# Azure Data Engineering Portfolio Project with Dashboard

*Author:*

Syed Faizan

March 28, 2025

# Contents

# Module 6: Final Project Report and Dashboard

## Introduction

The growing availability of big data presents exciting opportunities to uncover valuable insights that can inform business strategy, improve operational efficiency, and drive innovation. In this project, we explore a large-scale e-commerce dataset from the Olist platform using tools from the big data ecosystem. This dataset encompasses diverse data tables including customer information, order histories, seller profiles, payments, and product characteristics.

Leveraging a modern big data pipeline deployed on Microsoft Azure, we conducted end-to-end data engineering tasks—from data ingestion and transformation to visualization. Our architecture utilized Azure Data Factory for ingestion of raw data from GitHub and MySQL sources, ADLS Gen2 for data storage, Azure Databricks for data transformation using PySpark, and Power BI for dashboard creation. MongoDB was integrated to enrich the dataset for deeper analysis. The final outputs were visualized and interpreted in Power BI through interactive dashboards, enabling us to uncover meaningful trends, clusters, and correlations within the dataset.

## Project Summary

This final project report addresses a real-world problem in e-commerce logistics and consumer behavior using the Olist dataset—a comprehensive big data corpus derived from Brazil's leading marketplace platform. The central problem investigated in this study concerns delivery delays, product-level performance variability, and regional disparities in freight logistics. These issues affect customer satisfaction and operational efficiency, making them critical areas for business insight.

The project adopts an end-to-end big data analytics pipeline to examine these phenomena, combining scalable tools and advanced analytical methodologies. The dataset comprises transactional, customer, seller, and product-level information collected between 2016 and

2018. The methodology integrates Spark-based distributed processing, SQL-based aggregation, and visualization in Power BI.

- **Data Ingestion:** Source datasets were acquired from public repositories including GitHub, then ingested using parameterized workflows built within Azure Data Factory (ADF). Structured (MySQL), semi-structured (CSV), and NoSQL data from MongoDB were integrated for enrichment.

- **Storage and Transformation:** The raw data was stored in Azure Data Lake Storage Gen2 and transformed within Azure Databricks using PySpark. This included joins, filtering, null handling, feature engineering, delay calculation, and aggregation across temporal, geographic, and categorical dimensions.

- **Analytical Methods:** Techniques applied included Pearson correlation analysis, categorical aggregations, temporal trend decomposition, clustering, and influencer detection. Key metrics such as average delivery delay, freight value, payment installments, and customer density were examined across states, cities, and product categories.

- **Visualization and Reporting:** Processed outputs were transferred to Azure Synapse Analytics and Power BI for dynamic dashboard creation. The visual layer encapsulates geospatial insights, category-level trends, and delivery-time heatmaps that enable business users to quickly identify inefficiencies and anomalies.

This solution exemplifies the use of cloud-native platforms for real-world big data analysis. By synthesizing Spark processing, SQL-based modeling, and BI-layer interactivity, the project presents actionable insights with business implications. The advanced analytics revealed delivery delays up to -37 days, strong positive correlations between freight and payment value ($r = 0.74$), and disparities across states such as SP and AC in customer engagement and order fulfillment. These findings inform strategic recommendations around freight pricing, logistics coordination, and predictive risk modeling, positioning this report as a blueprint for data-driven decision-making in e-commerce environments.

## Architecture Overview: End-to-End Big Data Engineering with Azure Ecosystem

**The below figure** illustrates a comprehensive big data engineering architecture employing the Microsoft Azure ecosystem for ingesting, transforming, storing, and visualizing data from the Olist e-commerce dataset. Data ingestion is initiated via Azure Data Factory using parameterized JSON to extract raw data from SQL tables and HTTP sources hosted on GitHub, subsequently landing in Azure Data Lake Storage (ADLS) Gen2. Data transformation is executed within Azure Databricks, leveraging enriched MongoDB tables. The transformed data is re-stored in ADLS Gen2 and then queried through Azure Synapse Analytics. Visualization and reporting are performed in Power BI. This architecture adheres to modern data lakehouse principles, ensuring scalability, modularity, and real-time analytics capabilities.

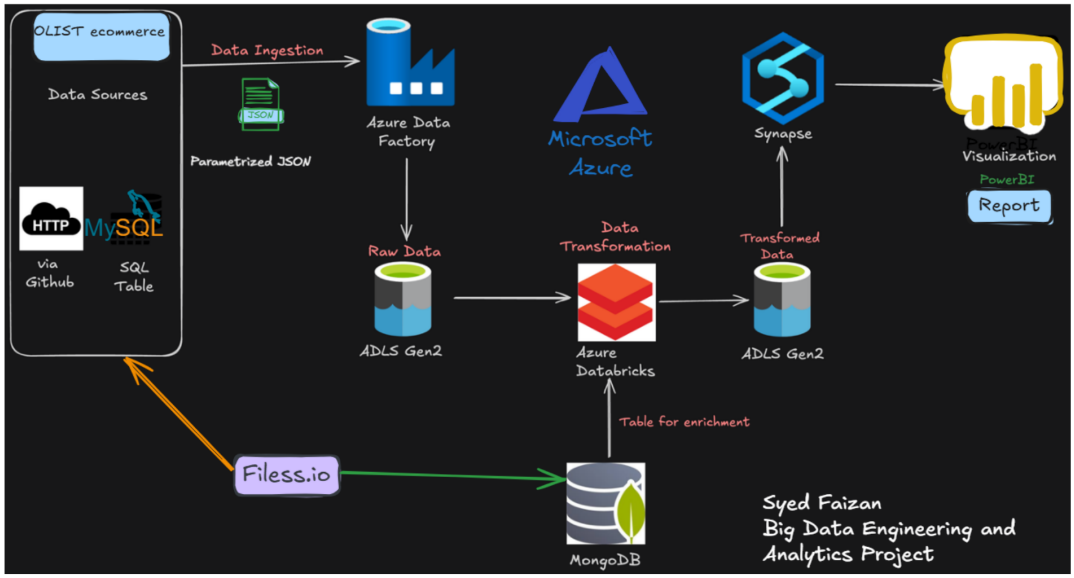


Figure 1: Architecture of the Olist Big Data project using Azure Data Factory, Databricks, Synapse, and Power BI

## GitHub Repository as a Versioned Source for E-Commerce Datasets

**The below figure** displays the GitHub repository hosting raw datasets integral to the Olist e-commerce pipeline. Each file, including customer, order, product, payment, review, and geolocation datasets, is uploaded in CSV format for reproducibility and transparency. The repository acts as a version-controlled source facilitating collaborative data engineering, reproducible ETL development, and integration with Azure Data Factory. Its publicly accessible nature enhances academic and industrial utility by serving as a reliable and centralized dataset for research and instructional purposes.

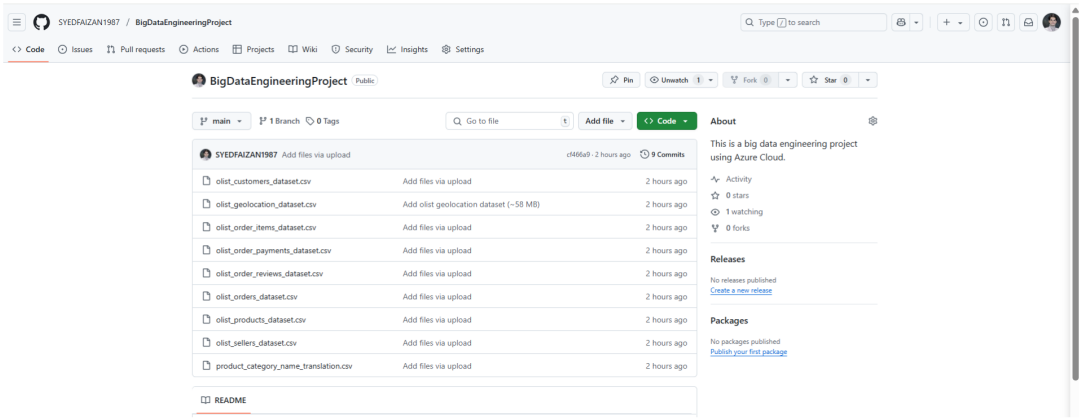


Figure 2: GitHub repository containing Olist e-commerce datasets used in the data engineering pipeline

## Database Infrastructure on Filess.io for Hybrid SQL/NoSQL Integration

**The below figure** highlights the dual-database deployment on Filess.io, demonstrating the availability of a MySQL relational database and a NoSQL MongoDB instance. This architecture facilitates both transactional and analytical workloads. The MySQL database handles structured tabular data ingestion, while the MongoDB instance supports schema-less enrich-

ment operations within Azure Databricks. This hybrid model enhances flexibility, enabling interoperability across polyglot persistence architectures for large-scale data workflows.

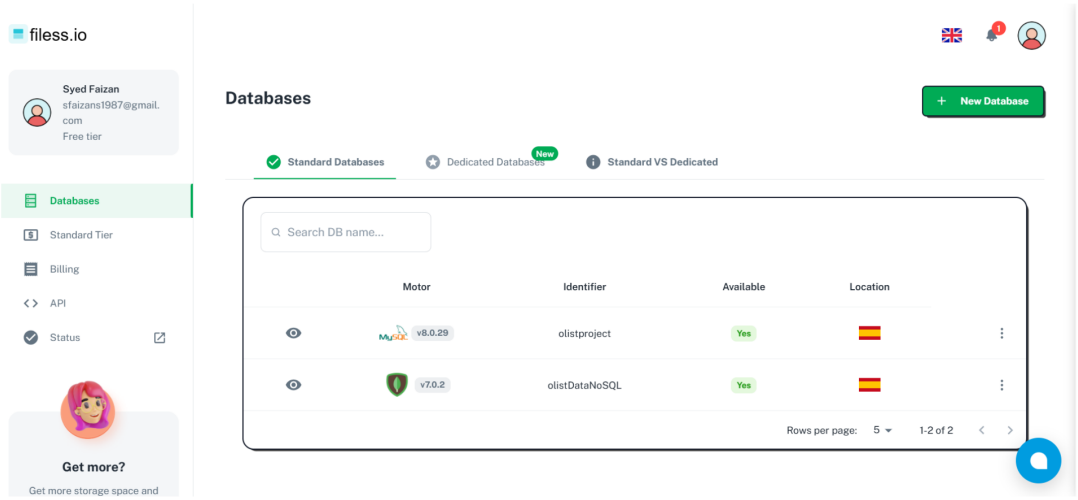


Figure 3: Dashboard from Filess.io showing availability of MySQL and NoSQL databases used in the project

## Entity Relationship Diagram (ERD) for Olist Dataset Schema

**The below figure** presents a normalized entity relationship diagram (ERD) for the Olist dataset, encompassing core and auxiliary tables linked via primary-foreign key associations. The diagram exhibits how orders interface with reviews, payments, items, customers, sellers, and geolocation data. Centralized around the *olist_orders_dataset*, this ERD facilitates relational joins critical for OLAP operations, ETL processing, and integrity checks during transformation workflows in Databricks.

