Visualization Assignment 1

Student Name: Ngoc Chien Le

Student ID:1057045

Tutor Name: Bruno Luis Mendivez Vasquez (Lab Monday 12pm)

Activity Number: 26

Contents

1. Introduction

a. *Problem description:*

Taking the view as a manager of an e-commerce platform, this analysis will focus on the questions aiming to the grasp the overview of the business as well as understand more about the sellers who are joining the platform, thanks to that we could find the way to improve seller performance, boost the their cohesion to the e-commerce platform.

b. *Question*

Who is making money on our e-commerce platform?

- Who are those sellers? What do they sell? Where are they?
- Where are their customers?
- What are the marketing channels the company used to acquire them?

c. *Motivation*

From those the Info and analysis, we could:

- See the bigger picture from the process of seller recruitment into tracking sellers' performance.
- Up sales those sellers by providing them more services (sell strategy consultancy, advertisements.).
- Find more similar sellers (based on location, industry - could be similar or relatively alike, or marketing channels).

2. Data Wrangling

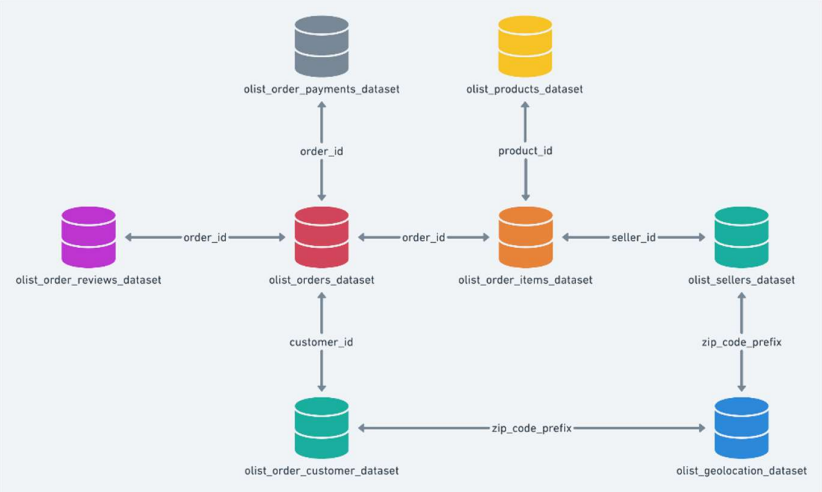
a. *Data Overview*

The data is from Olist, a Brazilian department store that operates in e-commerce segment, separated into multiple files therefore we need to merge them together based on database structure provided.

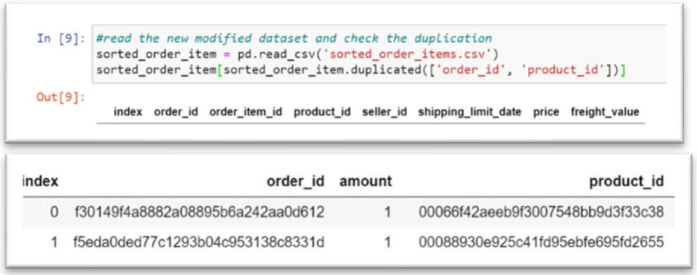
Data of 100k orders from 2016 to 2018

Link to the data source:

- Olist - e-commerce
- Olist - Marketing funel



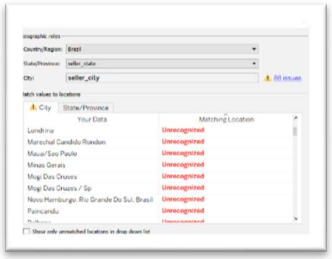
b. Data wrangling in detail

Problems	Actions	Methods	Software	Outcome
The data is separated as can be seen in data structure image above, included many columns that are redundant to answer our question	Merge datafiles and drop the unnecessary columns	Based on analysing the question, we could filter the features (necessary vs unnecessary) Merge the files by look at the database structure and the key that used to connect files together	**Tool:** Python **Package:** Pandas	Merged_ecommerce_data file
Order Item Dataset Instead of using one record as total number of a product, it has been separated into multiple records as an example below which led to many duplications with the data  **order_item_id = sequential number - identifying number of items included in the same order (**link to the explanation**)**	Sort the data and remove the unnecessary rows, only keep the last row represents the total amount	Run the loop to check if product ID repeat in the 2 consecutive rows and then remove the previous one since the last one is last number (total amount) **Note:** after sorting and removing the duplication, saving to a new file to avoid rerunning In order to reflect the true meaning of the **order_item_id** in the modified data, it is going to be renamed into **amount**	**Tool:** Python Run a for loop	A new "order item Dataset" with no duplication 

