Olist – Customer Segmentation

Author: Kimberly Cable (November 2022)

Introduction

Olist is a Brazilian e-commerce company that connects small businesses to a larger marketplace. It gives these small businesses a way to manage their products, shipping, and online payments. They have approximately 200,000 users in about 180 countries.

With any online retailer, retaining customers is key but knowing who may leave and who stays could be guesswork. Understanding how and why customers stay and shop and why they leave is pivotal to a company's business.

Knowing who your customers are and how and why they shop or do not return plays a big part in customer service which ultimately enhances a customer's satisfaction level. A high satisfaction level could lead to overall better reviews and with these good reviews, their sellers will see more new sales.

This study will hope to find similar traits among its customers. This will help the marketing team know who best to send offers to or to whom might they need to send a discount coupon because it looks like they haven't purchased in a while.

The data has been sourced from Kaggle which has 100,000 online orders from 2016 to 2018. The data consists of records with products, customers, and review information for each transaction provided by Olist.

The data consists of the following datasets (see Figure 1):

- **Customers**: This dataset has information about the customer and their location.
- **Geolocation**: This dataset has information about the Brazilian zip codes and their latitude/longitude coordinates.
- **Order Items**: This dataset has information about the items purchased within each order.
- **Order Payments**: This dataset has information about the order payment options.
- **Order Reviews**: This dataset has information about the reviews made by the customers.
- **Orders**: This dataset has information about all customer orders.
- **Products**: This dataset has information about all the products sold by Olist.
- **Sellers**: This dataset has information about the sellers that fulfilled the orders made at Olist.
- **Category Name Translation**: This dataset has English/Portuguese translations for all products sold at Olist.

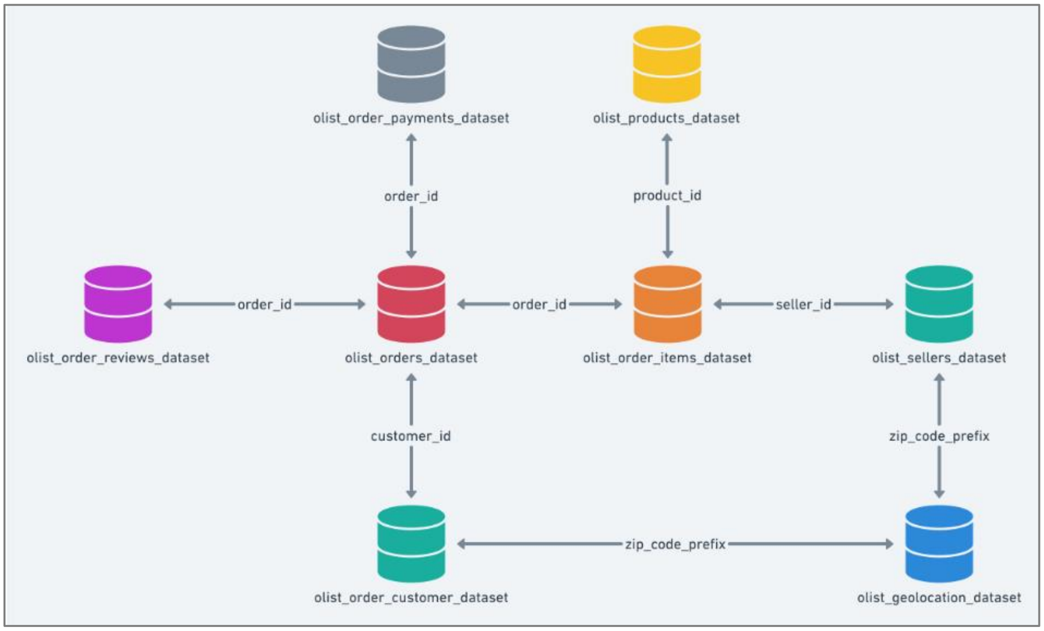


Figure 1: Data Schema

Methods/Results

Data Prepping:

Customers:

The customer's dataset consists of a unique customer identifier and the city/state/zip of each customer. As all customers resided in Brazil, I only kept the customer identifier and state for determining where most of the customers originated from.

Orders:

The order's dataset consists of a unique customer identifier along with order details such as the status of the order, the date of purchase, the order approved date, and the order delivered date. All dates were converted to a DateTime type.

Looking at the graph of orders over time, I noticed little to no orders in 2016 and little after August 2018 so I dropped those orders (see Figure 2).

