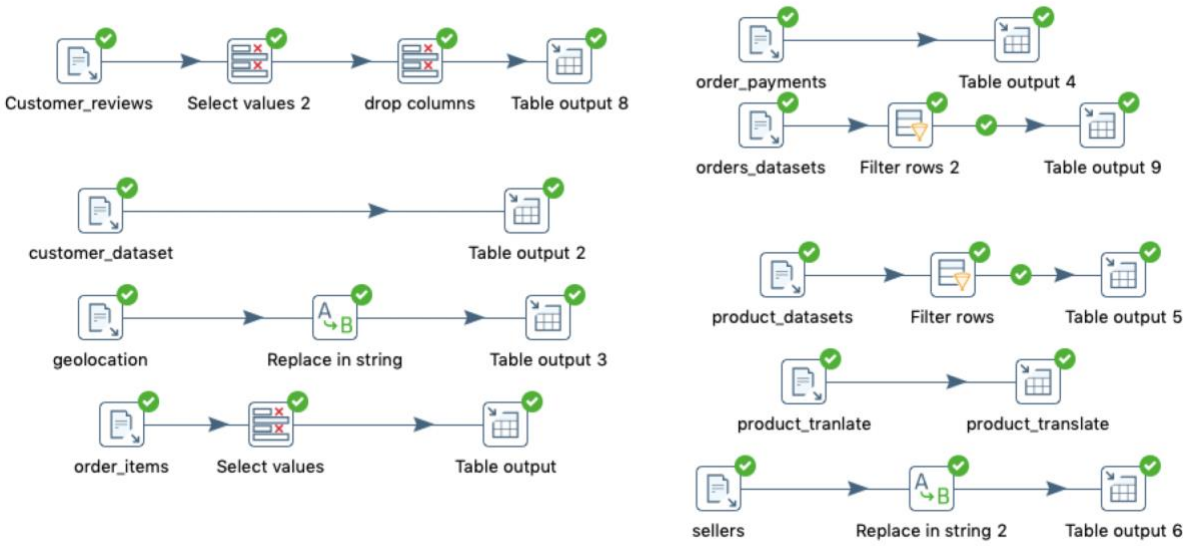
Ecommerce Final Project Report

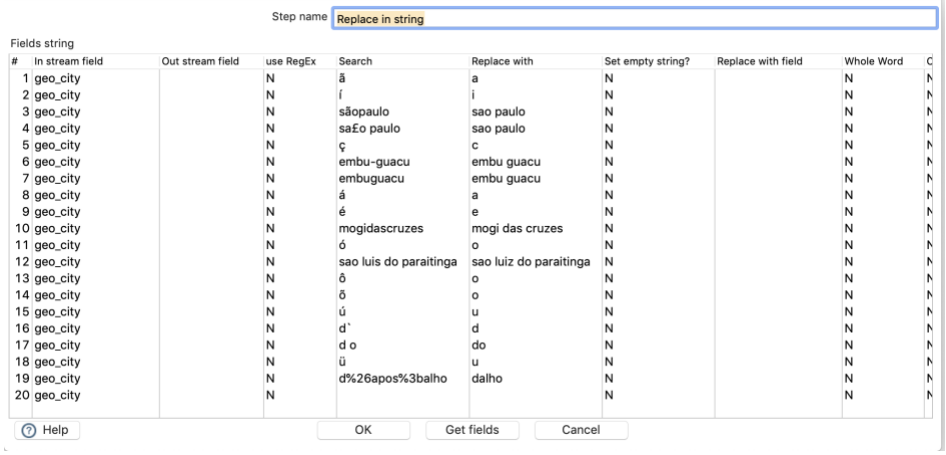
Team 13: Yizhen Lou, Yu-Wei Tang, Yuyan Wang

1. ETL

First of all, we used Data Integration to load data into local postgres and do some basic data pre-process to make later analysis easier. This is our kettle script.



Basically, we did three alterations. First, we changed some data type. For example, in customer review dataset, we changed the data type of survey score from factor to integer to convenient for customer ratings analysis. Then we filtered and eliminated some nulls, such as delivery date in order dataset and product category name from product dataset. Since this is a dataset from Brazil, we also need to translate some city names from Portuguese to English, such as geo_city from geolocation dataset and seller city from sellers dataset through Replace in String process as following.



