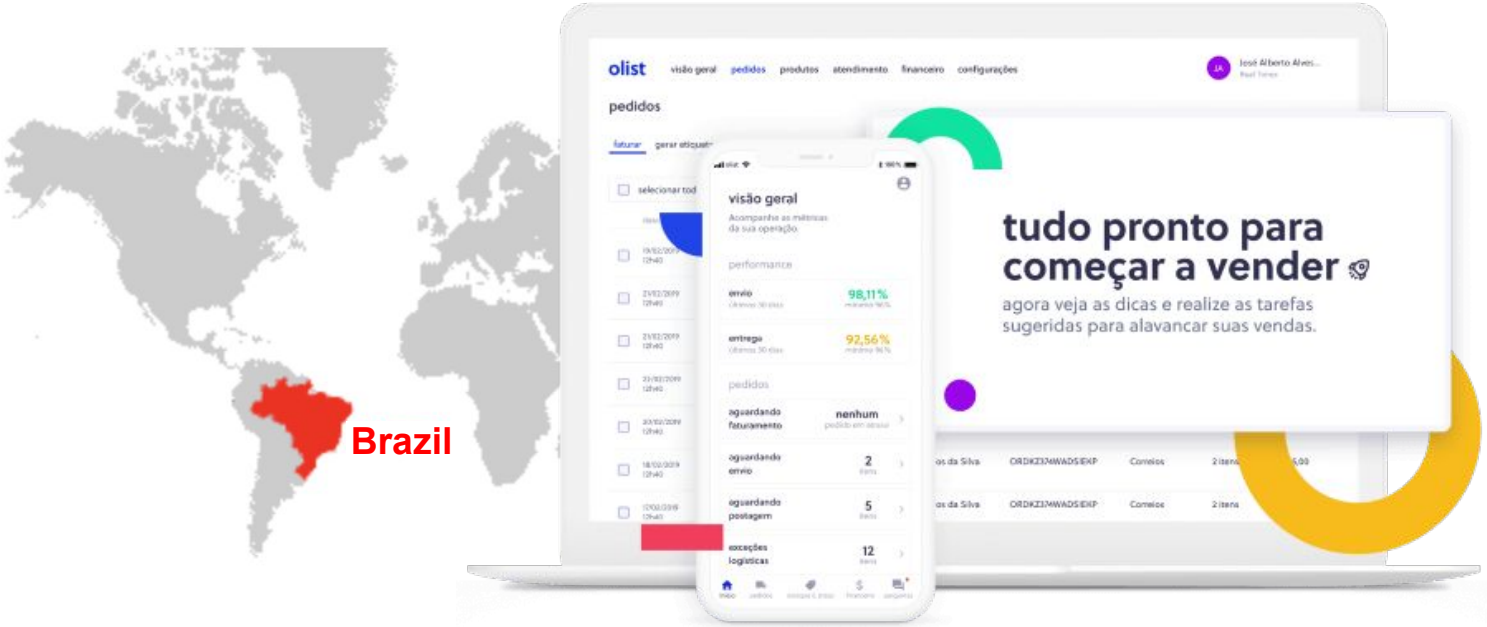
Olist E-Commerce

Final Visualisation Project Presentation

Group 16

<u>Recap:</u> Dataset from Olist e-commerce in Brazil, to provide insights and recommendations to sellers on e-commerce



Selected dataset is on Olist orders, **an e-commerce platform in Brazil on orders between Oct 2016 to Oct 2018**. Our objective is to provide insights to improve e-commerce performance of our target audiences, i.e. **the potential and existing sellers** based on three proposed key hypotheses.

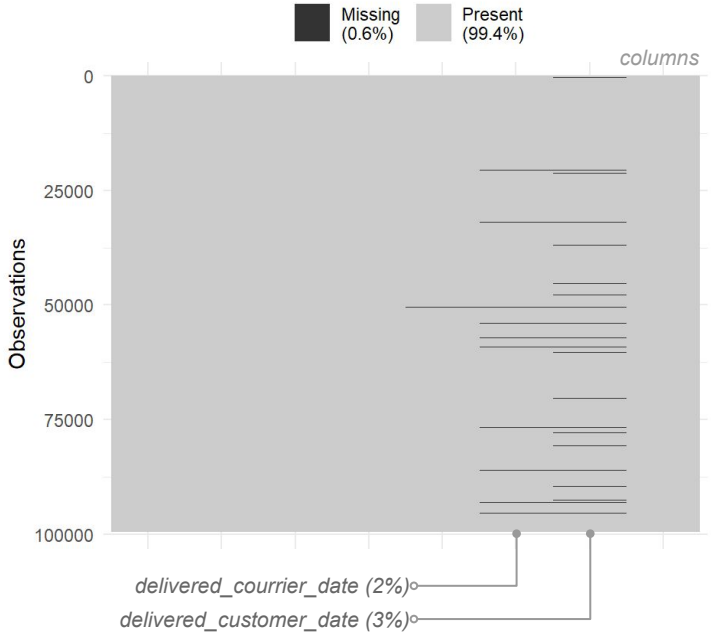
Data Wrangling & Pre-processing

- Missing values
- Orders data across time period
- Geolocation data
- Product category data

<u>Missing Values:</u> Found in Orders and Reviews data at ~0.6% and ~20.9% respectively; to remove missing Orders data only

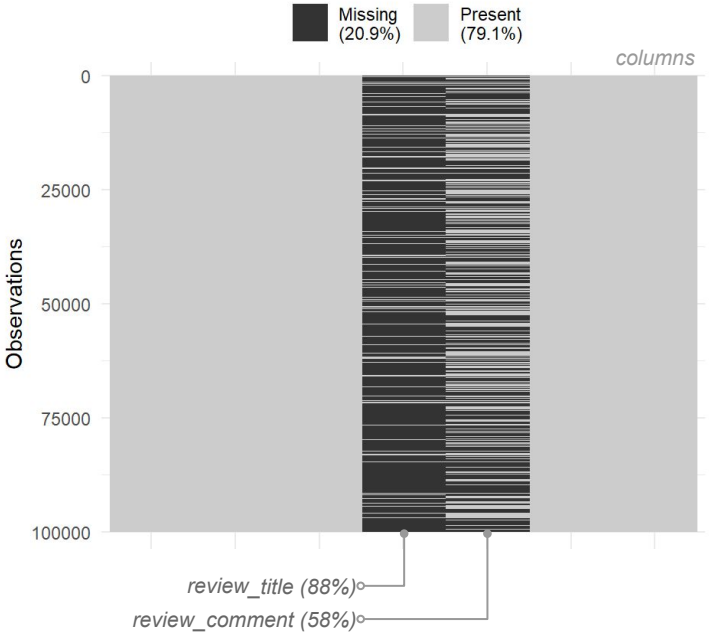
Missing values in Orders data found ~0.6%

Rows with missing values to be removed for data cleanliness

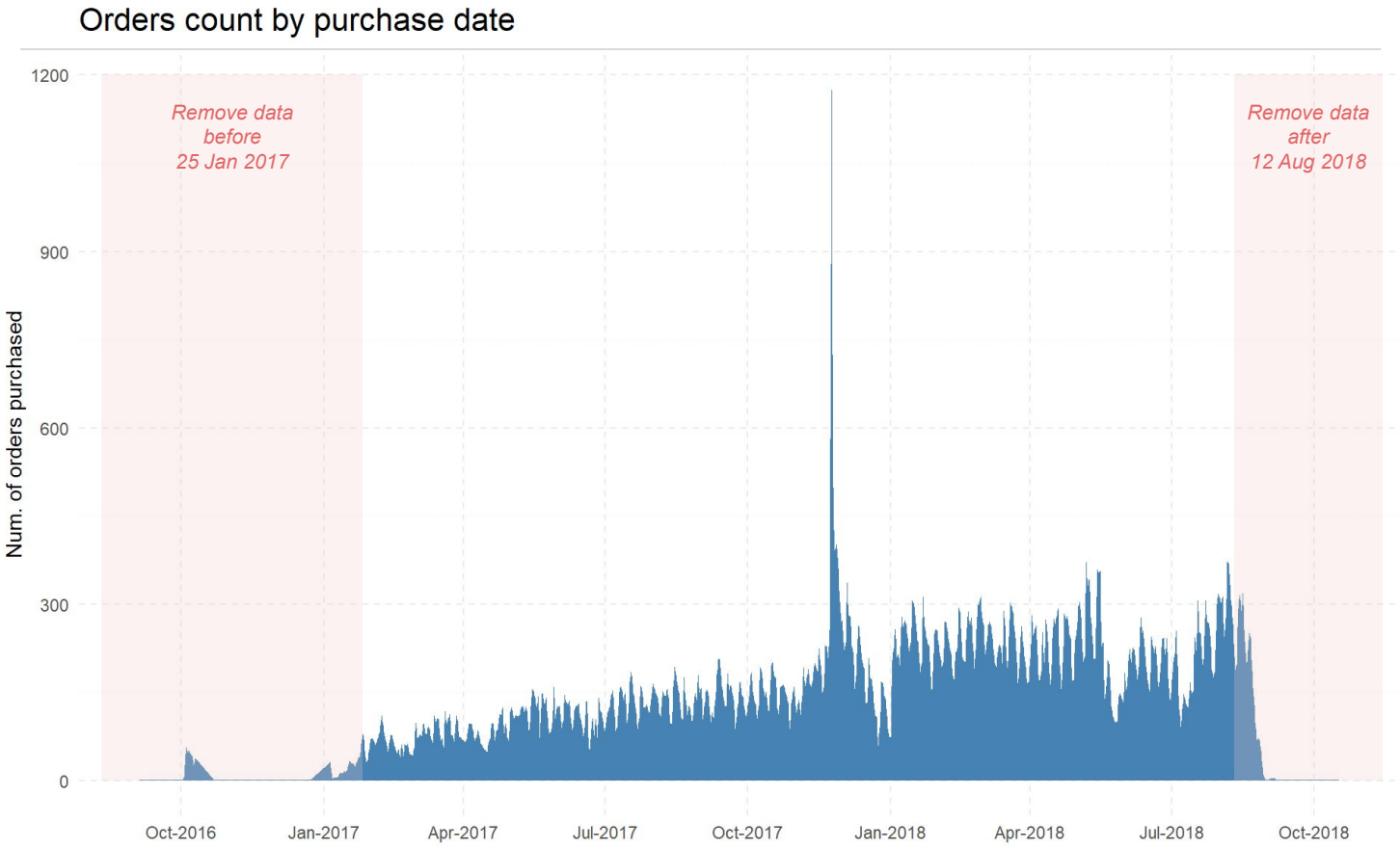


Missing values in Reviews data found ~20.9%

Missing values are expected and require no data cleaning



<u>Orders:</u> Suspicious order counts before 25 Jan 2017 and after 12 Aug 2018; to remove data points outside the time period (~4.0%)



<u>Geolocation:</u> Duplicated geolocation zip codes around ~98.1% and inconsistent spelling variations on city names

Zip Code Prefix	City	State	Latitude	Longitude
01001	são paulo	SP	-23.55064	-46.63441
01001	sao paulo	SP	-23.55050	-46.63434
01001	sao paulo	SP	-23.54978	-46.63396
01001	saopaulo	SP	-23.55050	-46.63434
01001	sao£ paulo	SP	-23.55143	-46.63407
01001	são paulo	SP	-23.55143	-46.63407

Geolocation data issues:

- Duplicated zip code (key) around **~98.1%**
- Inconsistent city names such as **sao£ paulo** (random characters), **saopaulo** (no spacing) and **são paulo** (accent)

01001	saopaulo	SP	-23.55064	-46.63441
01001	saopaulo	SP	-23.55050	-46.63434
01001	saopaulo	SP	-23.54978	-46.63396
01001	saopaulo	SP	-23.55143	-46.63407

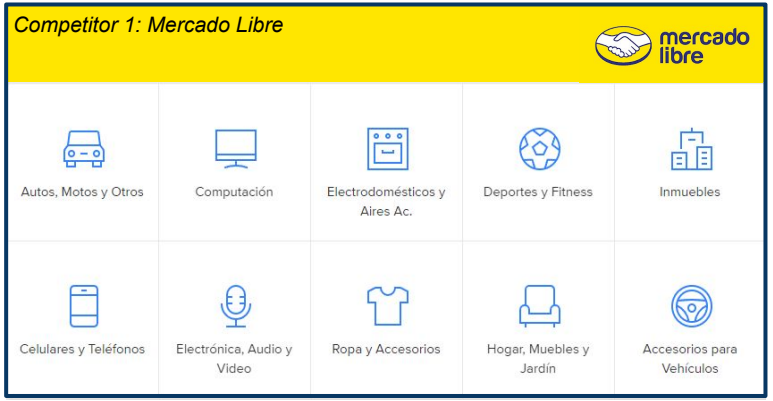
Clean geolocation data:

- Remove accents, special characters and strip the spaces
- Remove duplicated rows to prevent skew in centroids

01001	saopaulo	SP	-23.55059	-46.63420

Aggregate into centroid

Category: Product category data are re-categorised by referencing product categories from main competitors



