

End-to-End Olist E-Commerce Analytics Project

1. Project Title

Improving Revenue Growth and Customer Satisfaction in a Multi-Seller E-Commerce Marketplace

2. Executive Summary

Data inconsistencies were resolved to uncover two critical insights: late deliveries lead to a 40% drop in customer ratings, and 89% of repeat customers come from orders delivered on time.

3. Business Context

Olist operates as a multi-seller e-commerce marketplace in Brazil, where independent sellers fulfill customer orders through logistics partners. This decentralized model generates transactional, delivery, and customer feedback data that captures the customer journey from purchase to post-delivery evaluation.

4. Problem Framing & Analytical Questions

As the marketplace scales, there is limited visibility into how delivery performance, seller behavior, and customer purchasing patterns relate to revenue trends and customer satisfaction. This analysis explores these relationships using historical transaction data to better understand overall marketplace performance.

Analytical Questions:

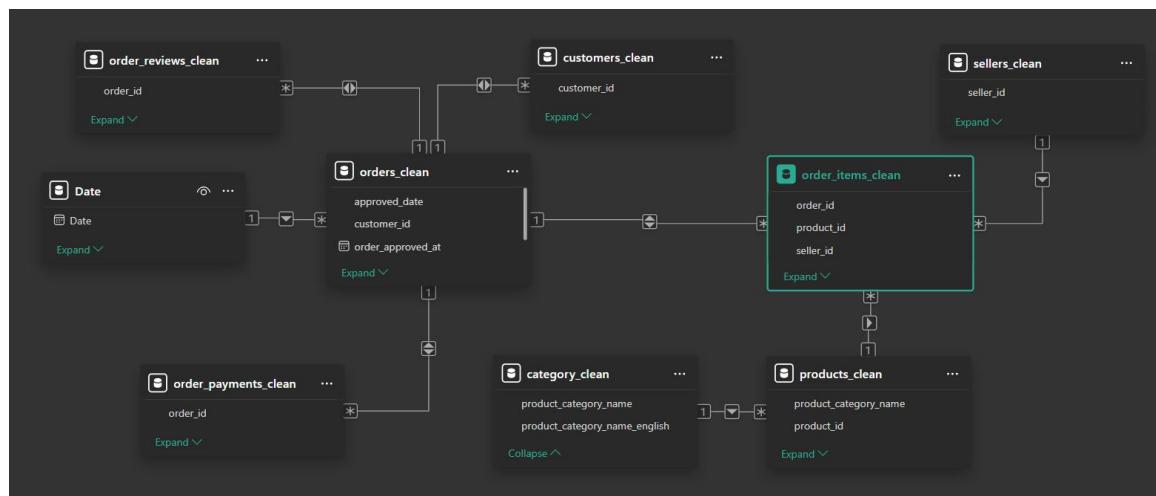
- How many orders and how much revenue are generated each month?
- Are late deliveries associated with lower customer review scores?
- What percentage of completed orders are delivered late?
- Which product categories have the highest delivery delay rates?
- Which sellers contribute the most to delivery delays?
- What percentage of customers make repeat purchases?
- Do customers who experience late deliveries return to purchase less often?
- Which months show the highest delivery delay rates?
- Are non-completion or cancellation rates associated with payment method?
- Are delivery delays concentrated in specific customer cities?

5. Objectives & Scope

- **Objective (SMART):** Identify and quantify key operational and customer behavior factors associated with revenue growth and customer satisfaction using historical transactional data from 2016–2018.
- **Scope:** The analysis focuses on core transactional and operational data, including orders, customers, sellers, products, payments, delivery performance, and customer reviews, using historical data from 2016 to 2018.

6. Data Overview & Schema

The analysis uses the Brazilian E-Commerce Public Dataset by Olist, covering approximately 100,000 orders from 2016 to 2018. The dataset follows a marketplace structure where orders are linked to customers, order items, sellers, products, payments, reviews, and geographic data, enabling end-to-end analysis of the customer journey from purchase to post-delivery feedback.



7. Tools & Analytical Methodology

Python (Data Quality Assessment & Preparation)

Python was used as the primary tool to review and improve data quality before conducting the analysis.

