Expanding E-Commerce Services in Brazil

Ammad Baber
Evan Canfield
Roma Dutta
Max Lawrence
Amit Pandit

Olist

- Brazilian E-Commerce Website
- Growing well, and internationally recognized as a leader in this new market
- Attempting to expand the number of small retailers that use the website as a platform
- Hope this expansion will allow them to remain ahead of new competitors (Amazon, other Brazilian startups, etc.)
- Gave out data on Kaggle competition in order to allow for others to help them

Our Research Question

How can Olist increase the number of sales they make in order to enhance their attractiveness to small scale retailers that they are competing for against Amazon and other e-commerce providers?

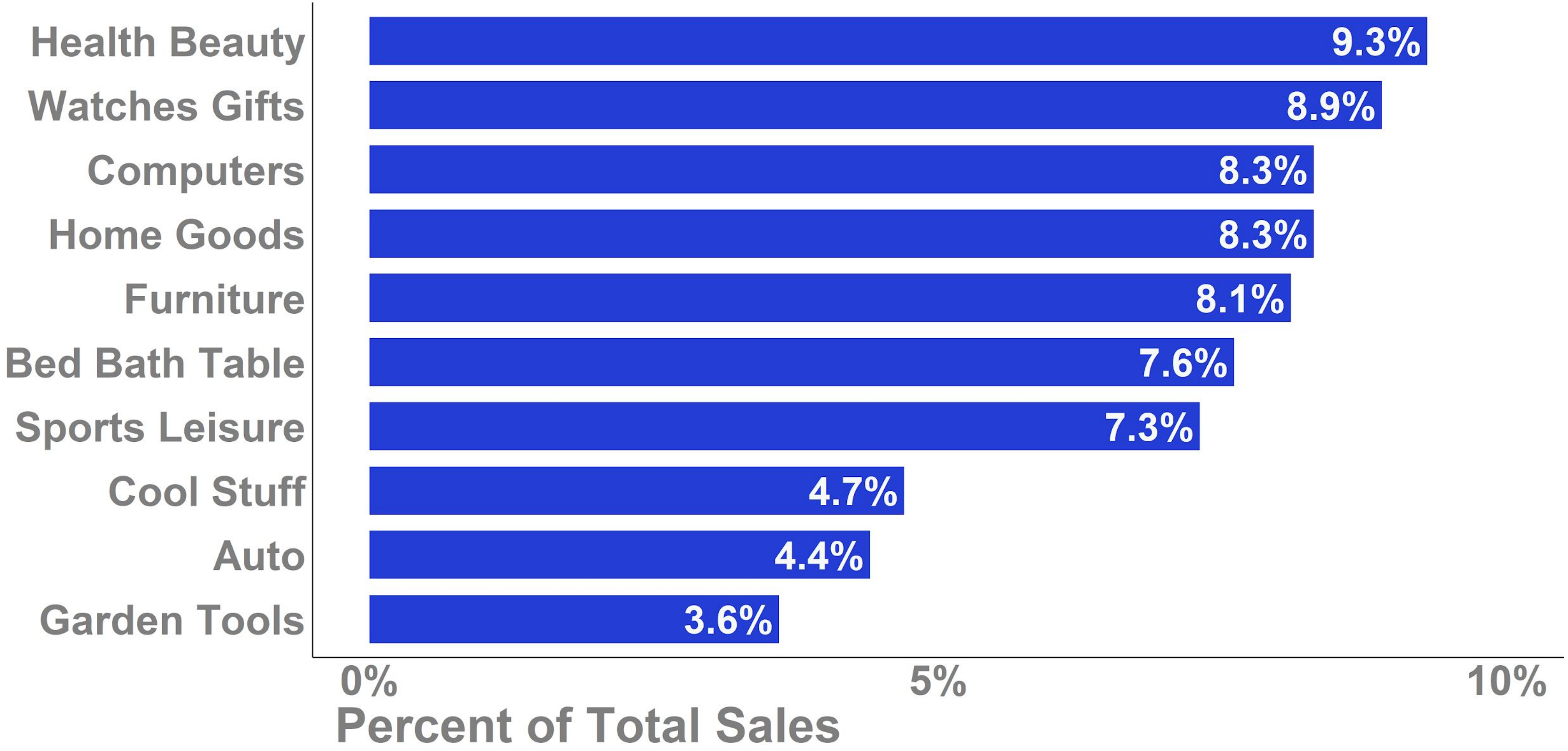