Step name: Replace in string

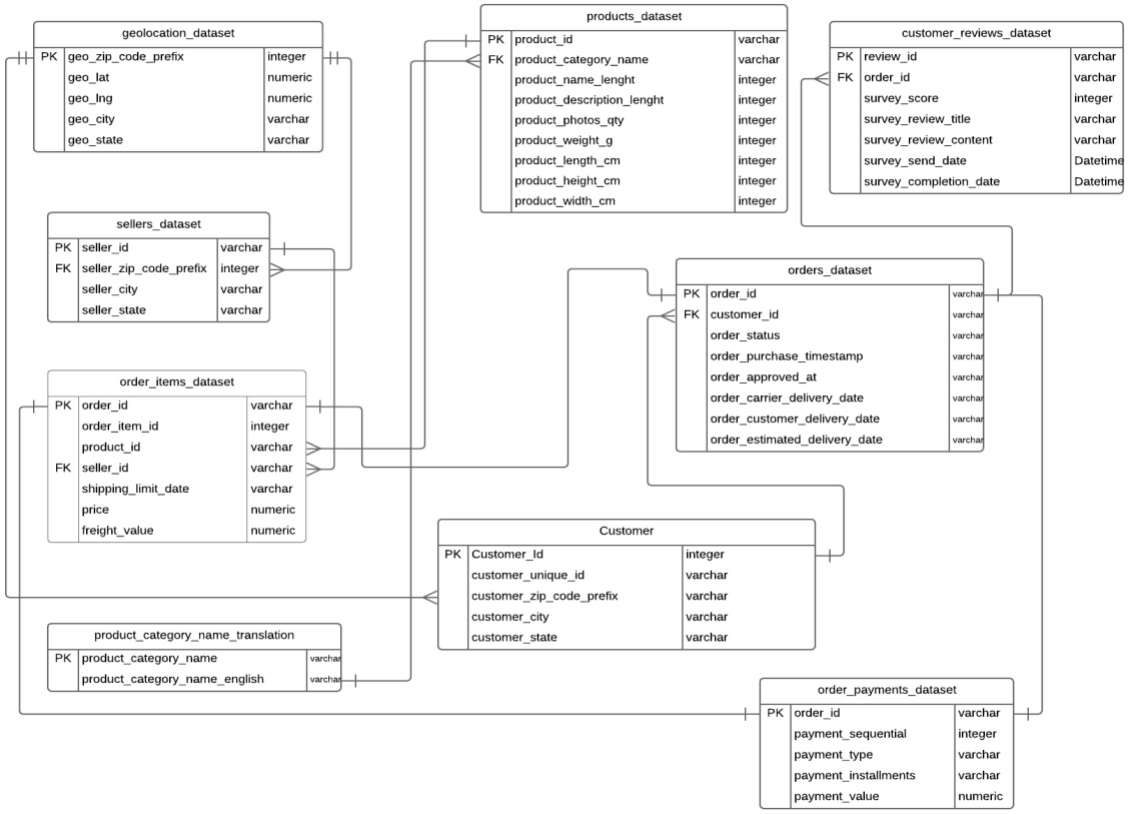
Fields string

#	In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	C
1	geo_city		N	ã	a	N		N	N
2	geo_city		N	í	i	N		N	N
3	geo_city		N	sãopaulo	sao paulo	N		N	N
4	geo_city		N	sa£o paulo	sao paulo	N		N	N
5	geo_city		N	ç	c	N		N	N
6	geo_city		N	embu-guacu	embu guacu	N		N	N
7	geo_city		N	embuguacu	embu guacu	N		N	N
8	geo_city		N	á	a	N		N	N
9	geo_city		N	é	e	N		N	N
10	geo_city		N	mogidascruzes	mogi das cruzes	N		N	N
11	geo_city		N	ó	o	N		N	N
12	geo_city		N	sao luis do paraitinga	sao luiz do paraitinga	N		N	N
13	geo_city		N	õ	o	N		N	N
14	geo_city		N	õ	o	N		N	N
15	geo_city		N	ú	u	N		N	N
16	geo_city		N	d`	d	N		N	N
17	geo_city		N	d o	do	N		N	N
18	geo_city		N	ü	u	N		N	N
19	geo_city		N	d%26apos%3balho	dalho	N		N	N
20	geo_city		N			N		N	N

Help | OK | Get fields | Cancel

2. SQL

For the remaining part of our project, we use the entity relational diagram we created as below to identify relationship between dataset. We join and aggregate different datasets to get the data we need.



2.1 Which category of products receive most 5 ratings (survey_score)?

I selected ratings, product English category names using customer reviews dataset, order items, product dataset and product translate dataset. Then we counted the number of five score and get the following table that shows top products category that received most 5 score ratings, and health beauty products is the one that receives most five ratings.

