

Olist Sales Analysis

Customer segmentation (RFM) and sales dashboard with Python, SQL Server and Power BI

Heejeong | [LinkedIn](#) | [GitHub](#)

93,357

customers segmented

R\$15.4M

sales analyzed (BRL)

96,477

delivered orders

What I built and why

Business questions

- 1 Which product categories sell the most?
- 2 How do sales change from month to month?
- 3 Which customer tiers generate the revenue?
- 4 Which states drive sales?

Approach: Python cleans the data and scores RFM, SQL Server turns it into views, and Power BI presents the dashboard.

99,441

orders from 2016 to 2018 across 9 tables

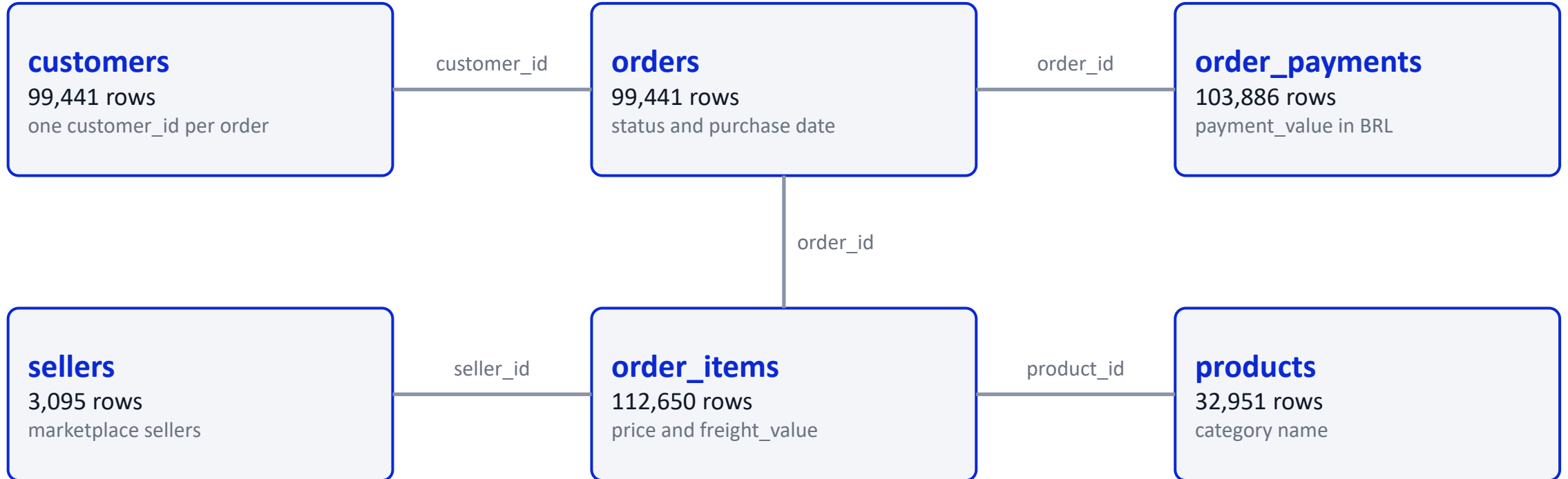
93,357

customers scored with RFM

6

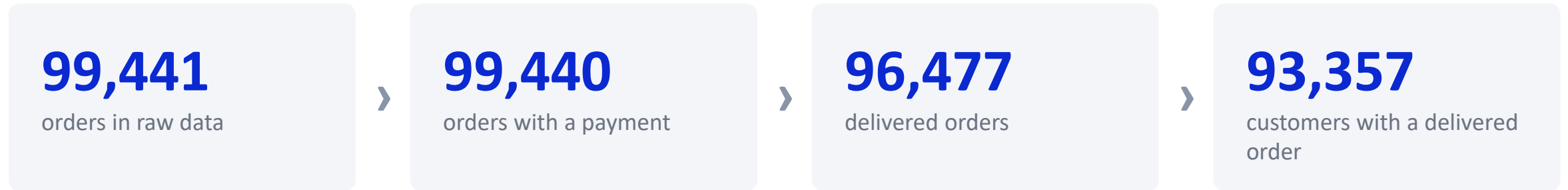
customer tiers, from VIP to Churned

Nine tables linked by order, customer and product IDs



Also in the dataset: `product_category_name_translation` (71 rows) gives English category names. `order_reviews` (99,224 rows) and `geolocation` (1,000,163 rows) are not used in this analysis.