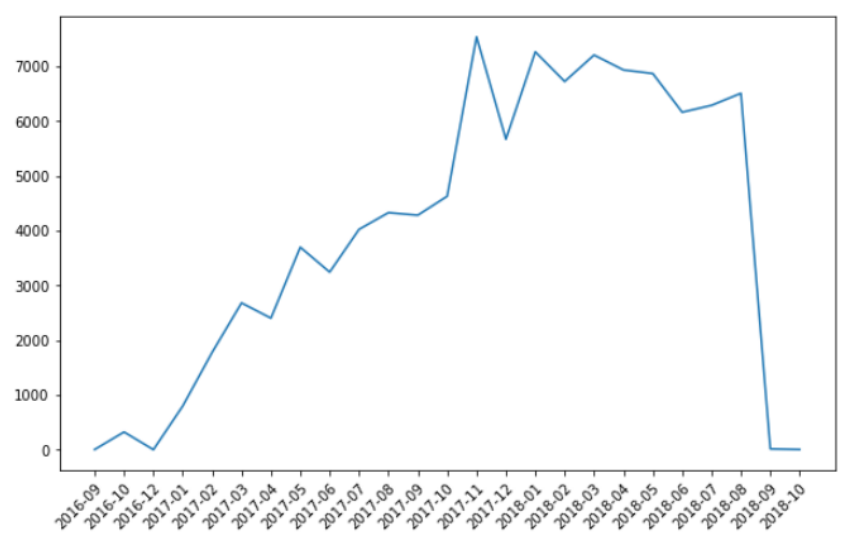


Figure 2: Number of Orders over Time

There were 82 empty values in the order_approved_at column. I looked at the order_status of each of the 82 values. 63 were ‘canceled’, 14 were ‘delivered’, and 5 were ‘created’. The ‘delivered’ status should not have any empty order_approved_at values so I replaced the null values with the order_purchase_timestamp.

There were 1,602 empty values in the order_delivered_carrier_date column. The order status for ‘delivered’ was only 2 and the rest had status’ that were okay for this field. I replaced the 2 empty values with the order_approved_at date.

There were 2,727 empty values in the order_delivered_customer_date column. The order status of ‘delivered’ was only 8 and the rest had status’ that were okay for that field. Looking at the histogram, I saw that it was skewed left so to replace this field, I got the difference between the order_delivered_customer_date and the order_delivered_carrier_date. Then I took the median number of days. I added the number of days to the order_delivered_carrier_date to get the new order_delivered_customer_date for those empty values. In this case, the median difference was 7 days which was added to the order_delivered_carrier_date.

Final Dataset: 92, 580 rows and 8 columns

Order Items:

The order items dataset consists of an order identifier along with specifics about each item purchased in the order like the product identifier, the seller identifier, and the freight value.

Order Reviews:

The order review dataset consists of a review identifier along with the review score and comments. Only the review identifier, order identifier, and review score were kept as the other information will not be needed for customer segmentation.

Products:

The product dataset consists of the product identifier and the category the product belongs to. The other information in the dataset, like specific product information, was dropped as it was not needed for customer segmentation.

Sellers, Order Payments, and Geolocation:

These datasets were not needed for the customer segmentation model and therefore dropped.

Final Dataset

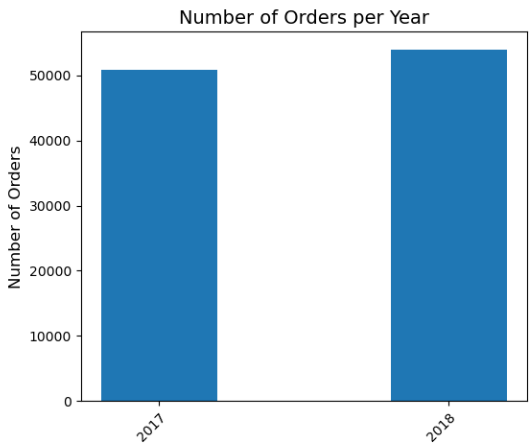
After initially prepping each dataset during the preliminary analysis, I merged all datasets using the following keys:

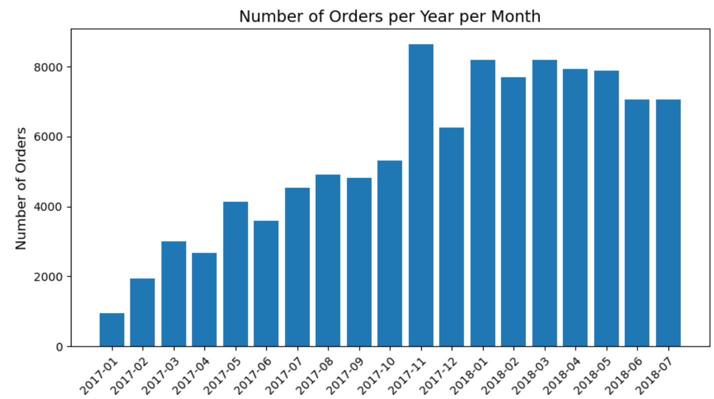
Order Reviews ← order_id → Orders ← order_id → Order Items ← product_id