Re-categorisation of product category

Drilled-down into 20 new category groups from 71

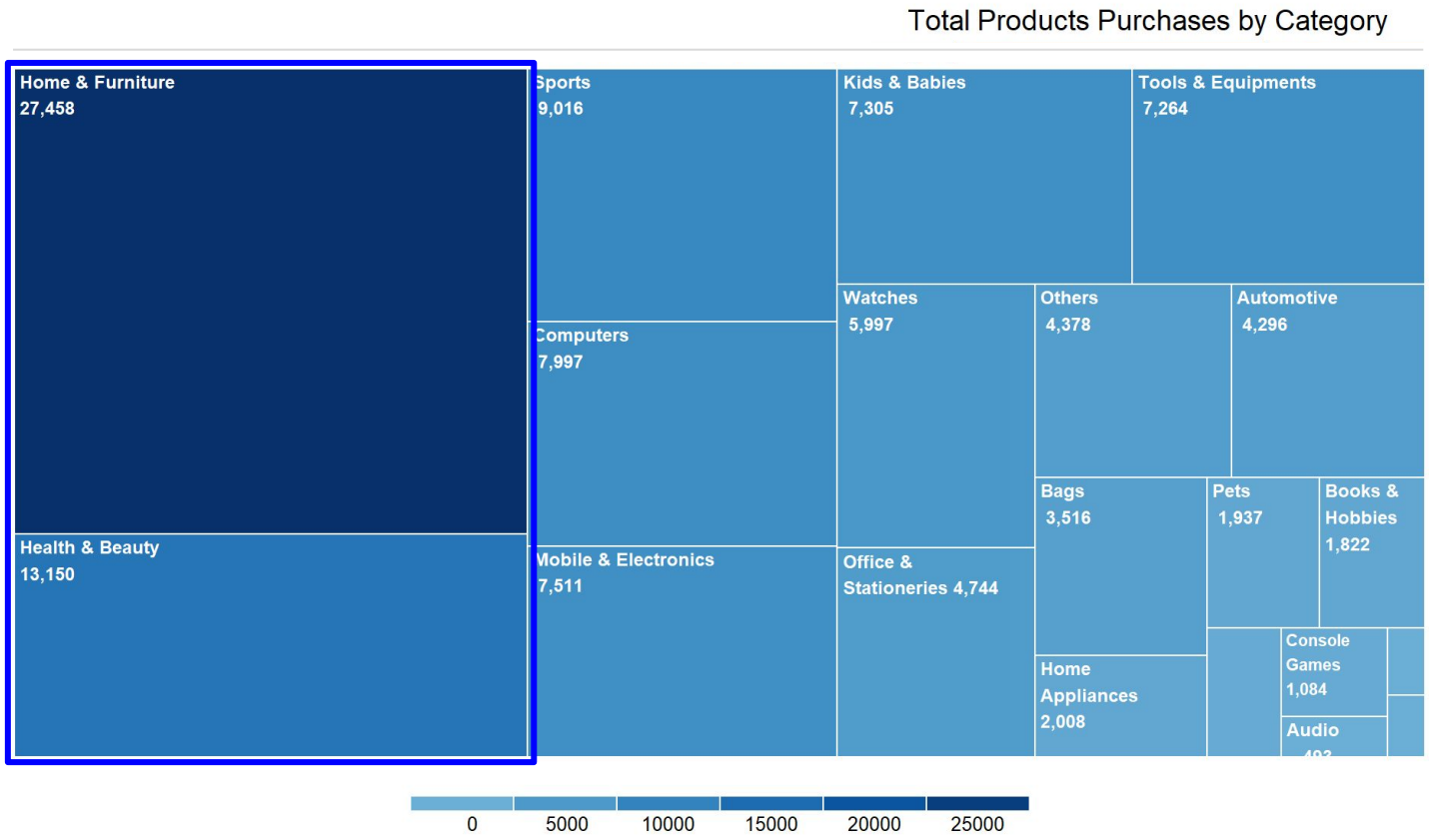
Original product category	Recategorised
home_appliances	Home Appliances
home_appliances_2	
small_appliances	
small_appliances_oven_and_coffee	
fashion_sport	Fashion
fashion_male_clothing	
fashion_female_clothing	

- New category groupings are referenced from **Olist's e-commerce competitors** in Brazil, i.e. Mercado Libre and Shopee
- Duplicated categories are grouped and recategorised manually from **71** into **20 new high-level categories**

Hypotheses & Insights

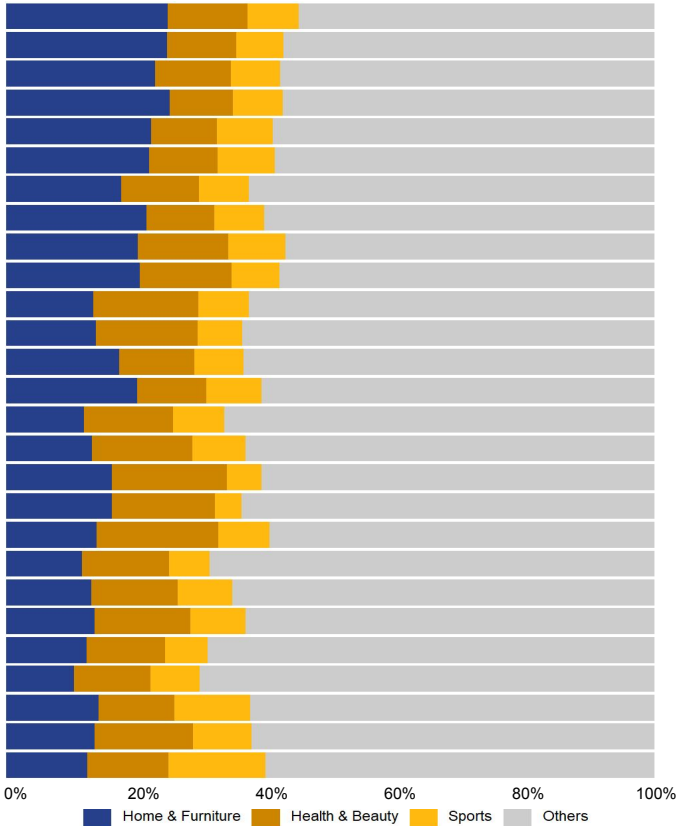
1. **Customers:** Customer's product preferences are different across states
2. **Orders:** Higher orders generated on products with greater product details
3. **Reviews:** Review ratings are significantly affected by delivery efficiency

1 Customers: Home & Furniture is the most preferred, generating ~27.5k orders, followed by Health & Beauty, at ~13.1k orders



1 Customers: Most popular products are Home & Furniture in 15 states and Health & Beauty in 10 states, takes up to 25% of orders

Brazil's Top 3 Category Proportion by States

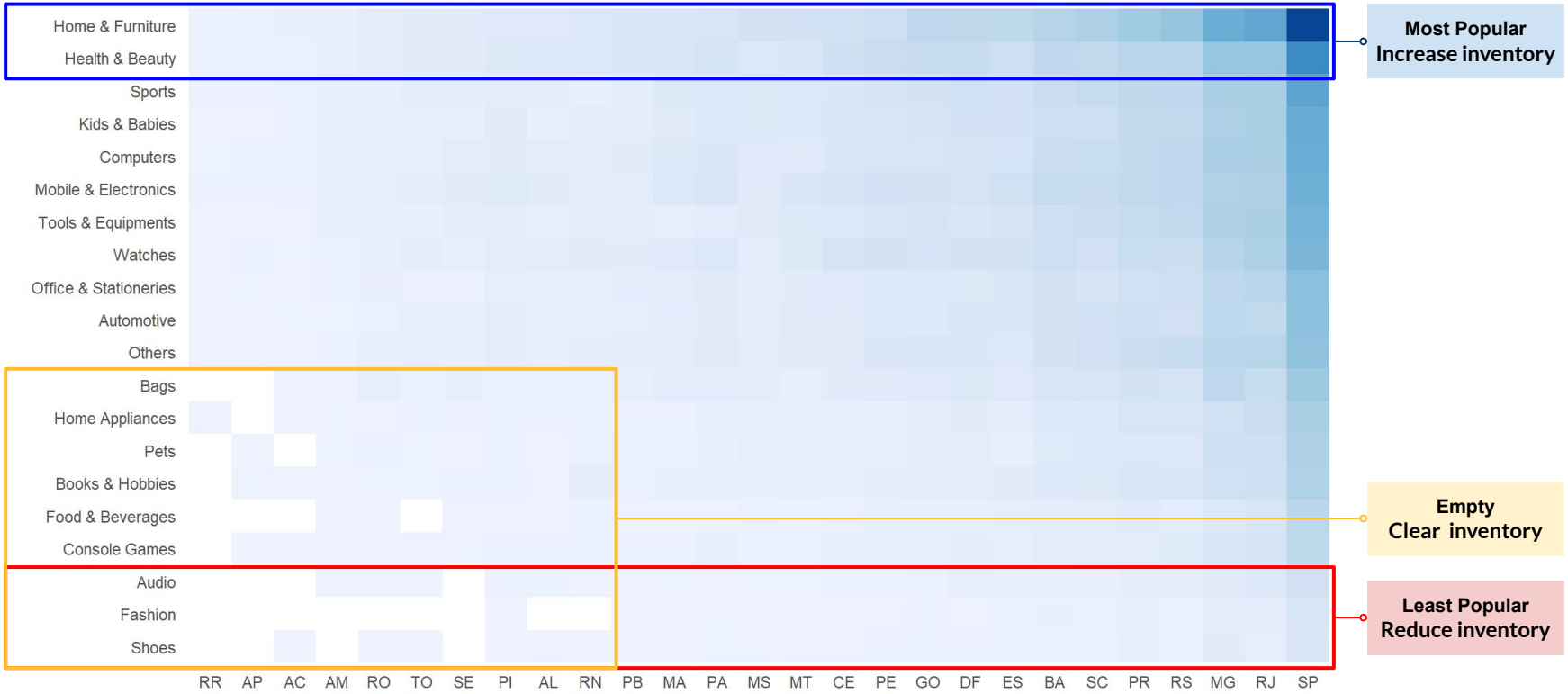


	States	Top category	Proportion
SP	Sao Paulo	Home & Furniture	24.9%
RJ	Rio de Janeiro	Home & Furniture	24.8%
MG	Mines Gerais	Home & Furniture	23.0%
RS	Rio Grande do Sul	Home & Furniture	25.2%
PR	Parana	Home & Furniture	22.3%
SC	Santa Catarina	Home & Furniture	22.1%
BA	Bahia	Home & Furniture	17.7%
ES	Espirito Santo	Home & Furniture	21.6%
DF	Distrito Federal	Home & Furniture	20.3%
GO	Golas	Home & Furniture	20.6%
PE	Pernambuco	Health & Beauty	16.1%
CE	Ceara	Health & Beauty	15.6%
MT	Mato Grosso	Home & Furniture	17.4%
MS	Mato Grosso do Sul	Home & Furniture	20.2%
PA	Para	Health & Beauty	13.7%
MA	Maranhao	Health & Beauty	15.5%
PB	Paraiba	Health & Beauty	17.8%
RN	Rio Grande do Norte	Home & Furniture	16.3%
AL	Alagoas	Health & Beauty	18.7%
PI	Piaui	Health & Beauty	13.4%
SE	Sergipe	Health & Beauty	13.4%
TO	Tocantins	Health & Beauty	14.8%
RO	Rondonia	Home & Furniture	12.4%
AM	Amazonas	Mobile & Electronics	12.5%
AC	Acre	Home & Furniture	14.2%
AP	Amapa	Health & Beauty	15.2%
RR	Roraima	Mobile & Electronics	15.0%

1 Customers: Similar preferences across states, most popular being Home & Furniture, and the least popular Audio, Fashion and Shoes

