OLIST BRAZILIAN E-COMMERCE

Presented on 5 June 2025

INTRODUCING THE ALCHEMISTS

DERRICK EE

Aspiring data analyst with background in Information Technology. Contributed to Exploratory Data Analysis.

NUR AIN

Aspiring data analyst with background in customer service and stakeholder engagement. Contributed to the development of seller reliability metrics using Power BI, with a focus on fair scoring and performance insights.

RANDALL VILLASENOR

Aspiring Power BI Data Analyst from a background in Semiconductor Manufacturing. Contributed to Python data cleaning and Power BI dashboards.

KIN MENG

Aspiring data engineer from background in Finance & Accounting. Contributed to ETL processes and Power BI visualisations and analysis.

https://images.app.goo.gl/t4LCanfjPU2ryMAV8

PROBLEM STATEMENT

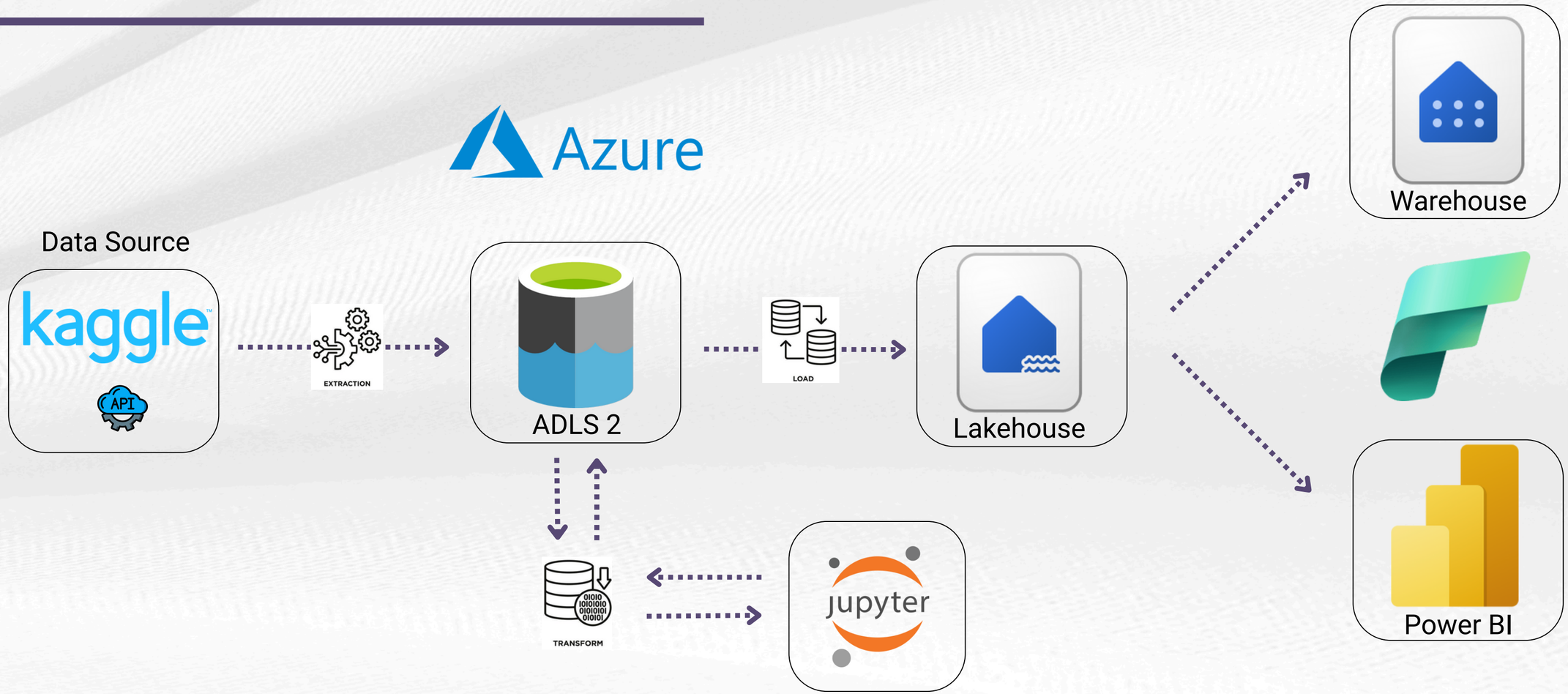
The raw Olist dataset was unstructured and inconsistent, with missing values, duplicates, and mismatched data types.

To make meaningful analysis possible, we had to clean, transform, and merge data from multiple sources into a reliable and queryable format.

OBJECTIVE

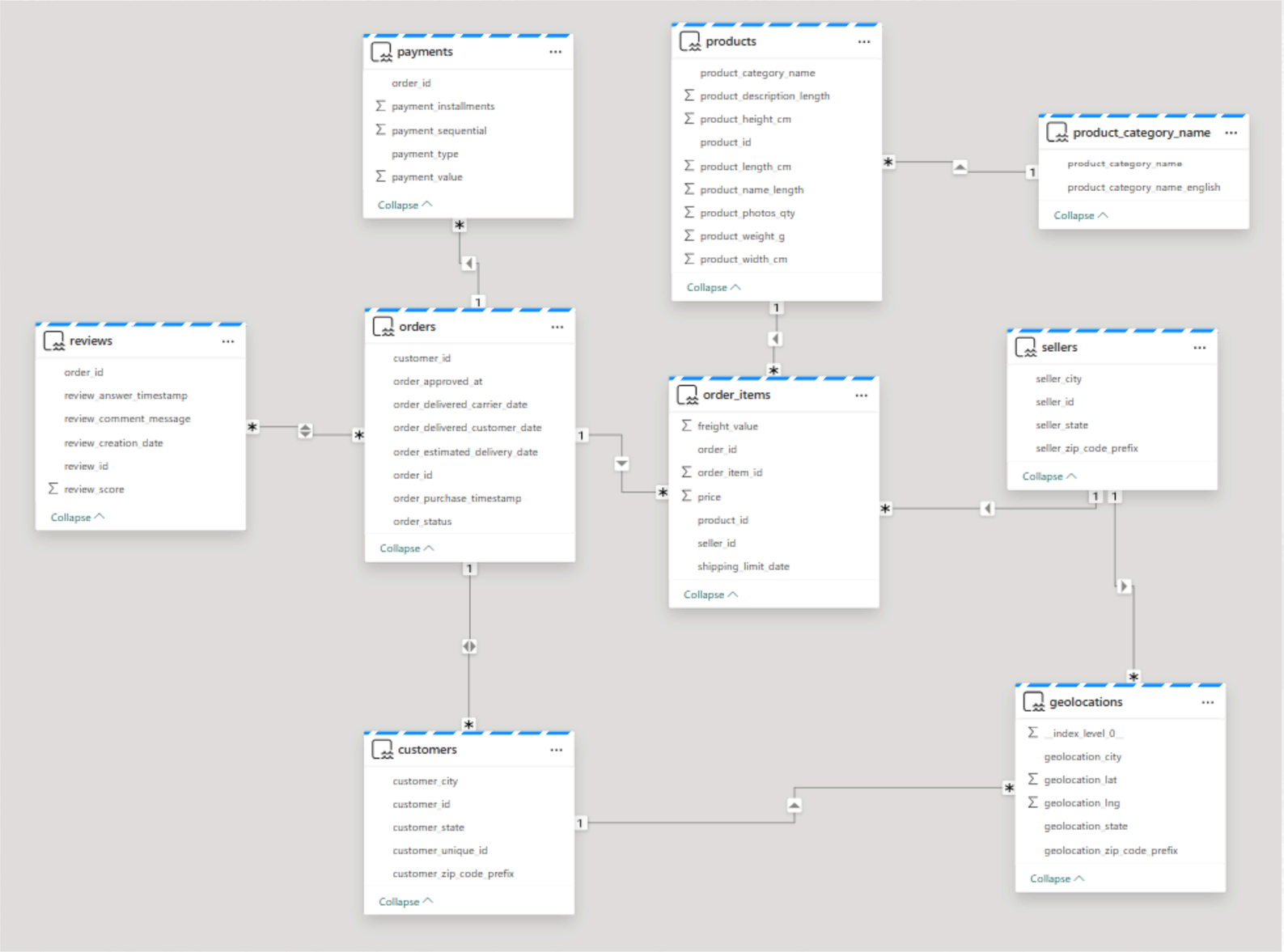
To uncover seller performance patterns and payment behavior insights from Olist's marketplace data — by cleaning and merging multiple sources (orders, reviews, deliveries, payments) into a unified format that powers a business intelligence dashboard.