Cleaning narrowed 99,441 orders to 96,477 delivered



Step	What I did	Result
1 Real customers	customer_id is issued per order, so I used customer_unique_id as the customer	96,096 unique customers
2 Missing payment	Removed the one order that has no payment record	99,440 orders
3 Delivered only	Kept order_status = delivered, dropped shipped, canceled and other statuses	96,477 orders (2,963 removed)
4 RFM population	Kept customers with at least one delivered order	93,357 customers (2,739 removed)
5 Readability	Translated category names to English and fixed misspelled column names	English labels throughout

RFM scores each customer on recency, frequency and value



Recency

Days since the customer's last purchase, measured from the latest order in the data (2018-08-29).

Score rule

- Quintiles of recency, scored 5 to 1
- 5 = most recent customers



Frequency

Number of distinct delivered orders per customer.

Score rule

- 1, 2, 3, 4 orders score 1 to 4
- 5 or more orders score 5



Monetary

Total payment value per customer, in BRL.

Score rule

- Under 20 = 1, under 50 = 2
- Under 100 = 3, under 500 = 4
- 500 or more = 5

RFM code = R score + F score + M score. Example: R4, F5, M5 gives code 455.

Six tiers are built from the RFM scores

VIP

18

customers

RFM code 455, 545, 554 or 555

Loyal

15

customers

R of 4 or 5 and F of 4 or 5 (not VIP)

Potentially Loyal

56,059

customers

Everyone not placed in another tier

New

18,106

customers

R = 5 and F = 1: very recent first-time buyers

At-Risk

549

customers

R = 2 and F = 2 or 3

Churned

18,610

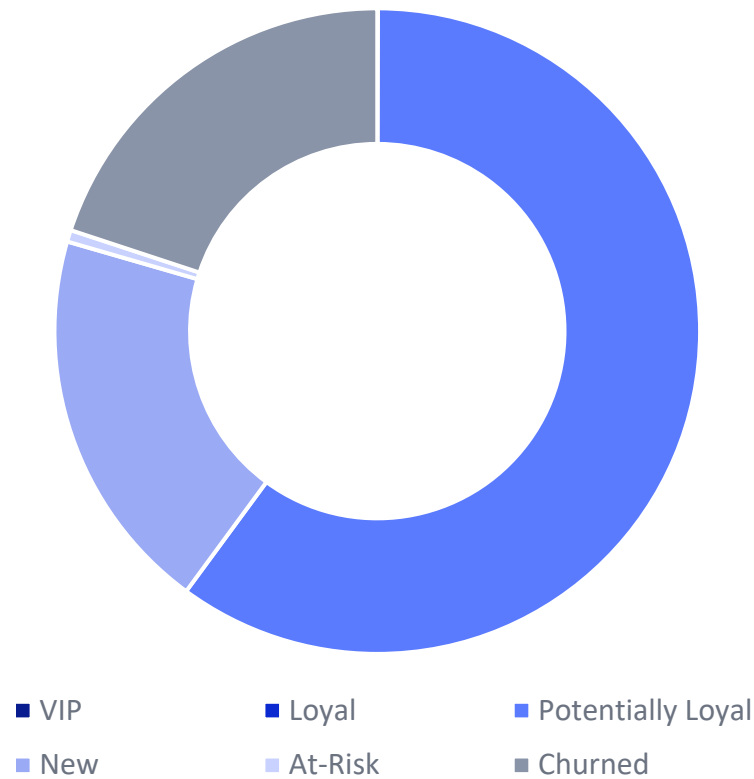
customers

R = 1 and F = 1 or 2: bought long ago

Rules are applied in this order: VIP, Loyal, New, Churned, At-Risk, then Potentially Loyal for the rest.

