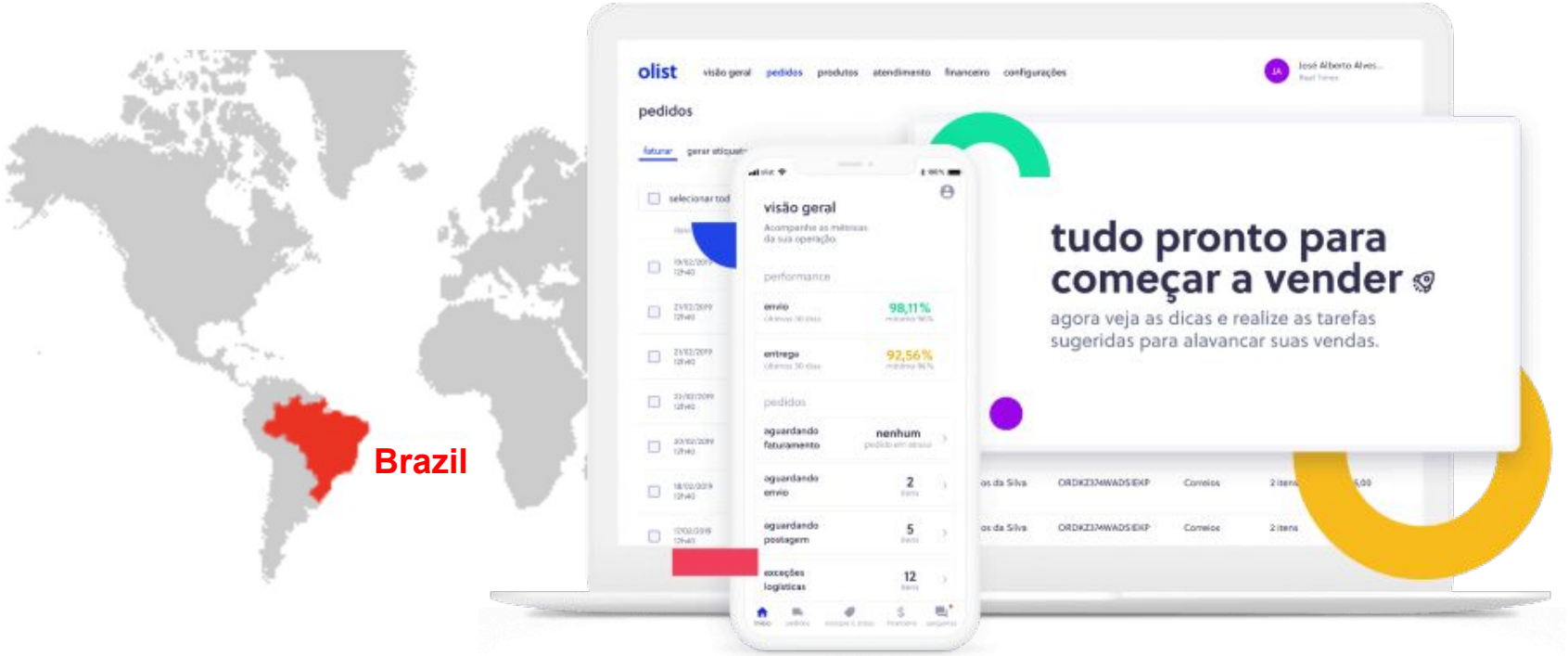
Olist E-Commerce

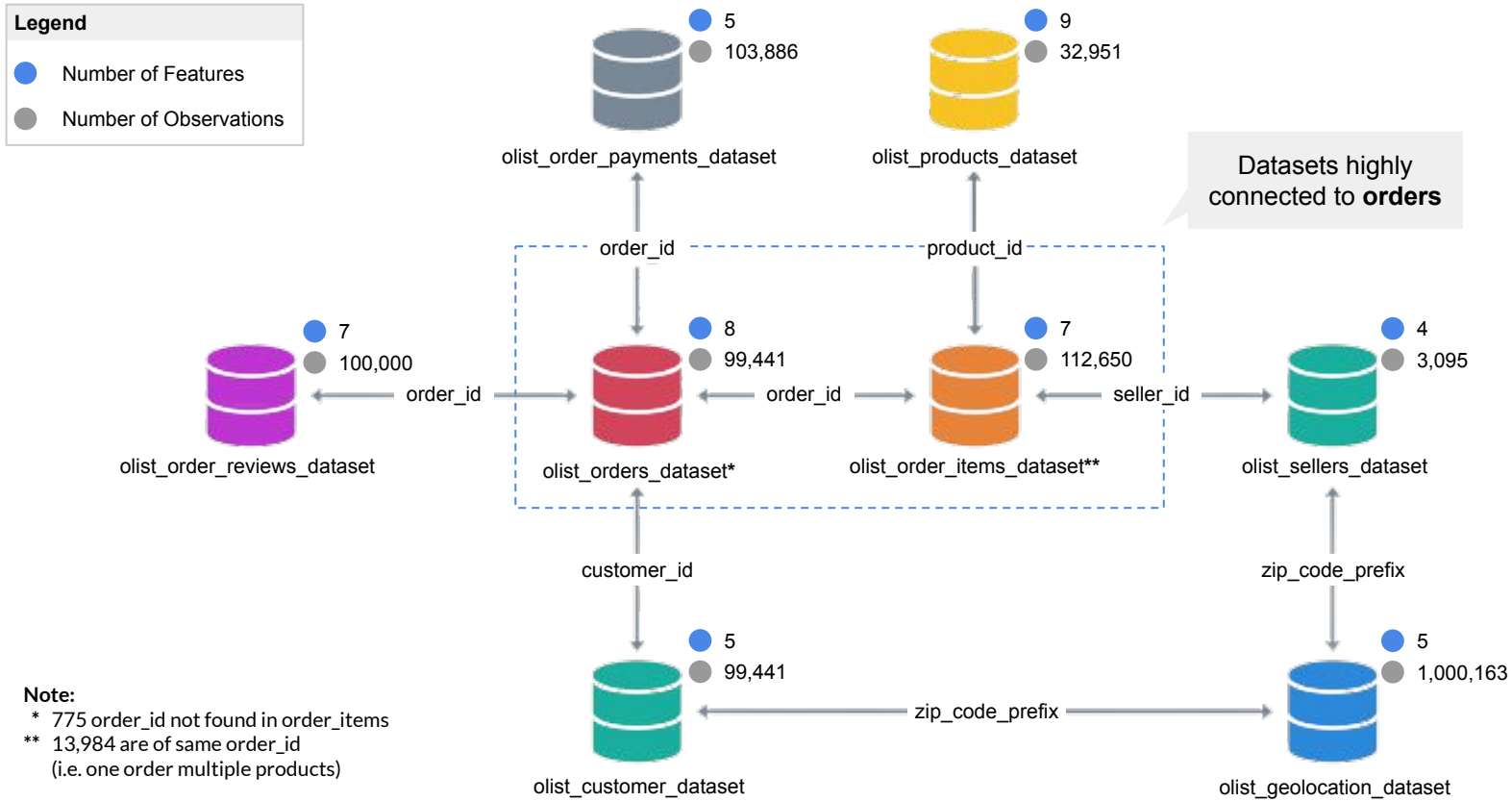
Exploratory Data Analysis & Visualisation

Group 16

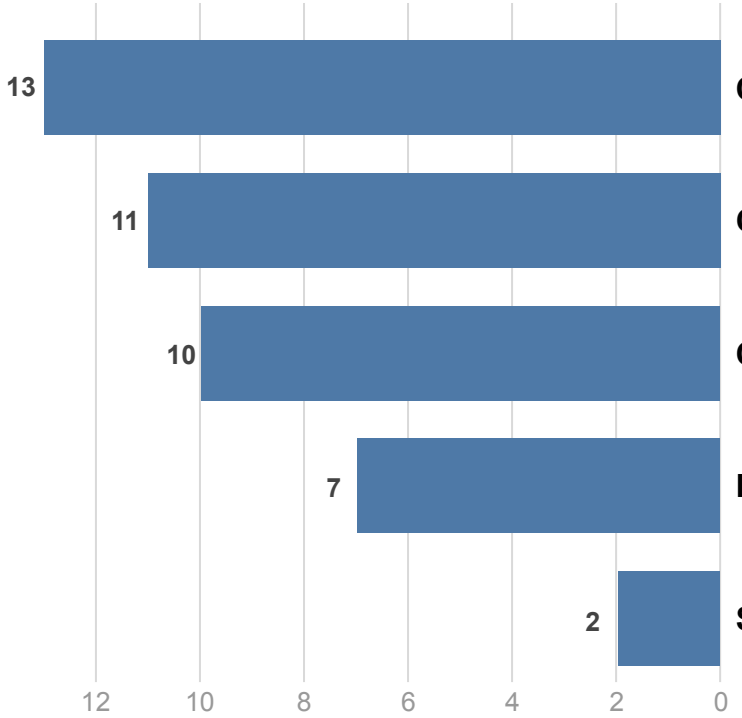
<u>Context & Target Audience:</u> Olist, a Brazilian e-commerce platform; target investors /sellers to improve performance through data



Data structure: Total 8 data sets focusing on customers, orders and products, but structure mainly centralised on orders



Data types: Total 5 types, with majority being quantitative data

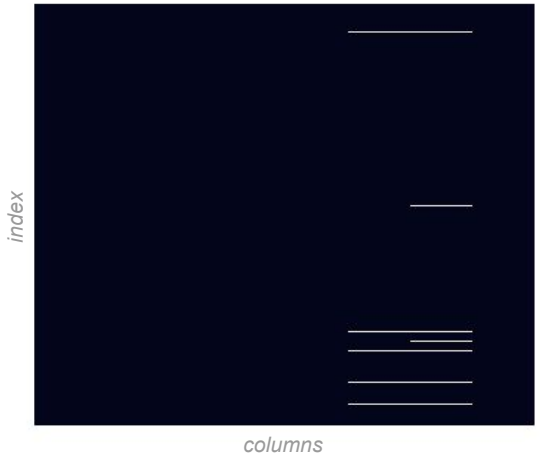


Sample Feature & Observations	
price	• $229
payment_type	• Credit card • Physical payment
review_score	• 5 → 4 → 3 → 2 → 1
delivery_date	• 2017-11-24 08:14 am • 2018-03-02 09:32 pm
review_message	• Super recomendo • MUITO BOM!

