https://www.kaggle.com/datasets/arjunbhasin2013/ccdata

Introduction of problem

Credit card users have different types of spending habits, such as making high one-time purchases, opting for installment-based buying, or relying on cash advances. It is important for credit card companies to understand their users on how credit cards are being used. This is so the companies can tailor their offerings, manage risks, personalize rewards programs, and create marketing strategies to meet the needs of different customers.

This project aims to explore the question: “Are there different user groups who prefer high one-time purchases, or installment-based buying, or rely on cash advances?”

What is clustering and how does it work?

Clustering is an unsupervised machine learning technique used to group similar data points into distinct clusters based on their features or characteristics. The data points in a cluster are more similar to each other than to those in other clusters. Clustering can help find underlying patterns and structures which are useful for data analysis.

Some clustering techniques:

- K-means clustering - Categorizes data points into clusters by using a mathematical distance measure, usually euclidean, from the cluster center. The goal is to minimize the sum of distances between data points and their assigned clusters. Data points that are closest to the centroid are grouped together within the same category. A higher k value means smaller clusters with greater detail, while a lower k value means larger clusters with less detail.

- Hierarchical clustering - Builds clusters by measuring the dissimilarities between data. It can be an agglomerative or bottom-up approach that repeatedly merges clusters into larger ones until a single cluster emerges. Or a divisive or top-down approach that starts with all data in a single cluster and continues to split out successive clusters until all clusters are singletons.

Introduce the data

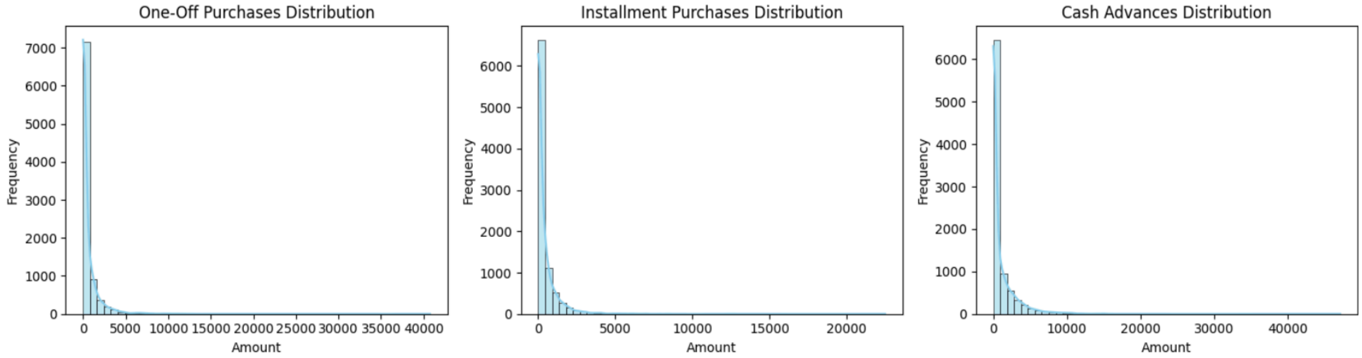
The data is sourced from Kaggle called "Credit Card Dataset" which contains 8950 rows and 18 columns. It contains information about the usage behavior of active credit card holders over a six month period.