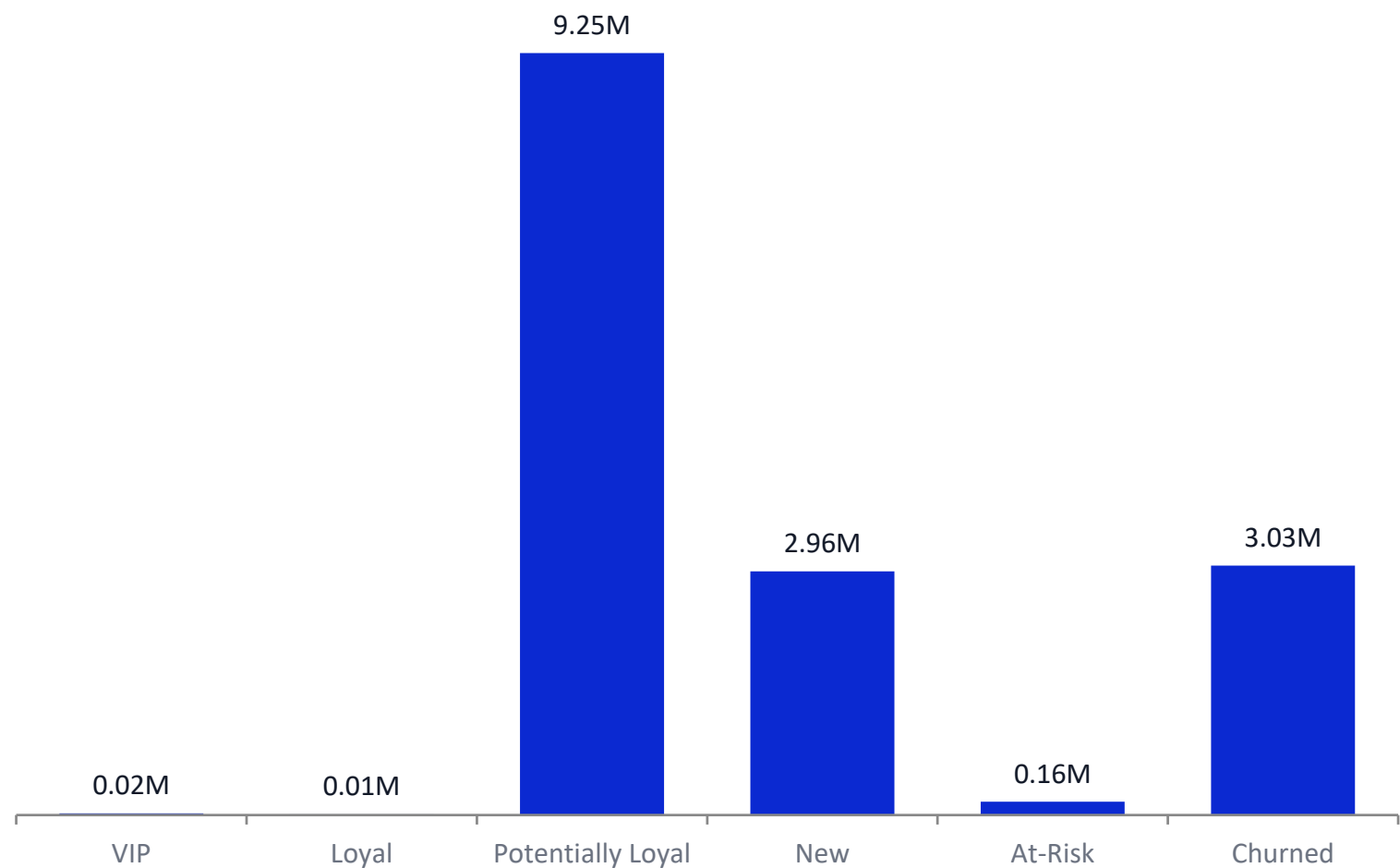
Potentially Loyal customers are 60% of the base

Tier	Customers	Share	Sales (BRL)
VIP	18	0.02%	20,696
Loyal	15	0.02%	7,899
Potentially Loyal	56,059	60.05%	9,246,892
New	18,106	19.39%	2,957,020
At-Risk	549	0.59%	161,856
Churned	18,610	19.93%	3,028,098
Total	93,357	100.00%	15,422,462



Only 33 customers reach VIP or Loyal. New and Churned customers each make up about 19 to 20% of the base.