ARCHITECTURE PIPELINE



DATABASE SCHEMA

Structured blueprint of a database that defines its tables, fields, relationships, and rules for how data is stored and connected.



DATA CLEANING & TRANSFORMATION

We divided the 9 datasets among ourselves and performed key data cleaning tasks:

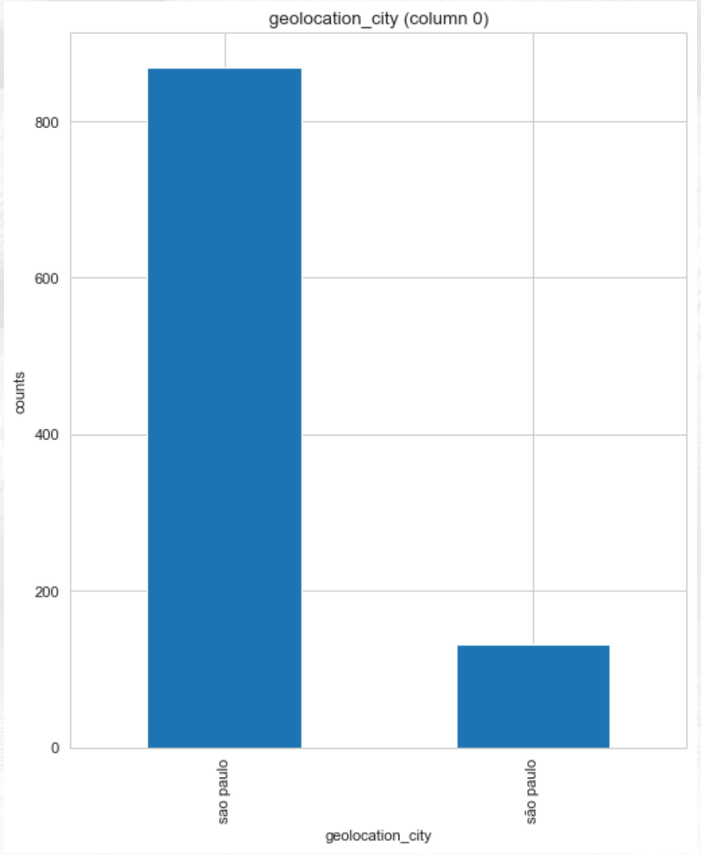
- Handled Missing Values
- Removed Duplicates
- Standardized Data Types
- Filtered Invalid Rows
- Previewed & Validated Cleaned Tables

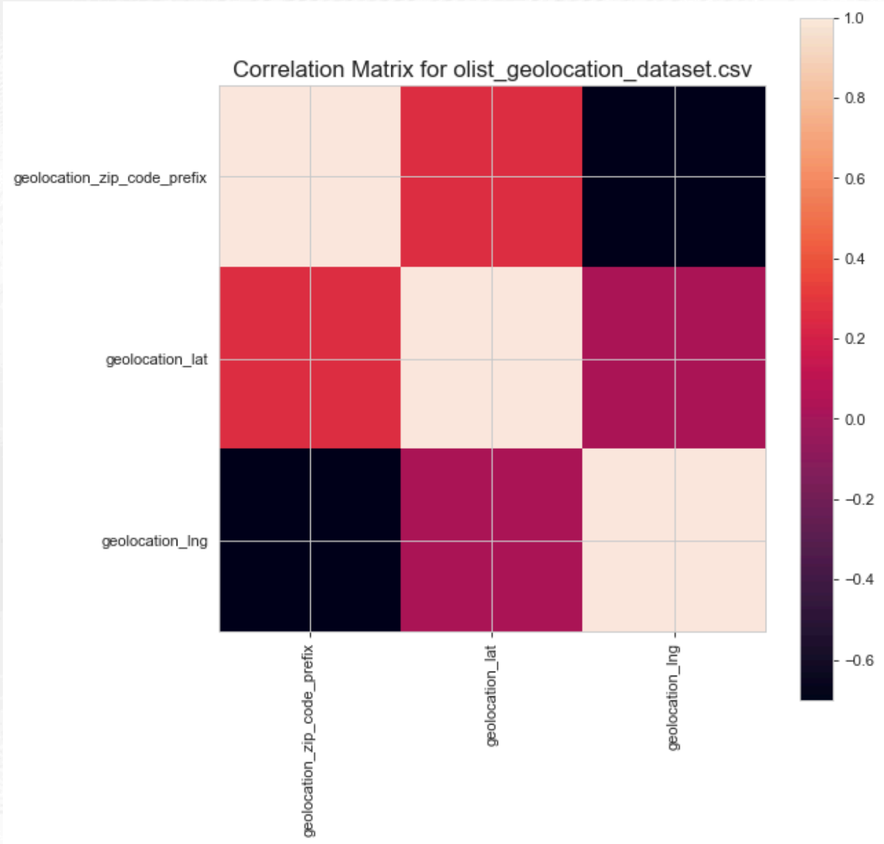
EXPLORATORY DATA ANALYSIS (EDA)

Initial data exploration was conducted on the dataset using the following:

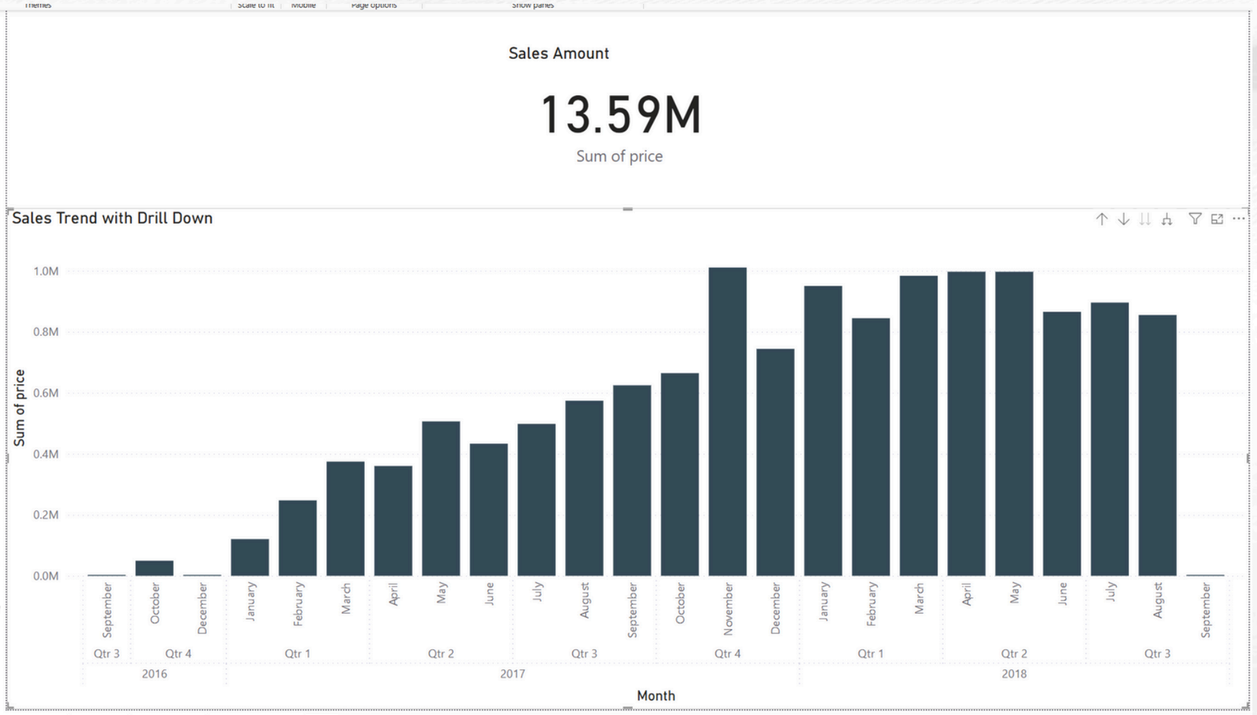
Method	Function
head([n])	Returns the first *n* rows.
info([verbose, buf, max_cols, memory_usage, ...])	Prints a concise summary of a DataFrame.
describe([percentiles, include, exclude])	Generates descriptive statistics.
value_counts([subset, normalize, sort, ...])	Returns a Series containing the frequency of each distinct row in the DataFrame.

