MATH 156 FINAL PROJECT
UNIVERSITY OF CALIFORNIA, LOS ANGELES
FALL 2023

Credit Card User Segmentation Using Machine Learning Techniques

Jacob Titcomb, Fang Shen,
Alicia Ying, Yixin Chen

In this project, we explore a comprehensive data set of credit card users. Our goal is to uncover underlying patterns in customer behavior—particularly customer risk—based on the given data. Utilizing machine learning techniques, primarily clustering, we aim to segment customers based on their credit card usage. This analysis is crucial for developing targeted marketing strategies and understanding consumer spending habits.

Code: `https://github.com/glassfox15/credit_card_segmentation`
Data Source: `https://www.kaggle.com/datasets/arjunbhasin2013/ccdata/data`

1 INTRODUCTION

The rapid growth of credit card usage has generated vast amounts of data, offering invaluable insights into consumer behavior. Our dataset, entitled 'CC GENERAL.csv,' is comprised of variables such as account balances, purchase amounts, cash advances, and credit payment patterns. As these features pertain to credit card usage and not the items themselves being purchased, we approach this customer segmentation from the point of view of a credit issuing company. We want to assess the spending patterns and financial risk associated with certain users: do certain groups accumulate debt while others pay their credit charges in advance? do groups with high credit limits and account balances buy high-value items more frequently?

In this project, we aim to address these questions and other related ones by applying unsupervised machine learning algorithms to the credit card data in order to identify distinct clusters of customers. We will compare K-Means, Hierarchical Clustering, Gaussian Mixture Models, and Spectral Clustering, choosing one of the four models and relating our findings to the original data.

2 Data Pre-Processing

For this project, we used the following packages: Pandas, Numpy, Scikit-Learn, Matplotlib, Seaborn, and Plotly. The dataset, initially loaded in as a Pandas DataFrame, consists of 18 variables:

> `CUST_ID`, `BALANCE`, `PURCHASES`, `ONEOFF_PURCHASES`, `INSTALLMENTS_PURCHASES`, `BALANCE_FREQUENCY`, `CASH_ADVANCE`, `PURCHASE_FREQUENCY`,`CREDIT_LIMIT`, `PAYMENTS`, `PURCHASES_INSTALLMENTS_FREQUENCY`, `CASH_ADVANCE_FREQUENCY`, `CASH_ADVANCE_TRX`, `PURCHASES_TRX`, `MINIMUM_PAYMENTS`, `PRC_FULL_PAYMENT`, `ONEOFF_PURCHASE_FREQUENCY`, `TENURE`

For variables labeled "frequency," values closer to 0 indicate infrequent occurrences and values closer to 1 indicate higher occurrences. Variables with the label "trx" are related to transactions; for example, `CASH_ADVANCE_TRX` is the number of transactions paid for in advance. Finally, the features `TENURE` and `PRC_FULL_PAYMENT` are, respectively, how long the user has been with the credit card service, and the percent of full payment the user made.

An initial data analysis revealed that the features consist of continuous variables, with only one—tenure—having relatively discrete (though still numerically informative) values.

2.1 Data Cleaning

Our first step was to address missing values. We noted missing entries in `CREDIT_LIMIT` and `MINIMUM_PAYMENTS`, opting to fill the former's single missing entry with the median value and dropping the latter feature for simplicity. We also removed observations that showed a cash advance frequency of greater than 1, which is impossible.

From here, we observed in histograms for each feature (figure 2.1) that most of the variables that range beyond 0 to 1 are heavily skewed. This makes sense since financial data often has very high outliers yet most observations tend to stay closer to 0. To address this issue, we performed an inverse hyperbolic sine transformation on the highly-skewed variables, i.e., for x_{ij} the i-th observation for variable j,

$$x_{ij}^{new} = \sinh^{-1}(x_{ij}) = \log\left(x_{ij} + \sqrt{x_{ij}^2 + 1}\right)$$

The benefit of this transformation is that we get the skew-reducing properties of the log transformation while still using observations with a value of 0. The result of this transformation can be seen in the EDA section (figure 3.1).

Lastly, we removed `CUST_ID`, as customer ID numbers were irrelevant for our analysis.