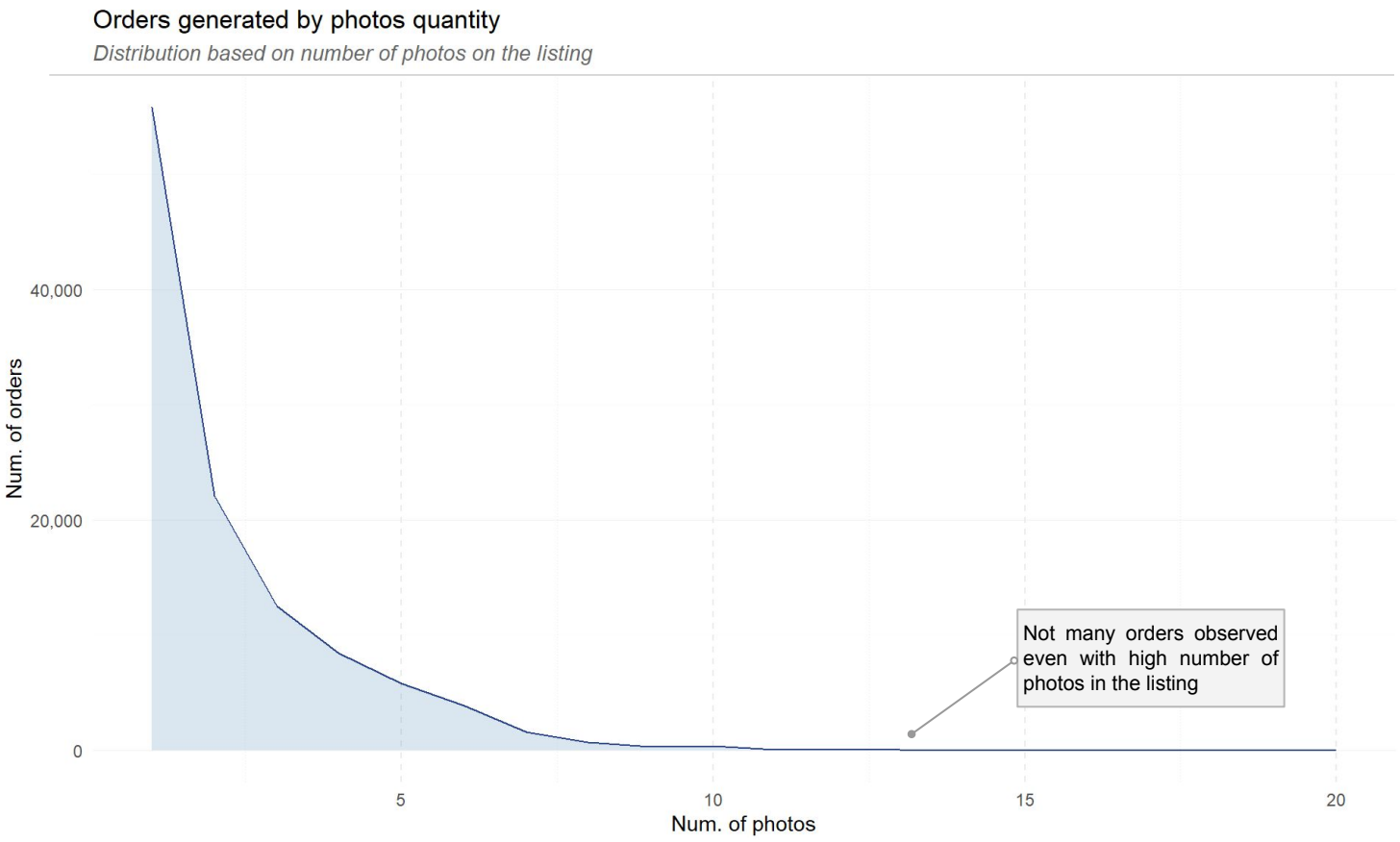
Total Orders Generated

Orders by product category and customer's state

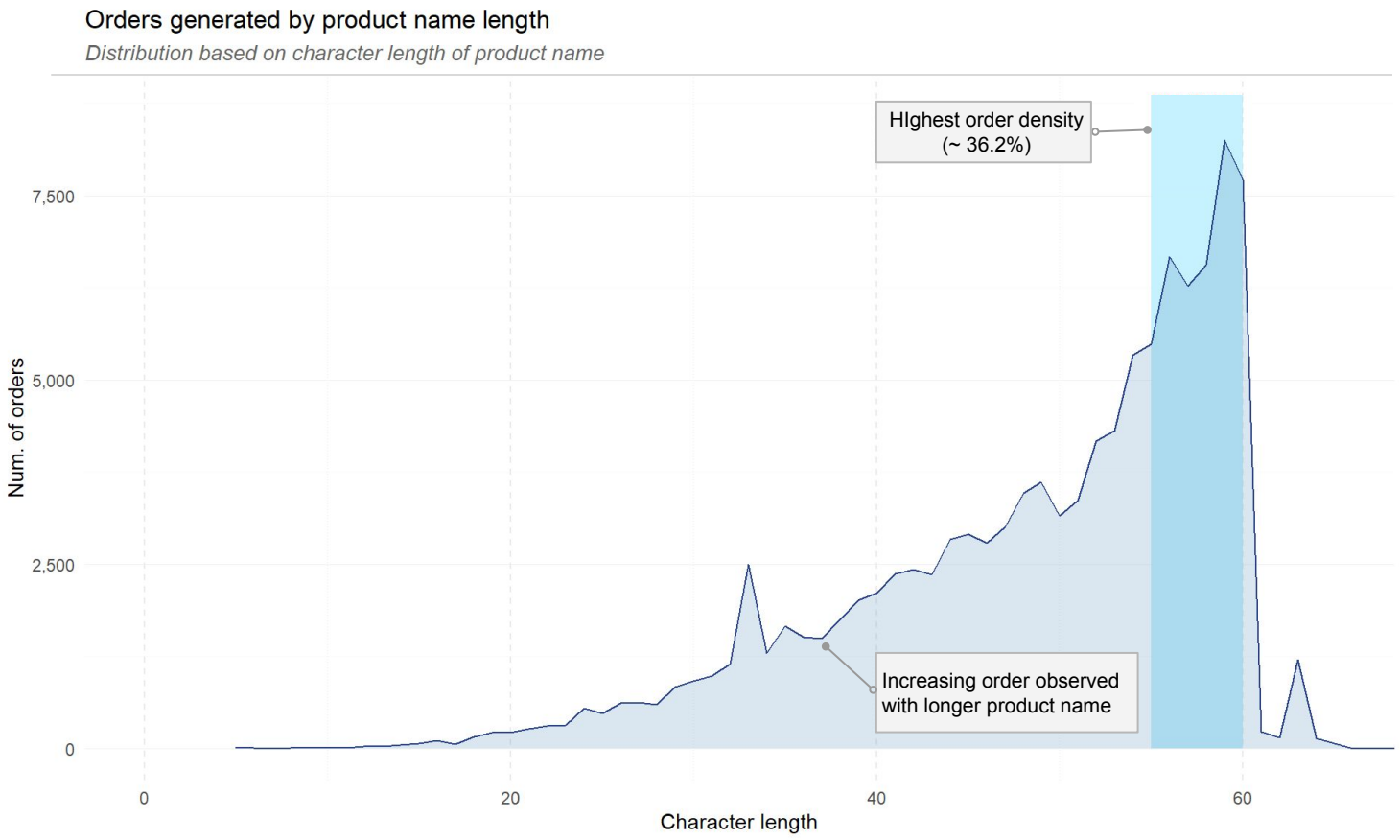


Data: Olist Public Dataset Feb 2017 to Aug 2018

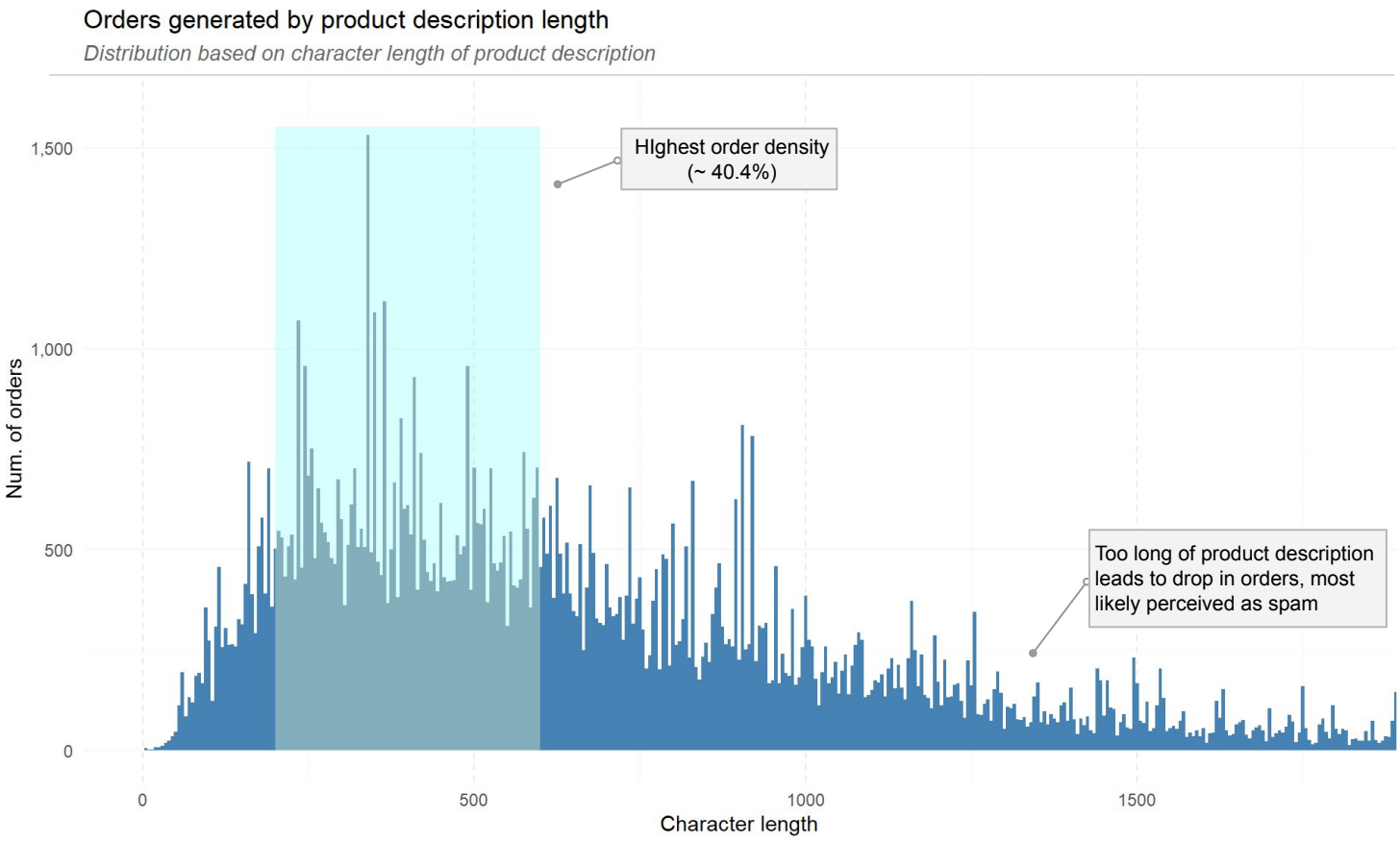
2 Orders: Increasing photos quantity does not affect the number of orders generated



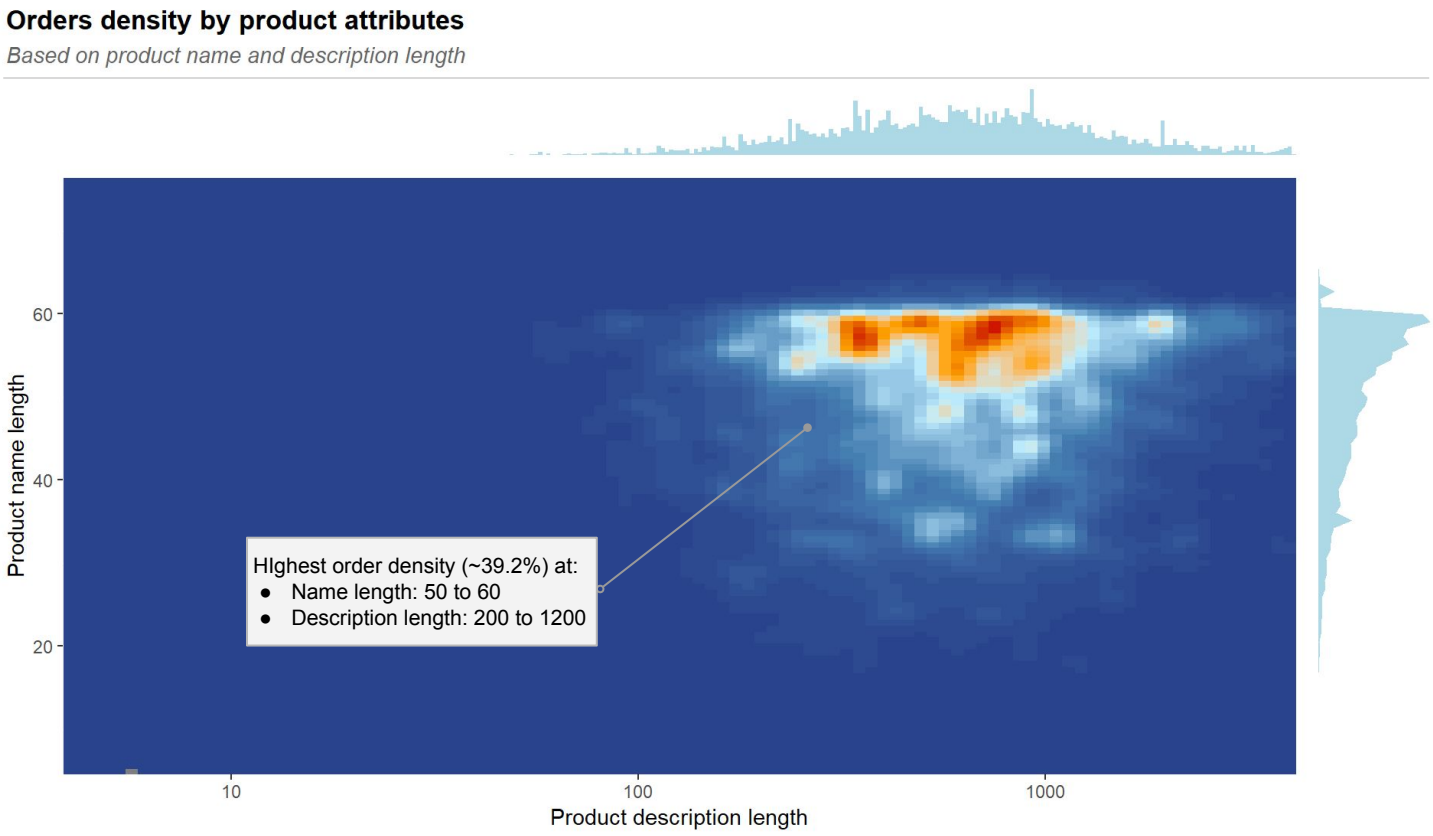
2 <u>Orders:</u> Increase in orders with longer product name, highest orders observed at approximately 55 to 60 product name length



2 Orders: Highest orders observed at around 200 to 600 product description length, and drops when description gets too long

