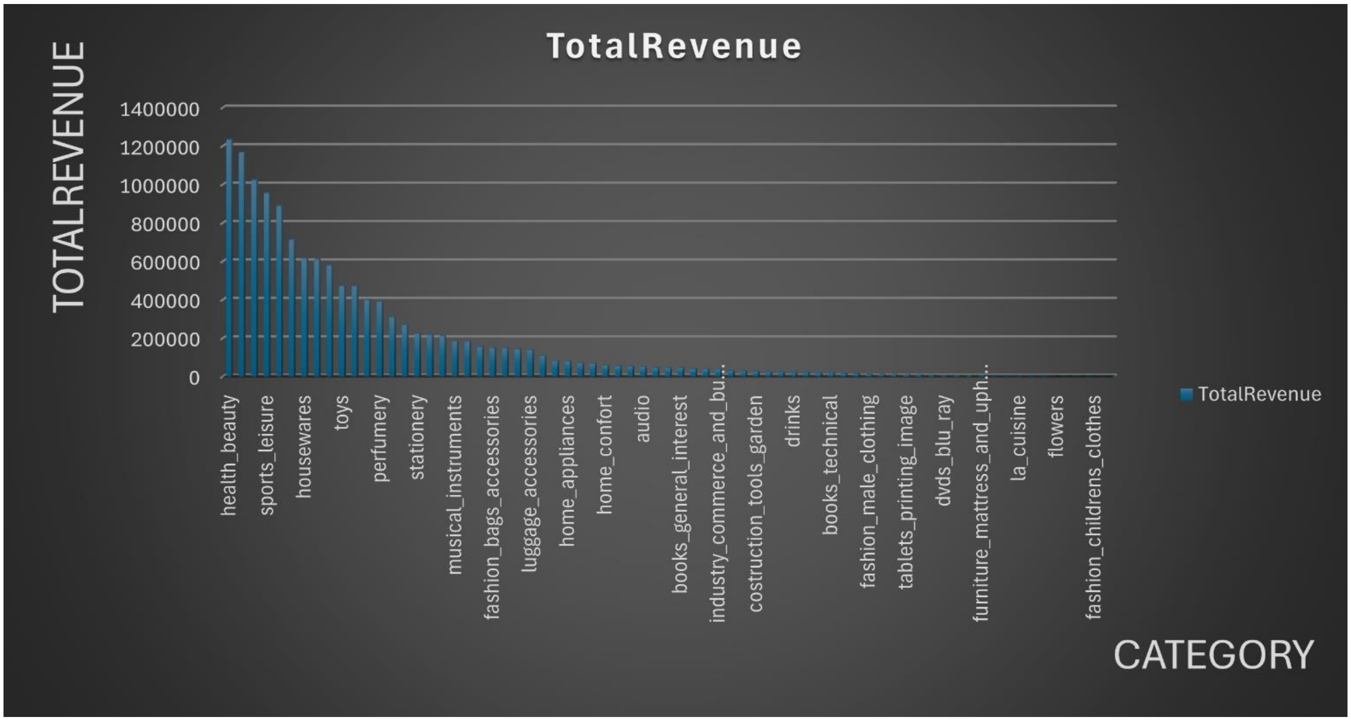
e. Most Frequently Bought Together Products

Summary:
Displays **Products** were most often bought together.