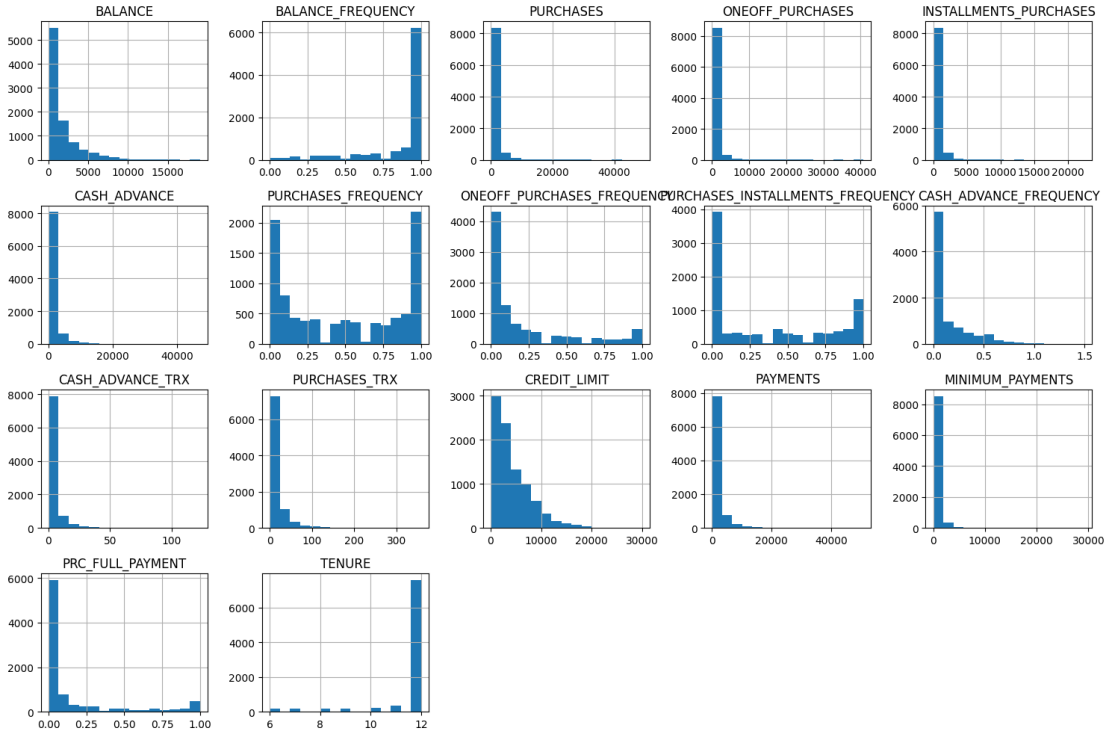


Figure 2.1: Feature Histograms, Pre-Transformation

3 Exploratory Data Analysis (EDA)

Our first mode of analysis was a visual analysis conducted with histograms (figure 3.1). For each variable, we graphed the distribution to get a general understanding of the overall data. We can see the benefits of our earlier transformation of the data in that the variables that were previously heavily skewed are now in somewhat of a Gaussian shape. For variables like `PURCHASES` and `INSTALLMENTS_PURCHASES`, we observe many data points at or near 0, already hinting at possible similar behavior.

Features `BALANCE` and `PAYMENTS` show strong unimodality with a slight left skew. `PURCHASES`, `ONEOFF_PURCHASES`, `INSTALLMENS_PURCHASES`, `CASH_ADVANCE`, `CASH_ADVANCE_TRX`, and `PURCHASES_TRX` exhibit unique bimodal behavior with one somewhat Gaussian peak and another larger, separate bar consisting of just values of 0 (or values close to it).

The variables with more extreme distributions—exhibiting bimodality or extreme skew—are features which are either discrete (i.e. `TENURE`) or a proportion between 0 and 1. Features with bimodal distributions, `PURCHASES_FREQUENCY` and `PURCHASES_INSTALLMENTS_FREQUENCY`, could also indicate underlying commonalities within the observations.

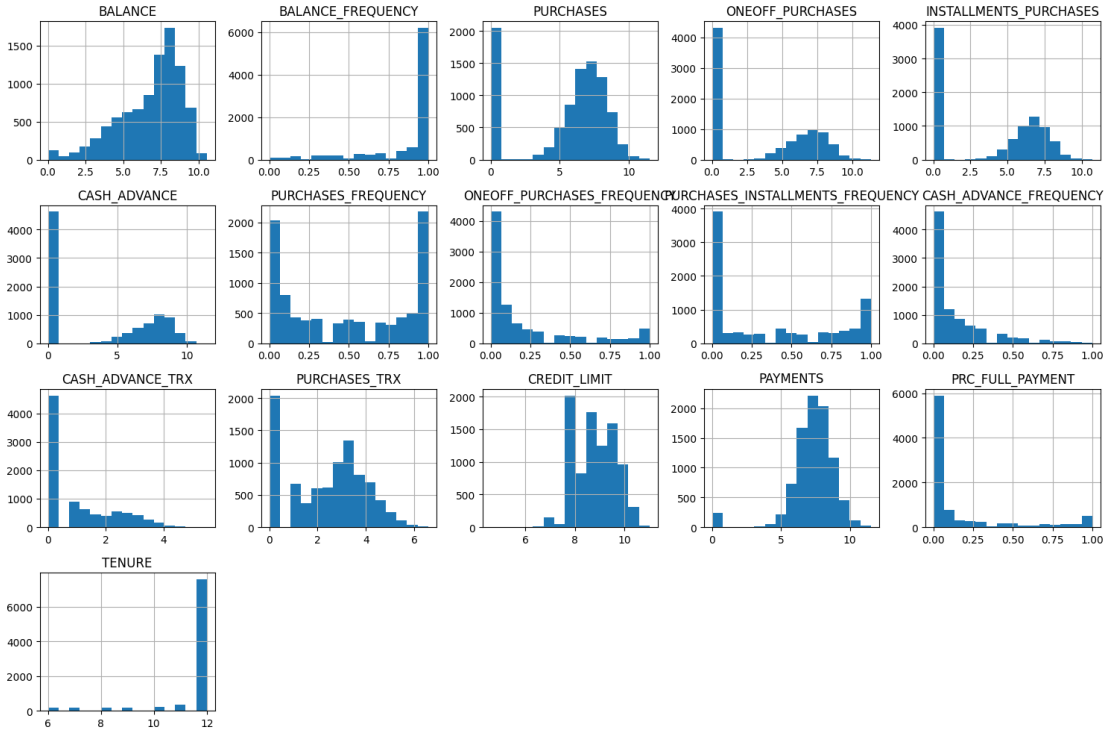


Figure 3.1: Transformed Feature Histograms

Our second method of exploratory data analysis was creating a correlation matrix (figure 3.2). The correlation matrix suggests strong relationships between certain credit card behaviors. Notably, `CASH_ADVANCE` and `CASH_ADVANCE_TRX` are highly correlated, indicating that users who frequently take cash advances also tend to have higher cash advance amounts. `PURCHASES` and `ONEOFF_PURCHASES` also show a strong positive correlation, suggesting that individuals who make one-time, large-value purchases contribute significantly to the total purchase volume. In contrast, a notable negative correlation between `BALANCE` and `PRC_FULL_PAYMENT` implies that customers who pay off their balances are likely to maintain lower overall account balances. These findings can offer financial institutions actionable insights into spending patterns and risk profiles. While these values give us ideas about linear relationships between the variables, we will look at deeper connections by applying clustering algorithms.