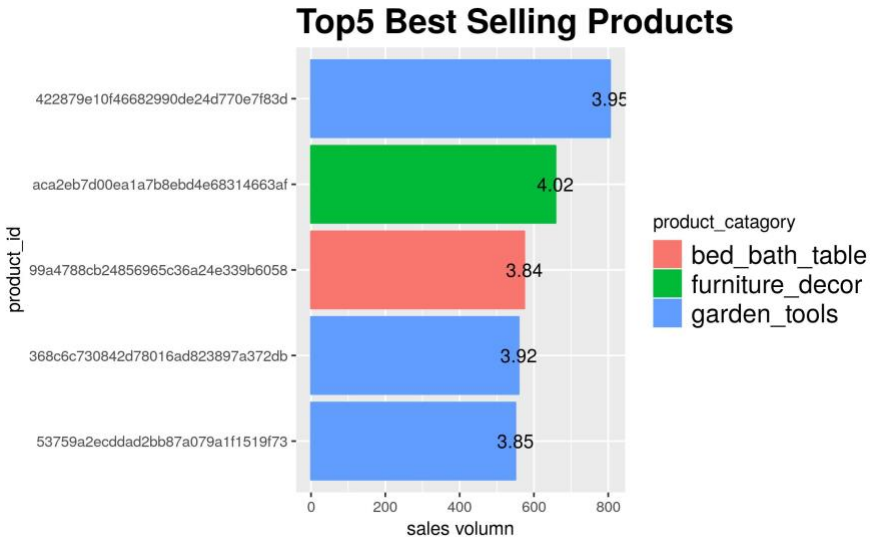
	product_e	count
1	health_beauty	667032
2	bed_bath_table	655056
3	sports_leisure	582180
4	furniture_decor	505044
5	computers_accessories	473076

2.2 What are the top 5 best selling products and what are their categories? What's the average rating of each of these products?

To answer this question we chose product id, product English category names, rating and order quantity from 4 different tables. Then we summed up order quantities and calculated the average ratings using subqueries and get the following results table.

	product_id	product_category_name_english	sales	ave_rating
1	422879e10f46682990de24d770e7f83d	garden_tools	87984	3.9482874412357287
2	aca2eb7d00ea1a7b8ebd4e68314663af	furniture_decor	70500	4.021887259962833
3	99a4788cb24856965c36a24e339b6058	bed_bath_table	61092	3.8529218647406435
4	368c6c730842d78016ad823897a372db	garden_tools	60984	3.9213389121338912
5	53759a2ecddad2bb87a079a1f1519f73	garden_tools	59652	3.8578947368421053

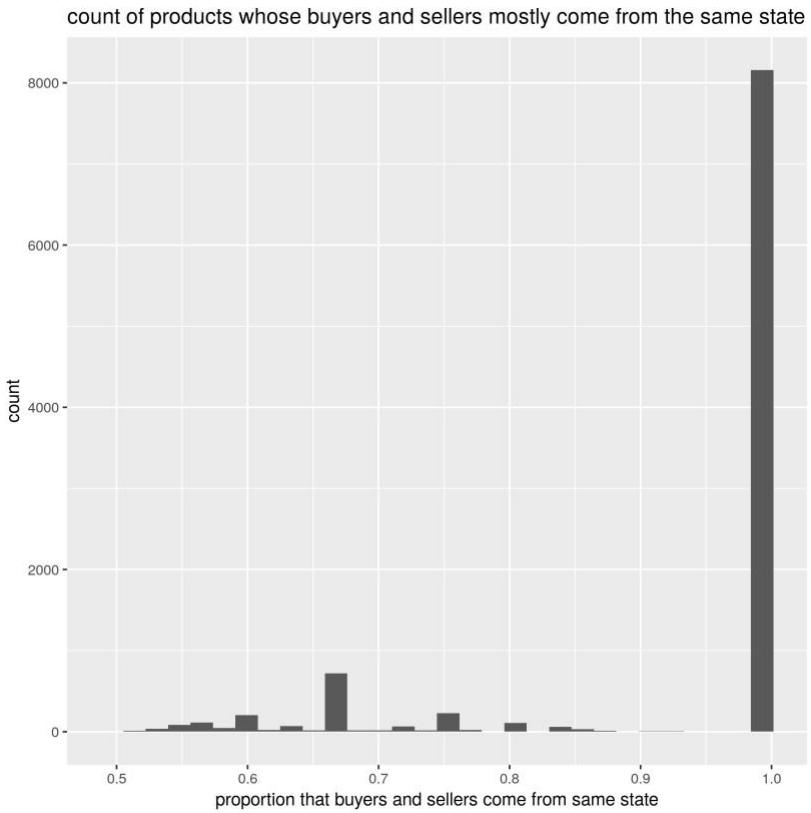
For better understanding, we visualized it in R got the following graph. The top 5 products came from three categories, Bed Beth table, furniture decoration and garden tools. The one that earns the highest average rating is second best-selling product.