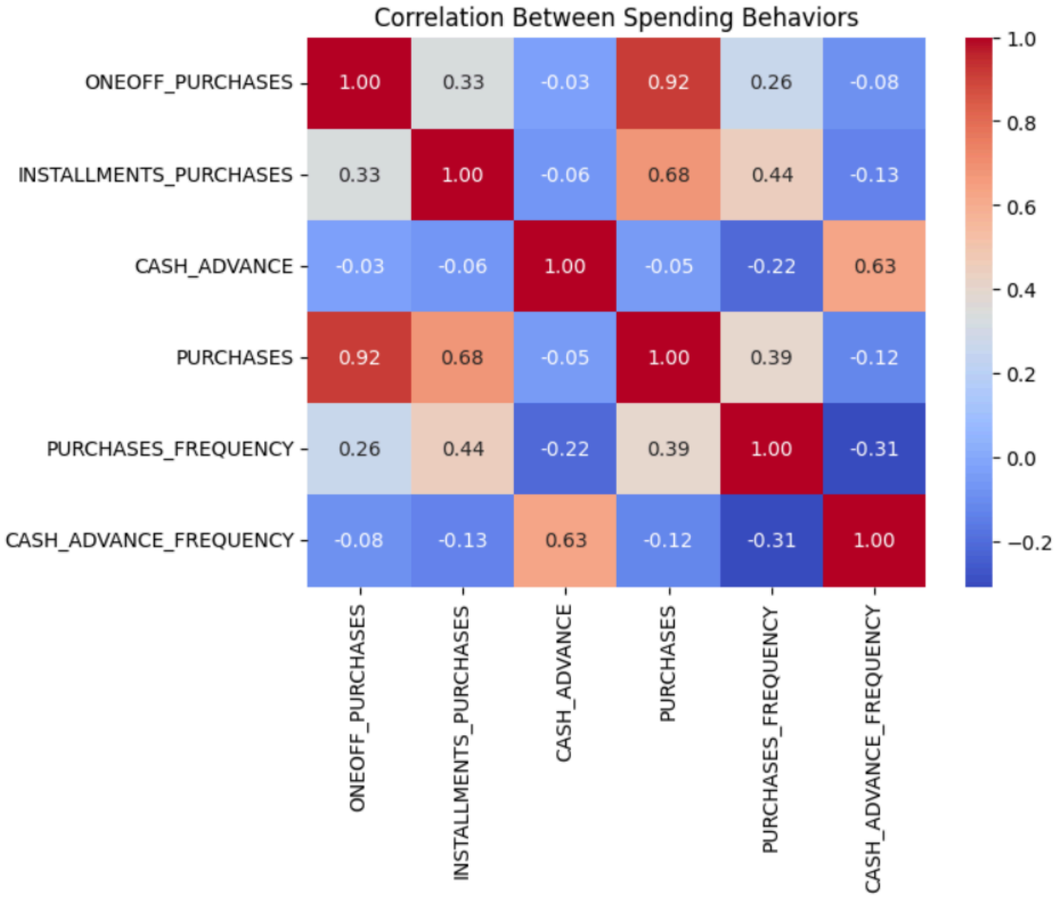
Columns:

- CUST_ID : Identification of Credit Card holder (Categorical)
- BALANCE : Balance amount left in their account to make purchases
- BALANCE_FREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- PURCHASES : Amount of purchases made from account
- ONEOFF_PURCHASES : Maximum purchase amount done in one-go
- INSTALLMENTS_PURCHASES : Amount of purchase done in installment
- CASH_ADVANCE : Cash in advance given by the user
- PURCHASES_FREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
- ONE OFF PURCHASE FREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)

- PURCHASE INSTALLMENTS FREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- CASH ADVANCE FREQUENCY : How frequently the cash in advance being paid
- CASH ADVANCE TRX : Number of Transactions made with "Cash in Advance"
- PURCHASES_TRX : Number of purchase transactions made
- CREDIT_LIMIT : Limit of Credit Card for user
- PAYMENTS : Amount of Payment done by user
- MINIMUM_PAYMENTS : Minimum amount of payments made by user
- PRC FULL PAYMENT : Percent of full payment paid by user
- TENURE : Tenure of credit card service for user