4 Feature Engineering and Standardization

To prepare our data for machine learning, we first standardized the features using StandardScaler. This step is crucial for algorithms like K-Means, which rely on distance calculations. This scaling will also be beneficial for data visualization at the end, since PCA requires the data to be standardized as well. Note that our clustering algorithms will not run on the principal components of the data, but the actual high-dimensional, standardized dataset.

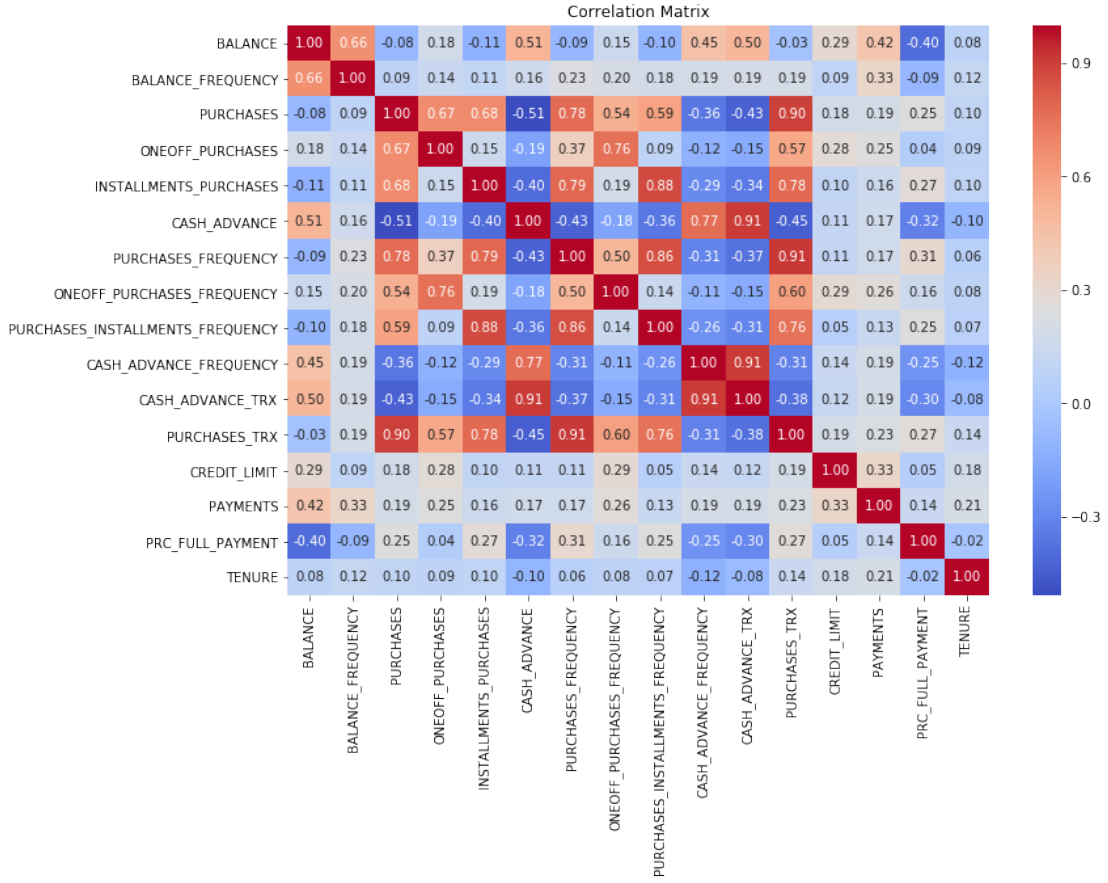


Figure 3.2: Correlation Matrix

5 PCA FOR VISUALIZATION

To get more of a sense of the data, we performed PCA in order to visualize the data in two dimensions. As noted earlier, when we use our algorithms, we will *not* be using the principal components, but rather the dataset with all its dimensions in full. This PCA is just to visualize the data, both as exploration and after the clusters have been assigned. To further emphasize that note, figure (5.1) shows the variance captured by each principal component; just using the first two principal components only accounts for around 50% of the variance—not enough for meaningful conclusions to be extracted.

For our graph with two principal components, figure (5.2), there appears to be a dense area in the lower left that could suggest one cluster. The spread along the PC1 axis shows a change in density from left to right which might indicate another cluster. The upper part of the plot seems to have a slightly lower density, hinting at another possible cluster. So we expect at least 3 clusters.

