Credit Card Dataset Analysis

Data Analysis and Mining 2023-24

Made by:

João Lopes 60055

José Romano 59241

FACULDADE DE
CIÊNCIAS E TECNOLOGIA

Index

FACULDADE DE
CIÊNCIAS E TECNOLOGIA

Introduction

Motivation:

We believe that the nature of our choices comes very much from who we are, so the best way to explain why we chose a dataset like this is to talk a little about who we are.

We're both computer engineering students and during our degree we always felt that the course didn't deal with subjects that fulfilled us, either because of the work or the topics covered. Over time, driven by a passion for emerging artificial intelligence technologies and a fascination with human interactions on a macro and micro scale, we decided to enter the world of data science, which is why we chose the subject of data mining and analysis.

Something that marks our society today is undoubtedly consumerism, and whether out of necessity or lust, the most widely used technology in the world for transactions and payments is credit cards. Considering the United States alone, by the end of 2023, it is (1) estimated that Americans had around **$1.129 trillion** in credit card debt. For perspective, that's the equivalent of the (2)GDP of Portugal, Denmark, Finland and Hungary **combined**, i.e. Americans owe so much on credit cards that we'd have to use all the wealth produced by Portugal, Denmark, Finland and Hungary to pay it off. This comparison highlights the impressive financial weight of credit card debt in the most powerful economy in the world.

Combining our interest in the macro view of human behavior with the inferences made in the previous paragraph, we really think that this dataset will give us good experience in applying data mining and analysis techniques.

(1) https://www.lendingtree.com/credit-cards/credit-card-debt-statistics/
(2) https://www.worldometers.info/gdp/gdp-by-country/

Information of the data:

The dataset comprises information related to credit card holders' usage behavior over the past six months. It includes 18 behavioral variables, covering aspects such as balances, purchase behavior, cash advances, credit limits, and payment patterns. Each feature provides insights into different aspects of the customers' interactions with their credit

cards, allowing for a comprehensive analysis of their behavior and preferences. Here is a comprehensive description of each feature:

Table with all the features:

Feature	Description
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)
ONEOFFPURCHASESFREQUENCY	How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
PURCHASESINSTALLMENTSFREQUENCY	How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
CASHADVANCEFREQUENCY	How frequently the cash in advance being paid
CASHADVANCETRX	Number of Transactions made with "Cash in Advanced"
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
PRCFULLPAYMENT	Percent of full payment paid by user
TENURE	Tenure of credit card service for user

The dataset contains data for approximately 9000 active credit card holders. These **entities represent individual customers** whose behavior and transactions are captured and analyzed to gain insights into their preferences and habits regarding credit card usage.

The specific source address of the dataset is not provided in the information provided. However, it is noted that the dataset is used for developing customer segmentation

strategies to define marketing approaches, indicating its relevance to marketing and customer analytics.

Possible problems to be addressed and our proposal:

In the domain of customer segmentation and marketing strategy in the credit card industry, several problems can be addressed using data analysis techniques. Here are some examples:

1. **Customer Segmentation:** Segmenting customers based on their spending behavior, frequency of purchases, payment patterns, and credit utilization can help tailor marketing strategies to specific customer groups. This can lead to more targeted promotional campaigns and personalized offerings, ultimately increasing customer satisfaction and loyalty.

2. **Identifying High-Value Customers:** Analyzing customer data to identify high-value customers who generate significant revenue for the credit card company can help prioritize resources and focus retention efforts on retaining these valuable customers. Understanding the characteristics and behaviors of high-value customers can also inform acquisition strategies to attract similar prospects.

3. **Risk Assessment and Fraud Detection:** Utilizing data analytics to identify patterns of fraudulent activities and unusual spending behavior can help mitigate risks associated with credit card fraud. By developing predictive models that flag suspicious transactions in real-time, credit card companies can enhance security measures and protect both customers and the organization from financial losses.

4. **Credit Limit Optimization:** Analyzing customers' credit utilization patterns and payment behaviors can inform decisions regarding credit limit adjustments. Optimizing credit limits based on individual customer profiles can help minimize credit risk while maximizing customer satisfaction and spending potential.

5. **Promotion Effectiveness Analysis:** Evaluating the effectiveness of marketing promotions, such as cashback offers, rewards programs, or discounts, by analyzing their impact on customer spending behavior and retention rates. This analysis can provide insights into which promotions resonate most with customers and generate the highest return on investment.

Given the dataset provided, we propose to explore the problem of **Customer Segmentation**. By leveraging the various behavioral variables available in the dataset, such as purchase frequency, types of purchases, payment patterns, and credit limits, we can cluster customers into distinct segments based on their spending habits and preferences. Understanding these segments can guide the development of targeted marketing strategies tailored to the specific needs and preferences of each group. Customer segmentation is a fundamental aspect of marketing strategy, and by addressing this problem, we can unlock

opportunities to enhance customer engagement, increase retention, and drive business growth.

0. Data Preparation

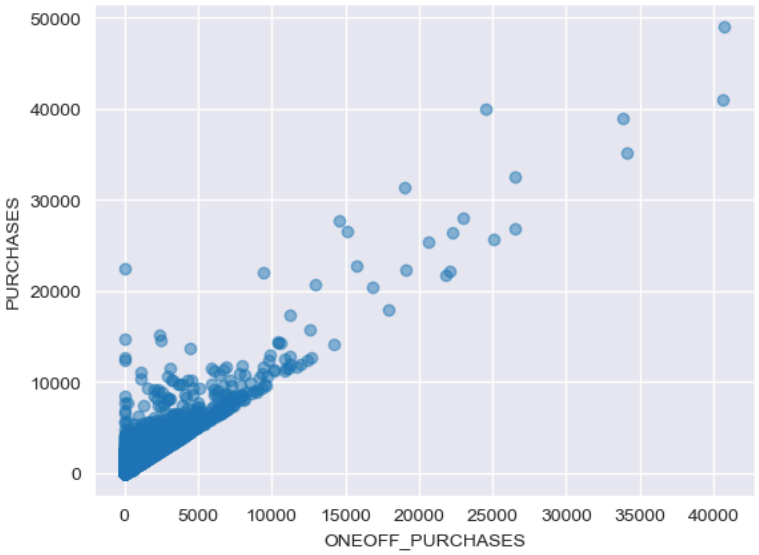
Like in every data science and analysis project, there's always data that has issues, whether it's missing values, outliers, or inconsistencies, and this dataset was no exception. Upon thorough examination, we identified that a particular feature in our dataset contained missing values. To address this issue, we opted for a common strategy: calculating the mean of the feature and using it to fill in the NaN (Not a Number) values. Fortunately, the number of missing values was relatively low, so employing this method will not significantly skew our data or compromise the integrity of our analysis. This approach allows us to maintain the overall structure and trends within our dataset while effectively handling the missing data.