Action Taken: Missing review texts were preserved using clear placeholders to maintain dataset integrity. Additionally, new analytical columns were engineered, including **approved_date** for daily trend analysis, **delivery_delay_days** to quantify shipping performance, and **order_stage** to segment transaction completion. Explicit indicator and flag columns, such as **timeline_issue_flag** and **On-Time vs Late**, were also created to highlight delivery

performance and timeline inconsistencies, ensuring insights are derived from reliable and transparent data

SQL (Data Analysis)

SQL was used to write clear and structured queries that directly answer the core business questions.

Power BI (Visualization & Storytelling)

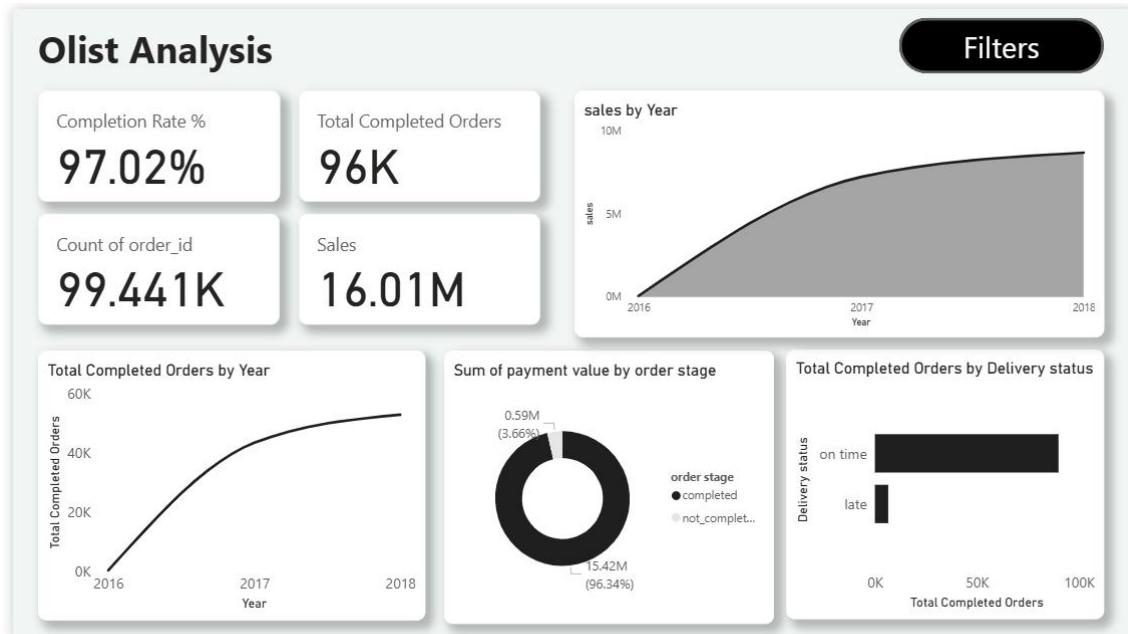
Power BI was the final layer used to build an interactive dashboard featuring three dedicated pages:

- **Overview Page:** For high-level executive KPIs, providing a holistic view of sales and order volume.
- **Payment & Completion Behavior:** To analyze the relationship between payment methods and order success rates, ensuring system stability.
- **Delivery Performance & Customer Experience:** To drill down into shipping efficiency