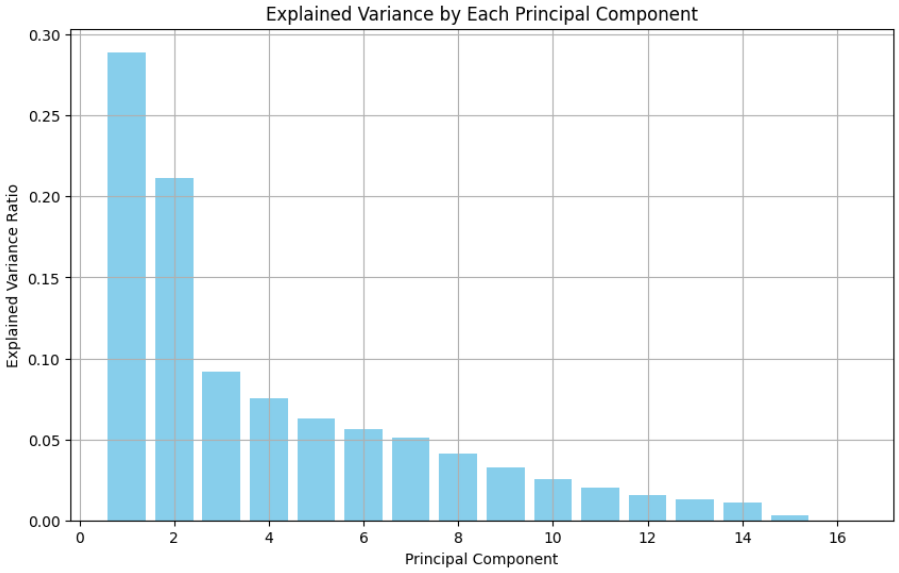


Figure 5.1: Explained Variance by Each Principal Component

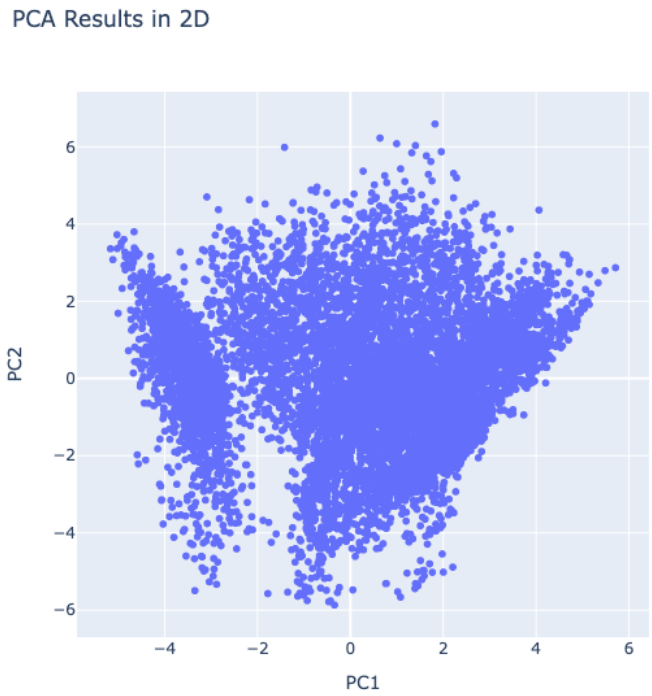
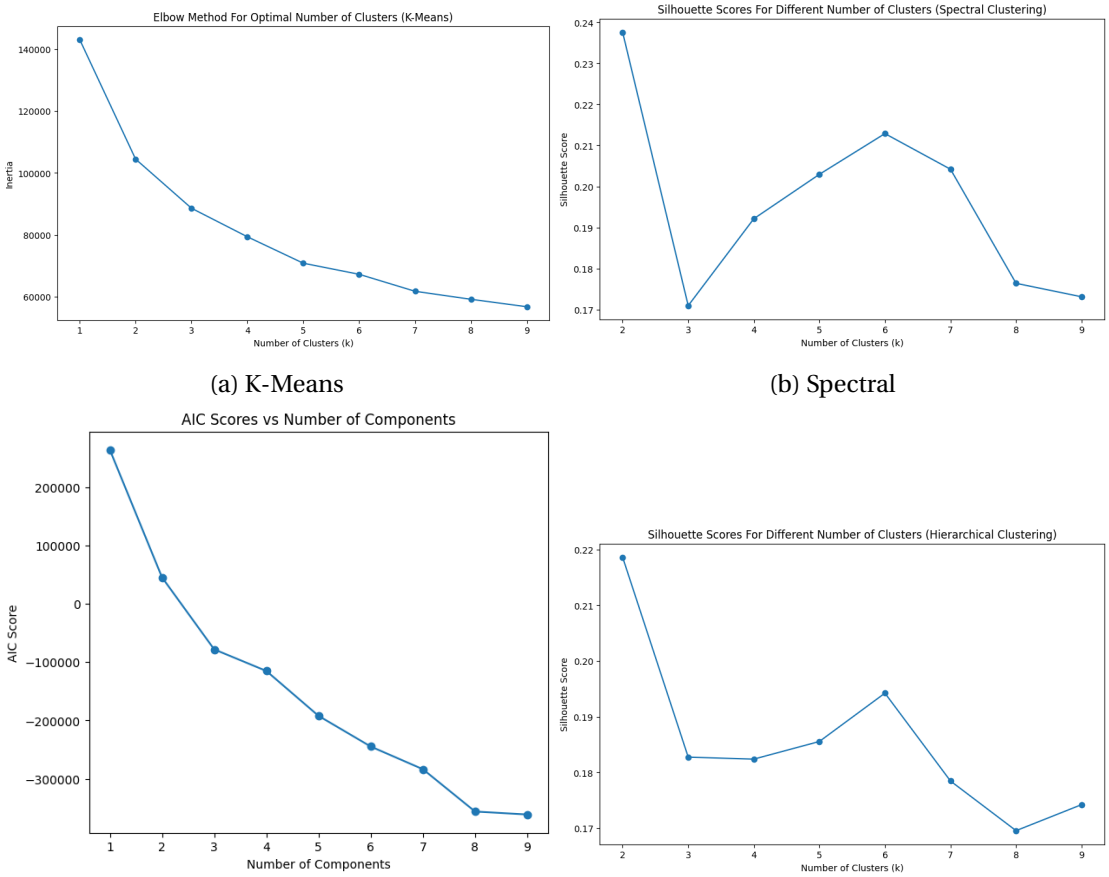


Figure 5.2: 2D Visualization (PCA)

6 Determine the Number of Clusters

We denote k to be the number of clusters of our credit card user data. To determine the optimal k between 2 and 10, we consider measures like inertia, for K-Means; silhouette score, for Spectral Clustering and Hierarchical Clustering; and Akaike Information Criterion (AIC), for Gaussian Mixture Models. These choices will be discussed further.



(a) K-Means

(b) Spectral

(c) GMM

(d) Hierarchical

Figure 6.1: Choosing the Optimal k

- **K-Means Clustering**
  - Inertia in KMeans clustering measures the coherence of clusters, representing the sum of squared distances between data points and their nearest centroid. Minimizing inertia aligns with KMeans' objective to form tight clusters, so this measure is specific to the K-Means algorithm.
  - An elbow plot helps determine the optimal number of clusters by showing where the decrease in inertia lessens with increasing k. At this 'elbow' point, the marginal

benefit of adding more clusters reduces, indicating the balance between cluster tightness and model simplicity.