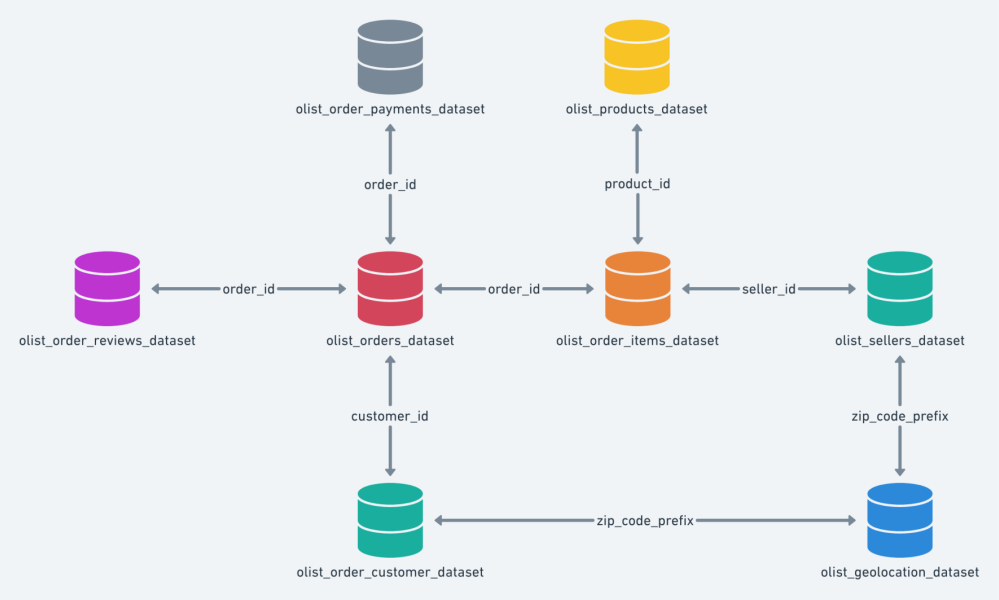


Figure 4: Entity relationship diagram (ERD) of the Olist datasets showing joins between customer, order, product, and seller tables

## Delta Lake Architecture for Multi-Zone Data Processing

**The below figure** depicts the layered data architecture implemented using Delta Lake to organize data into bronze, silver, and gold tiers. Raw ingestion is first captured in the *bronze* layer, followed by structural validation and cleaning in the *silver* layer, and finalized into business-ready aggregates in the *gold* layer. This architecture enhances data quality, streamlines query optimization, and supports high-throughput access for BI and ML workloads. The data lakehouse paradigm promotes governance, schema evolution, and scalable analytics.

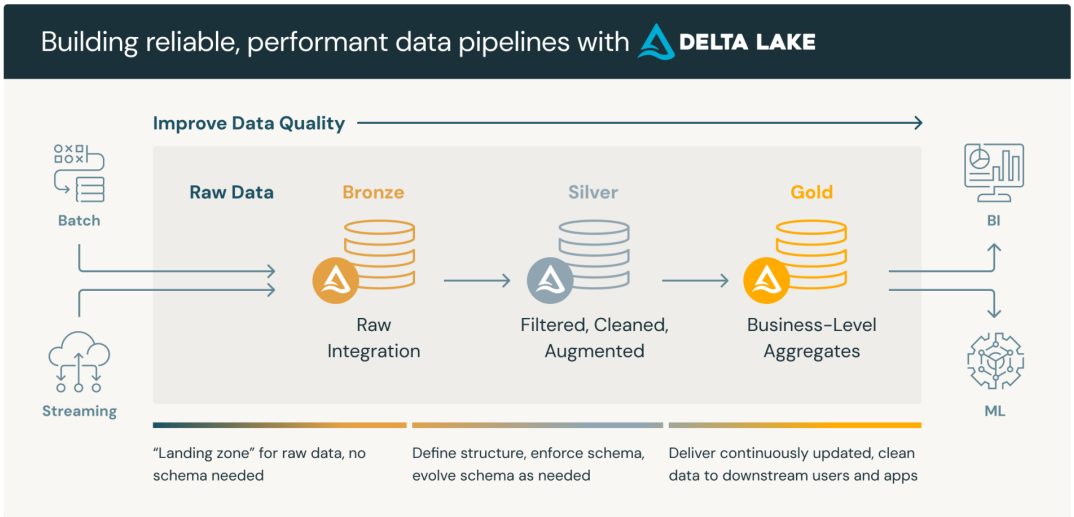


Figure 5: Delta Lake architecture showing raw (bronze), cleaned (silver), and aggregated (gold) layers in the data pipeline

## Parameterized Ingestion Pipeline with Azure Data Factory

**The below figure** illustrates an active Azure Data Factory pipeline employing a parameterized Lookup-ForEach-Copy construct to automate ingestion of structured data from SQL tables. Each iteration dynamically references metadata-driven inputs defined in a JSON configuration, supporting scalable and reusable ingestion patterns. The pipeline showcases successful execution logs, demonstrating reliability and integration runtime efficiency across east-US nodes. It serves as a foundational orchestration layer in the broader data engineering architecture.

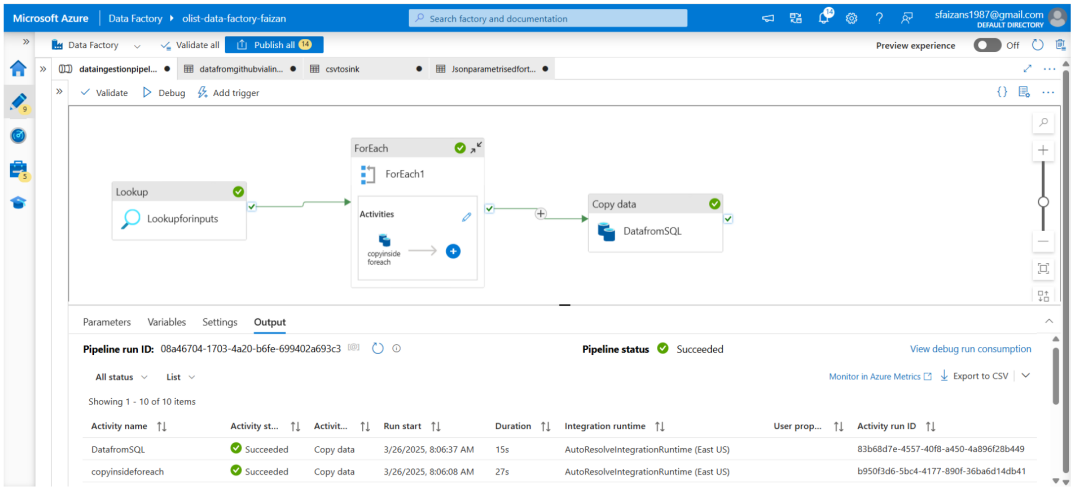


Figure 6: Azure Data Factory pipeline using Lookup, ForEach, and Copy activities to transfer SQL data

## Bronze Layer: Storage of Raw Data in Azure Blob Containers

**The below figure** displays the Azure Blob Storage interface where the bronze layer of the Delta Lake pipeline is instantiated. This container holds raw CSV files ingested from various sources, representing the uncurated and schema-flexible input for transformation operations. The metadata, including file sizes and timestamps, provides traceability and data lineage. It serves as the initial staging zone prior to schema enforcement and cleansing in the silver layer.

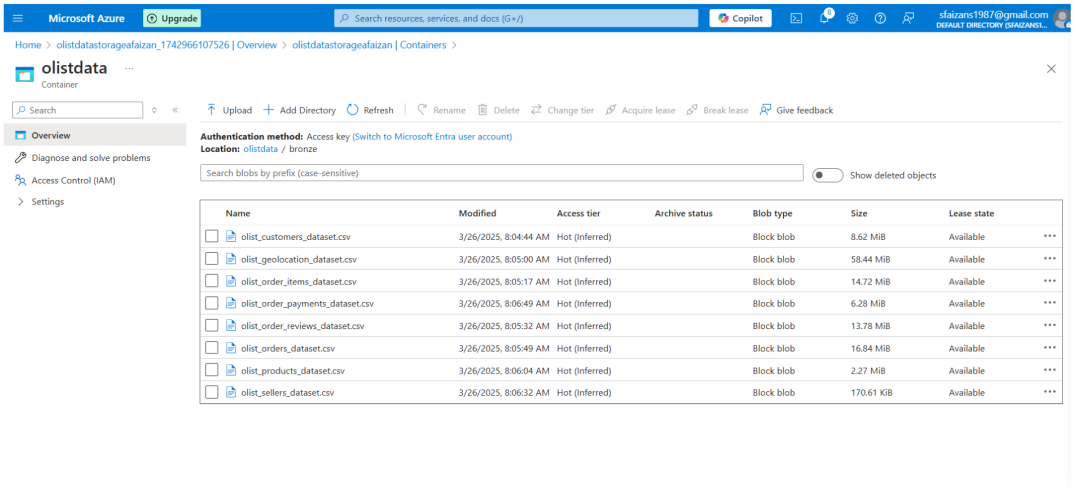


Figure 7: Azure Blob Storage container showing raw Olist datasets stored in the Bronze layer

## Gold Layer: Curated Aggregates for BI and Reporting

**The below figure** shows the gold layer within the Azure Blob Storage hierarchy. It consists of refined, aggregated datasets derived from the upstream bronze and silver layers. These curated files—organized by themes such as payment types, product performance, and delivery status—serve as the foundation for business intelligence, visualization, and executive reporting. The naming conventions and folder structure reflect domain-specific logic, facilitating efficient data access and governance.

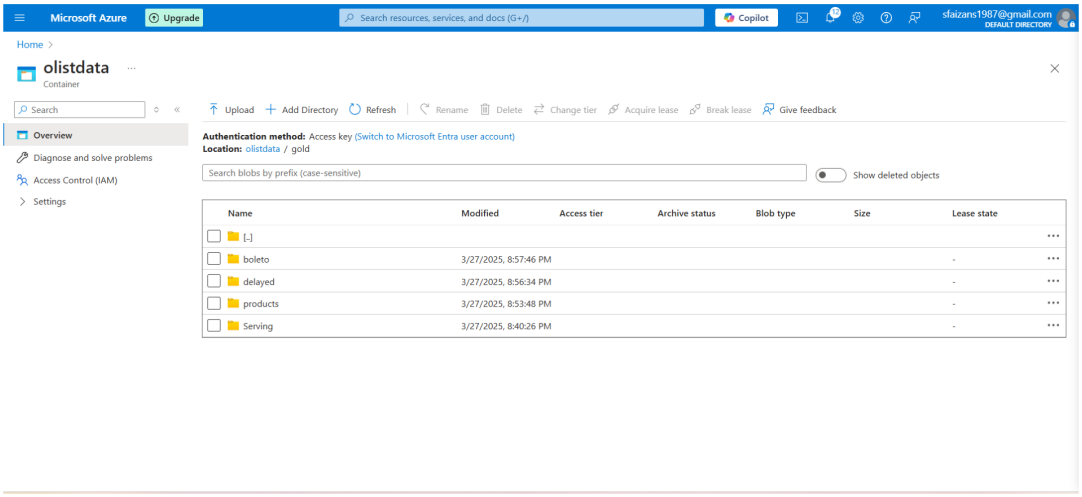


Figure 8: Azure Blob Storage container showing curated datasets in the Gold layer of the architecture

## Power BI Key Performance Indicator (KPI) Dashboard

**The below figure** features a Power BI dashboard presenting essential KPIs derived from the Olist dataset. It displays metrics such as average delivery delays, distinct product and customer counts, average payment values, and pricing variations among delayed shipments. These KPIs reflect the gold-layer aggregates and enable stakeholders to monitor operational performance and identify bottlenecks in fulfillment and payment processes. The visualization emphasizes decision support through clear, real-time summarization.

