CLUSTERING AND FITTING REPORT ON CREDIT CARD USERS

Name: Vishal Kumar Senthil Kumar

Student Number: 23076841

Github:
https://github.com/vishalkumar041298/clustering_and_fitting/tree/main

Introduction

This report explores the relationships, distributions, and clustering patterns present in the dataset of credit card customers. By analyzing key features such as the average credit limit, total number of credit cards, and the frequency of visits to the bank or online services, the goal is to identify meaningful patterns in customer behavior. The analysis involves correlation statistics, regression modeling, and clustering methods, with the results visualized to provide clarity on customer segmentation.

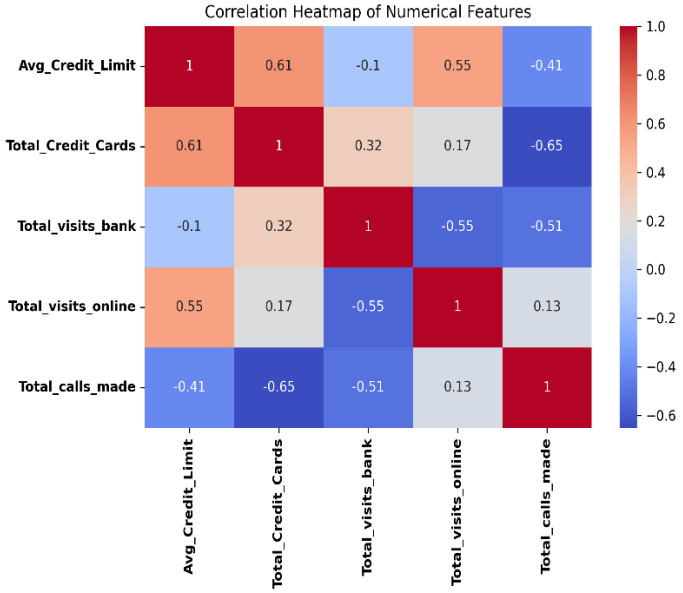

Figure 1: Correlation Heatmap

The correlation heatmap reveals the relationships between key numerical features in the dataset. The following are key correlations:

- **Avg_Credit_Limit vs. Total_Credit_Cards (0.61):** This positive correlation suggests a moderate relationship between a customer's credit limit and the number of credit cards they hold. As the average credit limit increases, the total number of credit cards tends to increase as well, though the correlation is not perfect.
- **Total_Credit_Cards vs. Total_calls_made (-0.65):** A negative correlation between the total number of credit cards and the total number of calls made suggests that customers with more credit cards are less likely to make calls. This could imply that customers with multiple credit cards are more self-sufficient or have access to digital services that reduce the need for calling customer support.
- **Total_visits_online vs. Total_visits_bank (-0.55):** A moderate negative correlation between online visits and bank visits indicates that customers who engage more frequently with online banking are less likely to visit the bank physically, pointing towards a shift in consumer behavior towards digital banking.

This heatmap provides valuable insights into how different variables are related, with moderate to strong correlations observed primarily between **Avg_Credit_Limit and Total_Credit_Cards.**

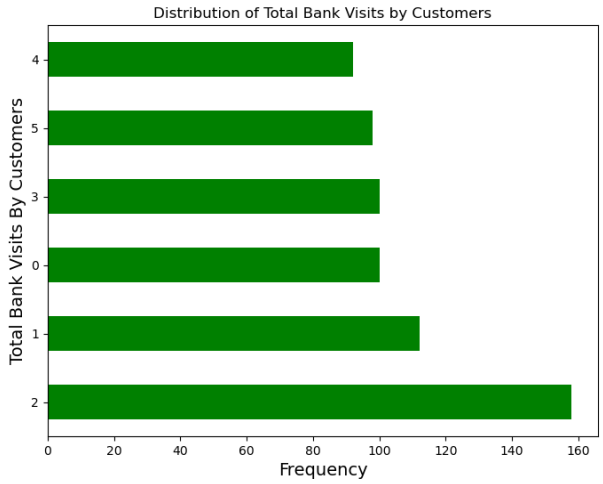

Figure 2: Distribution of Total Bank Visits by Customers

The bar plot displaying the distribution of total bank visits by customers reveals the following:

- **Highest frequency at 2 visits:** Most customers have visited the bank two times, which could suggest that customers tend to make few in-person visits and rely on other channels for banking needs.
- **Other visit frequencies (1, 4, 5):** The frequency of visits decreases as the number of visits increases, with relatively few customers making 4 or 5 visits. This further suggests that a significant proportion of the customer base prefers minimal interaction with physical bank branches.