<u>Missing values:</u> Found in 3 datasets through heatmaps; around 2-3% on orders and products, and around 89% on reviews

Orders

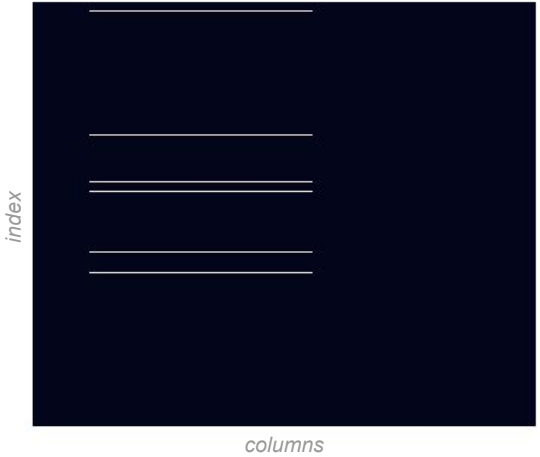
*Around 3% missing data on **delivered dates***



Potentially due to multiple reasons
1. Canceled orders
2. Incomplete data slicing
3. Software latency / API issue

Products

*Around 2% missing data on **product category, title, description, image***



*Potentially due to **incomplete/ unpublished listings** commonly listed for testing purposes*

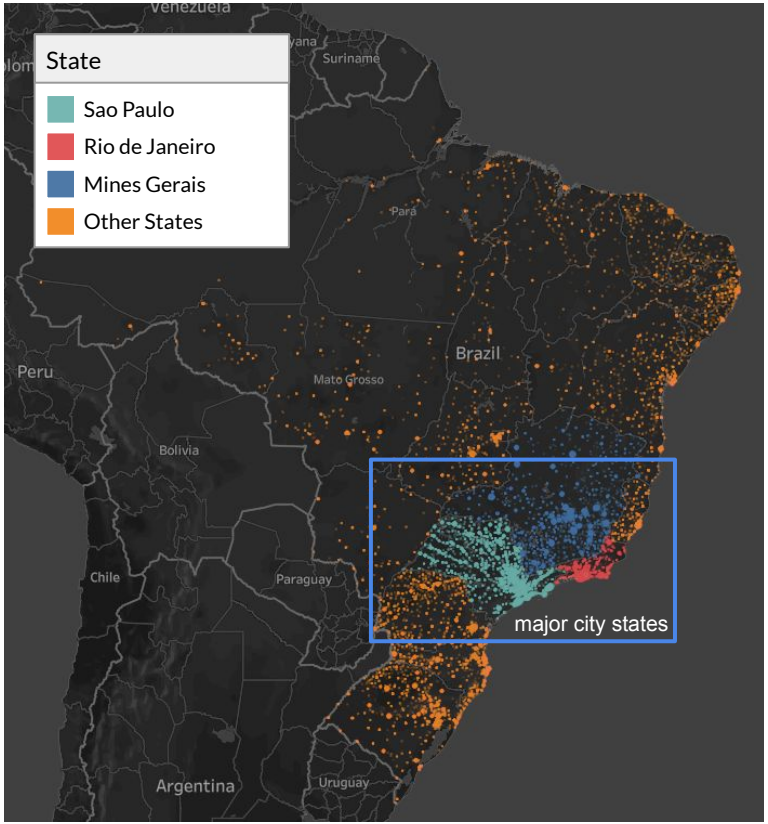
Reviews

*Around 89% missing data on **review titles and review messages***

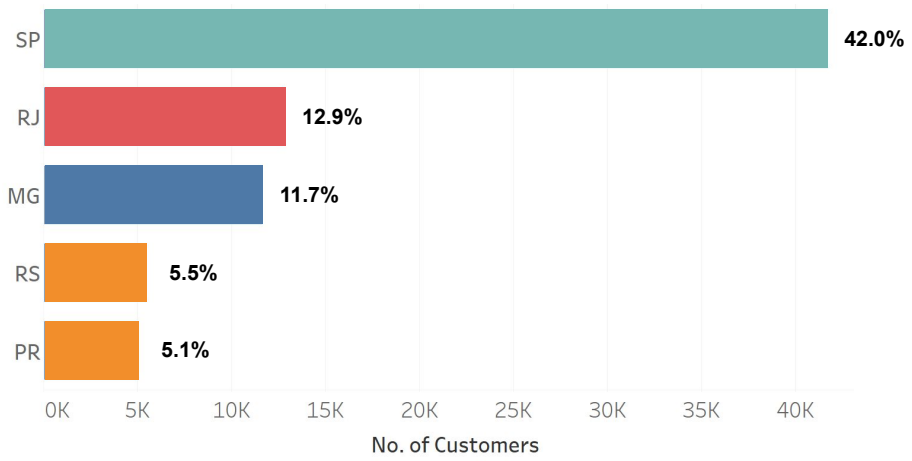
***Missing data expected** as customers often leave ratings but do not write any review*