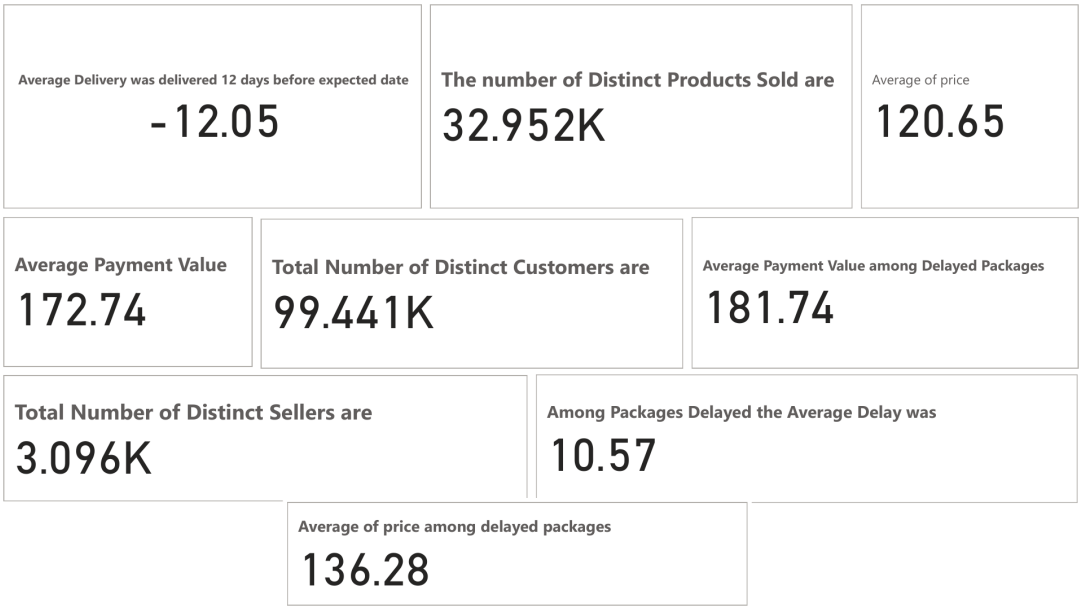


Figure 9: Power BI summary tiles showing KPIs such as average delay, average payment value, and distinct product/customer counts

## Dashboard Tool Used and Justification

**The below figure** showcases interactive visualizations developed using *Microsoft Power BI*, a widely adopted business intelligence platform renowned for its integration with Azure services, support for real-time analytics, and intuitive user interface. Power BI was chosen for its native compatibility with Azure Synapse and ADLS Gen2, enabling seamless ingestion of curated gold-layer data. Its robust data modeling and DAX capabilities facilitated dynamic analysis of key performance indicators (KPIs) such as delivery duration, order volume, and review-based freight trends. Furthermore, Power BI's slicers and filters empowered users to segment delivery insights across time, geography, and customer behavior dimensions, aligning with modern decision intelligence paradigms.

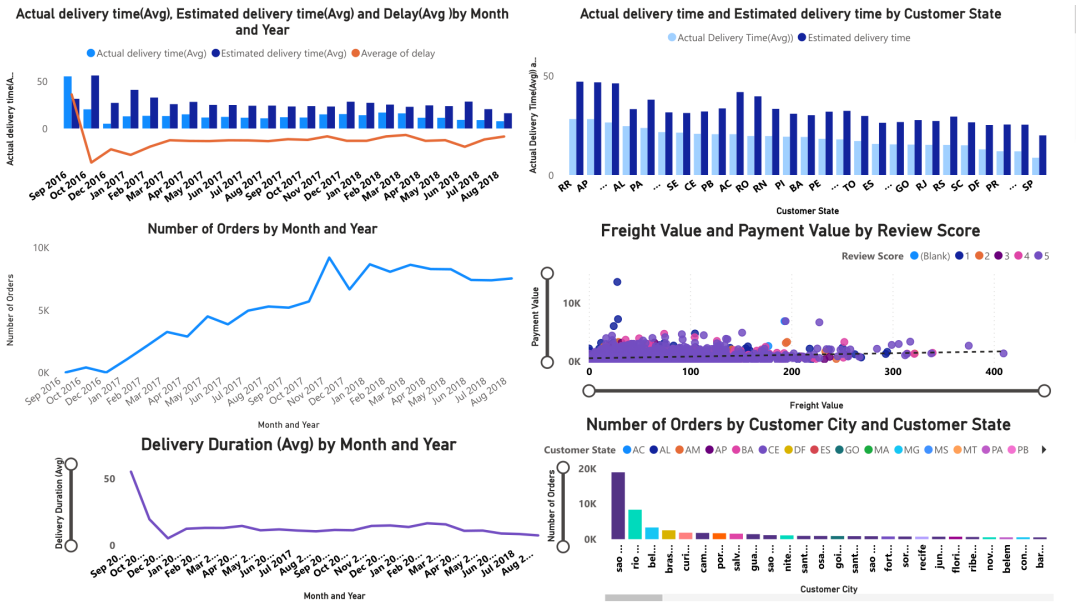


Figure 10: Power BI Dashboard

# Exploratory Data Analysis with Descriptive Statistics

## Product Categories with the Highest Average Delay

**The below figure** provides an analytical breakdown of product categories exhibiting the highest mean delivery delays, computed using PySpark in Databricks. The dataframe `final_df` was grouped by `product_category_name_english` and aggregated using the rounded average of the delivery delay. The top categories with significant delays include *fashion_female_clothing* with an average delay of -11.35 days, followed by *cine_photo* (-11.22 days), *construction_tools_lights* (-11.15 days), and *electronics* (-11.15 days). These negative values indicate that deliveries were completed, on average, earlier than expected. However, the magnitude and consistency of such negative deviations could point to inaccuracies in estimated delivery times or unusually early order fulfillments, warranting further scrutiny in delivery pipeline estimations.

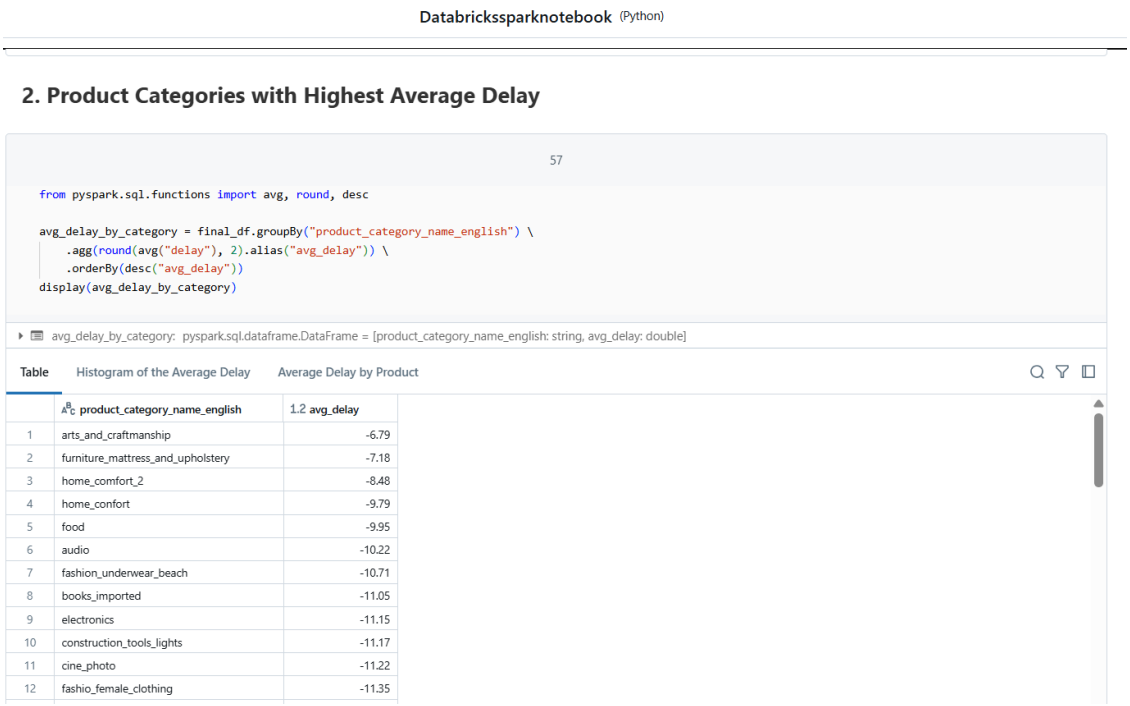


Figure 11: PySpark aggregation showing top product categories by average delivery delay

## Top 5 Product Categories by Total Sales

**The below figure** represents a ranked summary of the top five revenue-generating product categories, computed by aggregating `payment_value` over grouped categories using the `spark_sum()` function in PySpark. The leading category, *bed_bath_table*, recorded a total sales value of approximately 1.74 million BRL, followed by *health_beauty* (1.66 million BRL), *computers_accessories* (1.60 million BRL), *furniture_decor* (1.44 million BRL), and *watches_gifts* (1.43 million BRL). These categories contribute disproportionately to the platform's revenue, highlighting key verticals for investment and marketing focus. Such insights facilitate demand forecasting, category expansion decisions, and revenue-driven business strategies.