Data

- The main data we used came from Olist providing it to Kaggle
- The dataset was a sample of the sales data from 2016 through 2018. The dataset included information on 100,000 unique sales.
- The dataset also included additional information:
    - Geographic Location (Customer and Seller)
    - Detailed Order Description
    - Logistical Operation Information for each order
- We also used used Brazil’s 2019 population projections for secondary analysis

Data Description (Summary Statistics)

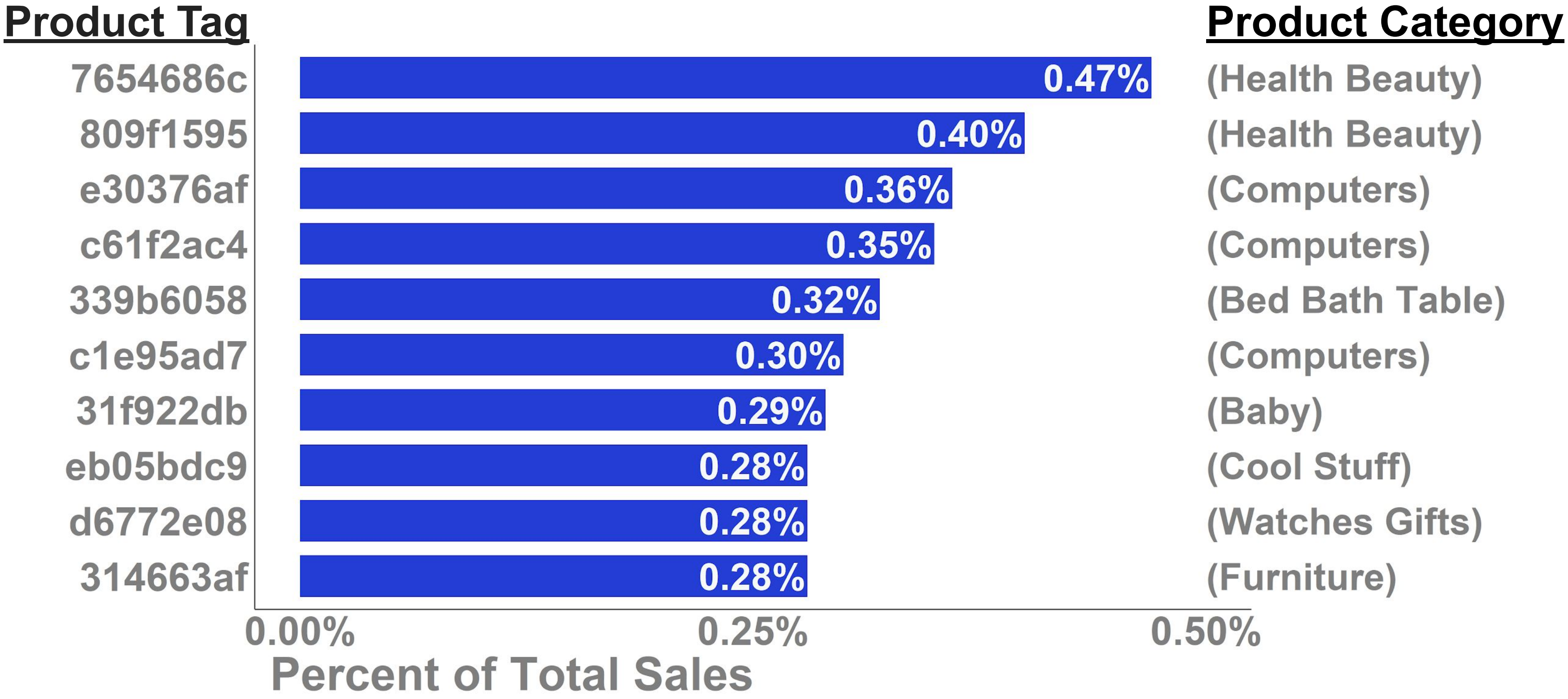
- Purchases per Customer
  - Minimum: 1
  - Median: 1
  - Maximum: 17
- Price by Order($R)
  - Minimum: 0.01
  - Mean: 154
  - Maximum: 13,664
- Delivery Time (Days)
  - Minimum: 0.5
  - Mean: 13
  - Maximum: 210