Order Item Dataset
 	
Missing total value of invoice (or seller_revenue from the seller perspective) we need to do a calculation
	Calculation	Calculate the invoice value by using the formula:
		
		Seller_revenue = amount * price
			Tool: Python	Seller Revenue Column
				
				`#calculate revenue for each seller`
				`sorted_order_item["seller_revenue"] = sorted_order_item["amount"] * sorted_order_item["price"]`
				`sorted_order_item.head()`
				
				product_id | seller_id | shipping_limit_date | price | freight_value | seller_revenue
				20066f42aeeb9f3007548bb9d3f33c38 | 5670f4db5b62c43d542e1b2d56b0cf7c | 2018-05-24 18:58:59 | 101.65 | 18.59 | 101.65
				20088930e925c41fd95ebfe695fd2655 | 7142540dd4c91e2237acb7e911c4eba2 | 2017-12-18 19:32:19 | 129.90 | 13.93 | 129.90
				2009406fd7479715e4bef61dd91f2462 | 4a3ca9315b744ce9f8e9374361493884 | 2017-12-29 16:12:36 | 229.00 | 13.10 | 229.00
				200b8f95fcb9e0096488278317764d19 | 40ec8ab6cdafbcc4f544da38c67da39a | 2018-08-16 13:35:21 | 58.90 | 19.60 | 58.90
				20d9be29b5207b54e86aa1b1ac54872 | 8ae520247981aa06bc94abddf5f46d34 | 2018-04-09 10:09:40 | 199.00 | 19.27 | 199.00

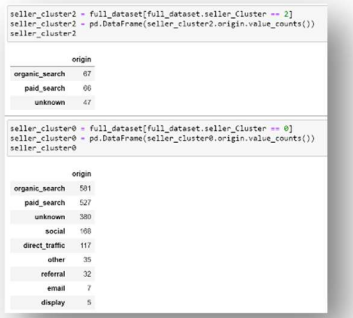
Merged marketing data Including missing values and a vague category(unknow)	Fill the missing values	179 observations have unknown origin. So, we will turn all the missing value at origin feature into unknown value	**Tool:** Python	All the missing values in origin of marketing channel fill with "unknown" value `merged_marketing_data.origin.value_counts()` organic_search 271 paid_search 195 unknown 193 social 75 direct_traffic 56 referral 24 email 15 display 6 other 4 other_publicities 3 Name: origin, dtype: int64
Merged_ecommerce_data (data file attained after merging other data files together) The "order purchase date" is not in a good format to analyse the trend by month and year customer_id | order_purchase_timestamp 92de9528e4b8f2568745e095beb94aeb | 2017-05-14 15:56:59 a0dd74c26f554c5ffd5bb25aab74fa1b | 2017-12-16 14:20:01 6ba6cff11767c3af78bebf58fe6e6cba | 2017-08-28 19:47:20 81f0639fc90123a98822afe9608e16fa | 2017-03-15 10:51:33	Convert to a year – month format	**Step 1**: Convert the column to datetime datatype **Step 2**: Using the map(lamda date: 100 * date.year + date.month) to convert to the new format **Step 3**: Sum over the month	**Tool:** Python **Package:** Pandas	We could calculate the sum of value by month and year order_year_and_month | seller_revenue 0 | 201609 | 215.28 1 | 201610 | 32407.19 2 | 201701 | 54771.28 3 | 201702 | 117664.95

3. Data checking in detail

Problems	Actions	Methods (How - Why)	Software	Outcome
Olist Seller Dataset Containing improper location format that seem fine at first during the cleaning process, however, cannot be recognized in Tableau geographic map  **Example:** +/ Using district name instead of city +/ Incorrect city name, typo +/Using state name instead city name	Matching location	Google the location to identify the mistakes District name replaced by city name Matching with the default location in Tableau Unidentified name will be matched by using the nearby city or provide the latitude + longitude English name turns into Latino name	Tool: Tableau	Location format is corrected and recognized by Tableau Your Data / Matching Location Andira-Pr → **Andirá** Andradas → **Andradas** Angra Dos Reis Rj → **Angra Dos Reis** Auriflama/Sp → **Auriflama** Balenario Camboriu → **Balneário Camboriú** Picarras → **Balneário Piçarras** Barbacena/ Minas Gerais → **Barbacena** Barro Alto → **Barro Alto** Mogi Das Cruzes / Sp → **Mogi Das Cruzes** Novo Hamburgo, Rio Grande Do Sul, Brasil → **Novo Hamburgo**