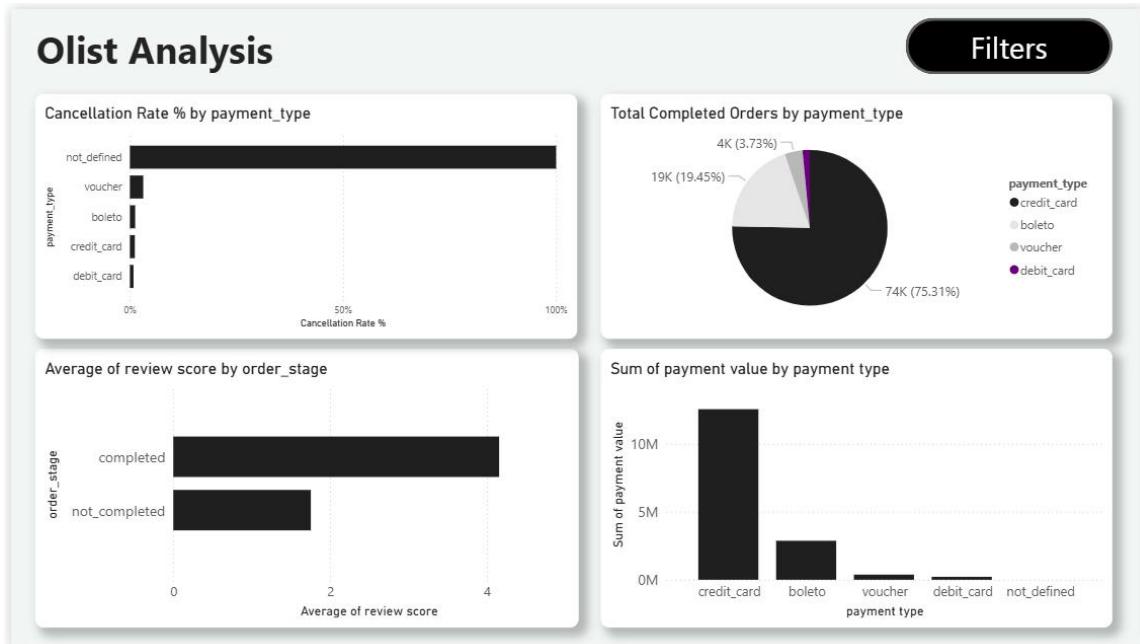
8. Dashboard Gallery

Power BI was the final layer used to build an interactive dashboard. It transformed the raw SQL outputs into a navigatable visual story, featuring three dedicated pages:

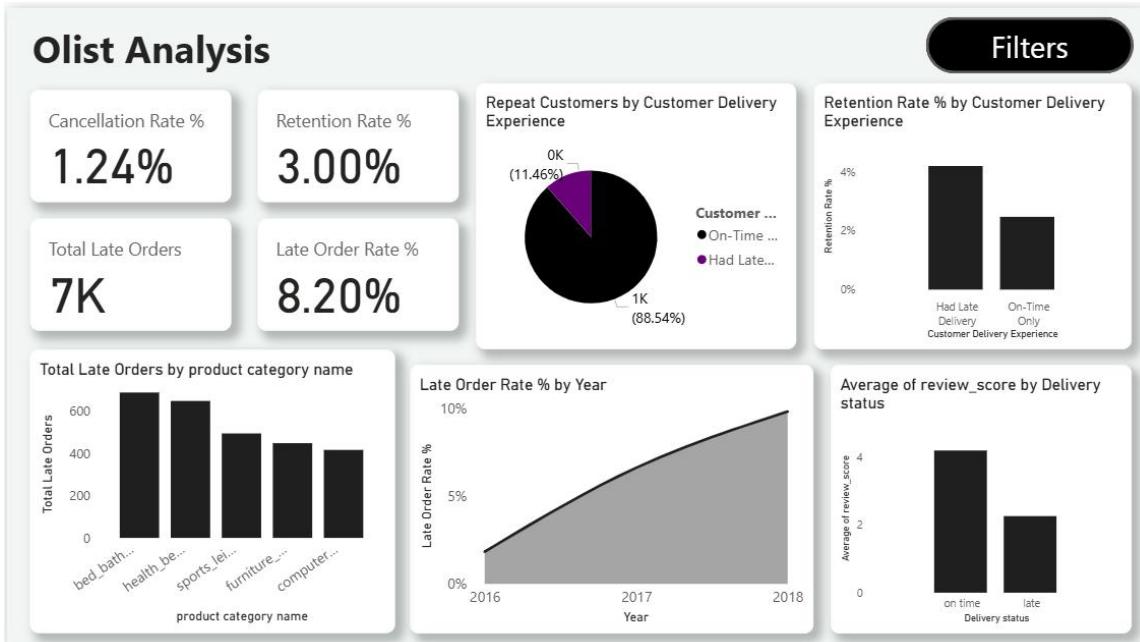
- **Overview Page:** Provides a holistic view of total sales trends, monthly order volume, and high-level KPIs to track the marketplace's growth trajectory.



- **Payment & Completion Behavior:** Analyzes the relationship between payment methods and order success rates, ensuring system stability and identifying any payment-related bottlenecks.