Top 10 Product Categories



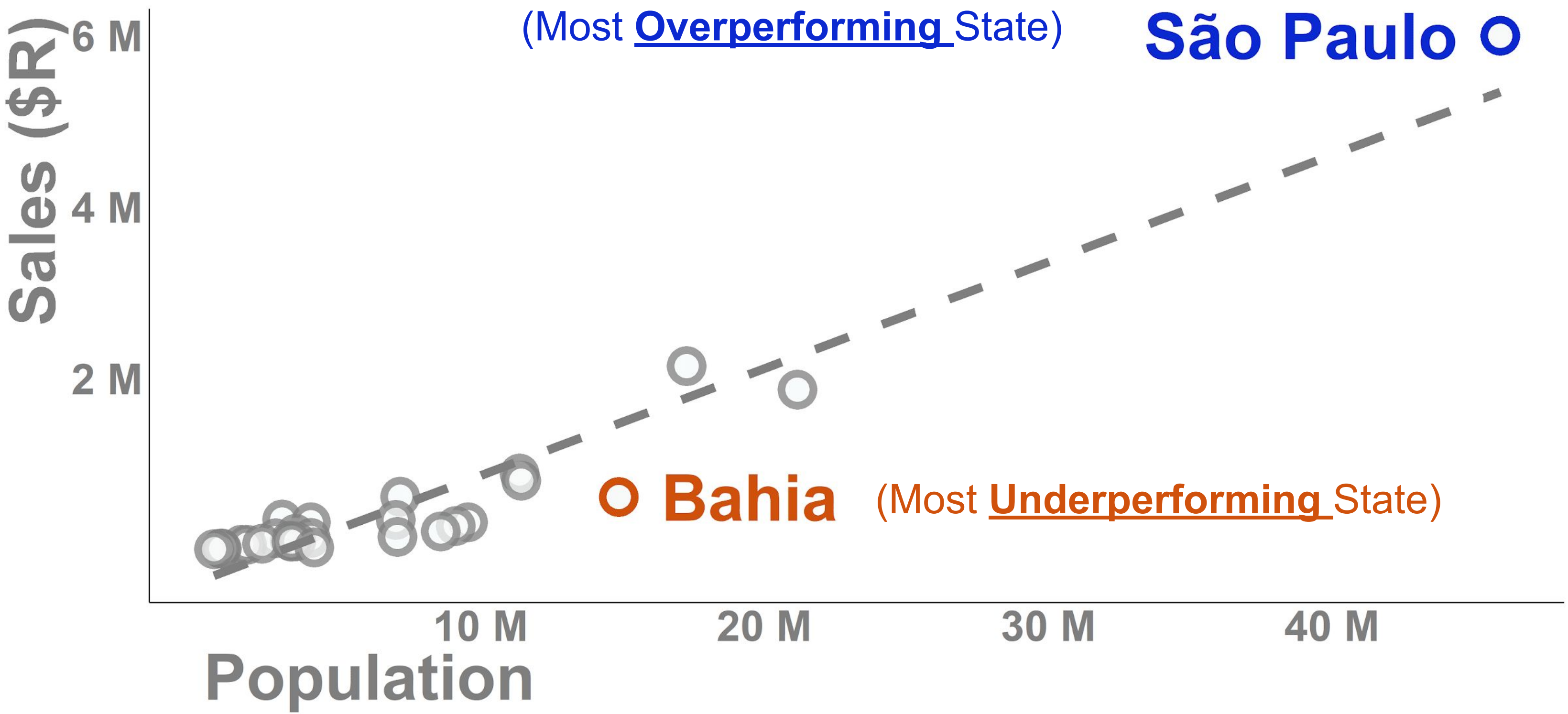
Source: Kaggle

Top 10 Products



Source: Kaggle

State Sales is Correlated with Population

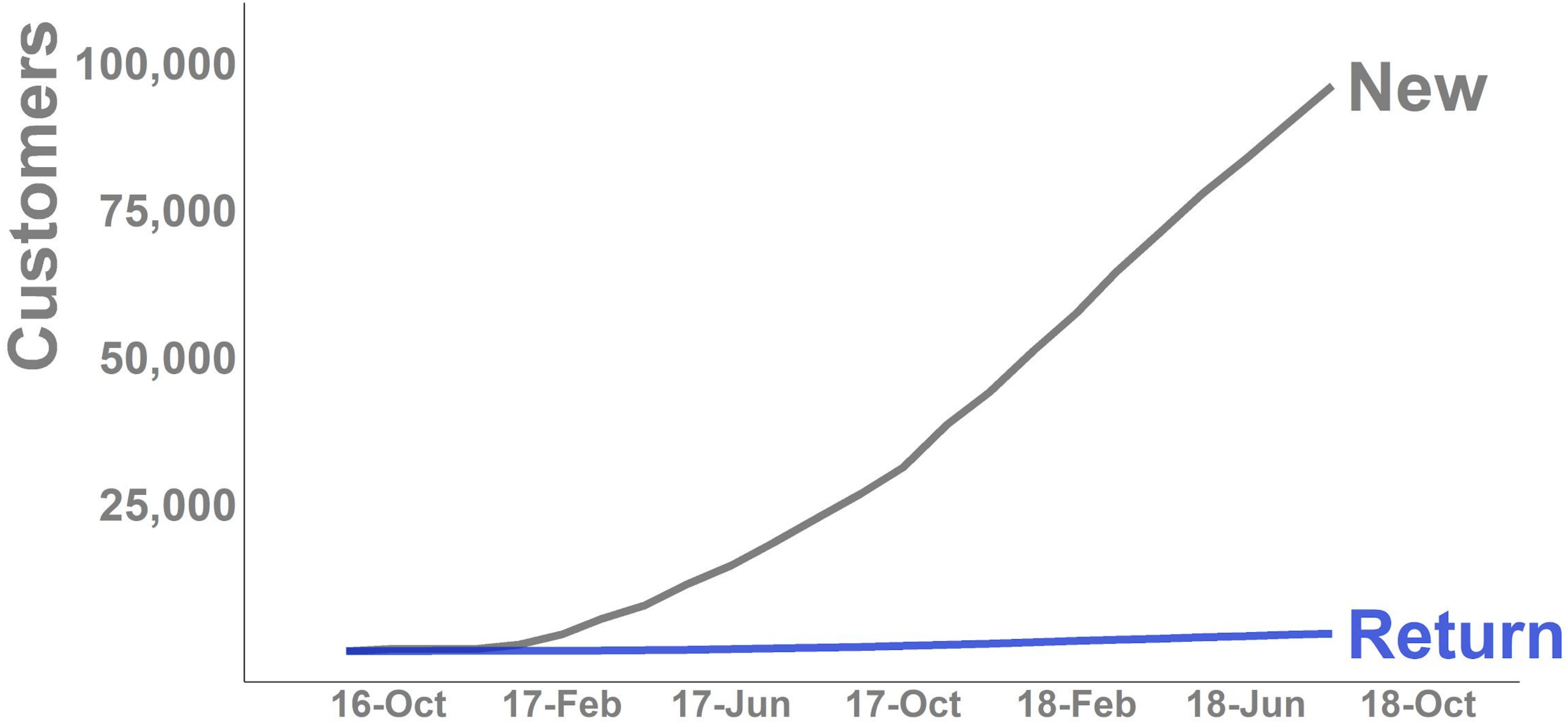


Source: Kaggle

Local Shipments Are Far Less Expensive

Source: Kaggle

Return Customer Growth is Sluggish



Source: Kaggle

Return Customer Breakdown

- While Brazil is a very large country, Olist cannot continually rely only on new customers as growth increases.
- To safely grow the total sales, Olist must increase the percentage of return customers.
- Current Return Customer Statistics:
  - Make up **3.1%** of all unique customers
  - Account for **6.7%** of all items ordered
  - Account for **5.7%** of all sales (R$)