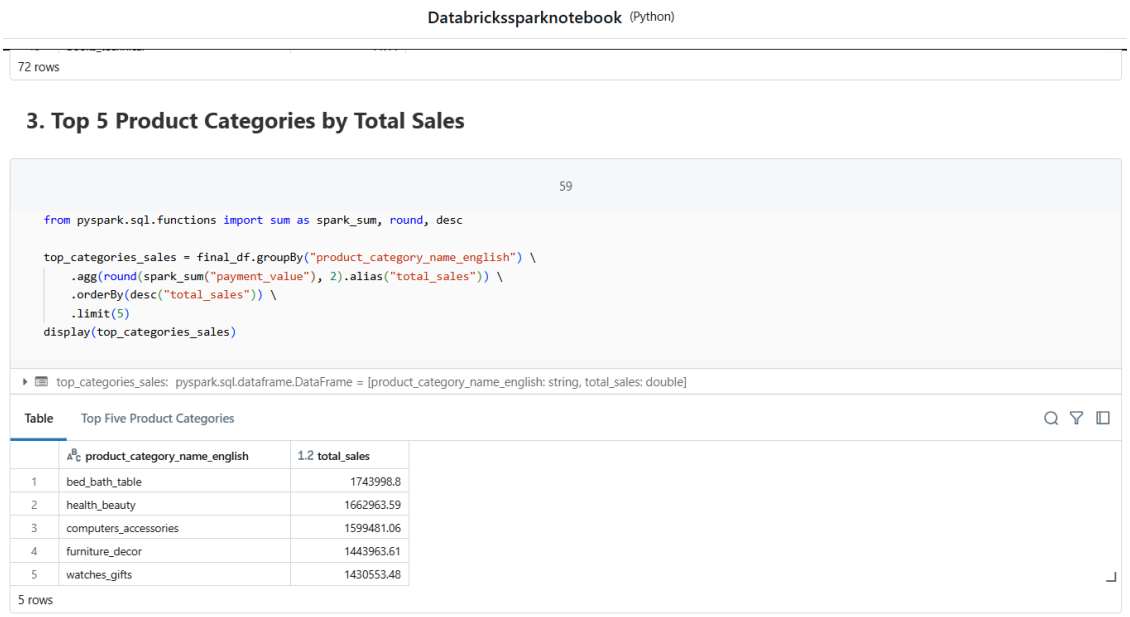


Figure 12: Top five product categories by total sales revenue computed in Databricks using PySpark

## Most Used Payment Types

**The below figure** presents the frequency distribution of payment types utilized by customers. The analysis, based on `count()` aggregation of the `payment_type` column, shows a clear dominance of *credit_card* transactions, which account for 87,776 payments. This is followed by *boleto* (23,190), *voucher* (6,465), and *debit_card* (1,706). Categories labeled *not_defined* and null values had negligible counts of 3 each. These statistics confirm a strong preference for credit cards, implicating their critical role in transaction success and customer convenience. Such information can guide payment gateway partnerships, fraud detection algorithms, and checkout process enhancements.

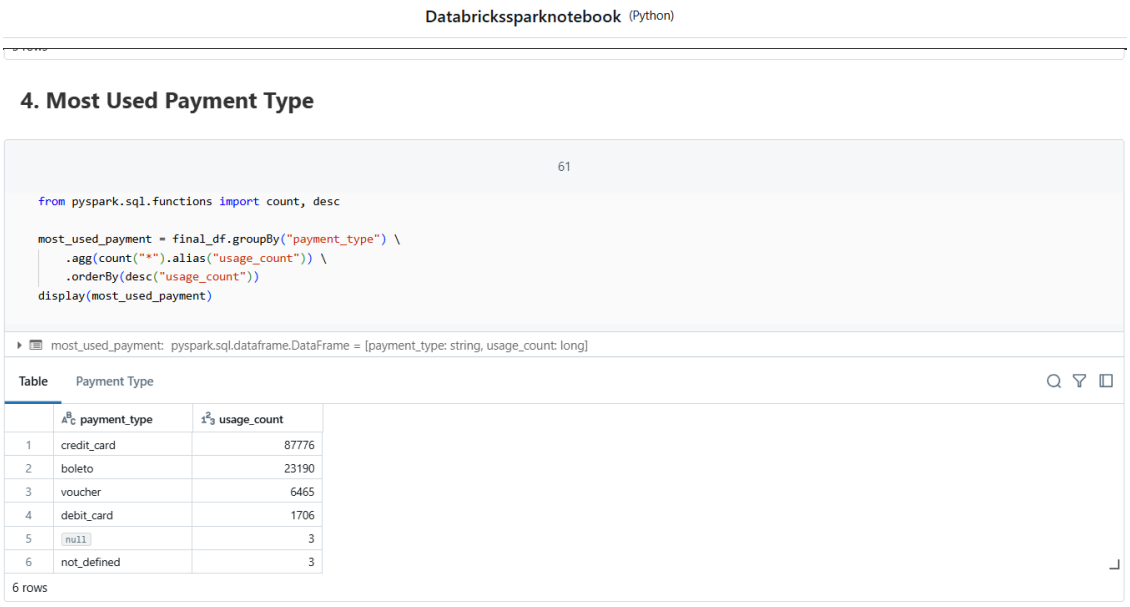


Figure 13: Distribution of payment types in Olist transactions based on usage count

## States with the Highest Number of Customers

**The below figure** outlines the geographical distribution of unique customers, aggregated by `customer_state` using the `countDistinct()` function. São Paulo (SP) is the leading state with 40,302 unique customers, substantially ahead of Rio de Janeiro (RJ) with 12,384 and Minas Gerais (MG) with 11,259. Other significant contributors include Rio Grande do Sul (RS) with 5,277 and Paraná (PR) with 4,882. This spatial customer density analysis is critical for region-specific marketing, logistics infrastructure planning, and personalized campaign targeting. The skewed concentration in southeastern Brazil reflects urban purchasing power and digital commerce penetration.

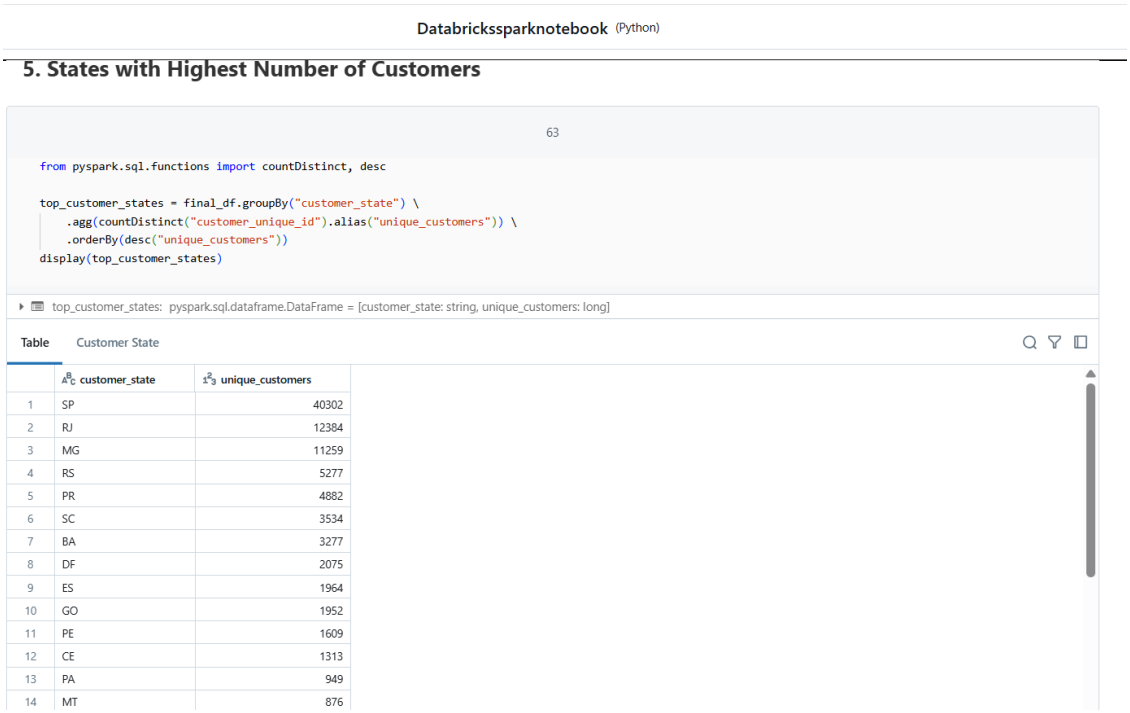


Figure 14: Top customer states ranked by number of unique customers in Olist platform

## Average Review Score per Product Category

**The below figure** ranks product categories by customer satisfaction, measured through the average review score. Utilizing `avg()` and `round()` functions, the dataset reveals that *cds_dvds_musicals* achieved the highest score of 4.64, followed by *fashion_childrens_clothes* (4.50) and *books_general_interest* (4.44). The consistently high scores (¿4.2) suggest robust customer satisfaction across these categories. These insights are essential for evaluating product quality, post-purchase experiences, and brand loyalty. Furthermore, the data may support the development of quality benchmarks and performance-based seller incentives.

Databrickssparknotebook (Python)

**6. Average Review Score per Product Category**

```
66

from pyspark.sql.functions import avg, round, col, desc

avg_review_by_category = final_df.groupBy("product_category_name_english") \
    .agg(round(avg(col("review_score").cast("double")), 2).alias("avg_review_score")) \
    .orderBy(desc("avg_review_score"))
display(avg_review_by_category)
```

avg_review_by_category: pyspark.sql.dataframe.DataFrame = [product_category_name_english: string, avg_review_score: double]

Table | Review Scores of Products

| | product_category_name_english | avg_review_score |
|---|---|---|
| 1 | cds_dvds_musicals | 4.64 |
| 2 | fashion_childrens_clothes | 4.5 |
| 3 | books_general_interest | 4.44 |
| 4 | books_imported | 4.42 |
| 5 | flowers | 4.42 |
| 6 | costruction_tools_tools | 4.42 |
| 7 | books_technical | 4.38 |
| 8 | small_appliances_home_oven_and_coffee | 4.32 |
| 9 | food_drink | 4.32 |
| 10 | luggage_accessories | 4.3 |
| 11 | fashion_sport | 4.26 |
| 12 | food | 4.23 |
| 13 | cine_photo | 4.22 |
| 14 | fashion_shoes | 4.22 |

Figure 15: Mean customer review score by product category in the Olist dataset

## Top 5 Sellers by Revenue

**The below figure** enumerates the top five sellers based on total revenue generation. The highest earning seller, with ID `7c67e1448b00f6e969d365cea6b01ab`, accrued 512,645.19 BRL. The next four sellers achieved revenues ranging from approximately 284,903 BRL to 312,456 BRL. These figures were derived using the `sum()` of `payment_value` grouped by `seller_id`. The identification of high-performing sellers is valuable for supply chain strengthening, loyalty programs, and volume-based partnership models. Revenue contribution at the seller level is a critical metric in marketplace sustainability and profit distribution modeling.