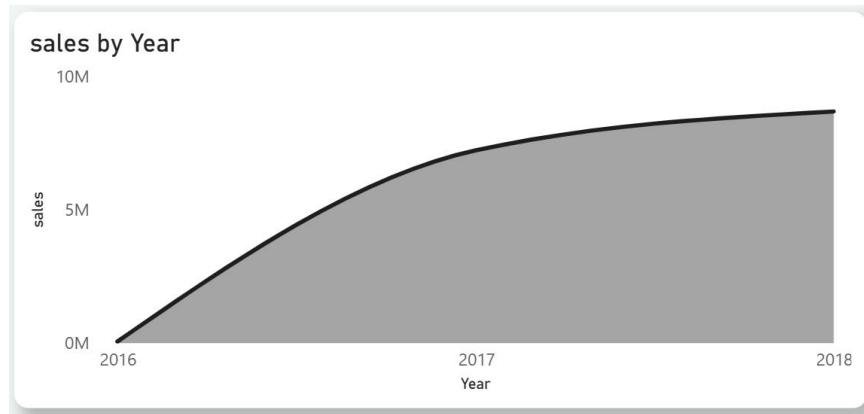


- **Delivery Performance & Customer Experience:** Focuses on shipping efficiency and its direct impact on review scores, highlighting geographic delay hotspots and satisfaction gaps.



9. Key Metrics & Insights

1. How many orders and how much revenue are generated each month?



	order_year	order_month	orders_count	monthly_revenue
1	2016	10	265	46566.71
2	2016	12	1	19.62
3	2017	1	750	127545.67
4	2017	2	1653	271298.65
5	2017	3	2546	414369.39
6	2017	4	2303	390952.18
7	2017	5	3546	567066.73
8	2017	6	3135	490225.6
9	2017	7	3872	566403.93
10	2017	8	4193	646000.61
11	2017	9	4150	701169.99
12	2017	10	4478	751140.27
13	2017	11	7289	1153528.05
14	2017	12	5513	843199.17
15	2018	1	7069	1078606.86
16	2018	2	6555	966510.88
17	2018	3	7003	1120678
18	2018	4	6798	1132933.95
19	2018	5	6749	1128836.69
20	2018	6	6099	1012090.68
21	2018	7	6159	1027903.86
22	2018	8	6351	985414.28

Key Insights:

- **Exponential Revenue & Order Growth:** Sales surged from near-zero in 2016 to over **16M** by 2018 , driven by a massive leap in total order volume reaching approximately **99K** orders.
- **Market scalability:** The continued upward trend reflects high operational scalability and sustainable market expansion throughout the analyzed period.

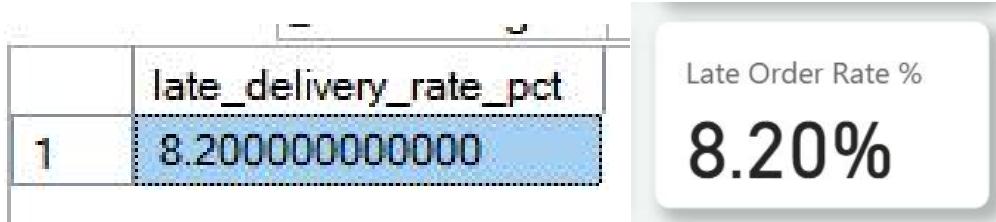
2. Are late deliveries associated with lower customer review scores?



Key Insights:

Delays Hurt Customer Satisfaction: Ratings drop sharply from 4.29

3. What percentage of completed orders are delivered late?



Key Insights:

- **8% Late Rate:** Approximately **8.2%** of all completed orders miss their delivery deadline.
- **Room for Improvement:** While the majority arrive on time, this **8.2%** represents thousands of orders, pointing to a specific bottleneck in the shipping process that needs fixing.

4. Which product categories have the highest delivery delay rates?



	category	late_orders_count	total_late_orders_marketwide	contribution_to_total_delays_pct
1	bed_bath_table	916	8542	10.7200000000000
2	health_beauty	854	8542	10.0000000000000
3	furniture_decor	685	8542	8.0200000000000
4	sports_leisure	625	8542	7.3200000000000
5	computers_accessories	594	8542	6.9500000000000

Key Insights:

- **Most Delayed Categories:** *Bed_Bath_Table* and *Health_Beauty* have the highest number of late deliveries (**916** and **854 orders**), making them the main sources of delivery delays.

- **Highest Business Impact:**

These two categories alone represent **over 20% of all delayed orders**. Improving delivery performance in just these areas would significantly reduce overall delays and customer complaints.