Data Understanding/Visualization





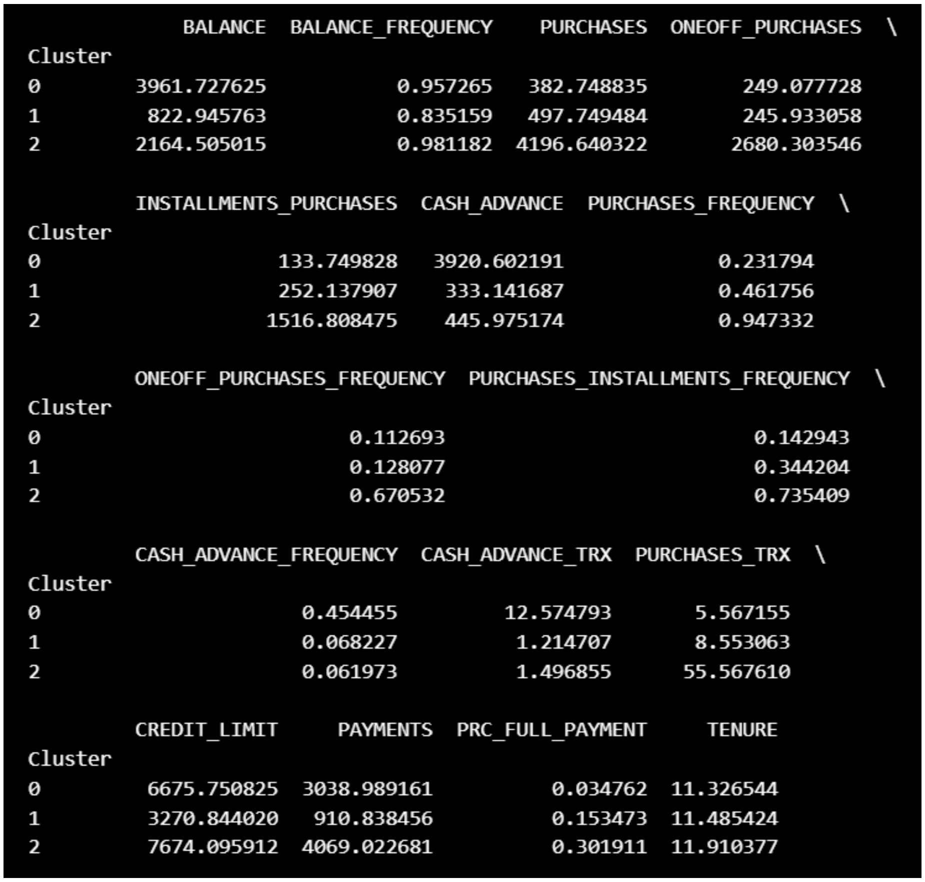
To better understand the dataset, distribution plots of one-off purchases, installment purchases, and cash advances were created. The one-off purchases plot shows most transactions are low in amount, also notice that there is a right skew indicating a few users make very large one time purchases. The installment purchases plot shows a right skew like one-off purchases. A majority of customers make small installment purchases, with some making significantly larger ones. The cash advances plot is similar to the other two plots showing high small advances, but a long tail of large cash advances. Overall all three plots are right skewed showing that the majority of credit card users spend smaller amounts, while a few have higher value transactions.

A heatmap was also created to better understand the dataset. The headmap shows a strong positive correlation between one-off purchases and total purchases, as well as between installment purchases and total purchases. Cash advances and total advance amounts also have a strong correlation. There is moderate correlation between installment purchases and purchase frequency, as well as purchase and purchase frequency. Cash advance has a negative correlation with one-off purchases and installment purchases.

Pre-processing

To prepare the dataset for clustering, several preprocessing steps were taken to clean and standardize the data. First, the dataset was checked for duplicate values, but none were found. Then the dataset was checked for null values, where it was seen that CREDIT_LIMIT had 1 null value and MINIMUM_PAYMENTS had 313 null values. To handle this, the row with the missing CREDIT_LIMIT value was removed because it wouldn’t make a drastic difference in the dataset, and the MINIMUM_PAYMENTS column was dropped because it was not essential for the research. The column CUST_ID was also removed because it’s an identifier and not essential for the research. Finally the dataset was standardized so the features are on the same scale.