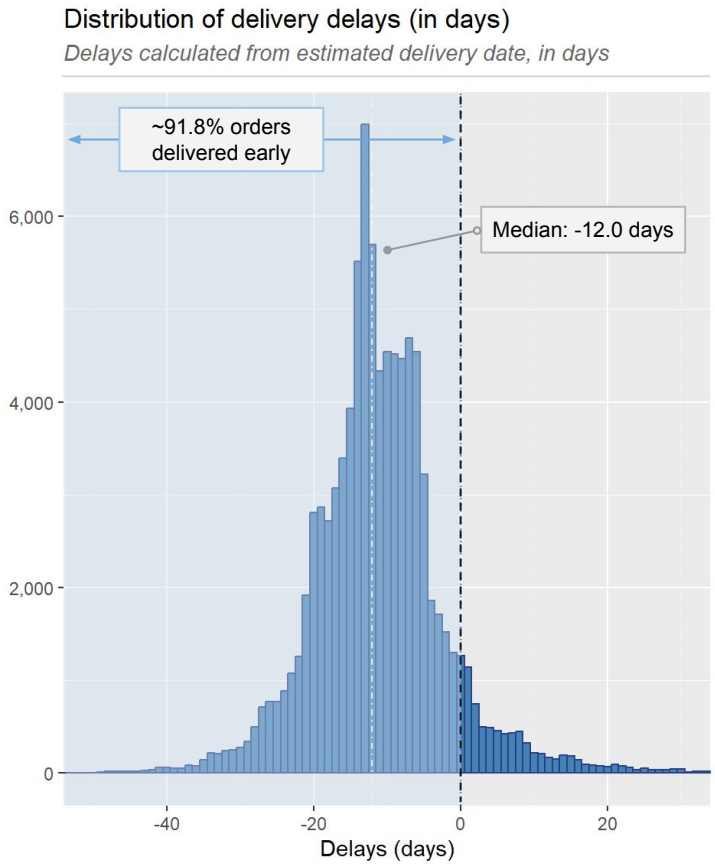


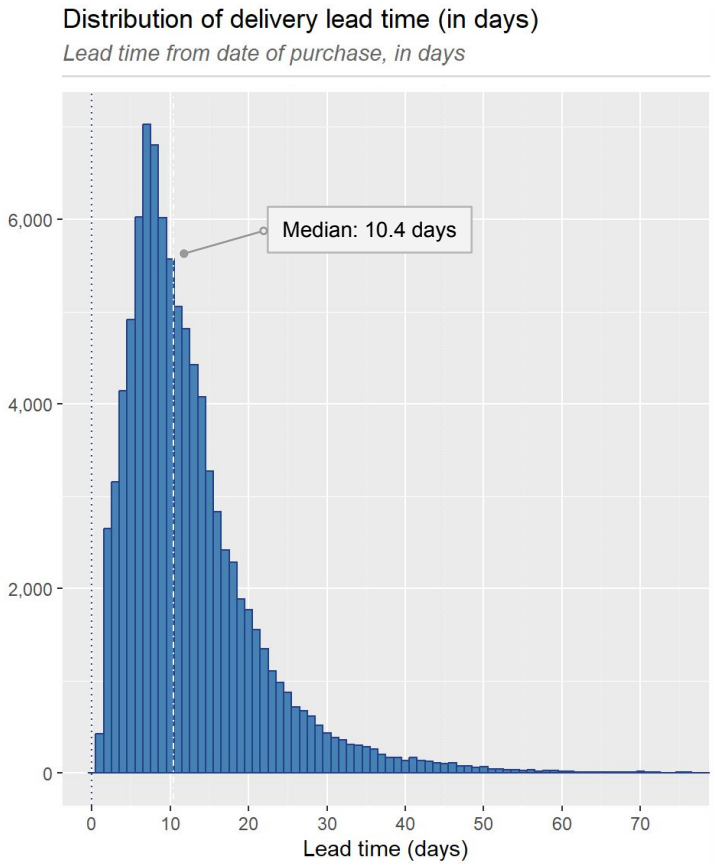
2 Orders: Highest order density observed when length of product name between 50-60 and description between 200-1200



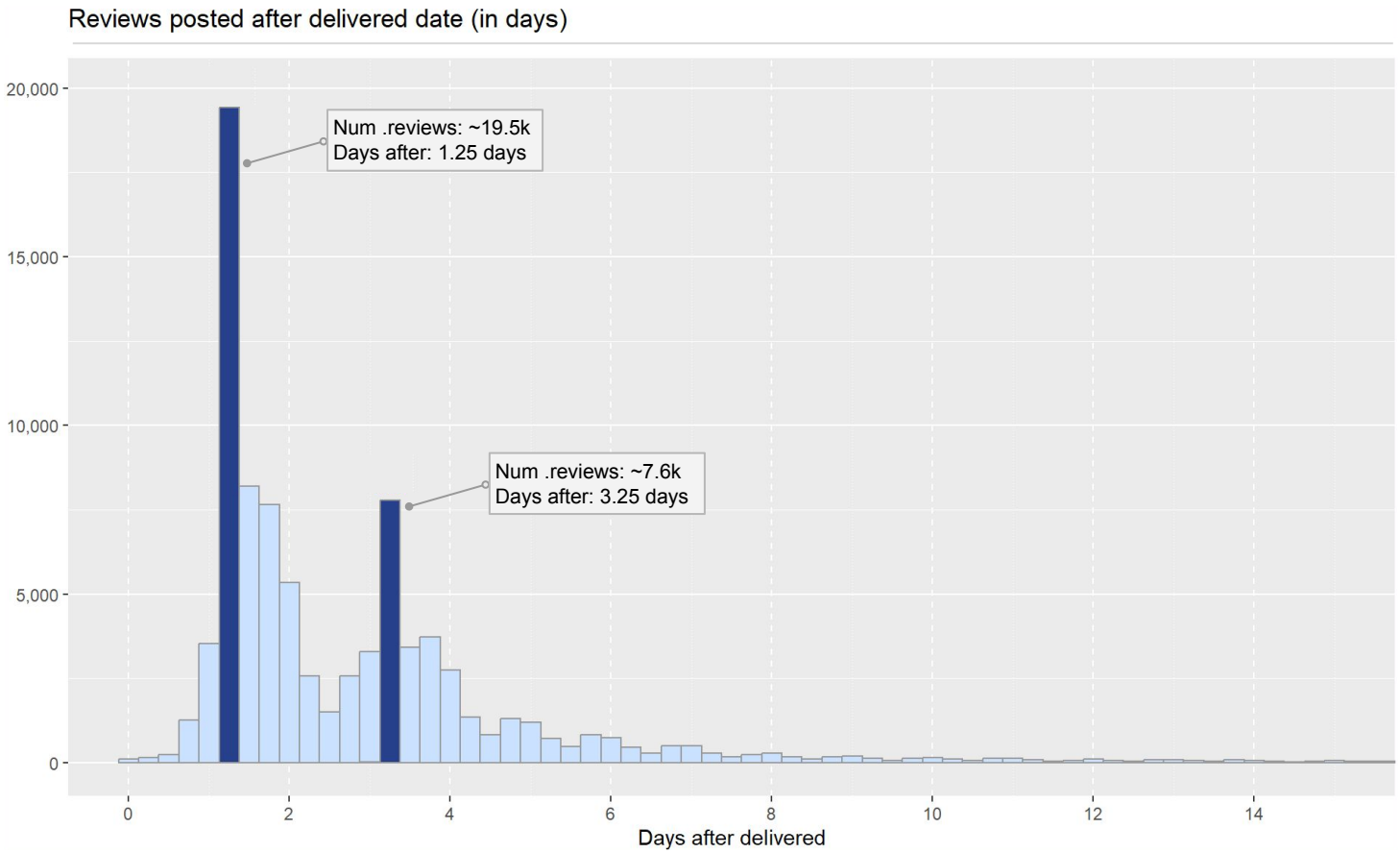
Data: Olist Public Dataset Feb 2017 to Aug 2018

3 Reviews: About ~92% orders arrived earlier than estimated, with median at ~12 days early; Median delivery lead time at ~10 days





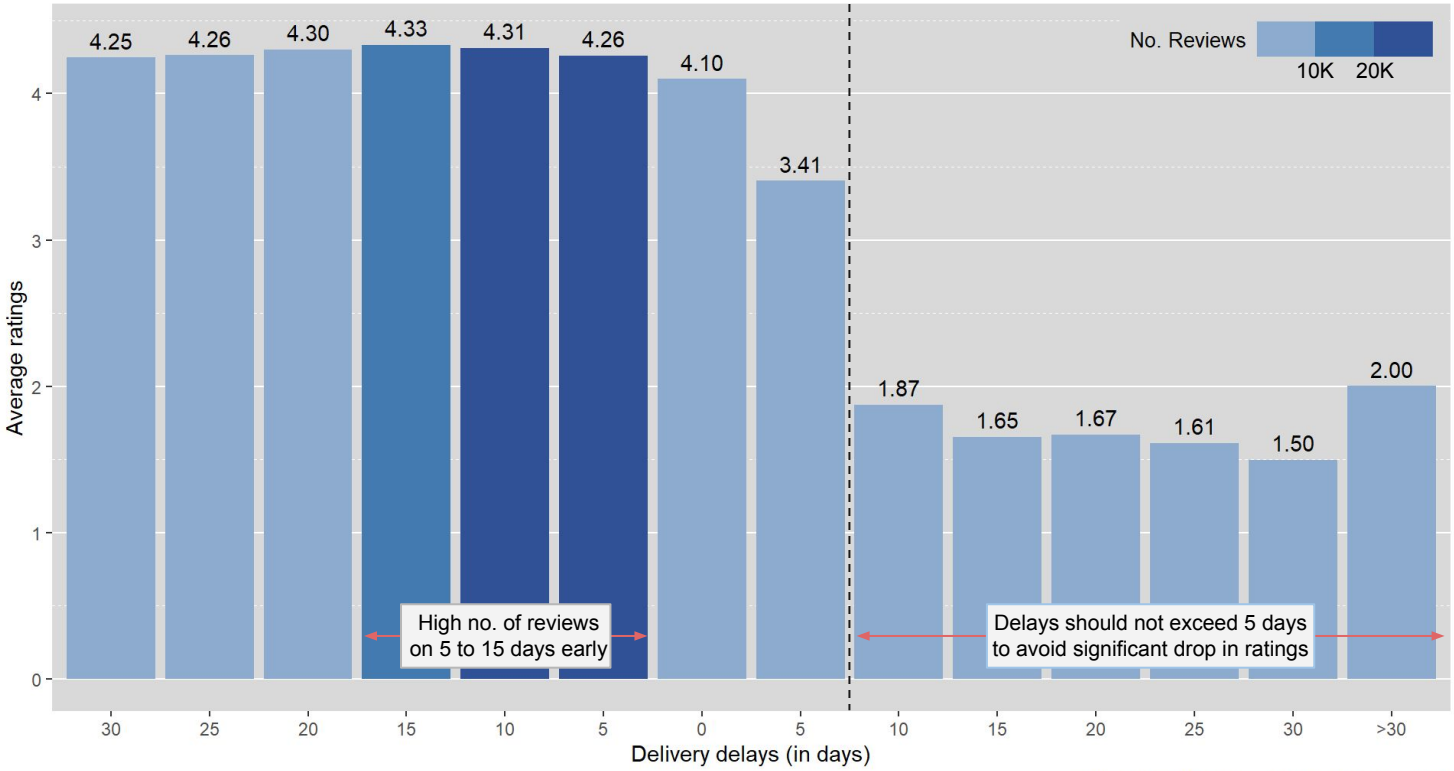
3 Reviews: Around 68% customers leave a review by day +3.25 after receiving the orders, peaking at +1.25 and +3.25 days post-delivery



3 Reviews: Review scores ranges between 4.1 - 4.3 when delivery is early, but drops to 1.5 - 2 when delayed longer than 5 days

Review score against delays (in days)

Average review scores against given delays in delivery, in days



Data: Olist Public Dataset Feb 2017 to Aug 2018

Conclusion: Sellers and investors to consider recommendations derived from the key insights over customer, review and orders

Analysis Scope	Key Insights Gained	Recommendations
1 Customer	• No big difference in preferences across states • Highest demand on Home & Furniture (H&F) and Health & Beauty (H&B) products • Lowest demand on Audio, Fashion and Shoes products	• Sellers should **expand or increase focus** on selling H&F and H&B products • **Minimise focus** on selling Audio, Fashion and Shoes products
2 Order	• Quantity of photos do not affect customer's purchasing decision • Highest order density observed on long product name (50 to 60) and medium description length (200 to 1200)	• Sellers set up **guidelines and best practices** on product name and description for listing creation
3 Review	• Majority of products ~92% arrived earlier than given estimated delivery time • Majority of customers left a review within ~3.25 days • Review ratings drops significantly when there are delays in delivery	• Sellers should avoid delivery delays by partnering with **reputable 3PLs** and improve **delivery estimation accuracy** • **Marketing campaigns** to promotes timely feedback

Archived Slides I
(from current slides)

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
1 Customer	**Customers have different product type preference in different 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 ecommerce product mix for different target states**
2 Review	**Fast-delivery products lead to better ratings and review scores** • Many factors affect reviews, among which, low priced and faster delivery time may be the most important	Advocate more reliable 3PL outsourcing to **maintain their reputation and quality of commerce**
3 Order	**More orders generated when products have complete attributes, i.e. longer titles, detailed descriptions and more photos quantity** • More information on the products allow customers to make more confident purchases	Provide **guidelines** for sellers to **potentially generate more orders**