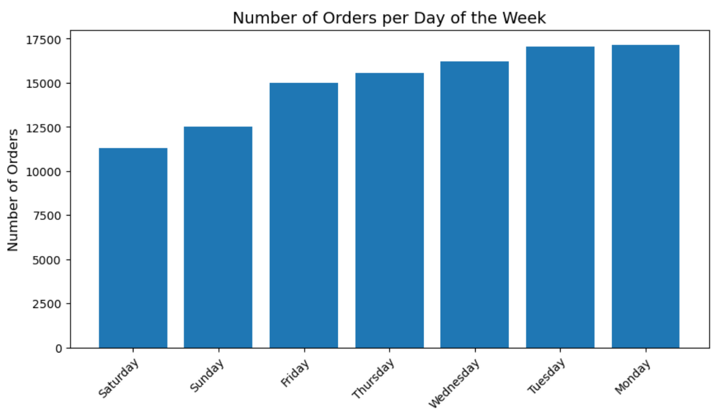
Once merged, the dataset had 104,782 rows and 15 columns.

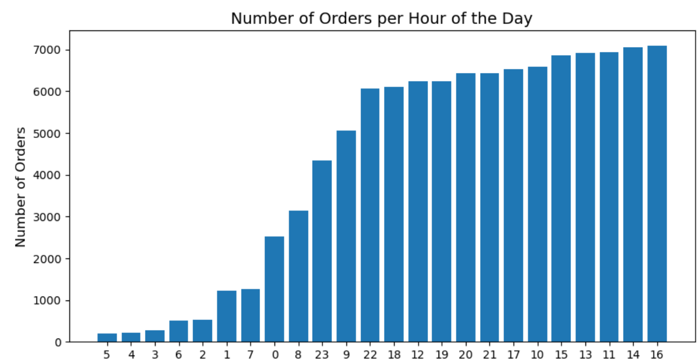
Visualizations:

Order Purchases Timeframes:



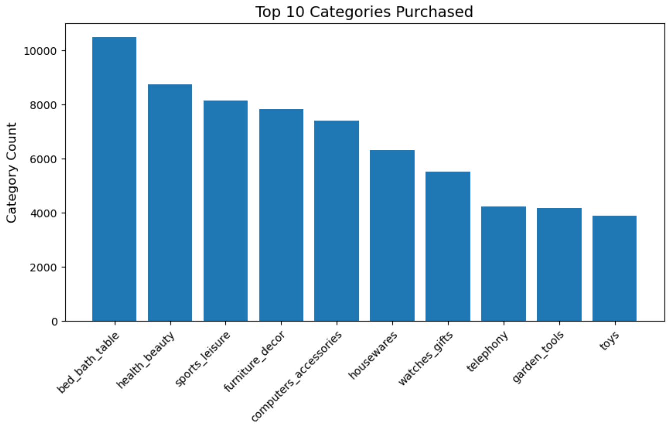


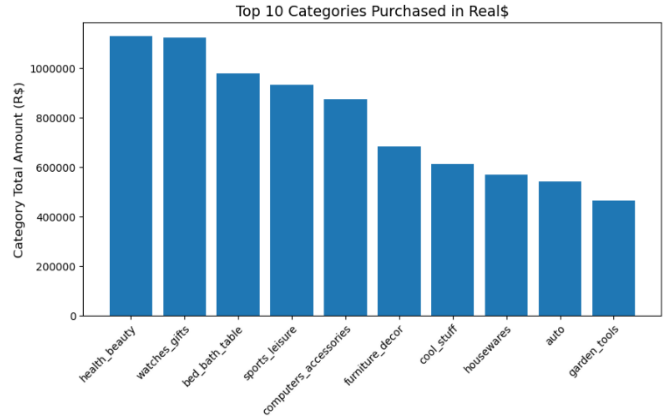




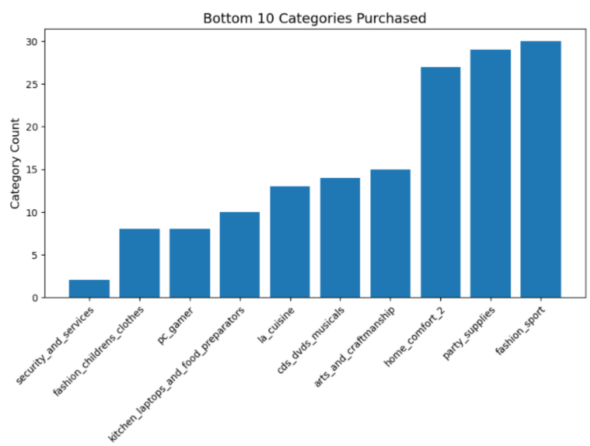
The number of orders grew between 2017 to 2018. November and January had the highest number of orders probably due to the holidays. March, April, and May were also good purchasing months. Weekdays were also more popular purchasing days as the beginning of the week was best probably due to items needed from the weekend. and early evenings.

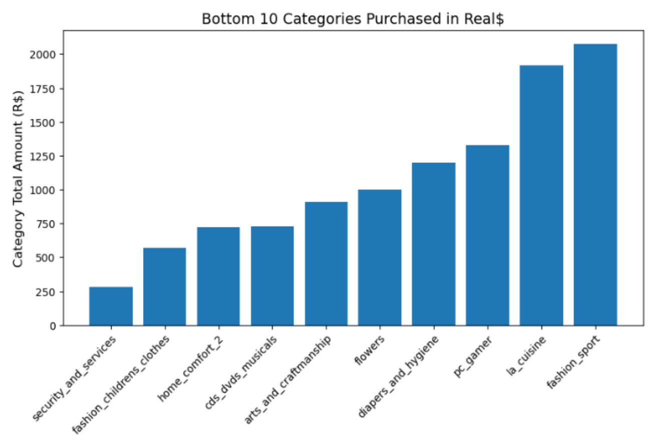
Categories





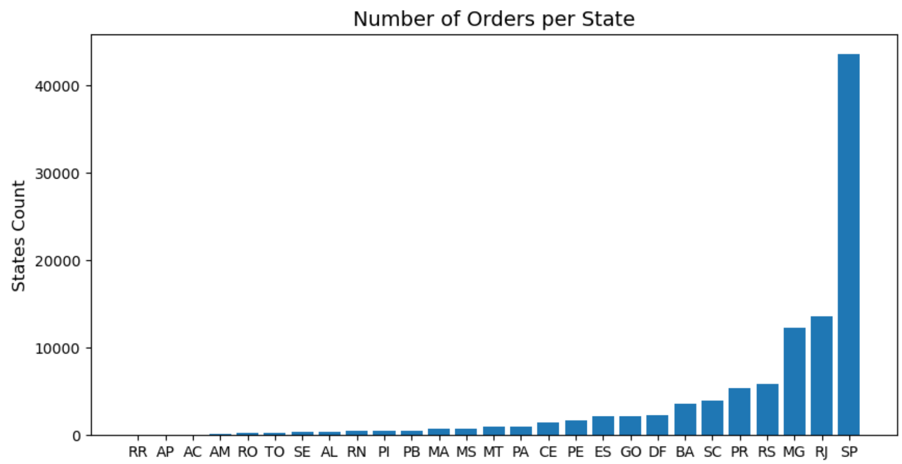
Most items that were purchased were from the Bed, Bath & Table category but the top category in the amount purchased was Health & Beauty. Most customers purchased items in the bedroom/bathroom area or that were used in the bedroom/bathroom area.





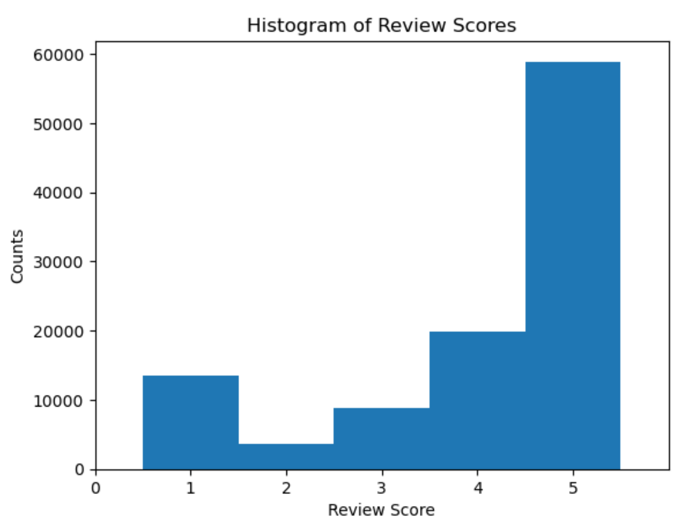
Looking at both the number of items purchased and the amount spent, the Security and Services and Children’s clothes categories are at the bottom. It could be that there are not many sellers of these items or more marketing needs to be allotted for these bottom categories.