Return Customer Modeling

- In order to identify factors important to Return Customers, we performed a logistic regression analysis with the data.
- This dataset included the following variables:
  - Target Variable
    - Return: Binary variable indicating if a customer made more than one unique orders
  - Predictor Variable - Continuous
    - Spend: Customer spend, average ($R)
    - Distance: Distance between customer and seller, average (km)
    - Shipping Time: Time from purchase to delivery, average (days)
  - Predictor Variable - Categorical
    - Late: Binary variable indicating if the package was delivered before or after projected date
    - Method of Payment (4 categories)
    - Product Category (39 categories)
    - Review Score: 1 - 5 Likert Scale (discrete values)

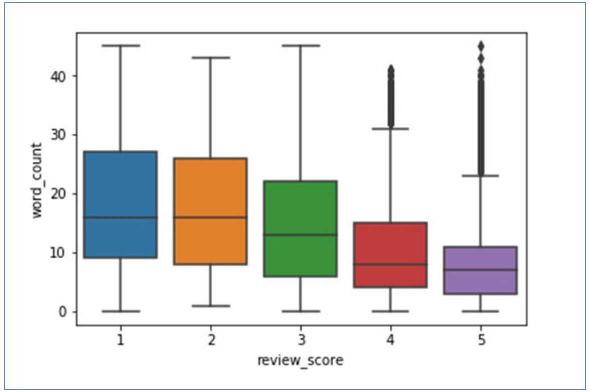
Return Customer Factors

Based on the the logistic regression targeted at return customers, the following factors were found to be important and significant:

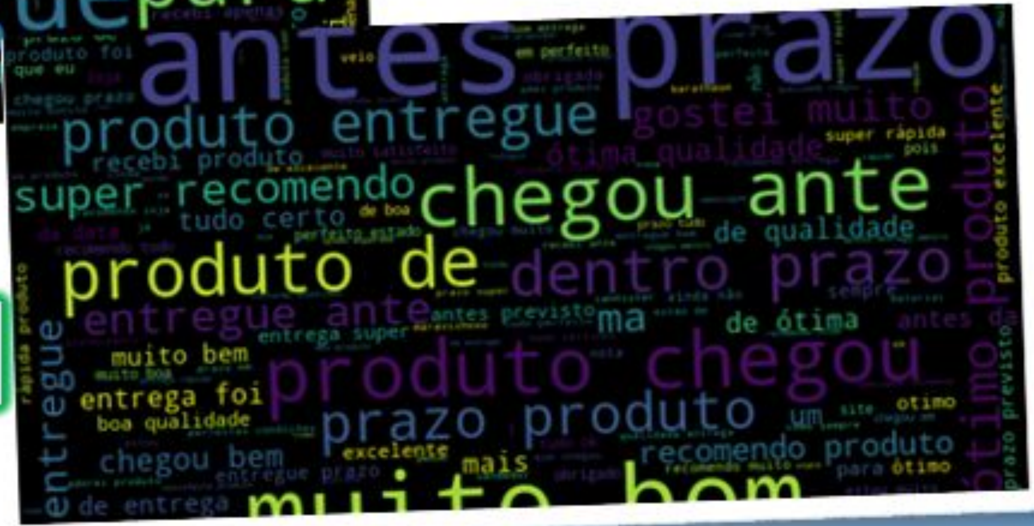
- The number of days from purchase to delivery have **no effect** on being a Return Customer (Odds Ratio of 1.02, p value < 0.05)
- BUT customers who received late orders were **76% less likely** to be Return Customers, compared to customers who did not receive late orders.
- Product Categories (compared to customers who purchased from *Health Body)*:
  - Customers who purchased from *Bed Bath Table* were **43% more likely** to be Return Customers
  - Customers who purchased from *Cool Stuff* were **37% less likely** to be Return Customers

Customer Sentiment

- Less satisfied customers are more likely to write review comments
- Less satisfied customers are more likely to write longer comments
- 1-star reviews are complaining about :
  - Not received their goods yet
  - Difficulty purchasing multiple products together
  - Received wrong and bad quality fake goods
- Customers with 5-star reviews enjoy most about the fast delivery