1 Hypothesis: Product preferences are different across states

- Home & Furniture is the most popular category in Brazil, generating ~27.5k orders over 1.5-year period
- Home & Furniture is the most popular product category in 15 states, while Health & Beauty in 10 states
- Preferences are similar across states, with the top being Home & Furniture and Health & Beauty, while the least popular are Audio, Fashion and Shoes

2 Hypothesis: Reviews are significantly affected by delivery efficiency

- 91.8% of orders arrived earlier than the estimated delivery time, with an average early arrival time of 13.4 days and median of 12.0 days
- 50% of shipments were received by customers within 10.4 days from date of purchase
- 68 % of customers tend to post reviews in 1.25 or 3.25 days after receiving their goods
- Review ratings average ~4.2 when delivered early, but drops to 3.4 when up to 5 days late and drops to approximately 1.7 when more than 5 days

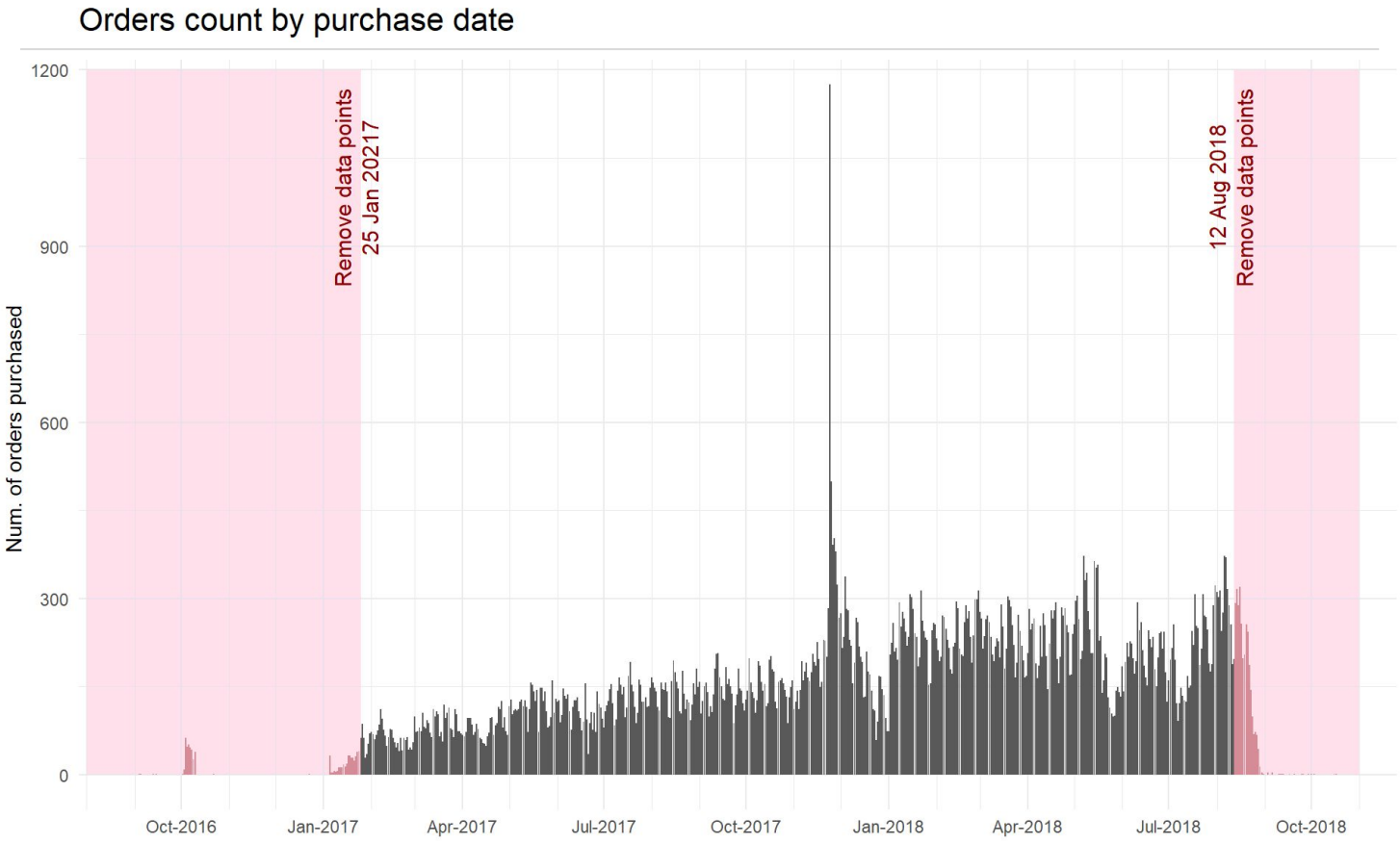
3 Hypothesis: Higher orders on products with greater details

- The number of photos included in the listing does not necessarily affect the orders generated, as we observe high orders with single-photo products
- Orders are generally higher with long product names around 50 to 60 characters and product descriptions at medium length at around 200 to 1200
- Product descriptions that are too long lead to lower orders

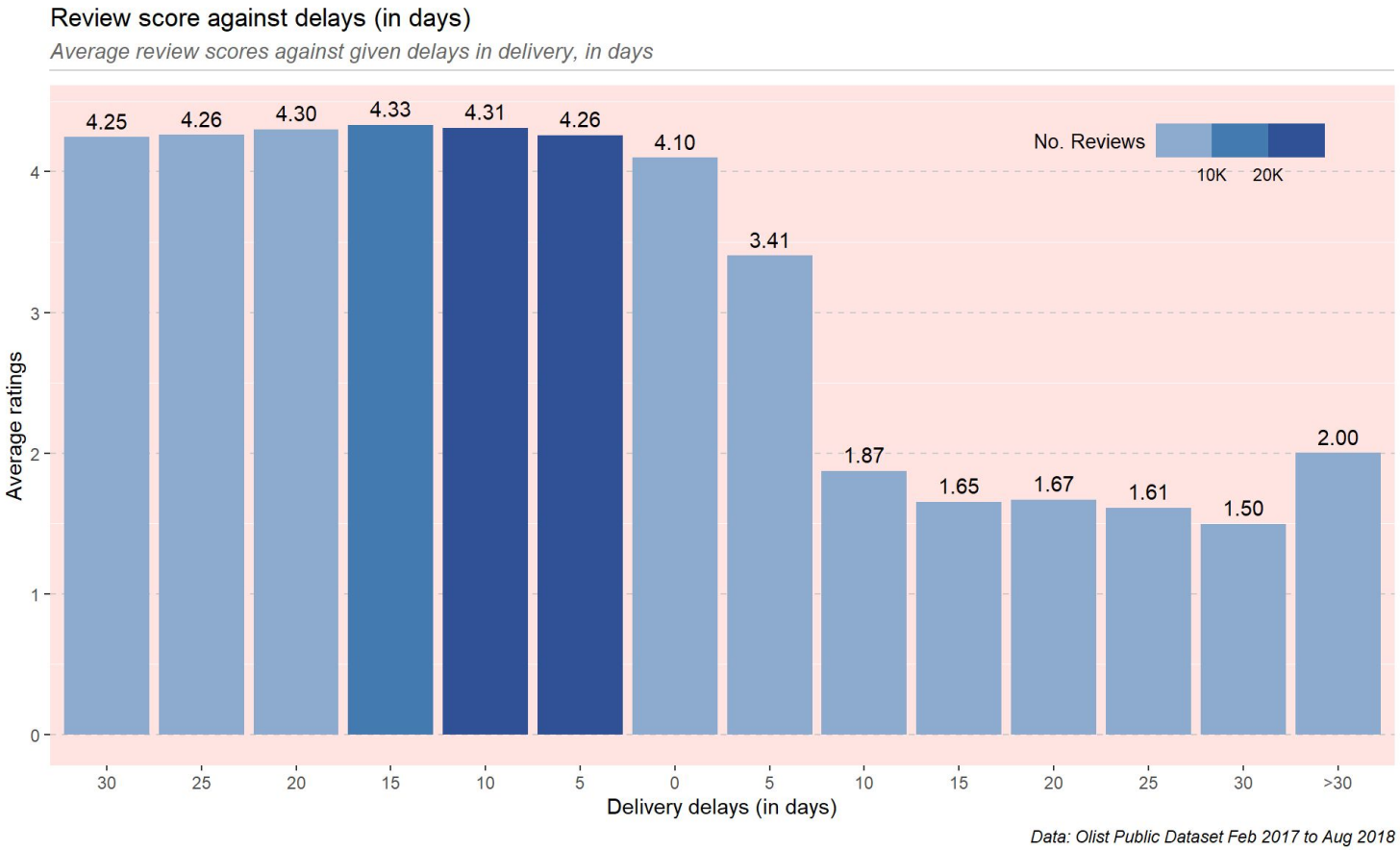
Conclusion & Recommendations

- **Customer:** Expand on selling **H&F and H&B products** in states like SP, RJ and MG, and avoid **Audio, Fashion and Shoes products** in states like RR, AP and AC to gain customer traction
- **Reviews:** Partner with reputable 3PLs and ensure proper pick-pack-and-ship SOP to avoid delivery delays
- **Orders:** Guidelines to ensure product name and description are within the desirable length

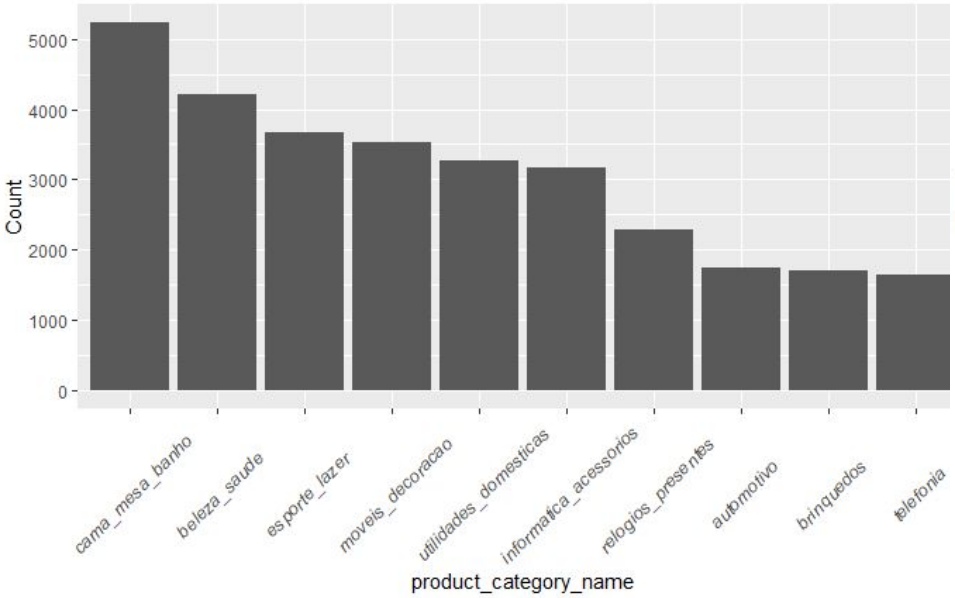
<u>Orders Data:</u> Suspicious order counts before 25 Jan 2017 and after 12 Aug 2018; to remove data points outside the time period (~4.0%)



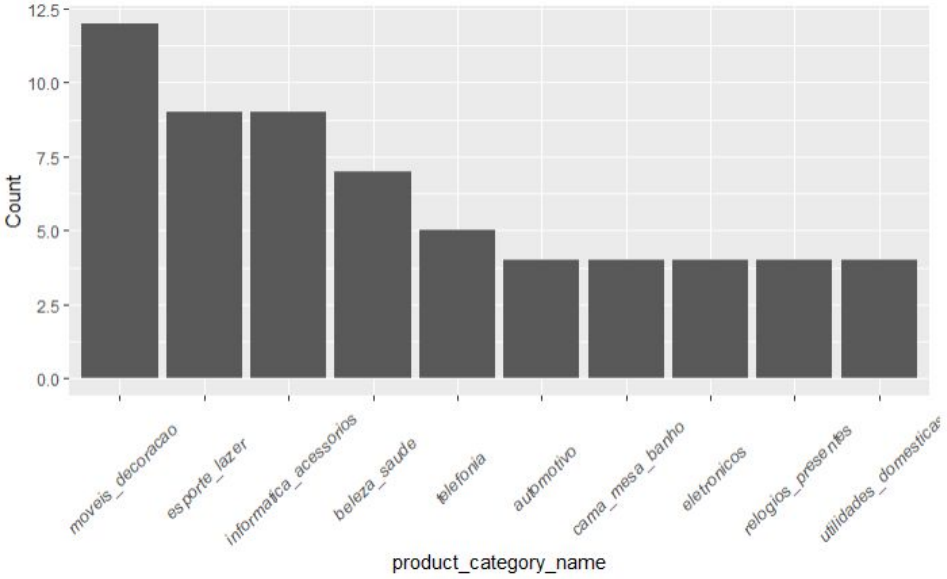
2 Insights: Review scores ranges between 4.1 - 4.3 when delivery is early, but drops to 1.5 - 2 when delayed longer than 5 days