Legend: Null values Non-null values

Customer: Buyer profiles heavily concentrated in major city states such as Sao Paulo, Rio de Janeiro and Minas Gerais

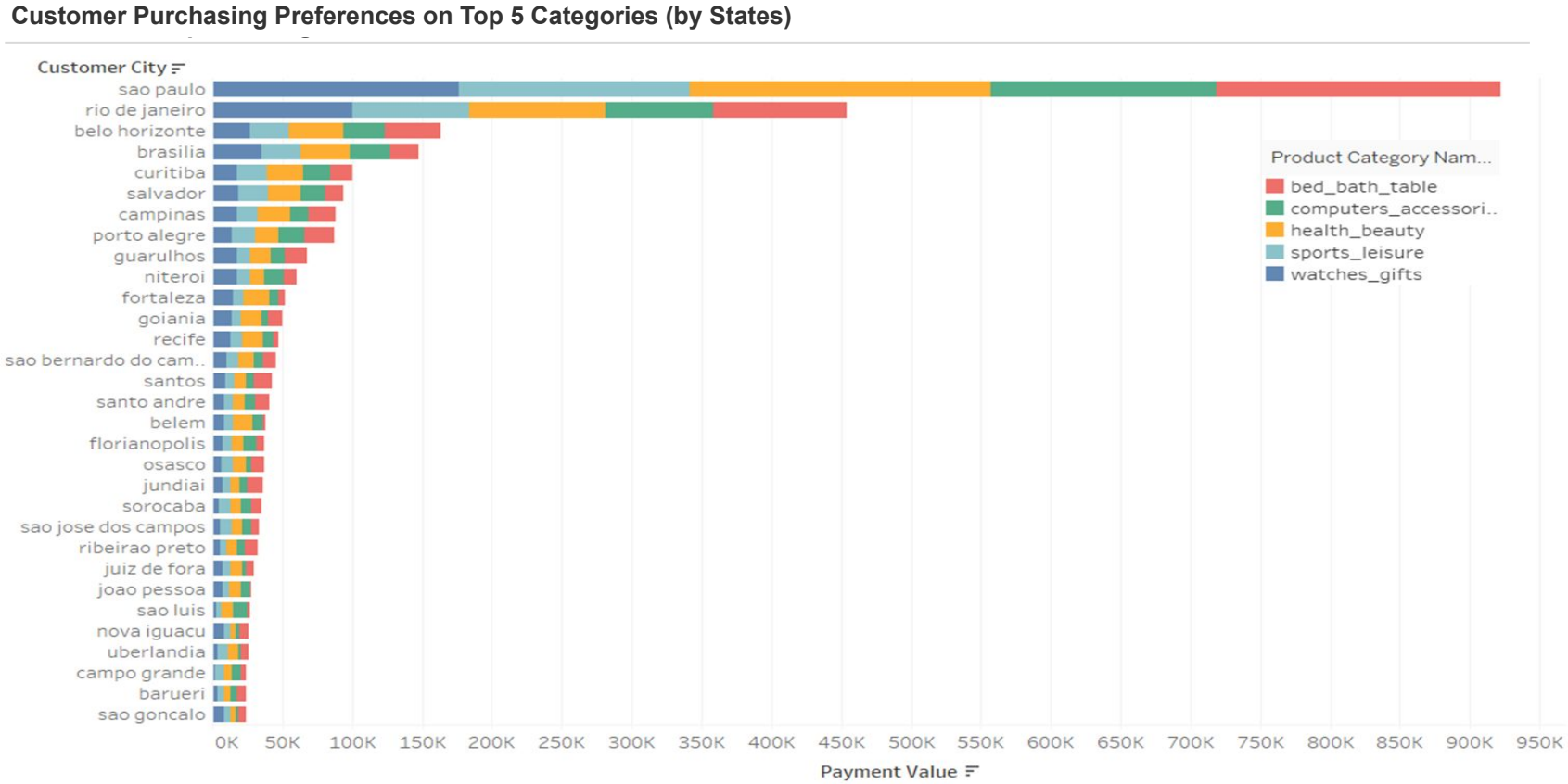


Customer distribution across the top 5 states in Brazil
(no. of customers in thousands)



