

Olist E-Commerce Analytics

NTU SCTP – Module 2 Project

Team Members:	Wilson Tan	Leslie Kang
	Phyllis Lim	Bradley Soh
	Charania Farida Sadruddin	Muhammad Taufiq Bin Idris

Agenda

- Business Context
- Business Questions
- Data Architecture
- Data Models
- Analysis & Insights
- Recommendations
- Learnings

Business Context

- Olist is a Brazilian e-commerce marketplace
- Connects small sellers to customers
- Generates rich transactional and operational data
- Goal: turn raw data into actionable insights

Business Questions

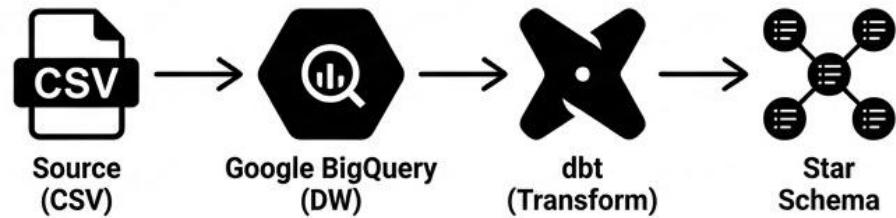
- How is revenue trending over time?
- Which product categories drive the most revenue?
- Which categories show the strongest growth potential?
- Insights on key categories for marketing efforts.
- How does delivery timeliness affect customer satisfaction and long-term customer value?



Data Architecture & Data Models

Data Architecture

- Source: Olist CSV datasets
- Data Warehouse: Google BigQuery
- Transformation: dbt (staging + marts)
- Analytics-ready star schema



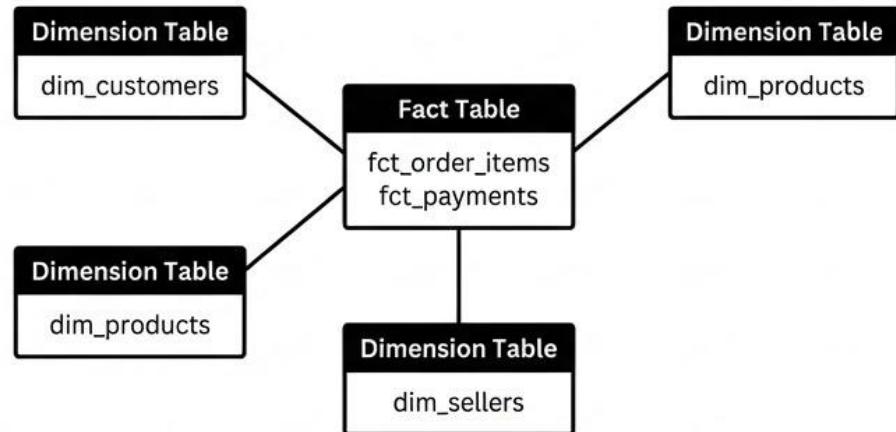
Data Model (Star Schema)

Fact Tables:

- fct_order_items
- fct_payments

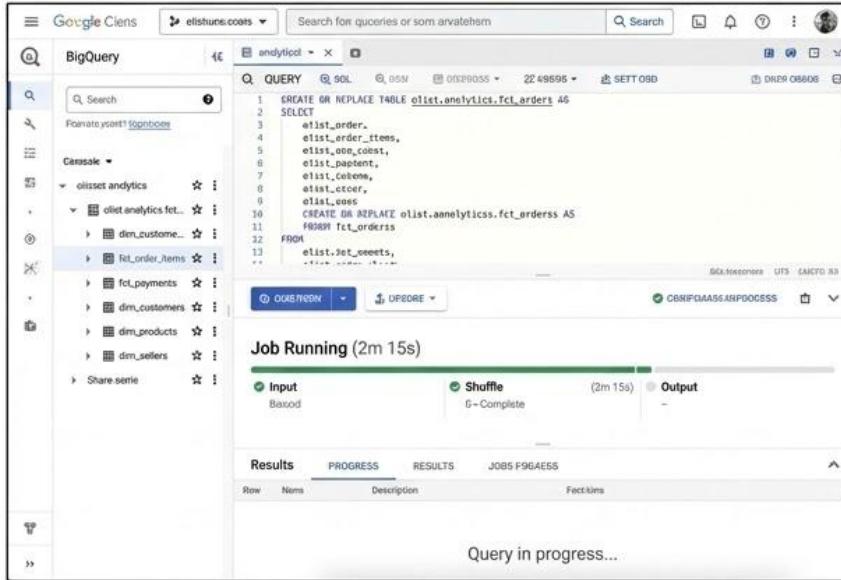
Dimension Tables:

- dim_customers
- dim_products
- dim_sellers



Data Architecture: Tools in Action

Google BigQuery (Data Warehouse)

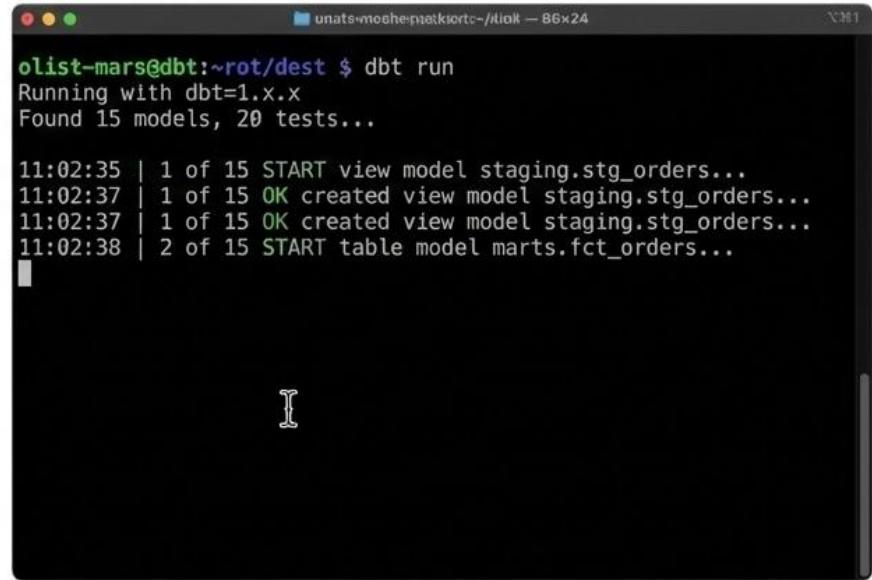


The screenshot shows the Google BigQuery web interface. On the left, the sidebar lists datasets like 'Console', 'oilstores', and 'oilstores analytics'. A query editor window is open with the following SQL code:

```
CREATE OR REPLACE TABLE olist.analytics.fct_orders AS
SELECT
    elist_order,
    elist_order_items,
    elist_order_consts,
    elist_order_pagent,
    elist_order_cetene,
    elist_order_ctcer,
    elist_order_sess,
    elist_order_seets,
    ...
FROM
    fct_order_items
JOIN
    dim_customers
JOIN
    dim_products
JOIN
    dim_sellers
ON
    fct_order_items.customer_id = dim_customers.customer_id
    AND fct_order_items.product_id = dim_products.product_id
    AND fct_order_items.seller_id = dim_sellers.seller_id;
```

The status bar at the bottom indicates 'Job Running (2m 15s)'. Below the editor, a results table shows columns: Input, Shuffle, and Output. The results section displays the message 'Query in progress...'. The top navigation bar includes tabs for 'QUERY', 'SQL', 'DSN', 'OSS', 'SET OSD', and 'DATA OSD'.

dbt (Transformation)



The terminal window shows the command `olist-mars@dbt:~rot/dest $ dbt run` being executed. The output shows the following log entries:

```
Running with dbt=1.x.x
Found 15 models, 20 tests...
11:02:35 | 1 of 15 START view model staging.stg_orders...
11:02:37 | 1 of 15 OK created view model staging.stg_orders...
11:02:37 | 1 of 15 OK created view model staging.stg_orders...
11:02:38 | 2 of 15 START table model marts.fct_orders...
```

Executing transformation query and loading data into analytic tables.