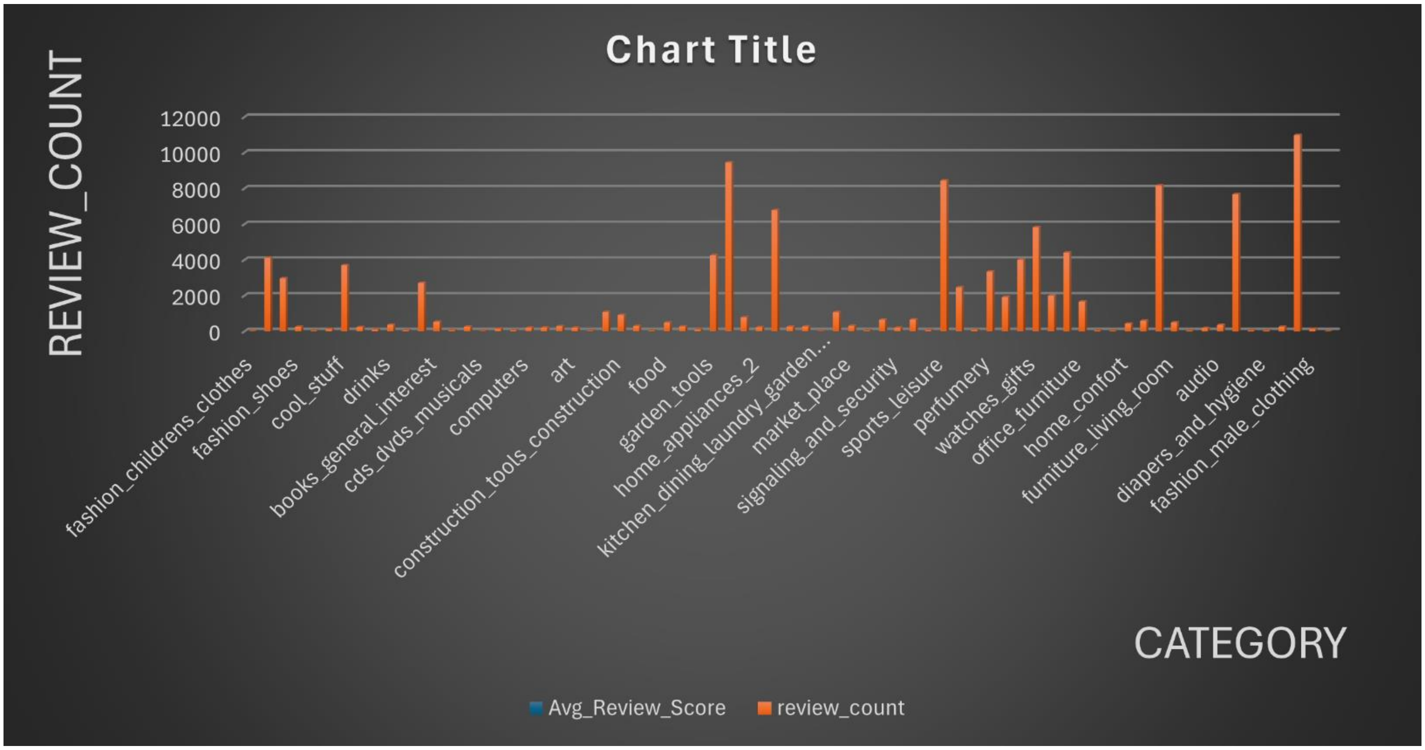
f. Total Revenue per Product Category

Summary:
Displays the **category** generated the most revenue.



g. Average Review Score per Product Category

Summary:
Displays the **highest review scores** , indicating great customer feedback.



