Customer Segmentation Using Clustering

Maryam Anvaripoor

Abstract

Customer segmentation gives useful business insights. We used different clustering algorithms such as k-means, DBSCAN and Agglomerative clustering to find meaningful clusters of customers. we divided customers into two and four groups based on evaluation metrics and business usefulness. We compared these groups in terms of their spending, review scores and other related metrics and provided actionable recommendations for each one of them.

Introduction

Customer segmentation is essential for businesses as it enables them to tailor marketing efforts and product offerings to specific customer groups, enhancing engagement and boosting conversion rates. By understanding distinct customer needs and preferences, companies can deliver personalized experiences that foster satisfaction and loyalty. It also helps identify high-value customers, allowing for smarter resource allocation and maximizing return on investment. Moreover, segmentation uncovers behavioral patterns that inform data-driven decisions in product development and service delivery.

In this analysis, first we used statistical tests to uncover patterns in data. We found that total spending of customers does not follow a normal distribution, customers with higher review score, spend different from those with less review score, customer's state effects customer's spending, customers with sequential payments spend less on average and there is a monotonic relationship between number of orders and customer's spending.

We used three clustering algorithms for our dataset and compared them based on evaluation metrics, visualization and business insight. We tried to find the best parameters for each model and applied them to both of initial dataset and projected dataset using PCA. Agglomerative clustering had the best Silhouette and Davies-Bouldin Score but did not provide useful business insights. K-means with k=2 (chosen by elbow method) and DBSCAN where more suitable. K-means with k=4 was added for more detailed insights.

The first two algorithms created two clusters of customers, higher spending minority and lower spending majority. For the first cluster we suggested loyalty rewards and personalized products and for second one, promotions and short-term installment options for affordability.

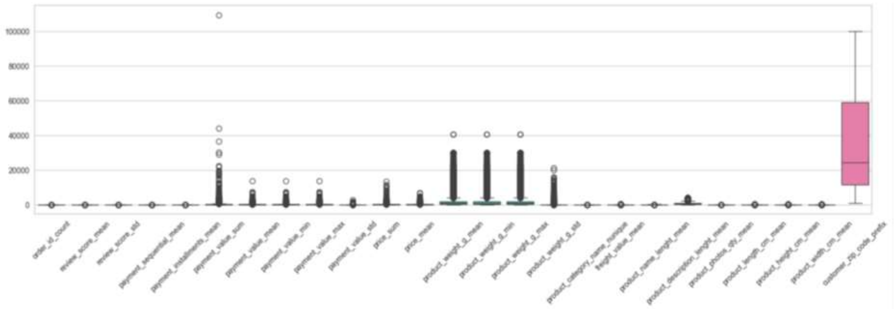
In case of k-means with k=4 we found these customer segments: High-Value Power Buyers, Mid-High Value Customers, Satisfied Budget Shoppers and Cautious Casual Buyers. We suggested VIP programs, enhancing personalization, good deals on essentials and gentle reactivation campaigns for each of these clusters respectively.

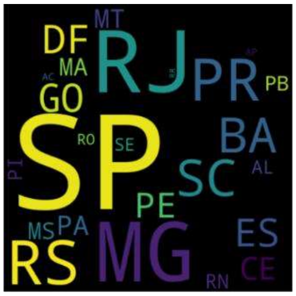
Data

This dataset is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. It also contains a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