States

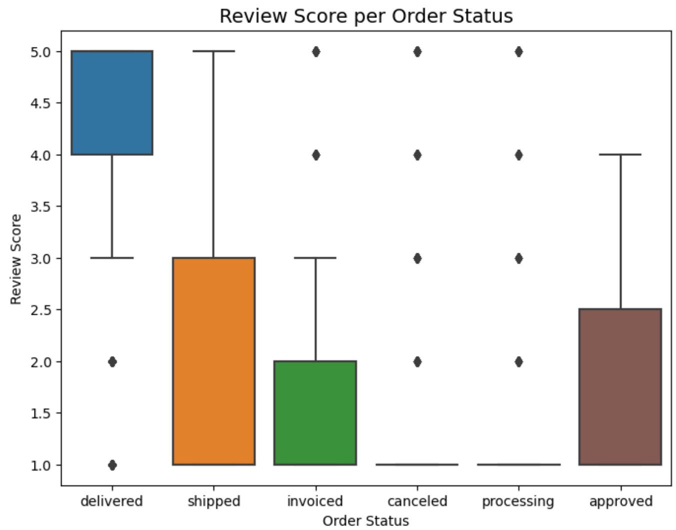


The top State was São Paulo by far compared to the other States in Brazil. Marketing may investigate advertising more in States other than Sao Paulo.

Review Score



The review score of 5 is by far the most used by Olist customers which is a good sign.



Looking at the review score based on the order status does show purchases that were delivered by far getting a review score of 5 and the other statuses, not surprisingly, were lower.

Customers

	customer_id	review_score	order_status	price	freight_value	product_category_name	cust_delivery_diff	est_delivery_diff
4201	d6646ea91d8cd9fc7e6882a7068779d4	5	delivered	81.99	14.51	computers_accessories	7	20
4350	679f84ceb2ee4ca5bca0c3ea34647746	5	delivered	59.90	17.67	garden_tools	20	4
13293	b4afeb58ac51bc903c5362286c6a5cfe	5	delivered	19.30	11.73	drinks	11	5
49414	10de381f8a8d23fff822753305f71cae	5	delivered	65.49	16.22	furniture_decor	19	5
54499	b7770073b02ed1d626a027ce86a4ff82	5	delivered	66.90	31.65	sports_leisure	10	44
67152	0d93f21f3e8543a9d0d8ece01561f5b2	5	delivered	20.70	16.11	housewares	8	8
67939	1ff773612ab8934db89fd5afa8afe506	5	delivered	284.99	16.87	drinks	14	18
77657	20c93357daf05d1c3a092be59aea2c2b	5	delivered	20.50	16.91	drinks	10	14
90355	0e772d9e02b17408e716f35cd1dcc222	5	delivered	36.99	11.85	bed_bath_table	10	13
96279	adb32467ecc74b53576d9d13a5a55891	5	delivered	51.00	1.20	garden_tools	14	20

This chart shows an example of customers that gave Olist a review score of 5 and how much they spent along with the difference in days from estimated delivery to actual delivery and the difference in days between estimated delivery and actual delivery.

	customer_id	review_score	order_status	price	freight_value	product_category_name	cust_delivery_diff	est_delivery_diff
6639	91f92cfee46b79581b05aa974dd57ce5	1	delivered	108.00	15.52	watches_gifts	11	13
19725	d5f2b3f597c7ccafbb5cac0bcc3d6024	1	delivered	59.00	13.43	garden_tools	14	10
24699	4a60b2ce1ee8c7b828e4bbcca5b86b41	1	delivered	137.90	38.81	computers_accessories	14	1
25887	be1c4e52bb71e0c54b11a26b8e8d59f2	1	delivered	49.99	7.10	bed_bath_table	5	11
37089	78fc46047c4a639e81ff65f0396e02fe	1	delivered	109.97	34.04	furniture_living_room	5	13
46421	be1b70680b9f9694d8c70f41fa3dc92b	1	delivered	100.00	10.12	computers_accessories	10	2
68674	cb87122c4871e202777cf243fbea2d12	1	delivered	149.91	0.14	computers_accessories	11	23
77621	a7693fba2ff9583c78751f2b66ecab9d	1	delivered	29.99	7.78	telephony	8	5
102470	fc3d1daec319d62d49bfb5e1f83123e9	1	delivered	1.20	7.89	health_beauty	14	-4
102724	9eb3d566e87289dcb0acf28e1407c839	1	delivered	5.31	15.23	housewares	10	9

This chart shows an example of customers that gave Olist a review score of 1.