1. Regression Analysis

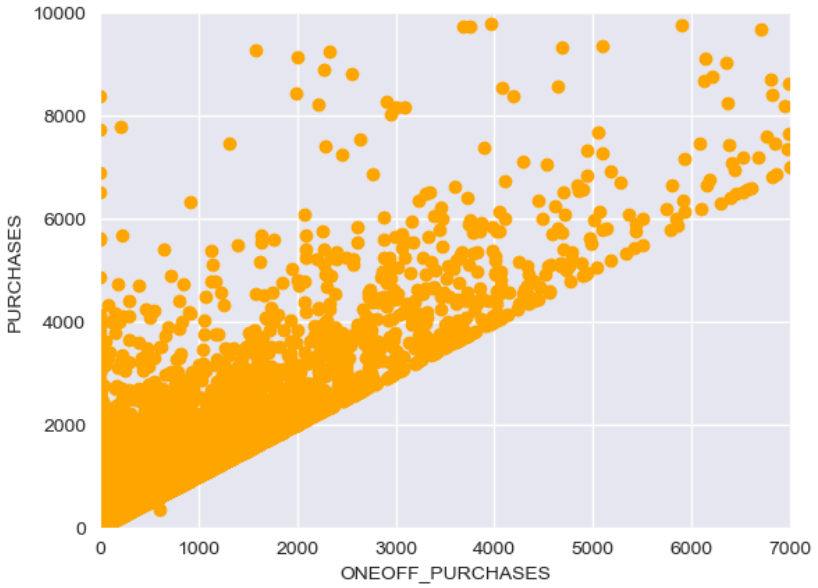
In this particular segment of the report, our focus will be on a thorough analysis of our dataset utilizing the linear regression method. Our primary objective is to dive into the viability of employing this method for our dataset and to uncover the insights and conclusions it can potentially offer. By applying the principles of linear regression, we aim to gain a comprehensive understanding of the relationships within our data and the extent to which this analytical approach can provide valuable insights for our objectives.

a) Choosing features to work with

To initiate our analysis, the initial step involves selecting two features for examination. These features ideally exhibit a "linear-like" scatterplot, indicating a potential linear relationship between them, which we will subsequently confirm mathematically.
After plotting all features against each other, our assessment led us to identify the PURCHASES and ONEOFF_PURCHASES as the most suitable features for our analysis. The corresponding scatterplot is depicted below:



Observing the plot, it's evident that there's a concentration of points within the range of [0; 10 000] on the x-axis of the graph. Additionally, these points appear relatively close together, indicating limited dispersion. However, this perception may be influenced by our current "zoomed-out" view. For clarity, we've included a plot below where we've restricted ONEOFF_PURHCASES to the interval [0;7000].



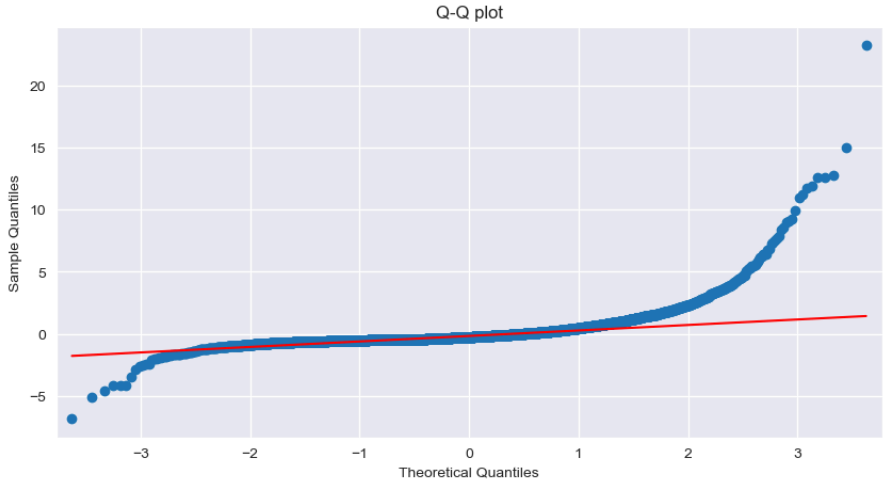
Upon comparing these plots with others, we have determined that these features exhibit the strongest linearity, prompting our decision to proceed with their analysis.

b) Create a linear regression between the features and generate a normal probability plot of the standardized residuals

Now let's generate the linear regression between the features that were chosen, the result is this:

Based on this regression, we will proceed to generate a QQ plot of the standardized residuals. The QQ plot allows us to visually assess the normality assumption of the residuals by comparing their distribution to that of a theoretical normal distribution. This graphical analysis will provide insights into the adequacy of the linear regression model and help us identify any potential departures from normality, such as skewness or heavy-tailedness, in the residuals. The result we got was:

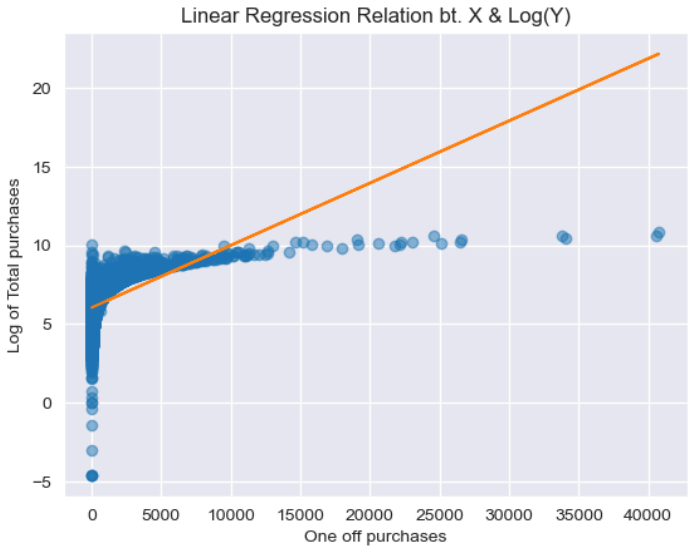
The normal probability plot suggests that the standardized residuals deviate from the assumption of normality. Specifically, the inverted S-shaped pattern, with a more pronounced



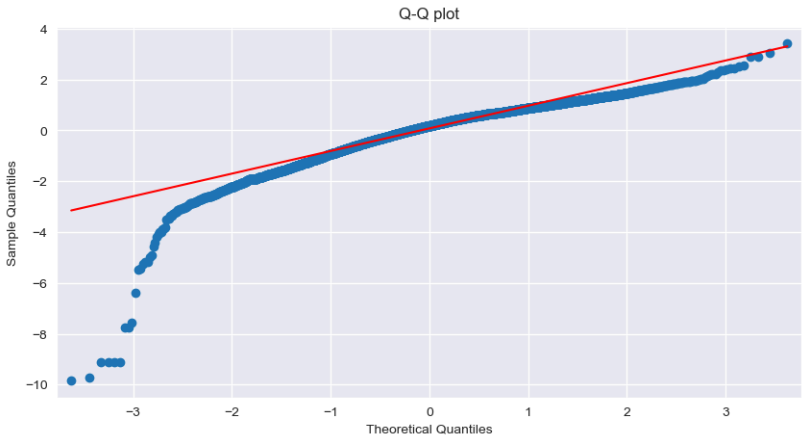
and skewed right tail, indicates the presence of positive skewness in the residuals. This skewness implies that the model tends to overestimate higher values more often than it underestimates them. Therefore, the normal probability plot does not indicate acceptable normality, and there is evidence of positive skewness in the standardized residuals.

c) Linear Regression with Natural Logarithm Transformation

Now we'll investigate the effectiveness of the natural logarithm transformation by performing a linear regression on the transformed features. By taking the natural logarithm of both variables, we aim to create a more linear relationship between them, making our regression model more interpretable and accurate. Here is the plot:



After performing the linear regression, we assess the normality of the residuals using a normal probability plot.