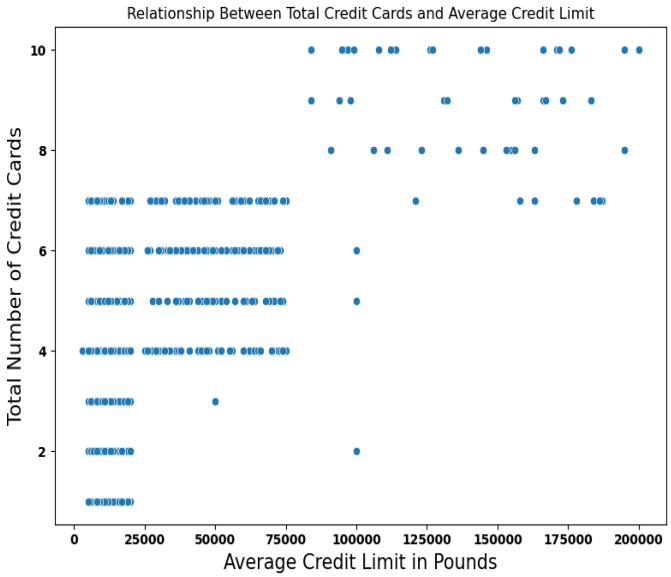

Figure 3: Relationship between Total Credit Cards and Average Credit Limit

This scatter plot reinforces the positive correlation observed in the heatmap. Each point represents a customer, and the plot shows that higher average credit limits tend to correlate with more credit cards. However, there are some key insights:

- **Clusters of points:** While the general trend is positive, there are clusters of points where customers with low credit limits have multiple credit cards. This suggests that other factors may be influencing credit card ownership, aside from just the credit limit.
- **Outliers:** A few outliers with high credit limits but a lower number of credit cards might represent a niche group of customers who, despite having access to significant credit, choose to hold fewer cards.

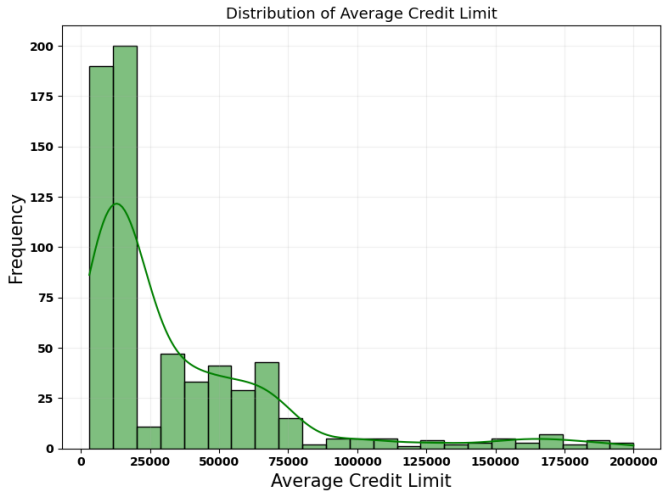

Figure 4: Distribution of Average Credit Limit

The histogram of the Avg_Credit_Limit shows a highly skewed distribution, with many customers having relatively low credit limits, typically under 25,000. This distribution is right-skewed, indicating that while most customers are in the lower credit limit range, a smaller proportion holds much higher limits. Specifically:

- **Peak at lower values (0 to 25,000):** The concentration of customers with low credit limits suggests that many credit card holders are in the consumer segment with limited credit.
- **Long tail to the right:** The tail extending toward higher credit limits suggests the presence of a small group of customers with significantly higher credit limits. This indicates a potential segmentation between average customers and premium customers.

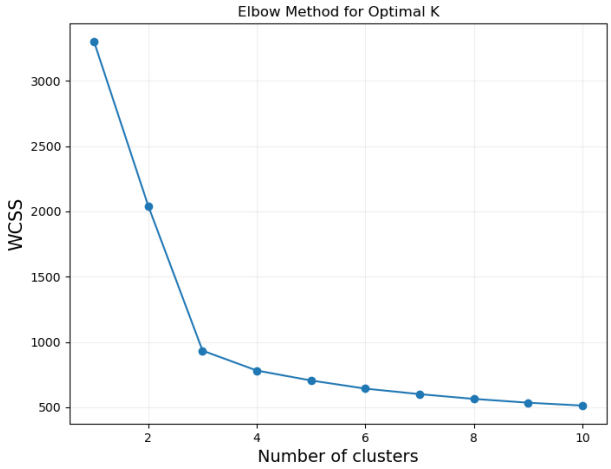

Figure 5: Elbow Curve to obtain K

The Elbow Method is used to identify the ideal number of clusters for K-Means clustering. The plot of the Within-Cluster Sum of Squares (WCSS) shows a significant drop in WCSS from K=1 to K=3, after which the rate of decrease slows down.