Looking at the two charts, you can't distinguish between those that gave a high review score to those that gave a low review score.

Final Dataset for Model

I dropped all columns that were not needed for the Recency, Frequency, Monetary Analysis, and K-Means Clustering analysis.

The final dataset includes the following columns:

customer_id, order_id, order_purchase_timestamp, and the price.

	customer_id	order_purchase_timestamp	order_id	price
0	9ef432eb6251297304e76186b10a928d	2017-10-02 10:56:33	e481f51cbdc54678b7cc49136f2d6af7	29.99
1	a20e8105f23924cd00833fd87daa0831	2017-08-15 18:29:31	128e10d95713541c87cd1a2e48201934	29.99
2	26c7ac168e1433912a51b924fbd34d34	2017-08-02 18:24:47	0e7e841ddf8f8f2de2bad69267ecfbcf	29.99
3	53904ddbea91e1e92b2b3f1d09a7af86	2017-10-23 23:26:46	bfc39df4f36c3693ff3b63fcbea9e90a	29.99
4	b0830fb4747a6c6d20dea0b8c802d7ef	2018-07-24 20:41:37	53cdb2fc8bc7dce0b6741e2150273451	118.70
...	...	...	...	...
104777	609b9fb8cad4fe0c7b376f77c8ab76ad	2017-08-10 21:21:07	e8fd20068b9f7e6ec07068bb7537f781	356.00
104778	609b9fb8cad4fe0c7b376f77c8ab76ad	2017-08-10 21:21:07	e8fd20068b9f7e6ec07068bb7537f781	356.00
104779	a2f7428f0cafbc8e59f20e1444b67315	2017-12-20 09:52:41	cfa78b997e329a5295b4ee6972c02979	55.90
104780	39bd1228ee8140590ac3aca26f2dfe00	2017-03-09 09:54:05	9c5dedf39a927c1b2549525ed64a053c	72.00
104781	edb027a75a1449115f6b43211ae02a24	2018-03-08 20:57:30	66dea50a8b16d9b4dee7af250b4be1a5	68.50

104782 rows × 4 columns

Model

I chose to use Recency, Frequency, and Monetary Value (RFM) Analysis and K-mean clustering analysis to group customers for the company to better distinguish between the various customers and better serve them.

RFM Analysis looks at historical customer behavior to predict how might a new customer act in the future. It looks at three key items:

1. How recently a customer has purchased
2. How many orders did a customer purchase
3. How much money a customer has spent

Once I have the three key factors of recency, frequency, and monetary value, I will use the three groups in a k-means clustering model to better distinguish the customer segments. This will allow the company to target different groups of customers depending on how much they spend, how frequently they shop, and how many items they typically purchase.

Recency: To get recency, I found the most recent purchase date and calculated the number of days other customers had purchased from and compared it to this date. I did this because the data I was using was from 2017 and 2018.

	customer_id	last_purchase_date	Recency
0	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	259
1	000161a058600d5901f007fab4c27140	2017-07-16 09:40:32	380
2	0001fd6190edaaf884bcaf3d49edf079	2017-02-28 11:06:43	518
3	0002414f95344307404f0ace7a26f1d5	2017-08-16 13:09:20	349
4	000379cdec625522490c315e70c7a9fb	2018-04-02 13:42:17	120

Frequency: To get frequency, I totaled up the number of orders a customer had purchased.

	customer_id	Frequency
0	00012a2ce6f8dcda20d059ce98491703	1
1	000161a058600d5901f007fab4c27140	1
2	0001fd6190edaaf884bcaf3d49edf079	1
3	0002414f95344307404f0ace7a26f1d5	1
4	000379cdec625522490c315e70c7a9fb	1

Monetary: To get monetary, I totaled up the amount of each order ordered.

