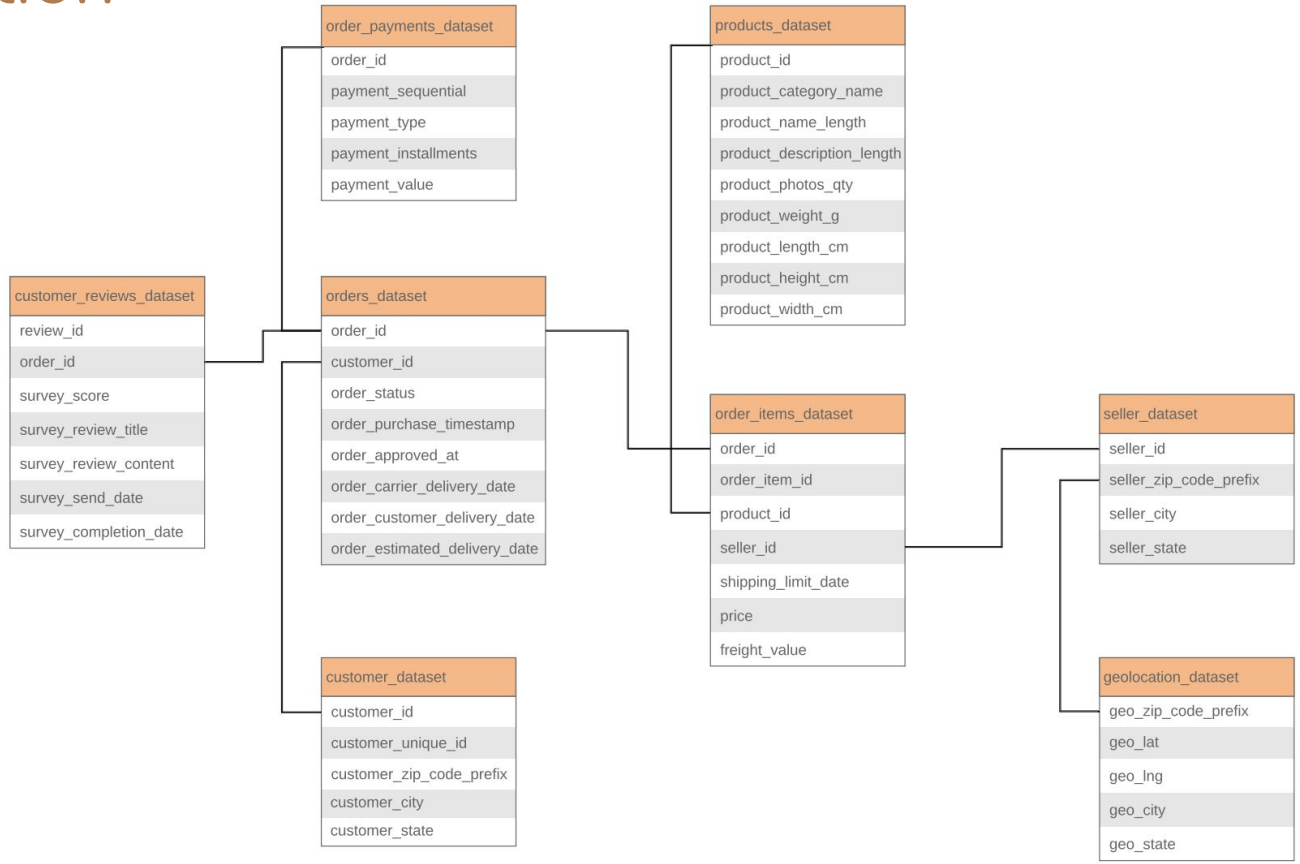
E-Commerce

ETL - Visualization - Analytics

Data Description

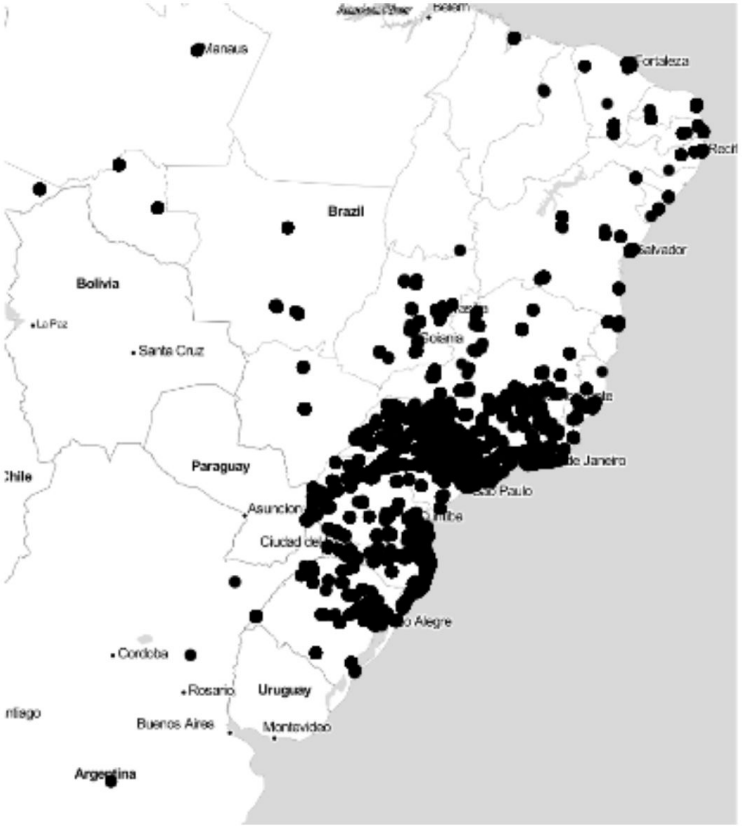


Basic Statistics

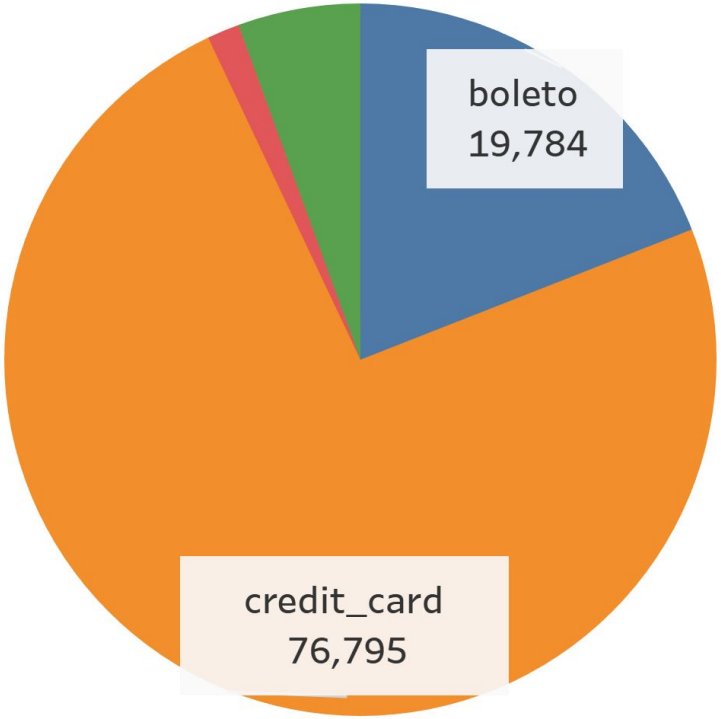
- Number of Customers 96,096
- Number of Orders 99,441
- Number of Categories 71
- Number of Products 32,951
- Percentage of Sequential Payments is 12.02%

Where do the buyers come from?

Where do the sellers come from?