Running dbt models to build staging and marts layers.

ELT Pipeline

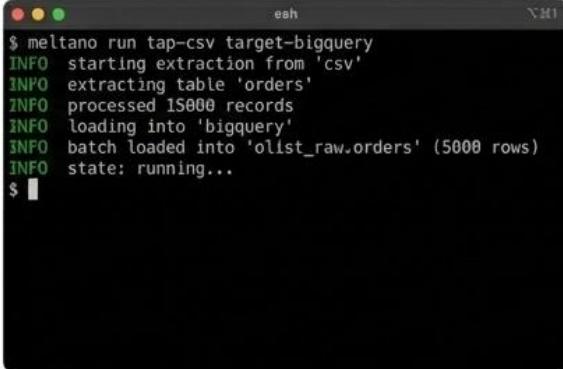


Extract: Raw CSVs **Extract & Load:** Stored in **Transform:** dbt staging **Output:** Analytics-ready
 Meltano olist_raw dataset + marts models tables in olist_analytics

- **Extract & Load:** Meltano extracts Raw CSVs and loads them
- **Stored:** Data is stored in olist_raw dataset in BigQuery
- **Transform:** dbt staging + marts models
- **Output:** Analytics-ready tables in olist_analytics

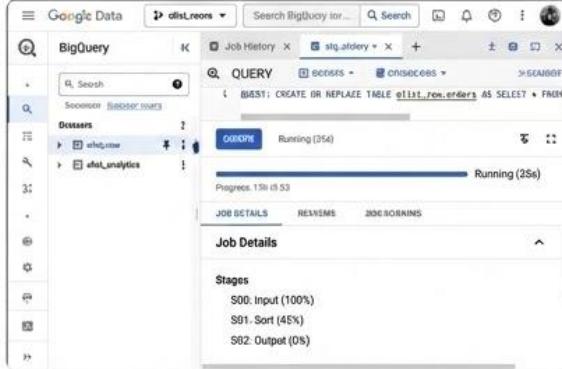
ELT Pipeline: Tools in Action

Meltano: Extract & Load

A terminal window titled 'esh' showing the execution of a Meltano command. The output indicates the start of an extraction from a CSV file, processing 15000 records, loading them into BigQuery, and batch loading into 'olist_raw.orders' (5000 rows). The state is shown as 'running...'.

```
$ meltano run tap-csv target-bigquery
INFO  starting extraction from 'csv'
INFO  extracting table 'orders'
INFO  processed 15000 records
INFO  loading into 'bigquery'
INFO  batch loaded into 'olist_raw.orders' (5000 rows)
INFO  state: running...
$
```

BigQuery: Storage & Processing

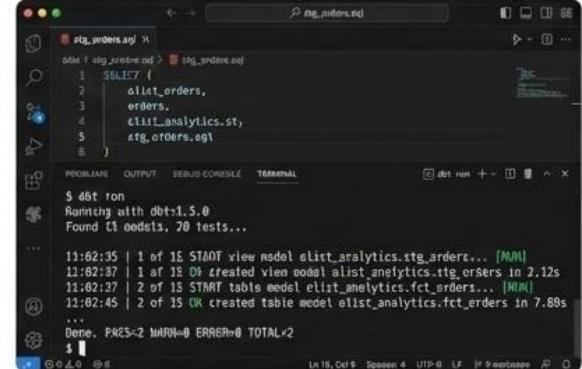
The BigQuery web interface showing the 'olist_raw' dataset. A job named 'sqldfairy' is running, processing a query to create or replace a table named 'elit_orders'. The job status is 'Running (D56)' with a progress bar at 15% and a duration of 25 seconds. The job details show stages: S00: Input (100%), S01: Sort (45%), and S02: Output (0%).

```
CREATE OR REPLACE TABLE `elit_orders` AS SELECT * FROM `olist_raw.orders`
```

Terminal showing Meltano extracting raw CSV data and loading it into BigQuery in batches.

BigQuery interface showing the 'olist_raw' dataset and a running job processing incoming data.

dbt: Transformation

A terminal window titled 'alg_yoders.01' showing the execution of a dbt run command. The output shows the successful creation of a view model, a staging table, and a fact table, along with the creation of a final analytics table. The total duration of the run is 7.89 seconds.

```
$ dbt run
Running with dbt:1.5.0
Found 11 models, 20 tests...
...
11:02:35 | 1 of 15 START view model elit_analytics.ttg_order...
11:02:37 | 1 of 15 OK created view model elit_analytics.ttg_orders...
11:02:37 | 2 of 15 START table model elit_analytics.fct_orders...
11:02:45 | 2 of 15 OK created table model elit_analytics.fct_orders in 7.89s
...
Done. PREC=2 MARN=0 ERRORS=0 TOTAL=2
$
```

dbt terminal output showing the successful run of staging and marts models to create analytics-ready tables.

Data Quality & Challenges

- Inconsistent data types in raw CSVs
- Non-numeric values in numeric fields
- Resolved using SAFE_CAST and staging filters
- Ensured reliable downstream analytics



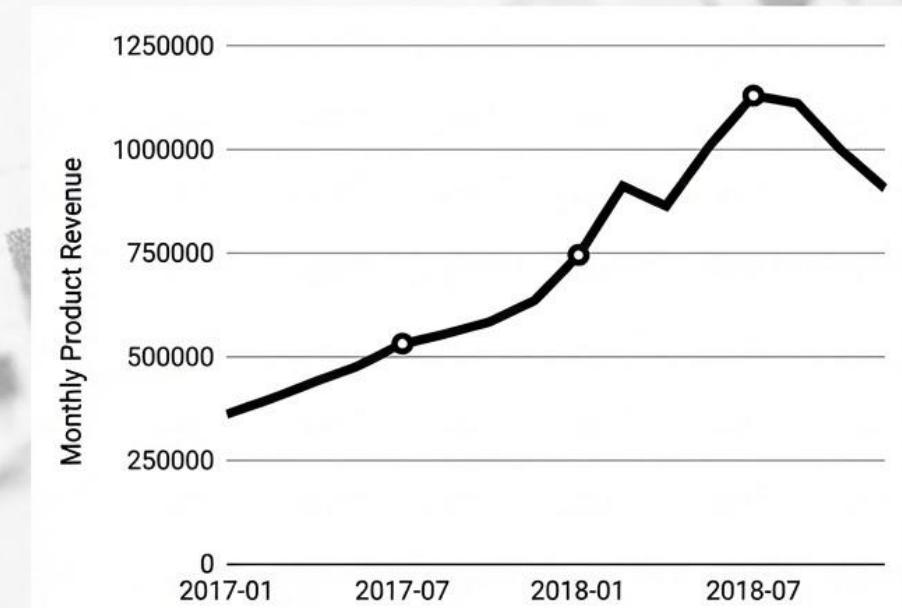
Analysis, Insights & Recommendations:

- How is revenue trending over time?
- Which product categories drive the most revenue?
- Which categories show the strongest growth potential?
- How does delivery timeliness affect customer satisfaction and long-term customer value?