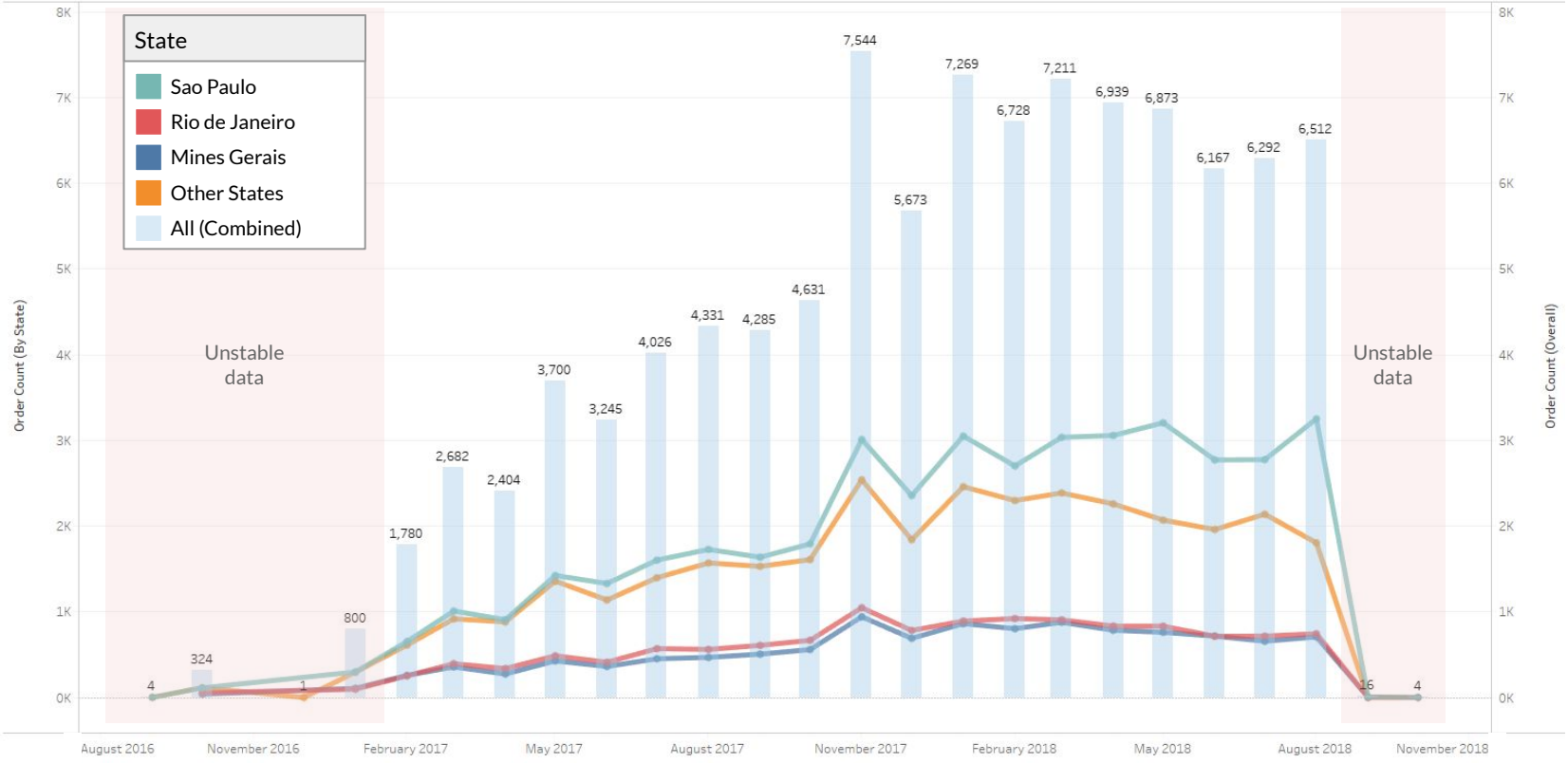
Top 5 states out of 27 in Brazil make up to **~77.2%** of the total customers, with the highest density in **Sao Paulo at ~42.0%**

Customer: Brazilian consumers in different states have highly similar preferences for types of goods, but rank them differently



