Third project Brazilian E-Commerce Public Dataset

Jeisson Steve Rojas Velasquez

June 2025.

Introduction

The purpose of this project is to analyze data from Brazilian e-commerce, based on customer orders placed through Olist Store across different online platforms. The analysis includes both categorical and numerical variables to better understand customer behavior, product sales, and other key aspects of the business. The main goal is to discover useful insights that can help improve sales and overall performance by looking at all these variables together in a more complete and connected way.

The project focuses on identifying the most sold and requested product categories, as well as the Brazilian states with the highest number of purchases. It also includes customer segmentation, analysis of sellers by their location, and evaluation of payment methods and installment plans. In addition, the analysis covers sales overtime, delivery performance, geographic trends, and customer satisfaction based on reviews—especially how reviews are related to delivery outcomes. Finally, a time-based analysis is carried out to explore order patterns, completed and canceled deliveries, and spending behavior from 2016 to 2018.

Database

The database used for this project was:

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

- ✓ This database is composed of the next tables:
- ✓ olist_customers_dataset.csv
- ✓ olist_geolocation_dataset.csv
- ✓ olist_order_items_dataset.csv
- ✓ olist_order_payments_dataset.csv
- ✓ olist_order_reviews_dataset.csv
- ✓ olist_orders_dataset.csv
- ✓ olist_products_dataset.csv
- ✓ olist_sellers_dataset.csv
- ✓ product_category_name_trans

Methodology

Methodology

The project was developed in three main phases: data preparation, including ETL, Exploratory Data Analysis, and result visualization.

1. **Data Preparation:**

In Excel, an initial cleaning and transformation of the data was performed. In the next steps

- ✓ Using Power Query, the data was cleaned by fixing errors, replacing null and blank values, and removing unnecessary columns. Tables were also merged using primary keys to optimize the database for analysis. In addition, data formats such as dates, numbers, and currency were adjusted to improve the quality and accuracy of the analysis.

- ✓ After that, Created power pivot tables to make a EDA, loaded tables in power pivot to create measures and calculated columns.

- ✓ Finally, the analysis and EDA, visualizations were created to highlight the most relevant data. Pivot tables were used as the main tool to generate charts, and shapes and design elements were added to create a visually appealing presentation. The goal was to communicate the insights clearly and show the value of the data for future decision-making, improving sales, and overall business performance.

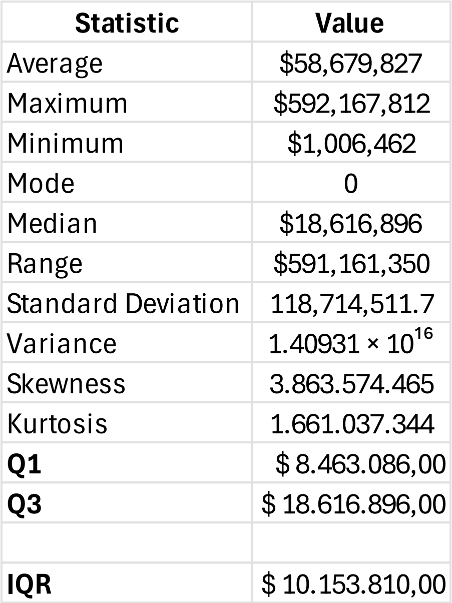
- ✓ Creating a dashboard in Excel is essential because it allows users to interact with data in a dynamic and organized way. It also helps summarize complex information visually, making it easier to monitor key metrics and support data-driven decisions.

On the other hand, for the SQL analysis, tables were imported from .csv files and cleaned to ensure consistency and reliability. Primary and foreign keys were created to define relationships between tables, data types were adjusted for accuracy, and constraints were applied to maintain data integrity. Several analytical queries were performed, including sales trends by product category, revenue by customer state, and order tracking by status and time. In addition, stored procedures were developed to calculate total sales by month and by payment type, and triggers were implemented to automate certain validations, such as flagging canceled orders or tracking late deliveries. This helped create a solid and optimized relational model, essential for extracting insights and supporting decision-making based on real business scenarios.

2. **Exploratory Data Analysis (EDA):**

- In Excel, pivot tables were used, and descriptive statistical analysis techniques were applied to key variables eda for every analisis like sales eda, customer eda, payment eda, product eda, sellers eda, geographic eda, review eda, temporal eda. For eda I calculated values such as: Promedio, Max, Min, Moda, Mediana, Rango, Desviacion estandar, Varianza, Asimetria. Curtosis

Exploratory Data Analysis for sales



Statistic	Value
Average	$58,679,827
Maximum	$592,167,812
Minimum	$1,006,462
Mode	0
Median	$18,616,896
Range	$591,161,350
Standard Deviation	118,714,511.7
Variance	1.40931×10^{16}
Skewness	3.863.574.465
Kurtosis	1.661.037.344
Q1	$ 8.463.086,00
Q3	$ 18.616.896,00
IQR	$ 10.153.810,00

Sales eda analisis

The sales variable shows a distribution with high variability, which means that sales changed a lot between the different months. In general, the average sales amount is **$58,679,827**, while the median is **$18,616,896**. The big difference between these two values suggests that the data is **left-skewed** — in other words, most months had lower sales, but there were a few months with very high sales that increased the average.

This can also be seen in the **sales range**, which goes from a minimum of **$1,006,462** to a maximum of **$592,167,812**, giving a total range of **$591,161,350**. This shows that sales changed a lot from month to month, which means the distribution is spread out and there were some peaks in certain months.

The **interquartile range (IQR)** is **$10,153,810**, and it also helps us understand the spread between the middle values of sales. Even though it's not too far from the average, it still shows that the middle values also have a lot of variation.

On the other hand, the **standard deviation** is **$118,714,511**, which means that the sales values are very far from the average. This supports the idea that sales don't follow a stable pattern and they change a lot each month.

The **skewness** is **3.86**, which means the sales distribution has a tail to the left. This tells us that there are a few months with extremely high sales, and they are affecting the shape of the distribution. Most months are closer to the lower values.

Finally, the **kurtosis** is **16.61**, which means the distribution is **leptokurtic**. In simple words, it has more extreme values than a normal distribution. This is related to the very high sales in some months, which happen more often than expected in a more regular distribution.

Exploratory Data Analysis for Customers

Statistic	
Average	$59.292.119
Maximum	$599.822.696
Minimum	$1.006.462
Mode	0
Median	$18.702.929
Range	$598.816.234
Standard Deviation	120160845,3
Variance	1,44386E+16
Skewness	3,872731654
Kurtosis	16,68514924
Q1	$ 8.610.416
Q3	$ 48.589.345
IQR	$ 39.978.930

Customer eda analisis

Customer expenses by state show a high level of variability, which means that the amounts spent were not consistent during the analyzed period. In general, the average expense is

$59,292,119, while the median is $18,702,929. The big difference between these two values suggests that the data is left-skewed — this means that most months had lower spending, but a few months had very high expenses that increased the overall average.

The range of expenses goes from a minimum of $1,006,462 to a maximum of $599,822,696, giving a total range of $598,816,234. This wide range shows that spending changed a lot between 2016 and 2018, according to the available data. It also suggests the presence of peaks and irregular spending in some months.

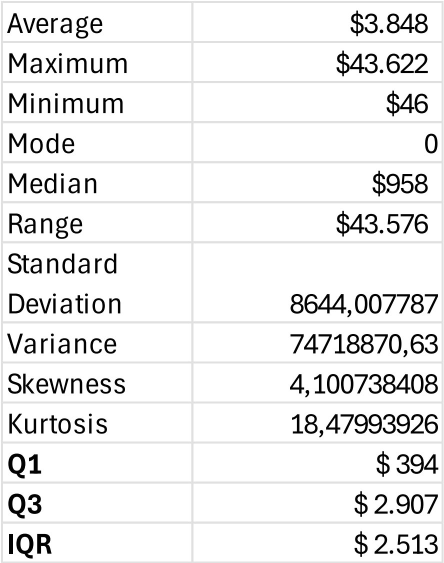
The interquartile range (IQR), which is $39,978,930, also shows the level of variation in the central values. Even though it is not very far from the average, it still indicates that the middle part of the data has noticeable differences.

We also observe the standard deviation, which is $120,160,845. This tells us that the values are very spread out around the average, and supports the idea that spending does not follow a stable trend, but changes a lot from one month to another.

The skewness is 3.87, which confirms that the distribution is skewed to the left. This means that there are some months with very high spending, while most other months show lower values. These few high values affect the shape of the distribution and make it unbalanced.

Finally, the kurtosis is 16.85, which suggests a leptokurtic distribution. In simple words, this means that there are more extreme values than we would normally expect. This is related to the very high expenses in specific months, which happened more often than usual in a standard distribution.

Exploratory Data Analysis for payment



Average	$3.848
Maximum	$43.622
Minimum	$46
Mode	0
Median	$958
Range	$43.576
Standard Deviation	8644,007787
Variance	74718870,63
Skewness	4,100738408
Kurtosis	18,47993926
Q1	$ 394
Q3	$ 2.907
IQR	$ 2.513

Payment eda analisis

For the payment analysis, the variable used was the number of installments by customer state, which shows how often customers used financing options for their purchases. This variable shows a high level of variability, meaning there are significant differences in the number of installment payments across different states.

In general, the average number of installments per state was 3,847.63, while the median was 958.00. This big difference suggests that the distribution is right-skewed (not left-skewed), meaning that most states used fewer installments, but a few states had very high values that increased the average.

The range of installments goes from a minimum of 46.00 to a maximum of 43,622.00, giving a total range of 43,576.00. This wide gap confirms the data is very spread out, with some clear peaks in certain states during the years 2016 to 2018.