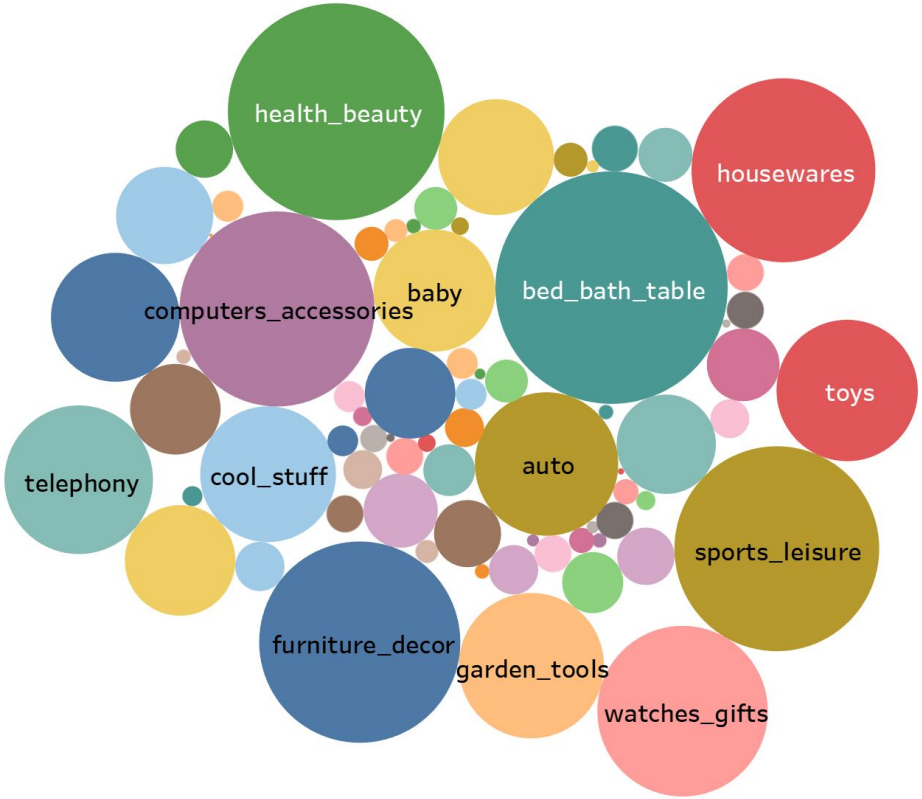


Payment Methods



Boleto Bancário, simply referred to as **Boleto** (English: Ticket) is a payment method in Brazil regulated by FEBRABAN, short for Brazilian Federation of Banks.

Top 10 popular category



How Many Different Categories Did Customers Order?

Number of Distinct Category per Customer	Number of Customer
1	98655
2	768
3	18

Top 10 Highest Rated category

cds_dvds_musicals	5.0000
fashion_childrens_clothes	5.0000
la_cuisine	5.0000
cine_photo	4.4242
books_imported	4.4000
costruction_tools_tools	4.3750
books_general_interest	4.3744
tablets_printing_image	4.3448
food_drink	4.3063
luggage_accessories	4.2897

Top 10 Categories where customer pay sequentially

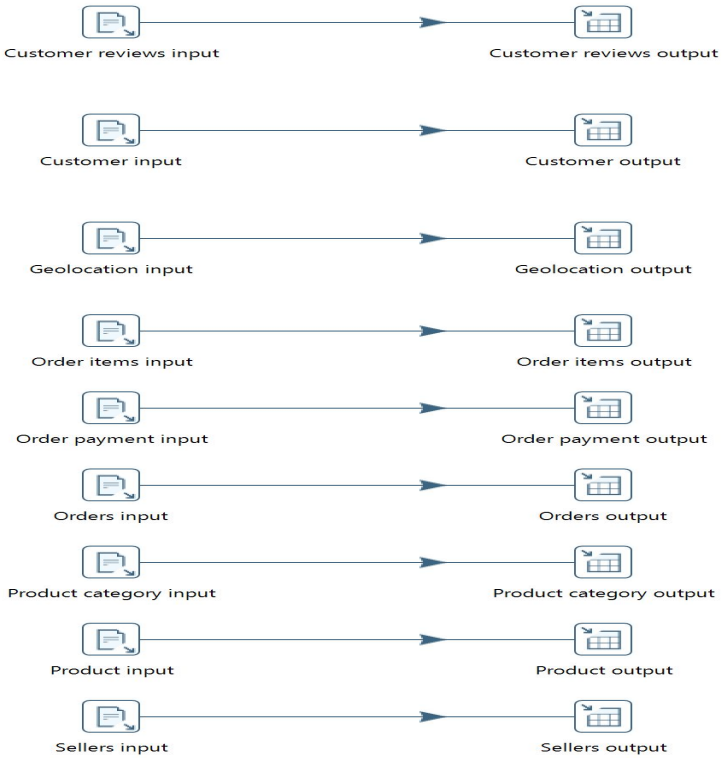
bed_bath_table	1,602
furniture_decor	1,397
computers_accessories	913
sports_leisure	903
housewares	868
health_beauty	797
garden_tools	644
watches_gifts	423
telephony	383
auto	341