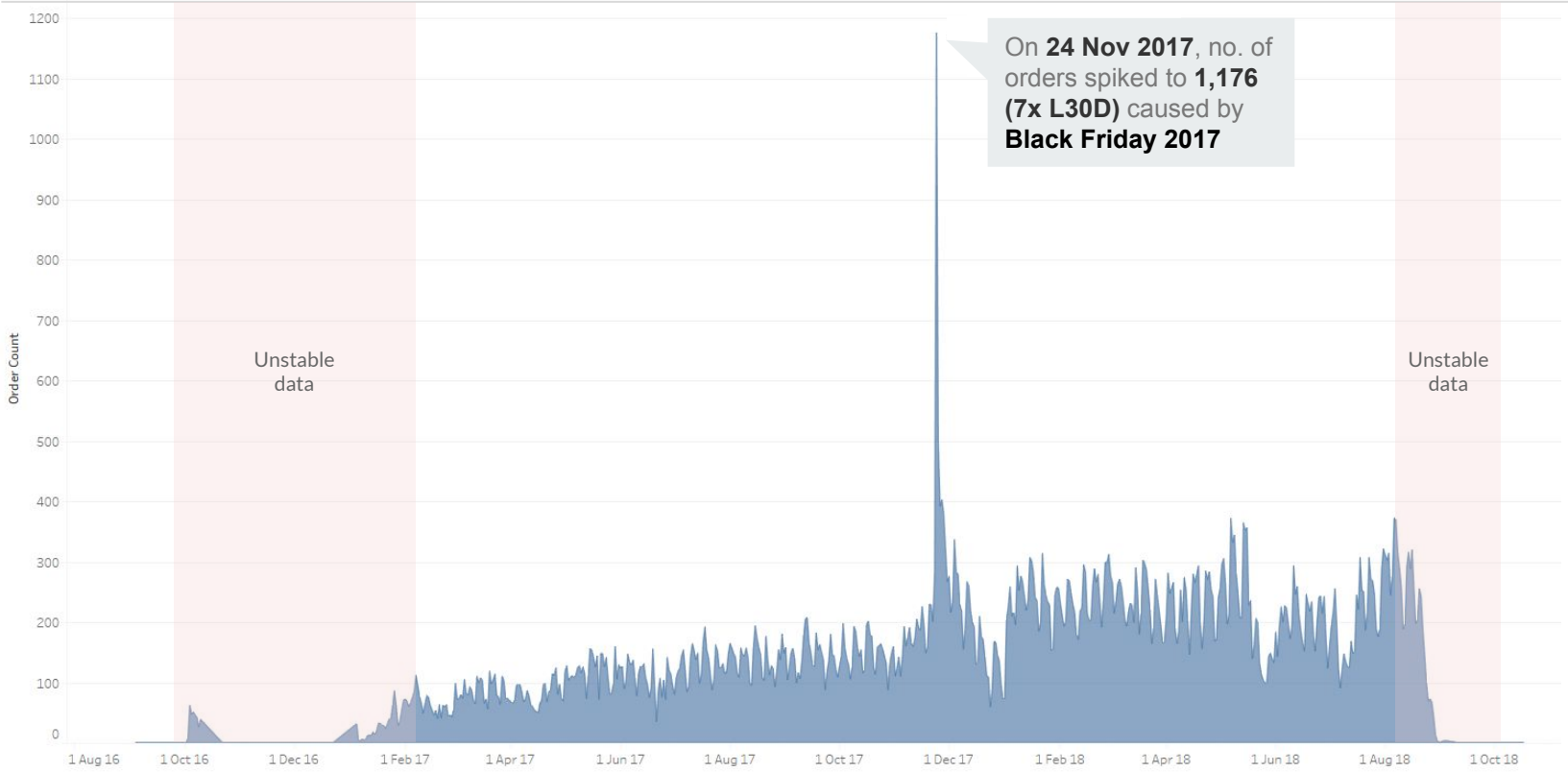
Orders: Upward trends observed on monthly orders; unreliable order counts before Feb 2017 and after Aug 2018 to be removed

Monthly Order Distribution (by states and combined)



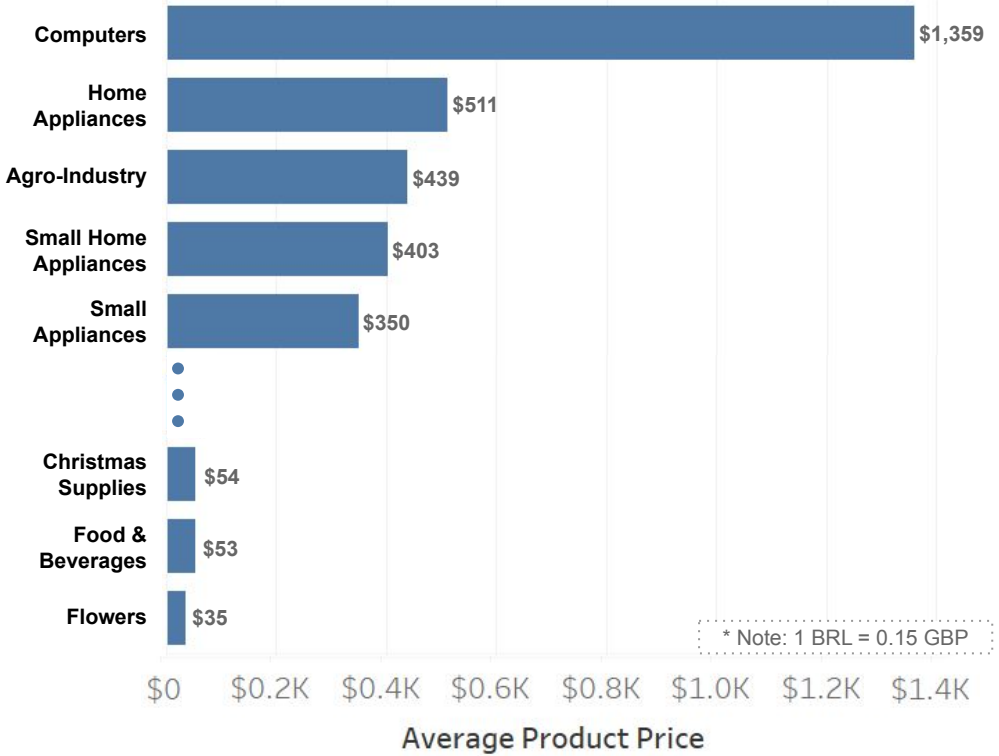
Orders: Upon further investigation, order spike uncovered on the 24 Nov 2017 due to 2017 Black Friday