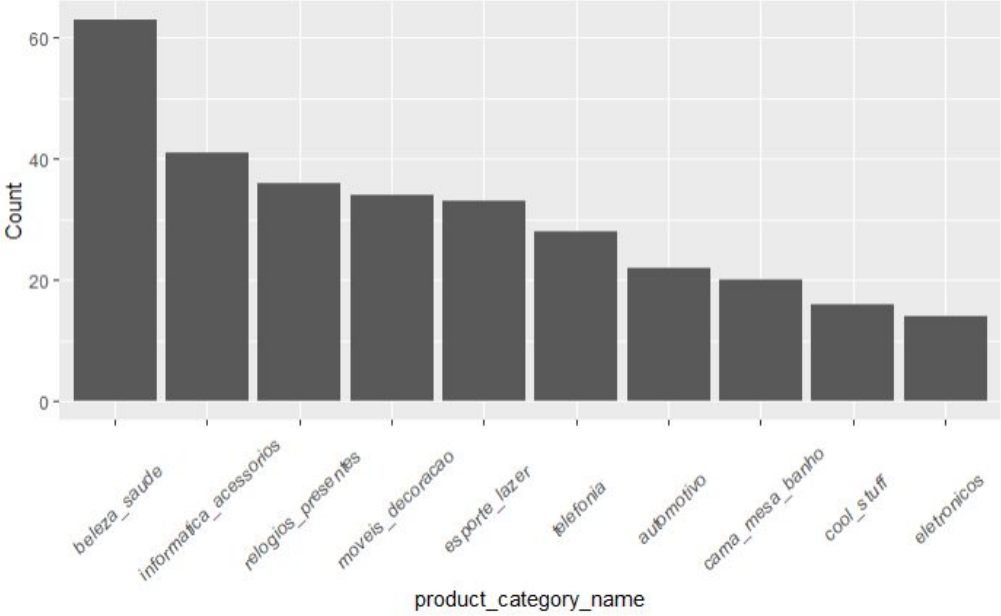
SP


Count
5000
4000
3000
2000
1000
0
cama_mesa_banho
beleza_saude
esporte_lazer
moveis_decoracao
utilidades_domesticas
informatica_acessorios
relogios_presentes
automotivo
brinquedos
telefonia
product_category_name

AC


Count
12.5
10.0
7.5
5.0
2.5
0.0
moveis_decoracao
esporte_lazer
informatica_acessorios
beleza_saude
telefonia
automotivo
cama_mesa_banho
eletronicos
relogios_presentes
product_category_name

AL


60
40
20
0
Count
beleza_saude
informatica_acessorios
relogios_presentes
moveis_decoracao
esporte_lazer
telefonia
automotivo
cama_mesa_banho
cool_stuff
eletronicos
product_category_name

1. Overall most popular product

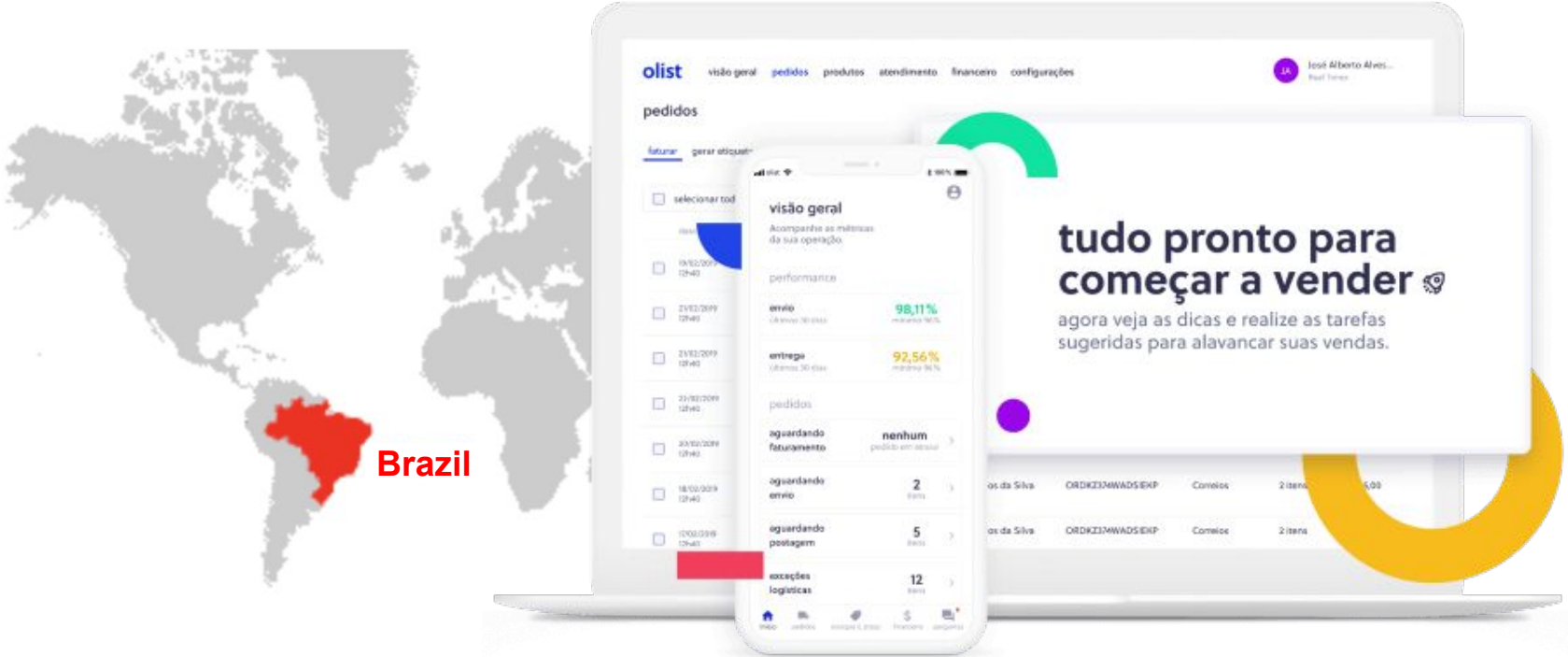
AC	AL	AM	AP	BA	CE	DF	ES	GO
furniture_decor	beleza_saude	beleza_saude	beleza_saude	beleza_saude	beleza_saude	beleza_saude	cama_mess_banha	cama_mess_banha

MA	MG	MS	MT	PA	PB	PE	PI	PR
beleza_saude	cama_mess_banha	esporte_lazer	beleza_saude	beleza_saude	beleza_saude	beleza_saude	beleza_saude	moveis_decoracao

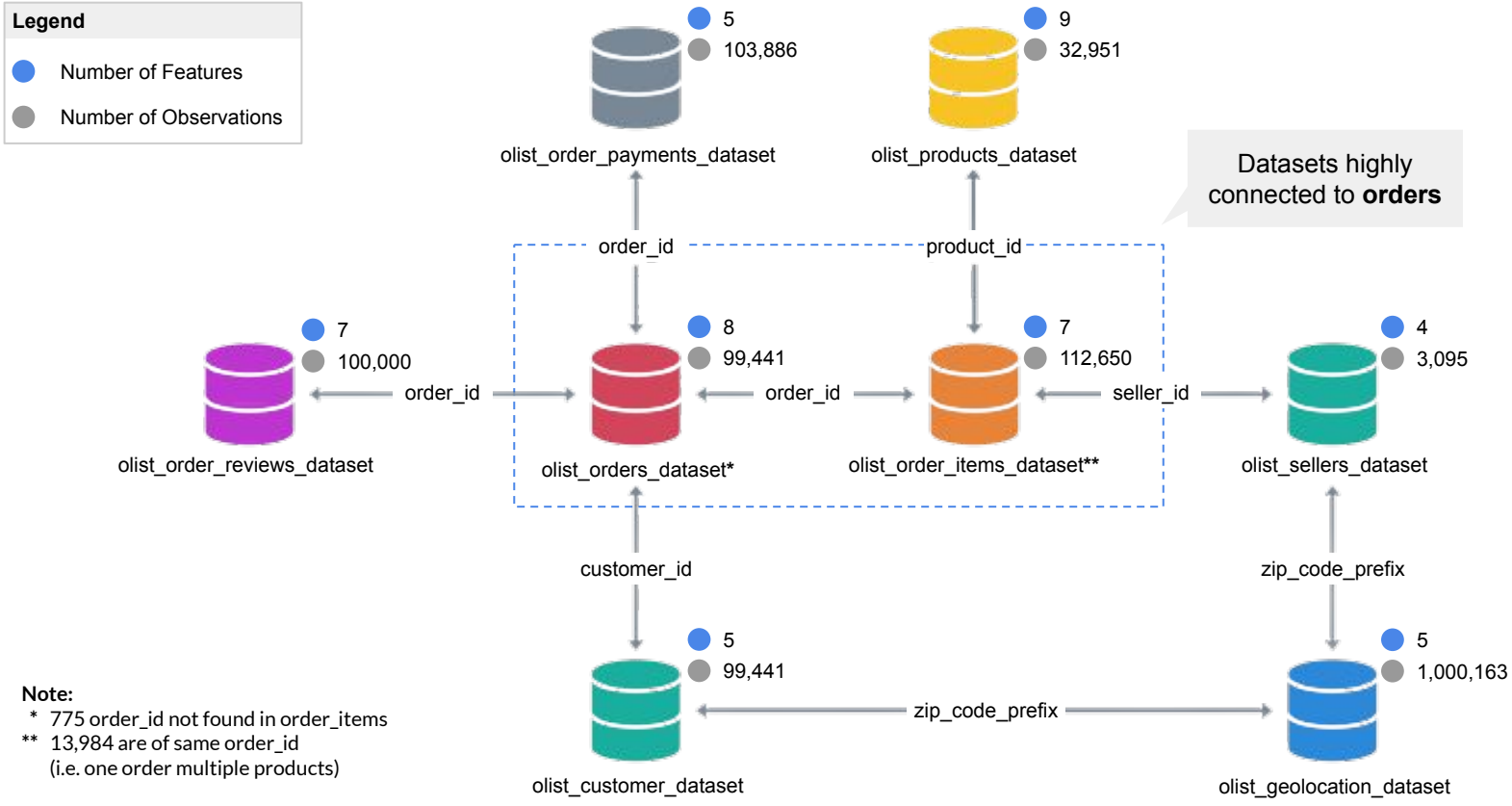
RJ	RN	RO	RR	RS	SC	SE	SP	TO
cama_mess_banha	beleza_saude	beleza_saude	esporte_lazer	cama_mess_banha	esporte_lazer	beleza_saude	cama_mess_banha	beleza_saude

Archived Slides II
(from hypotheses proposal)

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



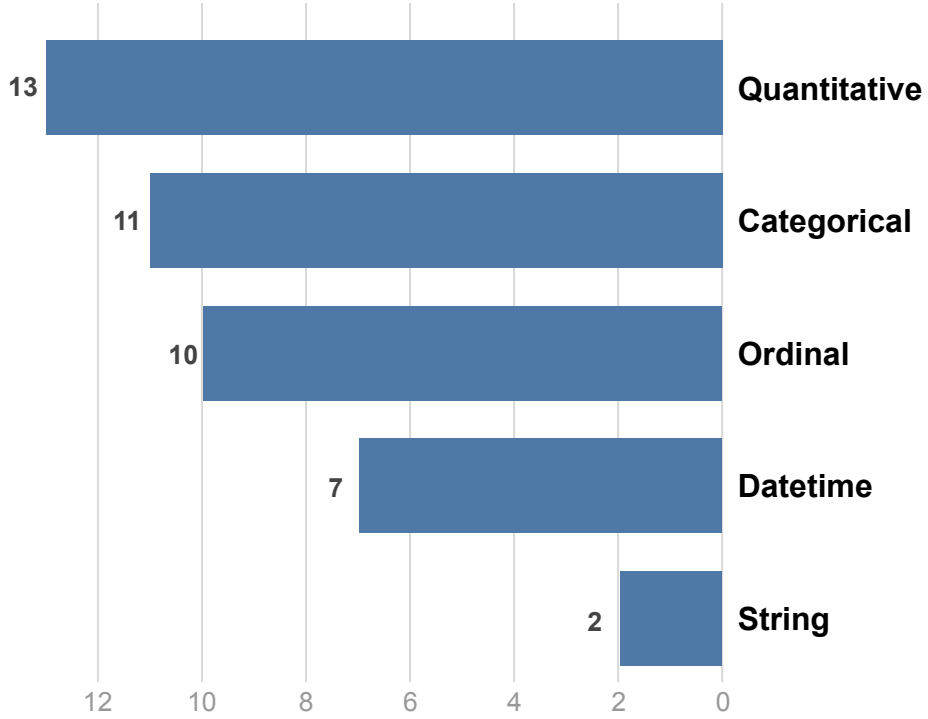
Data structure: Total 8 data sets focusing on customers, orders and products, but structure mainly centralised on orders



Note:
* 775 order_id not found in order_items
** 13,984 are of same order_id
(i.e. one order multiple products)

<u>Data types:</u> Total 5 types, with majority being quantitative data

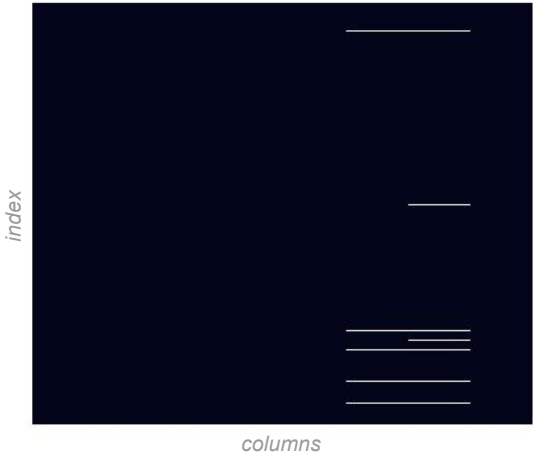
Number of Features by Data Types



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!