Kettle for Data Preparation

Actions Taken:

(1) Read files as csv input and load them to the database

(2) Basic data transformation: change data types



DB Question 1

Categories that Received Most 5 Ratings

	product_category_name	product_category_name_english	num
1	beleza_saude	health_beauty	5870
2	cama_mesa_banho	bed_bath_table	5795
3	esporte_lazer	sports_leisure	5129
4	moveis_decoracao	furniture_decor	4462
5	informatica_acessorios	computers_accessories	4204
6	utilidades_domesticas	housewares	3988
7	relogios_presentes	watches_gifts	3335
8	brinquedos	toys	2518
9	ferramentas_jardim	garden_tools	2482
10	automotivo	auto	2383

DB Question 2

Top 5 Best Selling Products

	product_id	product_category_name	product_category_name_english	num_sold	avg_rate
1	aca2eb7d00ea1a7b8ebd4e68314663af	moveis_decoracao	furniture_decor	527	4.007590132827324 5
2	99a4788cb24856965c36a24e339b6058	cama_mesa_banho	bed_bath_table	491	3.8615071283095723
3	422879e10f46682990de24d770e7f83d	ferramentas_jardim	garden_tools	487	3.9425051334702259
4	389d119b48cf3043d311335e499d9c6b	ferramentas_jardim	garden_tools	392	4.1122448979591837
5	368c6c730842d78016ad823897a372db	ferramentas_jardim	garden_tools	391	3.9156010230179028

DB Question 3

Top 5 Products whose Consumers and Sellers are from the Same States

Definition of "Mostly": The total number of times when customers and sellers are from the same state for each product.

	product_id	num_same
1	aca2eb7d00ea1a7b8ebd4e68314663af	265
2	99a4788cb24856965c36a24e339b6058	231
3	422879e10f46682990de24d770e7f83d	181
4	368c6c730842d78016ad823897a372db	135
5	389d119b48cf3043d311335e499d9c6b	133

DB Question 4

Regional Break-up of Total Sales of Heavy Products

Total Number of Products Sold > 10000:

SP

Total Number of Products Sold > 1000:

MG, PR, SC, RJ

Total Number of Products Sold > 100:

RS, ES, DF, BA, GO

Total Number of Products Sold < 100:

MT, CE, MS, PE, RN, PB, RO, PI, MA, SE, PA, AC

	seller_state	total_sales
1	SP	31760
2	MG	4249
3	PR	3423
4	SC	1752
5	RJ	1347
6	RS	928
7	ES	296
8	DF	241
9	BA	200
10	GO	161
11	MT	86
12	CE	63
13	MS	49
14	PE	33
15	RN	21
16	PB	11
17	RO	10
18	PI	9
19	MA	6
20	SE	5
21	PA	1
22	AC	1

Analytics: Customer Segmentation

Objective: Segment customers based on their purchase preferences

Modeling: PySpark Clustering - Kmeans

Features: a) Purchase times of each category that made by each customer

b) Average rating(survey score) on each category that they purchased

customer_id	agro_industry_and_commerce_rate	air_conditioning_rate	art_rate	arts_and_craftmanship_rate	audio_rate	auto_rate	baby_rate
00012a2ce6f8dcda20d059ce98491703	0.0	0.0	0.0	0.0	0.0	0.0	0.0
000161a058600d5901f007fab4c27140	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0001fd6190edaaf884bcaf3d49edf079	0.0	0.0	0.0	0.0	0.0	0.0	5.0
0002414f95344307404f0ace7a26f1d5	0.0	0.0	0.0	0.0	0.0	0.0	0.0

...	signaling_and_security_count	small_appliances_count	small_appliances_home_oven_and_coffee_count	sports_leisure_count	stationery_count	tablets_printing_image_count	t
...	0	0	0	0	0	0	
...	0	0	0	0	0	0	
...	0	0	0	0	0	0	
...	0	0	0	0	0	0	

Analytics: Customer Segmentation

Load Data

- customer_reviews
- order_items
- orders_dataset
- products_dataset
- product_category_name_translation