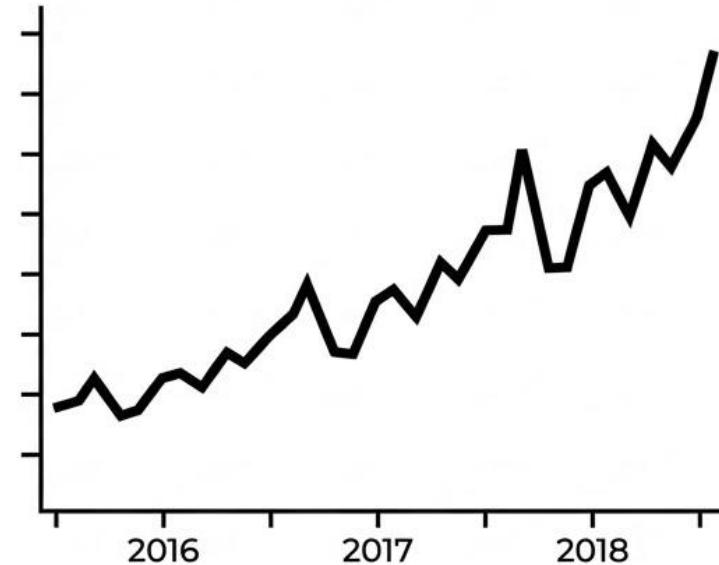
Overall Revenue Trend (2016–2018)

- Revenue increases steadily from 2016 through 2017.
- 2018 (Jan–Sep) already exceeds 2017 full-year revenue (note partial year coverage).
- Indicates strong growth momentum.



Order Volume Trend

- Order volumes broadly follow the revenue trend.
- Clear seasonal peaks are visible.
- Supports consistency of growth.



Revenue by Product Category (Top 10)

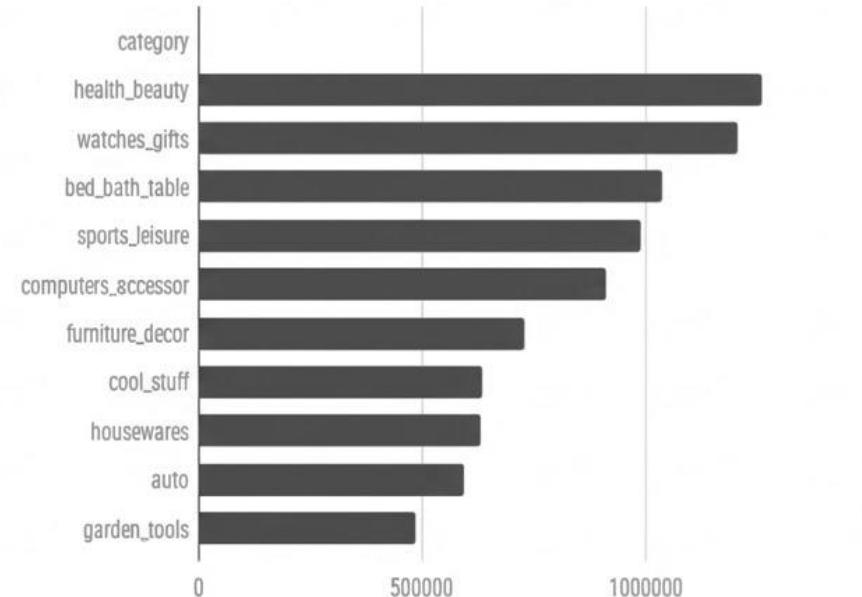
Revenue is concentrated in a small set of categories.

Top categories account for the majority of

Business focus should prioritize these seg



Top 10 Categories by Revenue



Key Business Insights & Implications

- Prioritize top categories for inventory and promotions.
- Plan operations and logistics ahead of peak months.
- Track category growth to guide assortment decisions.



Business Recommendations

1. Capitalize on “Health & Beauty” Momentum:

- **Recommendation:** Allocate the Q4 marketing budget specifically to the **Health & Beauty** category.
- **Why:** It has overtaken other categories to become the #1 revenue driver as of late 2018.



2. Operational Optimization for “Bed, Bath & Table”:

- **Recommendation:** Negotiate bulk shipping rates or dedicate specific warehousing space for this category.
- **Why:** It generates the highest order volume (e.g., 7,000+ orders/year pace). Small improvements in logistics margins here will scale significantly.

3. Prepare for Black Friday Volatility:

- **Recommendation:** Implement "Code Freeze" and inventory checks by October.
- **Why:** Historical data from Nov 2017 shows order volumes spike by ~80% in key categories. The platform must be technically ready to handle this load in Nov 2018.

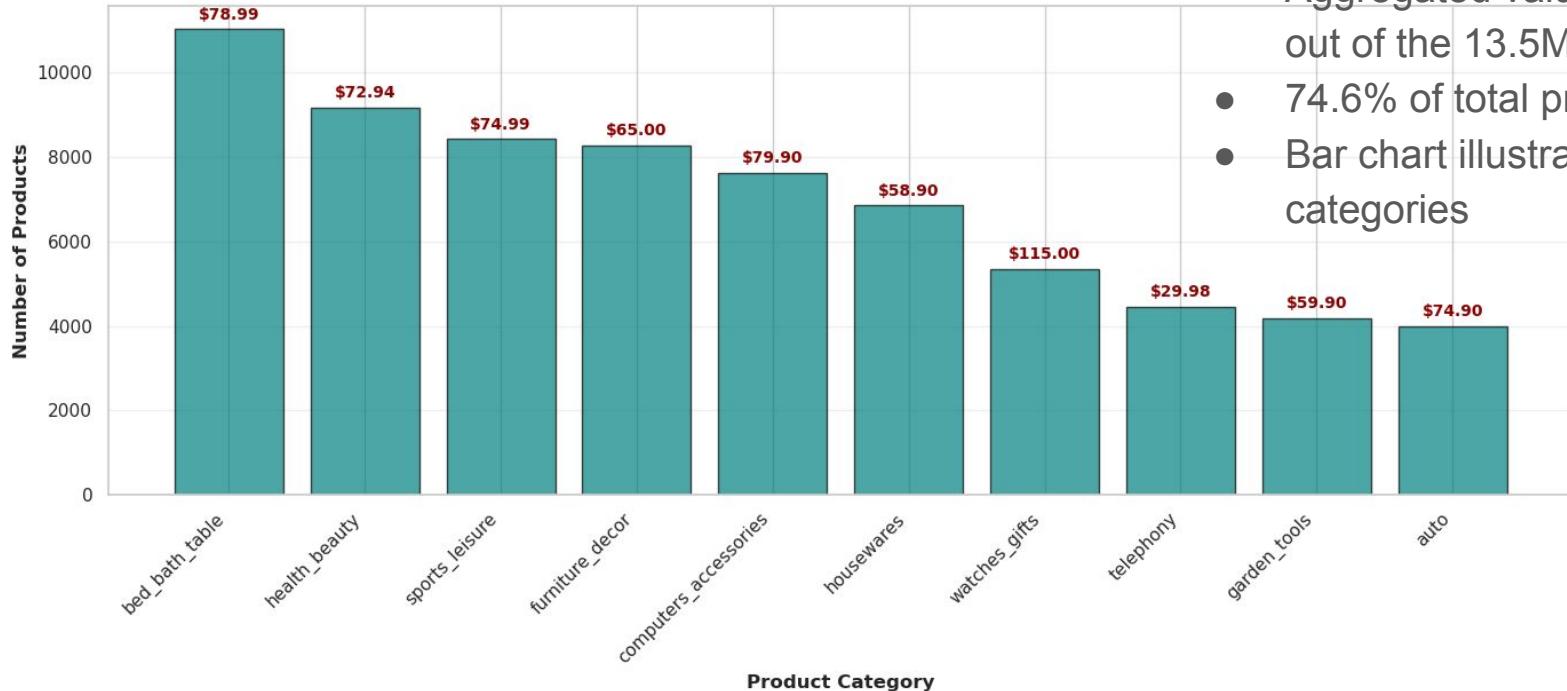
Analysis, Insights & Recommendations:

Insights on key categories for marketing efforts.