The nearly coinciding points with the QQ line suggest that the standardized residuals from the transformed regression model approximate a normal distribution quite closely. However, the presence of a small skewed left tail indicates a slight departure from perfect normality. This skewness suggests that there may be a slight asymmetry or non-normal behavior in the lower end of the residuals distribution.

Despite the small deviation, the overall alignment of the points with the QQ line suggests that the natural logarithm transformation has effectively addressed many of the non-linearities and variance heterogeneity (meaning the variance changes from observation to observation) present in the original data, resulting in residuals that closely approximate a normal distribution.

d) Population Regression Equation

Now we want to visualize the relationship between one-off purchases and total purchases by plotting the population regression equation. The population regression equation serves as a mathematical model that describes the average relationship between these two variables in the population. The equation that we got was:

$$Y = 1.16 * X + 409.45$$

The intercept, represented by Beta 0, suggests that even when there are no one-off purchases, there is a baseline level of total purchases, estimated to be approximately 409.45 units. This baseline value represents the minimum level of total purchases that can be expected, independent of the number of one-off purchases.

The slope coefficient, Beta 1, reveals the extent of change in total purchases for each unit increase in one-off purchases. With a value of 1.16, Beta 1 indicates a positive relationship between the two variables. This means that as the number of one-off purchases increases by one unit, the total purchases are expected to increase by approximately 1.16 units. This indicates that customers who make more one-off purchases tend to also make more purchases (in total).

e) Correlation and determinacy coefficients

In this subtopic we will calculate the Correlation (r) and the determinacy (R-Squared) and analyze its values. The correlation coefficient measures the strength and direction of the relationship between two variables, ranging from -1 to 1.

The determinacy coefficient, or R-squared, quantifies the proportion of variance in the dependent variable explained by the independent variables in a regression model. Ranging from 0 to 1, it helps assess the goodness of fit of the model and how well it captures the observed data.

$$r = 0.91 \; and \; R - Squared = 0.84$$

A correlation coefficient (r) of 0.91 indicates a strong positive linear relationship between the two variables. This value suggests that as one variable increases, the other tends to increase almost at the “same pace”.

An R-squared value of 0.84 indicates that approximately 84% of the variance in the dependent variable can be explained by the independent variable(s) in the regression model.

This high R-squared value suggests that the model provides a good fit to the data and effectively captures the relationship between the variables.

f) Test the statistical hypothesis : is there a linear relationship?

In our hypothesis testing, we aimed to determine whether there exists a linear relationship between the independent and dependent variables. Our null hypothesis (H0) stated that the coefficient of the independent variable (beta1 / One-Off Purchases) is equal to zero, suggesting no such relationship, while the alternative hypothesis (H1) proposed that beta1 is not equal to zero, indicating the presence of a linear relationship. We conducted our analysis using a significance level of 0.05, commonly accepted in statistical testing.

Upon conducting the hypothesis test, we obtained a p-value very close to absolute zero (as the image suggests) :

```
Test statistc for coefficient estimate  0 : [33.01969915] | P-value: [3.89118883e-222]
```

This exceptionally small p-value suggests that the observed results are highly unlikely to have occurred under the assumption of the null hypothesis. Therefore, we have strong evidence to **reject the null hypothesis** in favor of the alternative hypothesis. **This indicates that there is a statistically significant linear relationship between the independent and dependent variables.**

g) Confidence interval for the unknown true slope of the regression line

In our analysis, we constructed a confidence interval to estimate the unknown true slope of the regression line, representing the relationship between the independent variable (one-off purchases) and the dependent variable (total purchases). We chose a 95% confidence level for our interval, which is a commonly used threshold in statistical inference.

```
Confidence interval for unknown true slope of the line is: [1.14823815 1.17176185]
```

The confidence interval we obtained for the true slope of the regression line was [1.14823815, 1.17176185]. This interval provides a range of plausible values for the true slope with 95% confidence. Specifically, it suggests that we are 95% confident that the true slope of the regression line falls within this interval.

Interpreting the values within the interval, we can conclude that for every one-unit increase in one-off purchases, the total purchases are expected to increase by a value between approximately 1.14823815 and 1.17176185 units. This range of values represents the uncertainty associated with our estimate of the true slope.

h) Confidence interval for the population correlation coefficient

In our analysis, we constructed a 95% confidence interval to estimate the population correlation coefficient, which measures the strength and direction of the linear relationship between the two variables. Upon calculation, we obtained a correlation coefficient of 0.91. Subsequently, we constructed a confidence interval for the population correlation coefficient, which yielded the interval [0.9048562, 0.92395656]. This interval provides a range of plausible values for the population correlation coefficient with 95% confidence.

Interpreting the values within the interval, we can conclude that we are 95% confident that the true population correlation coefficient falls within the range of approximately 0.9048562 to 0.92395656. This range of values suggests a strong positive linear relationship between the variables, indicating that as one variable increases, the other variable tends to increase proportionally, and vice versa.

Our confidence interval serves as a valuable tool for quantifying the precision of our estimate and assessing the reliability of the relationship between the variables. By providing a range of plausible values, it allows us to account for variability in the data and make informed decisions based on the uncertainty inherent in our estimation process.

i) Confidence interval for the mean of the y-variable at a fixed value

As part of our study objective, aimed at creating customer profiles based on their spending behavior, we constructed a 95% confidence interval to estimate the mean value of the dependent variable (y) at a fixed value of our chosen independent variable. The resulting confidence interval for the mean value of y given x was calculated to be [11896.04110149, 12124.21872495]. This interval offers valuable insights into the expected variability in the mean spending of customers, providing a range of plausible values for the dependent variable with 95% confidence, given a specific level of the independent variable (One-off Purchases).

```
Confidence interval for mean value of y given x: [11896.04110149 12124.21872495]
```

Interpreting the confidence interval, we can assert with 95% confidence that the true mean value of the dependent variable (Total Purchases) falls within the range of approximately 11896.04110149 to 12124.21872495, for the chosen fixed value of the independent variable. This range offers valuable information about the range of plausible mean spending values for customers at the specified level of the independent variable.

As we aim to create customer profiles based on their spending patterns, this confidence interval gives us essential insights into the expected variability in customer spending behavior. By understanding the range of mean spending values at a predetermined level of the independent variable, we can better segment (later in this report) types of customers based on their spending habits.

j) Confidence interval for a randomly chosen value of the y-variable

Now that we've computed the interval using the mean of the y-variable, let's extend our analysis by constructing and interpreting a 95% confidence interval for a randomly chosen value of the y-variable (total purchases) at a fixed value of our chosen independent variable (one-off purchases).

```
Confidence interval for a randomly chosen value of y given x: [ 9368.35489726 14651.90492919]
```

If we consider this result, the confidence interval [9368.35489726, 14651.90492919] dollars illustrates the expected variability in total purchases for customers with a specified level of one-off purchases. With 95% confidence, we expect a randomly selected total purchase amount to fall within this range when one-off purchases are fixed at a certain level. This interval captures the diversity of individual spending behaviors among customers with similar levels of one-off purchases, providing insights into the potential range of total purchases for new observations.

2. PCA - Principal Component Analysis

In the Principal Component Analysis (PCA) phase, we aimed to extract essential information from our dataset to facilitate further analysis. When asked to select a subset of 3-6 features, we believe that the ones that really capture the main insights that we want to explore are: **balance** (remaining account balance), **purchases** (total credit card expenditures), **purchase frequency** (how often purchases are made), **one-off purchase frequency** (frequency of one-time purchases), and **installment purchase frequency** (frequency of installment payments).