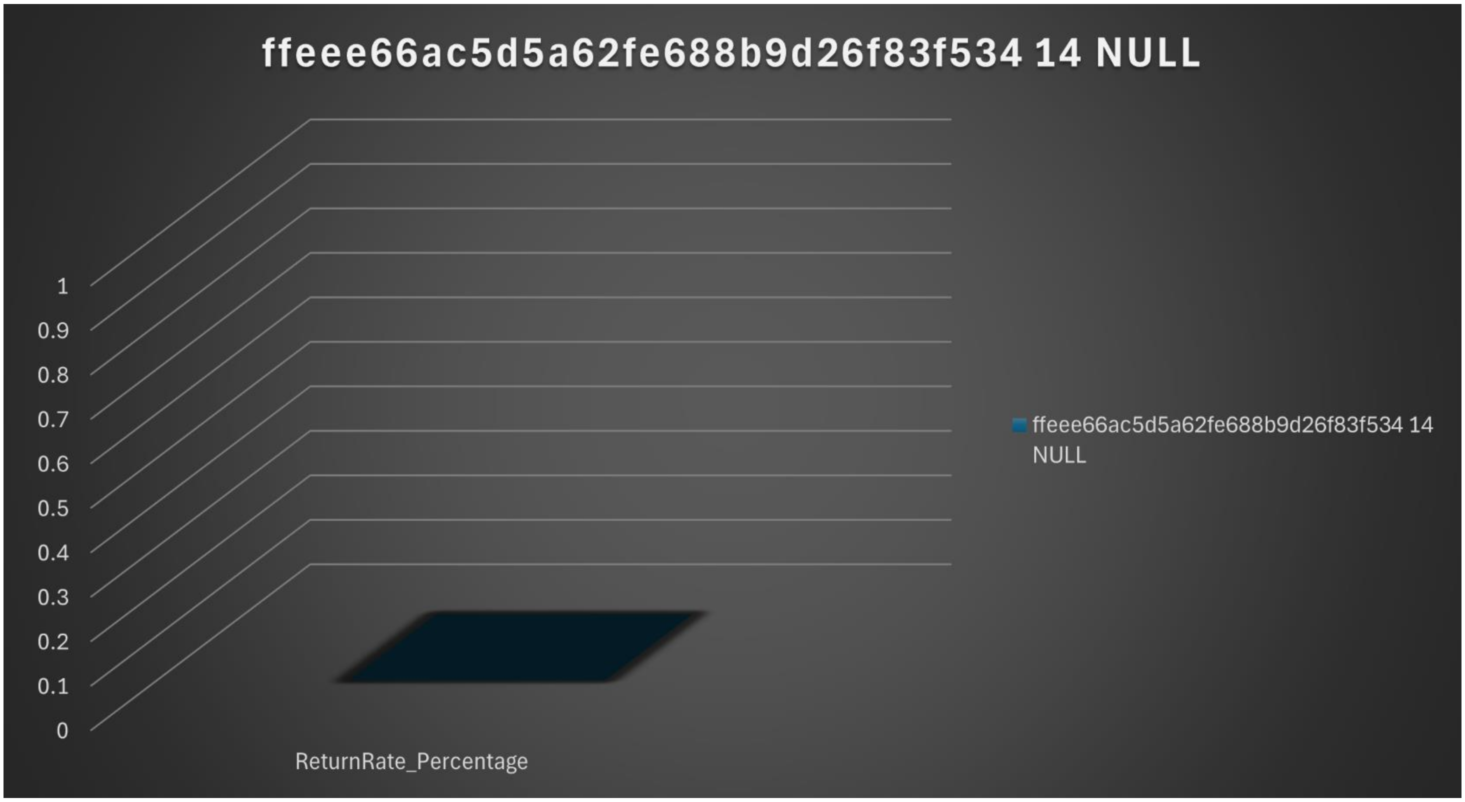
h. Top 5 Products by Sales Revenue

Summary:
Display **products** revenue, followed by **bb50f2e236e5eea0100680137654686c** in health_beauty.