Potentially Loyal customers bring in 60% of sales



60.0%

of sales (R\$9.25M) from Potentially Loyal

R\$3.03M

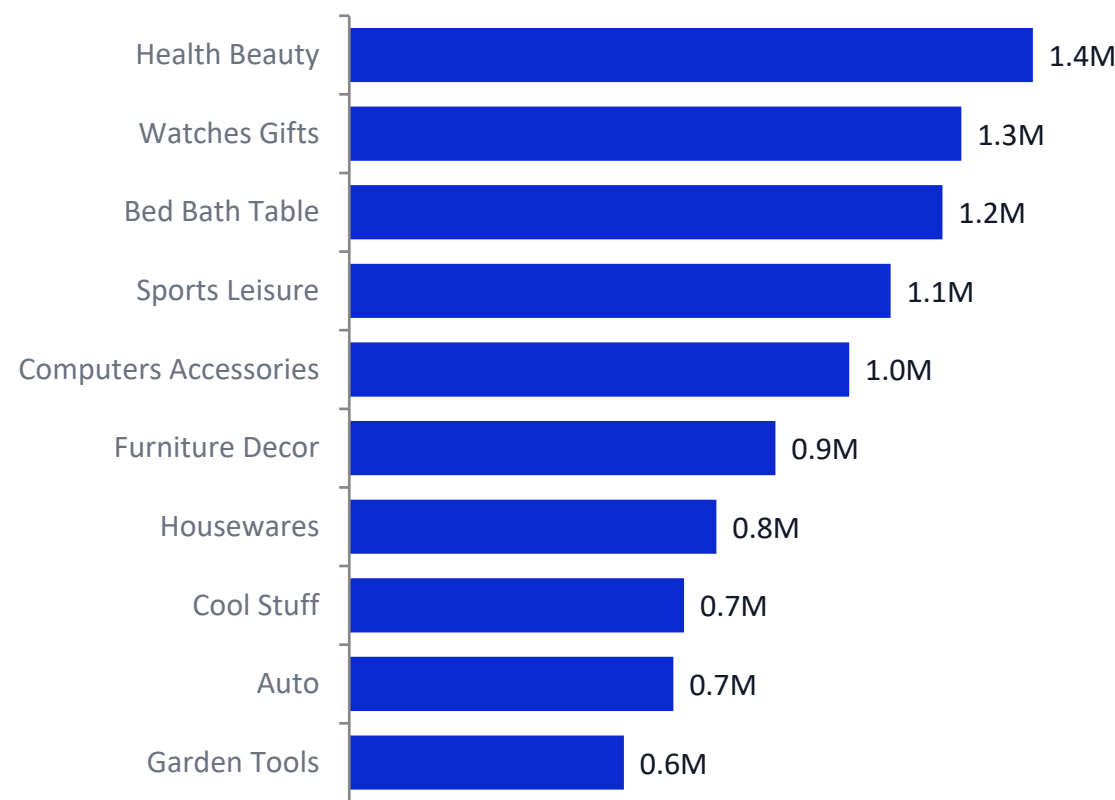
from Churned customers, a win-back target