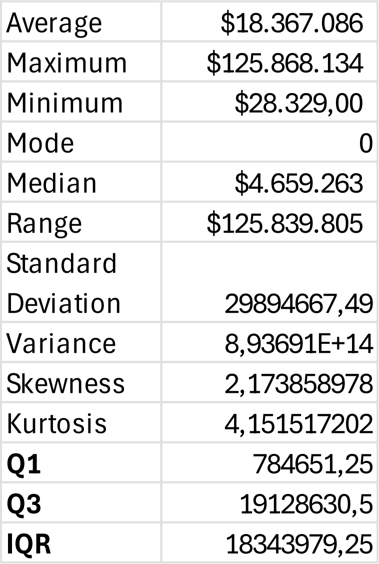
The interquartile range (IQR) was 2,513.00, which also shows that there is a lot of variation even in the middle 50% of the data.

In addition, the standard deviation was 8,644.01, meaning the values are widely spread around the mean, and that there is no stable or regular trend in how many installments were used in each state.

The skewness was 4.10, which confirms that the distribution is strongly right-skewed. This means a few states had a very high number of installments, while most had much lower values.

Finally, the kurtosis was 18.48, suggesting a leptokurtic distribution, which means there are many extreme values. This tells us that some states had a very different financial behavior compared to the overall average.

Exploratory Data Analysis for product



Average	$18.367.086
Maximum	$125.868.134
Minimum	$28.329,00
Mode	0
Median	$4.659.263
Range	$125.839.805
Standard Deviation	29894667,49
Variance	8,93691E+14
Skewness	2,173858978
Kurtosis	4,151517202
Q1	784651,25
Q3	19128630,5
IQR	18343979,25

Product eda analisis

For this analysis, the variable total revenue per product was examined to understand how sales are distributed across different product categories. This metric helps to assess which products contribute most significantly to the overall revenue and reveals key disparities in product performance.

The average revenue per product was $18,367,086, while the median value was $4,659,263. The large gap between the mean and median suggests that the data is not evenly distributed and that a small number of products earned exceptionally high revenue, pulling the average upwards.

The maximum revenue observed was $125,868,134, and the minimum was just $28,329, resulting in a range of $125,839,805. This large range highlights major differences in how products perform in terms of sales, indicating that only a few products dominate the revenue, while many others have relatively low contributions.

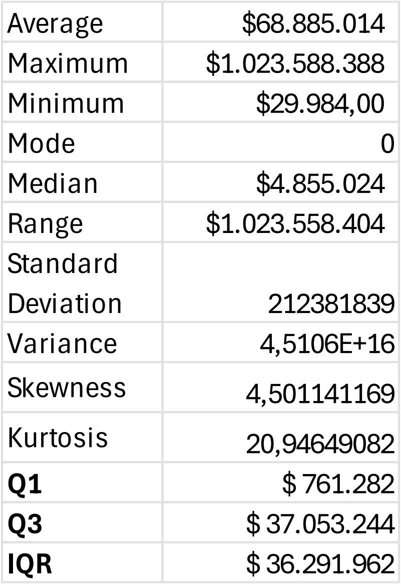
The interquartile range (IQR) was $18,343,979.25, with a first quartile (Q1) of $784,651.25 and a third quartile (Q3) of $19,128,630.5. This shows that even among the middle 50% of products, there is a considerable variation in revenue levels.

The standard deviation was $29,894,667.49, which reflects a high level of dispersion in the data. This means that revenue values for different products vary widely from the mean, making it difficult to describe a “typical” product in terms of sales performance.

The skewness was 2.17, which indicates a moderate right-skewed distribution. In other words, most products earned below-average revenue, while a few generated significantly higher amounts.

Finally, the kurtosis was 4.15, indicating a leptokurtic distribution. This suggests the presence of extreme values or outliers that are much higher than the rest, which may correspond to best-selling or highly popular products dominating the market.

Exploratory Data Analysis for sellers



Average	$68.885.014
Maximum	$1.023.588.388
Minimum	$29.984,00
Mode	0
Median	$4.855.024
Range	$1.023.558.404
Standard Deviation	212381839
Variance	4,5106E+16
Skewness	4,501141169
Kurtosis	20,94649082
Q1	$ 761.282
Q3	$ 37.053.244
IQR	$ 36.291.962

Seller eda analisis

For this analysis, the variable total revenue by orders grouped by seller state was used to observe the sales performance across the different states where sellers operate. This variable allows us to evaluate the geographical impact of sellers on total revenue, and it shows a highly variable and skewed distribution.

In general terms, the average revenue per state was $68,885,014, while the median was $4,855,024, a very significant difference. This gap indicates that the distribution is strongly right-skewed, meaning that most states generated relatively low revenue, but a few states had extremely high values, which increased the overall average.

The total revenue range was $1,023,558,404, with a minimum value of $29,984 and a maximum value of $1,023,588,388. This wide range shows a large disparity among states, suggesting that sales volume is not equally distributed among sellers in the country.

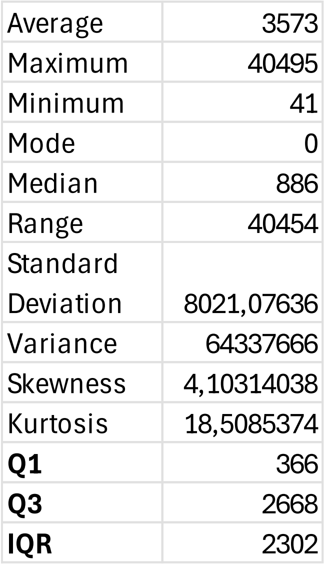
On the other hand, the interquartile range (IQR) was $36,291,962, with a Q1 of $761,282 and a Q3 of $37,053,244. This shows that even within the middle 50% of the distribution, there is a considerable variation in revenue between states.

The standard deviation was $212,381,839, which is very high and reinforces the idea that the data is widely spread around the mean. This makes it difficult to define a general average behavior for all seller states.

Regarding skewness, the value was 4.50, confirming that the distribution is highly skewed to the right. This means that a small number of states concentrate a large part of total revenue, while most states generated significantly lower amounts.

Finally, the kurtosis was 20.95, indicating that the distribution is leptokurtic, meaning there is a strong presence of extreme values. This pattern is associated with the existence of positive outliers, which could represent major commercial centers or economically active states that dominate sales.

Geographic Exploratory Data Analysis



Average	3573
Maximum	40495
Minimum	41
Mode	0
Median	886
Range	40454
Standard Deviation	8021,07636
Variance	64337666
Skewness	4,10314038
Kurtosis	18,5085374
Q1	366
Q3	2668
IQR	2302

Geographic eda analisis

This analysis focuses on the variable delivery time in days, grouped by customer state, with the goal of identifying regional differences in delivery performance and evaluating logistics efficiency across different parts of the country. This metric is important for understanding the customer experience and detecting possible distribution issues.

The average delivery time per state was 3,573 days, while the median was 886 days. The big gap between the mean and median shows that the data is highly skewed, with a few states having extremely long delivery times that raise the overall average.

The maximum delivery time recorded was 40,495 days, and the minimum was 41 days, giving a total range of 40,454 days. This extreme variation suggests that there may be states with serious delivery delays or unusual data points, possibly caused by errors or exceptional situations.

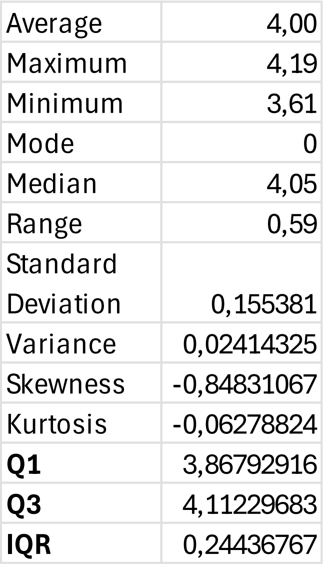
The interquartile range (IQR) was 2,302 days, with Q1 at 366 days and Q3 at 2,668 days. This means that even in the middle 50% of the distribution, there is a large difference in delivery times between states.

The standard deviation was 8,021 days, which shows that the values are very spread out from the average. This makes it difficult to define a common delivery pattern for all states, as the results vary a lot from region to region.

The skewness was 4.10, confirming that the distribution is strongly skewed to the right. This means most states had relatively short delivery times, while a small number had very high values.

Finally, the kurtosis was 18.51, which indicates a leptokurtic distribution. In other words, there are extreme outliers, especially high delivery times, that significantly affect the overall distribution. These outliers might represent states with special delivery problems or unusual cases.

Review Exploratory Data Analysis



Average	4,00
Maximum	4,19
Minimum	3,61
Mode	0
Median	4,05
Range	0,59
Standard Deviation	0,155381
Variance	0,02414325
Skewness	-0,84831067
Kurtosis	-0,06278824
Q1	3,86792916
Q3	4,11229683
IQR	0,24436767

Review eda analisis

This analysis uses the variable average review score grouped by customer state to identify potential satisfaction patterns across different regions. This metric reflects how customers perceive their experience, which can be influenced by delivery, product quality, or customer service.

The average review score across all states was 4.00, while the median was slightly higher at 4.05. This small difference suggests a slight negative skew, meaning some states had lower review scores that pulled the average down a bit.

The maximum score recorded was 4.19, and the minimum was 3.61, giving a total range of 0.59. Although the scale seems narrow, these small differences can be meaningful when measuring customer satisfaction, as scores closer to 5 indicate better service or product experiences.

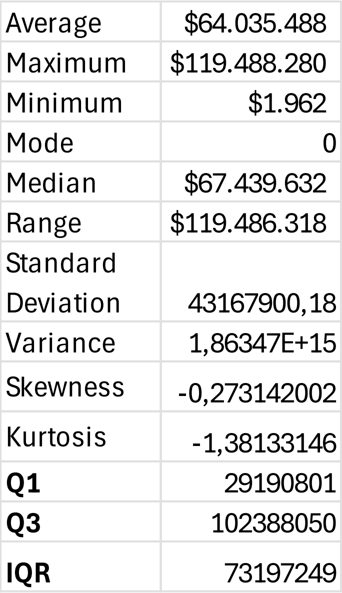
The interquartile range (IQR) was 0.244, with Q1 at 3.87 and Q3 at 4.11. This shows that the central 50% of states had fairly similar review scores, suggesting a certain consistency in customer satisfaction across regions.