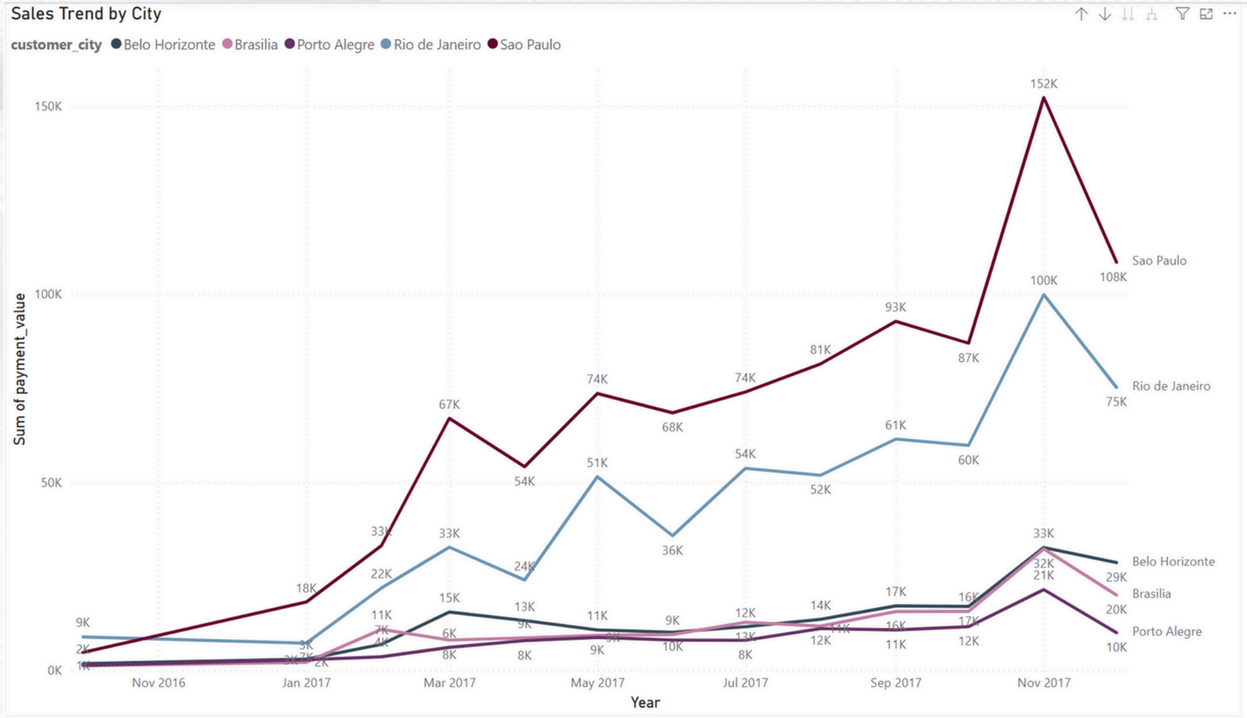
EXPLORATORY DATA ANALYSIS (EDA)