Modeling



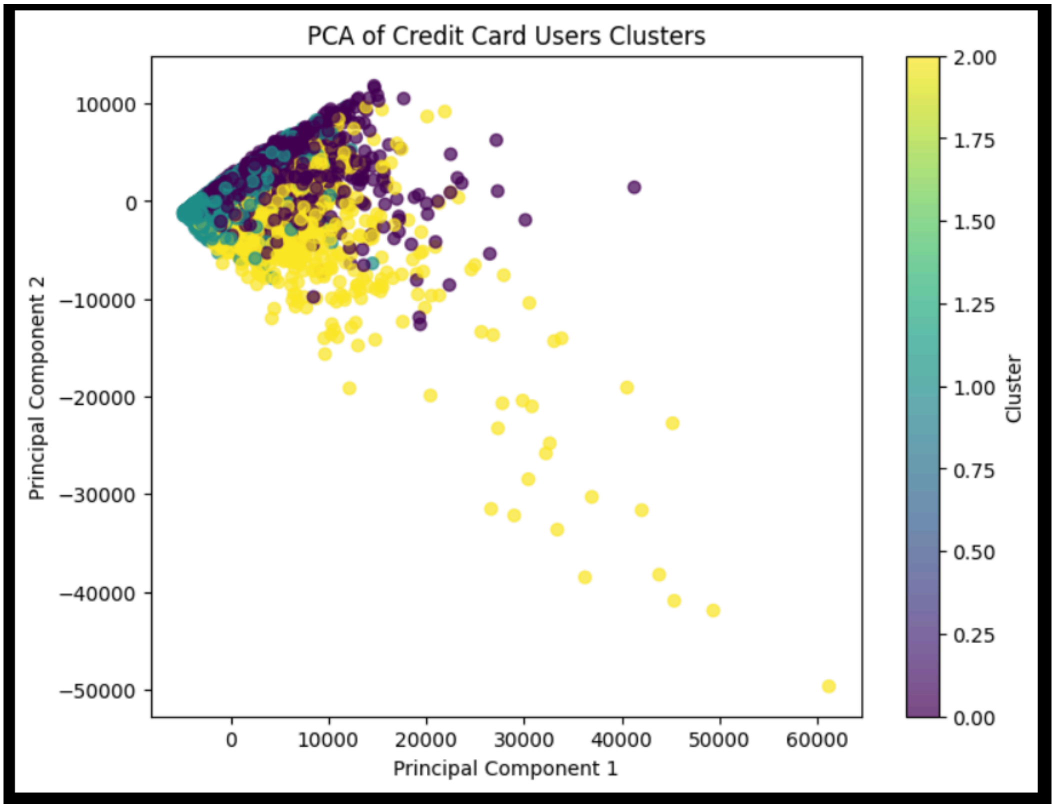
```
             BALANCE  BALANCE_FREQUENCY    PURCHASES  ONEOFF_PURCHASES  \
Cluster
0        3961.727625           0.957265   382.748835        249.077728
1         822.945763           0.835159   497.749484        245.933058
2        2164.505015           0.981182  4196.640322       2680.303546

         INSTALLMENTS_PURCHASES  CASH_ADVANCE  PURCHASES_FREQUENCY  \
Cluster
0                    133.749828   3920.602191             0.231794
1                    252.137907    333.141687             0.461756
2                   1516.808475    445.975174             0.947332

         ONEOFF_PURCHASES_FREQUENCY  PURCHASES_INSTALLMENTS_FREQUENCY  \
Cluster
0                          0.112693                          0.142943
1                          0.128077                          0.344204
2                          0.670532                          0.735409

         CASH_ADVANCE_FREQUENCY  CASH_ADVANCE_TRX  PURCHASES_TRX  \
Cluster
0                      0.454455         12.574793       5.567155
1                      0.068227          1.214707       8.553063
2                      0.061973          1.496855      55.567610

         CREDIT_LIMIT     PAYMENTS  PRC_FULL_PAYMENT     TENURE
Cluster
0         6675.750825  3038.989161          0.034762  11.326544
1         3270.844020   910.838456          0.153473  11.485424
2         7674.095912  4069.022681          0.301911  11.910377
```



K-means clustering was used for the modeling because it works well when the number of clusters is known in this case, three. It effectively separates credit card users into distinct groups based on their usage habits. The results show that Cluster 0 consists of users with high balances who rely heavily on cash advances but make relatively fewer purchases and rarely pay off their balances in full. Cluster 1 shows moderate users who use their credit cards more frequently for purchases, especially through installments. They make smaller payments but are slightly more likely to pay their balances in full. Cluster 2 represents users who make frequent and large purchases, particularly in both one-off and installment formats. They have the highest credit limits, make the largest payments, and are the most likely to pay off their balances completely, while rarely using cash advances.

Storytelling

Clustering has helped identify distinct patterns in how users interact with their credit cards. By using K-means clustering, I was able to group customers into three categories based on their spending habits. The analysis shows that users in Cluster 0 tend to rely heavily on cash advances and often struggle to pay off their balances. Cluster 1 shows users who use their credit cards moderately, mostly for installment purchases, and occasionally pay off their balances in full. Cluster 2 shows users who appear to have more financial stability, make high value purchases, and maintain strong repayment habits. By identifying these patterns, the initial question “Are there different user groups who prefer high one-time purchases, or installment-based buying, or rely on cash advances?” has been answered here.

Impact

This information can help credit card companies better understand how their users are spending money and identify which types of users are more likely to pay off their dues versus those who may struggle with repayment. By doing so credit card companies can tailor repayment plans to help users manage their credit more effectively.

Resources:

https://www.geeksforgeeks.org/clustering-in-machine-learning/

https://www.ibm.com/think/topics/k-means-clustering

https://www.ibm.com/think/topics/hierarchical-clustering