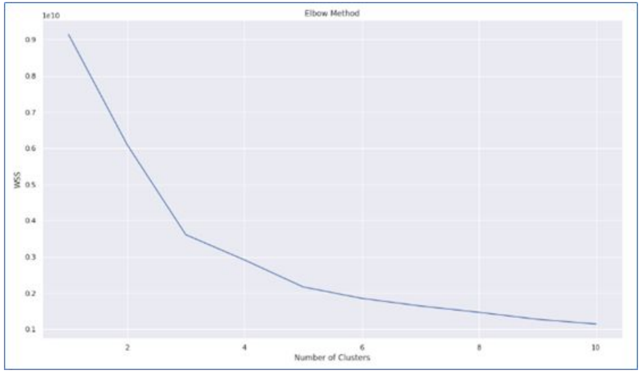
Wordcloud where review score is 1



Wordcloud where review score is 5

Customer Segmentation - Method 1

- In order to deduce necessary marketing action, we performed K-means clustering to broadly divide customers into different segments.
- Elbow method indicated to have K value as 5.
- Customers,Orders and Order Payments datasets were merged to derive the variables - Recency, Frequency, Monetary_Value and Client_Since.



	Recency				Frequency				Monetary_Value				Client_Since			
	count	min	mean	max	count	min	mean	max	count	min	mean	max	count	min	mean	max
prediction																
0	34718	18	265.898525	371	34718	1	1.070540	3	34718	10.07	138.288359	682.35	34718	165	269.843511	609
1	21503	92	469.956239	708	21503	1	1.067618	3	21503	11.63	140.114779	982.41	21503	369	472.006418	708
2	2239	17	241.198749	704	2239	1	1.132649	4	2239	660.44	1211.758276	13664.08	2239	17	248.707012	704
3	357	18	270.344538	708	357	4	5.941176	33	357	19.00	231.561008	1756.53	357	18	314.028011	708
4	34524	14	99.933843	185	34524	1	1.041536	3	34524	9.59	139.320631	754.46	34524	14	100.893842	318