2.3 For which products are the buyers and sellers mostly from the same sate?

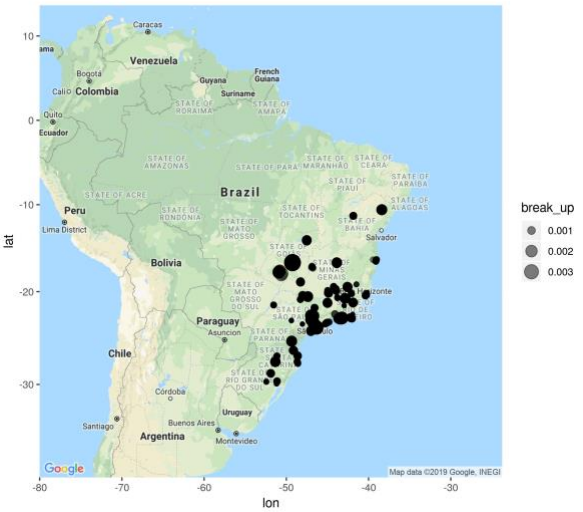
Before we answer this question, we first need to clarify that ‘mostly’ means that the proportion of buyers and sellers come from same state is higher than 0.5. To answer this question, apart from dataset mentioned before, we also need to get access to sellers dataset for seller location information. Eventually about 10,000 products satisfied our standards and As we can see from the following graph, more than 8000 products’s customers and buyers are from the same state.

	product_id	rate
1	00088930e925c41fd95ebfe695fd2655	1
2	0009406fd7479715e4bef61dd91f2462	1
3	001795ec6f1b187d37335e1c4704762e	0.66666666666666666667
4	001b237c0e9bb435f2e54071129237e9	1
5	0021a87d4997a48b6cef1665602be0f5	1
6	00250175f79f584c14ab5cecd80553cd	0.88235294117647058824
7	002c6dab60557c48cfd6c2222ef7fd76	1
8	002ec297b1b00fb9dde7ee6ac24b6771	1



2.4 Give a regional break-up of total sales of heavier (weighing at least 1 kg) products

In this question we calculated break-up of total sales of heavier products in terms of zipcode. We firstly joined product dataset, customer dataset, order items and order dataset together and created a view and then simply selected zipcode and calculated breakup from it. To draw a clearer map, we selected the top 50 zipcodes with the highest break-up and mapped them. We can conclude that regions that consumes more heavier products concentrate in southeast area of the country.



3. Modeling

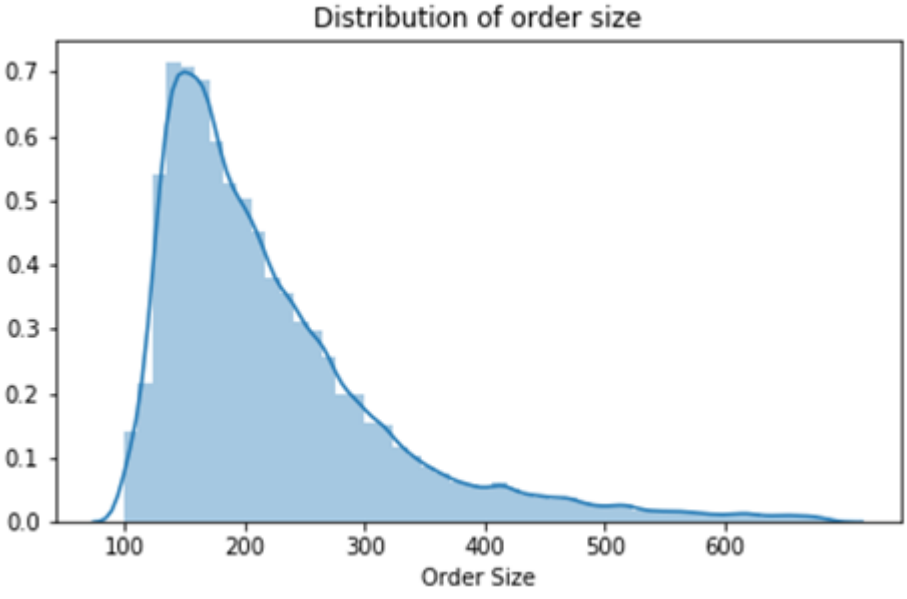
3.1 EDA

Before start modeling, we want to explore the data so as to find the characteristics of data and thus to decide the objective of our model. As data are ecommerce transaction, we could definitely gain knowledge about the overall performance of this ecommerce business.