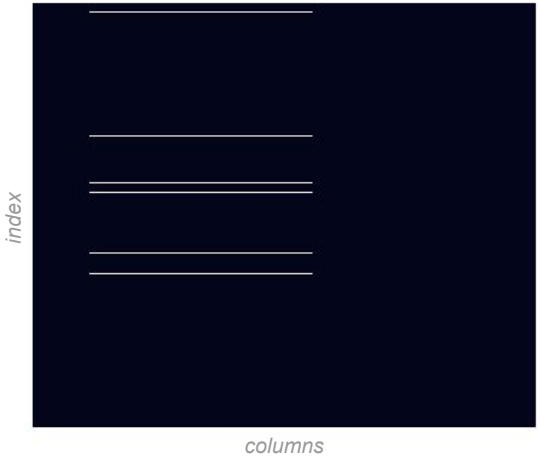
<u>Missing values:</u> Found in 3 datasets through heatmaps; around 2-3% on orders and products, and around 89% on reviews

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***



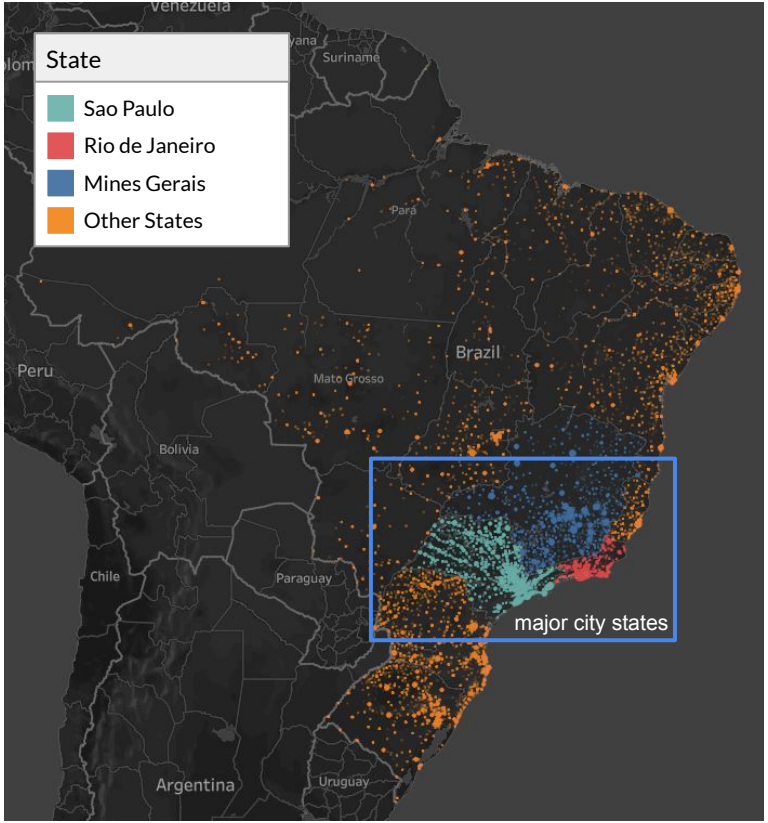
*Potentially due to **incomplete/ unpublished listings** commonly listed for testing purposes*

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

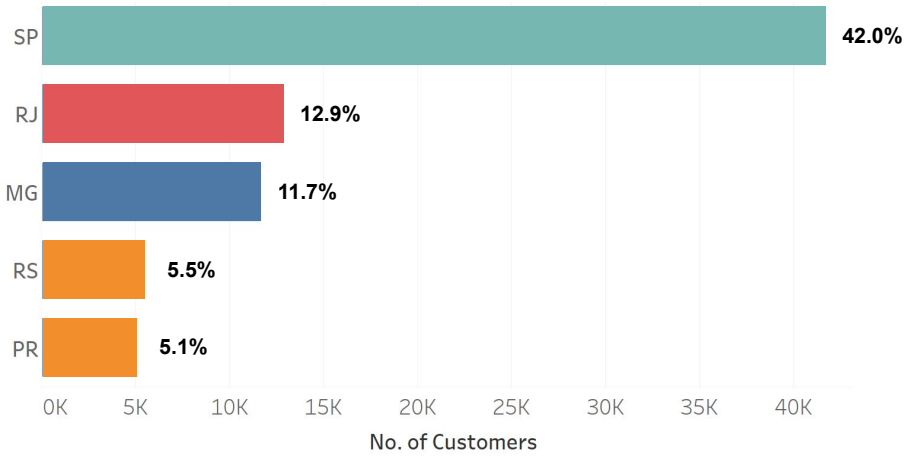
***Missing data expected** as customers often leave ratings but do not write any review*

Legend: Null values Non-null values

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



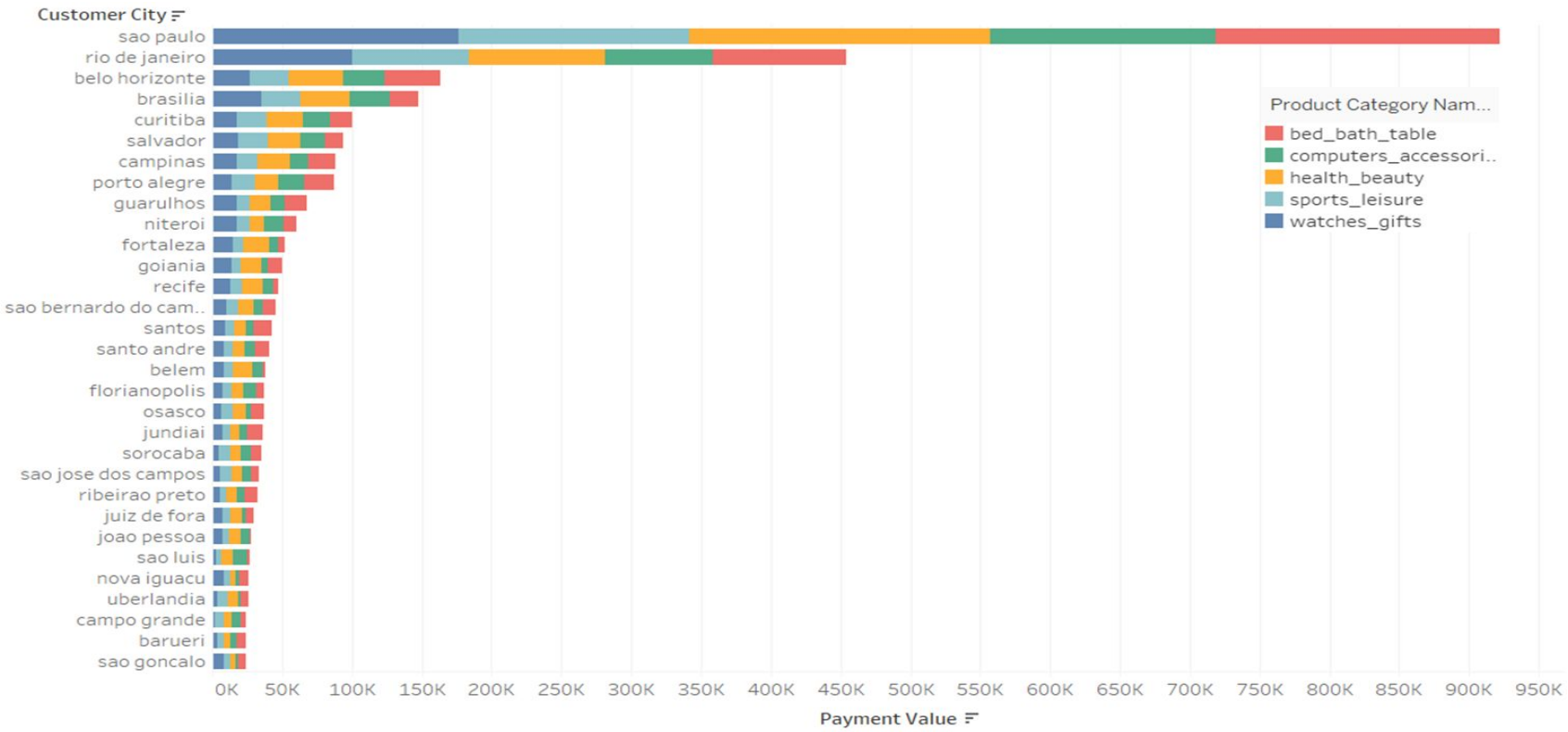
Customer distribution across the top 5 states in Brazil
(no. of customers in thousands)



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%**

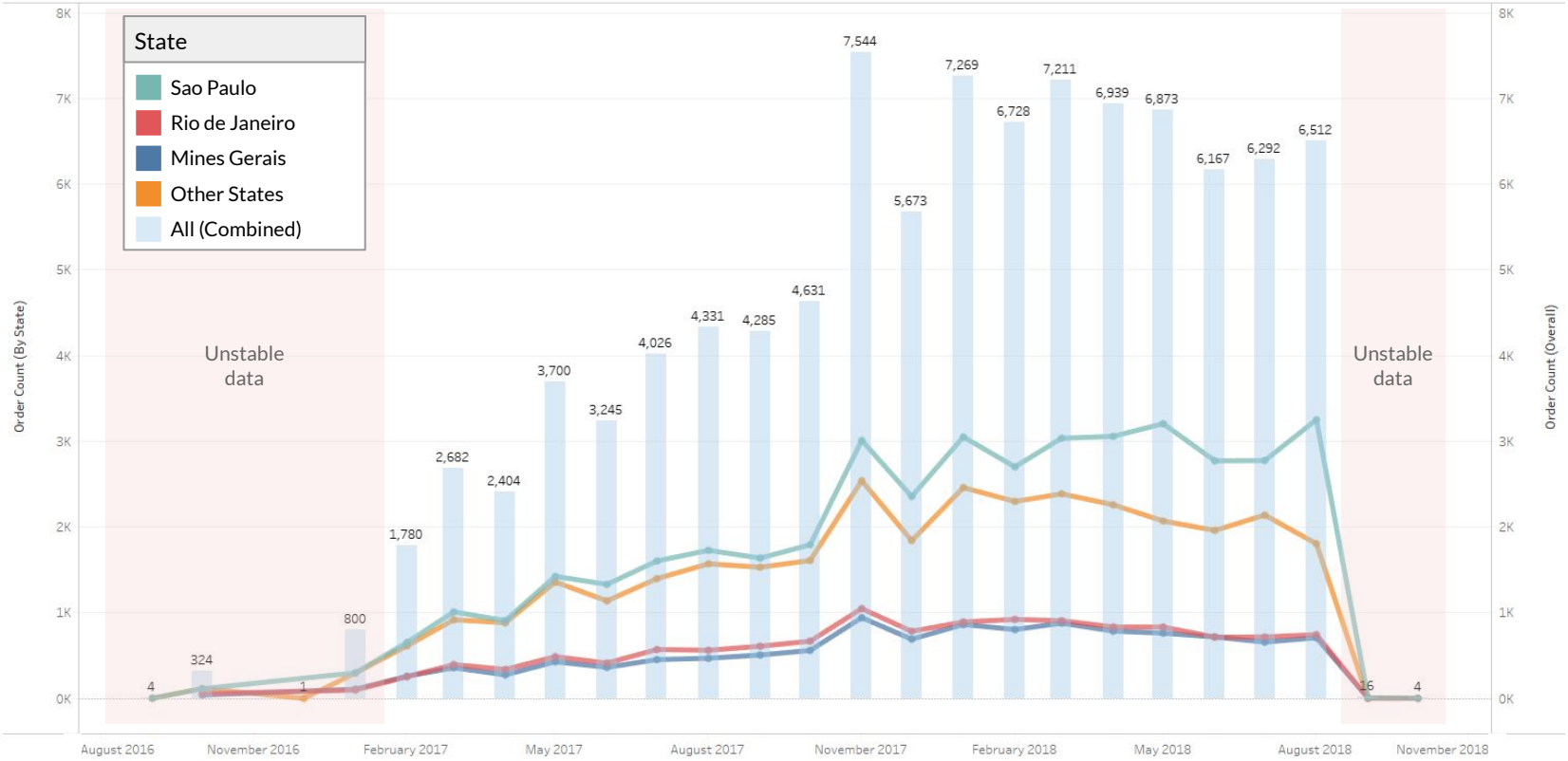
Customer: Brazilian consumers in different states have highly similar preferences for types of goods, but rank them differently

Customer Purchasing Preferences on Top 5 Categories (by States)