4. Data Exploration

What I do	Visualization	Tool	Outcome
Dashboard 1: "Product Category By Year Table" Visualize the product category by year to see the change of them year by year	Text table	Tableau	Many product categories were not in 2016, had been added in 2017, and some has been removed in 2018 such as fashion_children_clothers , cds_dvds. **Comment:** Those changes might reflect the changes of Olist (e-commerce company) business strategy in diversifying the product category in order to recruit more shops and attract more customers
Dashboard 1: "Top Popular Product Category" Using the groupby method in python to group the revenue by product categories Sort the product category by descending	Bubble chart with the presence of top 12 product categories ranked by revenue	Tableau And Python (pandas)	Not all the product Category perform the same, it could be seen that top 12 categories was all appeared in 2016 list.
Dashboard 1: "Relationship between Active Customer and Seller Revenue" Using the groupby method in python to compute the sum of customer and revenue by month during time period	Combine Charts: Line chart + Area Chart	Tableau And Python (pandas)	It seem like the extending of product category significantly impacted to the customer attration. The active customer maintaining a quite stable growth trend, which boost the revenue increse simutaneously.
Dashboard 2: "Who is making money" – Answer the question which are the top sellers? Where are they? What do they sell? *Apply K-NN algorithm:* 1. Using the elbow method to find the optimal clusters value 2. Fit the sum of seller revenue - segment the seller by their total revenue 3. Using kmeans.cluster_centers_ to see the average of total revenue in each cluster	Bar charts Symbol map	Tableau - Python – sklearn library	The purpose of this dashboard is to answer the question according to the seller as mentioned at the beginning: It could be seen that South regions of Brazil where the large cities and capital place has a dense distribution of seller. In fact, top sellers occupied about 3% in total number of sellers also located in this area (especially at Rio de Janeiro and Sao Paulo)

Filter and Group the total value by product category for each cluster	Pie Chart		The Top seller, in cluster 3, earning money by specilized in watches_gifts category while the lower rank one provide many different product types
Dashboard 3: "Customer distribution by Segment" – Answer the question: Where are the purchasers of top seller?	Symbol map Horizontal bar	Tableau	The purchaser spread over the coastal side of Brazil where have a high population density. However , the most active one are located around Sao Paulo and Rio de Janeiro Unsupringly, the majority of top customers come form both cluster 1 and 3 spend money on watches gifts
Answer the question: What are the marketing channels the company used to acquire them? After merging e-commerce data and the marketing data, I used value_counts() function in pandas to count the origin observation (marketing channel that use to recruit sellers) for each cluster 	Dataframe	Python	It could be seen that the higher rank sellers do not join the platform from email, social, display marketing, so we could reduce the budget for those marketing, instead put more money on SEO(Search Engine Optimization) and SEM(Search Engine Marketing) to increase the organic_search and paid search

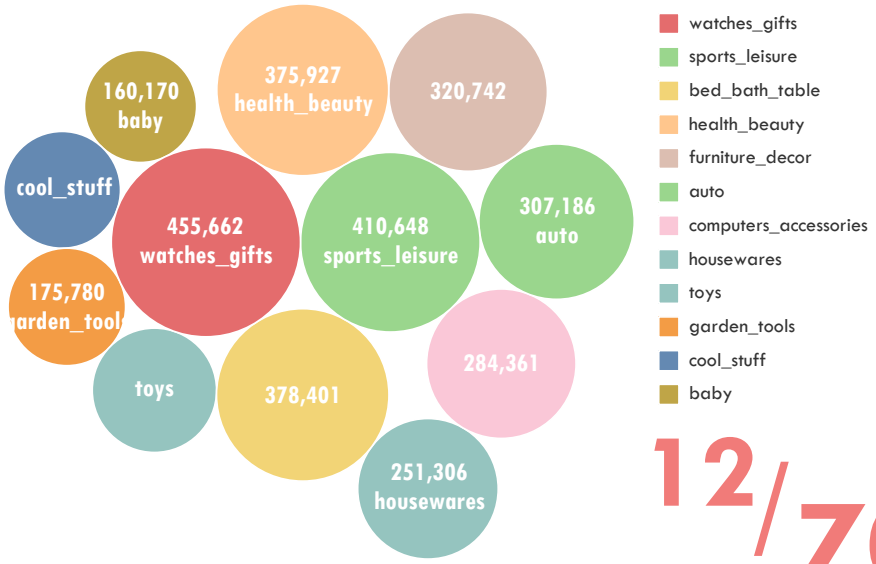
```
seller_cluster2 = full_dataset[full_dataset.seller_Cluster == 2]
seller_cluster2 = pd.DataFrame(seller_cluster2.origin.value_counts())
seller_cluster2
```