3.1.1 Number of orders

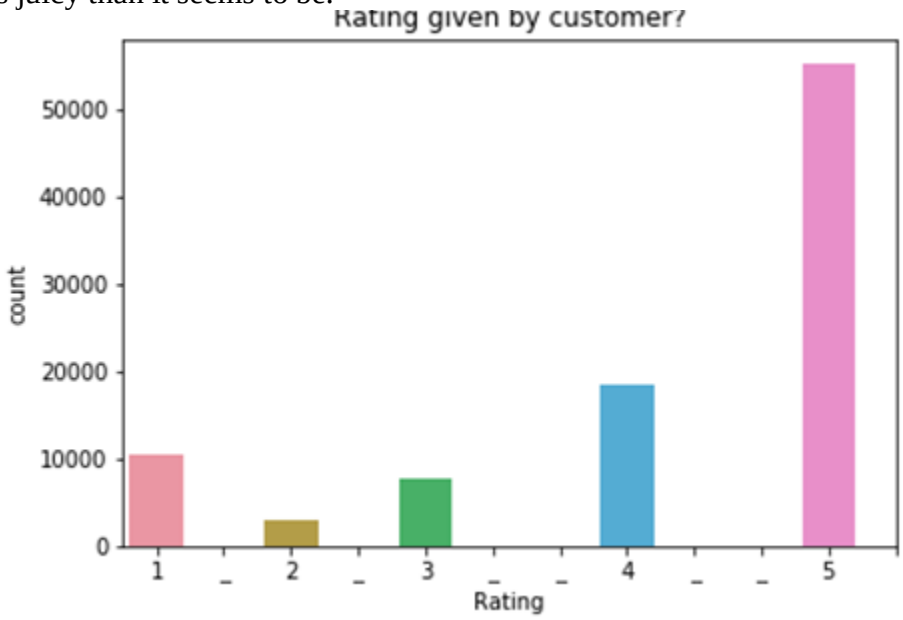
Using order data, we want to see how many orders a unique customer has made. In fact, we find that most customers only made 1 orders and only 3% of customers order more than twice.

In 3.1.2 Order size order payment data, we select order size and group by customer, finding that most people spend between 100BRL - 300BRL on their order but there are much variance on order size.



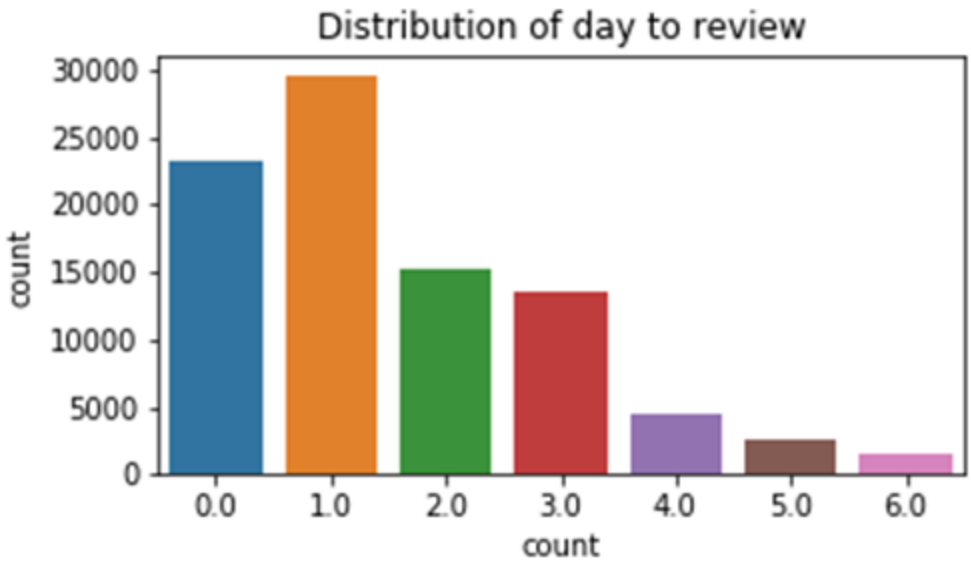
3.1.3Rating

In customer review data, we group by customer and select count of each average score, finding that 55125 out of 105188 observations have rating score of 5 and 18510 rate their product 4. Customers tend to rate high score on product. Having a score of 4 or even 5 becomes less juicy than it seems to be.