  - The silhouette score and Akaike Information Criterion (AIC) are not typically used for KMeans. For silhouette score, while it assesses how well data is clustered, it can be computationally intensive for large datasets and does not always provide a clear cut-off for the number of clusters like the inertia in the elbow method does. As for AIC, it is designed for models based on maximum likelihood estimations, which does not apply to K-Means.
  - For the given plot, the elbow is identified at $k = 5$, guiding us to select five clusters for an optimal K-Means model.

- **Spectral Clustering**
  - When implementing this method, we used K-Nearest Neighbors to identify similarity. After testing with multiple values for number of neighbors, we decided to use 10, since that value balanced higher silhouette scores and cleaner visualizations in 2 and 3 dimensions (from PCA).
  - Silhouette score evaluates the consistency within clusters compared to between clusters. It is more flexible and can be used with any clustering method, Spectral Clustering included.
  - The inertia and Akaike Information Criterion (AIC) are not typically used for Spectral Clustering because, with regard to each metric, Spectral Clustering does not seek to minimize the variance within clusters like for K-Means, and Spectral Clustering does not use maximum likelihood estimation to capture the distribution of the data.
  - From the silhouette score plot, despite the initial rise for $k = 2$, which could suggest a significant separation between two broad groups, the further increase for $k = 6$ indicates that within those two broad groups, there are more refined subgroups that provide an even better clustering solution according to the silhouette score metric. Therefore, 6 is the optimal number of clusters for Spectral CLustering.

- **GMM Clustering**
  - For Gaussian Mixture Models (GMM), the Akaike Information Criterion (AIC) helps in selecting a model that fits the data well while penalizing for increasing complexity to avoid overfitting. The optimal number of components (k) is usually where the AIC is lowest because it suggests the model is complex enough to fit the data well but not so complex that it's fitting the noise.
  - Inertia and silhouette scores are not typically used for determining the number of clusters in GMM. The former because it is specific to K-Means. While a silhouette score *can* be used, it is generally more common to take advantage of the fact that GMM utilizes probability distributions, so metrics like AIC and BIC are better fitted to the type of model, moreso than the more generic silhouette score.

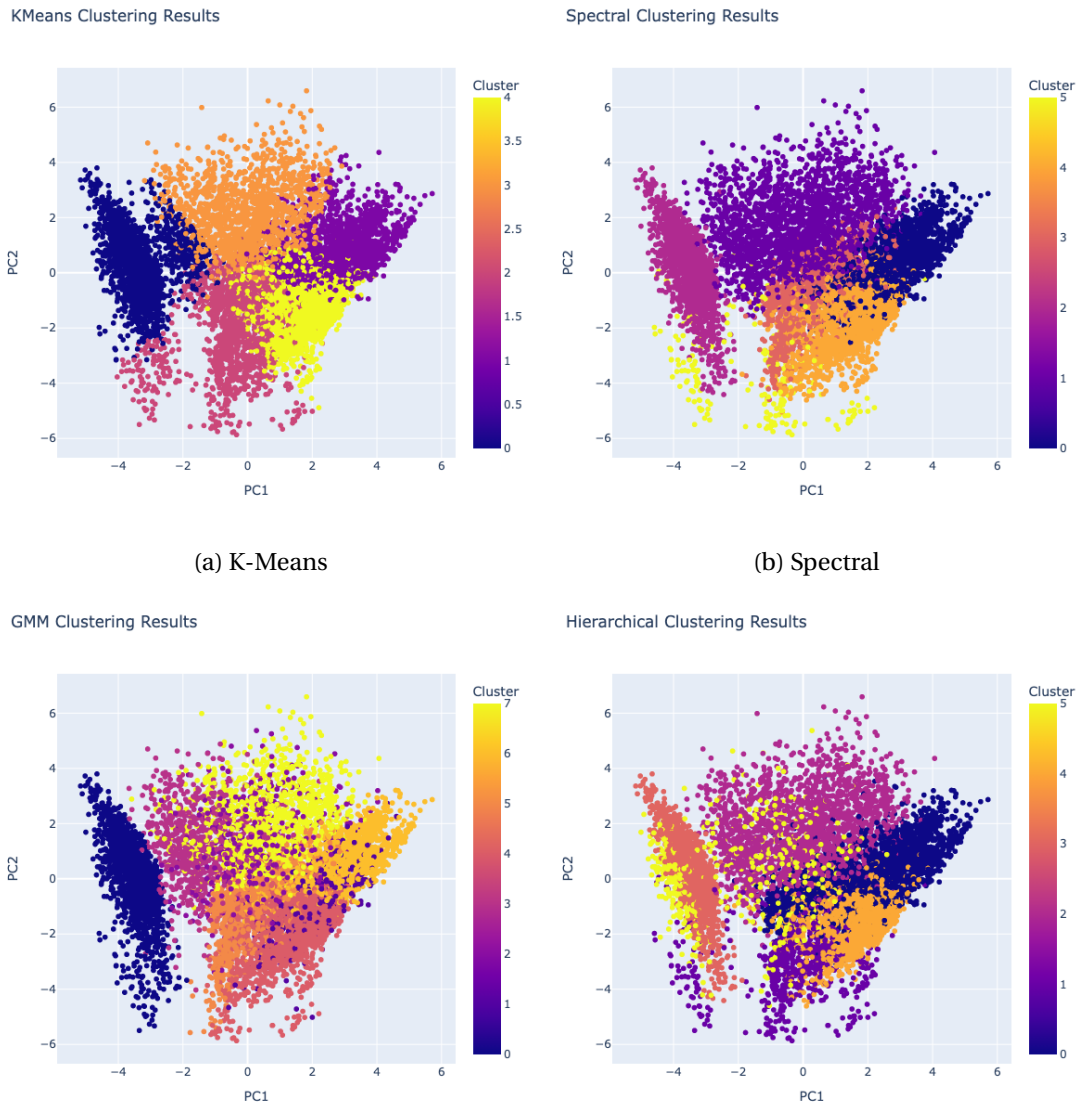
- From the plot, the AIC curve appears to be flattening as it approaches $k = 8$, suggesting diminishing returns in model improvement with the addition of more components. While an argument could be made for $k = 9$, we feel that the AIC values are close enough between 8 and 9 that we feel comfortable choosing the lesser of the two. Therefore, we choose $k = 8$ as the optimal number of clusters for the GMM.