Daily Order Distribution (all states)



Products: High priced products dominated by electronic-related categories, further data cleaning required to properly categorise

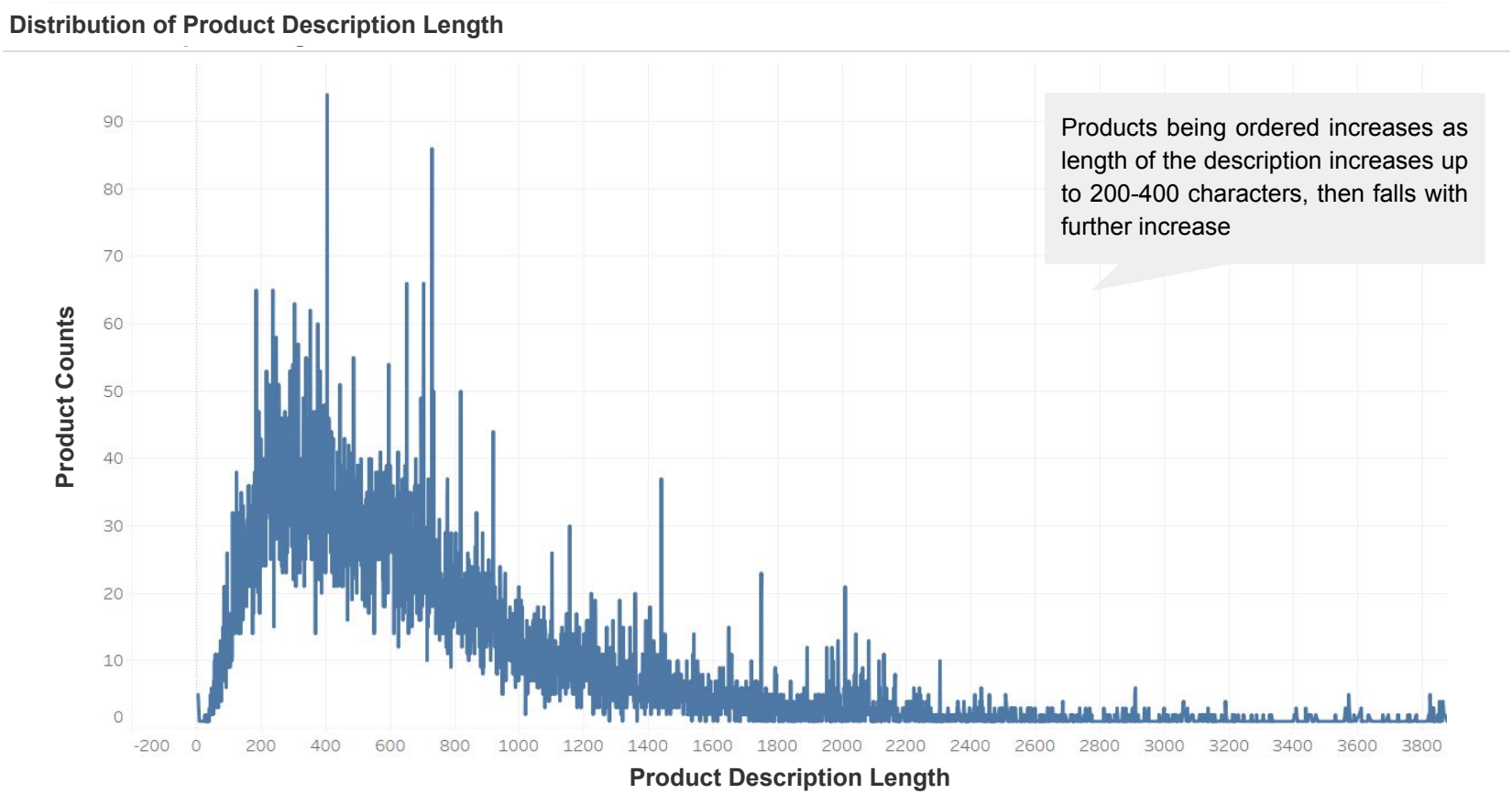
Product price distribution across different categories
(Average price in thousands of Brazilian Reals)



- Out of 77 available product categories, the highest priced are mostly **electronics**, with **computers** averaging at **~R$1,359**
- Duplicated or similar categories observed, **data cleaning and re-categorising** will be required