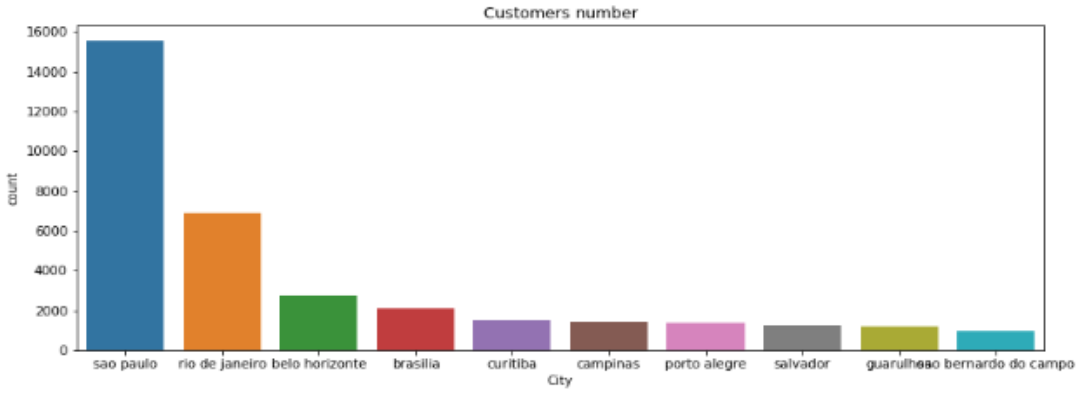
3.1.4 Review day used

Review day is extracted from survey completion data subtracted by survey sent day. It could be used to evaluate the efficiency of getting feedback from specific customer. We join review data with order data and group by customer to find average time used to complete survey. The average review time used by customer is 2.58 days with high variation. Maximum time used to complete a survey is 518 days.



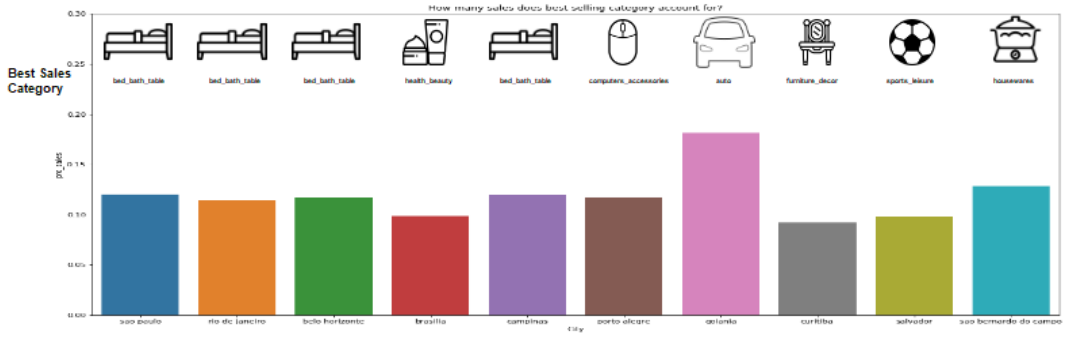
3.1.5 Customer base from each city

Customer base of this ecommerce company covers 4119 cities. However, the majority of customers are from top 10 cities. In customers data, by grouping city and selecting count of customers from each city, we find that 15% of customers are from Sao Paulo and 7% from Rio de Janeiro. Customers are not overly concentrated in one of the cities, which is normally happened in a healthy ecommerce company since ecommerce business has less geographical restrictions compare to physical store.



3.1.6 Famous genre in top city

Later, we want to find the most popular category in each city by percentage of sales of all items under that category account for total sales. We join customers data with order data and product data to find sales of items under each category and find that different city has really different famous category. However, 4 of the top 10 cities have the highest number of sales on bed/bath/table category and all of them account for 11.5% - 12%. Also, in goiania percentage of sales on auto category is 18%, which is way higher than percentage of sales on same category in any other top 10 cities.



3.2 K means modeling

In the modeling part, the objective is to break all our customers into different clusters that customers in the same cluster have similar features. In this case, the e-commerce platform can apply customized service and recommendation to difference customer clusters.

3.2.1 Data preparation process

Since we are doing clustering on customers, so the unit of observation is customer. From datasets we have, we choose number of orders customer placed, average rating they gave, average number of days they took to complete survey and their longitude and latitude of geolocation as features to train model. We combine and aggregate data to get the features we need.