Features and preprocessing

From order reviews dataset, we kept order id and review score since other columns had a lot of missing values. Then we merged order reviews, order payments and orders datasets based on order id. Also we merged order items and products based on product id and merged the final two datasets based on order id. We dropped duplicates. For aggregating numerical columns, we chose mean and for categorical columns we chose mode. For more important features like payment value, we added min, max, std and sum and for product category name, we added the number of unique categories. To aggregate order purchase timestamp, we used max, this indicates the last time a customer purchased something. Then we merged the new dataset with customers dataset based on customer id. We compared the distribution of categorical columns using box plot and found the common categories in categorical variables using pie and word cloud plot. By plotting heatmap for correlation between numerical variables we found high correlation between min, max and mean, so we dropped min and max columns. To handle 1379 missing values in product name, description and photos we used KNN Imputer and for 16 missing values in product length, height, width and weight we replaced them with mean of each these features. To encode order purchase timestamp, we replaced timestamps with number of days passed between that time and the max time available in the dataset. For order status and payment type, since these columns did not have so many categories, we used one-hot encoding and for customer city and customer state we used frequency encoding with normalizing. Finally, we used standard scaler since each feature had a different range of values.





Box plot for numerical features, we dropped the datapoint that created a large gap in payment value sum

Word cloud of customer state

Methods

K-Means:

K-Means Clustering partitions data into *k* clusters by minimizing intra-cluster variance. It starts with randomly placed centroids, assigns points to the nearest centroid, and updates centroids iteratively. It's efficient but requires specifying *k* and struggles with non-spherical clusters or noise.

DBSCAN:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups points that are closely packed and marks outliers as noise. It doesn't need the number of clusters beforehand and handles arbitrarily shaped clusters well. However, it struggles with varying densities and requires careful tuning of its parameters: epsilon and minPoints.

Agglomerative Clustering:

Agglomerative Clustering is a bottom-up hierarchical method that starts with each data point as its own cluster and merges the closest pairs step-by-step. It builds a dendrogram representing the merging process and doesn't need a fixed number of clusters at first. It's flexible but computationally intensive for large datasets.

Results and Discussion

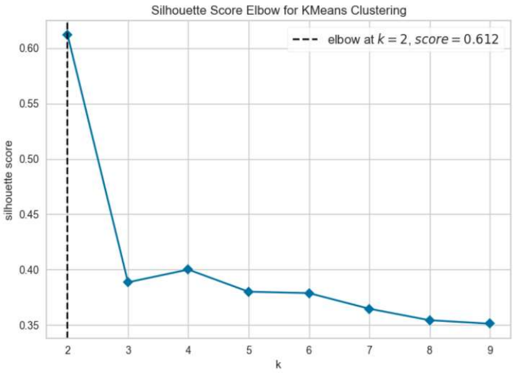
First we asked a few questions and used statistical tests to answer them.

n	question	test	answer
1	Does payment value sum follow a normal distribution?	Kolmogorov-Smirnov	Since p-value is 0 we reject the null hypothesis and conclude that it does not follow a normal distribution
2	Do customers with higher review score mean, spend different from those with less review score?	Mann Whitney U-test	p-value=0 confirms that the spending between these two groups is different.
3	Does customer's state effect customer's spending?	Kruskal-Wallis H-test	p-value=0 mean that it does
4	Do customers with sequential payments, spend higher on average?	Mann Whitney U-test	No they spend less
5	Is there a relationship between number of orders and payment value sum?	Spearman correlation coefficient	Yes there is

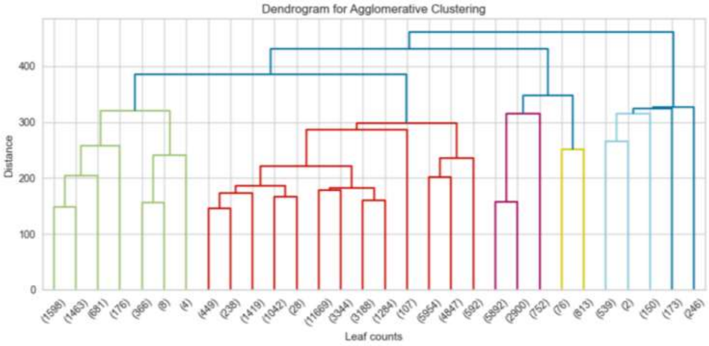
Then we applied clustering algorithms to our dataset. For each algorithm we calculated WCSS, Silhouette Score and Davies Bouldin Score. We tried to get better results for each algorithm by using elbow method for k-means, dendrogram for agglomerative clustering and experimenting with different parameters for DBSCAN. Also we applied clustering algorithms after using PCA and compared results. Summary of results are available in the next table.