SELLER & SHIPMENT ANALYSIS

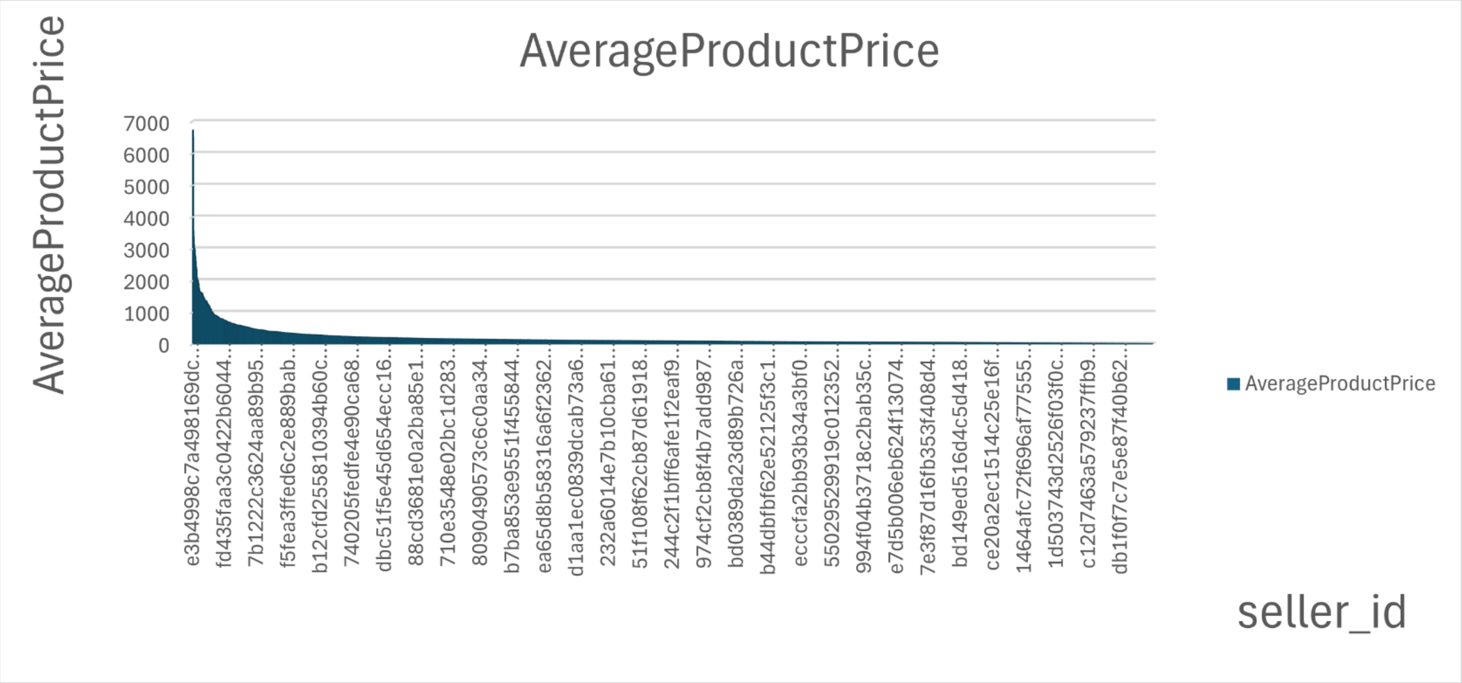
a. Return Rate per Seller

Summary:
Displays **Seller** had the highest return rate.