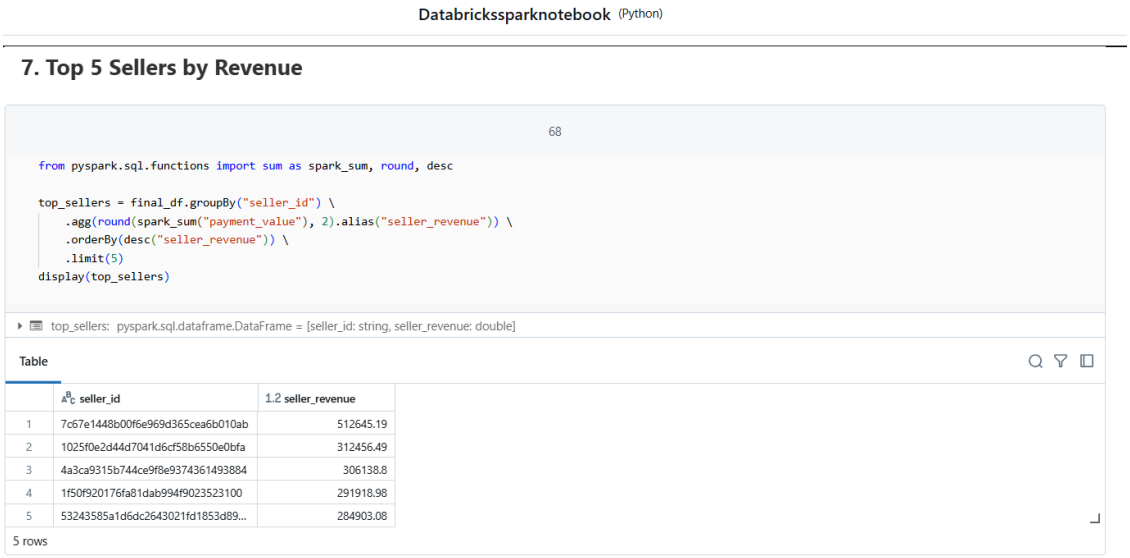


Figure 16: Top revenue-generating sellers identified using PySpark aggregation

## Product Categories with Most Reviews

**The below figure** displays the five product categories with the highest number of customer reviews. The `bed_bath_table` category received the maximum feedback with 11,847 reviews, followed by `health_beauty` (9,947), `sports_leisure` (8,942), `furniture_decor` (8,743), and `computers_accessories` (8,105). These metrics, based on `count()` of `review_id`, serve as a proxy for user engagement and brand reach. High review counts also suggest higher traffic, stronger visibility, and a more dynamic buyer-seller feedback loop.

Databrickssparknotebook (Python)

**8. Product Categories with Most Reviews**

```
70

from pyspark.sql.functions import count, desc

reviews_by_category = final_df.groupBy("product_category_name_english") \
    .agg(count("review_id").alias("review_count")) \
    .orderBy(desc("review_count")) \
    .limit(5)
display(reviews_by_category)
```

reviews_by_category: pyspark.sql.dataframe.DataFrame = [product_category_name_english: string, review_count: long]

Table | Product Category by Number of Reviews

| | product_category_name_english | review_count |
|---|---|---|
| 1 | bed_bath_table | 11847 |
| 2 | health_beauty | 9947 |
| 3 | sports_leisure | 8942 |
| 4 | furniture_decor | 8743 |
| 5 | computers_accessories | 8105 |

5 rows

Figure 17: Product categories with highest number of customer reviews

## Average Freight Cost by State

**The below figure** captures the average logistics expenditure per customer state by computing the mean of `freight_value`. The state of Paraíba (PB) had the highest average freight cost at 43.23 BRL, followed by Roraima (RR) at 42.98 BRL and Rondônia (RO) at 40.97 BRL. These higher costs correlate with geographic remoteness or logistical inefficiencies. States like Amapá (AP) and Rio Grande do Norte (RN) also show elevated freight averages, underscoring the need for supply chain cost optimization. These results are vital for strategic warehousing, route optimization, and regional pricing models.

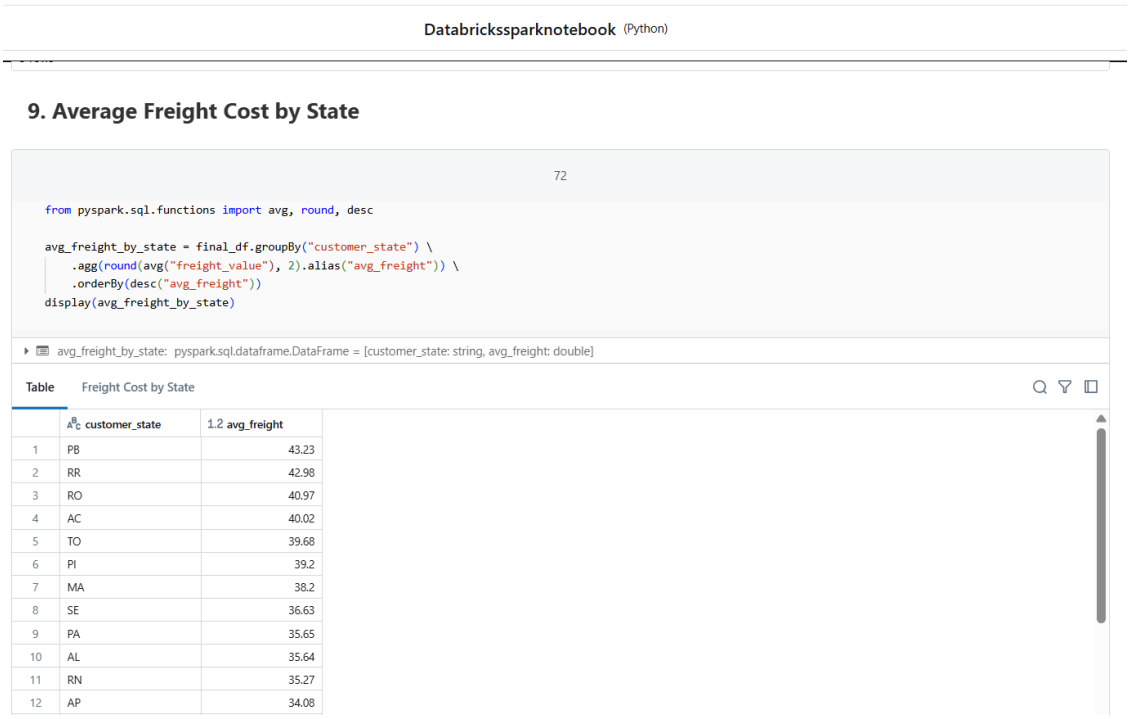


Figure 18: Average freight value per state in Brazilian e-commerce logistics

# Insights and Visualizations

## Insight 1: Relationship Between Freight Value, Payment Value, and Review Scores

**The below figure** visualizes the interaction between *freight value*, *payment value*, and *customer review scores* using a multi-dimensional scatterplot. Review scores, ranging from 1 to 5, are encoded with color, while the x-axis and y-axis represent freight and payment values respectively. A substantial concentration of data points lies below a freight value of 100 and payment value of 2000. Notably, high payment values (e.g., ¿ 10,000) are present across multiple score levels, including lower review scores, suggesting that price alone does not guarantee higher satisfaction. There appears to be a moderate positive trendline, albeit with high variance, implying weak linear correlation. This visualization aids in identifying outliers and understanding customer sentiment relative to transactional logistics.

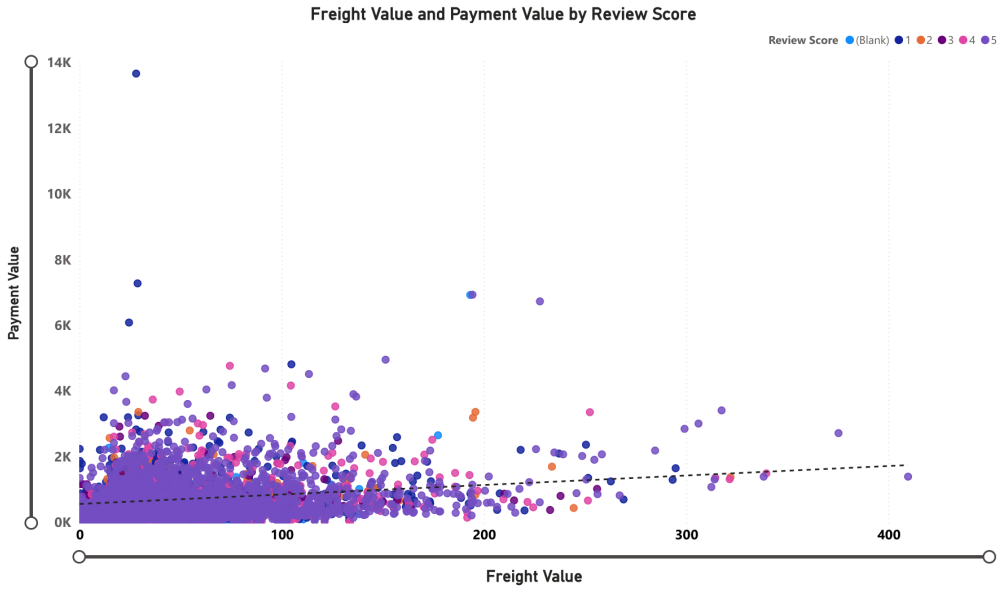


Figure 19: Scatter plot of freight value vs payment value, segmented by review score

## Insight 2: Correlation Analysis Among Numerical Features

**The below figure** presents a Pearson correlation matrix to evaluate linear relationships among 18 numerical variables from the Olist dataset. Strong positive correlations are observed between *price* and *payment value* (0.74), *payment value* and *freight value* (0.42), and among product dimensions such as *length, height, width*, and *weight* (ranging from 0.46 to 0.61). Interestingly, the variable *delay* shows a moderate positive correlation with *actual delivery time* (0.60) and a moderate negative correlation with *estimated delivery time* (-0.51), validating the accuracy gap in delivery estimates. The matrix offers an essential feature selection reference and supports regression modeling by identifying multicollinearity.

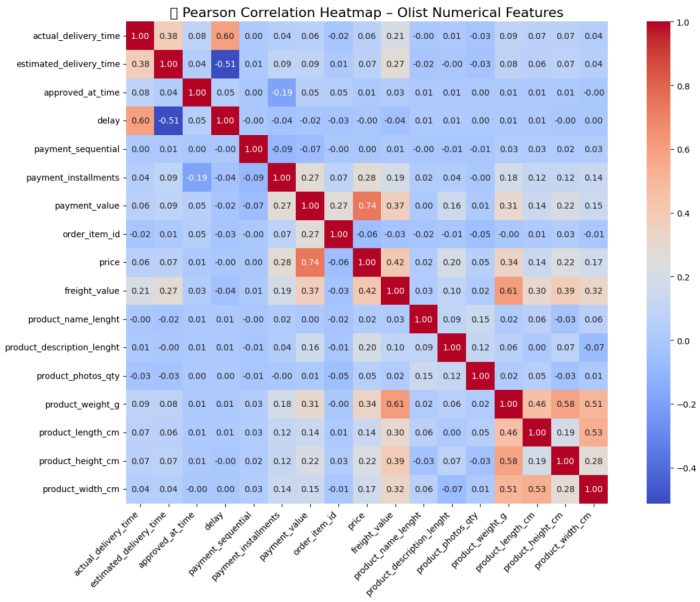


Figure 20: Pearson correlation heatmap of key numerical features in the Olist dataset

## Insight 3: Negative Correlation Between Delay and Freight Value

**The below figure** shows a distinct negative linear correlation between *delay* and *freight value*, indicated by a red regression line. As delays become more negative (earlier deliveries), the freight value appears to rise, with the trend extending toward freight values up to 200,000 for significantly negative delays. This inverse relationship implies that higher logistics expenditure might contribute to early or on-time deliveries. The plot suggests the potential of *freight value* as a predictive feature in modeling delivery performance or delay classification.