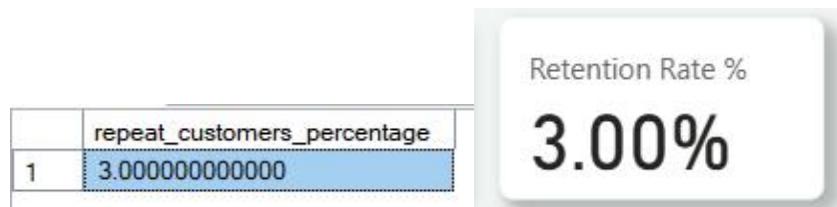
5. Which sellers contribute the most to delivery delays?

	seller_id	total_orders	late_orders	late_rate_pct
1	06a2c3af7b3aee5d69171b0e14f0ee87	370	90	24.32000000000000
2	8160255418d5aaa7dbdc9f4c64ebda44	377	63	16.71000000000000
3	7d13fca15225358621be4086e1eb0964	549	68	12.39000000000000
4	1835b56ce799e6a4dc4eddc053f04066	404	49	12.13000000000000
5	897060da8b9a21f655304d50fd935913	302	36	11.92000000000000
6	4869f7a5dfa277a7dca6462dcf3b52b2	1112	130	11.69000000000000
7	77530e9772f57a62c906e1c21538ab82	305	35	11.48000000000000
8	4a3ca9315b744ce9f8e9374361493884	1756	195	11.10000000000000
9	16090f2ca825584b5a147ab24aa30c86	392	43	10.97000000000000
10	70a12e78e608ac31179aea7f8422044b	310	34	10.97000000000000

Key Insights:

Seller Performance Gap: A clear outlier exists, with the top underperforming seller delaying nearly 1 in 4 orders (24.3%), identifying a critical target for immediate intervention or suspension.

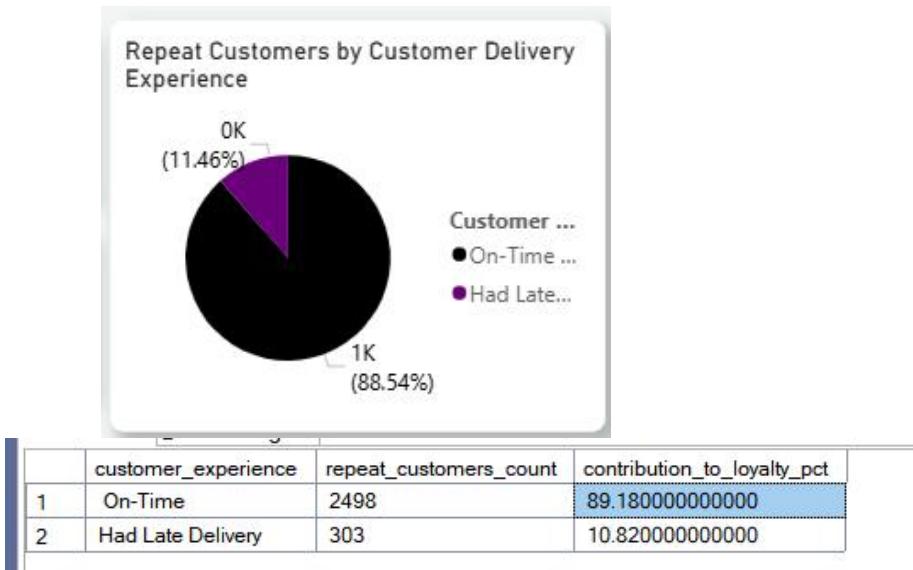
6. What percentage of customers make repeat purchases?



Key Insights:

Low Retention Warning: With only 3% of customers returning for a second purchase, the business relies heavily on one-time transactions, signaling a critical need for loyalty programs to boost Customer Lifetime Value.

7. Do customers who experience late deliveries return to purchase less often?



Key Insights:

The analysis reveals a clear relationship between delivery speed and customer retention.