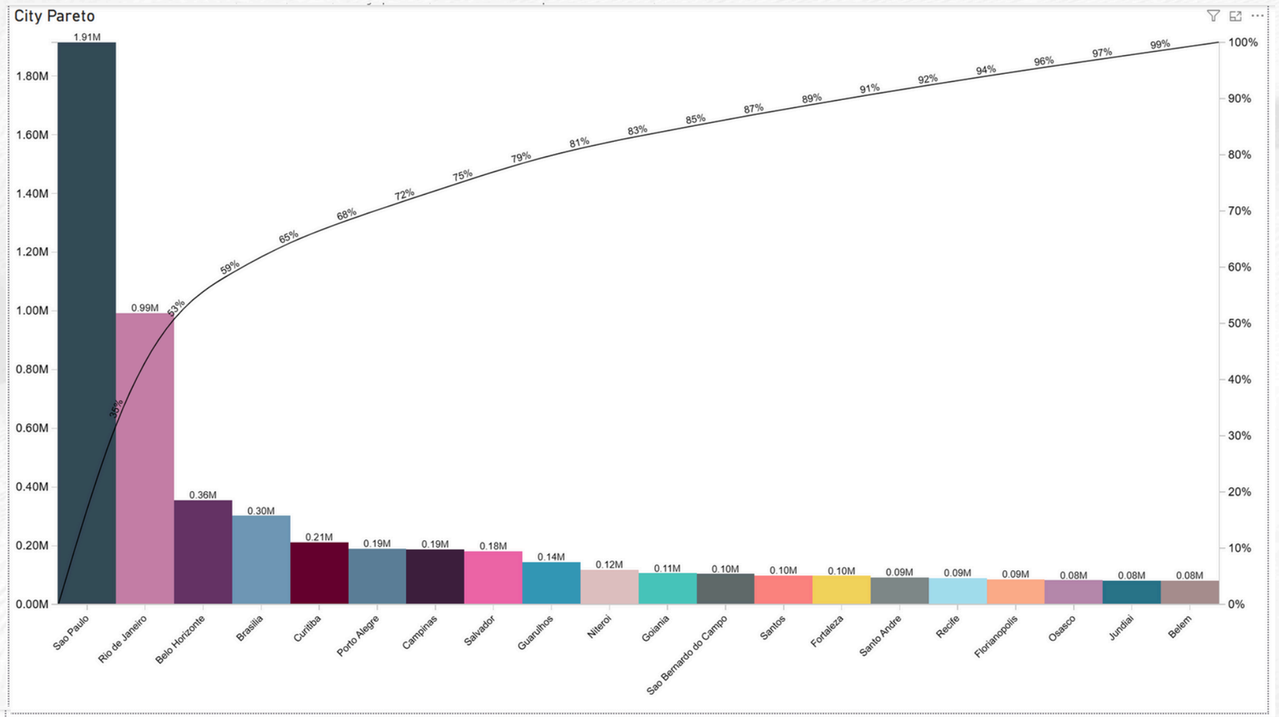
OVERALL ANALYSIS

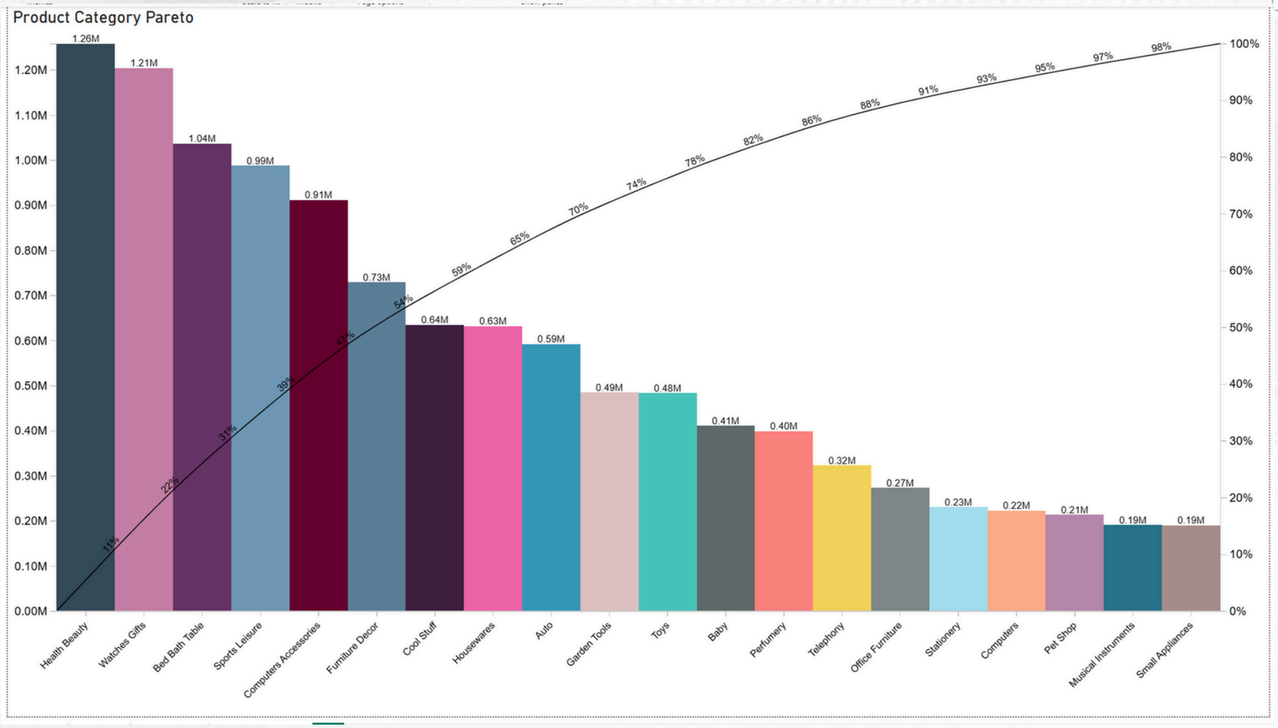




Cover the basics such as sales trend over time

OVERALL ANALYSIS

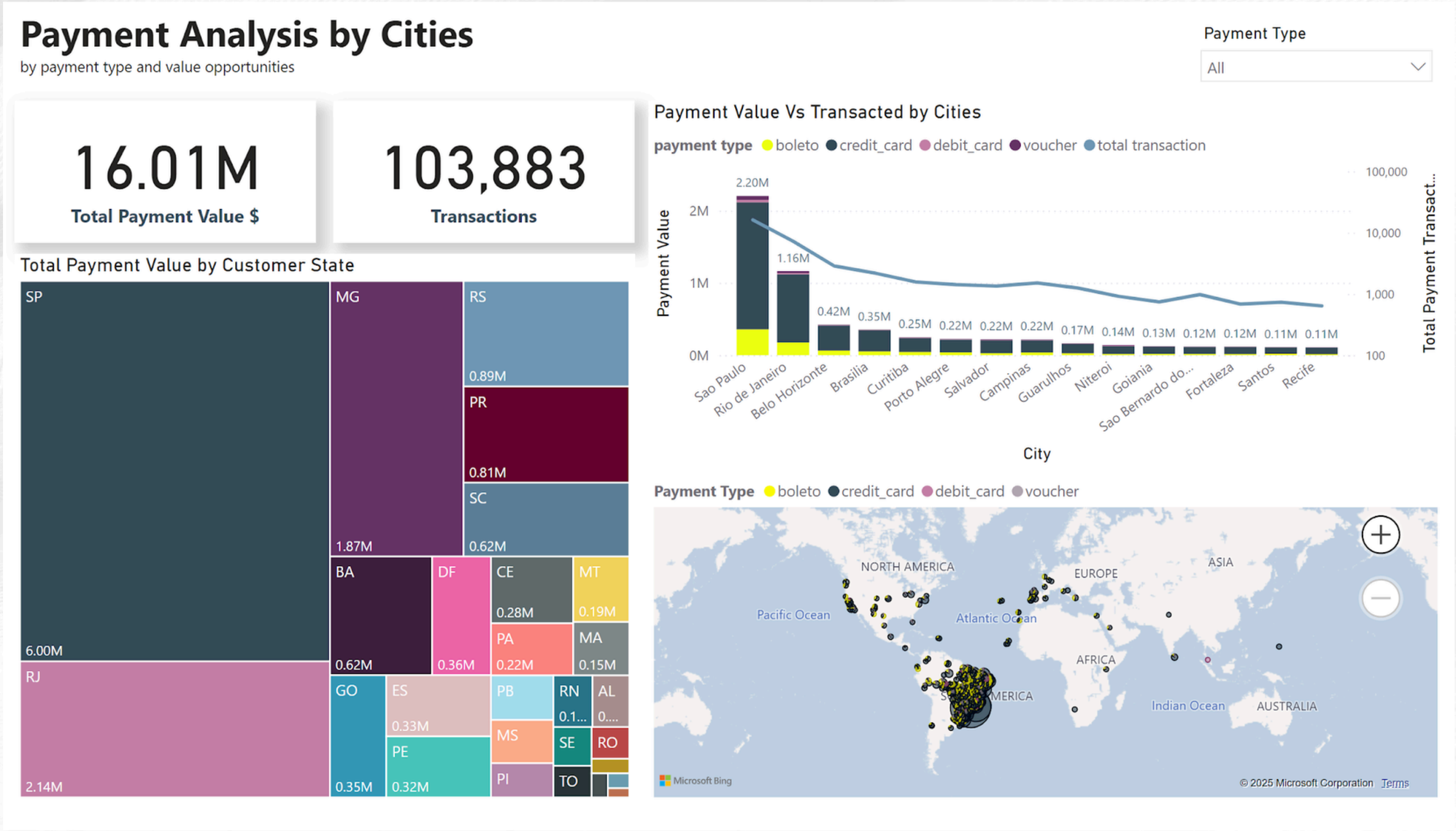




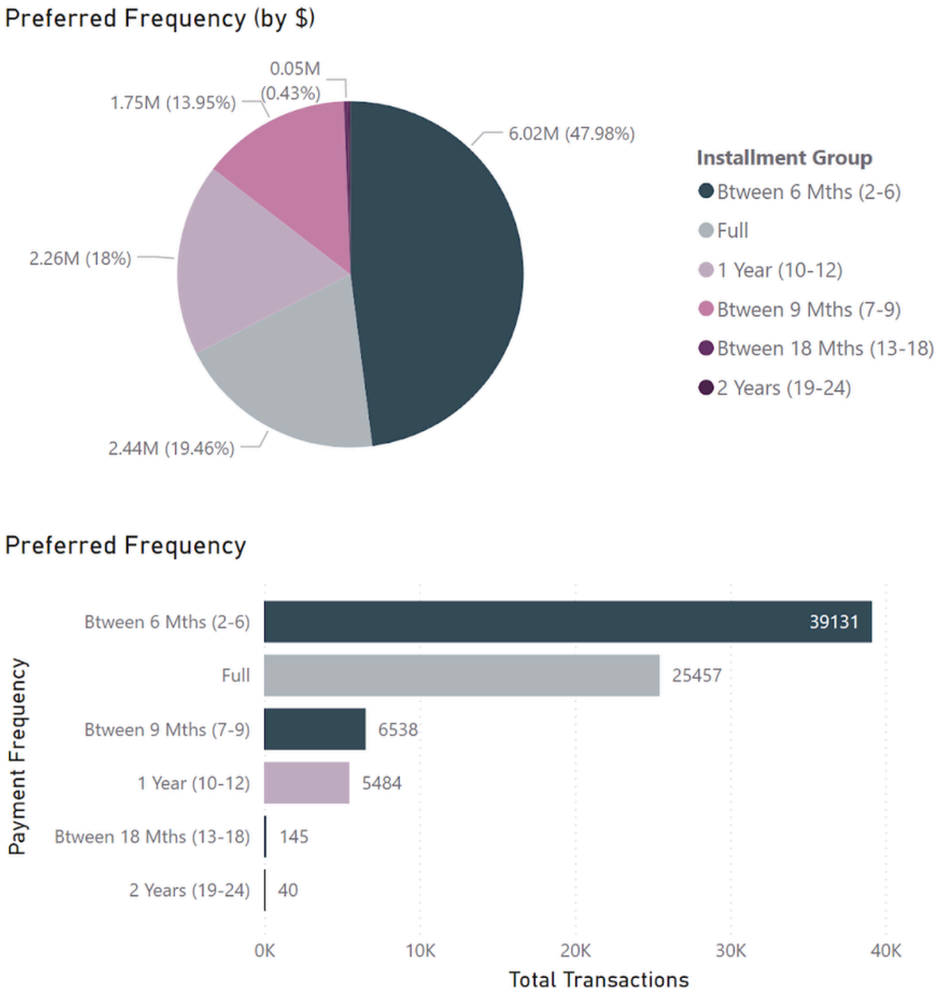
Pareto charts to zoom-in on the “vital few” such as key products or key cities that give the most impact to the business