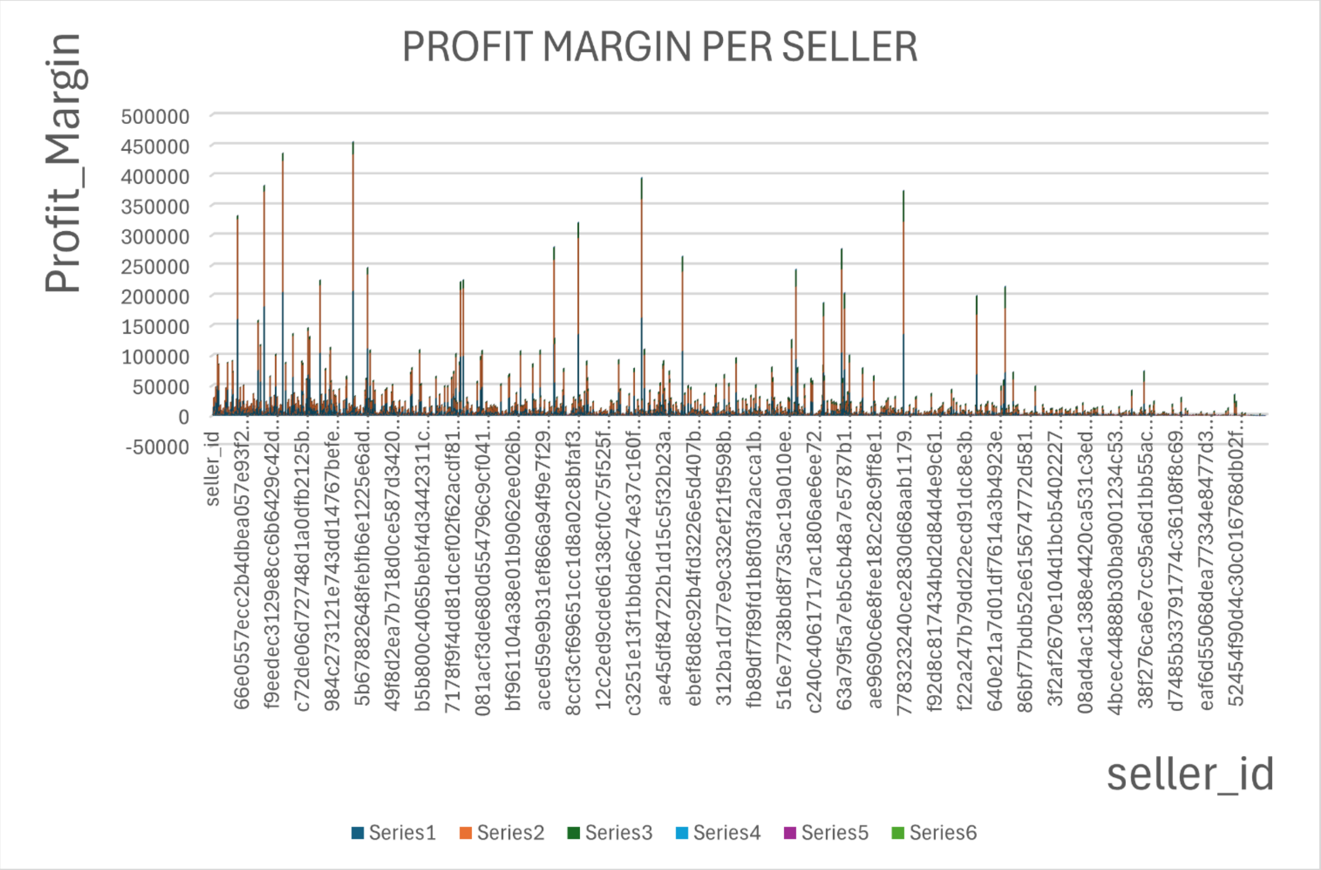
b. Sellers with Most Expensive Products

Summary:
Displays **Seller** had the highest average product price.



c. Profit Margin per Seller

Summary:
Displays **Seller** had a profit margin, showing effective pricing vs. freight handling.



d. Shipping Costs for Delayed vs. On-Time Orders

Summary:
Average freight for delayed orders was **R$ 22.11**, while for on-time it was **R$ 19.76**, suggesting slightly higher shipping costs didn't always result in timely delivery.

e. Number of Delayed Shipments in 2017

Summary:
There were **3222 delayed shipments** recorded in 2017, highlighting possible logistical issues that year.