algorithm	WCSS	WCSS (PCA)	Silhouette	Silhouette (PCA)	Davies Bouldin	Davies Bouldin (PCA)
k-means (k=2)	2783551	379130	0.3738	0.6116	2.1833	0.8508
k-means (k=4)	-	199056	-	0.4086	-	0.8251
DBSCAN	2995962	474344	-0.2771	0.8078	1.3033	0.7588
Agglomerative clustering	2790528	566628	0.9325	0.9257	0.0503	0.2686

Even though Agglomerative clustering has the best Silhouette and Davies-Bouldin Score, since it generates one big cluster (almost containing all datapoints) and a few tiny other clusters, it does not provide useful business insights. So we used k-means with k=2 and DBSCAN after PCA for our customer segmentation. Each of these algorithms created two clusters. Also we included k-means with k=4 on dataset after PCA for more business insights.



Elbow method for k-means on the dataset after applying PCA



Dendrogram for agglomerative clustering

Next we examine insights achieved from each clustering method. We are going compare mean of important features for each cluster.

1. K-means with k=2:

n	Cluster name	Number of customers	Number of orders	review score	payment installments	payment value sum	payment value mean	payment value std	product description length	product photos
1	high spending minority	7567	1.5321	3.8507	5.6088	894.7766	597.0090	7.8402	1056.249	2.4442
2	lower spending majority	80556	1.1622	4.1302	2.6575	140.3121	116.351	0.8714	769.028	2.2322

recommendations: To retain and grow cluster 1, the business should offer loyalty rewards, personalized product recommendations, and flexible installment plans that cater to their spending habits. Additionally, highlighting premium offerings with rich content and photos can further encourage their purchasing behavior. For cluster 2, the business should focus on scaling sales volume through promotions, while offering short-term installment options to support affordability. Optimizing product listings to be clear but concise, and targeting them with lower-cost products can encourage them to buy more.

2. DBSCAN (eps=0.7 and minpts=30):

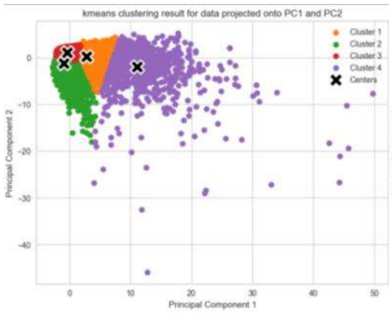
n	Cluster name	Number of customers	Number of orders	review score	payment installments	payment value sum	payment value mean	payment value std	product description length	product photos
1	premium gadget lovers	761	3.8975	3.7523	4.9181	2713.4632	1250.2662	47.9510	1056.249	1223.4234
2	lower spending majority	87,362	1.1705	4.1093	2.8934	183.2469	148.1072	1.0650	790.1061	2.2494

recommendations: the underlying reasons for this clustering is the same with the previous model. so same recommendations apply. However, we need to be more careful with cluster 1 since these customers are more concerned with their products (lower review scores). we should only offer them products with high quality.

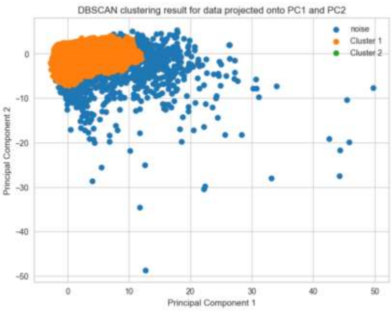
3. K-means with k=4:

n	Cluster name	Number of customers	Number of orders	review score	payment installments	payment value sum	payment value mean	payment value std	product description length	product photos
1	Mid-High Value Customers	11,728	1.3028	3.9226	5.1920	479.2393	360.8992	3.0658	953.7602	2.4070
2	Satisfied Budget Shoppers	36,541	1.2469	4.1648	2.1121	130.2057	99.2602	1.3973	741.9772	2.1594
3	Cautious Casual Buyers	38,572	1.0812	4.1163	2.8575	124.2534	112.4300	0.2600	772.1579	2.2725
4	High-Value Power Buyers	1282	2.0905	3.8128	6.4202	2264.1932	1321.4142	25.3386	1451.048	2.7544

