COLLEGE OF BUSINESS
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

MSC 692 - Business Analytics Practicum

Redstone Federal Credit Union

REDSTONE FEDERAL CREDIT UNION
Live Brighter!

Customer Segmentation Using Transactional Credit Card Data

Course Instructor

Dr. Hieu Pham

Sponsors

Raj Prasad
Chelsea Guo

Team Members

Manju Bhargavi Penumarthi
Aditya Gude
Prapul Nanjala
Vageesh Raghavendra

Contents

Introduction

Scope of work

The scope of this project is to develop an analytical model for Redstone Federal Credit Union (RFCU) that segments their customers into different categories based on their daily transaction data from 2022-01 to 2022-12. The model will provide insights and recommendations to inform RFCU's decisions related to credit line management, marketing, and fraud detection.

The team's responsibilities include:

1. Acquiring and preprocessing the transaction data from 2022-01 to 2022-12 at the account level.
2. Conduct a thorough analysis of the transaction data to identify similar transaction patterns based on transaction amount, frequency, merchant, location, and seasonality.
3. Implementing clustering algorithms in Python to segment customers into different categories.
4. Present the results and recommendations to the RFCU team in a clear and concise manner.
5. Assisting the RFCU team in the implementation of the recommended segments as needed.

Project Objectives

The main objective of this project is to segment Redstone Federal Credit Union's customers into different categories based on their transaction patterns. This will be achieved by conducting an analysis of the transaction data and identifying accounts with similar spending behavior. The data will be evaluated based on transaction amount, frequency, merchant, location, and seasonality.

Using Python programming, the team will implement various clustering algorithms to efficiently identify the different customer segments. These segments may include high-spending customers, low-frequency/high-value customers, seasonal spenders, channel-specific spenders, and category-specific spenders, among others.

The final outcome of this project will be a set of recommendations for RFCU on which customer segments to target for various financial purposes. These recommendations will be based on the results of the clustering analysis and will be presented to the RFCU team for their consideration and implementation

Project Progress

- The team held regular meetings with Chelsea & Raj for input and to show the current progress.
- The team extracted daily transaction data at the account level from 2022-01 to 2022-12 to allow for real-time predictions.
- The team cleaned and shaped the transaction data and documented the process for repeatability, and assisted in building the model.
- The team performed EDA (Exploratory Data Analysis) to yield insights into trends and summary statistics of the dataset provided.
- The team evaluated the importance of features by various methods such as statistical tests, feature selection techniques, or domain knowledge.
- The team worked on implementing different clustering algorithms to efficiently identify accounts with similar transaction patterns and follow a trial-and-error approach to find the best algorithm suitable for the data provided.
- The team segmented the customers into different categories and provide recommendations to the RFCU on credit line management, marketing, fraud detection, and so on.
- The team provided the Python code and all associated data with the sponsor in a way that may lead to real, impactful decisions and that can be maintained by RFCU going forward.
- The team provided a report of analysis methods and the most impactful factors involved in the recommendations.
- The team explored visualization methods and used them to view historical trends alongside real-time data.
- The team provided a midterm presentation to the sponsor and associated UAH professors, which will include a comprehensive summary of progress.
- The team delivered the final analytical model, the report of methods and results, and any other applicable materials to both the Sponsor and associated UAH professors.

- The team created a poster of results, predictions, and findings, and will participate in a UAH-sponsored poster session.

Data Summary

Initial Data

Our initial data consisted of 16 million rows of customer transactions made using RFCU cards. The features include:

NEW_MCID:	Unique Customer Identifier
RDT_MRCH_SIC_CODE:	Merchant Industry Code
RDT_TRANSACTION_AMOUNT:	Transaction Amount
TRX_DATE:	Transaction Date
RDT_TRANSACTION_CODE:	Transaction Code
RDT_CHD_EXT_STATUS:	External Authorization Status
RDT_CHD_INT_STATUS:	Internal Authorization Status
RDT_MERCHANT_CITY:	Merchant City
RDT_MERCHANT_NAME:	Merchant Name
RDT_MERCHANT_STATE:	Merchant State
BIN:	Bank Identification Number

In total, there were around 85k customers analyzed in our project.

Data Preparation

First, we preprocessed the 16 million customer transaction records that made up the data that was provided and then grouped the customers based on the unique customer identifier.

Steps:

Handling Missing Data: Checked the dataset for null values and appended proper values where possible.
The dataset consists of ~16 million rows and 11 columns. We replaced the missing values in the Merchant Name, Merchant City, and Merchant State columns with the most frequent value (**mode**).