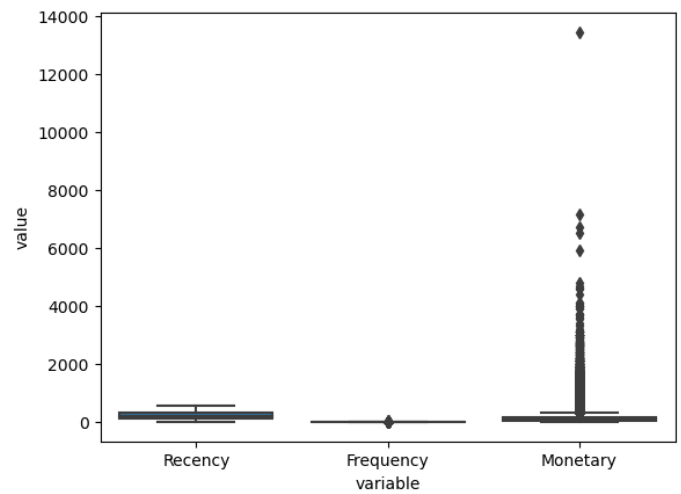
	customer_id	Monetary
0	00012a2ce6f8dcda20d059ce98491703	89.80
1	000161a058600d5901f007fab4c27140	54.90
2	0001fd6190edaaf884bcaf3d49edf079	179.99
3	0002414f95344307404f0ace7a26f1d5	149.90
4	000379cdec625522490c315e70c7a9fb	93.00

I then merged the recency, frequency, and monetary datasets into one to get a better understanding of the different groups.

	Recency	Frequency	Monetary
customer_id			
00012a2ce6f8dcda20d059ce98491703	259	1	89.80
000161a058600d5901f007fab4c27140	380	1	54.90
0001fd6190edaaf884bcaf3d49edf079	518	1	179.99
0002414f95344307404f0ace7a26f1d5	349	1	149.90
000379cdec625522490c315e70c7a9fb	120	1	93.00

Examining the RFM Analysis

	Recency	Frequency	Monetary
count	91182.000000	91182.000000	91182.000000
mean	225.794137	1.149152	138.491396
std	144.446134	0.554204	210.737977
min	0.000000	1.000000	0.850000
25%	107.000000	1.000000	45.950000
50%	204.000000	1.000000	87.990000
75%	331.000000	1.000000	149.990000
max	572.000000	22.000000	13440.000000



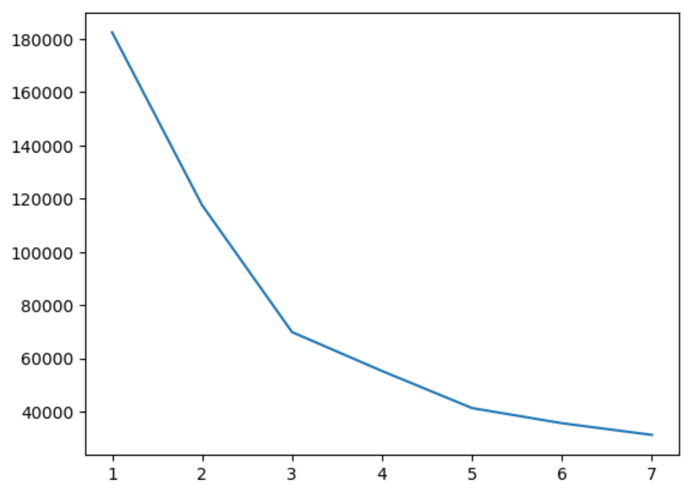
The average number of days customers recently purchased is 226 days, the customers mostly only purchased one time and the average amount a customer spent was R$138.50. Since most of the customers only purchased one time, I dropped the frequency column and only used the recency and monetary columns for the k-means clustering analysis.

I chose to use the k-means clustering method because it will allow me to identify groups of customers within the recency and monetary columns that have similar traits. This should help marketing or other business units better understand their customers and better serve them.

The k-means clustering method requires us to first determine the optimal number of clusters to group the customers into. To do this I will use the silhouette method and the elbow method. The silhouette score determines how similar an object is to its current cluster and the other clusters. Scores range from -1 to 1 where the higher value indicates that the object is similar to its own cluster. I chose to test out the silhouette method with clusters from 2 – 6 to see how the scores measured. Looking at the scores, clusters 3 and 4 were both high and looked promising.

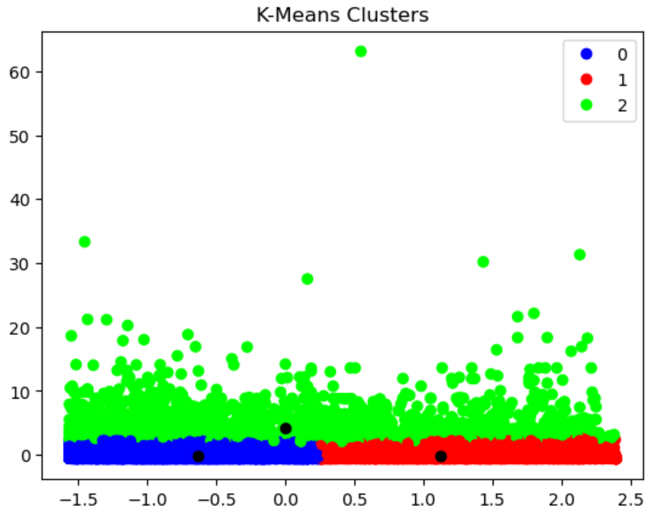
```
For n_clusters = 2. The average silhouette score is 0.46731442416207347
For n_clusters = 3. The average silhouette score is 0.5010325943820231
For n_clusters = 4. The average silhouette score is 0.5111099455792638
For n_clusters = 5. The average silhouette score is 0.4311208168994719
For n_clusters = 6. The average silhouette score is 0.43634171713433667
```