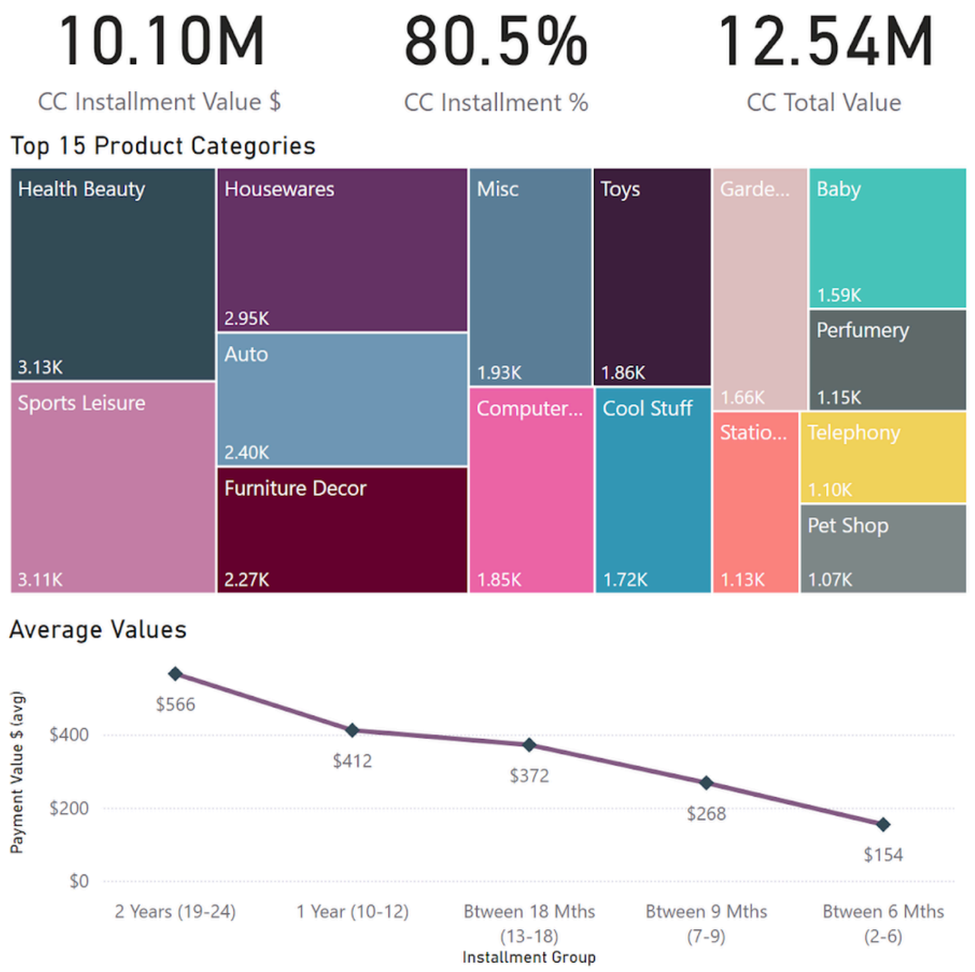
PAYMENT TRENDS ANALYSIS

- São Paulo, Rio de Janeiro, and other major urban centers account for the largest share of payment value and transaction volume.
- Credit cards are the overwhelmingly preferred payment method across all cities, with boleto and debit card lagging behind.



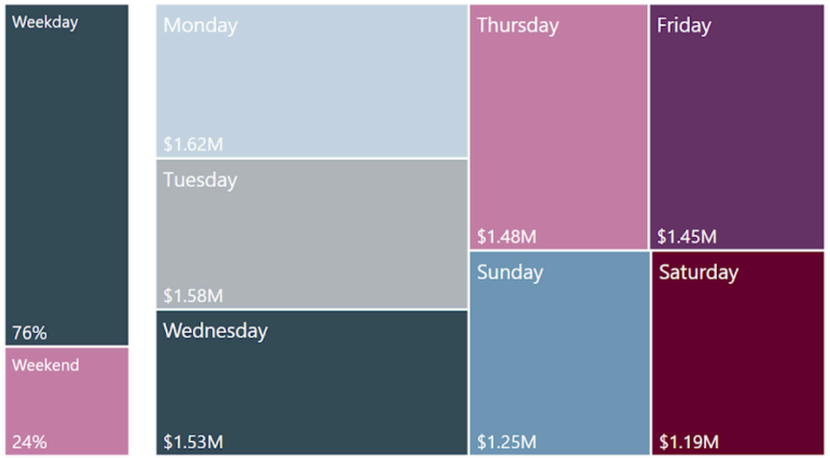
Installment Trends Analysis (1)

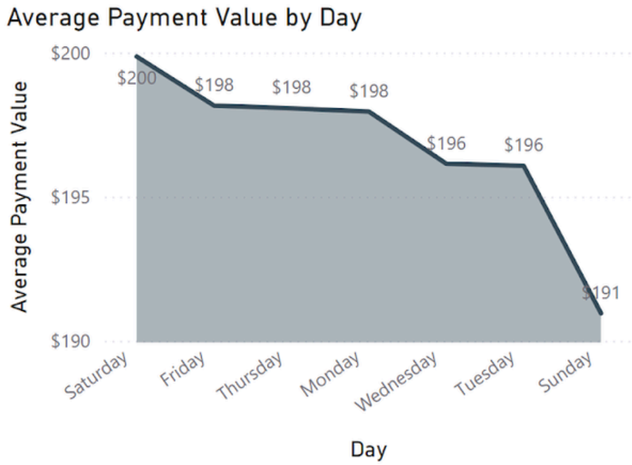




- Most Brazilians prefer to split their payments and over 80% of credit card payments are made in installments, with 2–6 months being the most popular option.
- Big ticket and lifestyle categories such as Health & Beauty, Housewares, and Sports Leisure that lead in both sales and installment use.
- The average value of purchases drops as the number of installments increases, but longer installment plans still attract sizable spend.

Installment Trend Analysis (2)