Insights on key categories for marketing efforts.

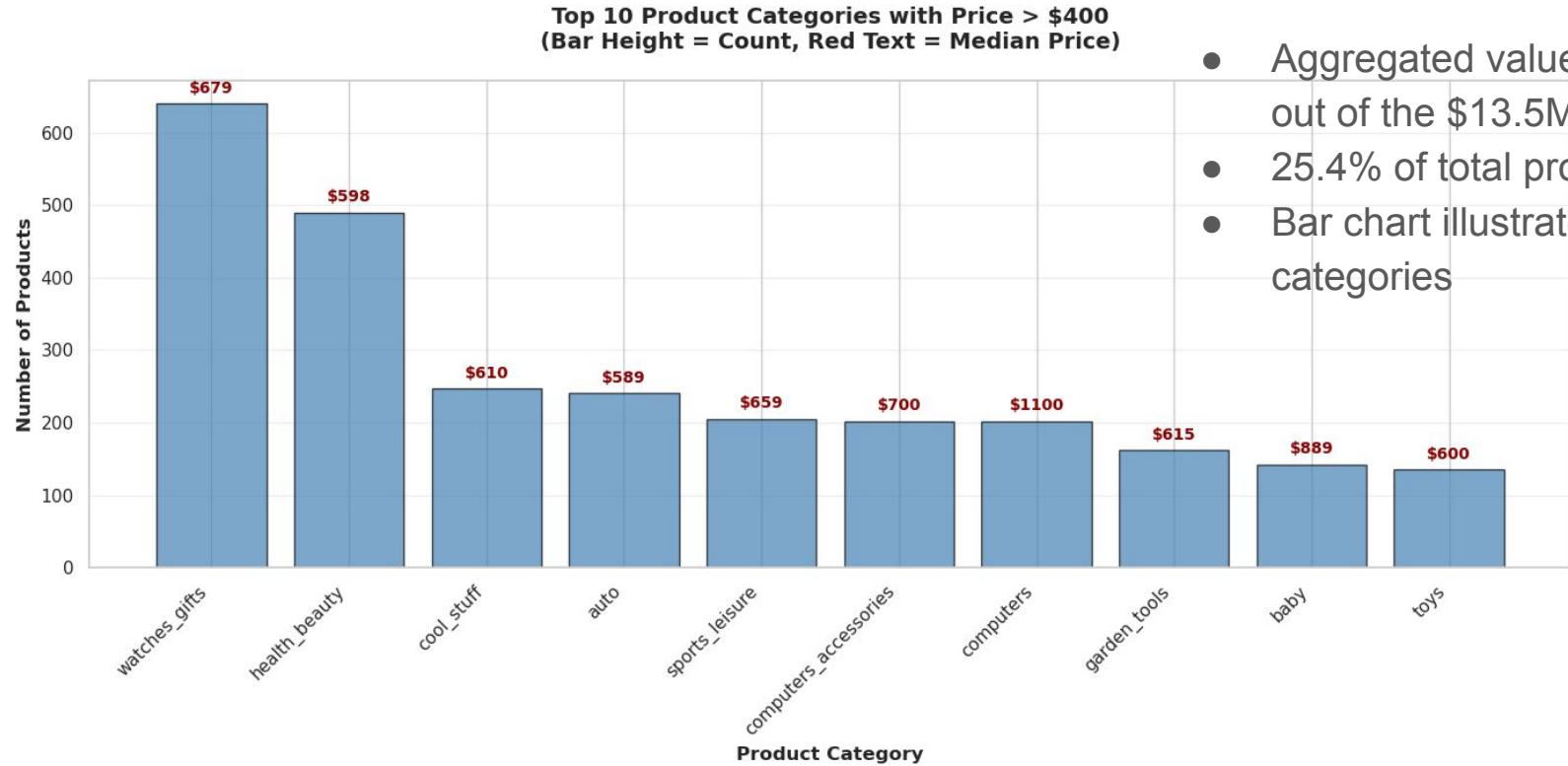
**Top 10 Product Categories with Price \leq \$400
(Bar Height = Count, Red Text = Median Price)**



Products \leq \$400/pc

- Aggregated value of 10.1M out of the 13.5M total value
- 74.6% of total product value
- Bar chart illustrates top 10 categories

Insights on key categories for marketing efforts.



Products > \$400/pc

- Aggregated value of \$3.4M out of the \$13.5M total value
- 25.4% of total product value
- Bar chart illustrates top 10 categories

Insights on key categories for marketing efforts.

Strategic Focus on Key Categories



Categories	Products <= \$400/pc	Products > \$400/pc
Watches Gifts	\$694,285	\$510,721
Health & Beauty	\$904,672	\$354,009
Sports & Leisure	\$809,061	\$178,988
Computer Accessories	\$733,126	\$178,828
Automotive items	\$402,001	\$190,710
Computers	-	\$222,929
Bed, Bath & Table	\$984,930	-

\$6.1M. 45% of all sales value!

Insights on key categories for marketing efforts.

“Now that we have identified the strategic product categories, WHEN and WHERE should we target the marketing efforts?”

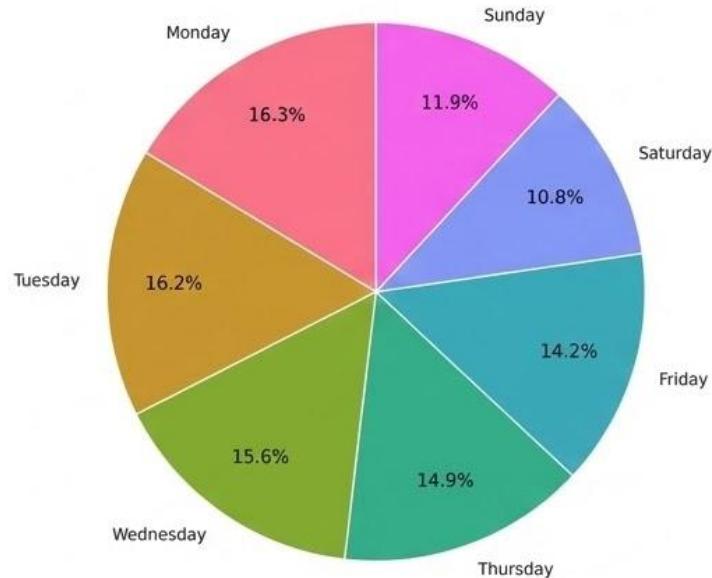


Insights on key categories for marketing efforts.