	origin
organic_search	67
paid_search	66
unknown	47

```
seller_cluster0 = full_dataset[full_dataset.seller_Cluster == 0]
seller_cluster0 = pd.DataFrame(seller_cluster0.origin.value_counts())
seller_cluster0
```

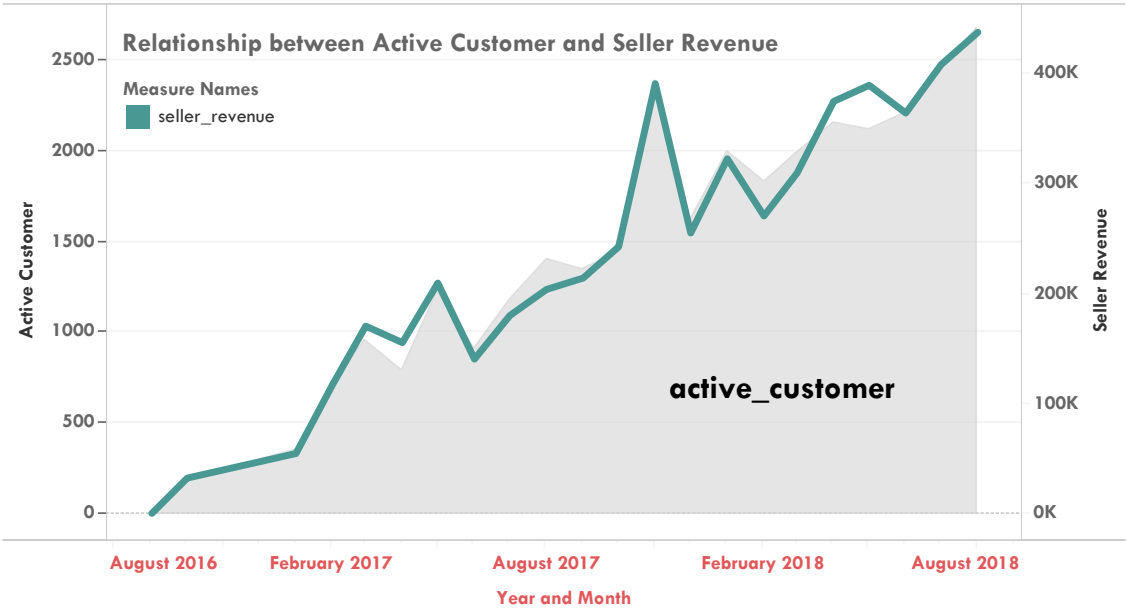
	origin
organic_search	581
paid_search	527
unknown	380
social	168
direct_traffic	117
other	35
referral	32
email	7
display	5