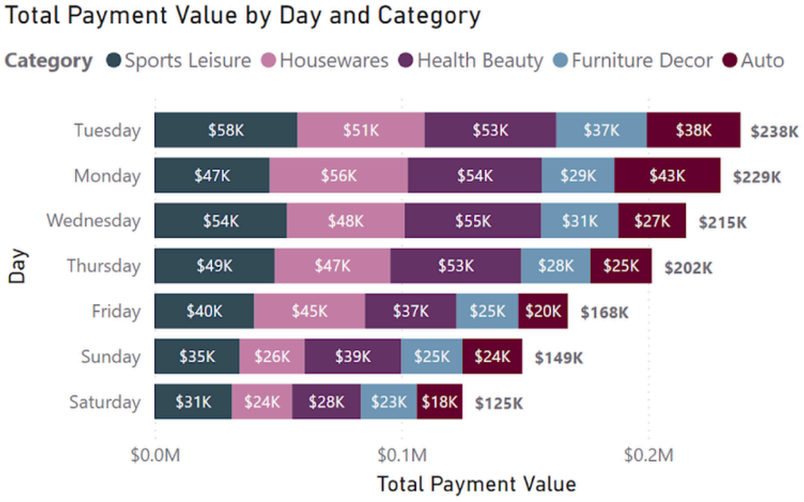
$7.66M
Weekday Installment PV

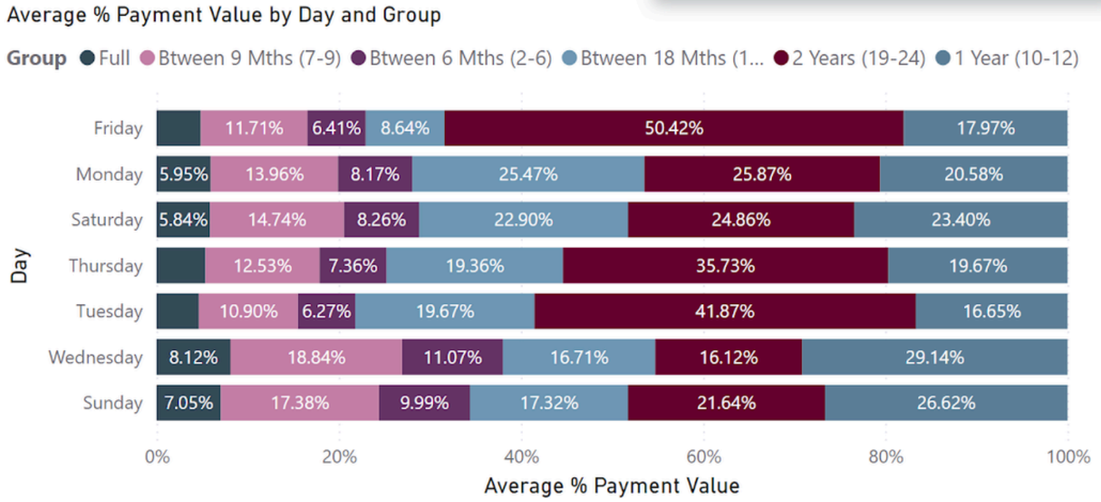
$2.44M
Weekend Installment PV

Monday
Peak Day

Sports Leisure
Top Category

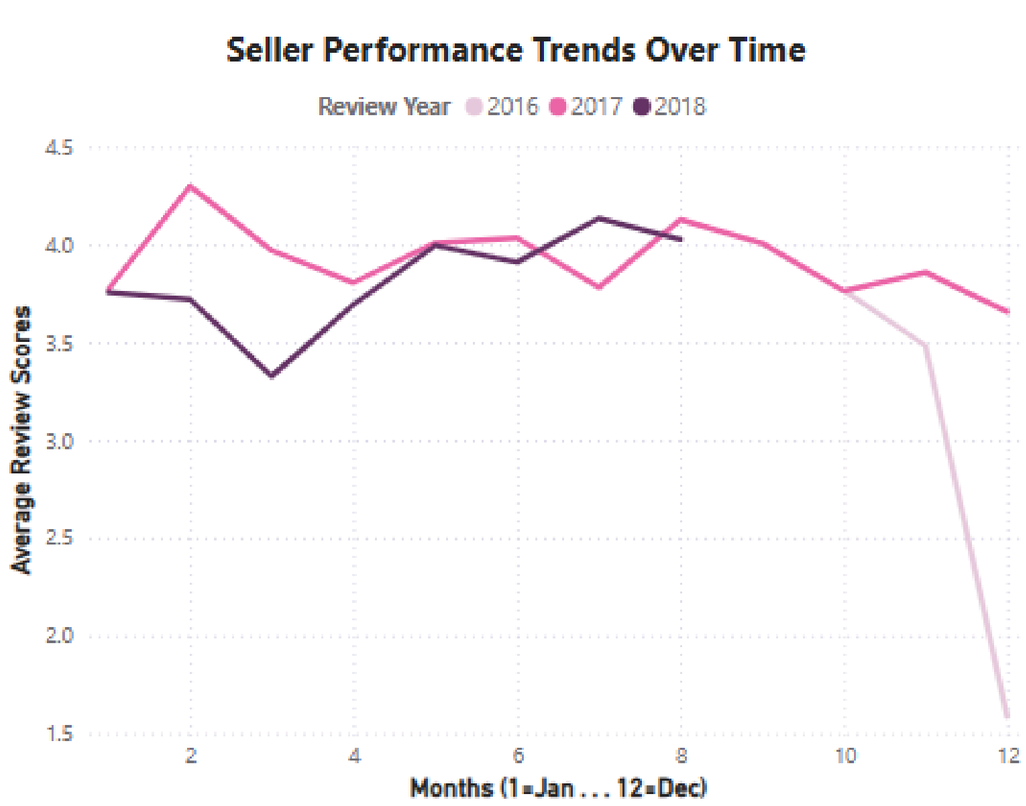
2 Years (19-24)
Top Installment Group

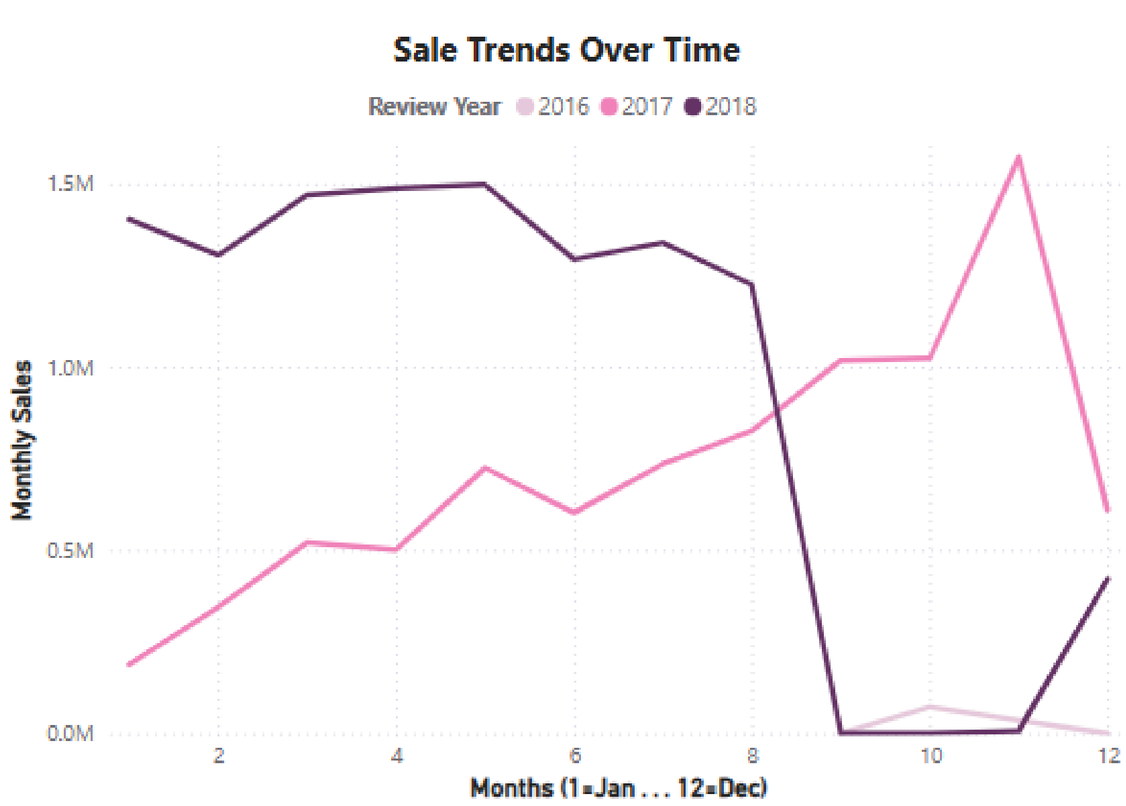




- Over three-quarters (76%) of installment transactions occur on weekdays, with Mondays leading as the top day for both order count and payment value.
- Weekday purchases favor longer-term installment plans, while weekends see fewer transactions and a preference for shorter plans or full payment.
- Sports Leisure remains the top category across all days, but other categories like Health Beauty and Housewares also maintain strong weekday traction.

SELLER PERFORMANCE ANALYSIS





Looking at Olist's sales trend across the dataset, two things stand out:

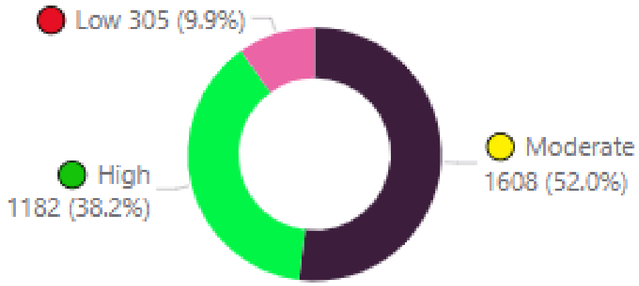
- A clear spike in sales between Nov 24–27, 2017 — aligned with their Black Friday Mega Sale.
- A sustained sales decline from August to November 2018 — which we believe may be due to a mix of economic slowdown, logistics strain, seasonality, or even increased competition.

How did Olist's sellers actually perform during these periods?

SELLER PERFORMANCE ANALYSIS

3095
Total No. of Sellers

Sellers Reliability Distribution



Seller Reliability	Seller Count	% over Total
Low	305	9.9%
Moderate	1608	52.0%
High	1182	38.2%
Total	**3095**	**100.0%**

Top 3 Best Performing City-Tier

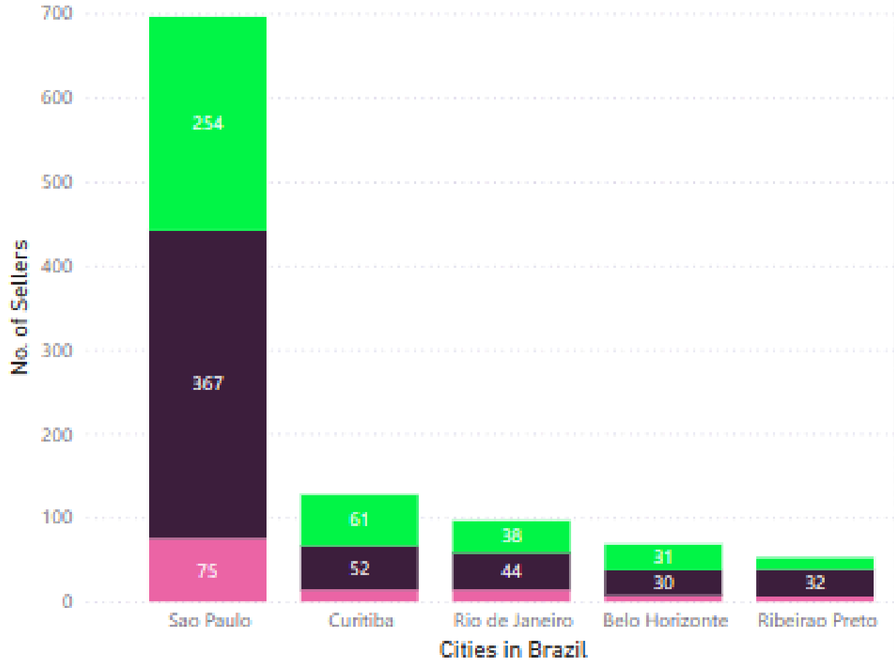
City - Tier	No. of Reviews	Review Score	Fulfillment Rate	Delay Rate	Reliability Score
Sao Paulo - High	4492	5	0.97	63.45%	84.18%
Curitiba - High	1350	5	0.97	63.45%	83.91%
Curitiba - Moderate	2536	5	0.97	63.45%	74.58%

Worst 3 Performing City-Tier

City - Tier	No. of Reviews	Review Score	Fulfillment Rate	Delay Rate	Reliability Score
Aguas Claras Df - Low	1	1	0.97	63.45%	46.42%
Alfenas - Low	2	1	0.97	63.45%	46.42%
Almirante Tamandare - Low	4	1	0.97	63.45%	46.42%

Note: *Due to high standardization in customer review scores and centralized processing, fulfillment and delay rates may appear similar across groups. However, the computed reliability score still reflects relative performance based on complete seller activity.*

Seller Reliability Across Locations



For the Composite Seller Reliability Score, we tested different weight combinations — from equal weights to delay-sensitive models. In the end, we chose this relatively balanced approach: 50% weight for customer reviews, 30% for order fulfillment rate, and 20% for delivery delay rate.