In the realm of credit card businesses, defining user profiles and segmenting users are essential tasks. To achieve this, understanding the main characteristics of potential clients becomes paramount. In our case, we seek to identify clients who don't spend beyond their account balance but still use their credit cards to make some purchases, thereby targeting possible clients that might be appealed by products with special conditions. Subsequently, we aim to discern between clients who prefer one-off purchases over installment payments in order to understand what kind of products they could be looking for : lower interest rates or possibly cashback options.

Given these objectives, our selection of features is well-aligned with our analytical pursuits, paving the way for insightful analyses and informed decision-making.

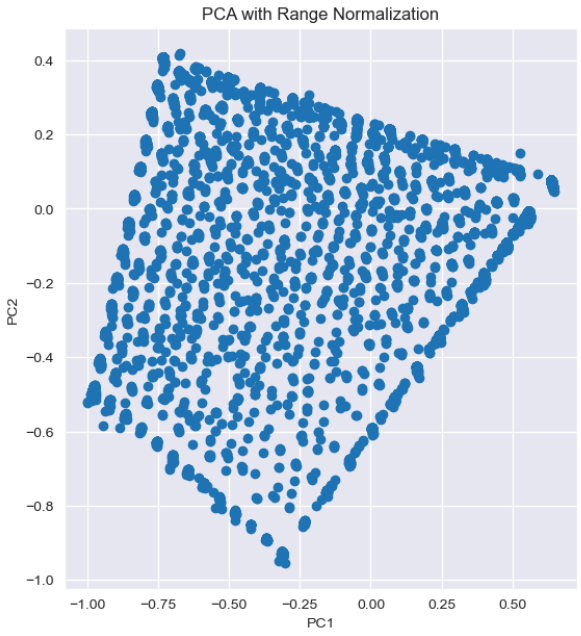
a) Visualize the data over these features using two normalization methods: range and Standard deviation

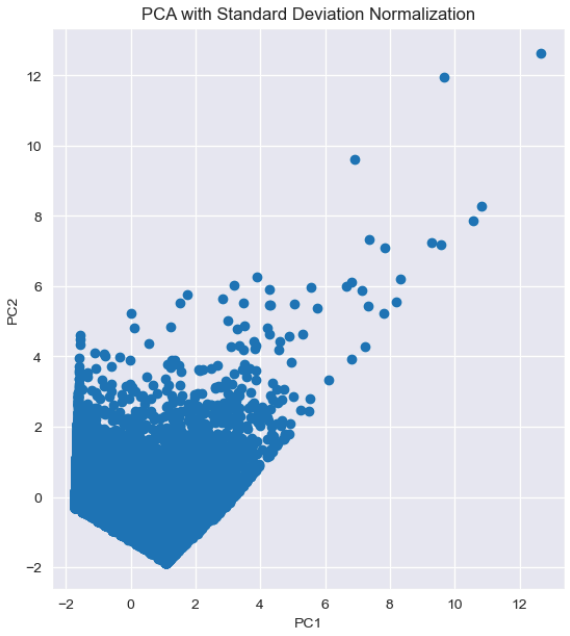
To help us better understand the data, we used two different ways to make it easier to see patterns. First, we normalized the data by range, which means we put it all on a scale from 0 to 1. Then, we normalized it by standard deviations, which involves adjusting the data based on its average and how much it varies from that average.

We then plotted this data in a 2D or 3D graph, using principal component analysis (PCA). This method helps simplify the data so we can see any trends or groupings more clearly. By doing this with both normalization methods, we could compare which one made the data easier to understand. This step is important because it helps us choose the best way to work with the data for our further analyses.

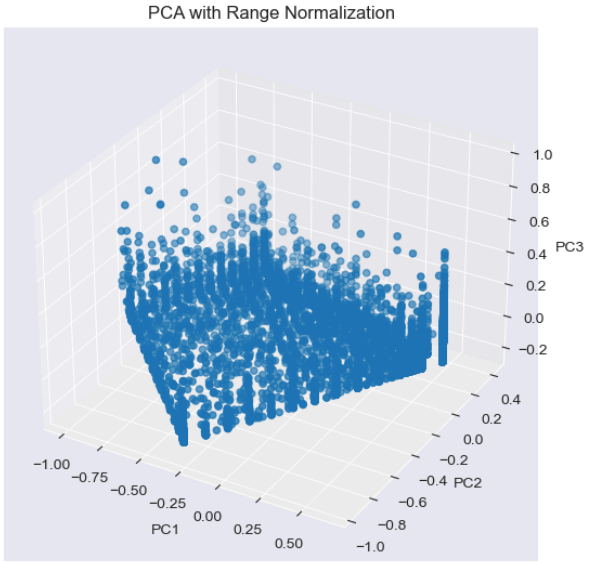
Given all of this, we got the following plots:

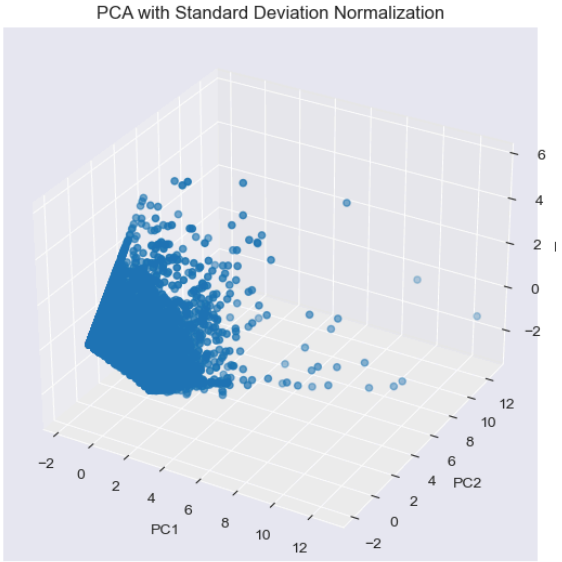
2D:





3D:





Considering that one of the main objectives of normalization is to create a less dispersed plot from data - a more concentrated distribution indicates that the data points are closer together, suggesting less variability in the data set - we can conclude from this topic that Normalization with **Standard deviation** is the best option for our data.