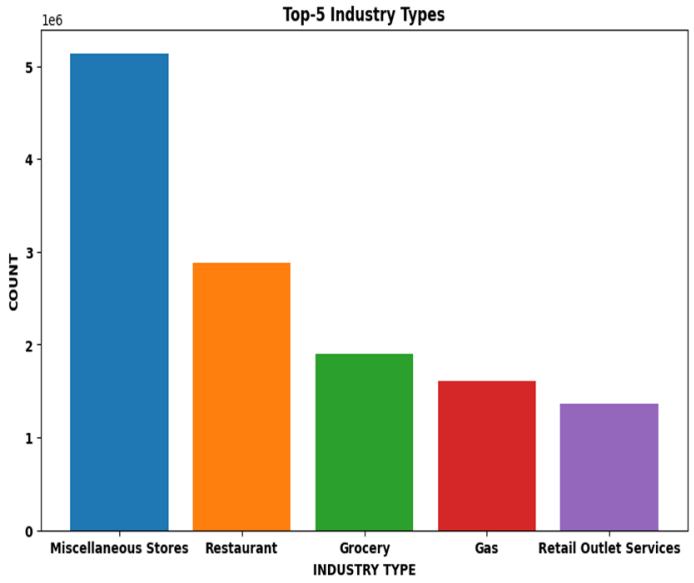
Feature	Number Missing	Percentage Missing
NEW_MCID	0	0
TRX_DATE	0	0
RDT_MRCH_SIC_CODE	0	0
RDT_TRANSACTION_AMOUNT	0	0
RDT_TRANSACTION_CODE	0	0
RDT_CHD_EXT_STATUS	0	0
RDT_CHD_INT_STATUS	0	0
RDT_MERCHANT_CITY	3816	0.02
RDT_MERCHANT_NAME	6	0
RDT_MERCHANT_STATE	7290	0.05
BIN	0	0

Feature Engineering

Selected and transformed the relevant features in the dataset to improve the performance of a machine learning algorithm.

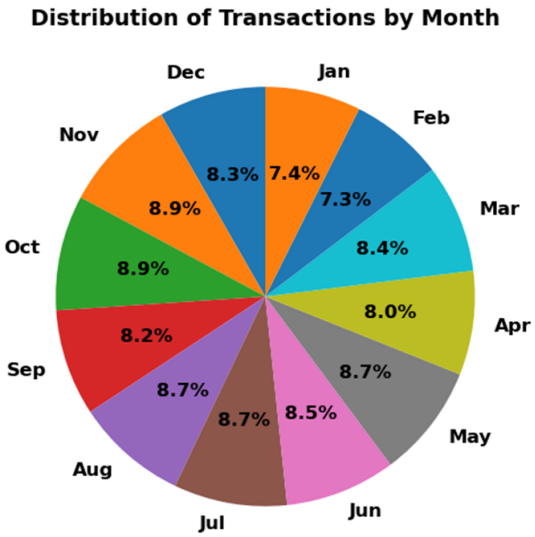
1. Industry Types:

- **SIC Codes:** The SIC codes represent the merchant category. There are a total of 537 unique SIC codes in the dataset provided.
- **Converting SIC Codes to Industry Types:** Extracted industry names from the SIC codes to identify the type of products purchased by each customer.



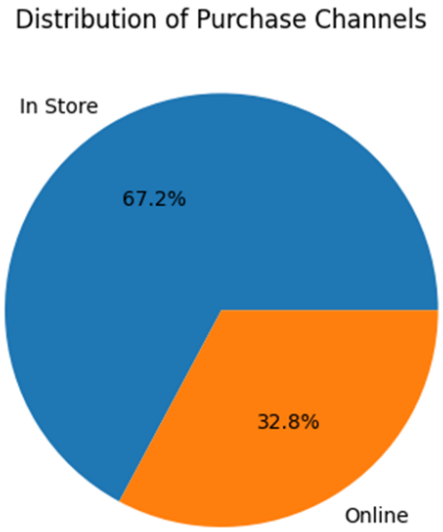
2. Latest_Month:

- We have added a new column to the dataset called '*Latest_Month*', which represents the month of the most recent transaction made by each customer.
- This additional information allows us to identify **seasonal customers** who may have a higher frequency of transactions during certain times of the year.
- By analyzing the patterns of these seasonal customers, we can gain insights into their behavior and preferences and use this information to inform our marketing strategies and product offerings.



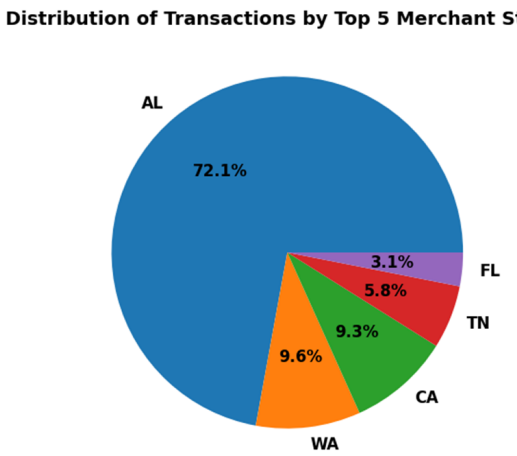
3. Purchase Channel, Percent-Online Columns:

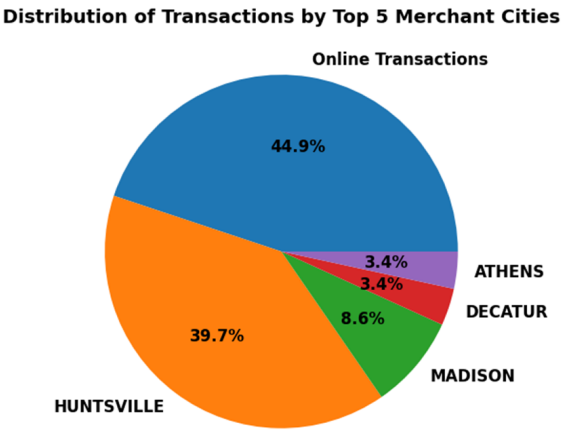
- **City Column:** The city column contains around 5M values which include website links or app information from which customers are buying the products.
- **Purchase Channel Column:** We replaced these unknown city values with online transactions. Using this info, we created a new column called Purchase Channel, which tells us if the transaction is made online or in-store.



- **Percent Online Column:** Aggregated the original dataset by unique customer ID, calculated the percentage of online transactions for each customer, and stored it in a new column called '*Percent_Online*'.

Top 5 Merchant States and Cities





4. New Features:

We extracted the most informative features from the raw data and created a new set of features that can be used to train a machine-learning model. The features are as follows:

- ***Amount_Spent*** - Total amount spent by a customer (Purchases – Returns)
- ***Purchase_Frequency*** - How often a customer makes a purchase
- ***Max_Payment_Amount*** - Highest amount paid by a customer
- ***Min_Payment_Amount*** - The smallest amount paid by a customer
- ***Total_Payment_Amount*** - Sum of all payments made by a customer
- ***Payment_Frequency*** - How often a customer makes a payment
- ***Max_CashAdvance*** - The highest amount of cash advance by a customer
- ***Min_CashAdvance*** - The smallest amount of cash advance by a customer
- ***Total_CashAdvance*** - Sum of all cash advances made by a customer
- ***CashAdvance_Frequency*** - How often a customer makes a cash advance
- ***Return_Frequency*** - How often a customer returns

Data Encoding:

Encoding is the process of converting categorical data into a numerical format that can be easily understood by machine learning algorithms.

- **Frequency Encoding:** It is a technique used in feature engineering to transform categorical variables into numerical variables. It involves replacing the categories in a categorical variable with the frequency of their occurrence in the dataset. We applied the frequency encoding to the "***Merchant State***" and "***Merchant City***" columns in our dataset.
- **One-hot encoding:** It is a technique used in feature engineering to transform categorical variables into numerical variables that can be used in machine learning algorithms. It involves creating a binary vector for each category in the categorical variable, where only one element of the vector is. We applied the one-hot encoding to the "***Industry Types***" columns in our dataset.

Feature Selection: Correlation-Based Method

We used correlation-based feature selection to choose only one feature from each pair of highly correlated features, reducing redundancy in the data.

This method involves iterating over each feature, checking its correlation with all previously selected features, and adding it to the list of selected features if it meets the threshold.

The final selected final features are as follows:

Merchant_City
Merchant_State
Frequent_Month
Purchase_Frequency
Max_Payment_Amount
Min_CashAdvance
Total_CashAdvance
Industry_Type_Agricultural Services
Industry_Type_Business Services
Industry_Type_Clothing Stores
Industry_Type_Government Services
Industry_Type_Miscellaneous Stores
Industry_Type_Transportation Services
Industry_Type_Retail Outlet Services
Industry_Type_Professional Services and Membership Organizations
Return_Frequency
Percent_Online
Amount_Spent
Min_Payment_Amount
Payment_Frequency
Max_CashAdvance
CashAdvance_Frequency
Industry_Type_Airlines
Industry_Type_Car Rental
Industry_Type_Contracted Services
Industry_Type_Lodging
Industry_Type_Other
Industry_Type_Utility Services

Data Scaling:

Data scaling is an important preprocessing step in many machine learning applications. It involves transforming the data to a common scale or range so that features with different units or scales can be compared on the same level. It can also help to improve the performance of machine learning models, especially those that are sensitive to the magnitude of the input features.

We applied a Min-Max scaler to our dataset, which scales the data to a fixed range, usually between 0 and 1. It involves subtracting the minimum value of each feature and dividing it by the range (i.e., the difference between the maximum and minimum values).

Data Modeling:

We applied four different clustering algorithms to identify a similar group of customers.