- **K=3 is optimal:** Based on the plot, K=3 is the optimal number of clusters, as the WCSS drops sharply until this point and then levels off. This suggests that clustering the data into three groups provides the most meaningful segmentation of customers without overfitting.
- **WCSS behavior:** The flattening of the curve after K=3 indicates diminishing returns with additional clusters, implying that three clusters are sufficient to capture the underlying patterns in the data.

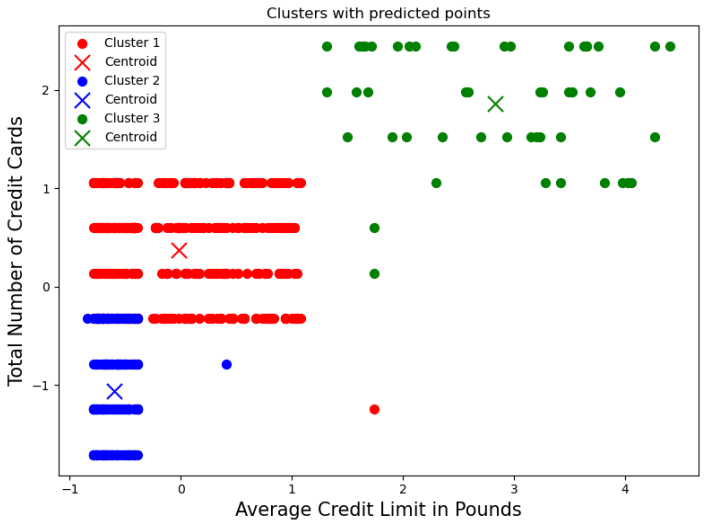

Figure 6: Clustering with predicted points

This plot shows the results of K-Means clustering with three clusters, based on the features Avg_Credit_Limit and Total_Credit_Cards. The clusters are clearly visible with different colors, and the centroids are marked with large crosses.

- **Cluster 1 (red):** This cluster likely represents customers with lower credit limits and fewer credit cards. The centroids suggest that customers in this group have average credit limits close to zero.
- **Cluster 2 (blue):** This cluster represents customers with moderate credit limits and a medium number of credit cards. The centroid of this group indicates that these customers have a slightly higher credit limit than Cluster 1.
- **Cluster 3 (green):** Customers in this cluster have the highest credit limits and hold the most credit cards, as seen by the position of the centroid and the spread of data points.

The clustering analysis provides valuable segmentation, which can be used for targeted marketing, personalized offers, or financial services aimed at each cluster's specific needs.

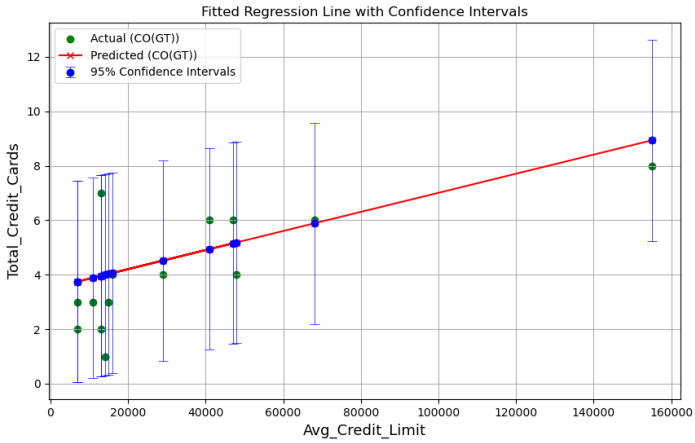

Figure 7: Fitted Regression Line

This plot displays the regression results for predicting **Total_Credit_Cards based on Avg_Credit_Limit**. The red regression line shows a positive trend, indicating that as the average credit limit increases, so does the total number of credit cards.

- **Confidence intervals:** The 95% confidence intervals, represented by blue error bars, are relatively narrow, indicating that the model's predictions for Total_Credit_Cards are reasonably precise for most values of Avg_Credit_Limit. The intervals widen slightly at higher credit limits, but overall, the model appears to fit the data well.
- **Slope of the regression line:** The positive slope indicates a direct relationship between credit limits and the number of credit cards. This suggests that customers with higher credit limits tend to hold more credit cards, which could be a characteristic of wealthier or more credit-worthy individuals.

This analysis supports the idea that credit limit is a strong predictor of the number of credit cards a customer holds.