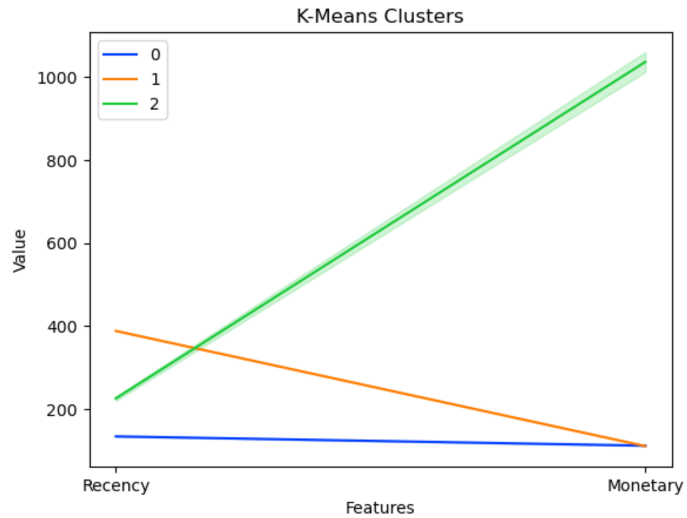
To verify the optimal number of clusters, I then used an elbow method to see what it determined the optimal number of clusters. Again, using clusters 2 – 7 this time the elbow method calculated the optimal number of clusters was 3.



Because the silhouette method had 3 very close to four, I determined that 3 would be the optimal number of clusters to use with the k-means clustering method.

The results of the k-means clustering method grouped the customers into the 3 optimal segments.





Looking at the two graphs, we can see the differences between the different clusters.

- **Group 1**: Customers that haven't purchased items in a long time and their purchases amount to a very low dollar amount.
- **Group 2**: Customers that have purchased recently but their purchases amount to a very low dollar amount
- **Group 3**: These are customers that have purchased items somewhat recently but these customers' purchases amount to a very high dollar amount.

Conclusion

With these 3 groups of customers, the marketing team will be able to target advertising, coupons, and new products to each group individually to help drive higher profits, customer satisfaction with high review scores, and better customer service. For example, to get the Group 1 customers back, marketing may target them with sales or coupons to draw them in and to purchase items. Group 2 could be targeted with advertising for items that go with their purchase of buy one get one sale. Finally, group 3 could be targeted with things like a frequent purchaser card hoping they will come back and purchase a lot more.

The current model deals with historical data from about 5 years ago with customers that have only purchased once. To create a deployment model, I think more data is needed for customers that purchased multiple times to get a better frequency feature to add to the k-means model. I think this will create more groups of customers with more specific traits to help marketing better serve the customers ultimately creating more sales and high review scores.

With this project, I have learned a lot about how to analyze retail customers based on how they purchase. Doing the explanatory data analysis showed me the different ways you can look at the data. After looking at customer segmentation, I would like to see about predicting review scores or even product sales. It was a very interesting experience and I'm glad I chose it.

References

Arvai, K. (n.d.). *K-Means Clustering in Python: A Practical Guide*. Retrieved from Real Python: https://realpython.com/k-means-clustering-python/

Begam, S. (2021). *Customer Profiling and Segmentation – An Analytical Approach To Business Strategy In Retail Banking*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/03/customer-profiling-and-segmentation-an-analytical-approach-to-business-strategy-in-retail-banking/

Kaggle. (n.d.). *Brazilian E-Commerce Public Dataset by Olist*. Retrieved from https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Pagan Research. (n.d.). Retrieved from olist: https://paganresearch.io/company/olist

Selvaraj, N. (2022). *How to Build Customer Segmentation Models in Python?* Retrieved from 365 Data Science: https://365datascience.com/tutorials/python-tutorials/build-customer-segmentation-models/

Wright, G. (n.d.). *RFM analysis (recency, frequency, monetary)*. Retrieved from TechTarget: https://www.techtarget.com/searchdatamanagement/definition/RFM-analysis