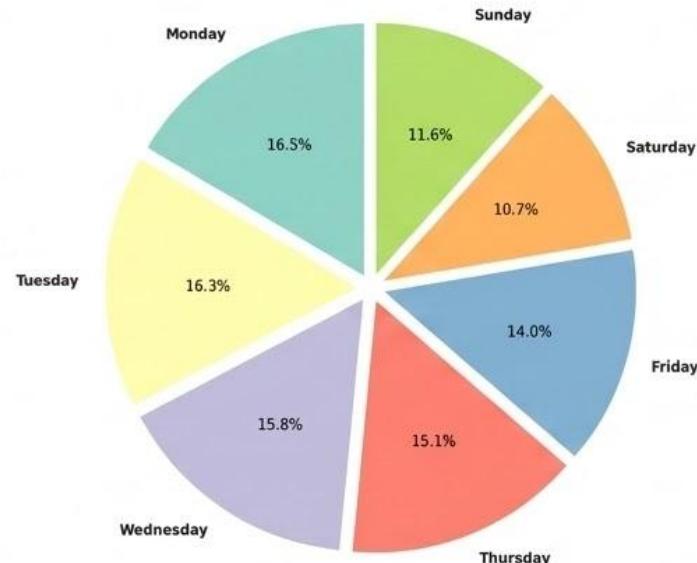
WHEN? Day of Week

On the contrary, weekends are NOT prime!

Order Volume by Day of Week

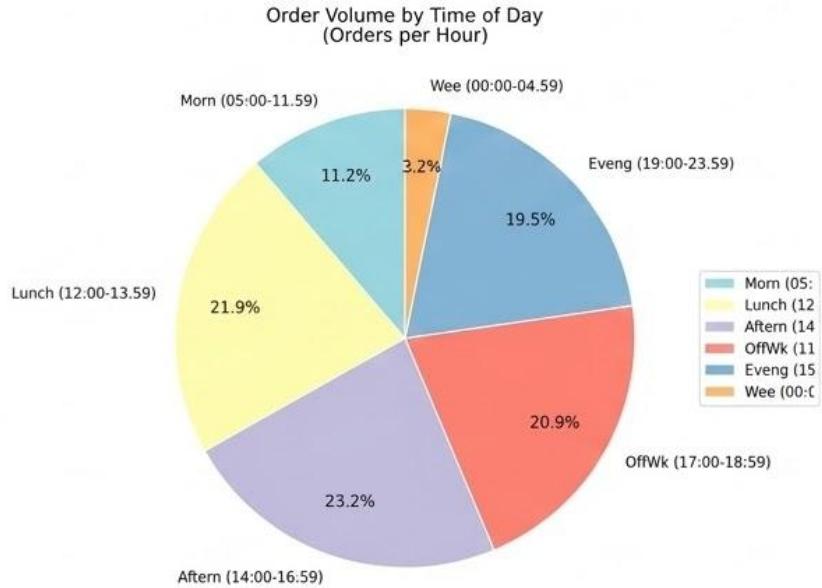


Distribution of Purchases by Day of Week
(Niche Categories)

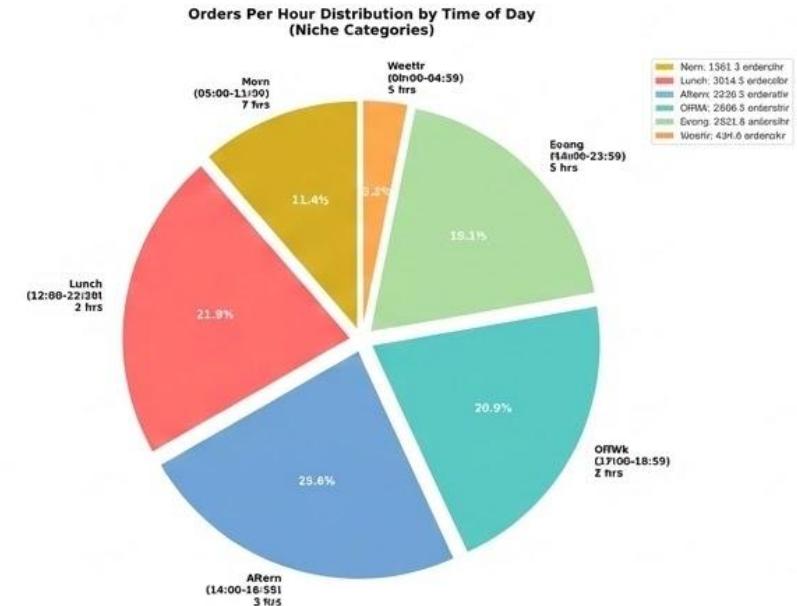


Insights on key categories for marketing efforts.

WHEN? Time of day



Lunch, Work-hours and Commuting hours are PRIME!



Insights on key categories for marketing efforts.

WHERE?



Insights on key categories for marketing efforts. Recommendations in a nutshell...



Watches, Gifts, Health & Beauty, Sports & Leisure, Computer Accessories, Automotive items, Computers, Bed, Bath & Table



Future Data EXPLORATIONS

E.g. width of product range, customer habits, time-based trending of sales (slice by time-periods) etc.



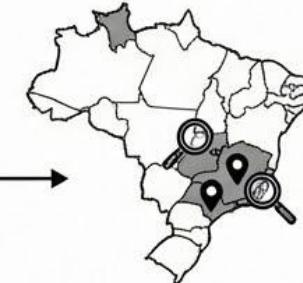
Prioritise Delivery Improvements

In high-delay states (RR, SE, AP) to protect customer satisfaction and prevent future churn.



Strategic Time Windows

Strategic marketing efforts on WEEKDAYS, between 11am to 5pm or 7pm



Markets for FOCUSED EFFORT

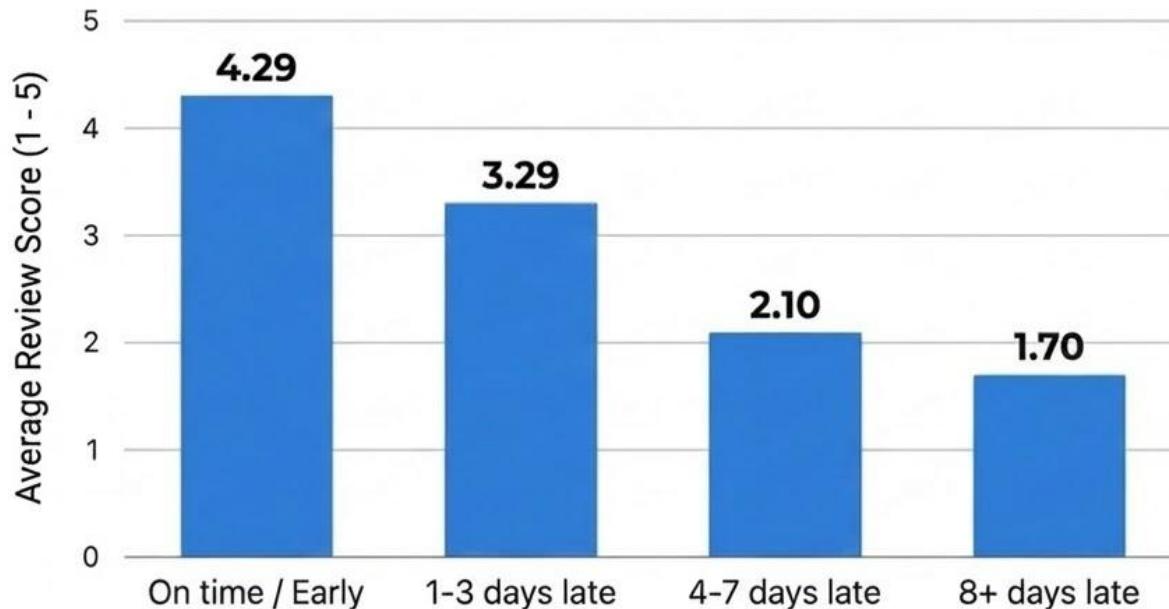
Sao Paulo, Rio De Janerio, Belo Horizonte & Brasilia

Analysis, Insights & Recomendations:

“How does delivery timeliness affect customer satisfaction and long-term customer value?”