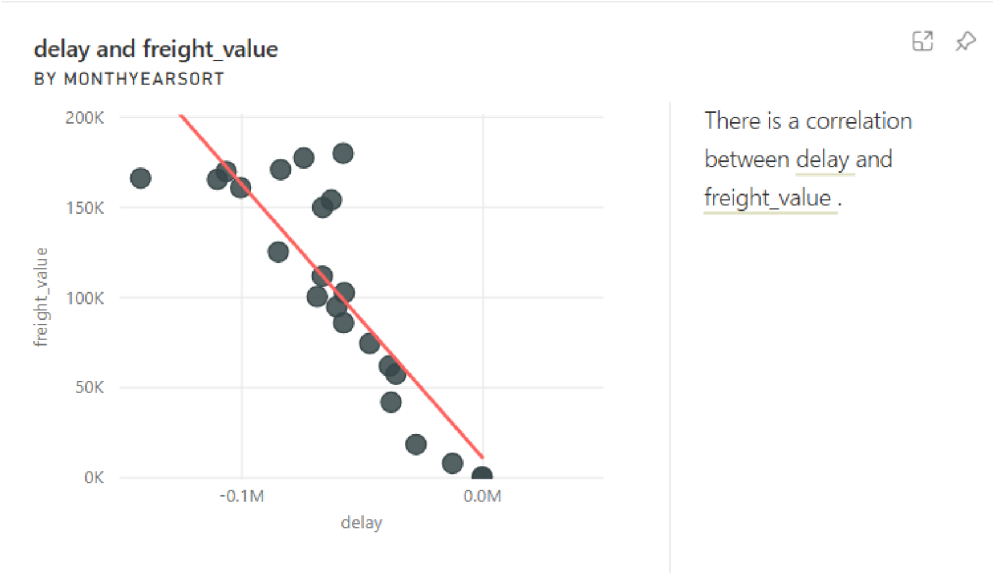


Figure 21: Scatter plot showing negative correlation between delivery delay and freight value

## Insight 4: Delay and Visual Presentation (Product Photos)

**The below figure** examines the relationship between *delay* and *product photos quantity*. The red trendline illustrates a negative correlation, where products with more images tend to have less delivery delay. For instance, products with nearly 20,000 photos are associated with the most negative delays (earliest deliveries), while those with fewer photos exhibit smaller negative delays. This insight suggests that better-presented products (those with richer media) are possibly from more professional or efficient sellers, enhancing operational predictability. This variable could be leveraged for seller segmentation or delivery risk prediction models.

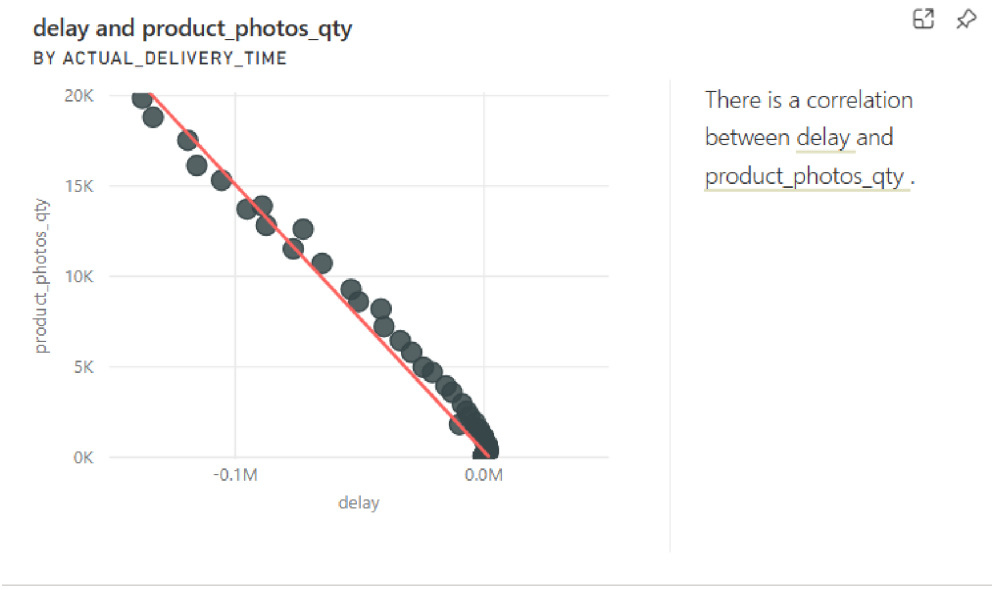


Figure 22: Scatter plot showing correlation between delivery delay and number of product photos

## Advanced Data Analysis

### Analysis of Delivery Time, Estimated Delivery, and Delay Over Time

The **below figure** presents a temporal analysis of delivery performance metrics from September 2016 to August 2018. The vertical bars depict average *actual delivery time* (light blue) and *estimated delivery time* (dark blue), while the overlaid line graph (in orange) illustrates the *average delay*. In September 2016, the average actual delivery time exceeded 55 days, while the estimated delivery was approximately 30 days, resulting in an average delay of over 25 days. Subsequently, both actual and estimated delivery times declined, stabilizing between 10–15 days for actual and 20–30 days for estimated deliveries. The delay consistently remained negative, ranging from approximately -37 days in October 2016 to around -10 days by mid-2018, indicating a significant early delivery pattern. These findings highlight improved logistics over time.

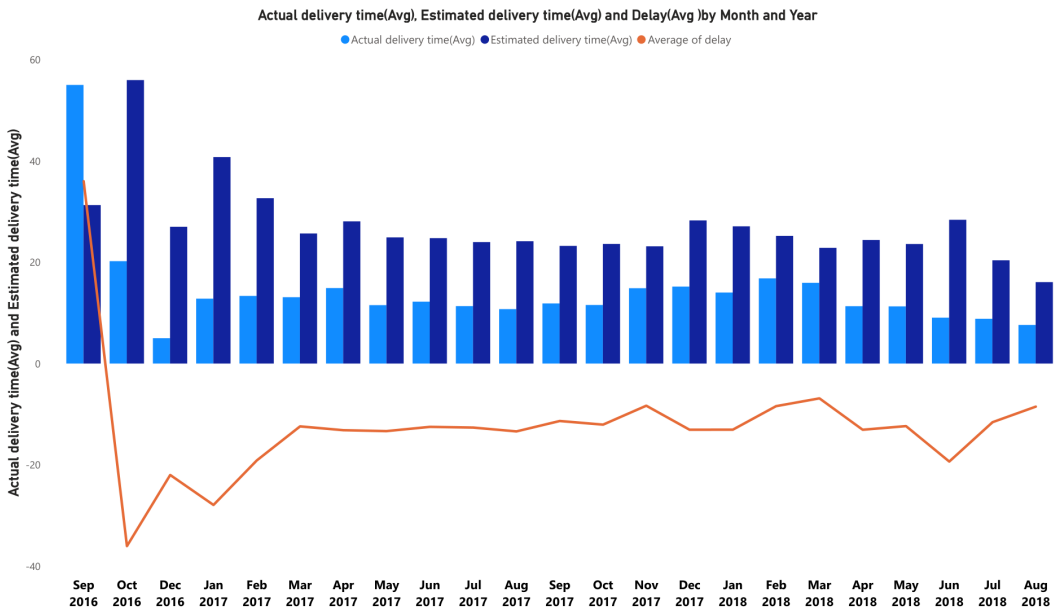


Figure 23: Average Actual Delivery Time, Estimated Delivery Time, and Delay by Month and Year

### Geographic Variation in Delivery Delays

The **below figure** illustrates the distribution of *average delivery delays* across Brazilian customer states. The states are arranged in descending order of delay. States such as AC and RO exhibit the highest negative delays, nearing -22 days, indicating earlier-than-estimated deliveries. In contrast, states like AL and MA show comparatively lower negative delays of approximately -10 days. This wide geographic variation implies significant logistical disparities, potentially influenced by regional infrastructure, urbanization, or warehouse proximity. Overall, all states demonstrate negative delay values, confirming a general trend of early deliveries.

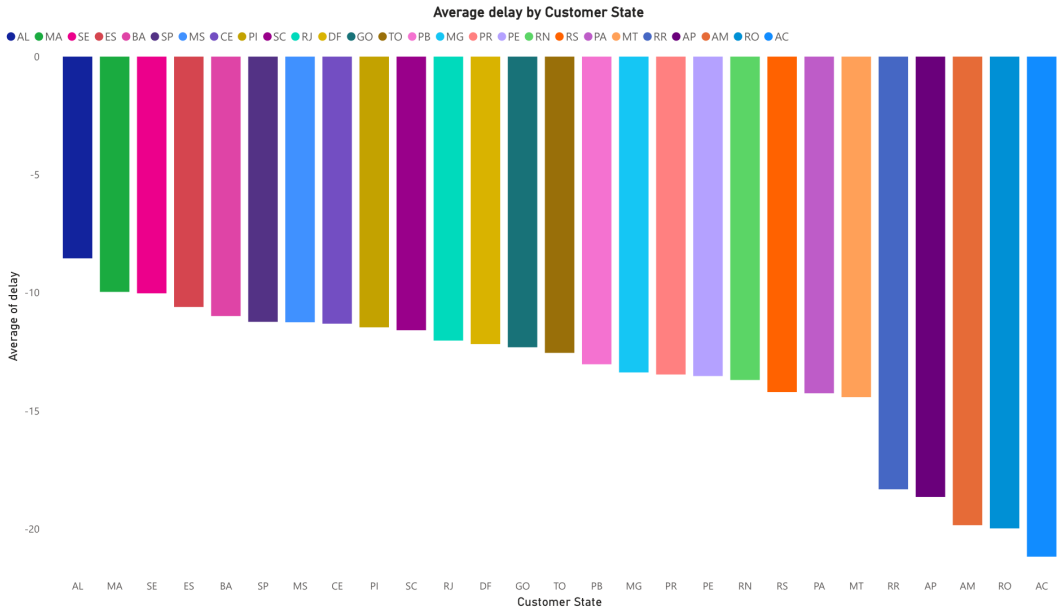


Figure 24: Average Delivery Delay by Customer State

### Delivery Time Metrics by State

The **below figure** examines state-wise comparison of *average actual delivery time* and *estimated delivery time*. Notably, states such as RR, AP, and AM show high estimated times (above 45 days) and actual times around 28 days, indicating substantial delivery advances. Meanwhile, states like SP and MG demonstrate lower delivery durations, with actual delivery times below 15 days. The data reflects a consistent trend of actual delivery times being shorter than the estimates across all states. This underscores the efficiency of logistical execution relative to projections, especially in densely populated or well-connected states.