b) Choose between conventional PCA or SVD

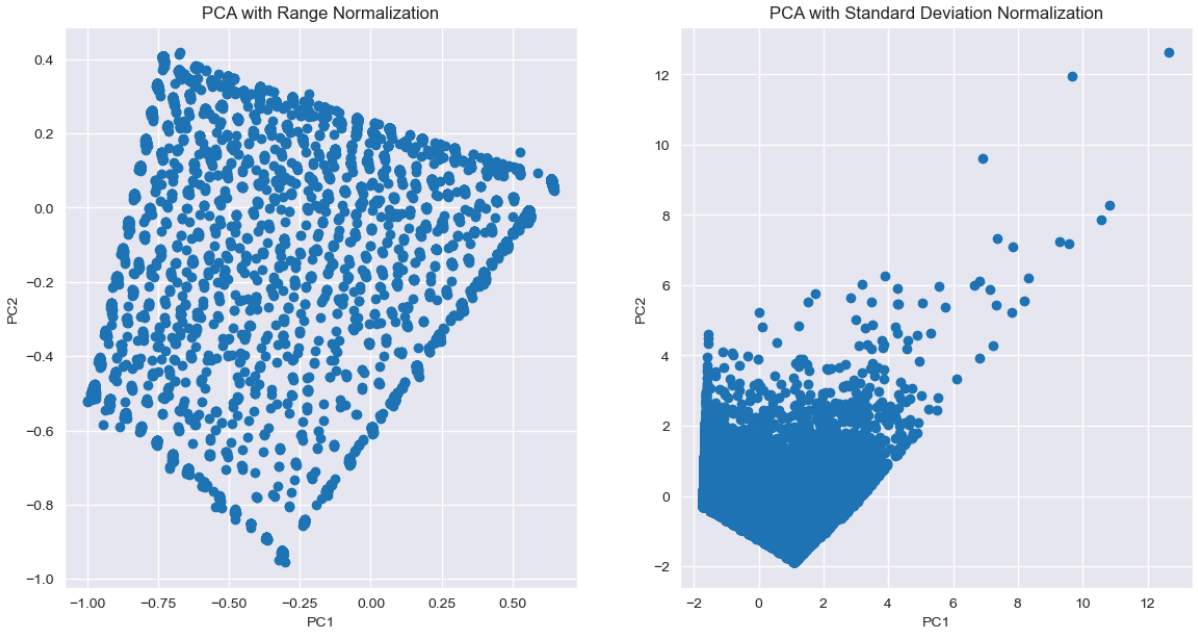
Principal Component Analysis (PCA) and Singular Vector Decomposition (SVD) are techniques used for dimensionality reduction in data analysis.

1. **PCA** reduces the dimensionality of a dataset while retaining its key features, making it easier to interpret and visualize complex data. It does this by transforming the original features into a new set of uncorrelated variables called principal components.

2. **SVD** decomposes a matrix into three constituent matrices, providing insights into the underlying structure and relationships within the data. It is widely used in various fields, including image processing, recommendation systems, and natural language processing.

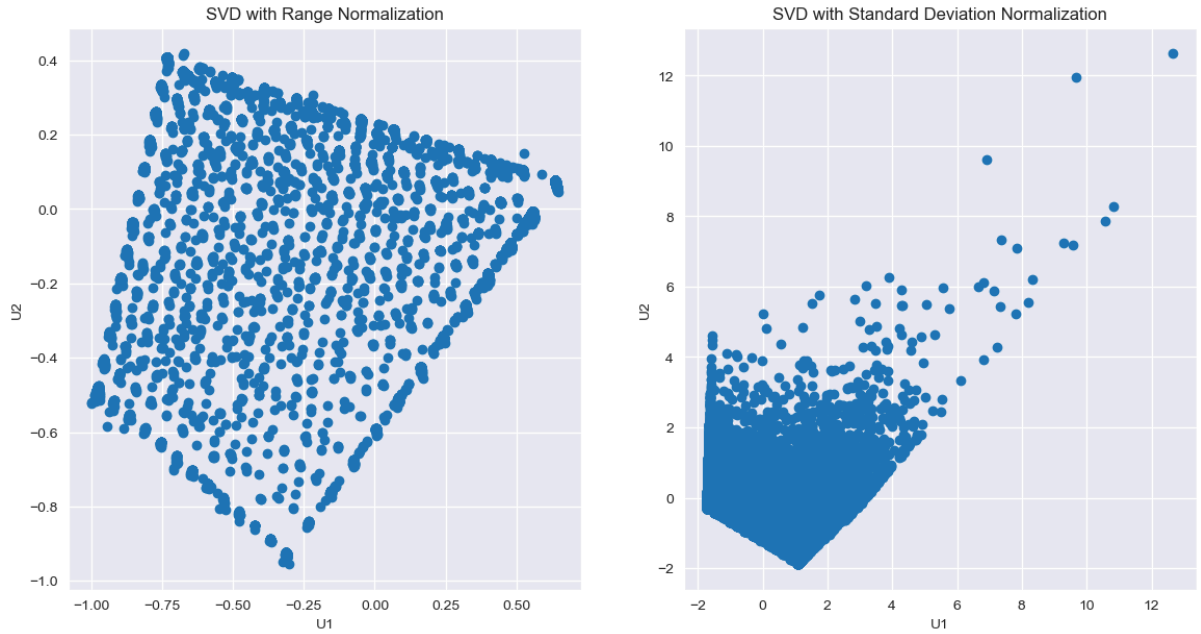
So, after making the calculations necessary for each method, we've come with the following plots (for the normalized data using range and standard deviation):

PCA- Visualizing normalization techniques: (1) range (2) Standard deviation



These plots are the same as we observed in the previous section, where we concluded that normalizing our data with standard deviation is more effective and better than normalizing by range, due to the fact that data gets more concentrated on the first one when compared to the second.

SVD- Visualizing normalization techniques: (1) range (2) Standard deviation



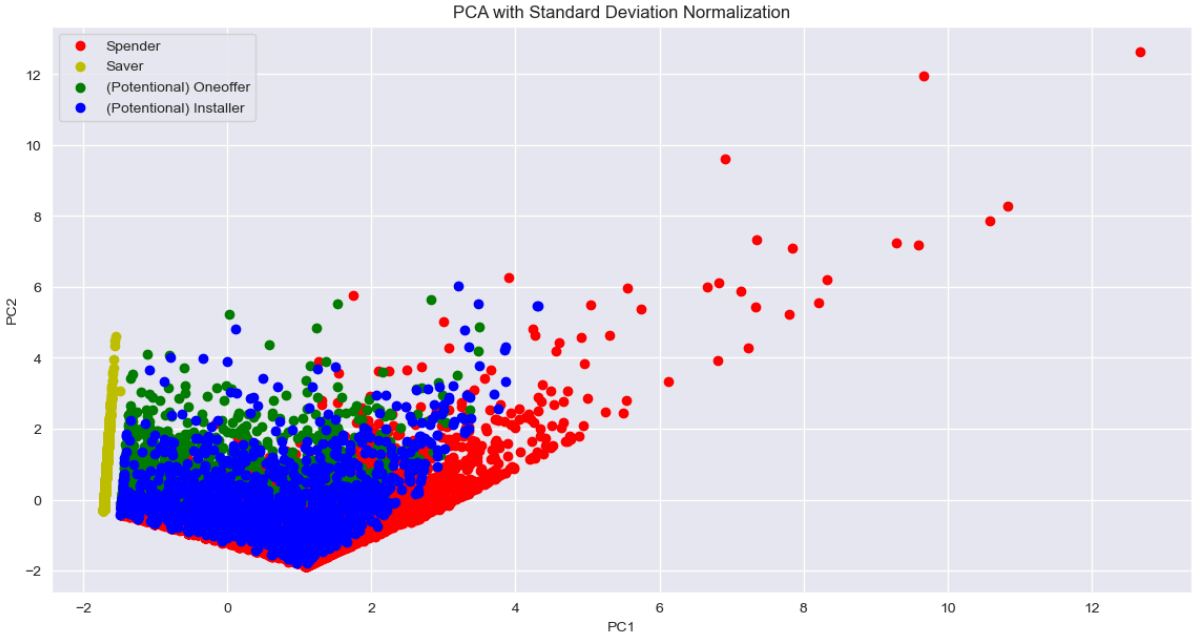
By analyzing the plots generated from the application of Singular Value Decomposition (SVD), we can confirm that the premise stating that **standard deviation normalization is better than range normalization** for our dataset holds **true**. The SVD plots demonstrate that standard deviation normalization effectively manages data dispersion.