K-means Clustering:

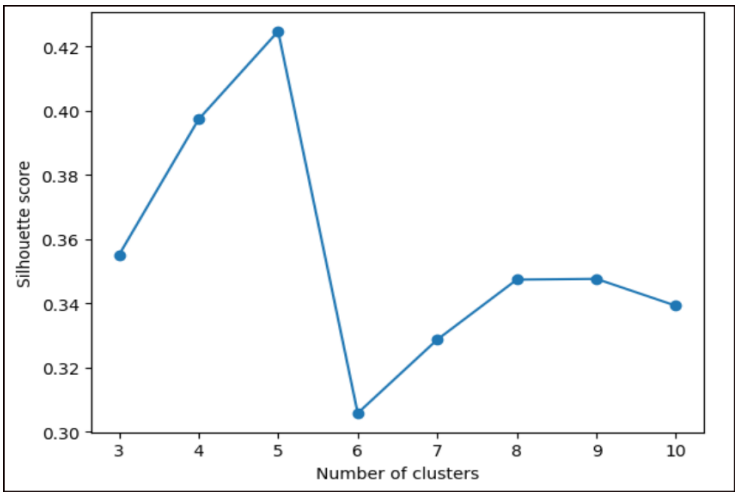
K-means clustering is a popular unsupervised machine learning algorithm used to partition a given dataset into K clusters or groups, based on the similarity of the data points within each cluster. The goal of the algorithm is to minimize the sum of squared distances between the data points and their assigned cluster centroids.

The algorithm works as follows:

1. Choose the number of clusters, K, that you want to partition the dataset into.
2. Randomly initialize K cluster centroids.
3. Assign each data point to the cluster whose centroid is closest to it.
4. Recalculate the centroids of each cluster as the mean of the data points assigned to it.
5. Repeat steps 3 and 4 until the algorithm converges, i.e., the cluster assignments no longer change.
6. Finally, the algorithm returns the K cluster centroids and the assignment of each data point to a particular cluster.

Silhouette Score to find the best number of clusters in K-means: After testing potential K values between 3 and 10; the higher silhouette score is shown at K=5. So, we chose 5 as the best value for K in this application in the K-means algorithm.

K	Silhouette Score
3	0.3551
4	0.3973
5	0.4247
6	0.3058
7	0.3229
8	0.3474
9	0.3481
10	0.3402



Mean Shift Clustering:

Mean shift clustering is another popular unsupervised machine learning algorithm used for clustering similar data points together. It works by finding the mode or the maximum density region of a probability density function, which represents the data distribution of the input data points.

The algorithm works as follows:

1. Initialize each data point as a cluster centroid.
2. For each data point, compute the mean shift vector by taking the weighted average of the difference between the data point and its neighboring points, where the weights are given by a kernel function (such as Gaussian kernel).
3. Update each data point to be the point to which the mean shift vector points, which is usually towards the maximum density region of the probability density function.
4. Merge clusters that are close to each other by setting their centroids to be the average of the data points in both clusters.
5. Repeat steps 2-4 until the centroids no longer move or the algorithm converges.
6. Finally, the algorithm returns the cluster assignments of each data point.

Gaussian Mixture Clustering:

Gaussian Mixture Clustering is a popular unsupervised machine learning algorithm used to model the probability distribution of a dataset using a mixture of Gaussian distributions. It works by estimating the parameters of the Gaussian distributions that best fit the data and then assigning each data point to the Gaussian distribution with the highest probability.

The algorithm works as follows:

1. Initialize the parameters of the Gaussian mixture model, such as the number of components, mean, and covariance matrix of each component.
2. E-Step: Assign each data point to a component by computing the posterior probability of the data point belonging to each component using Bayes' rule and the current model parameters.
3. M-Step: Update the model parameters, such as the mean and covariance matrix of each component, using the data points assigned to that component.
4. Repeat steps 2 and 3 until the algorithm converges or a stopping criterion is met.
5. Finally, the algorithm returns the model parameters and the cluster assignments of each data point.

BIRCH Clustering:

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering algorithm used to cluster large datasets. It was designed to be memory-efficient and able to handle large datasets without having to store all data points in memory. Instead, BIRCH builds a tree-based data structure, known as a Clustering Feature Tree (CFT), which allows for the efficient clustering of large datasets.

The algorithm works as follows:

1. Initialize the CFT with a user-defined threshold value for the maximum number of data points that can be stored in each leaf node.
2. For each data point, insert it into the CFT by updating the corresponding leaf node's parameters, such as the centroid and the number of data points in the node.
3. If a leaf node's number of data points exceeds the threshold value, split the node into two new child nodes and redistribute the data points between them.
4. Merge nodes that are close to each other and satisfy a user-defined merging criterion.
5. Repeat steps 2-4 until all data points are inserted into the CFT.
6. Finally, traverse the CFT to generate the hierarchical clusters.

Evaluation Metrics:

Silhouette Score: It measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, where 1 indicates a well-clustered sample, 0 indicates overlapping clusters and negative values indicate misclassified samples.

Davies-Bouldin Index: It measures the average similarity between each cluster and its most similar cluster, where lower values indicate better clustering performance.

Calinski-Harabasz Index: It measures the ratio of the between-cluster dispersion and within-cluster dispersion, where higher values indicate better clustering performance.