The standard deviation was 0.155, a low value that indicates the scores are closely clustered around the mean. This reinforces the idea that review scores are generally stable and do not vary much from state to state.

The skewness was -0.85, which confirms that the distribution is negatively skewed, meaning a few states had lower scores compared to the rest.

Finally, the kurtosis was -0.06, which suggests the distribution is approximately normal (mesokurtic), with few extreme values. In other words, most review scores fall within a common range, with no unusually high or low outliers.

Temporal Exploratory Data Analysis



Average	$64.035.488
Maximum	$119.488.280
Minimum	$1.962
Mode	0
Median	$67.439.632
Range	$119.486.318
Standard Deviation	43167900,18
Variance	1,86347E+15
Skewness	-0,273142002
Kurtosis	-1,38133146
Q1	29190801
Q3	102388050
IQR	73197249

Temporal eda analisis

This analysis is based on the variable Total Spending by Month and Year, with the purpose of identifying how much money was spent between 2016 and 2018. This metric helps reveal both the spending frequency and the total amount spent by customers over time.

The average monthly spending was around $64,035,048, while the median was slightly higher, at $67,439,632. This small difference suggests a slight negative skew, meaning that some months had lower spending values that pulled the mean down.

The maximum amount spent in a single month was $119,488,280, and the minimum was only $1,962, resulting in a total range of $119,486,318. This wide gap reflects strong variability in customer spending, which could be influenced by seasonal sales or special events.

The interquartile range (IQR) was $73,197,249, with a Q1 of $29,190,801 and a Q3 of $102,388,050. This shows that the middle 50% of monthly spending values were also spread out, indicating some inconsistency in customer behavior during the period.

The standard deviation was $43,167,900, which is quite high. This confirms that spending values were widely spread out around the average, making it difficult to define a typical monthly spending pattern.

The skewness value was -0.27, indicating a slight left skew. In this case, a few months with lower-than-average spending caused the distribution to lean slightly to the left.

Finally, the kurtosis was -1.38, which means the distribution is platykurtic. This type of distribution is flatter than a normal curve, with fewer extreme values and more evenly distributed data across the months.

3. **Visualization and Insight Generation:**

To provide insights and present the findings during this process, Excel and Power BI were used for visualization. The analysis was divided into eight dashboards:

- ✓ Sales Dashboard
- ✓ Customer Dashboard
- ✓ Payment Dashboard
- ✓ Product Dashboard
- ✓ Seller Dashboard
- ✓ Geographic Dashboard
- ✓ Review Dashboard
- ✓ Temporal Dashboard

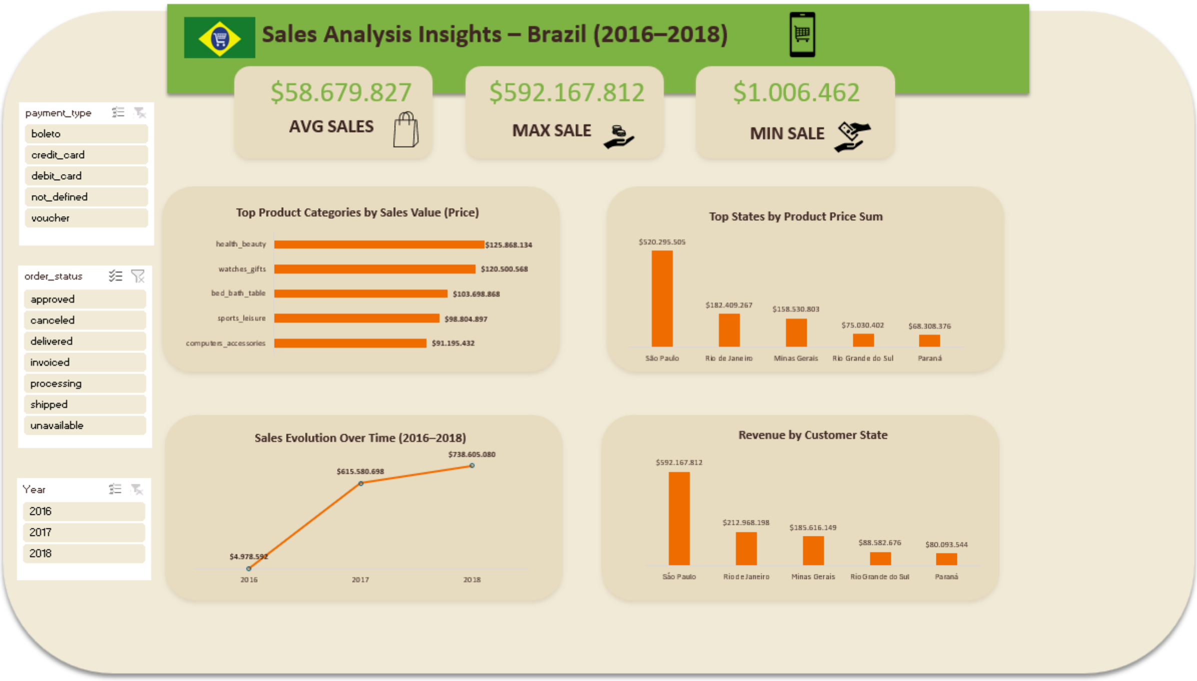
This structure was based on the variables found in the database, the type of data available, and the different analyses carried out.

These dashboards allow for customer segmentation, product description, identification of states and top-selling products, seller performance, and delivery optimization, among other insights.

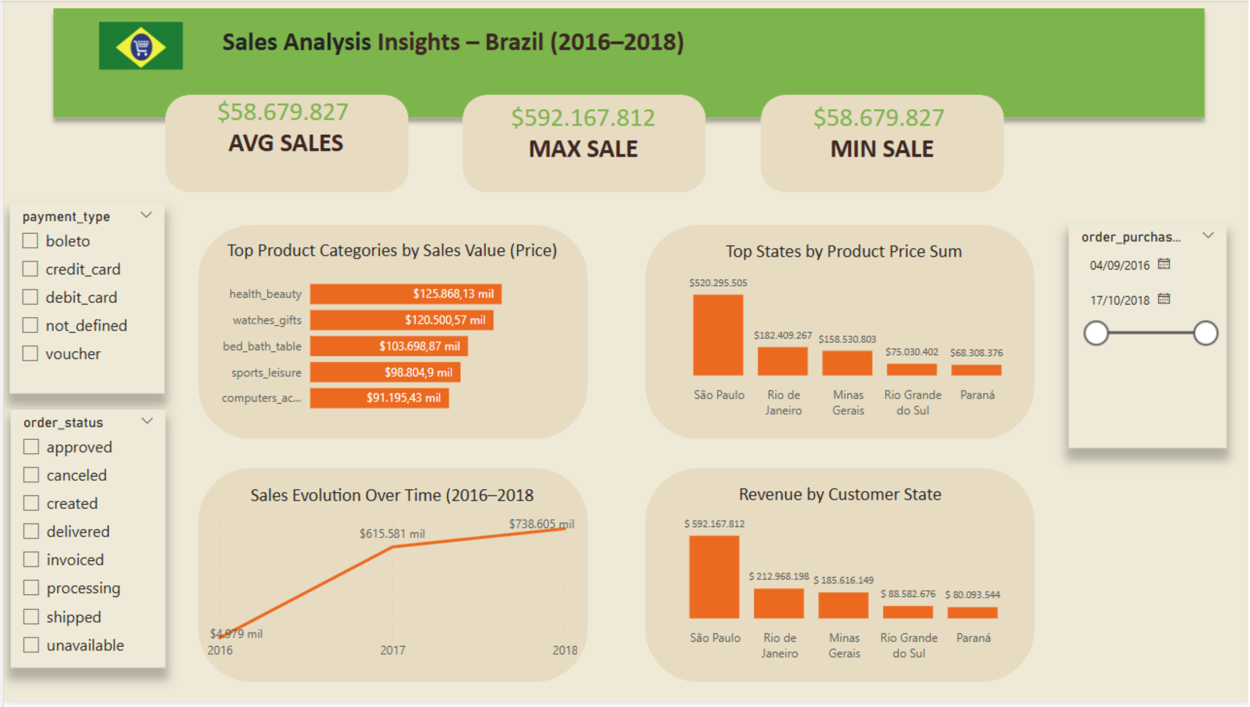
Dashboards

Sales Dashboard

Excel version



Power bi versión



Sales Dashboard Findings

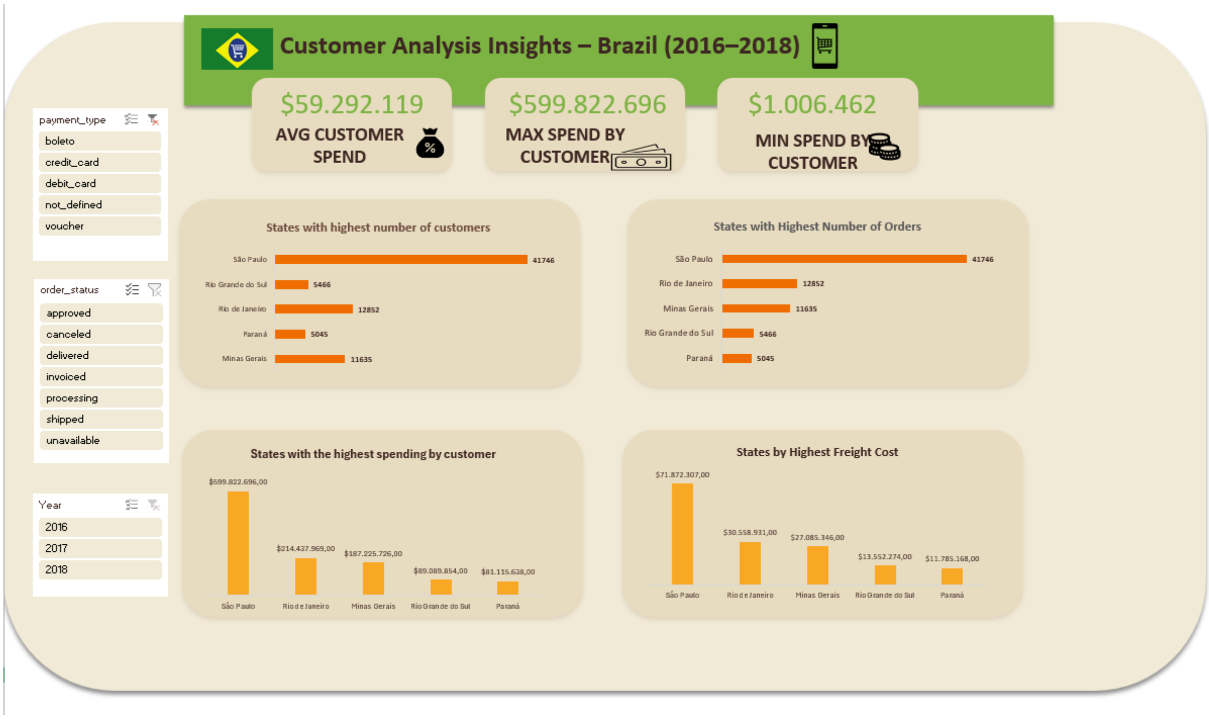
Regarding the sales analysis, the dashboard shows that the most sold product category was Health & Beauty, with total sales of $125,868,134.