Customer Satisfaction Drops Sharply With Late Delivery



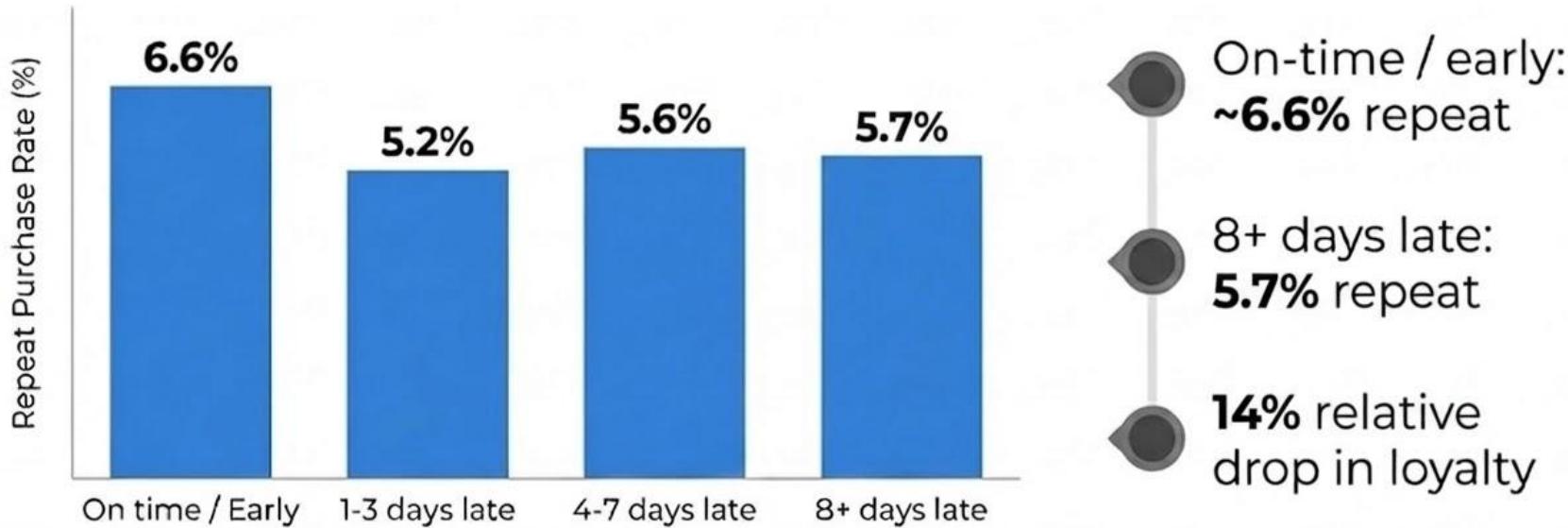
On-time / early deliveries:
~4.3 / 5

8+ days late deliveries:
~1.7 / 5

-60% satisfaction decline

Customer satisfaction collapses once delivery exceeds 4 days late

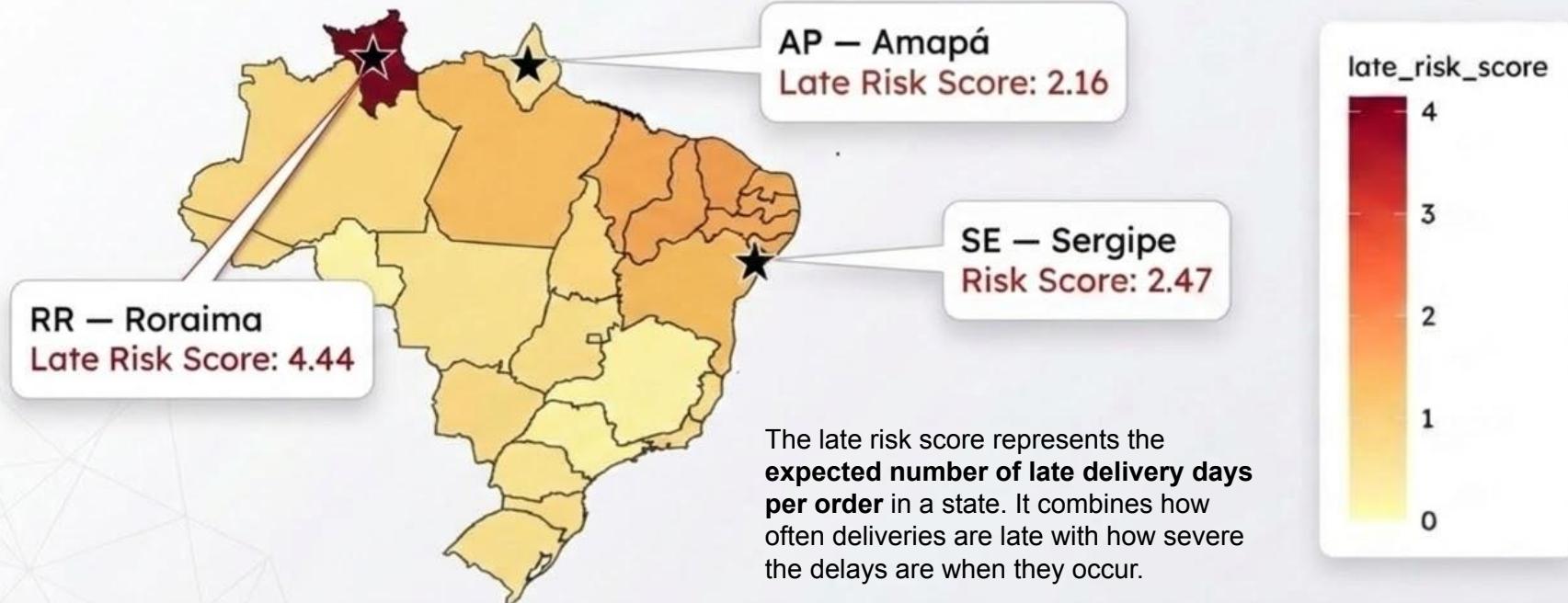
Late Delivery Gradually Erodes Customer Loyalty



Late delivery doesn't cause customers to churn immediately — it quietly reduces lifetime value.

Delivery Performance and Satisfaction Are Uneven Across Regions

Late Delivery Risk by State (Frequency × Severity)



Targeted interventions in 3 states could reduce national late-delivery risk disproportionately

Recommendations

Strategies for Improving Delivery Performance & Customer Experience



1. Intervene Early — Before Delays Exceed 3 Days

Trigger operational action as deliveries approach **3 days late**, not after severe delays occur.



2. Position Delivery Performance as a Customer Experience Driver

Elevate delivery performance into **customer experience monitoring**, not just logistics reporting.



3. Use Review Scores as an Early Warning Signal

Monitor review score declines as a **leading indicator** and intervene before loyalty metrics deteriorate.



4. Focus Targeted Improvements in High-Risk Regions

RR (Roraima), SE (Sergipe), AP (Amapá) show disproportionately high late-delivery risk.