We first join customers_dataset.csv and geolocation_dataset.csv by geo_zip_code_prefix, so that we have geolocation information for each customer. Then we join customers_dataset.csv and orders_dataset.csv by customer_id and "group by" customer to get total number of orders each customer made in the past. In order to get average dollars people paid: we first join orders_dataset.csv and order_payments_dataset.csv by order_id to get price of each order and customer who bought the product. Then we join the three result tables above by customer_id, so that we can get choose number of orders customer placed, average rating they gave, the average number of days they took to complete survey and their longitude and latitude of geolocation.

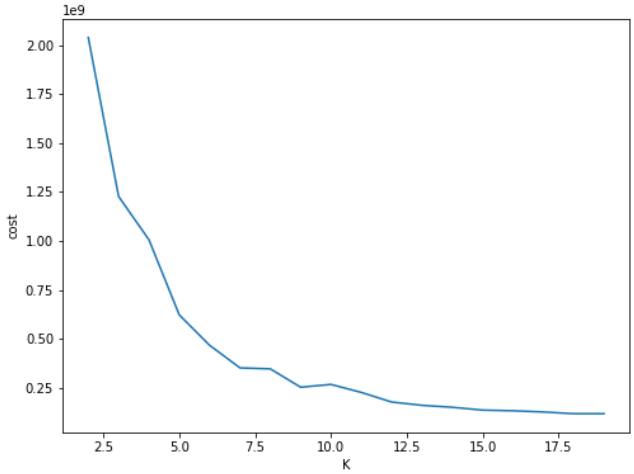
So the data after preparation will looks like the table below with customer_id and features:

customer_id	lat	lng	order_number	avg_price	avg_rating	avg_review_days
01d190d14b00073f76e0a5ec46166352	-23.50285023	-46.70118551	1	258.74	5	1
03a7750fc7a7bfbd7a84b2f4f26b92f1	-25.36823007	-49.1235043	1	198.1	3	3
04495037fc6899faffa41ba3bc4272b4	-23.64671331	-48.7999225	1	265.04	4	3
04b7d26bde4f2d2fee0043ef81e664b1	-19.73697289	-47.94263387	1	372.56	5	1
04cef6b920c0d8f16702cab269b59044	-23.56243003	-46.59556919	1	340.72	5	1
06199a7981ec145069afeb0baf66a49b	-23.5596494	-46.93765333	1	55.38	5	1
0632cb63610a5d7d5da9a4fb595dd101	-23.35942681	-51.18193361	1	113.42	5	8
068d6956b4f2f9a39cfb807516f55ecb	-22.95383244	-43.69196201	1	117.77	5	0

3.2.2 K-means

The model we choose is K-means unsupervised learning that there is no response variable, but explanatory variable we get from data preparation step: latitude, longitude, number of orders, average dollar, average rating and average review days.

In order to apply the model, we first use Vector Assembler to transfer features so that they're compatible for K-means model. Then we determine K(number of clusters) by comparing cost(sum of squared distance from each customer to cluster center) for different K. The plot below shows how cost changes when K increases.



We want to compute cost to be as small as possible which means that our model predict accurately. We first try K = 18, but 18 clusters is a complex model that takes a long time to train. 18 clusters also too much that does not make sense to our data. So by balancing compute cost and model complexity, we choose K = 9.