R\$28.6K

from VIP and Loyal combined

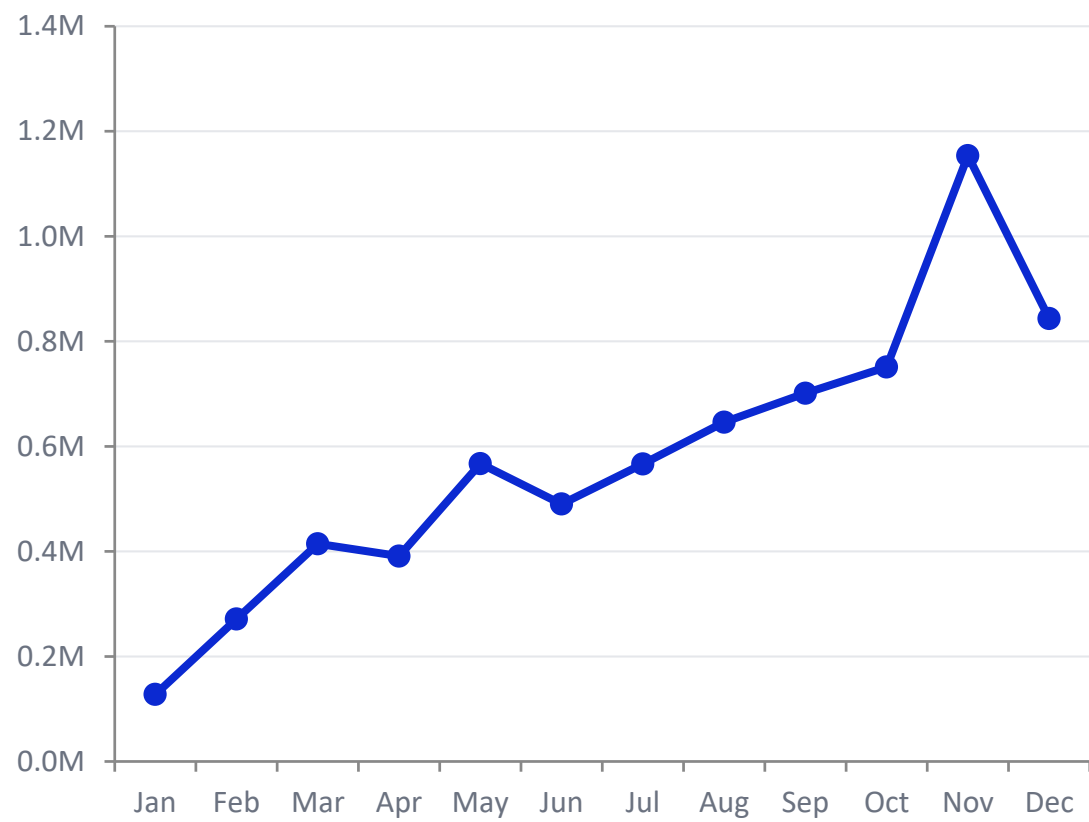
Health beauty leads categories; sales peak in November 2017

Top 10 categories by sales (BRL)



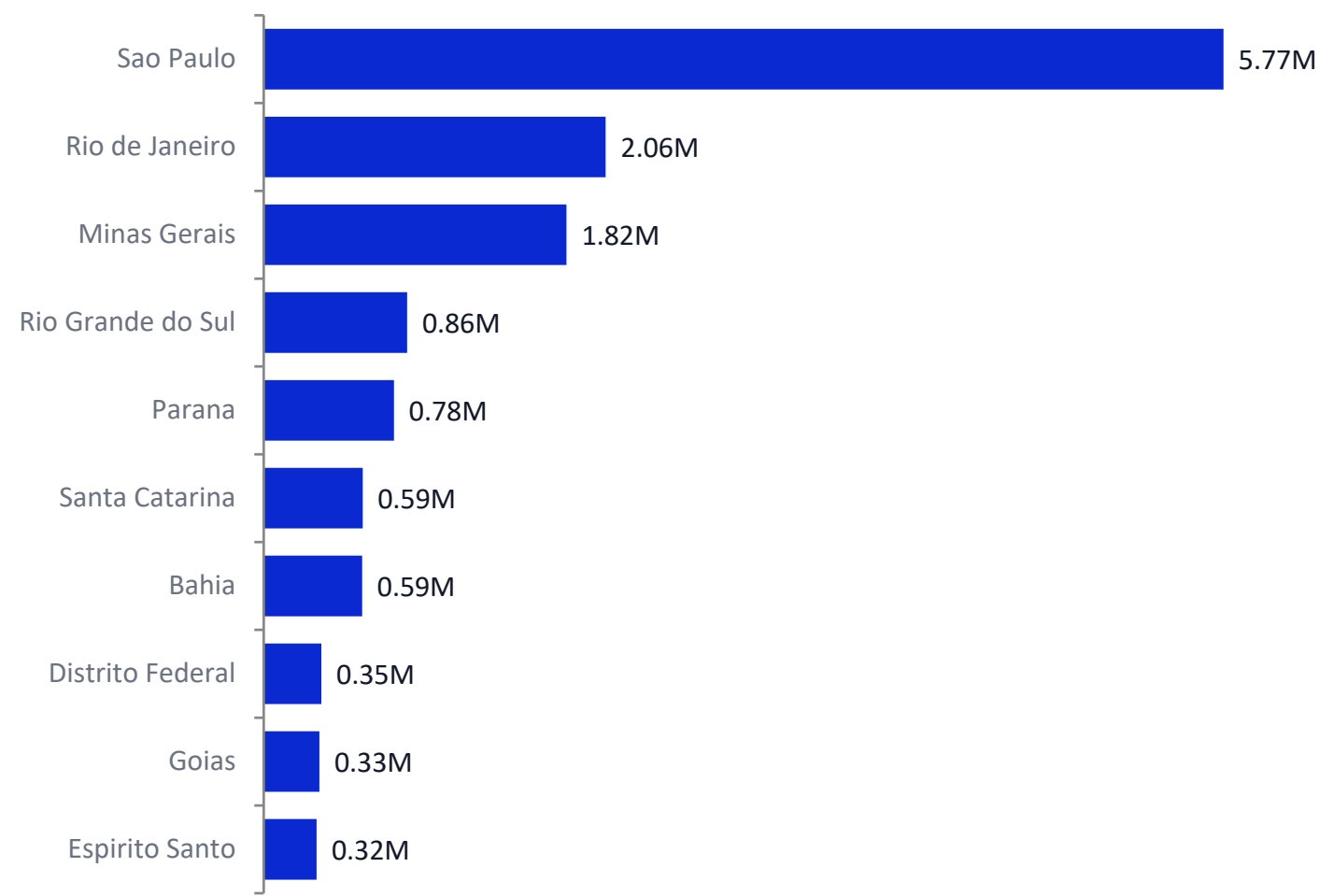
Monthly sales, 2017 (BRL)