This examination reveals also that there is **minimal difference** between the outcomes obtained from SVD and Principal Component Analysis (PCA). Both techniques yield comparable results, indicating that either approach can be chosen for dimensionality reduction. However, it is crucial to note that the superior performance of standard deviation normalization persists across both SVD and PCA analyses.

Then, based on these observations, we can confidently opt for either SVD or PCA for dimensionality reduction, always utilizing standard deviation normalization for optimal results.

c) Represent group of objects using distinct colors

To segment users effectively and uncover meaningful patterns, we employed distinct colors to represent predefined groups of customers on the PCA plot. Red was assigned to customers with a balance less than their total purchases, indicating **prudent spending behavior** - remember that we want to focus on potential clients that might not have found the perfect service for them. Yellow denoted customers with a balance greater than their purchases, who also exhibited a **preference for one-off purchases over installments** and made frequent purchases overall. Green represented customers with a balance greater than purchases but a **preference for installment payments over one-off purchases**. Lastly, blue was assigned to customers who did not engage in any purchases during the observation period.



This segmentation strategy allowed us to visually identify and differentiate between various customer profiles based on their spending habits and preferences. By focusing on specific groups, such as those who spend responsibly or favor certain payment methods, we can tailor marketing strategies and product offerings to better meet their needs and preferences.

Overall, the combination of feature selection and customer segmentation through distinct coloring on the PCA plot lays quite a solid foundation for insightful analyses and informed decision-making in the credit card business domain (even before applying cluster algorithms).

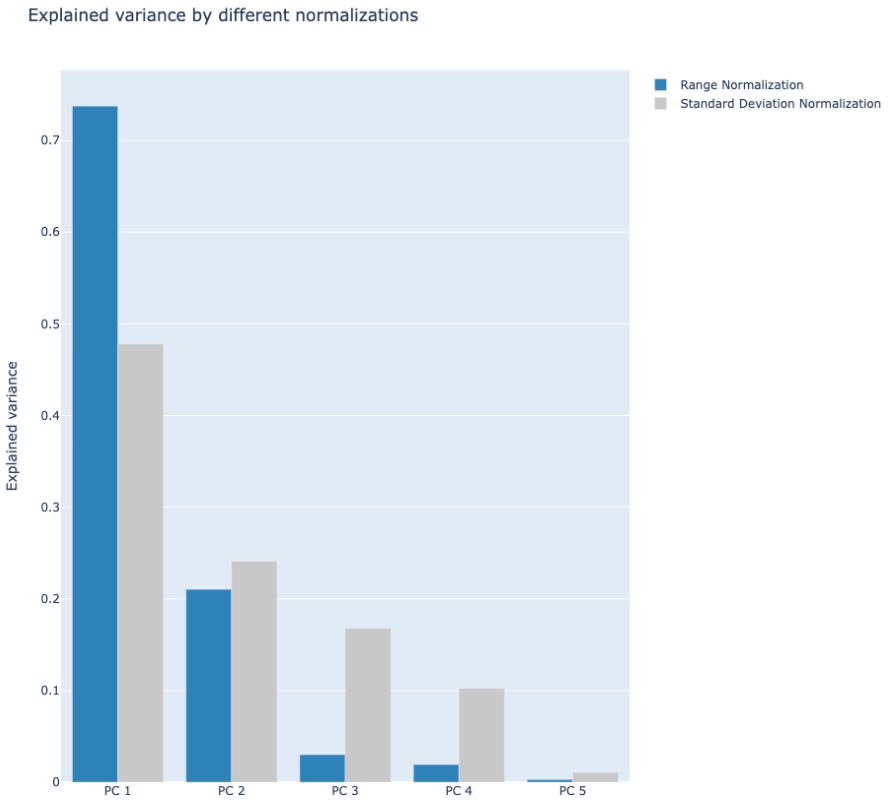
d) Discuss the “quality” of the PC Projection

Before proceeding to analyze the principal component (PC) projections, it's important to evaluate their quality first to ensure the effectiveness of our Principal Component Analysis (PCA) approach. To achieve this, we utilize various metrics and techniques to assess the fidelity of the projections and the preservation of essential information from the original dataset.

To analyze the quality of the PC projections, we have various key metrics at our disposal, but the one we will be using is:

- **Explained Variance**: Examining the percentage of variance explained by each principal component.

By examining the explained variance, we gain insight into how much of the original dataset's variability is captured by each principal component, allowing us to understand the significance of each PC in representing the data's essential characteristics.



Although range normalization may yield higher explained variance by component, it does not necessarily equate to better overall performance. Standard deviation normalization, despite potentially resulting in slightly lower explained variance, may still be superior in capturing the

underlying structure of the data and preserving essential relationships between variables. Thus, while range normalization excels in certain aspects, standard deviation normalization remains a more effective choice for principal component analysis in terms of data representation and interpretability.

Note: For a more in-depth quality assessment, we would need to employ additional techniques and metrics, such as scree plots, cumulative variance analysis, reconstruction error evaluation, and visualization of data clusters in the reduced-dimensional space.

3. Fuzzy Clustering with Anomalous Patterns

Fuzzy clustering with anomalous patterns is a data analysis technique used to group similar data points together while simultaneously identifying anomalous patterns within the dataset. Unlike traditional clustering methods that assign each data point to a single cluster, fuzzy clustering allows for partial membership, meaning a data point can belong to multiple clusters to varying degrees of membership.

Anomalous patterns, or outliers, are data points that deviate significantly from the rest of the dataset. These outliers can provide valuable insights into unusual behavior or unexpected trends within the data. In the context of fuzzy clustering, detecting and understanding these anomalous patterns alongside the clustering process can enhance the robustness and interpretability of the clustering results.

By applying fuzzy clustering with anomalous patterns, data scientists can gain a deeper understanding of complex datasets, uncover hidden patterns, and make more informed decisions based on the insights derived from both the clustering and anomaly detection processes.

This report delves into the application of fuzzy clustering with anomalous patterns on a credit card dataset, aiming to extract meaningful customer profiles and explore the potential for enhancing business strategies. Let's delve into our analysis and findings to understand the implications for real-world applications.

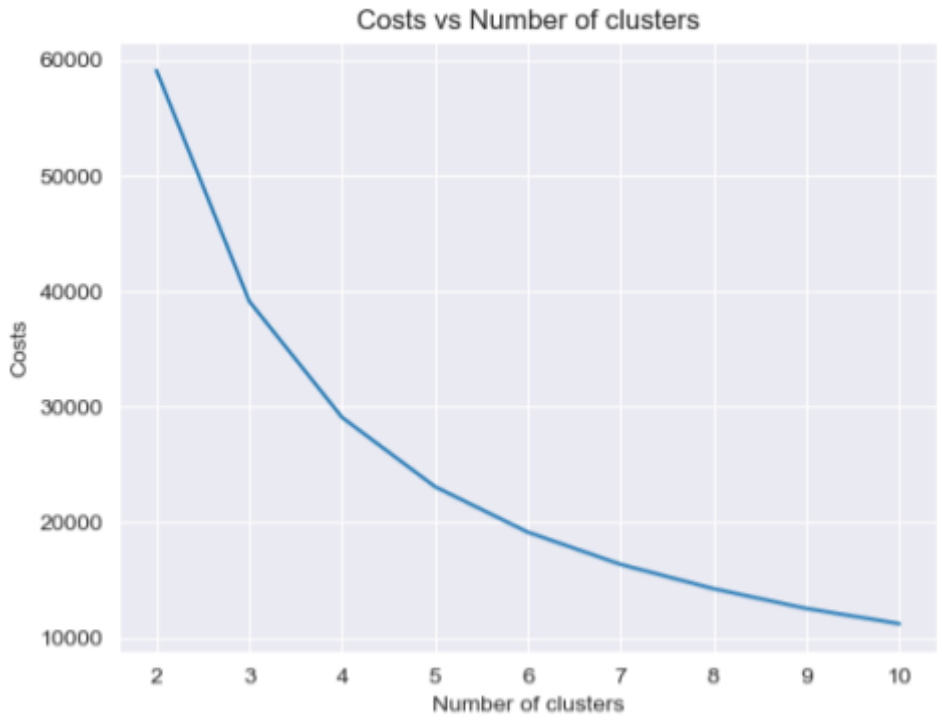
a) Finding the best number of clusters c