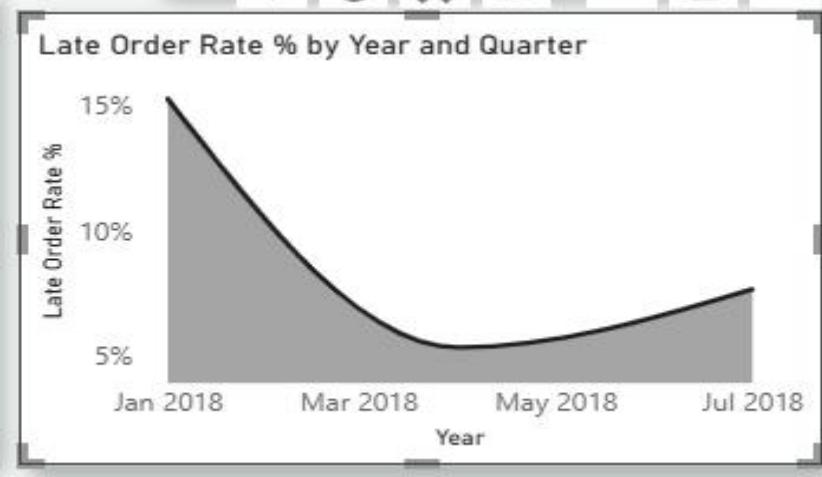
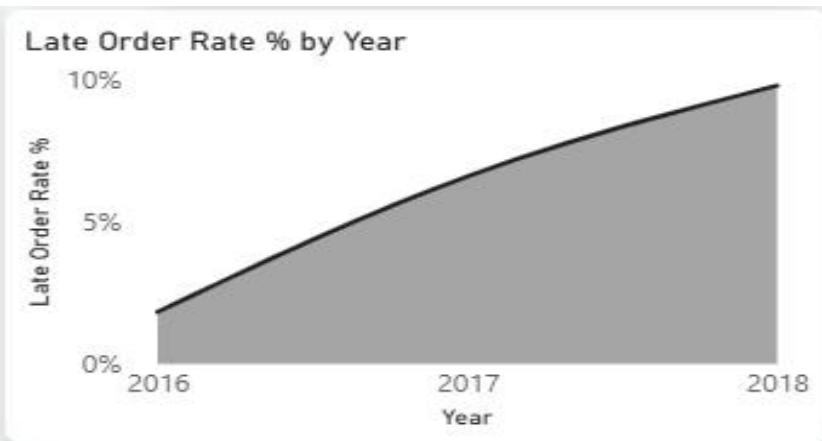
- **Customer Behavior Insight:**

Around **89% of repeat customers** come from orders delivered **on time**, indicating that positive delivery experiences drive loyalty.

- **Business Impact:**

Customers who experience delivery delays account for only **~11% of repeat purchases**, confirming that **logistics performance is a key driver of Customer Lifetime Value (CLV)**.

8. Which months show the highest delivery delay rates?



	order_year	order_month	total_orders	late_orders	late_rate_pct
1	2018	3	7081	1530	21.61000000000000
2	2018	2	6501	1008	15.51000000000000
3	2017	11	7150	1010	14.13000000000000
4	2018	8	6389	681	10.66000000000000
5	2017	12	5669	502	8.86000000000000
6	2018	5	6872	562	8.18000000000000
7	2017	4	2254	178	7.90000000000000
8	2018	1	6990	458	6.55000000000000
9	2018	4	6250	355	5.68000000000000
10	2017	3	2552	143	5.60000000000000

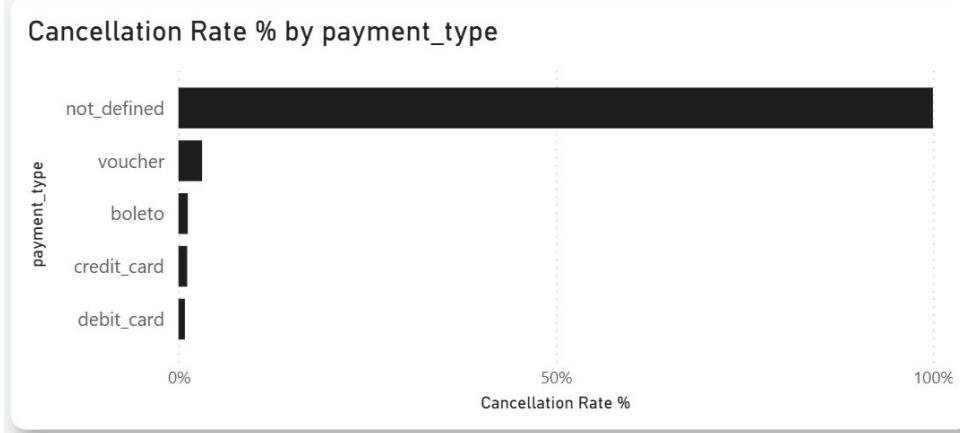
Key Insights:

- Logistics Disruption (Late 2017 – Early 2018): A major disruption impacted the supply chain during this period, peaking in March 2018 with a 21.6% delivery delay rate.