Additionally, São Paulo recorded the highest total product prices, reaching $520,295,505, and also had the highest customer revenue by state, with a total of $592,167,812.

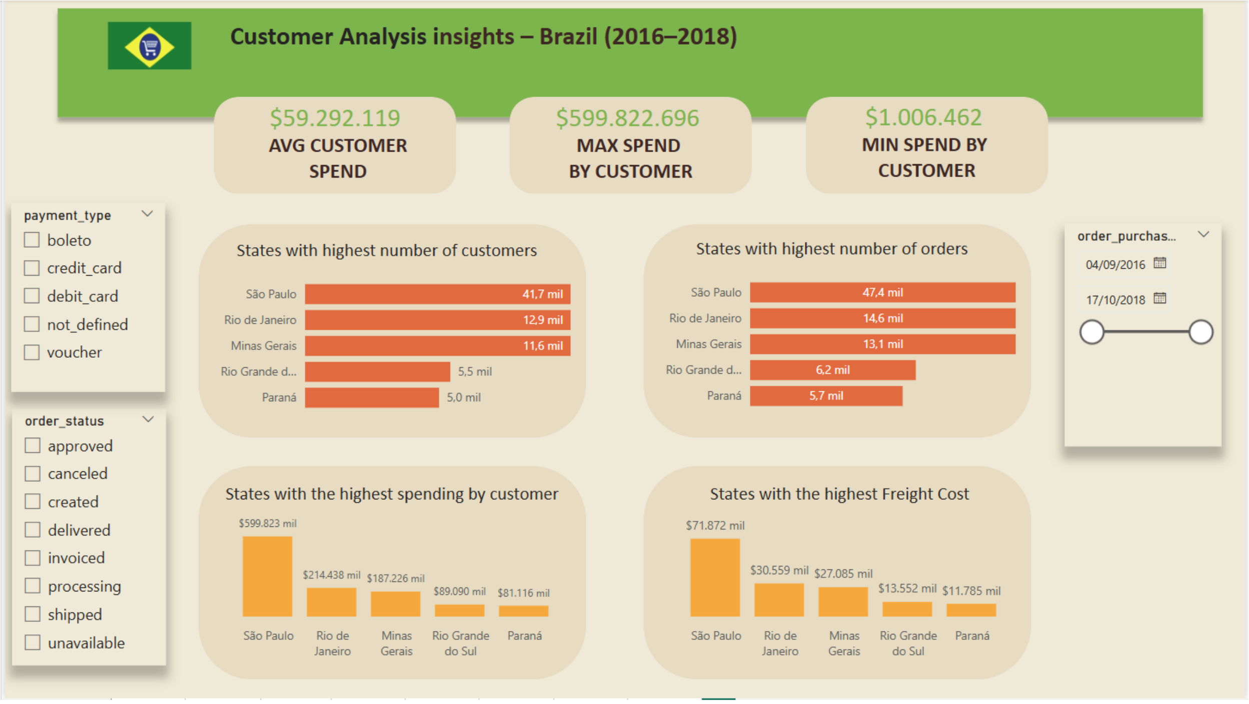
Finally, the dashboard reveals a remarkable sales growth, increasing from $4,978,592 in 2016 to $738,605,080 in 2018.

Customer Dashboard

Excel version



Power BI versión



Customer Dashboard Findings

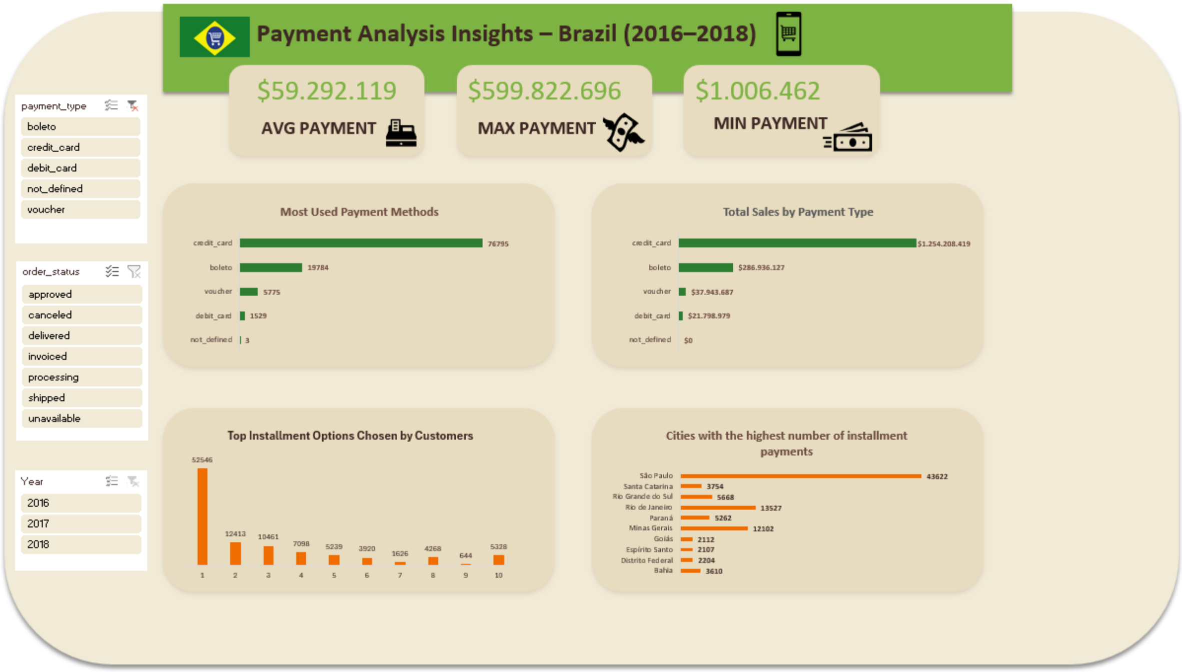
On the other hand, the customer-focused dashboard shows that, once again, São Paulo is the state with the highest number of customers, totaling 41,746. It is also the state with the highest number of orders, also 41,746, which suggests that the database recorded one order per customer.

In addition, São Paulo had the highest customer spending, with a total of $599,822,696.00, and also reported the highest freight cost, reaching $71,872,307.00.

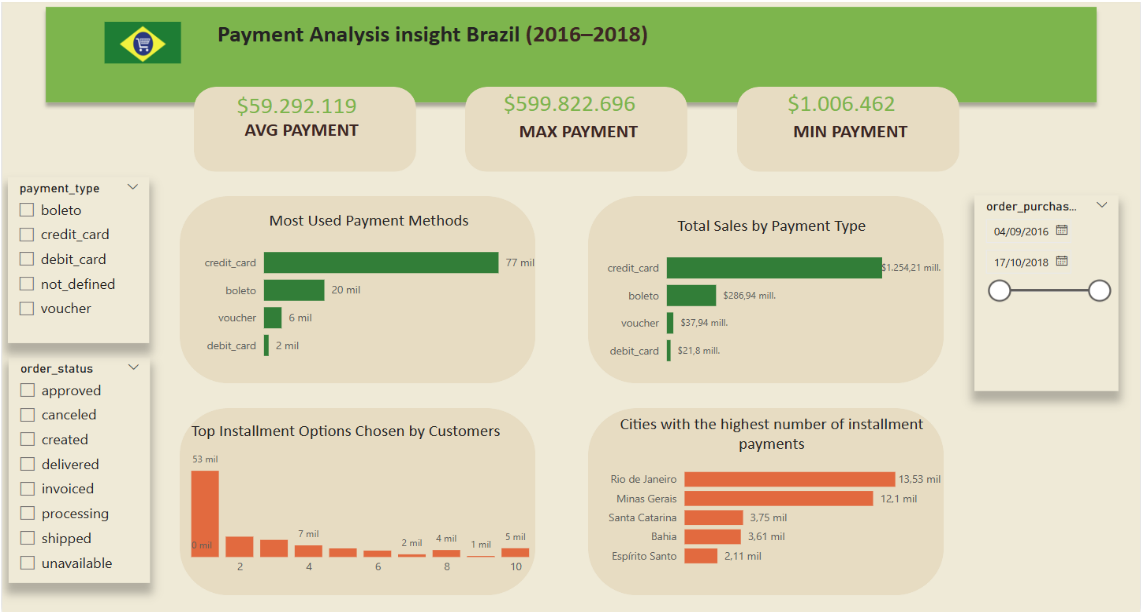
Excel Version

Payment Dashboard

Excel version



Power Bi versión



Payment Dashboard Findings:

Now, moving on to the payment analysis, the dashboard shows that credit card was the most used payment method, with a total of 76,795 transactions.