The Fuzzy C-Means (FCM) algorithm is utilized to perform fuzzy clustering on a dataset. In this analysis, the FCM clustering criterion is evaluated for different numbers of clusters (c) ranging from 2 to 10. The criterion is a measure of the fuzziness of the clustering, where lower values indicate better-defined clusters.

By iterating over each value of c and multiple random seeds, the FCM clustering algorithm is applied, and the final clustering criterion (JM) is recorded. This process is repeated for each value of c to obtain a comprehensive understanding of how the clustering criterion varies with the number of clusters.

Interpreting the FCM Loss Graphic (JM) VS Fuzzy Partitioning Coefficient(FPC):

Even though the *scikit-fuzz* outputs in its function *cmeans* the variable jm (objective function history), we were advised that we should try to plot the number of clusters against other variable that is generated inside the function but its not the main output -the fpc. This variable is the true fuzzy partitioning coefficient of each number of clusters.



Plot using jm against number of clusters

With this plot, we would need to apply the elbow method ("where's the shape that looks like an elbow?") and we would get c = 3 as the ideal number of clusters to apply fuzzy c-means. This method of inference is wrong due to the fact that this doesn't explain what we really want!

In order to really understand how the fuzziness changes as we change the number of clusters, we need to plot the FPC (fuzzy partitioning coefficient) against the number of clusters. Then, we get this result:

Plot of FPC against number of clusters

Now with this plot we can really assess how the fuzziness changes with the increasing number of clusters. Since we want to choose a number of clusters that minimizes the fuzziness, we clearly see that c = 9 corresponds to the lowest fpc value (and not 3, as we've seen before).

Although c = 9 minimizes fuzziness, this doesn't mean that it provides a more clear structure of clusters because considering the high number of points, visually this could lead to a harder graphic to interpret. This is something that c = 3 does better (providing a simpler structure, but with more fuzz).

b) Apply the Iterative Anomalous Pattern (IAP) as the initialization algorithm for FCM (CHANGES)

In our Fuzzy C-Means (FCM) implementation, we opted not to utilize the Iterative Anomaly Prediction (IAP) algorithm for initialization due to a **combination of external and internal circumstances.**

Externally, our project faced time constraints and resource limitations, which affected the feasibility of incorporating additional complexity into our FCM algorithm. Given the project deadlines and the need to prioritize other essential aspects of our analysis, such as data preprocessing and model evaluation, we made the decision to streamline our implementation to ensure timely completion and effective utilization of available resources.

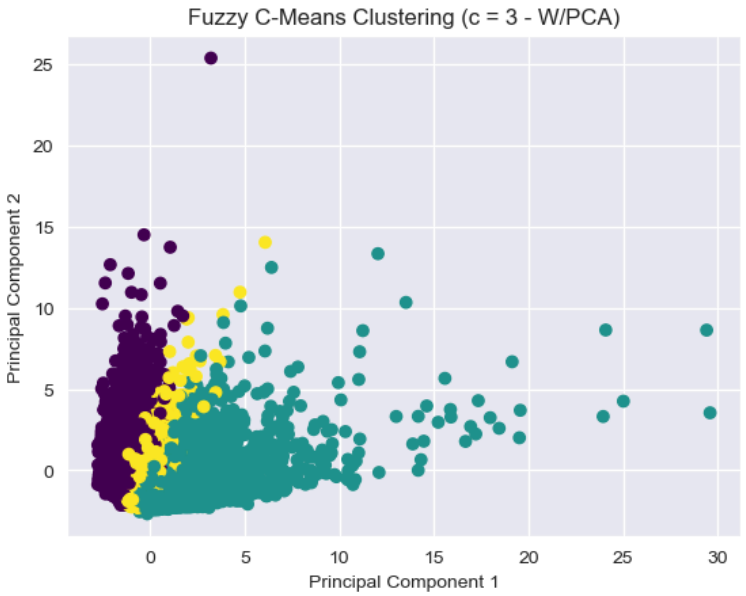
Internally, the dynamic nature of our dataset and evolving project requirements influenced our decision-making process. While IAP offers potential benefits in initializing cluster centroids by iteratively identifying anomalous data points, its implementation would have required careful parameter tuning and validation to achieve optimal results. However, dedicating resources to fine-tuning the IAP algorithm may have detracted from addressing other critical aspects of our research.

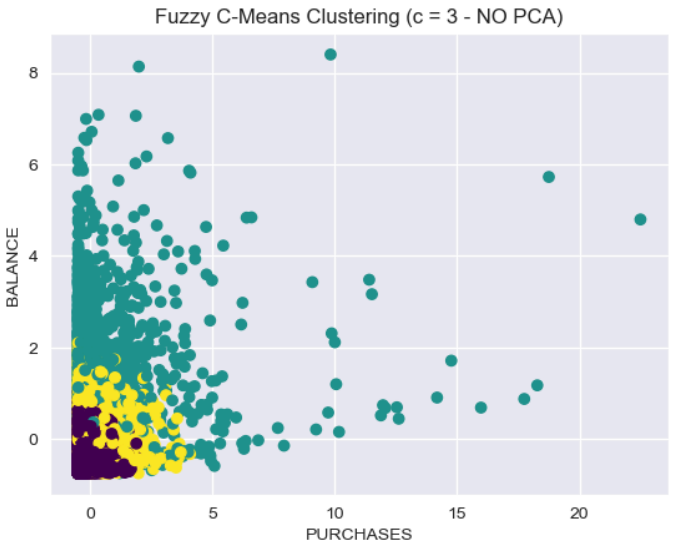
Despite the exclusion of IAP from our FCM implementation, we acknowledge its potential value in enhancing initialization strategies and improving clustering performance. Moving forward, we remain open to exploring alternative initialization methods, including IAP, in future iterations of our analysis, as circumstances permit.