- **System Overload:** The spike began in **Nov 2017 (14.1%)**, likely triggered by Black Friday volume, but the failure to recover for months suggests a deeper external issue or carrier collapse.

9. Are non-completion or cancellation rates associated with payment method?

	payment_type	total_orders	non_completed_orders	cancellation_rate
1	not_defined	3	3	100.000000000000
2	voucher	3866	162	4.190000000000
3	boleto	19784	245	1.240000000000
4	credit_card	76505	890	1.160000000000
5	debit_card	1528	13	0.850000000000



Key Insights:

- **System Stability:** There is no significant correlation between payment methods and order failure. Major methods like **Credit Cards** and **Boleto** show high reliability with cancellation rates staying below **1.5%**.
- **Data Artifacts:** The **100%** failure rate for "not_defined" is a statistical outlier driven by negligible volume (only 3 orders), likely representing system glitches rather than actual customer behavior.

10. Are delivery delays concentrated in specific customer cities?

	customer_city	total_orders	late_orders	late_rate_pct
1	sao goncalo	385	96	24.94000000000000
2	sao luis	329	75	22.80000000000000
3	fortaleza	609	111	18.23000000000000
4	salvador	1169	204	17.45000000000000
5	vila velha	325	50	15.38000000000000
6	nova iguacu	415	57	13.73000000000000
7	recife	584	79	13.53000000000000
8	florianopolis	547	71	12.98000000000000
9	niteroi	812	103	12.68000000000000
10	campo grande	311	38	12.22000000000000

Key Insights:

- **Critical Geographic Failures:** Logistics performance collapses in specific regions. **Sao Goncalo** is the worst-performing city with a massive **24.9%** delay rate, meaning 1 in every 4 customers there receives their order late.
- **High-Volume Risk:** **Salvador** represents a major operational threat; despite having high demand (1,169 orders), it suffers from a **17.5%** delay rate, generating the highest absolute number of dissatisfied customers (204) in this list.

10. Impact & Next Steps

Based on the data-driven insights gathered throughout this analysis, the following strategic actions are recommended to drive revenue growth and improve customer satisfaction:

- **Launch a Targeted Customer Loyalty Program:**
 - **The Issue:** Data indicates a critical retention challenge, with only **3%** of customers returning for a second purchase.
 - **The Action:** Implement a tiered loyalty program or "Prime-like" subscription that rewards frequent buyers. Since 89% of returning customers are those who experienced on-time delivery, focusing retention efforts on this satisfied segment will yield the highest ROI.
- **Strengthen Logistics Resilience against External Shocks:**
 - **The Issue:** The historical analysis revealed a massive spike in delays during **Late 2017 – Early 2018** (reaching 21.6%), indicating the supply chain's vulnerability to external disruptions or seasonal peaks.
 - **The Action:** Diversify logistics partners to avoid "single-point-of-failure" risks. Developing a predictive model to forecast demand surges (like Black Friday) will

- allow for proactive capacity planning, ensuring the network remains robust ("immune") against future volatility.
- **Enforce Stricter Seller Performance Standards:**
 - **The Issue:** A significant performance gap exists, with specific sellers delaying nearly **1 in 4 orders** (24%).
 - **The Action:** Establish a strict Service Level Agreement (SLA) for sellers. Implement a warning system where sellers exceeding a 10% delay rate face temporary suspension or reduced visibility until performance improves.
- **Regional Logistics Optimization (Focus on Northeast):**
 - **The Issue:** High-demand cities like **Salvador** and **Sao Goncalo** are suffering from severe delay rates (17% - 25%).
 - **The Action:** Investigate the root cause in these specific regions. This may require contracting local "last-mile" delivery specialists in these cities or setting up regional warehousing hubs to bypass long-haul shipping bottlenecks.

11. Project Notes

- **Timeframe:** This analysis is based on historical data from **2016 to 2018**. Results reflect the market conditions of that period and may differ from current trends.
- **Root Cause Scope:** While the data shows *where* and *when* delays happened , it doesn't include external factors (like weather or strikes) that explain the exact reason for each delay.