- **Hierarchical Clustering**
  - As with Spectral Clustering, the optimal number of clusters (k) according to the silhouette score plot should be chosen based on the highest value of the silhouette score, which indicates the best balance between within-cluster cohesion and between-cluster separation.
  - We use silhouette score here for the same reasons as for Spectral Clustering—inertia is specific to K-Means and Hierarchical Clustering does not employ maximum likelihood estimation in its algorithm.
  - While $k = 2$ shows the highest silhouette score, indicating strong separation and cohesion at this number of clusters, we opt for the second highest score of $k = 6$ over $k = 2$ since we believe (from visualizing the data with PCA) that there should be at least 3 clusters.

7 CLUSTERING

We utilized multiple clustering algorithms and their corresponding optimal number of clusters to find the best cluster representation for the data. After determining the clusters for each point, we performed PCA to reduce dimensions to 2 and visualized the data with cluster labels. We will choose our preferred clustering method based on qualitative analysis of the graphs (figure 7.1). We will pay particularly close attention to how well the clustering algorithm handled the body of points on the left which separate from the main body relatively significantly; we will call this group the left offshoot.

- **K-Means Clustering**
  - The clusters in the graph appear to have non-spherical shapes, which K-Means had difficulty handling the data because it works on the assumption that a cluster can be defined by a centroid and that all points closest to that centroid belong to the same cluster.
  - The boundaries between some of the clusters appear to be non-linear, whereas K-Means only works well with linear cluster boundaries.
  - The K-Means algorithm had trouble with the left offshoot, splitting it it two pieces and joining those pieces in clusters which include the main body of points.



(a) K-Means

(b) Spectral

(c) GMM

(d) Hierarchical

Figure 7.1: Different clustering results

- **Spectral Clustering**
  - As mentioned earlier, our Spectral Clustering algorithm used K-Nearest Neighbors to determine similarity, with 10 being the number of neighbors.
  - The clusters shown in the graph appear to be relatively compact, which implies that spectral clustering has done a good job at maximizing the intra-cluster similarity and minimizing the inter-cluster similarity, which are key goals in cluster analysis.

  - The left offshoot shows a distinctly separated cluster compared to the main body, which is indicative of strong performance. The main body of the data shows some—but not much—overlap, especially compared to GMM and Hierarchical Clustering. This overlap can most likely be attributed to the fact that the clustering was performed in high-dimensional space then projected onto a 2-dimensional plane, leading to a loss of information.

- **GMM Clustering**
  - The GMM algorithm performs surprisingly well when differentiating the left offshoot from the main body of data points.
  - However, the GMM clustering appears to have difficulty establishing clear separation between clusters in the main body of the plot, as evidenced by the noticeable intermixing of colors.

- **Hierarchical Clustering**
  - The visualization of the hierarchical clustering results indicates that the algorithm may have had difficulty in segregating the data into well-defined, distinct groups. The clusters as depicted do not exhibit clear boundaries, and there is a noticeable blend between different groups.
  - Comparing these messy 6 clusters to the neater 6 from Spectral Clustering highlights just how poorly this clustering algorithm performed when segmenting the data.

In conclusion, after reviewing the results of Hierarchical Clustering, Gaussian Mixture Models (GMM), Spectral Clustering, and K-Means applied to the credit card dataset, **Spectral Clustering emerges as the most suitable method for this particular analysis** based on the visual inspection of the scatter plots. It has demonstrated the ability to effectively capture the inherent groupings within the data, yielding well-delineated and coherent clusters that align with the natural distribution of the data points in the reduced principal component space. GMM, though it differentiated the left offshoot fairly well, shows a higher degree of overlap between clusters. Hierarchical clustering seems less effective for this dataset, as it produces overlapping clusters without clear boundaries. K-Means, with its assumption of spherical clusters, imposes somewhat arbitrary divisions that likely do not represent the true underlying structure of the data.

8 Inter-Cluster Analysis (for Spectral Clustering)

To analyze customer behavior within the clusters, we used the principal components to determine which factors contribute the most to the trends seen in the clustering. In particular, we prioritized the three most influential factors from the first principal component and the top two factors from the second principal component. We note that we determine the "most influential" features for principal component i to mean the columns with the largest absolute coefficient in the linear combination that defines the i-th principal component. This is a valid method of selection since all variables were standardized, so difference in scale is irrelevant. These components were selected based on their significant contribution to the dataset's variance, which is indicative of their potential to offer valuable insights into the data's structure.

With respect to each cluster (labeled as cluster 0 through 5, see horizontal axis of Figure 8.1), we dissect the behavior of the following factors to discern underlying patterns and relationships: `PURCHASES`, `PURCHASES_TRX`, `ONEOFF_PURCHASES`, `CASH_ADVANCE`, and `CASH_ADVANCE_FREQUENCY`. These 5 features were closely analyzed given their prominence in the PCA results.