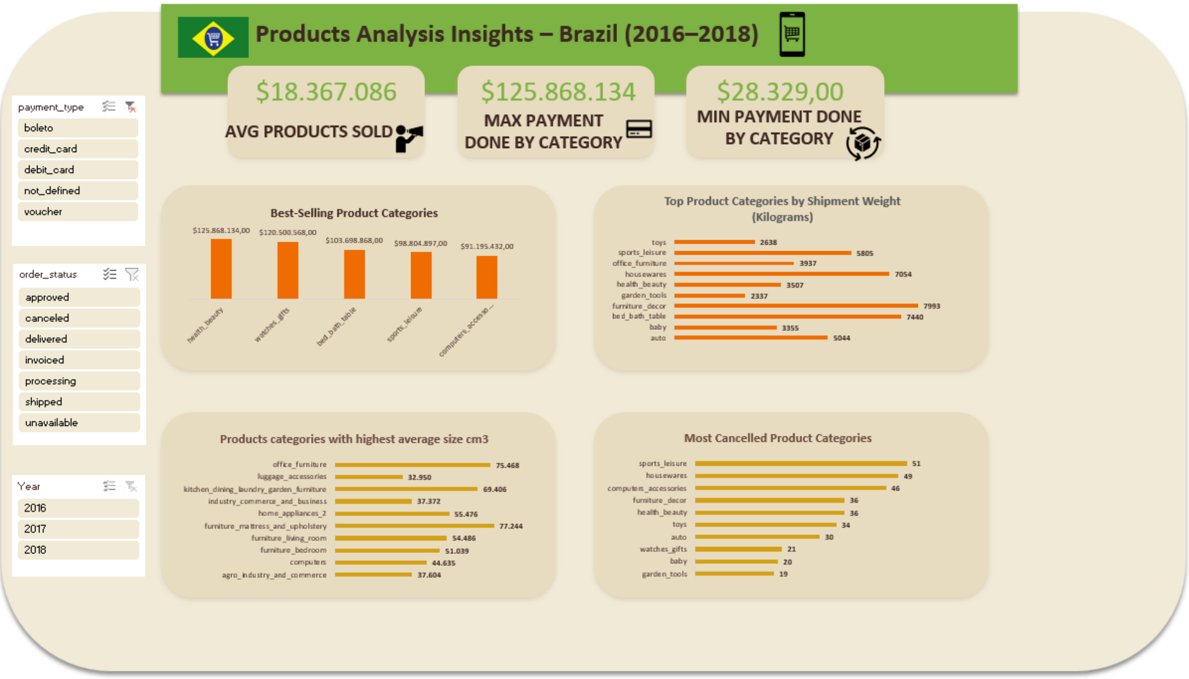
Likewise, the highest sales were also made using this method, collecting a total of $1,254,208,419 through credit card payments.

It was also found that most people preferred to pay in a single installment, with 52,546 transactions, compared to only 5,328 who chose to pay in 10 installments.

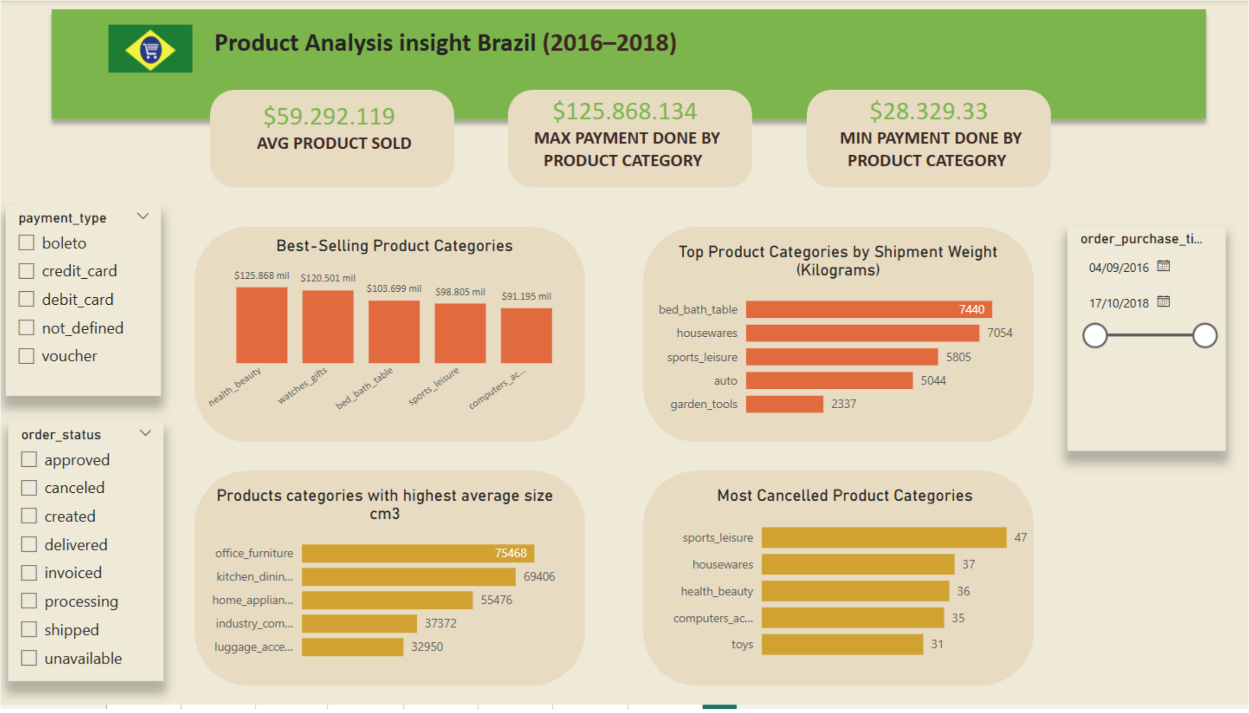
Finally, São Paulo stands out again as the state with the highest number of installment payment requests, totaling 43,622 orders.

Product Dashboard

Excel Version



Power bi version



Product Analysis insight Brazil (2016–2018)

$59.292.119
AVG PRODUCT SOLD

$125.868.134
MAX PAYMENT DONE BY PRODUCT CATEGORY

$28.329.33
MIN PAYMENT DONE BY PRODUCT CATEGORY

Product Dashboard Findings

By reviewing the product dashboard, we can see that the best-selling product category was Health & Beauty, with total sales of $125,868,134.00.

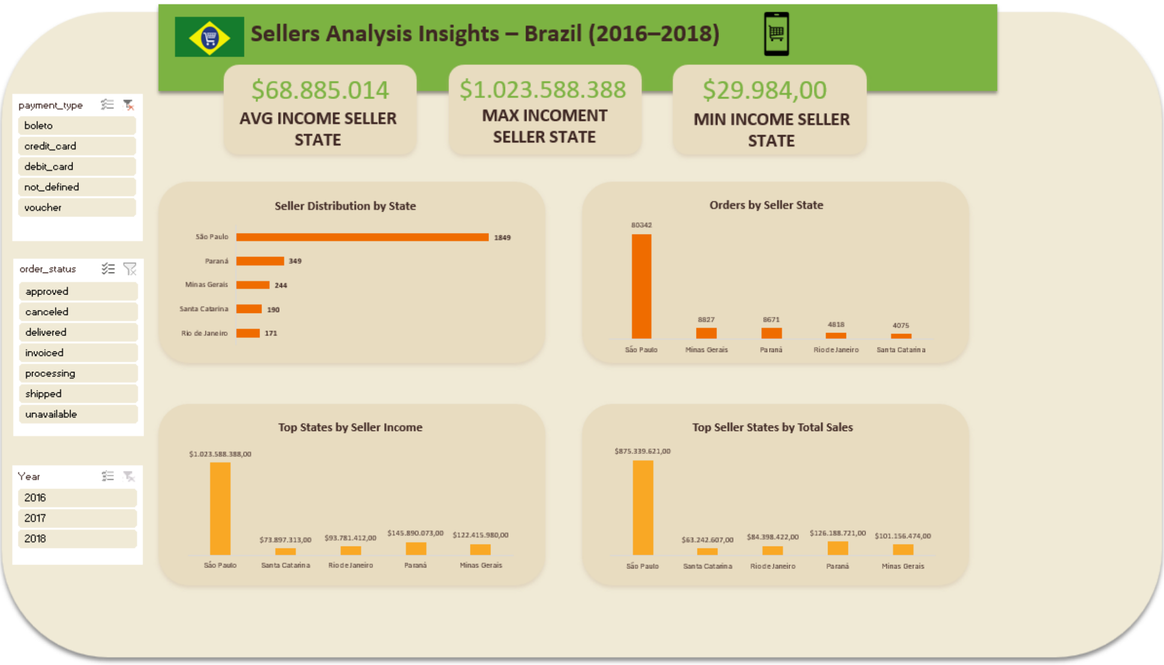
At the same time, the main product category by shipping weight was Furniture & Decor, with a total weight of 7,993 tons across all products in that category.

It is also important to highlight that, among the categories with the largest average product size, Office Furniture stood out with an average size of 75,468 cm³.