Dashboard 1: Overview

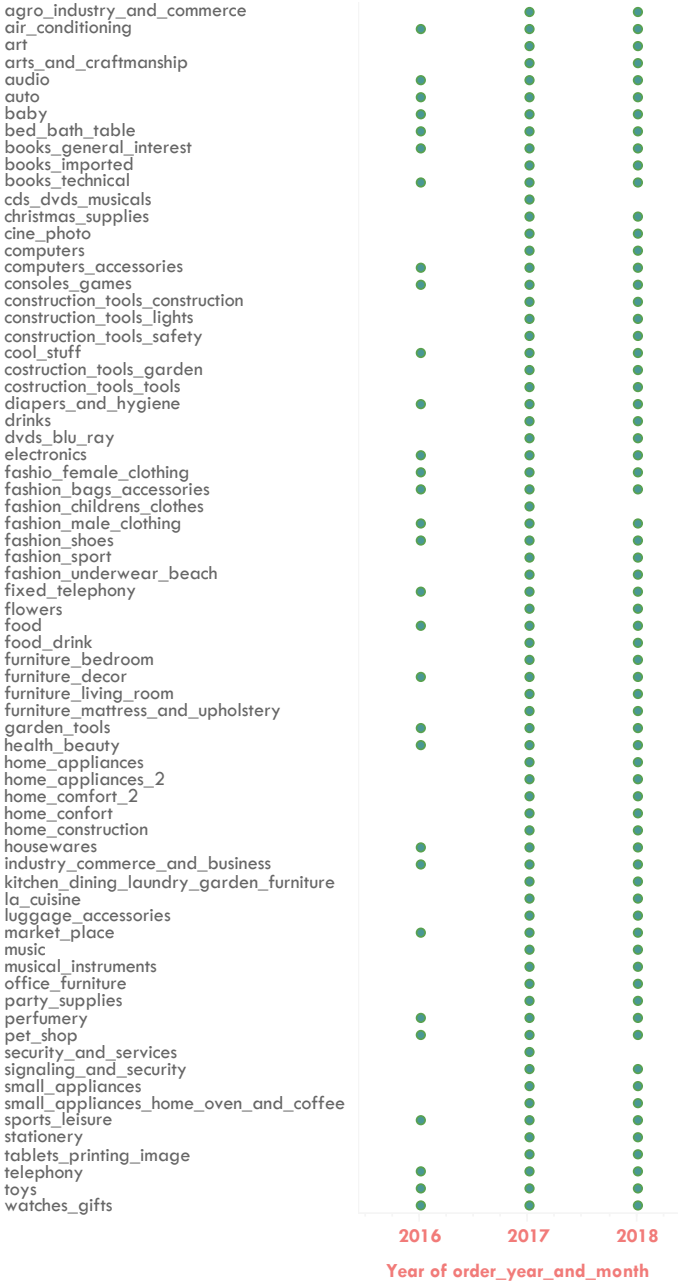


12/70 Product Category

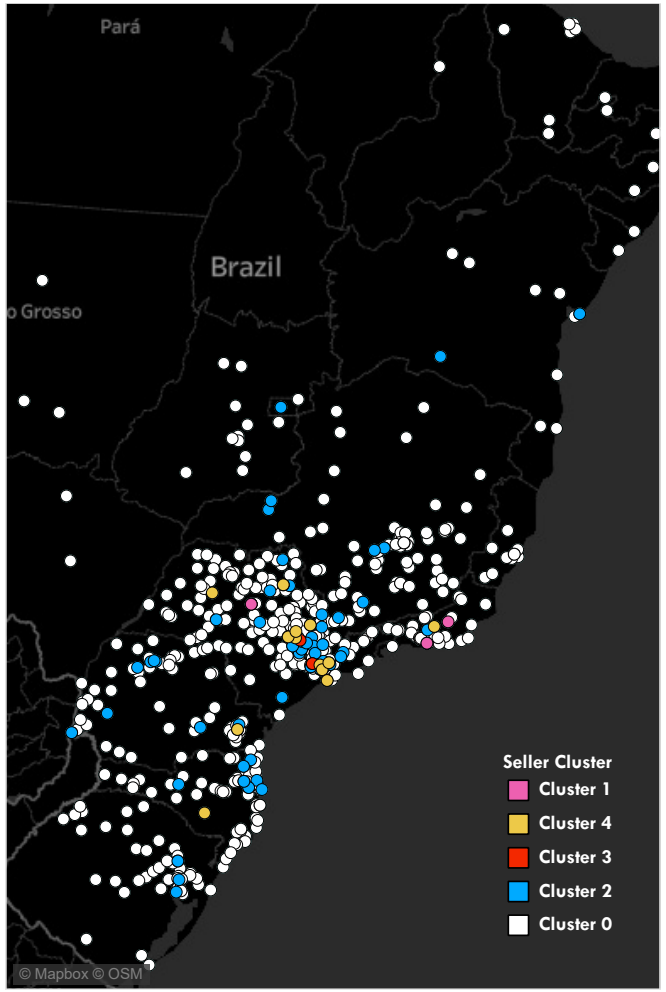
Top Popular Product Category



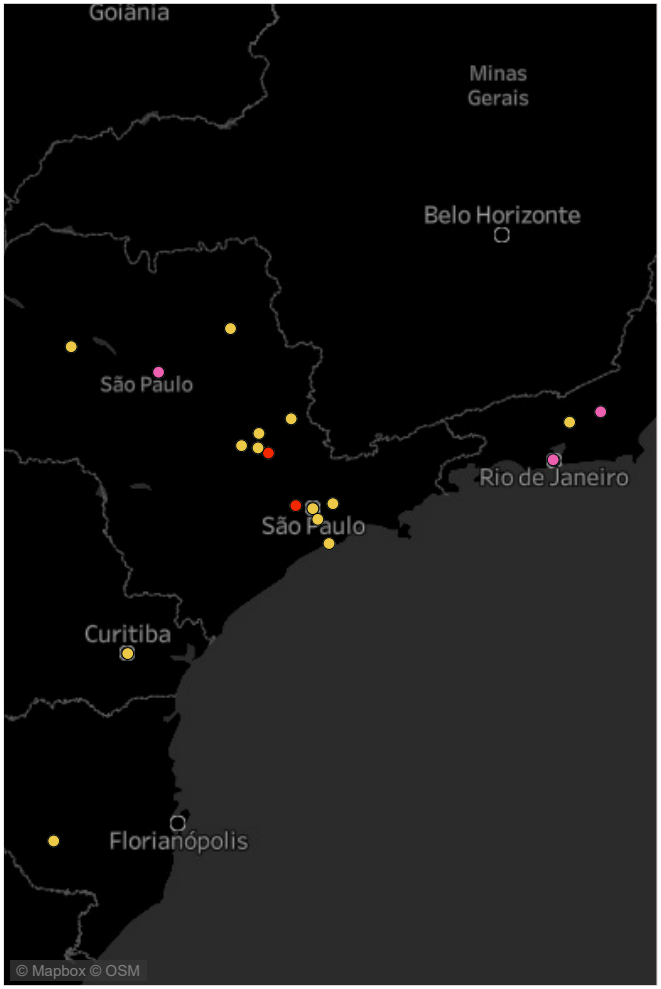
Product Category By Year



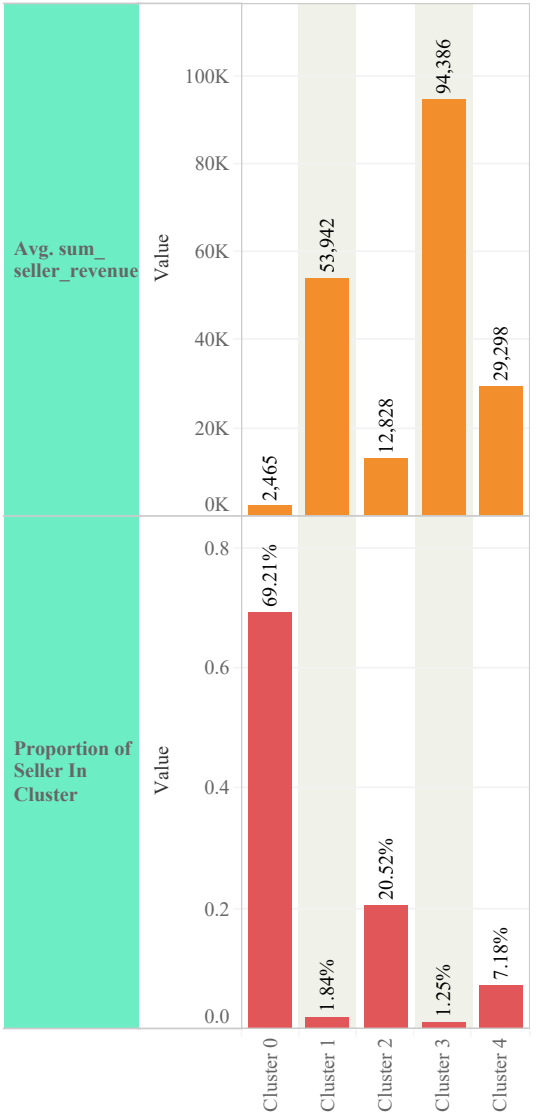
Seller Distribution - Brazil



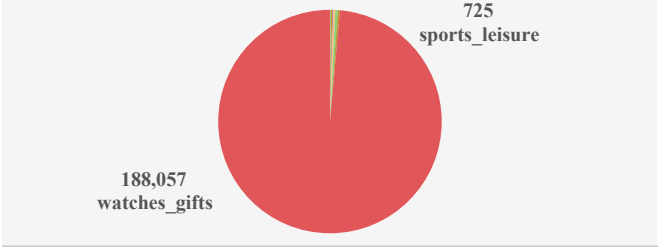
Top Seller Distribution



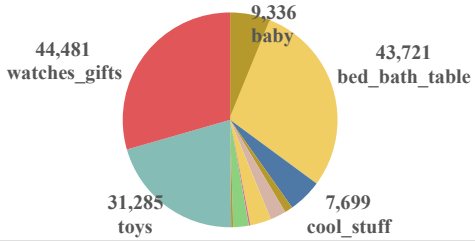
Cluster Segment Overview



Cluster 3 - Revenue by Product Category

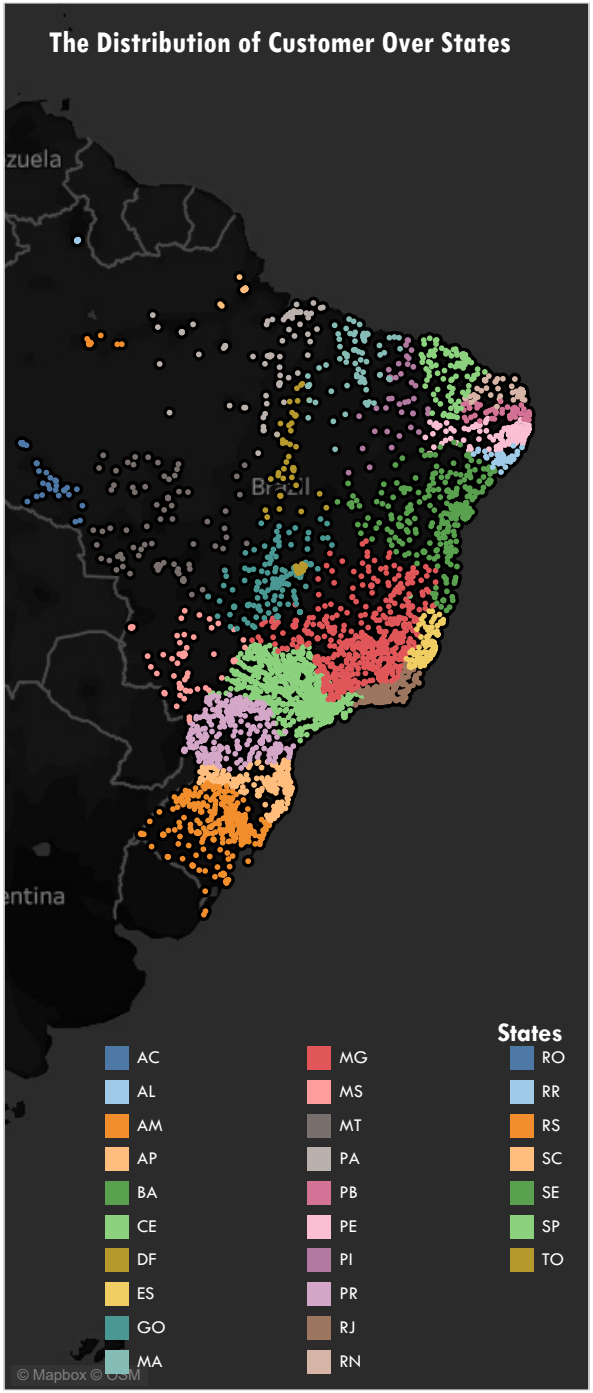


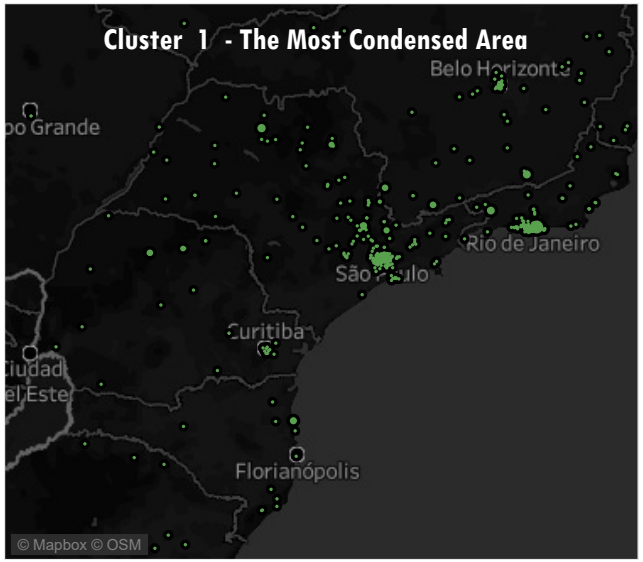
Cluster 1 - Revenue by Product Category

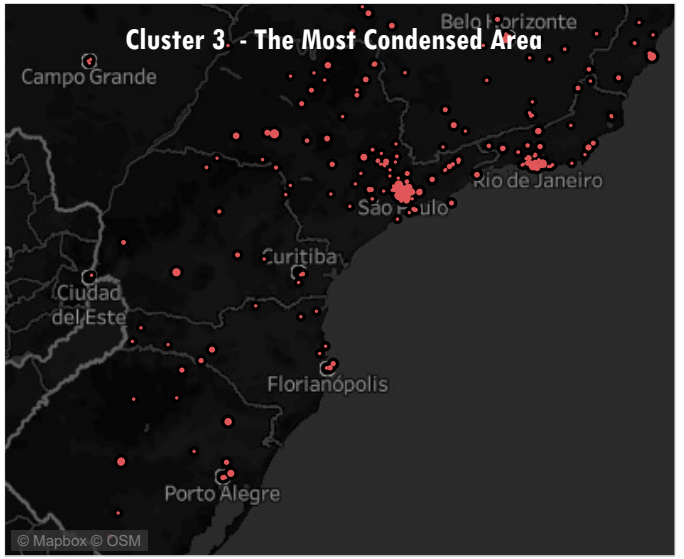