c) Final analysis

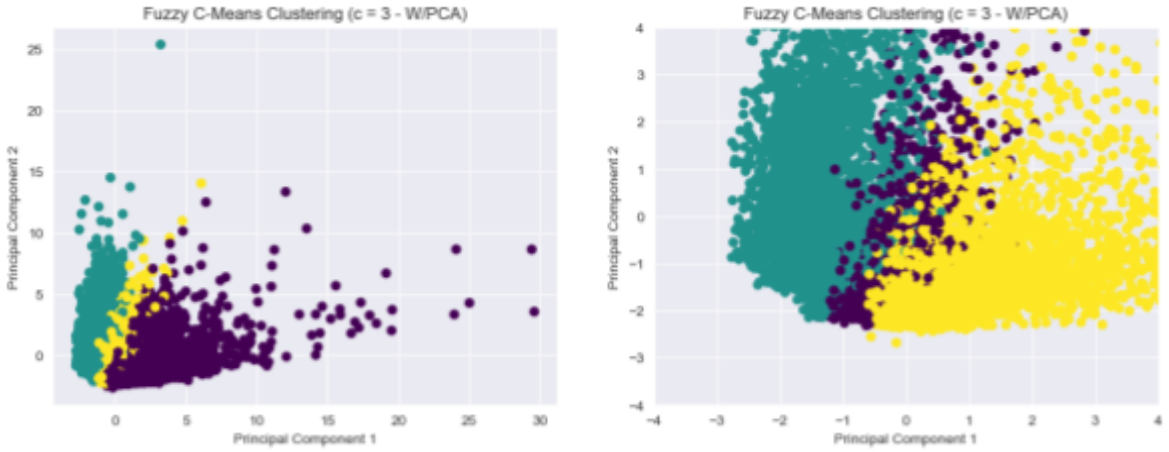
In our exploration of clustering techniques using Fuzzy C-Means (FCM), we conducted a comparative analysis of clustering outcomes obtained from datasets preprocessed with Principal Component Analysis (PCA) and the original datasets. Additionally, we investigated the clustering performance with varying numbers of clusters, specifically 3 clusters and 9 clusters.

When examining the clustering outcomes, we found that applying PCA preprocessing to the datasets had a notable effect on the clustering results. Clusters formed after PCA preprocessing exhibited a more cohesive shape, providing a clearer delineation of cluster boundaries and spatial distribution. The reduction in dimensionality facilitated by PCA contributed to a more interpretable representation of the data, enabling easier identification and understanding of cluster patterns.



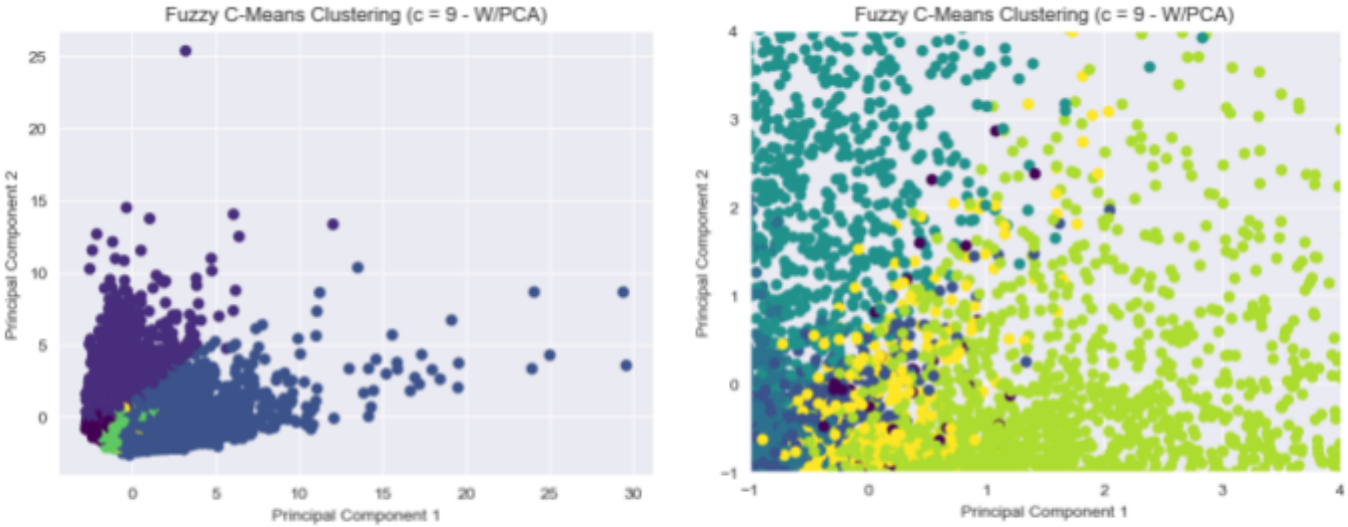


Interestingly, we observed a variation in the fuzziness partitioning coefficient between the 3-cluster and 9-cluster configurations. The 3-cluster configuration exhibited a higher fuzziness partitioning coefficient compared to the 9-cluster configuration, indicating a greater degree of overlap and ambiguity in cluster assignments. However, despite the higher fuzziness partitioning coefficient, the 3-cluster configuration yielded a significantly higher silhouette score compared to the 9-cluster configuration. This discrepancy suggests that while the 9-cluster configuration may provide a more granular segmentation of the data, the 3-cluster configuration offers a more cohesive and internally consistent clustering solution.



3 clusters - Plot and respective zoom for clearer view

As we can see, when we zoom right in the center where the clusters touch on each other, we see that due to the fact that there's less clusters, the shape is better (even though the points on one cluster might overlap on others).



9 clusters - Plot and respective zoom for clearer view

In addition to evaluating the clustering outcomes based on the fuzziness partitioning coefficient, we also examined the silhouette score as a measure of clustering quality. The silhouette score provides a measure of how well each data point lies within its assigned cluster, ranging from -1 to 1. A higher silhouette score indicates that the data point is well-matched to its own cluster and poorly matched to neighboring clusters.

When computing the silhouette scores, we got the following results:

PCA | c = 3	PCA | c = 9	No PCA | c = 3	No PCA | c = 9
0.12788	0.05280	0.07841	-0.00469

The higher silhouette score associated with the 3-cluster configuration implies that the clusters formed are more distinct and well-separated from each other, resulting in a clearer delineation of cluster boundaries and a higher degree of confidence in the cluster assignments. This observation underscores the importance of considering not only the number of clusters but also the quality and interpretability of the resulting clustering solution.

By incorporating the silhouette score into our evaluation of clustering outcomes, we gained deeper insights into the structure and characteristics of our dataset. This global view allows us to make informed decisions about the most suitable clustering configuration based on the specific requirements and objectives of our analysis.

d) How does this all relate to customer profiling?

Our exploration of clustering techniques using Fuzzy C-Means (FCM) revealed interesting insights into the optimal number of clusters for our dataset. We found that configuring FCM with three clusters resulted in the formation of more distinct and well-defined macro profiles of clients. This macro-level analysis provides a comprehensive understanding of the overarching characteristics and behaviors of client segments.

However, for a deeper understanding of the specific attributes and nuances within each client segment, we recommend utilizing nine clusters. By increasing the granularity of segmentation, we can delve into the finer details of each client subgroup and identify unique patterns and preferences. This micro-level analysis enables us to tailor strategies and interventions more effectively to meet the diverse needs and preferences of our clients.

In summary, while the three-cluster configuration offers a broad overview of client profiles, the nine-cluster configuration allows for a more detailed examination of individual characteristics within each segment. **By leveraging both macro and micro analyses, we can gain comprehensive insights into client behavior and preferences, empowering us to make informed decisions and optimize our strategies for greater effectiveness and impact.**

References

1. Throughout this project, artificial intelligence (AI) was used as a tool for gaining deeper insights into various concepts, assistance with python code for debugging and enhancing the quality of English language components.
2. https://pythonhosted.org/scikit-fuzzy/api/skfuzzy.cluster.html#cmeans
3. https://run.unl.pt/bitstream/10362/72311/4/TEGI0451.pdf
4. https://files.consumerfinance.gov/f/documents/cfpb_consumer-credit-card-market-report_2021.pdf