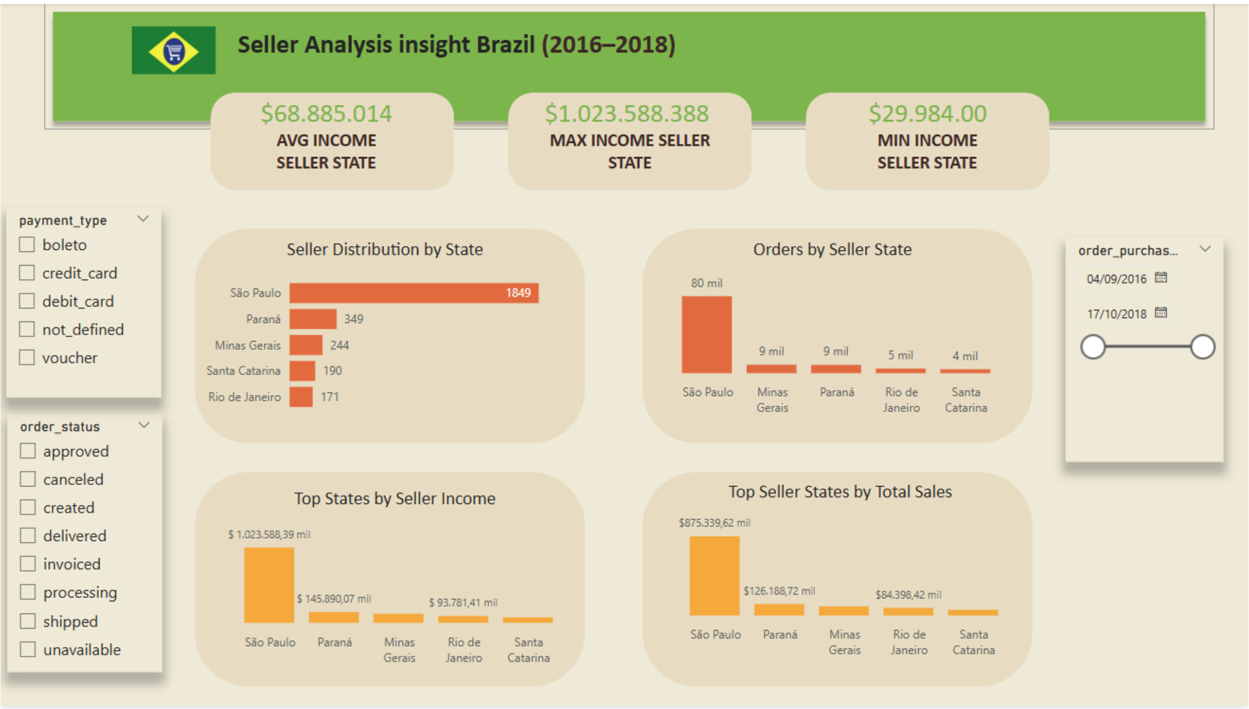
Finally, in terms of order cancellations, the category with the highest number of cancellations was Sports & Leisure, with a total of 51.

Seller Dashboard

Excel version



Power bi version



Seller Dashboard findings

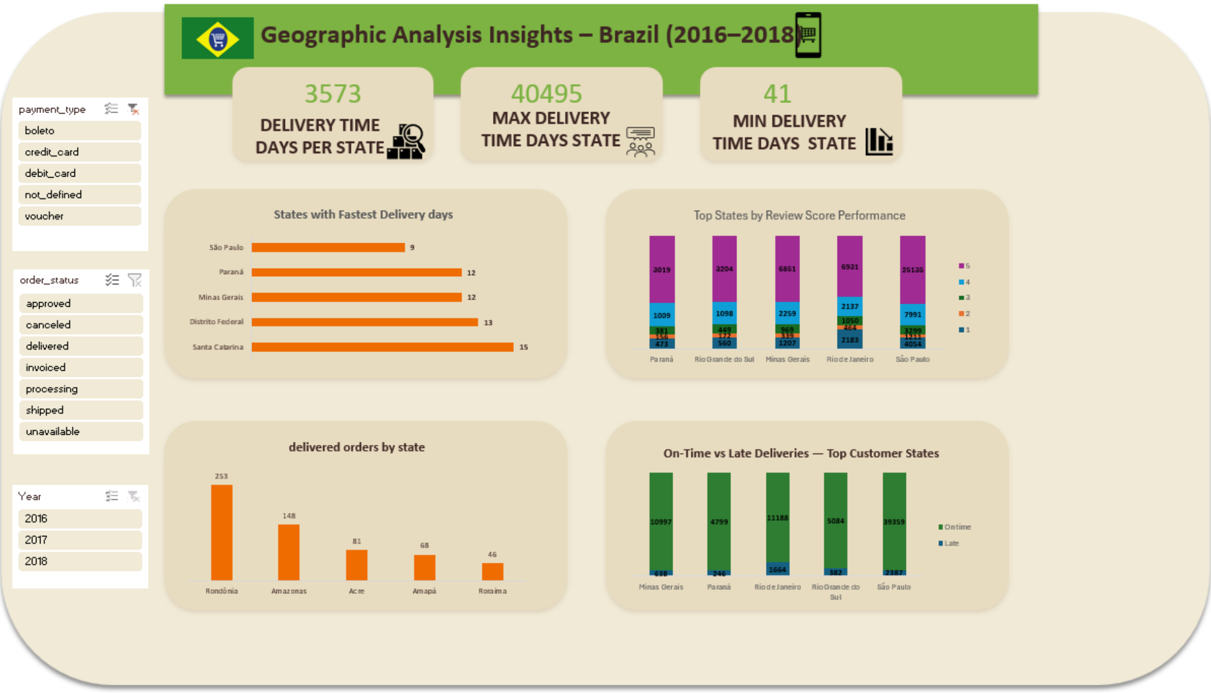
Regarding the seller analysis, the dashboard shows that most sellers are located in São Paulo, with a total of 1,849 sellers.

This state also generated a total of 80,342 orders.

In addition, São Paulo reported $1,023,588,388.00 in total revenue by seller and was also the state with the highest sales, reaching $875,339,621.00.

Geographic Dashboard

Excel Dashboard



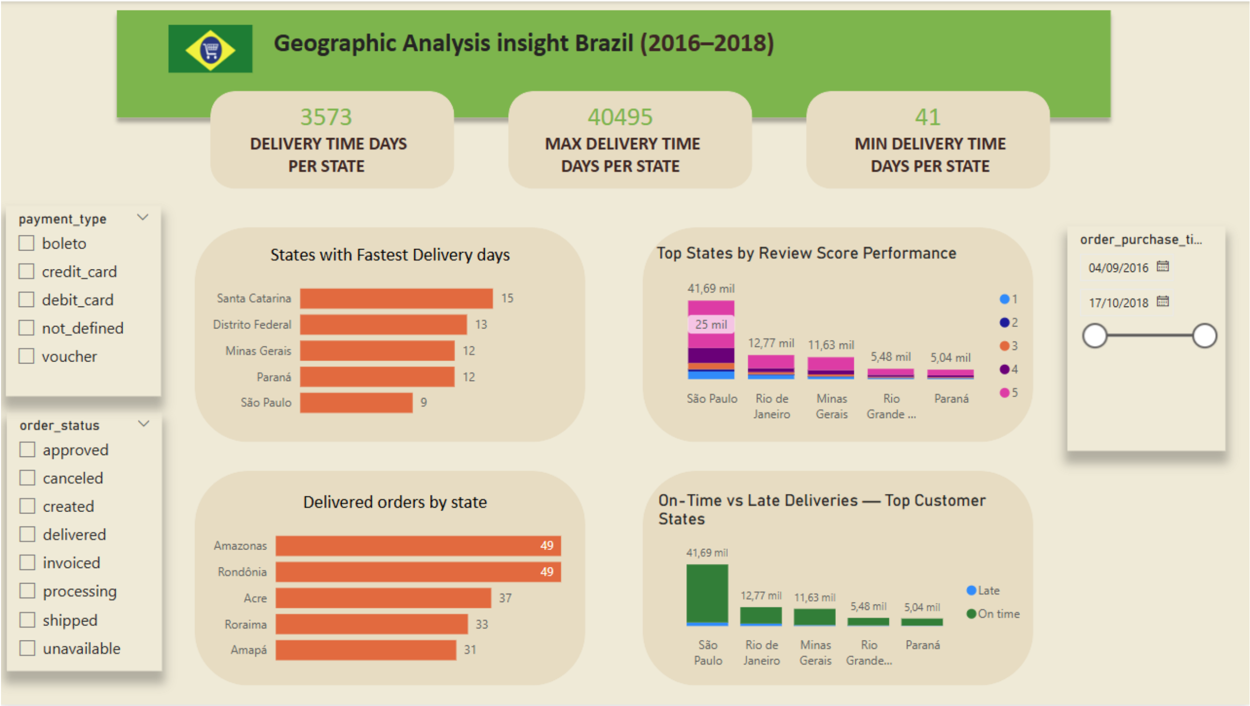
Geographic Analysis Insights – Brazil (2016–2018)

3573 DELIVERY TIME DAYS PER STATE

40495 MAX DELIVERY TIME DAYS STATE

41 MIN DELIVERY TIME DAYS STATE

Power Bi Dashboard



Geographic Dashboard Findings

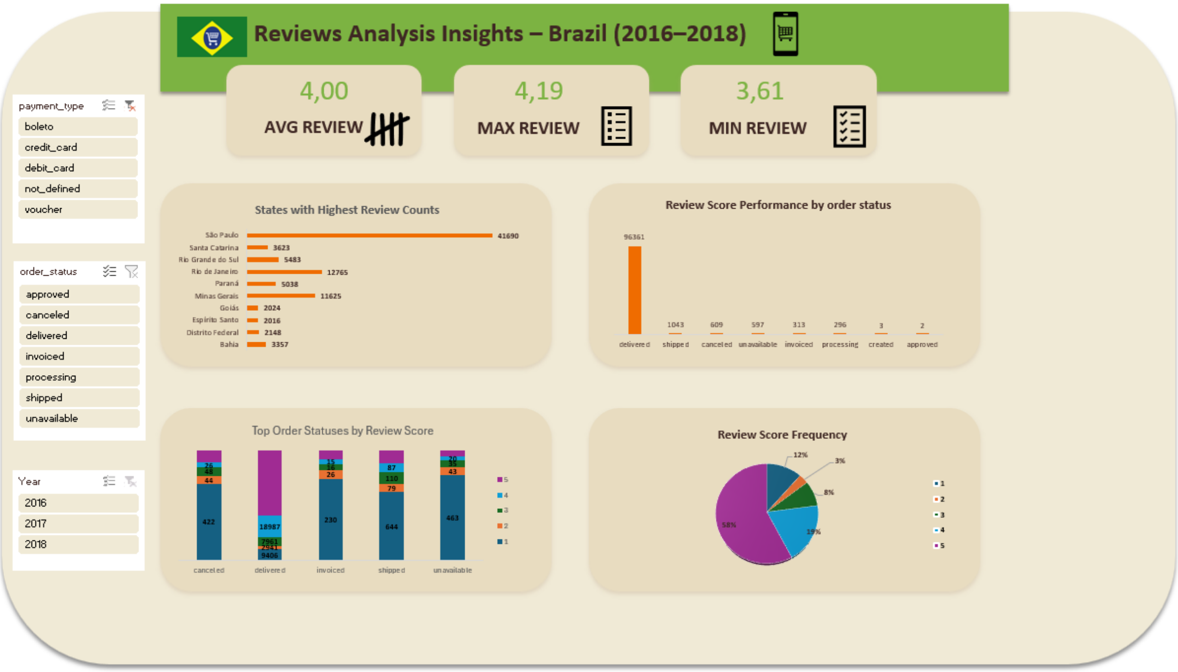
Regarding the geographic analysis, we can see that São Paulo was the state with the fastest deliveries, where the delivery time was no more than 9 days. It also showed the best review score performance, with a total of 25,135 five-star reviews, far above the others. However, it also had the highest number of negative reviews (1 star), with 1,211 in total.