Dashboard 2

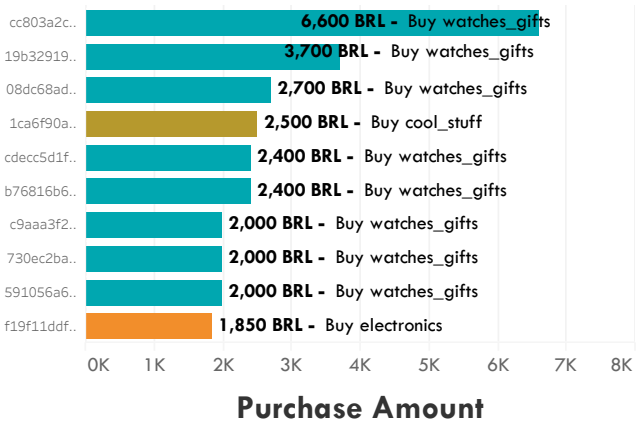
"Who is making money"







Top Ten Customers and Their Buying Category in Cluster 1 and 3



Dashboard 3

Customer Distribution by Segment

5. Conclusion

What I learnt from the data:

Expanding the product category draw more customers to the platform, so the company should maintain introduce the new category as well as improve the quality of product selling in the platform based on the customer reviews

Except the top 12 product categories, these other category performances are not as good as expected. To diversify product, we could search for the shops who has advantages in other categories as well. Firstly, we could start from the big states such as Rio or Sao Paulo where our top sellers located to find the potential candidates. The marketing method could be used are Search Engine Optimization and Search Engine Marketing, besides cut off other inefficient marketing channels such as email or display.

Watch gifts is the most popular product and the customer who spending the most are in the also gather around the big cities and states in Southern of Brazil. Obviously, the marketing budget should be allocated more to this area instead of spreading widely to another city and states.

6. Reflection

What I learnt in this project:

1. Understanding the meaning of each feature is a crucial step. In this data, amount of product is record as a sequence that created many duplications that does not help to answer our questions
2. The importance of data cleaning. The "city" feature includes many improper data formats and some of them is difficult to recognized such as name of district had been used instead of city name, or some of location cannot be found on the Tableau map
3. Using different type of charts and graphs in Tableau to obtain the insights

What in hindsight you might have done differently: select a smaller number of clusters maybe 3.