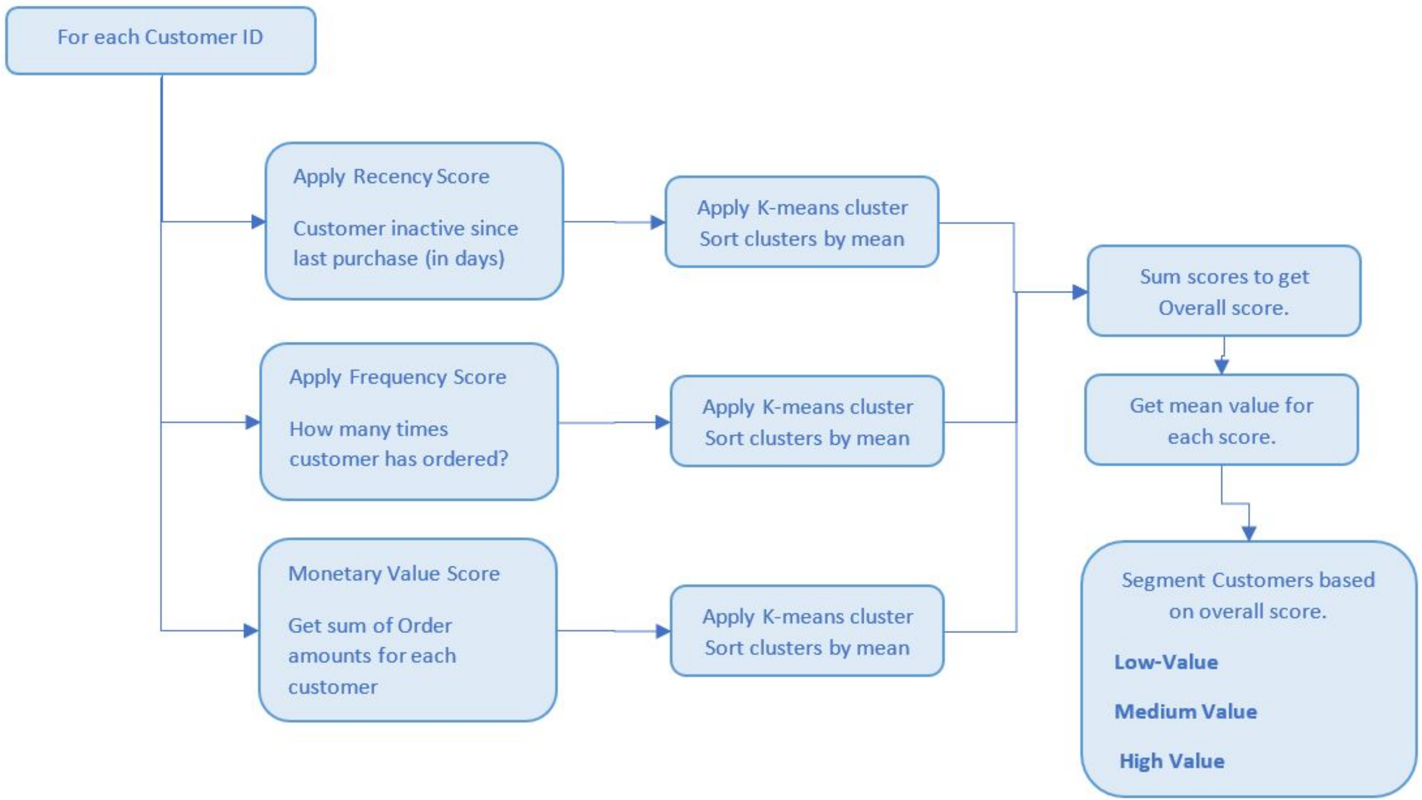
#CLUSTER	CLUSTER LABEL	DESCRIPTION	MARKETING STEP
2	Big Spenders	Spend the most	Olist can market their most expensive products
3	Best Customers	Bought most often	Need no price incentives or loyalty programs. Olist can easily campaign for new products
4	Loyal Customers	Buy most frequently, however, conservative spender	Little to medium price incentives
0	Almost Lost	Have not purchased for long, spend the least	Aggressive price incentives
1	Lost	Have not purchased for long	Aggressive price incentives in terms of voucher/coupon etc

Customer Segmentation - Method 2

Recency, Frequency, Monetary (RFM)

- **Low Value**: Customers who are less active than others, not very frequent buyer/visitor and generates very low - zero - maybe negative revenue.
- **Mid Value**: In the middle of everything. Often using our platform (but not as much as our High Values), fairly frequent and generates moderate revenue.
- **High Value**: The group we don't want to lose. High Revenue, Frequency and ow Inactivity.

Customer Segmentation - RFM



Customer Segmentation - RFM (Cont.)

	customer_unique_id	Recency	RecencyCluster	Frequency	FrequencyCluster	Monetary	RevenueCluster	OverallScore
0	708ab75d2a007f0564aedd11139c7708	125	3	1	0	99.33	0	3
1	0ae522661311f598df20272643d39ce6	160	3	1	0	157.45	0	3
2	d386a136dc889cf681443061874caad8	113	3	1	0	136.71	0	3
3	3af91c25f393301d27ccd39bce43f29d	112	3	1	0	103.49	0	3
4	1504b719eb728716ff7583dc7b86a0b0	132	3	1	0	64.03	0	3
...	...	...	...	...	...	...	...	...
93078	20a5257c01689ac69410a14cb51bb447	351	1	10	3	17671.00	2	6
93079	698e1cf81d01a3d389d96145f7fa6df8	371	1	20	3	45256.00	2	6
93080	93bc212addb25a5f5139fded3c2ee6b3	328	1	10	3	18667.00	2	6
93081	3d47f4368ccc8e1bb4c4a12dbda7111b	330	1	10	3	22346.60	2	6
93082	0a0a92112bd4c708ca5fde585afaa872	333	1	8	2	109312.64	3	6

93083 rows × 8 columns

```
df_user.groupby('OverallScore')['Recency','Frequency','Monetary'].mean()
```