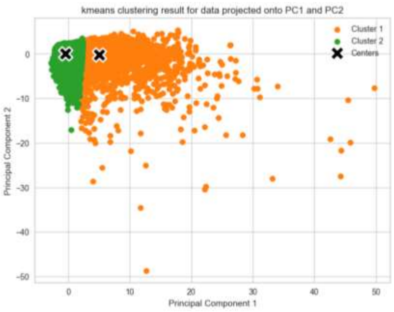
recommendations: Cluster 4 is a small but extremely valuable segment. The business should focus on retention through VIP programs, early access to products, premium support, and exclusive offers. Since their reviews are slightly lower, improving the post-purchase experience—like faster delivery or improved service follow-ups—could boost satisfaction. Cluster 1 shows a moderate number of orders and high average spend, with a good engagement level with product content (longer descriptions and more photos). Although their review score is slightly lower than other clusters, their spending potential is promising. The business should invest in enhanced personalization, targeted upselling, and informative product pages to nurture them further. Improving customer service and delivery tracking may help improve satisfaction (reflected in review scores). Cluster 2 customers are numerous, spend less, but leave high review scores, indicating strong satisfaction with smaller purchases. The focus here should be on volume-driven strategies: optimize for quick, low-cost transactions through simplified product listings, deals on essentials, and one-click purchases. This cluster is ideal for subscription models, flash sales, or automated reorder suggestions. Cluster 3 shows the lowest average number of orders and slightly less spending, with relatively low payment variation (low std), meaning their purchases are very consistent. Despite this, their review scores are high. They are potentially occasional but reliable buyers. The business should engage them with seasonal promotions, gentle reactivation campaigns, and reminders. Offering personalized discounts or highlighting products similar to their past purchases could encourage repeat buying.



k-means with k=4



DBSCAN with eps=0.7 and minpts=30



k-means with k=2

Conclusion

We found that total spending of customers does not follow a normal distribution, customers with higher review score, spend less from those with less review score, customer's state effects customer's spending, customers with sequential payments spend less on average and there is a relationship between number of orders and customers spending.

by using three clustering algorithms and comparing them based on evaluation metrics, visualization, business insights, finding best parameters for each model and applying them to both of initial dataset and projected dataset using PCA, we found that Agglomerative clustering had the best Silhouette and Davies-Bouldin Score but did not provide useful business insights, it formed one big cluster containing almost all datapoints and a few tiny clusters. K-means with k=2 (chosen by elbow method) and DBSCAN where more suitable, k-means with k=4 was added for more detailed insights.

The first two algorithms created two clusters of customers, higher spending minority and lower spending majority. For the first cluster we suggested loyalty rewards and personalized products and for second one, promotions and short-term installment options for affordability.

In case of k-means with k=4 we found these customer segments: High-Value Power Buyers, Mid-High Value Customers, Satisfied Budget Shoppers and Cautious Casual Buyers. We suggested retention through VIP programs, enhancing personalization to encourage buying more products, good deals on essentials and gentle reactivation campaigns for each of these clusters respectively.

This analysis comes with its limitations. It was found that removing categorical features results in better silhouette and Davis-Bouldin score, also adding interactions will improve results, trying different scaling methods might also land in better results. Another suggestion could be using other clustering algorithms and taking advantage of new informative features both for training the model and interpreting the results for useful business insights.