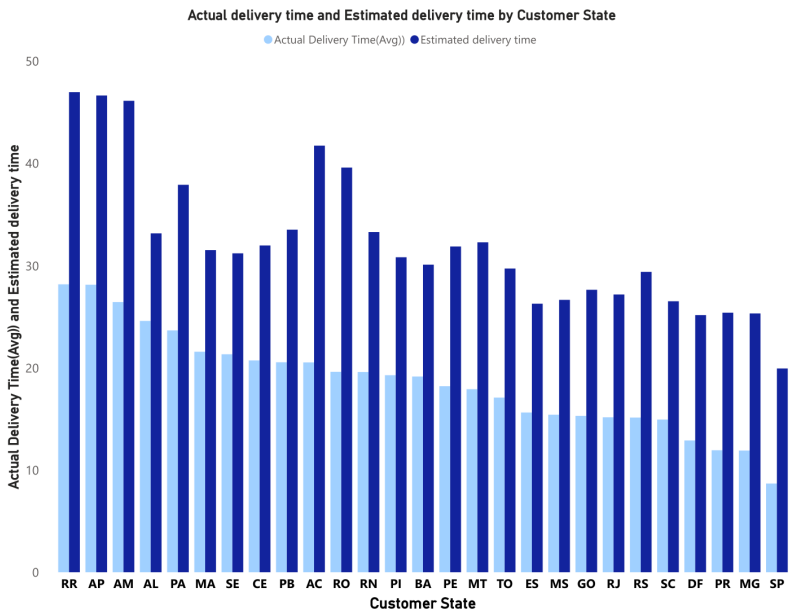


Figure 25: Actual and Estimated Delivery Time by Customer State

### Payment Value Distribution by State

The **below figure** presents a bar graph of average *payment value* by customer state. States like PB and AC exhibit the highest average payment values (above 280 units), whereas SP, MG, and ES reflect lower values, averaging around 150 units. This disparity may be linked to purchasing behavior, income levels, or product availability by region. Interestingly, many lower-population states exhibit higher per-transaction values, indicating concentrated high-value purchases or bulk orders. The visualization aids in identifying high-value regions for targeted marketing or service upgrades.

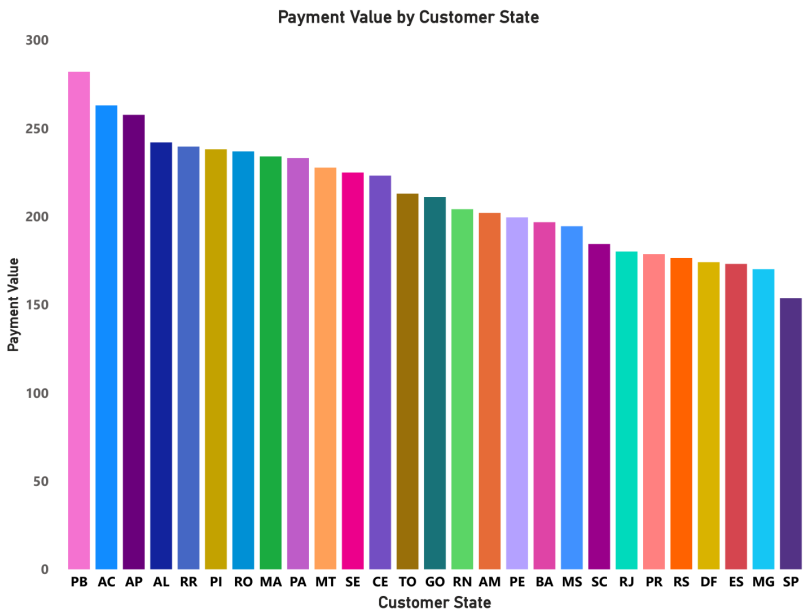


Figure 26: Payment Value by Customer State

### Monthly Trends in Order Volume

The **below figure** shows a time series line graph representing the *number of orders* placed monthly from September 2016 to August 2018. The trajectory reveals a rapid growth trend, peaking at over 9,000 orders in December 2017. Order volumes remained relatively stable between 6,000 and 9,000 orders monthly from January 2018 onward. This escalation in transaction volume over time reflects a scaling customer base and improved operational reach. Seasonal variations and promotional campaigns may explain periodic peaks, making this insight valuable for supply chain and inventory planning.

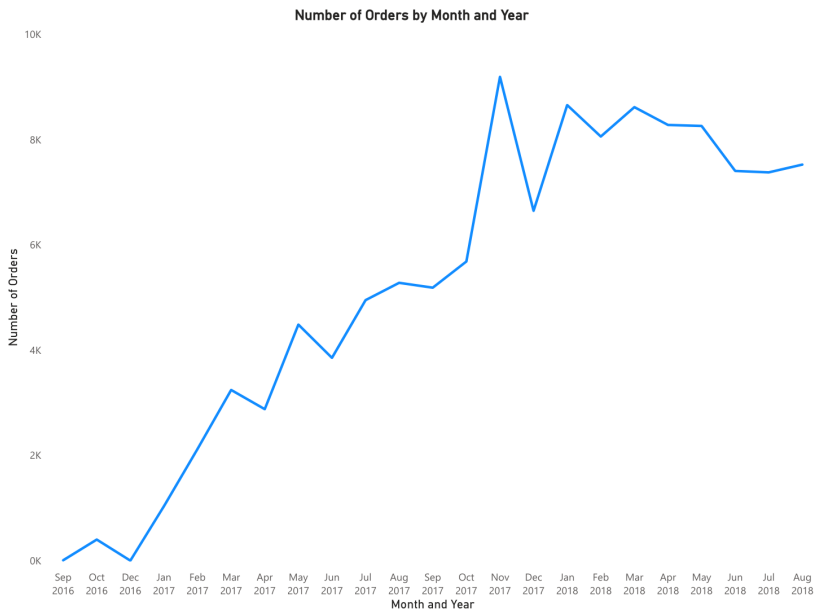


Figure 27: Number of Orders by Month and Year

### Order Status Distribution

The **below figure** visualizes the breakdown of *order statuses* using a donut chart. A dominant 97.13% of orders are marked as "delivered," followed by smaller proportions of "shipped," "canceled," "unavailable," "invoiced," and other statuses. The exceptionally high delivery rate underscores logistical reliability and customer fulfillment. However, the presence of canceled and unavailable orders, though under 2%, still warrants investigation into their root causes, such as stockouts or order rejections. This categorical breakdown is vital for assessing operational health and service quality.

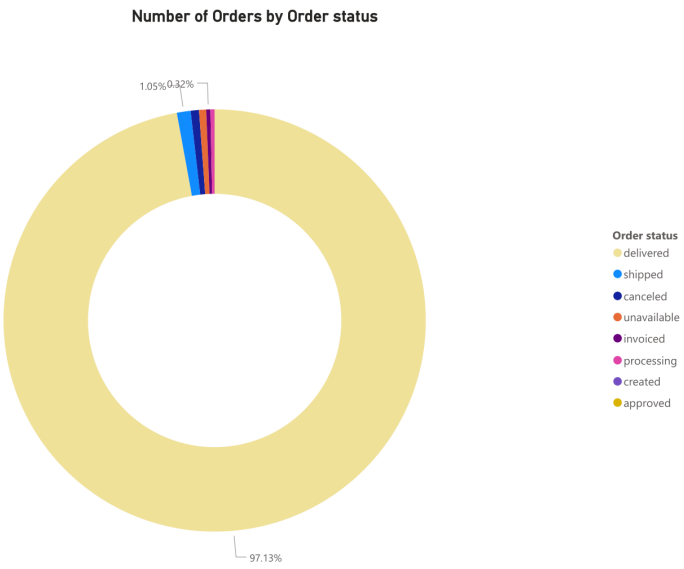


Figure 28: Distribution of Orders by Status

### City and State-level Order Distribution

The **below figure** visualizes the volume of orders by customer city and corresponding state. São Paulo leads with approximately 19,000 orders, followed by Rio de Janeiro with over 8,000, and Belo Horizonte with more than 3,000. Other notable cities include Brasília, Curitiba, and Campinas. The long-tail distribution indicates a high concentration of demand in urban centers, with smaller cities contributing modestly. This spatial pattern provides critical insights for logistics optimization, regional warehouse placement, and targeted customer engagement strategies.

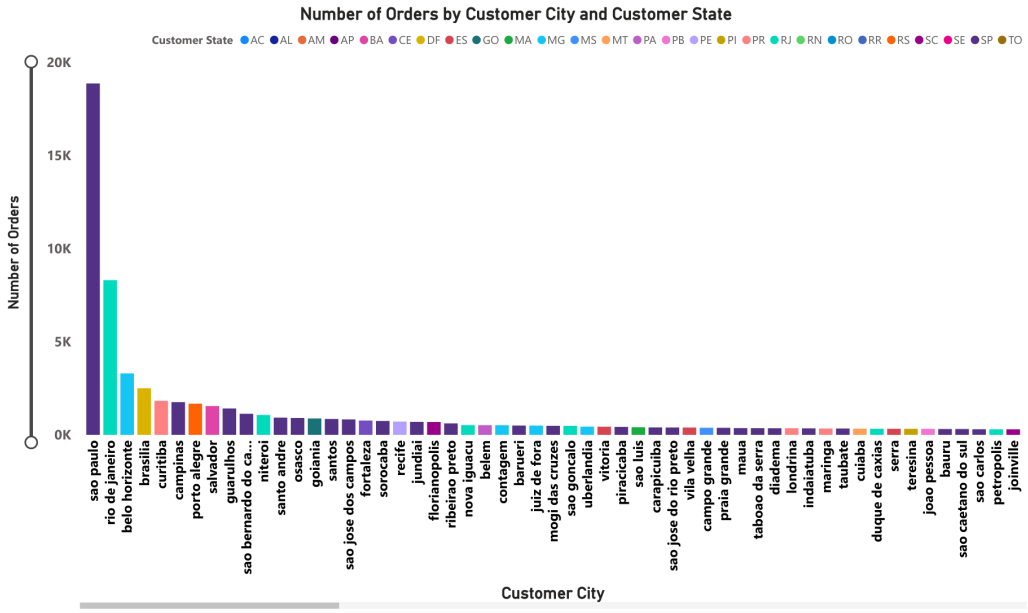


Figure 29: Number of Orders by Customer City and State

### Key Influencer Analysis: Payment Installments

The **below figure** performs a key influencer analysis on factors affecting *payment installments*. It reveals that when the *payment value* exceeds 220.85 units, the average installment count increases by 2.25. The visual clearly shows higher bars corresponding to higher payment bins, with payment value being a primary determinant of installment behavior. This insight suggests that high-value orders are often financed in more installments, which has implications for financial planning and consumer credit offerings.