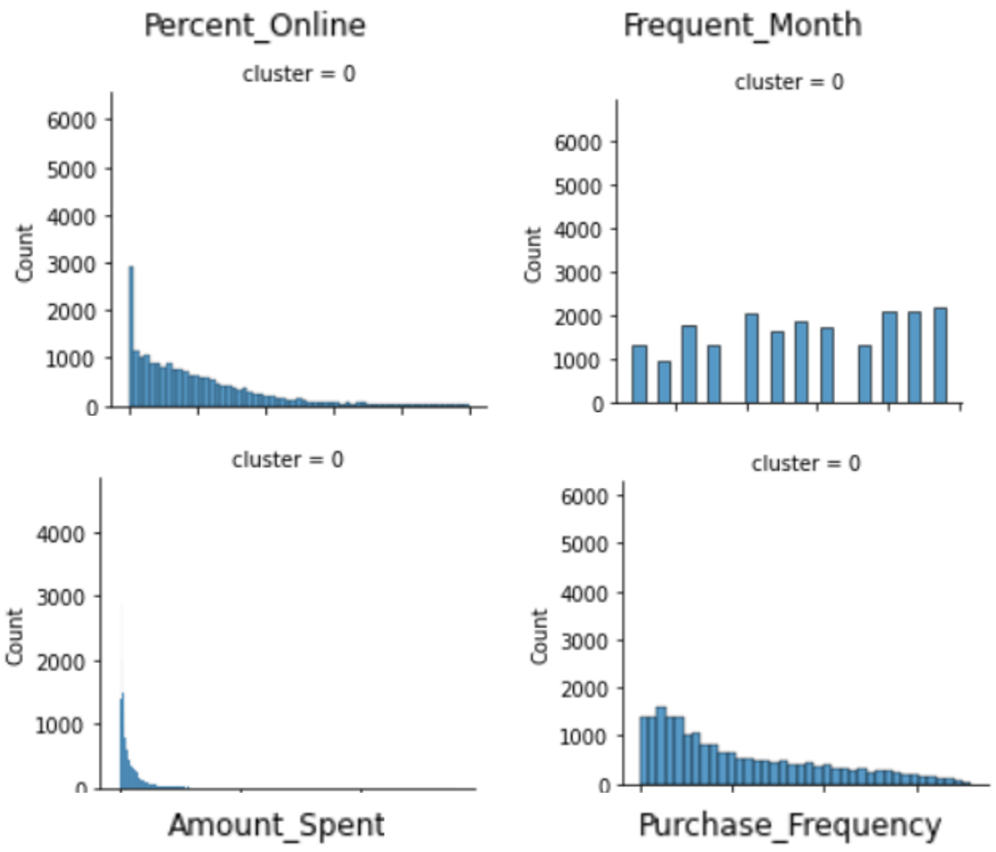
	Silhouette Score	**Davies-Bouldin Index**	**Calinski-Harabasz Index**
K-means	0.42	1.34	28844.16
MeanShift	0.31	0.82	1496.27
Gaussian Mixture	0.38	1.95	24205.68
BIRCH	0.45	1.37	9403.77

From the above results we can say that K-means performed well in clustering the customers. Below are the K-means segmentation results followed by recommendations for the RFCU team.

K-means Segmentation Results

Segment 1:

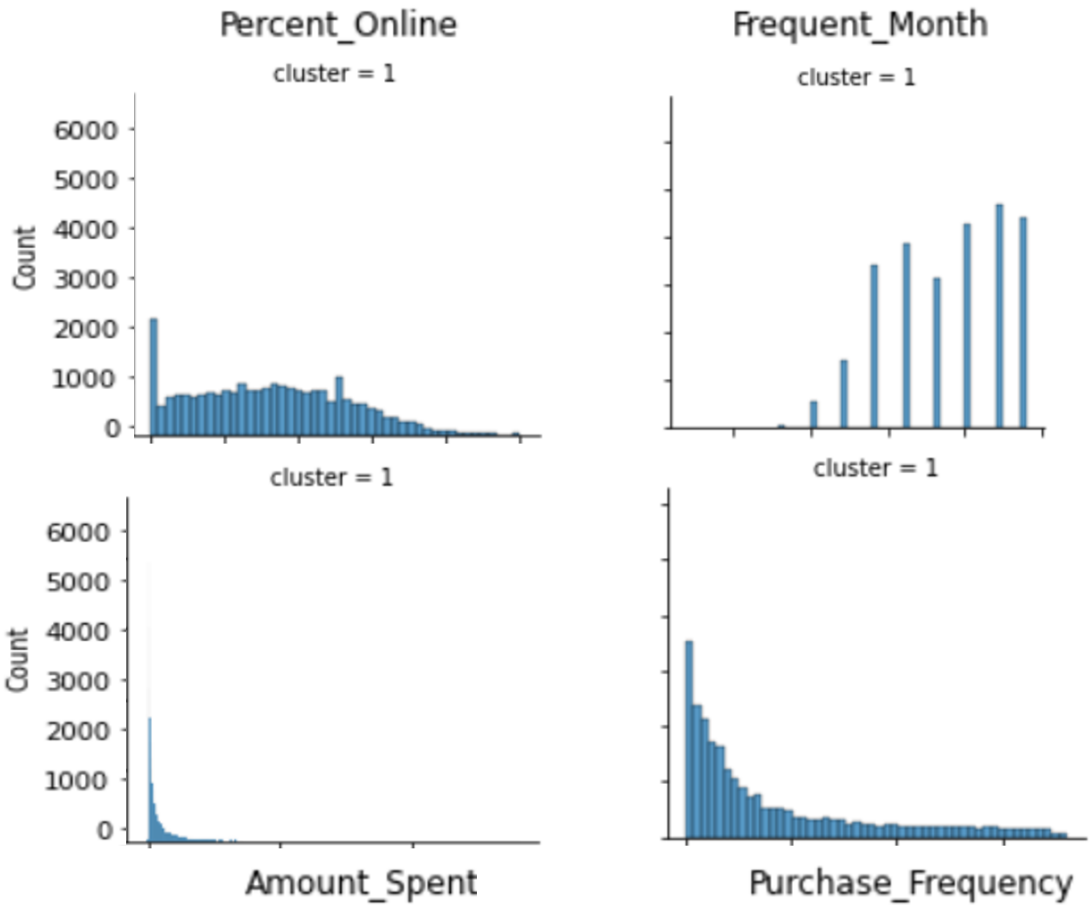
Column	Mean	Median	St Deviation	Min	Max
Percent_Online	18.46	13.69	18.17	0.0	100
Frequent_Month	7.03	7.0	3.40	1	12
Amount_Spent	13453.23	6759.27	18376.37	-685.99	561118.59
Purchase_Frequency	105.77	77.0	88.81	0.0	360