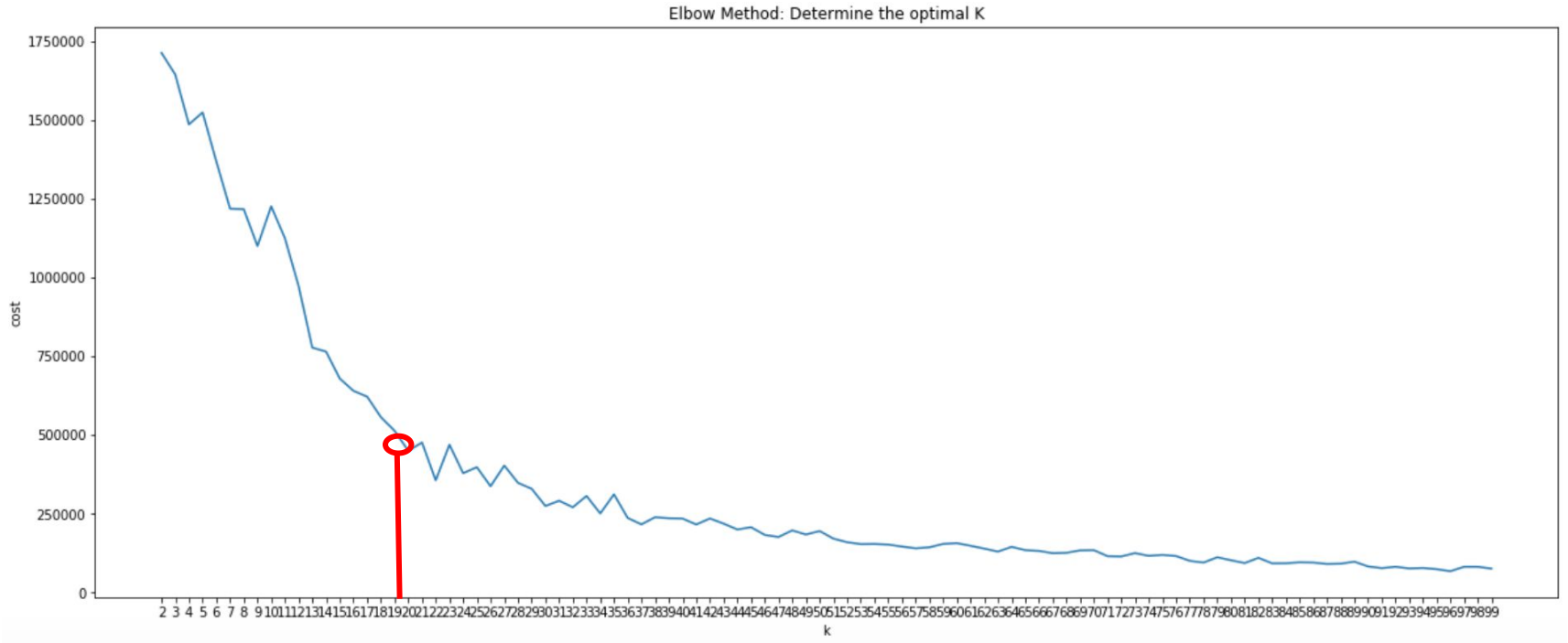
Prepare Data

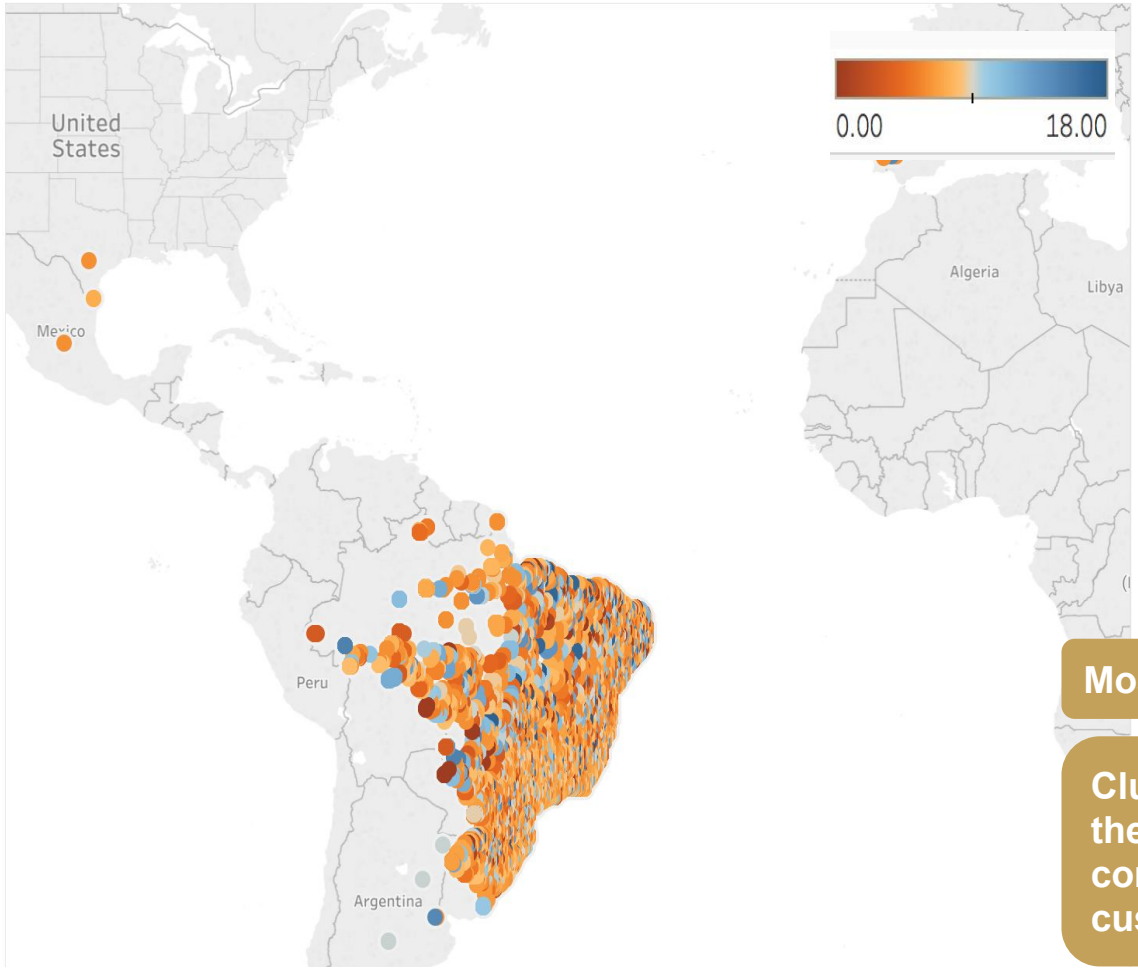
- Merge into one dataset
- Drop NAs (Total: 97256)
- Use VectorAssembler to prepare features into one vector