OverallScore	Recency	Frequency	Monetary
0	501.960153	1.000000	142.786789
1	385.694601	1.116281	170.250610
2	268.408445	1.137355	168.483845
3	166.046097	1.194145	189.164205
4	67.238617	1.251121	209.367601
5	68.585628	2.545332	584.127740
6	86.768085	4.855319	2301.905277
7	72.064000	7.064000	3719.160080
8	79.227273	10.454545	11435.360455
9	21.000000	24.000000	27935.460000

```
In [325]: df_user.groupby('Segment').customer_unique_id.count()

Out[325]: Segment
          High-Value      148
          Low-Value     69413
          Mid-Value     23522
          Name: customer_unique_id, dtype: int64
```

Customer Segmentation - Analysis

	customer_unique_id	Recency	RecencyCluster	Frequency	FrequencyCluster	Monetary	RevenueCluster	OverallScore	Segment
0	708ab75d2a007f0564aedd11139c7708	125	3	1	0	99.33	0	3	Low-Value
1	0ae522661311f598df20272643d39ce6	160	3	1	0	157.45	0	3	Low-Value
2	d386a136dc889cf681443061874caad8	113	3	1	0	136.71	0	3	Low-Value
3	3af91c25f393301d27ccd39bce43f29d	112	3	1	0	103.49	0	3	Low-Value
4	1504b719eb728716ff7583dc7b86a0b0	132	3	1	0	64.03	0	3	Low-Value
...	...	...	...	...	...	...	...	...	...
93078	20a5257c01689ac69410a14cb51bb447	351	1	10	3	17671.00	2	6	Mid-Value
93079	698e1cf81d01a3d389d96145f7fa6df8	371	1	20	3	45256.00	2	6	Mid-Value
93080	93bc212addb25a5f5139fded3c2ee6b3	328	1	10	3	18667.00	2	6	Mid-Value
93081	3d47f4368ccc8e1bb4c4a12dbda7111b	330	1	10	3	22346.60	2	6	Mid-Value
93082	0a0a92112bd4c708ca5fde585afaa872	333	1	8	2	109312.64	3	6	Mid-Value

93083 rows × 9 columns

We can align different marketing campaigns/strategies based on the different customer segments.

Retention Rate

	month_y	TotalUserCount	RetainedUserCount	RetentionRate
0	201703	2508	3	0.001196
1	201704	2274	11	0.004837
2	201705	3478	14	0.004025
3	201706	3076	16	0.005202
4	201707	3802	16	0.004208
5	201708	4114	23	0.005591
6	201709	4082	32	0.007839
7	201710	4417	32	0.007245
8	201711	7182	37	0.005152
9	201712	5450	41	0.007523
10	201801	6974	16	0.002294
11	201802	6401	27	0.004218
12	201803	6914	23	0.003327
13	201804	6744	31	0.004597
14	201805	6693	45	0.006723
15	201806	6058	38	0.006273
16	201807	6097	26	0.004264
17	201808	6310	37	0.005864

Retained User Count: Customers for the month who also visited previous month.

Total User Count: Customers for the current month. We see that retention rate is very small, which is a issue since there are very few repeat customers.

Sellers have to address these issues in future.

Future Considerations for Analysis

1. We lack standard store sales data (compared to e-commerce data)
2. We lack data from other ecommerce sites (Amazon or otherwise)
3. Expanded internal dataset. Research indicates that there may be implicit bias in the creation of the available dataset.
4. We weren't provided demographic info on customers (Age, sex, income, etc.)
5. Data set is biased to those who leave reviews

Conclusion

- Olist is not building a loyal customer base of returning shoppers to attract small retailers with
- Reliable shipping and operational efficiency seem to be the traits that attract customers to return
- Customers and sellers are mostly located in the more developed sections of Brazil (South-East Coast, etc.)
- There are clear segments of customers based on how much they spend, and what types of products they buy

Recommendation

- Olist should invest in assisting retailers in delivering the product (monitoring, delivery company contacts, etc.)
- Olist and/or sellers should offer customers discounts and sales based on their background and type (Coupons for products that often are bought by return, etc.)
- Marketing campaign catering to differing kinds of customers based on customer segmentation
- Invest in more marketing in states that are underperforming based on population (will also allow for building brand recognition)
- Olist should explore the possibility of creating its own delivery service in the future to combat Amazon

Questions?