In addition, we found that the state with the highest number of delivered orders was Rondônia, with a total of 253. Finally, São Paulo was the state with the most on-time deliveries (delivered on or before the estimated delivery date), with a total of 39,539.

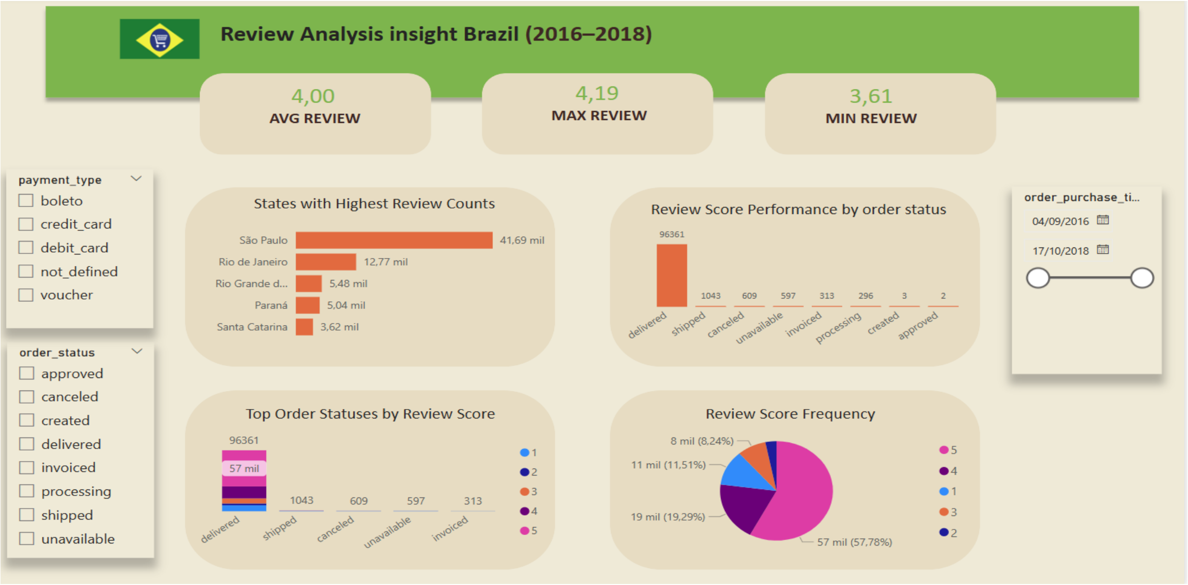
It is important to highlight that in all states, most reviews have a score of 4 or 5, which reflects a generally satisfactory customer experience. However, São Paulo also has a higher sales volume, which is confirmed in this analysis.

Review Dashboard

Excel version



Power Bi Version



Review Dashboard Findings

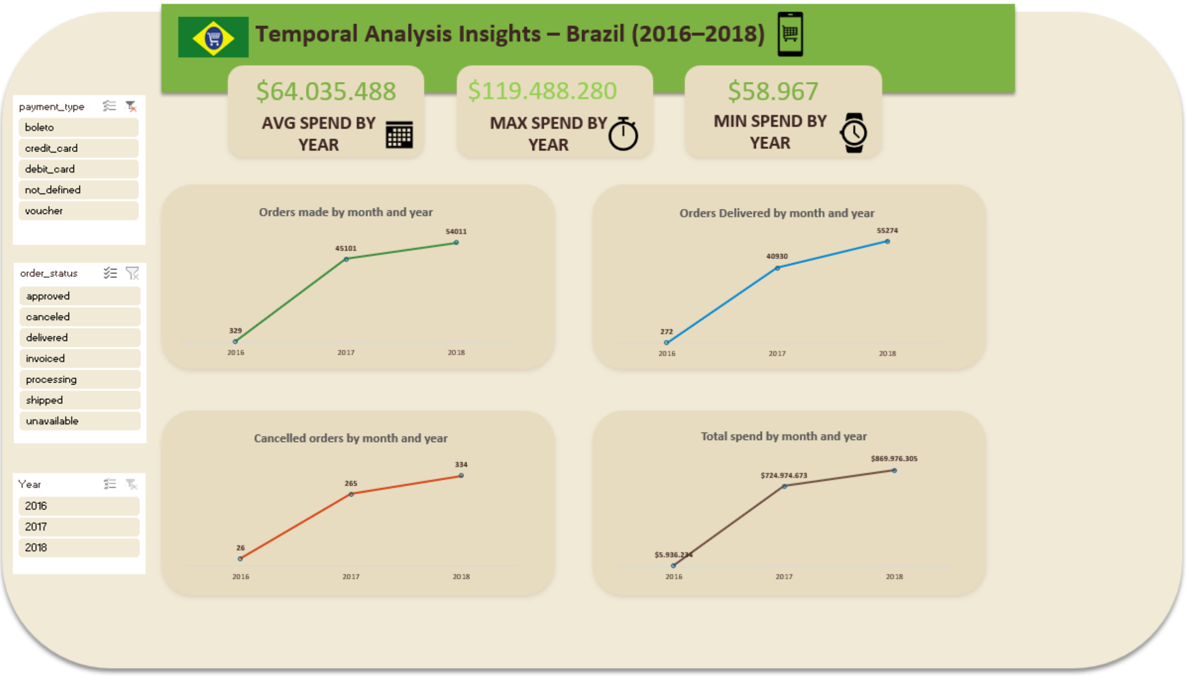
Now, going into the review analysis, it is important to mention that, according to the dashboard, the state with the highest number of reviews is São Paulo, with a total of 41,690. The dashboard also shows that delivered orders received the most reviews.

We can also see that the orders with a review score of 1 mostly come from states where the delivery was not completed or is still in progress, with "Shipped" being the most frequent status with score 1, totaling 644 reviews. On the other hand, the "Delivered" status received mostly 5-star reviews, with a total of 57,066.

Finally, regarding Review Score Frequency, most of the reviews received a score of 5, which represents 56% of all reviews.

Temporal Dashboard

Excel version



Temporal Analysis Insights – Brazil (2016–2018)

$64.035.488
AVG SPEND BY YEAR

$119.488.280
MAX SPEND BY YEAR

$58.967
MIN SPEND BY YEAR

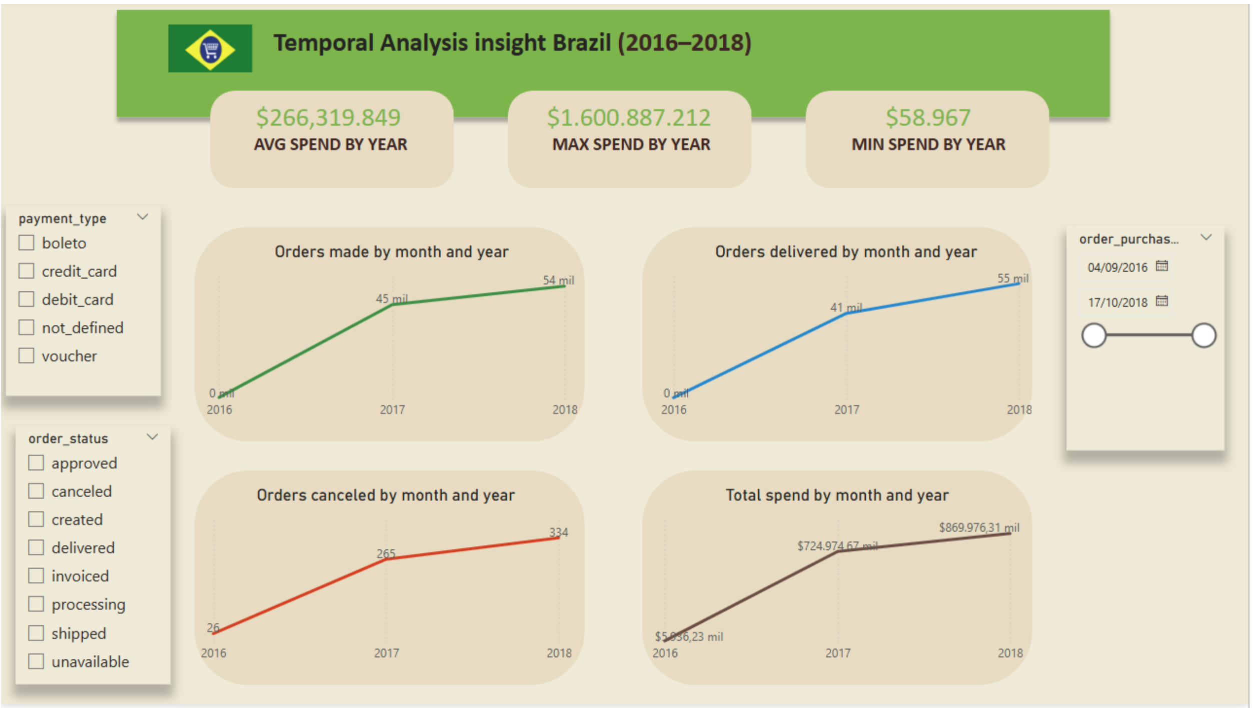
payment_type
- boleto
- credit_card
- debit_card
- not_defined
- voucher

order_status
- approved
- canceled
- delivered
- invoiced
- processing
- shipped
- unavailable