This segment includes customers who have a high spending level and a high purchase frequency. They make transactions throughout the year and do not show any seasonal purchasing behavior. They are making lesser online transactions which means they are channel specific and prefer to purchase in-store.

Segment 2:

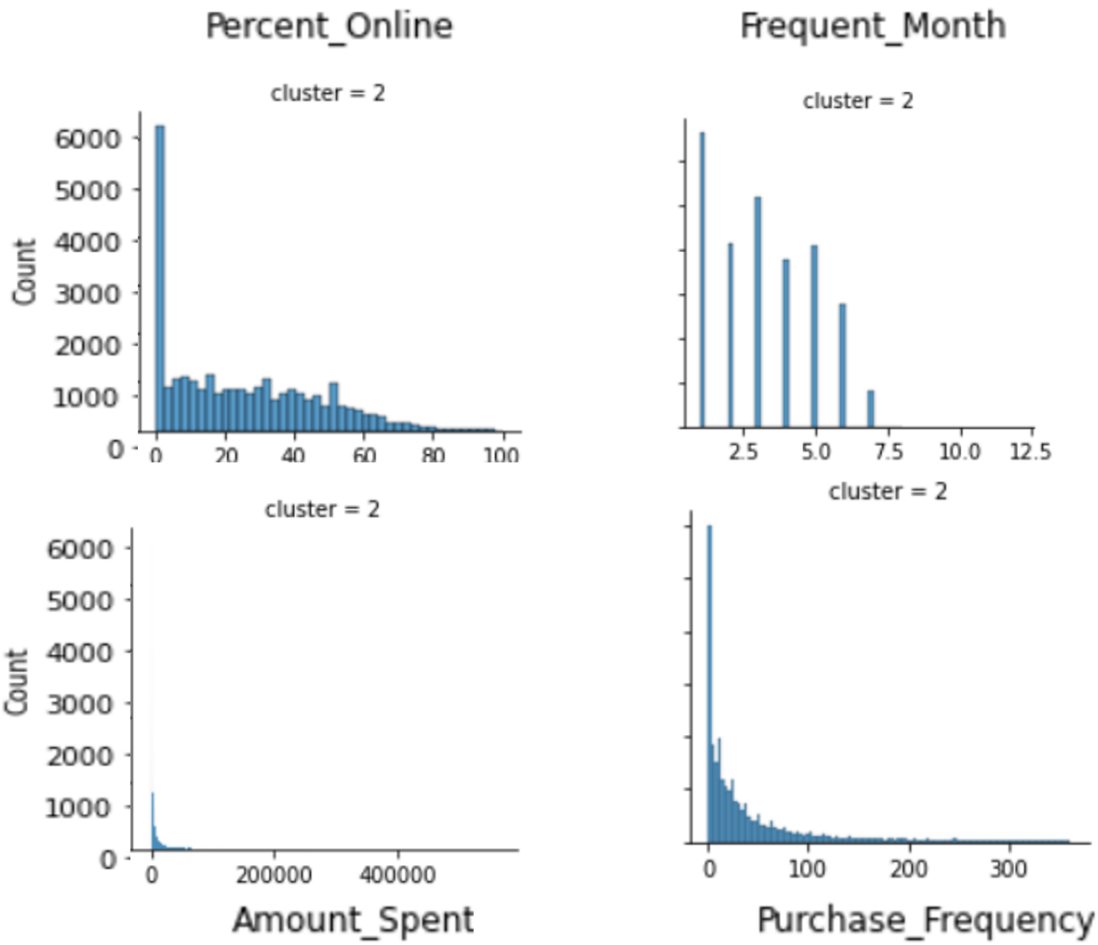
Column	Mean	Median	St Deviation	Min	Max
Percent_Online	31.81	31.65	20.51	0.0	100
Frequent_Month	9.37	10.0	1.96	4	12
Amount_Spent	12377.11	4833.0.	19422.74	-1046.80	565067.21
Purchase_Frequency	84.02	45.0	91.91	0	360



These are customers with high spending levels who make a medium percentage of online transactions and exhibit seasonal purchasing behavior, with more frequent purchases from June to December.