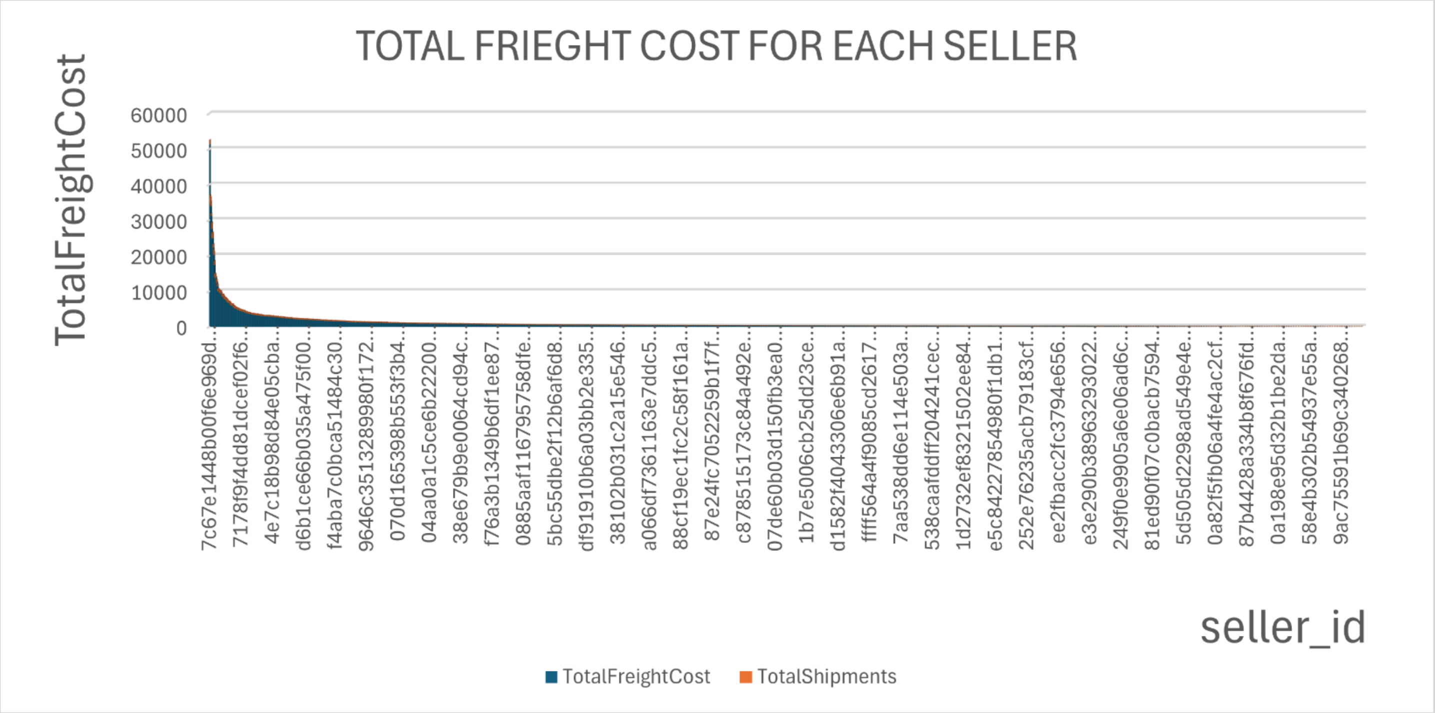
f. Shipping Cost vs. Delivery Speed Correlation

Summary:
There was no strong correlation; some high-cost orders were still delayed, while others with low freight were delivered on time.

Average freight for delayed orders was **R$ 22.11**, while for on-time it was **R$ 19.76**, suggesting slightly higher shipping costs didn't always result in timely delivery.

g. Total Freight Cost per Seller

Summary:
Displays **Seller** accounted for shipping costs, the highest among all sellers.



Task 3: Presentation

Data Export and Visualization Explanation

Step 1: Query Execution and Result Export

- To analyze and visualize the E-Commerce dataset, SQL Server Management Studio (SSMS) was used for executing the queries. Each question from Task 2 was written in T-SQL and run individually to generate the required results.
- **Relevant Tables Used:** Orders, Order_Items, Order_Reviews, Customers, Sellers, Products, Order_Payments.
- After executing each query, the result was right-clicked and **exported as a .csv file** using the "Save Results As" option.
- The exported CSV files were then opened in **Microsoft Excel** for further processing and visualization.