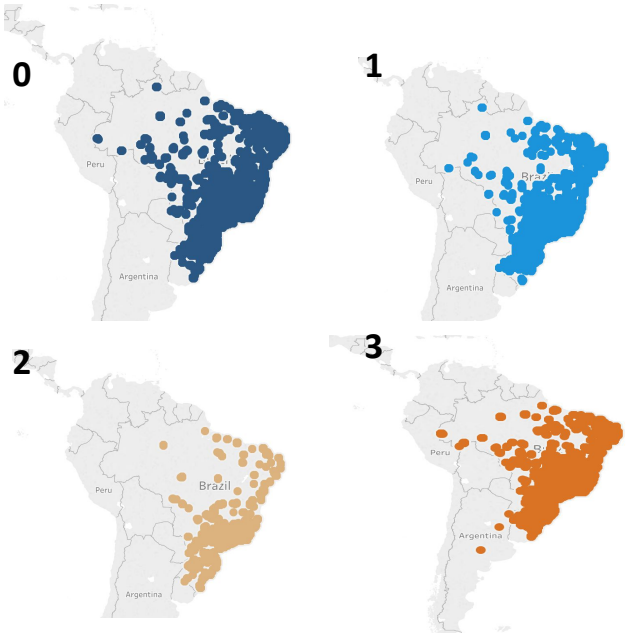
Data Modeling

- KMeans Clustering
- Use Elbow Plot to determine the optimal K

Determine the optimal K: 19



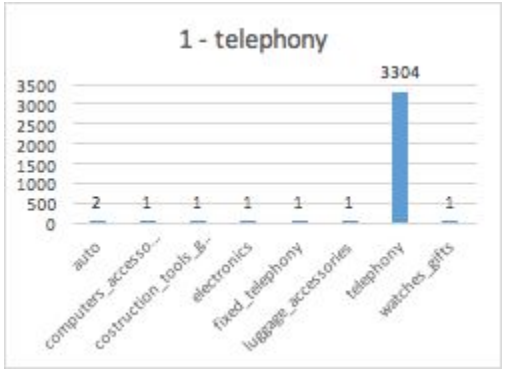


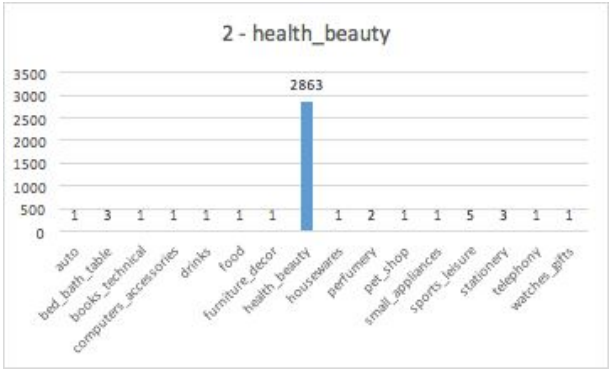


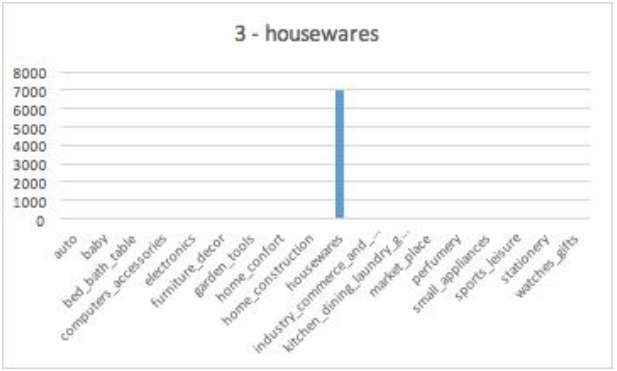
Most customers are in Brazil.

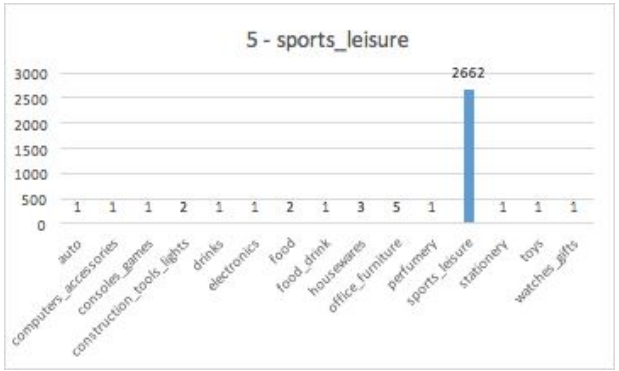
Clusters are distributed evenly on the map, which means there is no correlation between geography and customer segmentation.

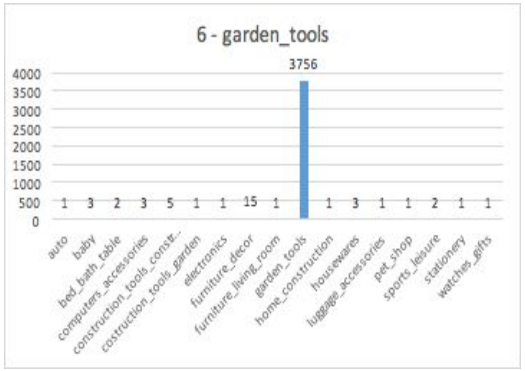
Category Distribution in clusters











Cluster	Top Category	Cluster	Top Category
0	telephony	10	bed-bath-table
1	health-beauty	11	health-beauty
2	housewares	12	electronic
3	home-confort	13	stationery
4	bed-bath-table	14	cool_stuff
5	sports-leisure	15	sports-leisure
6	furniture/home_appliance/construction_tools	16	cool_stuff
7	garden_tools	17	cool_stuff
8	computor_accessories	18	watches_gift
9	fashion_male_clothing		

Application

- Able to divide customers into different segmentation
- Better to target customers in the marketing

Things we can do further

- More demographic data about customers
- More data points on customer purchases
- Sample to have ground truth

THE END