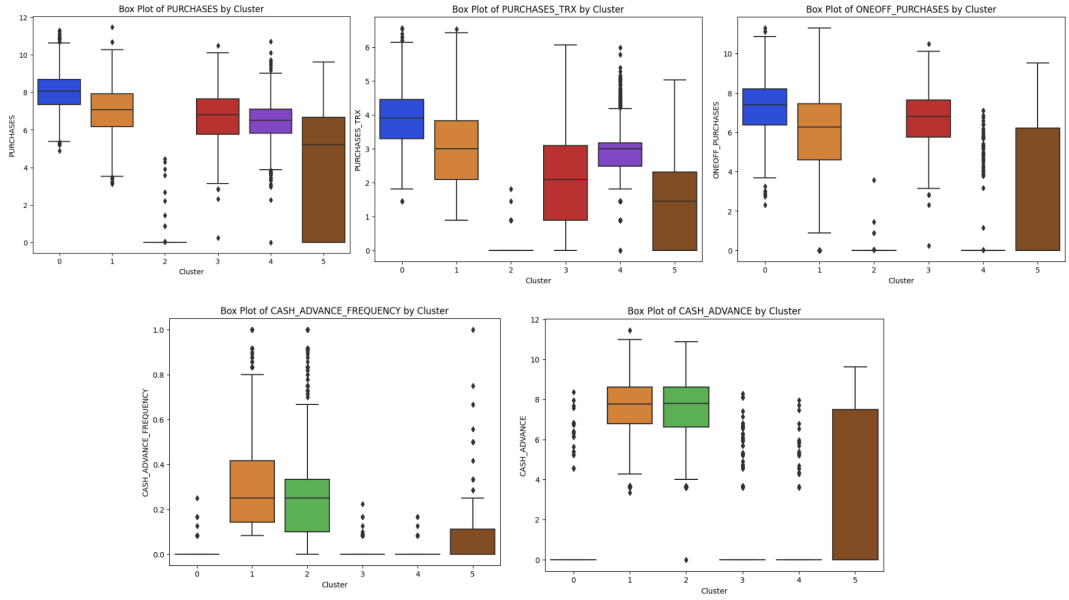


Figure 8.1: Box Plots by Clusters: PURCHASES, PURCHASES_TRX, ONEOFF_PURCHASES, CASH_ADVANCE_FREQUENCY, CASH_ADVANCE

9 Cluster Analysis Conclusions

Based on the box plots for the top five variables, namely `PURCHASES_TRX`, `ONEOFF_PURCHASES`, `CASH_ADVANCE`, `PURCHASES`, and `CASH_ADVANCE_FREQUENCY`, the characteristics of the users in each of the six clusters can be summarized as follows (clusters are labeled as cluster 0 through 5, see figure 9.1):

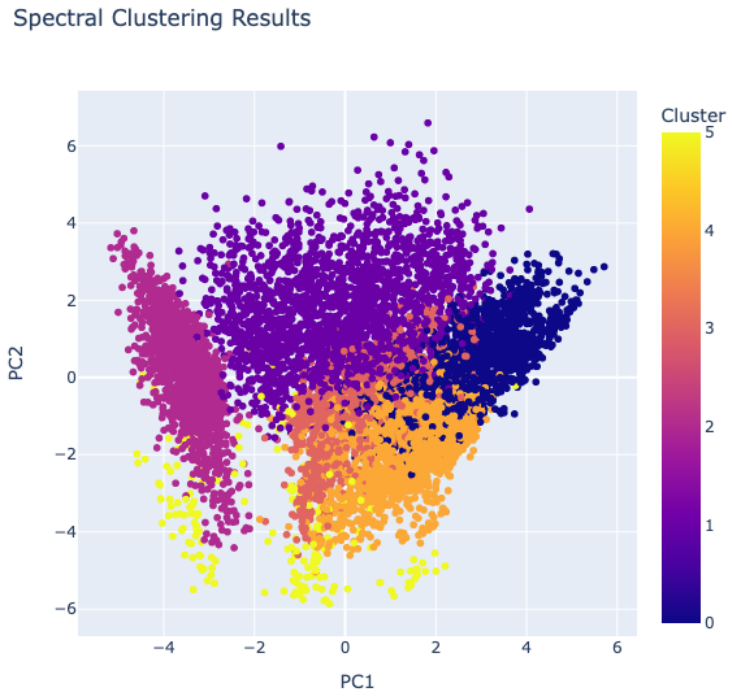


Figure 9.1: Spectral Clustering Results

Cluster 0: *High-Spenders with Direct Purchasing Power (dark blue)*

This group seems to be heavy spenders with the highest average purchases. They conduct frequent transactions and have a strong inclination towards one-off purchases, suggesting they may engage in significant, possibly discretionary, spending. Their use of cash advances is almost negligible, which could indicate a healthy financial status or a preference for direct purchases over cash borrowing.

Cluster 1: *Moderate Shoppers with Cash Advance Needs (dark purple)*

Members of this cluster have moderate levels of purchases and transactions but are notable for their relatively high use of cash advances. This suggests a group that, while they do make purchases, might also be relying on credit for cash liquidity. Their need

for cash advances could reflect a specific lifestyle or the management of cash flow gaps. Individuals in this cluster would likely be considered high-risk credit investments.

Cluster 2: *Cash Advance Reliants with Minimal Purchases (pink)*

This is a unique segment with virtually no purchasing activity but a high frequency of cash advances. It may represent individuals who primarily use their credit for cash rather than direct purchases, possibly due to personal preference or because they are in a tight financial situation where liquid cash is a necessity. While the use of cash advance in this group indicates relatively higher credit risk, their low volume and quantity of purchases could indicate less risk compared to cluster 1.

Cluster 3: *Selective Big-Ticket Purchasers (dark orange)*

Users in this cluster have a moderate level of purchasing, characterized almost exclusively by one-off transactions, which might suggest infrequent but substantial spending. The very low cash advance activity can indicate that these users are less likely to borrow against their credit line, perhaps due to better financial planning or sufficient liquidity.