Order Status Distribution with Total Price

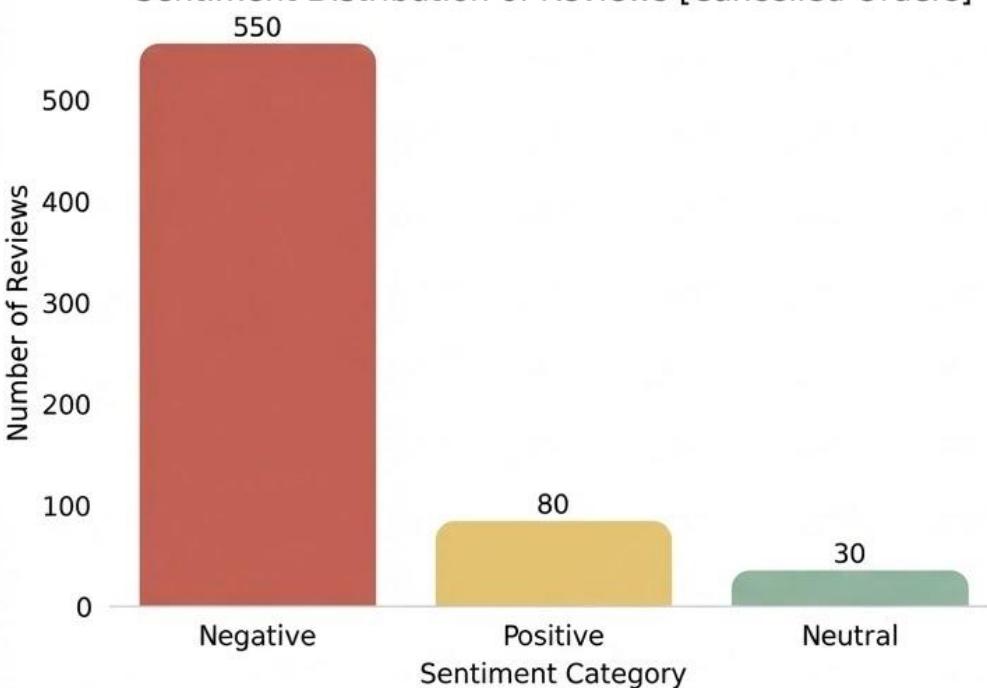
Order Status Distribution with Total Price



Order Status	
delivered:	97.82% - \$13,279,836.59
shipped:	1.06% - \$151,710.65
canceled:	0.48% - \$95,720.47
invoiced:	0.32% - \$61,969.25
processing:	0.32% - \$60,469.22
unavailable:	0.01% - \$2,007.69
approved:	0.0% - \$209.60

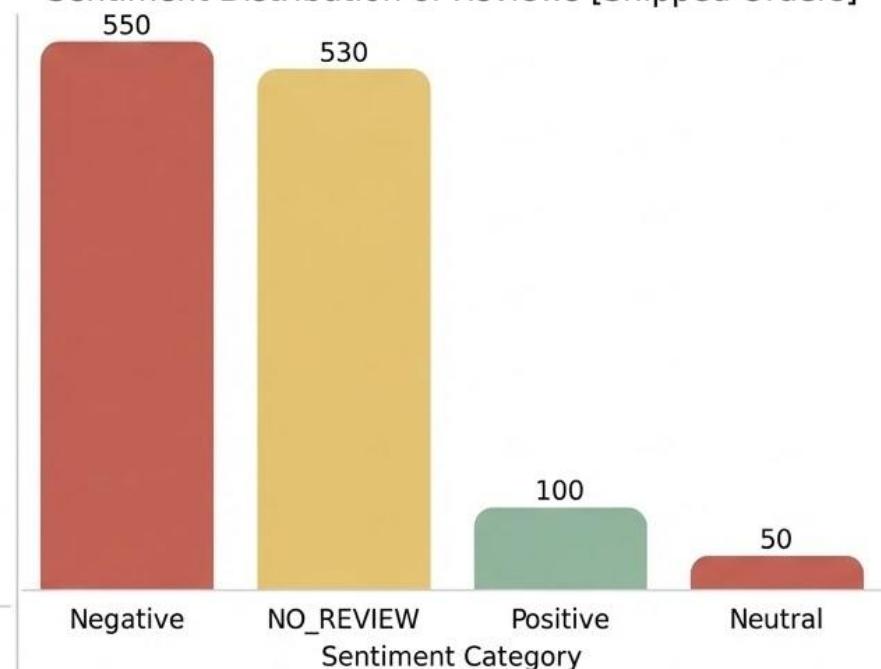
Sentiment Distribution of Reviews [Cancelled Orders]

Sentiment Distribution of Reviews [Cancelled Orders]



Sentiment Distribution of Reviews [Shipped Orders]

Sentiment Distribution of Reviews [Shipped Orders]



Courier issues? Shipping Issues? Store fulfillment issues?

Sentiment Distribution Summary

Negative:	307
NO REVIEW:	196
Positive:	30
Neutral:	13

Key Negative Review Examples (1 Star)

-  They didn't meet the delivery deadline!
-  The company did not deliver my product. They waited for...
-  I still haven't received the product, can I cancel the order?
-  The store was irresponsible because it didn't issue the invoice...

Courier issues? Shipping Issues? Store fulfilment issues?

Raw Review Data (Code Snippet)

```
review_comment_message
0 fazem 2 meses a espera do produto, ja entrei em contato com a baratheon
1 baratheon meu produto não chegou veja aí que
2 Infelizmente o produto ainda não chegou. Hoje
3 Não entregaram no prazo, já tá com mais de 30 dias
4 Fiz a solicitação de cancelamento na mesma
```

sentiment_category

```
0 Negative
1 Negative
2 Negative
3 Negative
4 Negative
```

sentiment_category

```
Negative      554
```

```
NO REVIEW     527
```

```
Positive      82
```

```
Neutral       33
```

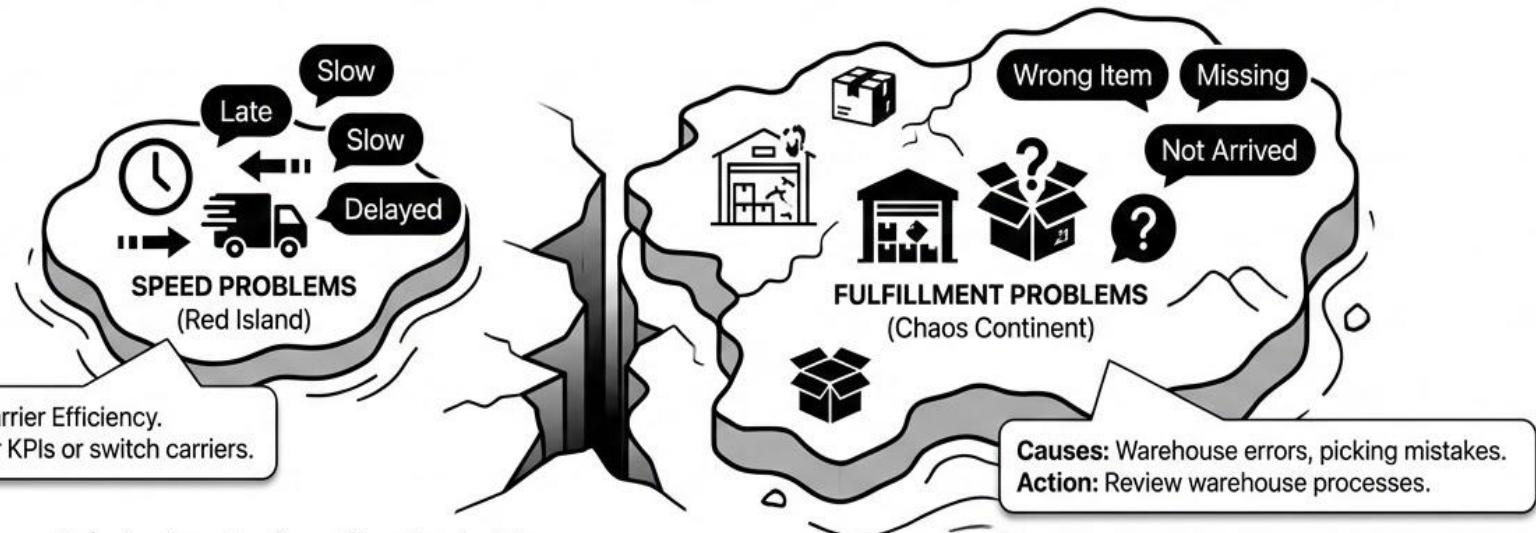
```
Name: count, dtype: int64
```

Key Negative Review Examples (1 Star)