Year
- 2016
- 2017
- 2018

Orders made by month and year
329
45101
54011
2016
2017
2018
Orders Delivered by month and year
272
40930
55274
2016
2017
2018
Cancelled orders by month and year
26
265
334
2016
2017
2018
Total spend by month and year
$5.936.234
$724.974.673
$869.976.305
2016
2017
2018

Power Bi Version



Temporal Analysis insight Brazil (2016–2018)

$266,319.849
AVG SPEND BY YEAR

$1.600.887.212
MAX SPEND BY YEAR

$58.967
MIN SPEND BY YEAR

payment_type
□ boleto
□ credit_card
□ debit_card
□ not_defined
□ voucher

order_status
□ approved
□ canceled
□ created
□ delivered
□ invoiced
□ processing
□ shipped
□ unavailable

order_purchas...
04/09/2016
17/10/2018

Temporal Dashboard Findings

To finish, we carried out the time analysis. In this regard, the dashboard shows that for most variables, there was exponential growth in the number of orders per year. It went from 0 in 2016, to 45,000 in 2017, and then increased to a total of 54,000 in 2018.

Regarding the number of delivered orders, it also grew: from 0 in 2016, to 41,000 in 2017, and reached 55,000 in 2018.

The dashboard also shows that, in parallel, canceled orders increased, from just 26 in 2016 to 334 in 2018.

Finally, the last chart shows that customer spending per year also increased significantly, from $5,936,234 in 2016 to $869,976,305 in 2018, representing a large growth in e-commerce revenue in Brazil.

Insights and recommendations

Product Diversification by Category / Category Combination

It is recommended to combine categories to create promotions, offers, or bundles that include both high-demand products and those with lower demand. This strategy could help increase sales of less popular products and generate greater interest among customers.

Improvement of Logistics Processes

It is essential to identify the causes of delivery delays and work on solutions. This may involve evaluating and changing courier services, as well as implementing alternatives such as pickup points—e.g., local stores in neighborhoods.

In addition, automating processes through artificial intelligence could help track deliveries, predict delays, and plan dynamic routes to optimize delivery times.

Reducing Shipping Costs

Shipping costs tend to be higher in states with more customers, making purchases more expensive for them. Reducing these costs could encourage higher purchase volumes and increase company revenue. This can be achieved through strategic delivery partnerships or by offering pickup locations.

Increasing Sales in Low-Demand States

To boost sales in states with lower commercial activity, it is advisable to offer targeted promotions, introduce more payment methods, and provide options like installment payments or discount cards after a certain number of purchases.

Encouraging the Use of Installments

Promoting payment in installments can increase the average order value and improve revenue flow. This strategy is particularly useful in regions with lower purchase volumes.

Advanced Customer Segmentation

Identifying characteristics such as age, gender, interests, purchase frequency, average order value, and preferred purchase channel is key. This segmentation will help personalize offers, identify high-potential products, and develop more effective marketing strategies. It will also allow classification of customers into groups such as frequent buyers, novelty seekers, occasional visitors, and more.

Referral Programs

Implementing a referral program that offers rewards or discount coupons for every new referred customer can boost both sales and customer acquisition.

Customized Strategies by State and Season

Marketing campaigns should be adapted to the characteristics of each city and time of year. It's recommended to design offers based on relevant local dates, such as Children's Day, Mother's Day, national holidays, or regional celebrations. It is also useful to take advantage of the unique traits or appeal of each state.

Customer Loyalty Through Programs and Benefits

Loyalty programs should be implemented, including gift cards, exclusive discounts, or special promotions. These initiatives help strengthen customer retention and increase purchase frequency.

Management of Canceled Orders and Bulky Products

Some products are difficult to ship due to their weight or volume, which may lead to order cancellations. More cost-effective and efficient delivery methods should be considered, including shipping bulky items along with lighter products to reduce total shipping costs. It's also important to analyze the reasons behind cancellations and returns, which may stem from product issues, delivery delays, unclear descriptions, or courier service deficiencies.

Opening Physical Stores

Due to the high demand and volume of sales in cities like São Paulo, opening physical stores in strategic locations could be a profitable move. It would improve customer experience and expand sales channels.

Conducting Customer Satisfaction Surveys

Surveys should be carried out at different stages of the process (purchase, shipping, delivery) to identify possible problems or areas for improvement in both customer service and logistics.

Evaluation Indicators for Recommendations

It is advisable to define key performance indicators (KPIs) to measure the impact of each proposed strategy. These might include sales growth, return rates, orders delivered on time vs. delays, and more. This will allow for precise tracking of effectiveness and support informed decision-making.

Sales Channel Analysis

It is important to analyze where customers feel most comfortable and satisfied when making purchases (website, WhatsApp, phone call, mobile app, etc.). This insight will help optimize user experience and guide investments toward the most effective sales channels.

Increase local vendors in states with lower sales.

Conduct surveys of canceled products and orders to determine reasons for returns and cancellations.

Storytelling – Project 3: Brazilian E-Commerce Analysis

E-commerce in Brazil has been growing fast in recent years. However, it still faces some challenges—especially in logistics, customer satisfaction, and regional differences in sales. While thinking about these issues, I started asking myself:

Which products are sold the most, and why?
What payment methods do people prefer?
Do customers use credit cards, cash, or pay in installments?

As someone who values accurate and objective information, I knew the best way to explore these questions was by using real data. That's how I discovered the Brazilian E-Commerce Public Dataset by Olist, which contains over 100,000 orders from 2016 to 2018. This dataset covers the full buying process—from order and payment to shipping and customer reviews. Its snowflake-style structure gave me a great opportunity to apply my data skills and explore customer behavior, logistics, segmentation, and sales performance.

1. Customer location and spending

São Paulo clearly stood out as the top region: it had the most customers (41,476), the highest number of orders (41,476), and total spending of more than $599 million. However, it also had the highest shipping costs, with $71 million in freight charges. Other regions with a strong customer base included Rio de Janeiro (12,852) and Minas Gerais (11,365).

2. Payment behavior

The most popular payment method was the credit card, with 76,795 orders and over $1.254 billion in sales. The second most used method was the boleto bancário, with 19,784 transactions and $286 million in sales. Regarding installment payments, most customers (52,546) paid in just one installment—showing that paying in parts is not widely used yet.

3. Delivery and logistics

São Paulo had the highest number of on-time deliveries (39,359), while Rondonia had the most completed deliveries (253), even surpassing bigger states. However, the number of canceled orders increased over time—from 26 in 2016 to 334 in 2018. This could be related to product delays, incorrect descriptions, or shipping issues.

4. Customer satisfaction

Customer reviews showed that 57.78% of users gave 4 or 5 stars. The average rating was 4.00, with a minimum of 3.61 and a maximum of 4.19. One common review that summarizes this positive experience was:

"The product arrived earlier than expected and in great condition. I would definitely buy again. Olist does a great job."

These reviews reflect a high level of trust and satisfaction among customers.

5. Top-selling products

The best-selling categories were Health & Beauty ($125 million), Watches & Gifts ($120 million), Bed, Bath & Table ($103 million), and Sports & Leisure ($98 million). However, Sports & Leisure also had the most order cancellations (51), followed by Housewares (49) and Computer Accessories (46). This might show problems with product quality or customer expectations.

Strategic Recommendations

Based on the data, I suggest the following improvements:

Improve logistics in key areas: Focus on regions with high demand, like São Paulo, to reduce shipping times and delivery costs.

Encourage installment payments: Promoting this option could help increase average order value, especially in low-demand regions.

Use text analysis on product reviews: This can help identify quality issues or common complaints faster.

Create state- or season-based promotions: For example, campaigns for holidays like Children's Day or Christmas could boost sales.

Offer product bundles: Combining top-selling products with slower-moving items can improve overall sales.

Send satisfaction surveys at each step: From purchase to delivery, this can help detect pain points in the customer journey.

Expand the seller network in low-sales regions: This can help reach more customers and improve delivery performance.

Conclusion

This project, my third data analysis case study—was a great opportunity to work with a complex and realistic dataset. It helped me apply technical skills like data modeling, cleaning, and visualization, while also understanding the business cycle of a large e-commerce company.

By analyzing customer behavior, product performance, logistics, and regional trends, I was able to generate insights that can help companies like Olist improve the customer experience, grow sales, and optimize their operations. The recommendations are based directly on the data and reflect real opportunities for action.

Challenges during the process

This project presented several technical challenges, starting with the creation of relationships between the dataset tables. The complexity of these relationships directly affected the performance of filters in Power BI and slicers in Excel. Even after rebuilding the model multiple times, filters did not always work as expected, which forced me to redesign some parts of the dashboards.

Another important challenge was loading the data into SQL Server. Some issues appeared due to missing values or inconsistencies between the data types in the original files and the structure defined in the database tables. This required several validations and manual corrections.

During the exploratory analysis, I found some values that seemed inconsistent, such as a maximum review average of 4.19. At first, this raised questions, but after checking the individual records, I saw that many reviews between 4.0 and 4.9 were treated as “5”, which helped me adjust the interpretation.

Creating eight dashboards was also demanding due to the large amount of data and the visual decisions involved. However, as the data was cleaned and transformed (ETL), the structure became more clear and useful for analysis and visualization.

It was also challenging but interesting to work with Power Pivot, especially when creating calculated columns and measures using DAX language. This experience showed how powerful and important DAX is for building deeper insights.

Finally, I noticed some small differences between values shown in Excel and Power BI, even though both used Power Query for data cleaning. This highlighted the need for better control during calculations and review in both tools.