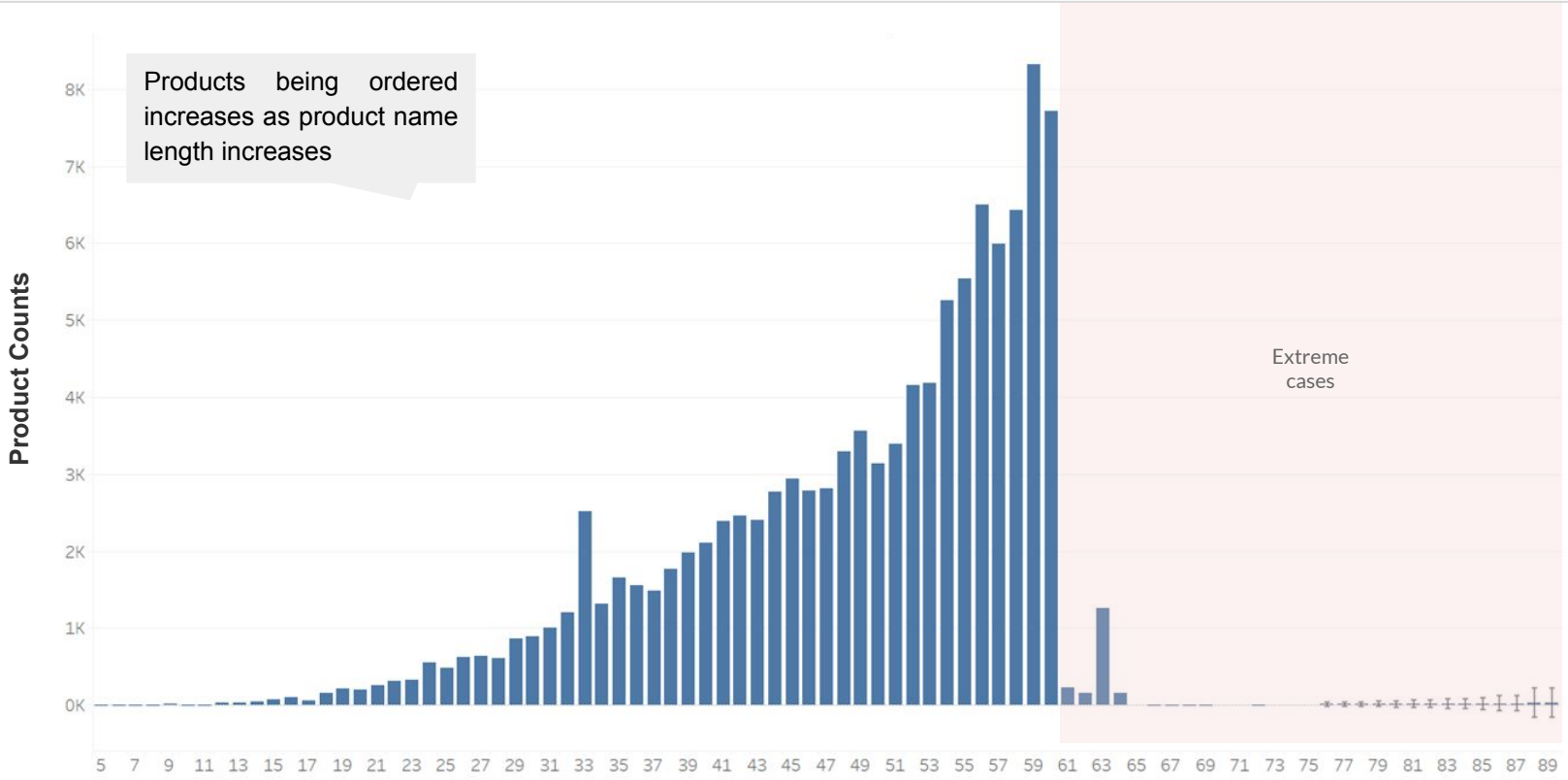
Original product category	Cleaned
home_appliances	Home Appliances
home_appliances_2	
small_appliances	
small_appliances_oven_and_coffee	
fashion_sport	Fashion
fashion_male_clothing	
fashion_female_clothing	

Products: Descriptions within 200-400 length range tend to generate more orders, but drops when too long



<u>Products:</u> Longer product names facilitate sales, further data cleaning required to eliminate effects of extreme cases

Distribution of Product Name Length



Derived features: Delivery lead time and freight-volumetric weight ratio can also be derived to better understand the order profiles

Delivery Lead Time
Delivered date - Purchased date

Freight-Volumetric Weight Ratio
Freight value ÷ (Volume ✕ Weight)

Hypotheses: Aims to value-add our investors and sellers through key explorations on customer, review and orders data

Analysis Scope	Key Hypotheses / Questions	Actionable Insights
Customer	**Customers prefer certain product types more in different city/ states** • States with high percentage of younger population have higher demand on electronics, • States with high percentage of older population have higher demand on health products	Insight used to support sellers in finding the **ideal product mix for different target states**
Review	**Customers tend to give better ratings and reviews on low-priced, fast-delivery products** • Many factors affect reviews, among which, low priced and faster delivery time may be the most important	Advocate competitive pricing to sellers and reliable 3PL outsourcing to **maintain their reputation and quality of commerce**
Order	**Products with longer titles, descriptions and more image quantity tend to generate more orders** • More information on the products allow customers to make more confident purchases	Provide **guidelines** for sellers to **potentially generate more orders**