In this earlier version of the report, we used:

- Average scores for aggregation
- And a fixed delivery window of 21 days as the “late delivery” cutoff

⚠️ But this approach assumed all sellers had the same delivery expectations, which isn’t always true — and averages can be skewed by a few extreme scores.

SELLER PERFORMANCE ANALYSIS

In this refined analysis, we made two key improvements:

1. We used median instead of average to reduce bias from outliers.
2. We measured delay based on Olist's own promised delivery date (estimated_delivery_date) — a fairer and more realistic benchmark.

Conclusion:
This version provides a more representative, context-aware analysis of seller performance — helping Olist better target seller support, rewards, and interventions.

No. of Sellers:

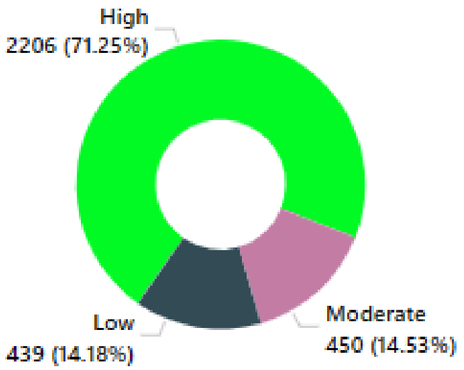
3096

100.00%

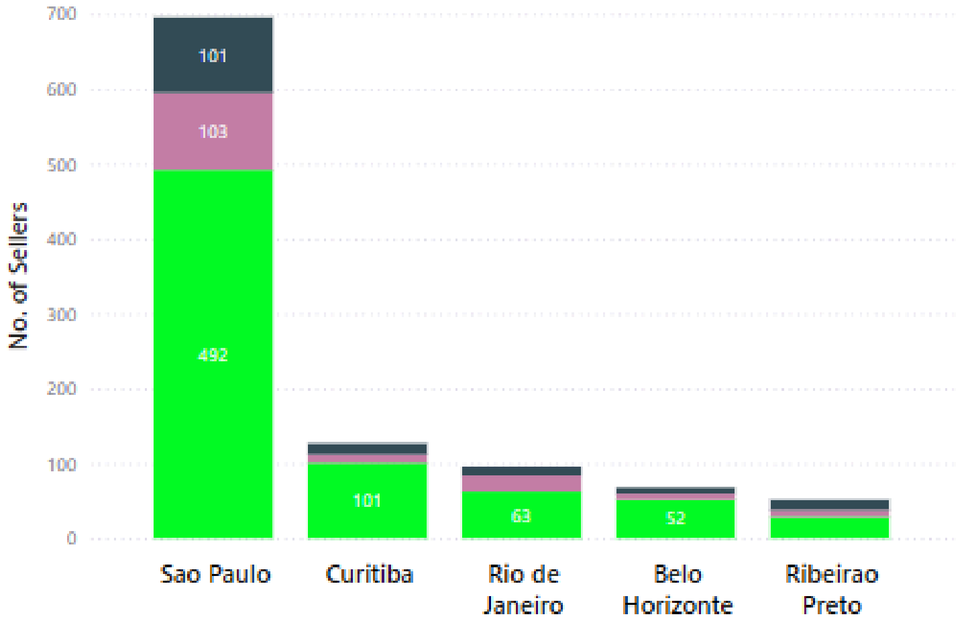
% Distribution

Reliability Tier	No. of Sellers	% Distribution
Low	439	14.18%
Moderate	450	14.54%
High	2206	71.28%
Total	3095	100.00%

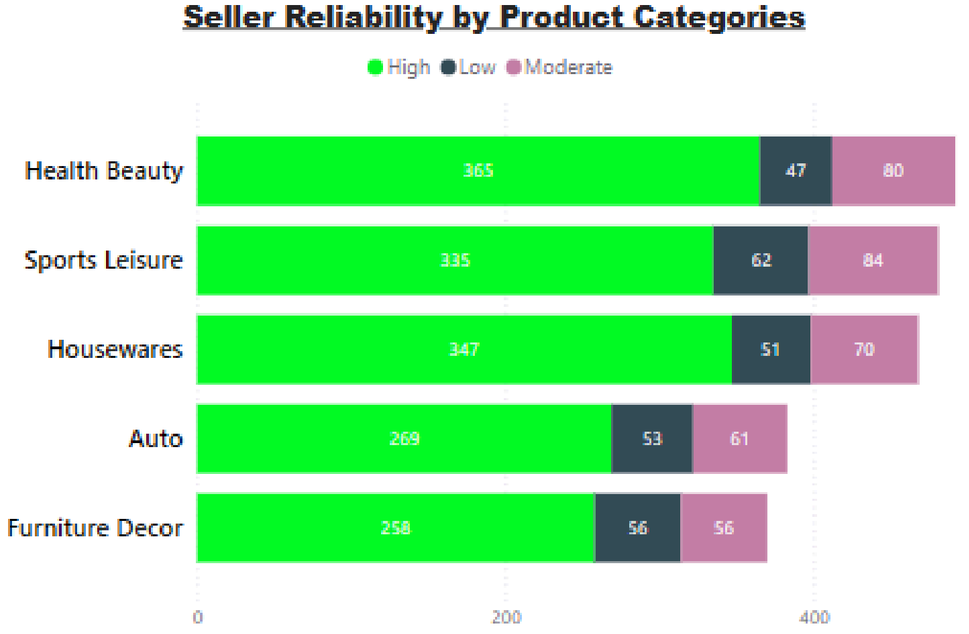
% of Sellers by Reliability Tier:



Seller Reliability by Cities



Seller Reliability by Product Categories



BUSINESS RECOMMENDATIONS

🛒 1. Maintain and Strengthen Seller Performance During Peak Seasons

- Observation: Seller performance is generally consistent, but a slight drop in review scores and an uptick in delays was observed during Q4 (e.g. Black Friday).
- Recommendations:
  - Implement a Flash Sale Readiness Program focused on SOPs and proactive support
  - Prioritise high-volume but moderate/low-tier sellers for coaching and logistics aid
  - Maintain service quality by pre-emptively supporting sellers before peak events

BUSINESS RECOMMENDATIONS

💳 2. Leverage Installment Payment Behavior

- Key Insights:
  - 80.5% of credit card value comes from installments.
  - Most popular plans: 2–6 months (39,131 transactions).
  - Higher installment lengths = higher average value
  - (e.g. 19–24 months → $566 avg)
  - Weekday installment payments total $7.66M vs weekends $2.44M
- Recommendations:
  - Promote flexible installment plans for higher-value products.
  - Time promotions around weekday behavior patterns.
  - Explore financing partnerships or optimised payment offers.

CHALLENGES/LIMITATIONS

🧩 Technical Setup

- Faced package and dependency issues during Python environment setup (e.g., Azure SDK conflicts, build warnings).
- Resolved installation errors to ensure stable execution of scripts.

🧹 Post-Merge Data Gaps

- Merging datasets introduced new missing values across products, sellers, and orders.
- Applied median imputation for fields like geolocation and product measurements.

🔁 Data Type Issues

- Several timestamp columns were misread as strings (object) instead of datetime.
- Manual conversion was needed for consistency and proper analysis.