Segment 3:

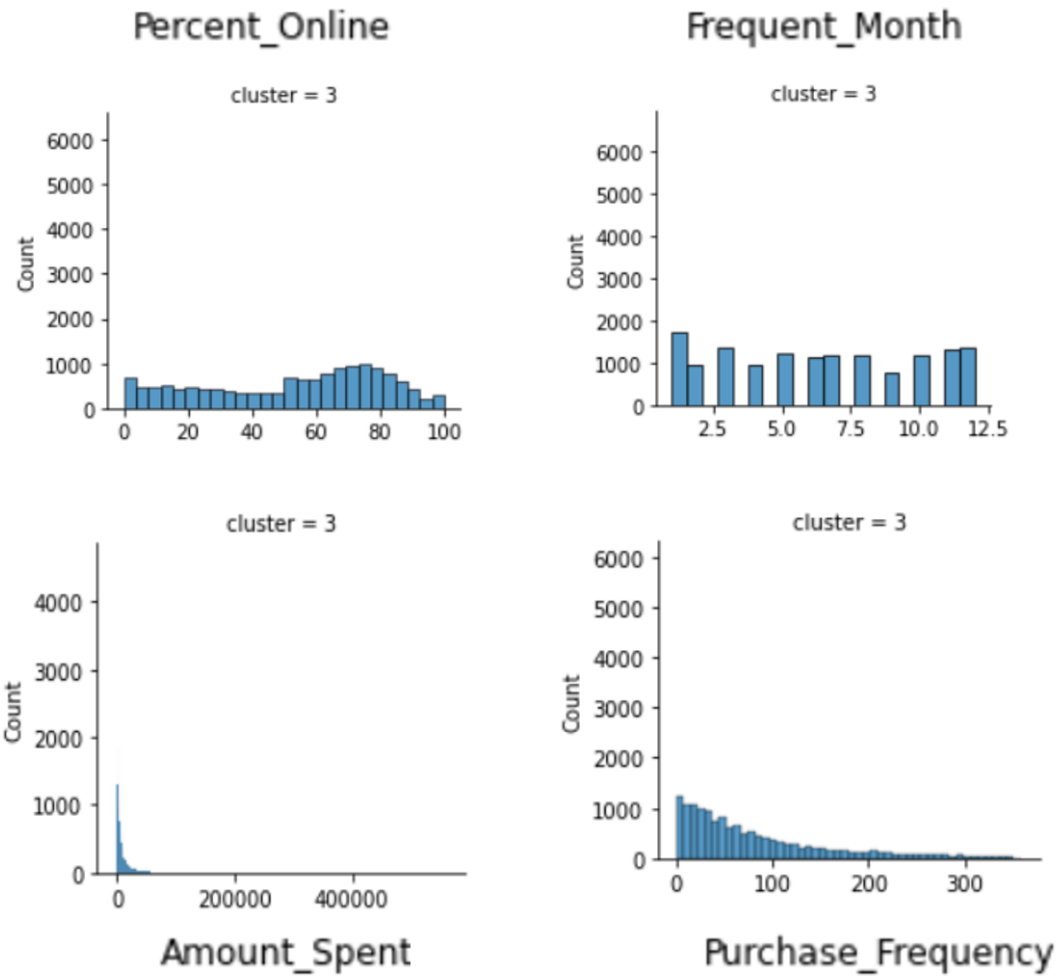
Column	Mean	Median	St Deviation	Min	Max
Percent_Online	24.11	20.51	21.20	0	100
Frequent_Month	3.22	3	1.78	1	8
Amount_Spent	5469.35	1692.72	10739.64	-1627.73	254796.88
Purchase_Frequency	43.88	20	62.57	0	361



These are customers with low spending levels who make a medium percentage of online transactions and exhibit seasonal purchasing behavior, with more frequent purchases from January to June.