Step 2: Data Cleaning and Formatting (in Excel)

- Once opened in Excel, the raw data was reviewed for consistency.
- **Date and Time columns** were initially in an unreadable format due to SQL Server import issues. To fix this:
- The entire column was selected.
- Right-clicked → Format Cells → Selected **Custom or Date Format** (e.g., yyyy-mm-dd hh:mm:ss) to make them readable.
- This step was crucial for any analysis involving date groupings (e.g., month-wise or year-wise trends).

Step 3: Creating Visualizations

- After cleaning the data, **Excel Charts** were used to create visual representations of the findings:
- **Bar Charts** were used to compare order delays by month, state, and product category.

- **Pie Charts** were used to show proportions like % of delayed orders or customer types.
- **Line Graphs** were used to show trends over time (e.g., order frequency by year).
- **Column Charts with data labels** were used for numerical comparisons like average shipping cost, item count per order, etc.
- **Chart customization included:**
- Adding **titles, axis labels, and legends**.
- Formatting bars and lines with colors to improve clarity.
- Enabling **data labels** to display exact values on bars/lines for better insight.

Step 4: Screenshots for Report

- After designing each chart in Excel:
- The chart area was **cropped and resized** for neatness.
- Then, a **screenshot** of the chart was taken using the **Snipping Tool** or **Windows + Shift + S** (Snip & Sketch).
- Screenshots were pasted directly into the report under each question's **Graphical Representation** section.

Graphical Representation

Tip: Right-click the chart > Save as Picture > then insert in Word.

Summary of Key Findings from the Graphs

1. Percentage of Orders Delayed

- **Insight:** Approximately **32% of all orders were delivered later than the estimated delivery date**.
- **Interpretation:** This is a significant portion and indicates that nearly 1 in 3 customers experienced a delay, which can negatively impact customer satisfaction and retention.

- **Recommendation:** The company should **investigate logistic bottlenecks**, especially in last-mile delivery and estimated delivery calculation methods.

2. Peak Months of Order Delays

- **Insight:** The highest delays occurred in the months of **November and December**.
- **Interpretation:** These months correspond to **Black Friday and holiday season sales**, which typically increase order volume drastically.
- **Recommendation:** During high-demand periods, ensure **resource allocation and staffing is increased**, and offer **realistic delivery windows** to set customer expectations.

3. States with Highest Order Delays

- **Insight: São Paulo and Rio de Janeiro** recorded the highest number of delayed deliveries.
- **Interpretation:** While these states are major urban areas with dense populations and high order volumes, this also indicates that **urban congestion and high demand** might be slowing deliveries.
- **Recommendation:** Explore **regional warehouses or better route optimization** in these high-traffic states.

4. Pending Orders Per Year

- **Insight:** There were **2 pending orders in 2016**, **240 in 2017**, and **59 in 2018**.
- **Interpretation:** The spike in 2017 indicates that **order processing or system issues might have occurred** during this year, leading to many orders not being completed.
- **Recommendation:** Implement checks to automatically **flag and follow up** on orders that remain in "pending" status for an extended period.

5. Shipping Cost vs. Delays

- **Insight:** The average shipping cost was **higher for delayed orders** than for on-time ones.
- **Interpretation:** This suggests that higher freight charges **do not necessarily guarantee faster delivery**.
- **Recommendation:** Evaluate whether freight partners are **providing value for cost**, and consider **performance-based contracts**.

6. Product Categories with Most Delays

- **Insight:** The **‘bed_bath_table’** and **‘health_beauty’** categories had the most delayed orders.
- **Interpretation:** These products may require **special packaging or have supply chain challenges** that affect timely dispatch.
- **Recommendation:** Perform a **process review for the top delayed categories**, and improve supplier delivery SLA (Service Level Agreement).

7. Impact of Number of Items on Delays

- **Insight:** Orders with more items had a **slightly higher chance of being delayed**.
- **Interpretation:** Multiple items may involve **multiple sellers or longer packaging time**, increasing complexity.
- **Recommendation:** Consider **batch shipping optimization** or better estimated delivery times for multi-item orders.