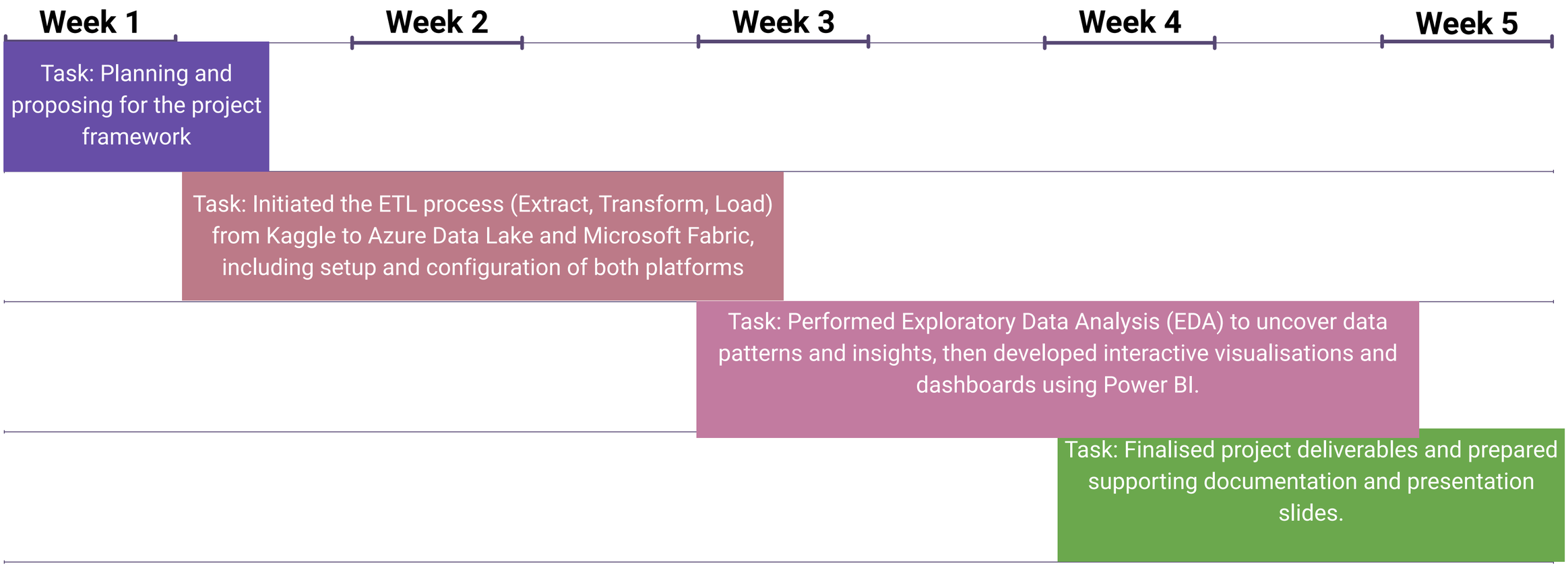
🚧 Limitations

- Project scope focused on ETL and early analysis — in-depth business insights were out of scope.
- Median imputation was chosen for simplicity but may introduce bias.
- Geolocation precision was limited (used zip code prefix, not full code).

🧪 Simplifications

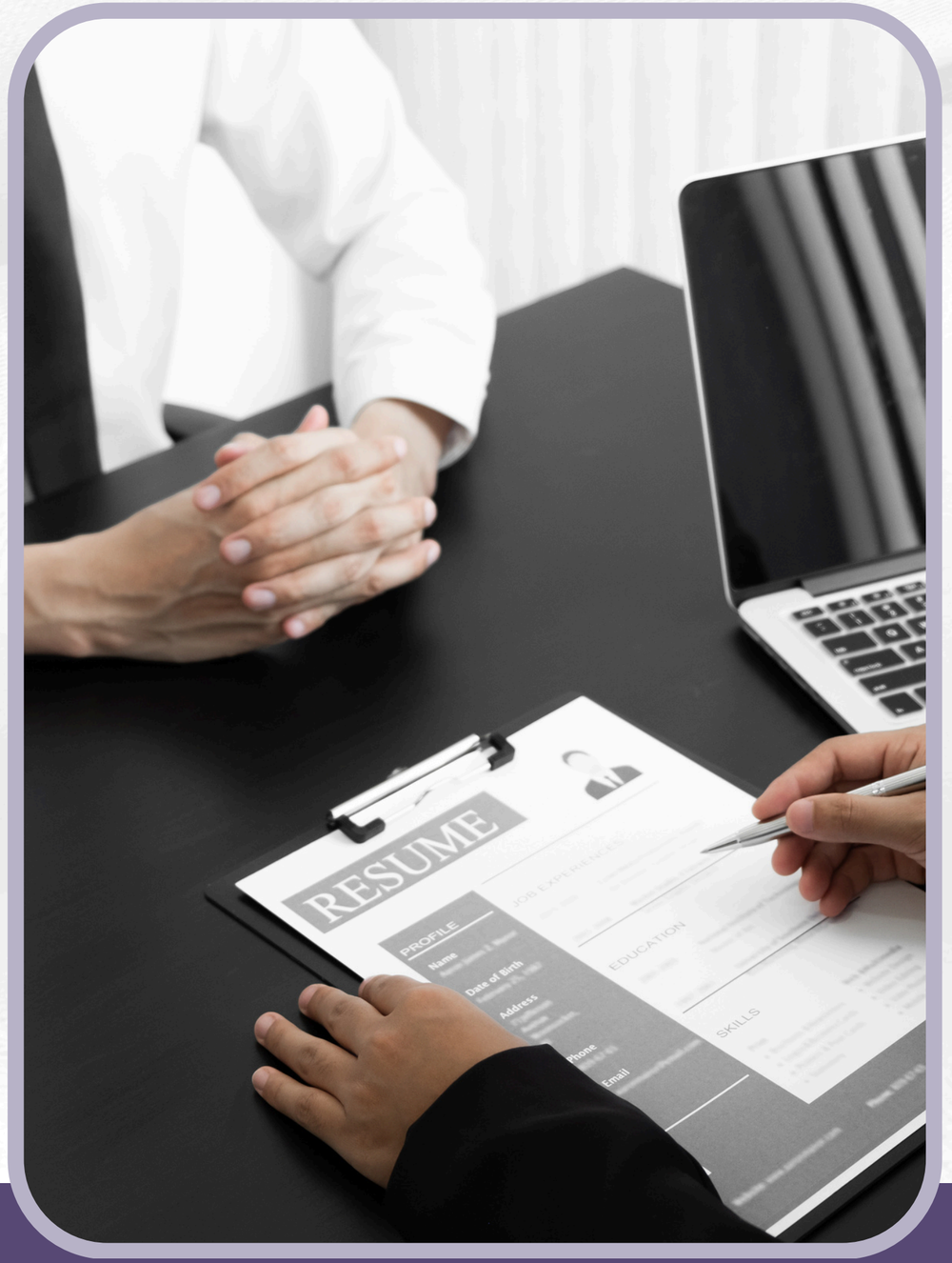
- Dropped unused fields (e.g., review_comment_title)
- Initial EDA based on a sample (first 1000 rows) to explore distributions

TIMELINE


Week 1
Week 2
Week 3
Week 4
Week 5
Task: Planning and proposing for the project framework
Task: Initiated the ETL process (Extract, Transform, Load) from Kaggle to Azure Data Lake and Microsoft Fabric, including setup and configuration of both platforms
Task: Performed Exploratory Data Analysis (EDA) to uncover data patterns and insights, then developed interactive visualisations and dashboards using Power BI.
Task: Finalised project deliverables and prepared supporting documentation and presentation slides.

CONCLUSION

- Successfully built a full ETL pipeline to extract, clean, and transform Olist’s e-commerce data from Kaggle.
- Used Python and Azure Data Lake to process and standardise across multiple datasets.
- Final cleaned data was loaded into Microsoft Fabric for reporting and visualisation.
- Dataset is now ready for deeper business insights and advanced analytics.



WHAT'S NEXT?

Deeper Business Analysis

- Advanced customer segmentation
- Product performance by region and category
- Geo-level sales insights using location data

Predictive & Advanced Analytics

- Forecasting sales and delivery delays
- CLTV and churn prediction
- Market basket analysis and recommendation engines

Future Enhancements

- Add COGS/Margin for profitability tracking
- Track customer behavior (browsing, cart, wishlist)
- Integrate external factors (holidays, inflation, weather)
- Refine logistics with inventory and carrier data
- Link to seller marketing funnel for deeper seller insights

RESOURCE PAGE

jupyter	PostgreSQL
Azure	
GitHub	Canva

OBRIGADO PELA SUA ATENÇÃO
THANK YOU FOR YOUR ATTENTION