Peak: Nov, R\$1.15M



Category sales = item price + freight of delivered orders. Monthly sales = payment value of delivered orders by purchase month.

Sao Paulo alone generates 37% of sales



37.4%

of sales (R\$5.77M) in Sao Paulo

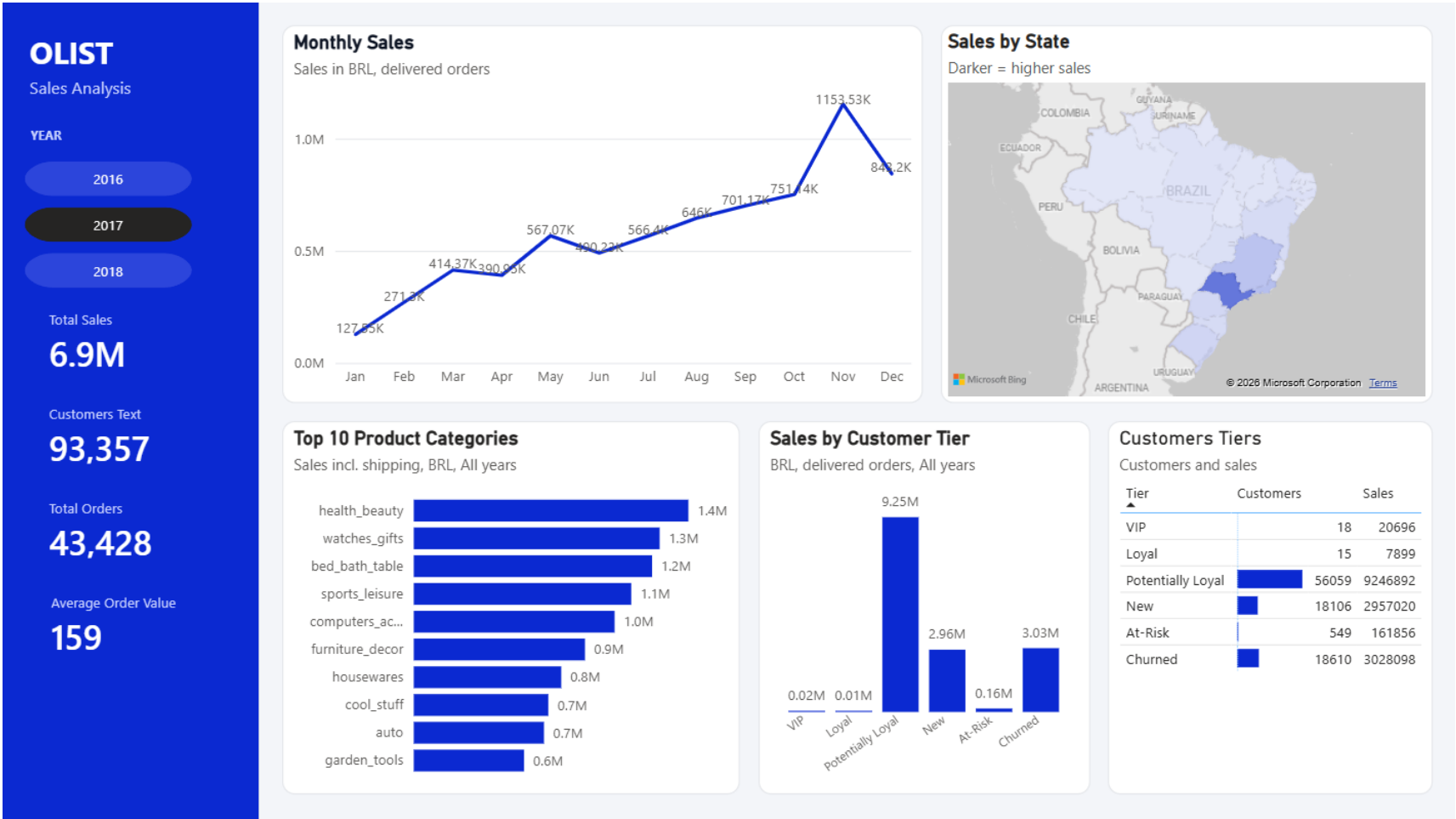
62.5%

of sales in Sao Paulo, Rio de Janeiro and Minas Gerais

27 states

all appear in the data; smallest is Roraima (R\$9K)

The Power BI dashboard brings the four views together



KPI sidebar

Total sales, customers, orders and average order value

Four views

Monthly sales, sales by state, top 10 categories, customer tiers

Data source

SQL Server views loaded in Import mode

Key takeaways

Insights

- Most sales come from one group, "Potentially Loyal" customers: they are 60% of customers and bring 60% of sales.
- Almost everyone (97%) bought only once. Getting customers to come back is the biggest chance to grow.
- Three states bring in most of the sales: Sao Paulo, Rio de Janeiro and Minas Gerais (62.5%).
- November 2017 was the best month (R\$1.15M), about 1.5 times more than October.

Appendix: columns used from each table

Table	Key columns	Notes
customers	customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, customer_state	customer_id is issued per order; customer_unique_id is the real customer
orders	order_id, customer_id, order_status, order_purchase_timestamp, approval and delivery dates	One row per order; only delivered orders are analyzed
order_items	order_id, order_item_id, product_id, seller_id, price, freight_value	One row per item; used for category sales
order_payments	order_id, payment_sequential, payment_type, payment_installments, payment_value	An order can have several payments; summed per order