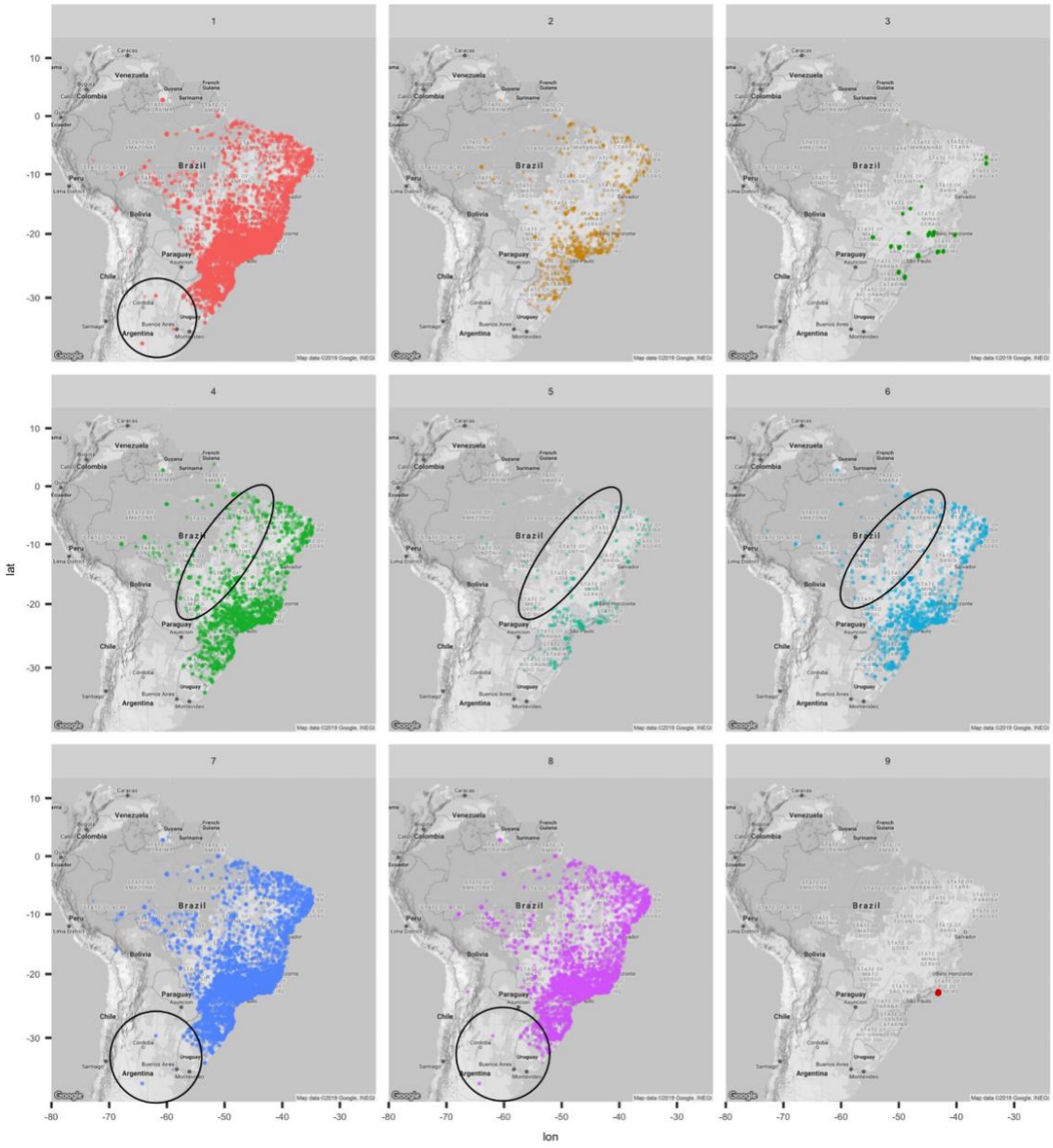
The model returns 9 clusters. The table below shows summarize information of 9 clusters and in each cluster the product category with the largest number of sales.

Cluster	# of customers	(average) Order number	(average) Dollars	(average) rating	Top Categories
1	43353	1.08	56.70	4.16	Bed bath table
2	765	1.00	1253.81	3.91	Furniture decor
3	25	1.00	4450.06	4.00	Agro industry and commerce
4	5558	1.01	400.44	3.98	Furniture decor

5	264	1.01	2176.36	3.73	Computer accessories
6	2241	1.01	723.49	3.92	Computer accessories
7	29231	1.01	131.32	4.10	Bed bath table
8	13711	1.01	229.44	4.00	Bed bath table
9	1	1.00	13664.08	1	Fixed telephony

From order number column, we can see that no matter which cluster, order numbers are always around 1. This means that most customers purchased only once on this website and never come back again. Getting more return customers is a big problem for the website now. Meanwhile, cluster 1, 7 and 8 have large population so that they are the website's main market. And the top category for these three clusters are all bed bath table, which indicates that website can focus more on this category to get more and better quality this category product. And for cluster 9, there is only one customer. This person is an outlier that with 13664 dollars purchase for once. But he or she only gave 1-star rating. Regarding the future demand of this customer, the website may have good after sales service to keep the customer.

Finally, we visualize geolocation of 9 clusters. From plot below, we can see that firstly for all clusters most customer population located in southwestern region of Brazil. This is because this region has a large population and it contributes 60% of Brazilian GDP. For cluster 4, 5, 6 with high average dollar, there are relatively many customers locate in inner region of Brazil. This indicates that there is a potential high-end market in this region. From cluster 1, 7, 8, there are some customers outside Brazil, from Argentina for example. So there are foreign demand, foreign markets we can develop.


1
2
3
4
5
6
7
8
9
lat
lon
10
0
-10
-20
-30
-80
-70
-60
-50
-40
-30
Brazil
Venezuela
Colombia
Peru
Bolivia
Paraguay
Chile
Argentina
Uruguay