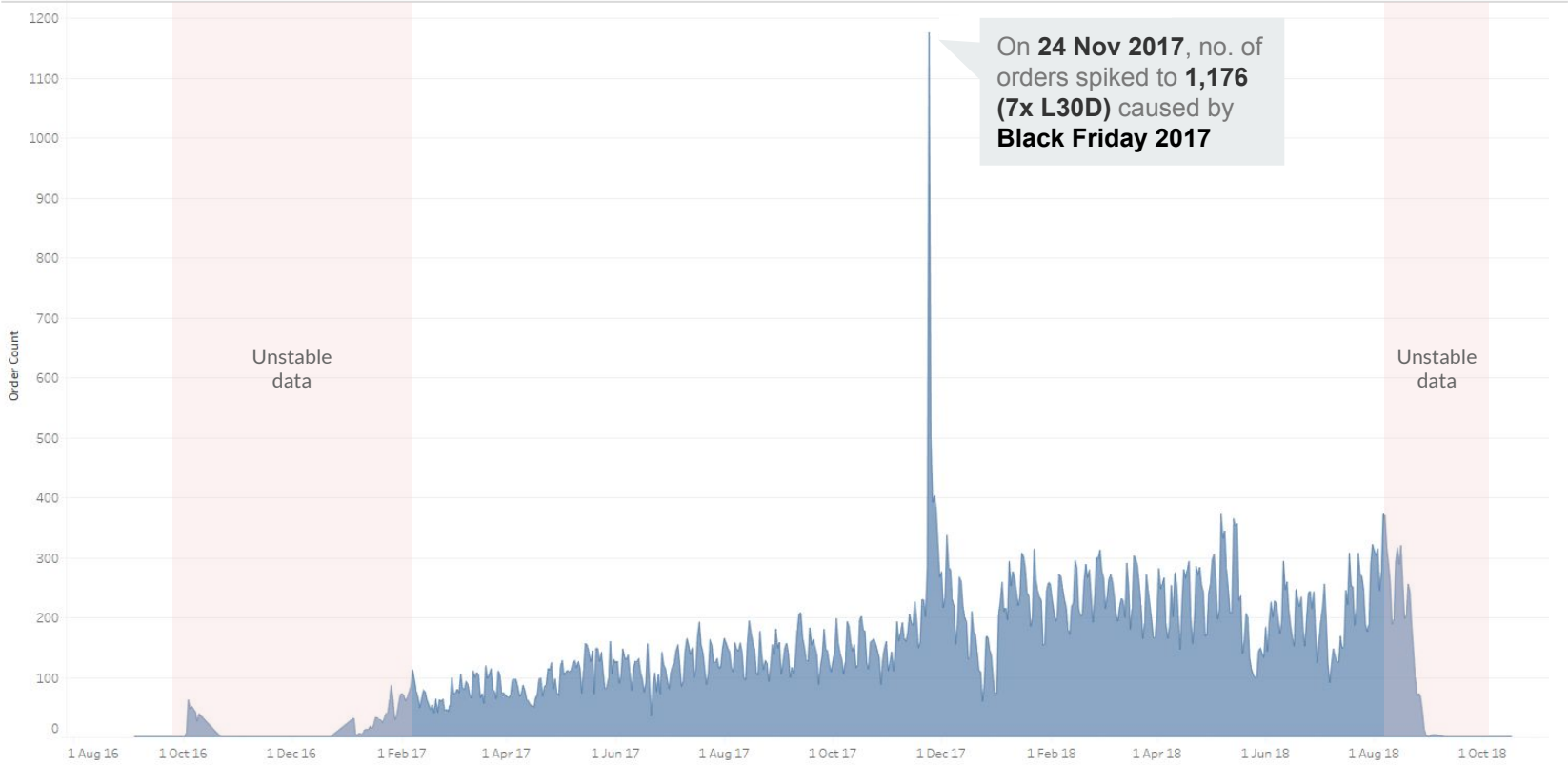
Orders: Upward trends observed on monthly orders; unreliable order counts before Feb 2017 and after Aug 2018 to be removed

Monthly Order Distribution (by states and combined)



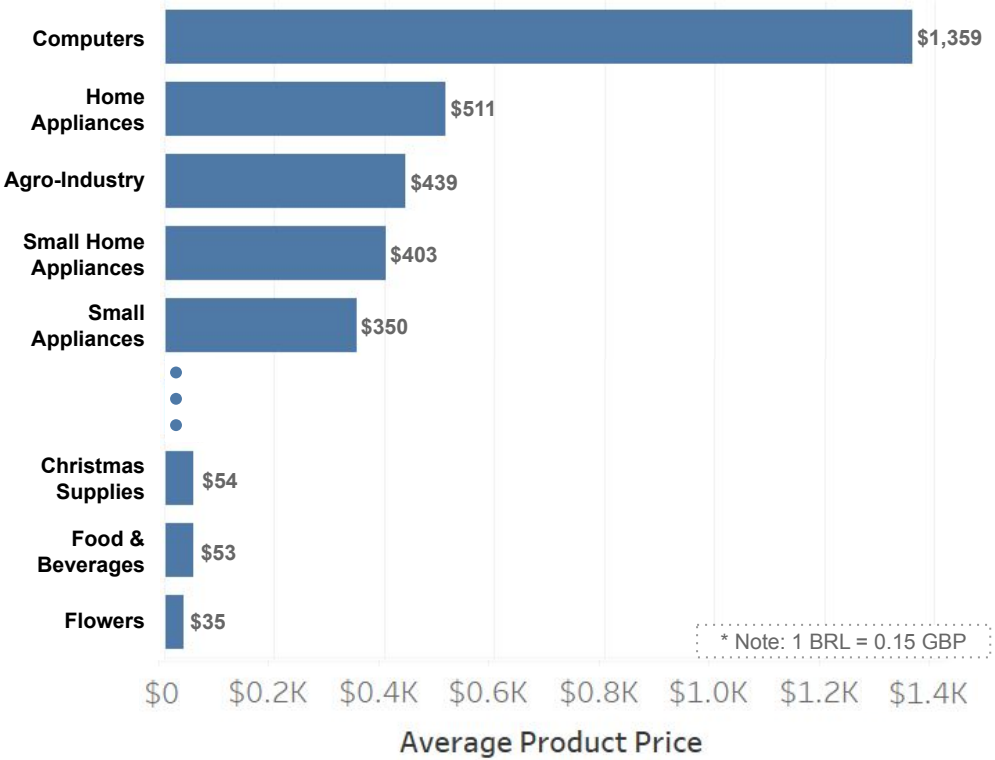
Orders: Upon further investigation, order spike uncovered on the 24 Nov 2017 due to 2017 Black Friday

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)

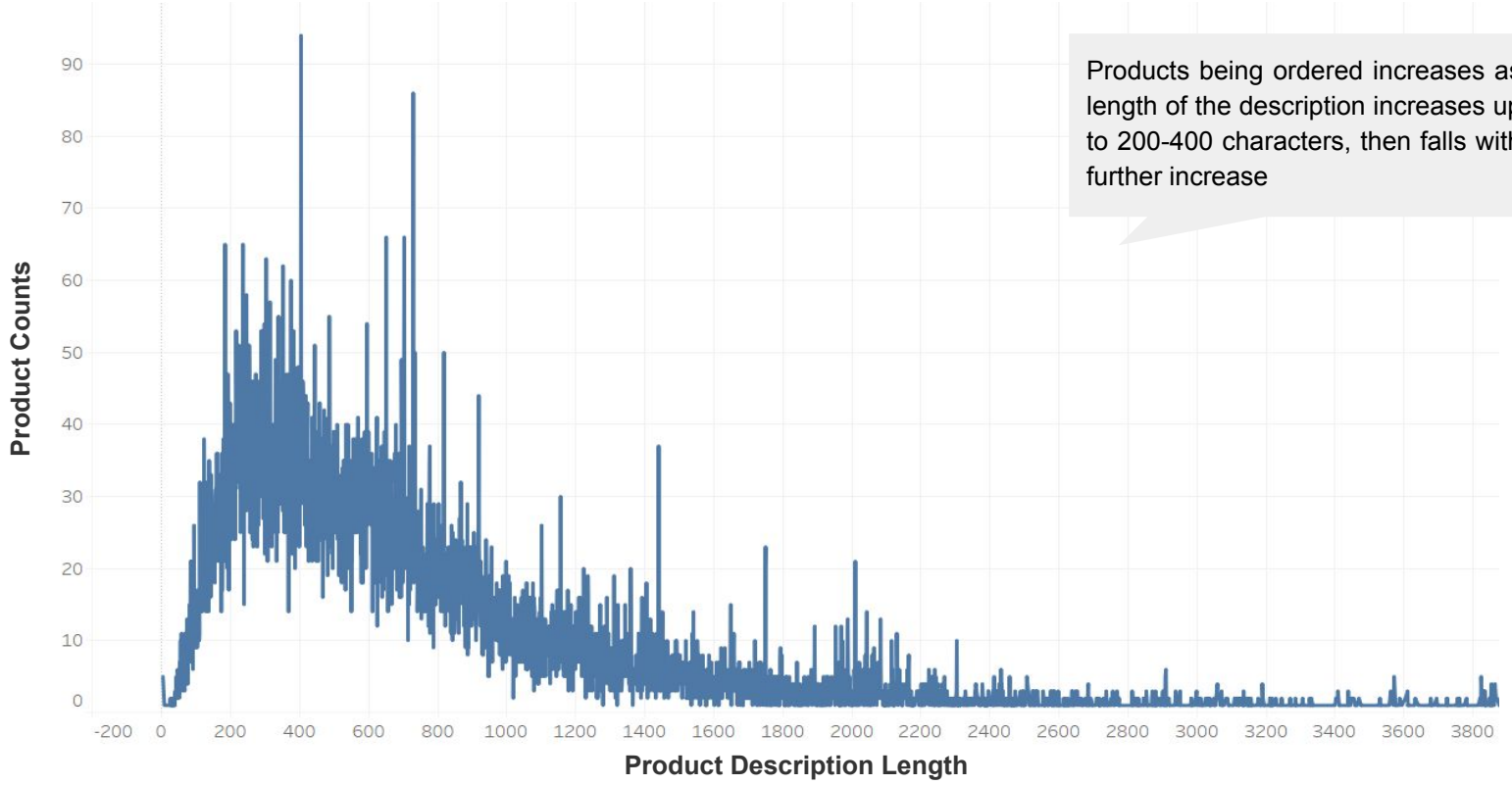


- Out of 77 available product categories, the highest priced are mostly **electronics**, with **computers** averaging at **~R$1,359**
- Duplicated or similar categories observed, **data cleaning and re-categorising** will be required

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

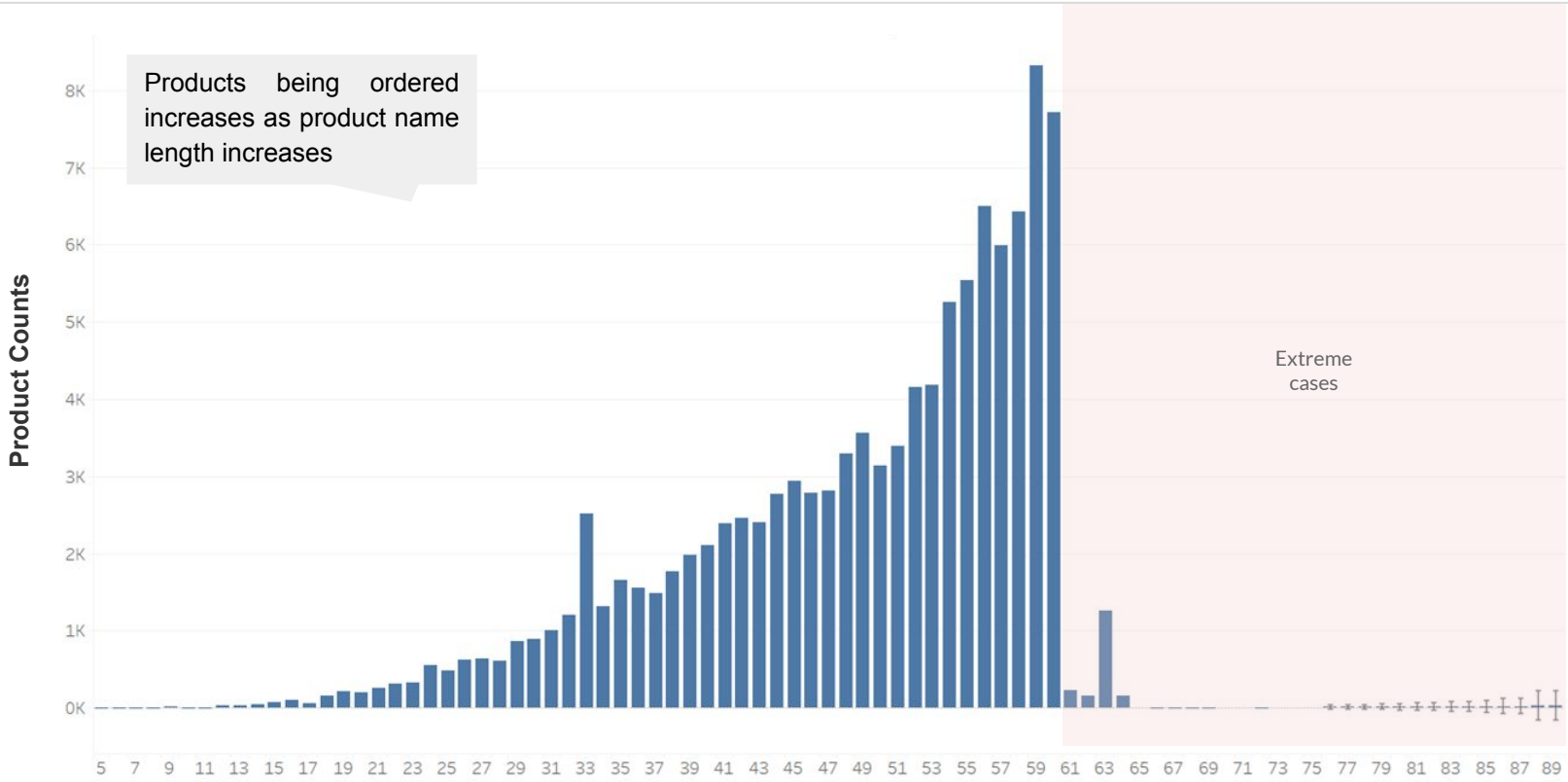
Distribution of Product Description Length



Products being ordered increases as length of the description increases up to 200-400 characters, then falls with further increase

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

Distribution of Product Name Length



<u>Derived features:</u> 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**