Cluster 4: *Conservative Credit Users (bright orange)*

This cluster's users engage in lower levels of purchasing and transaction activity with almost no one-off purchases or cash advances. This could be indicative of conservative spenders who use credit sparingly and are cautious about accruing debt. This type of credit card user would likely carry the least associated credit risk of all the groups.

Cluster 5: *Occasional Buyers with Periodic Cash Needs (yellow)*

Individuals in this cluster have lower purchasing amounts and fewer transactions, with some tendency towards one-off purchases and a moderate frequency of cash advances. The members of this group show the widest variety in credit use practices, exhibiting a relatively wide interquartile range for each of the five features. This group might represent individuals with occasional significant expenses who occasionally rely on cash advances, potentially pointing to an inconsistent cash flow or the need to cover unexpected expenses.

10 Limitations and Future Questions

10.1 Limitations

In our project, dimensionality reduction was accomplished using Principal Component Analysis (PCA), where we made the deliberate choice to use only two principal components for visualization purposes. While this enabled us to generate clear, two-dimensional plots that can be easily interpreted by a non-technical audience, it also imposed a significant constraint: a substantial portion of the variance within the dataset was not captured, which could potentially omit important characteristics of the data.

For a more nuanced analysis, future studies should consider incorporating more principal components, specifically aiming for five to account for approximately 80% of the data's variance. This would likely provide a richer, more detailed understanding of the dataset, though at the cost of more complex visualizations that may be challenging to interpret for those outside the field of data analysis.

Moreover, our approach to hyperparameter selection in the clustering process relied on heuristic methods. To enhance the rigor of our analysis, subsequent research could employ systematic techniques for hyperparameter tuning, such as grid search or random search, as well as Bayesian optimization. These methods could refine the selection of model parameters, leading to potentially improved clustering results.

We also acknowledge that our initial strategies for algorithms like K-Means and GMM may have affected the outcome due to their sensitivity to starting conditions. Therefore, exploring various initialization techniques and different distance or affinity metrics in clustering algorithms could yield insights for particular types of data.

In the context of GMM, a deeper examination of the choice of covariance structure for the Gaussian components is warranted. Different covariance types can dramatically impact the clustering results, and the current choice may not have been optimal.

Lastly, implementing cross-validation techniques could offer a measure of the stability and robustness of the clustering solutions across different data subsets.

By addressing these limitations, future work can build upon our findings, employing more advanced, automated tuning and validation techniques to enhance both the quality and the applicability of the clustering models.

10.2 Future Questions

- *Can the identified customer segments be further analyzed to predict susceptibility to credit card fraud?*

  Future research could explore how the customer segments identified in our project might exhibit different behaviors or patterns that could indicate susceptibility to credit card fraud. This would involve analyzing transactional behaviors, spending patterns, and account activity within each segment to identify markers that are predictive of fraud.

- *Was the method of Spectral Clustering the optimal way to categorize credit card consumer behavior?*

  While spectral clustering yields a satisfactory result when looking at the 2-dimensional principal component representation the data set, no one could conclude that it is the best approach.

  - Given that the first two principal components (PCs) already encompass the essential features we aimed to analyze, and are sufficient for a general categorization of customers, a 2D format was intentionally chosen for simplicity in visualization. Calculations indicate that a 6D PCA would capture over 80% of the variance in the original data. Although spectral clustering demonstrates the most striking results in 2D PCA among the four methods used in this project, it may not be the

most effective in higher dimensions. Future research could benefit from applying higher-dimensional analysis for a more detailed examination of customer behaviors. Furthermore, exploring other clustering methods not mentioned in this study could also yield valuable insights into the data set.

  - Spectral clustering, as an unsupervised learning technique, results in outcomes that are less predictable and controllable compared to supervised methods. Researchers using spectral clustering analyze the resulting clusters, and attempt to infer the underlying reasons for their formation. While this method can reveal hidden patterns in the data, it requires careful interpretation to ensure meaningful insights, as the clusters may not always align with initial hypotheses. Future research could examine data within the same field but with slightly different aspects, yielding a completely different conclusion.

- *How might integrating transaction time-series data enhance the evaluation of risk or detection of fraudulent patterns within each customer segment?*

  Integrating transaction time-series data could provide a more dynamic and comprehensive view of customer behavior. This approach would allow for the analysis of transaction patterns over time, potentially uncovering periods of high credit risk (e.g. high spending and high cash advance) or subtle, irregular patterns indicative of fraudulent activity. This information could provide dynamic context for what is otherwise static data, uncovering even more behavioral patterns within the customer groups.

- *How does the introduction of geographic and demographic data enhance the understanding of each segment's unique characteristics?*

  Future research could explore the enhancement of customer segmentation by integrating geographic and demographic data. This would involve analyzing how factors like location, age, income, and education level contribute to differing consumer behaviors and preferences within each segment. Such an approach could unveil more nuanced segment distinctions, improving the effectiveness of targeted marketing strategies and product development. It's also essential to consider how these additional data layers might intersect with existing behavioral and transactional data to provide a more holistic understanding of customer profiles. While our dataset was relatively neutral to demographic and geographic features, including that information in the data could discern more patterns but then raises concerns about the ethical use of that information.

11 Final Remarks

In conclusion, this study showed us significant insights regarding credit card user behavior, showcasing the diverse spending and payment patterns among customers. Understanding insights into consumer behavior can help companies target advertisements to the appropriate groups of people. Financial institutions would also benefit from these findings, as the delineated customer behaviors inform credit risk for the companies. Though the cluster labels may overgeneralize the different populations of credit card users, whose behaviors may vary

between time, the developments of this project allow for a basic understanding of how credit card users can be grouped. More complex analysis involving integrating additional variables and other data analysis techniques—perhaps other clustering methods or dimensionality reduction techniques—would also help to analyze sub-groupings of user behavior. Furthermore, if the data had a variable indicating fraudulent activity or the user's credit score, future work could integrate some form of classification or regression algorithm to predict those feature values.