	review_comment_message	sentiment
0	I've been waiting for the product for 2 months, I've already logged in and...	1 star
1	Baratheon, my product hasn't arrived, please check it out...	1 star
2	Unfortunately, the product hasn't arrived yet. There was...	1 star
3	They didn't deliver on time, it's been over 30 days already.	1 star
4	I submitted the cancellation request at the same time...	1 star

Qualitative/Insight Analyst based on Hybrid RAG

Analysis between Reviews - Orders - Customers - Products | Topic: Delivery Complaint Analysis: Semantic Map Insights



Delivery Complaints Analysis – Key Insights

- Two Distinct Delivery Issues (Red Island vs. Chaos Continent)
- Red Island:** Customers complain about 'slow'/late' deliveries; purely carrier efficiency
- Chaos Continent:** Complaints about wrong/missing items; warehouse & inventory issues

RAG Insight

- Semantic mapping separates complaints by root cause
- Warehouse issues are more complex and high-priority than speed issues

Reasoning and Fact check

»

Olist Smart Data Assistant

Model: gemini-2.0-flash-001 Status: Online (Gemini 2.0 Flash) Region: us-central1 Deploy :

RAG Chat Assistant Dynamic Semantic Galaxy

Why do people complain about the delivery?

Based on the provided context, customers complain about delivery for the following reasons:

- Non-delivery of products:** Some customers report not receiving their products at all. For example, a customer for the "alimentos" (food) category said, "Como devo proceder?" (My product was not delivered. How should I proceed?). Another customer also stated "Não recebi o produto entregue." (I did not receive the product, but the website says it was delivered).
- Excessive delays:** Many customers mention significant delays in the delivery timeframe. Examples include: "Um prazo enorme e não entregaram." (A huge deadline and they didn't deliver); "Muita demora para a entrega do produto" (Too long for the delivery of the product); "Um absurdo esse atraso" (This delay is absurd).
- Lack of communication about delays:** Customers are frustrated by the absence of updates or feedback from the company regarding delivery status. One customer mentioned, "A ENTREGA ESTAVA MARCADA PARA O DIA 27/03, E NÃO HOUVE NENHUM TIPO DE FEEDBACK POR PARTE EMPRESA SOBRE O ATRASO." (THE DELIVERY WAS SCHEDULED FOR 03/27, AND THERE WAS NO FEEDBACK FROM THE COMPANY ABOUT THE DELAY.). Another customer stated "Não cumpriram com a entrega." (They did not meet the delivery deadline and did not notify the delay.).
- Problems during transport:** One customer mentioned a problem during transport that caused a significant delay. The customer in "cordo" said, "Houve problema no transporte e em consequência disso um grande atraso. Mas foi a primeira vez que ocorreu. Os produtos sempre são entregues dentro do prazo." (There was a problem in transport and as a consequence a great delay. But it was the first time it happened. The products are always delivered before the deadline).
- Having to pick up the product:** One customer complained about having to go to the post office to pick up the product, which generated dissatisfaction. The customer in "franca" said "Me deslocar para retirar o produto nos correios . Não consegui rastrear o produto." (I had to travel to the post office to pick up the product. I couldn't track the product).

list down top 5 most selling items over all time

Based on the provided data, here are the top 5 best-selling product categories of all time:

1. cama_mesa_banho (Bed, Table & Bath)
2. beleza_saude (Health & Beauty)
3. esporte_lazer (Sports & Leisure)
4. informatica_acessorios (Computers & Accessories)
5. moveis_decoracao (Furniture & Decor)

Generated SQL

```
SELECT
    product_category_name,
    COUNT(*) AS total_orders
FROM
    my-project-sctp-module-2.olist_dbt_dataset.init_search_unioned
GROUP BY
    product_category_name
ORDER BY
    total_orders DESC
LIMIT 5
```

Data Result

product_category_name	total_orders
cama_mesa_banho	1234567890
beleza_saude	9876543210
esporte_lazer	8765432109
informatica_acessorios	7654321098
moveis_decoracao	6543210987



Semantic Galaxy & VoC Analytics

AI-powered visualization. Filter the map by typing a context below.

Filter Map by Context

Why do people complain about the delivery?

Topics

5

View



2D



3D

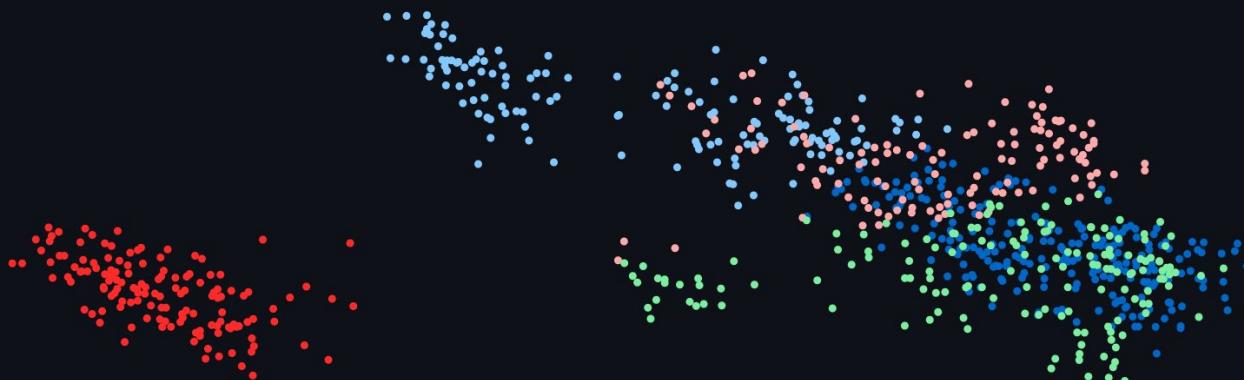
Color

cluster_label



2D Semantic Map for 'Why do people complain about the delivery?'

- cluster_label
- Late delivery/quality
 - Product not received
 - Delivery delay/issues
 - Delivery issues/leakage
 - Shipping/packaging problems



Key Learnings



Importance of clean
staging models



Value of star schema
design



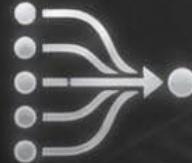
Real-world data requires
defensive engineering



dbt enables scalable
analytics workflows



Understanding raw data
should not be undermined
e.g. customer_id versus
customer_unique_id.



Progressive amalgamation
of data enable checks to
ensure data accuracy

THANK YOU



QUESTIONS?

Let's discuss our approach and findings. →