products	product_id, product_category_name, size, weight and photo columns	Category name starts in Portuguese; English added in cleaning
sellers	seller_id, seller_zip_code_prefix, seller_city, seller_state	State names expanded to full names

Appendix: how the raw data was cleaned

1 Fix column names

In products, `product_name_lenght` and `product_description_lenght` were renamed to `product_name_length` and `product_description_length`.

2 Find the real customers

`customer_id` has 99,441 values but `customer_unique_id` has 96,096, so 3,345 ids belong to repeat buyers.

3 Map state names

27 abbreviations became full names (SP to Sao Paulo) in customers and sellers. A check confirmed none were left unmapped.

4 Translate categories

Portuguese names were mapped to English with the translation table. Two categories missing from it were translated by hand.

5 Clean orders

One order with no payment was removed (99,441 to 99,440). Then only delivered orders were kept (96,477).

6 Re-key to the customer

Delivered orders were joined to customers and `customer_id` was dropped, so each order is keyed by `customer_unique_id`.

Appendix: order status before the delivered filter

Order status	Orders	Share
delivered	96,477	97.02%
shipped	1,107	1.11%
canceled	625	0.63%
unavailable	609	0.61%
invoiced	314	0.32%
processing	301	0.30%
created	5	0.01%
approved	2	0.00%
Total	99,440	100.00%

Why only delivered orders

- Canceled and unavailable orders are not real revenue.
- Shipped, invoiced and processing orders are not complete yet.
- Delivered orders are 97% of all orders, so the filter keeps almost all of the data.