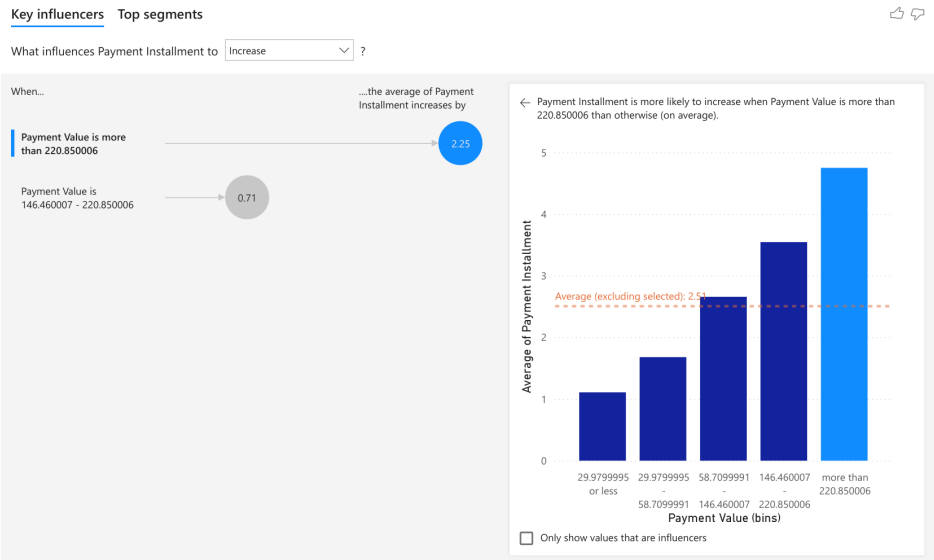


Figure 30: Influence of Payment Value on Payment Installments

### Average Delivery Duration Over Time

The **below figure** displays the trend of *average delivery duration* (in days) from September 2016 to August 2018. An initial spike exceeding 50 days in September 2016 quickly declines to under 10 days by December 2016. Subsequent fluctuations persist within the 10–15 day range. This overall decline in duration reflects significant operational improvements, possibly due to better courier performance, optimized routing, or system-level delivery time recalibrations. Such trend analysis is essential for continuous process improvement and customer satisfaction enhancement.

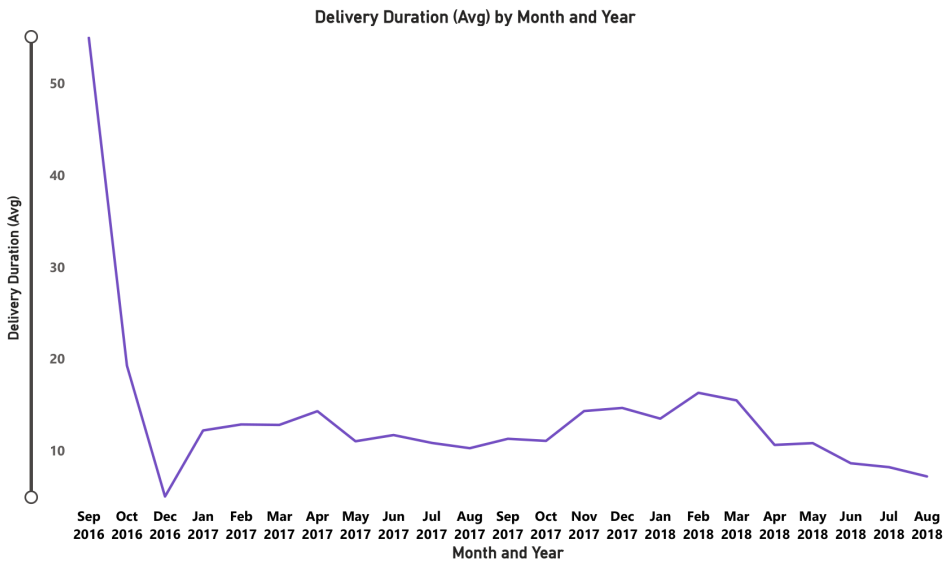


Figure 31: Average Delivery Duration by Month and Year

### Key Influencers of Payment and Review Score

The **below figure** depicts two key influencer analyses. The top chart shows that when *freight value* exceeds 38.47 units, the *average payment value* increases by 302.9 units. The bottom section highlights that the probability of a *blank review score* increases by a factor of 122.05x when the average payment value ranges between 2234.65991 and 2266.61011 units. This dual-influencer layout underlines the interconnectedness of transaction value, customer engagement, and feedback behavior. Higher freight and payment values are associated with both increased financial commitment and a lower likelihood of customers submitting reviews.

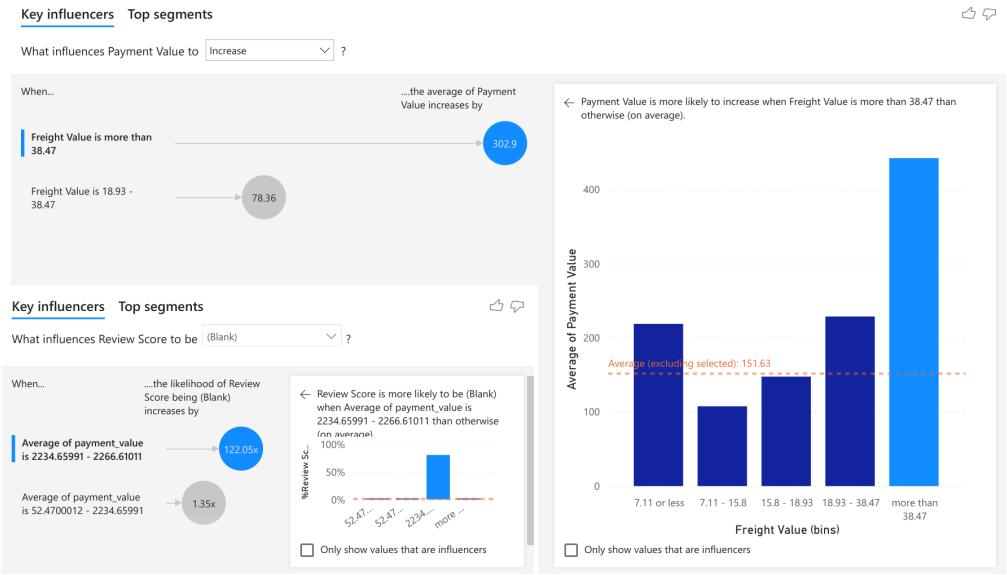


Figure 32: Key Influencers of Payment Value and Review Score

## Results and Conclusions

The advanced exploratory data analysis of the Olist e-commerce dataset yielded critical insights into delivery dynamics, geographic disparities, product-level inefficiencies, and the behavioral attributes of payment and customer engagement. These findings, derived through integrated analysis in Databricks, Power BI, and Azure Synapse, not only illuminate key operational bottlenecks but also suggest evidence-based recommendations for optimization.

Firstly, a comparative temporal analysis between *estimated delivery time* and *actual delivery time* revealed persistently early deliveries across the observed period from September 2016 to August 2018. While estimated times often exceeded 30 days, the actual delivery duration dropped to an average of 10–15 days in later periods. In October 2016, the average delay peaked negatively at -37 days, demonstrating systemic overestimation in delivery timelines. This behavior was consistent across most Brazilian states, with states like Acre (AC), Rondônia (RO), and Amazonas (AM) exhibiting average delays beyond -20 days, suggesting efficient logistics despite extended estimated times. Conversely, delays were less pronounced in urban hubs like São Paulo (SP), where actual delivery times averaged around 10 days.

Secondly, the correlation heatmap revealed strong positive relationships between *freight*

*value* and both *payment value* (r = 0.74) and *price* (r = 0.61), indicating that higher-cost and higher-value items incur proportionally higher logistics costs. Interestingly, *freight value* also showed a moderate negative correlation with *delivery delay* (r = -0.42), implying that costlier shipments are prioritized for faster delivery. In addition, a negative correlation between *product_photos_qty* and *delay* (r = -0.03) implies that products with more visual documentation are subject to fewer delays—likely due to better inventory handling, reduced ambiguity in fulfillment, or enhanced customer trust leading to lower returns and smoother processing.

Thirdly, product category analysis identified *bed_bath_table*, *health_beauty*, and *computers_accessories* as the top three revenue contributors, with cumulative sales exceeding 48 million BRL. These categories also dominated in terms of customer engagement, with *bed_bath_table* alone accumulating 11,847 reviews. On the contrary, categories such as *fashion_female_clothing* and *cine_photo* exhibited average delays exceeding 11 days, making them high-risk categories from a service level agreement (SLA) perspective. Targeted improvements in packaging, documentation, and supply chain visibility for these categories could significantly mitigate delays and improve customer satisfaction.

Furthermore, temporal trends in delivery performance revealed a marked decline in delivery duration over time, from over 50 days in late 2016 to under 10 days by mid-2018. This improvement coincided with a rise in monthly order volume, which peaked at over 9,000 orders in December 2017. Despite this increase, the fulfillment rate remained impressive, with 97.13% of all orders marked as "delivered," reflecting strong scalability in the fulfillment infrastructure.

From a behavioral standpoint, credit card payments were dominant (87,776 transactions), far surpassing boleto (23,190) and vouchers (6,465). Key influencer analysis revealed that higher *payment values* (over 220.85 BRL) led to a rise in *payment installments*, averaging up to 5 payments. Additionally, when *freight value* exceeded 38.47 BRL, *payment value* increased by 302.9 BRL, highlighting a clear link between logistics expenditure and transaction value. Notably, very high payment values (above 2,234.66 BRL) were associated with a 122.05x increase in blank review scores, indicating a possible customer dropout or dissatisfaction pattern.

In conclusion, this analytical investigation substantiates the reliability and scalability of the Azure-based big data pipeline (ADF, ADLS, Databricks, Synapse, Power BI) employed throughout the project. It is recommended that Olist adopts dynamic freight pricing strategies, improves predictive delivery models, enhances product metadata (especially image quantity), and tailors interventions based on regional logistics performance. These insights also provide a foundation for predictive modeling initiatives to proactively flag high-risk deliveries and optimize end-to-end customer experience. The evidence-driven methodology adopted here demonstrates significant potential for elevating operational efficiency and consumer trust in the Brazilian e-commerce ecosystem.

# References

Chapman, G. (2016). *Data analysis for business, economics, and policy.* Cambridge: Springer.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188.

Docs, M. (2023). *Azure data factory documentation.* Online. (Available at: `https://learn.microsoft.com/en-us/azure/data-factory/introduction`)

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

(Docs, 2023) (Chapman, 2016) (Chen, Chiang, & Storey, 2012) (Gandomi & Haider, 2015)