Segment 4:

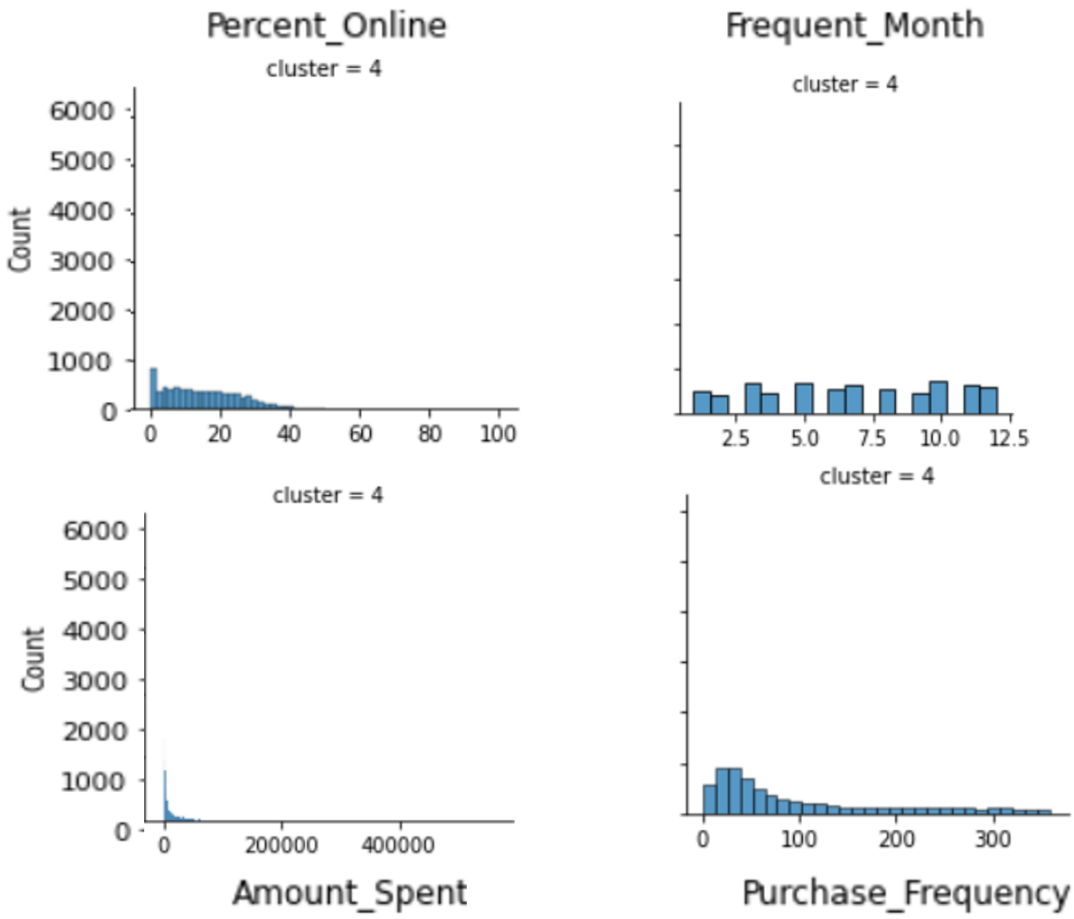
Column	Mean	Median	St Deviation	Min	Max
Percent_Online	52.80	59.59	27.86	0	100
Frequent_Month	6.39	6	3.61	1	12
Amount_Spent	9259.21	4268.27	14303.05	-6063.16	411426.88
Purchase_Frequency	80.87	56	76.18	0	359



This segment includes customers who have a medium spending level and a medium purchase frequency and who make a high percentage of online transactions making them channel-specific customers. They make transactions throughout the year and do not show any seasonal purchasing behavior.

Segment 5:

Column	Mean	Median	St Deviation	Min	Max
Percent_Online	14.47	13.33	10.25	0	50
Frequent_Month	6.73	7	3.40	1	12
Amount_Spent	11417.57	4093.56	17668.75	-1752.58	205941.77
Purchase_Frequency	101.55	60	96.50	0	359



This segment includes customers who have a medium spending level and a high purchase frequency. They make transactions throughout the year and do not show any seasonal purchasing behavior. They are less likely to make purchases online, indicating that they prefer to shop in-store.

Recommendations

Segment 1 - High Spending, High Frequency, Low Online Transactions
One effective way to encourage repeat purchases is by offering loyalty programs and in-store rewards. Assigning an executive to high-value customers for exclusive deals can also boost retention. Additionally, personalized promotions through targeted email or direct mail campaigns can encourage online purchases.

Segment 2 - High Spending, Seasonal, Average No. of Online Transactions
To drive sales during peak seasons, optimize the online shopping experience by promoting seasonal products and offering targeted promotions. Additionally, encourage repeat purchases during the off-season with personalized recommendations based on past purchase history.

Segment 3 - Low Spending, Low Frequency, Seasonal Customers, Average No. of Online Transactions
To attract cost-conscious shoppers, offer low-cost, high-value items and consider offering BuyNow PayLater or No-cost APRs to increase their purchasing power. Incentivize purchases by offering exclusive seasonal or limited-time products.

Segment 4 - Medium Spending, Medium Frequency, High Online Transactions
Improving the online shopping experience with easy product search and checkout processes, while offering personalized upgrade and exchange options, can increase customer spending frequency.

Segment 5 - Medium Spending, High Frequency, Low Online Transactions
Tailor in-store experiences to attract and retain customers by offering personalized promotions and exchange deals based on their usage. Encourage repeat purchases and boost online transactions by allowing customers to redeem these offers online.