Name: Sandeep Kumar

Student No: 23087422

Github: https://github.com/sandeep1993kumar04/clustering-and-fitting.git

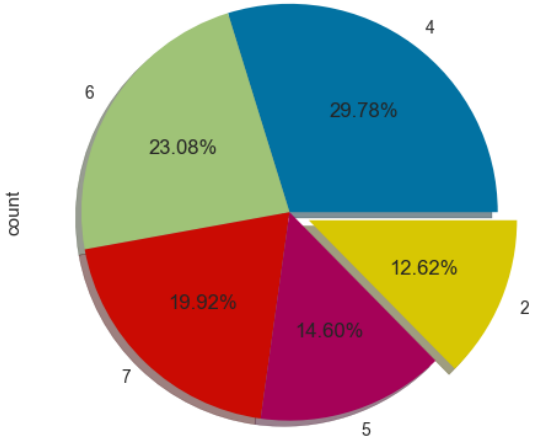
Assignment: Clustering and Fitting

Credit Card Customer Data Analysis

This document presents an analysis of the Credit Card Customer dataset. The dataset includes variables such as average credit limit, total credit cards, total visits to the bank, online visits, and calls made to customer service. The analysis uses linear regression to predict the average credit limit based on other features.

This **pie chart** represents the distribution of categories, likely labeled 2, 4, 5, 6, and 7, with percentages indicating their share of the total.

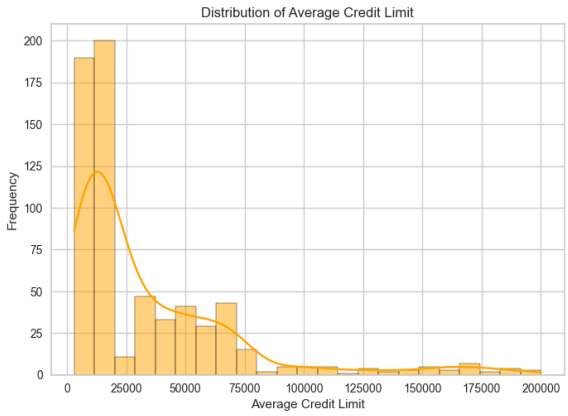
- **Category 4**: Largest share (29.78%).
- **Category 2**: Smallest share (12.62%).
- Other categories (6, 7, and 5) have moderate shares, with Category 6 contributing 23.08%.



This histogram shows the **distribution of average credit limits**:

- Most customers have an average credit limit between **0 and 25,000**.
- The distribution is **right-skewed**, with fewer customers having higher credit limits beyond **75,000**.
- A **density curve** overlays the histogram, highlighting the peak and the gradual decline in frequency.

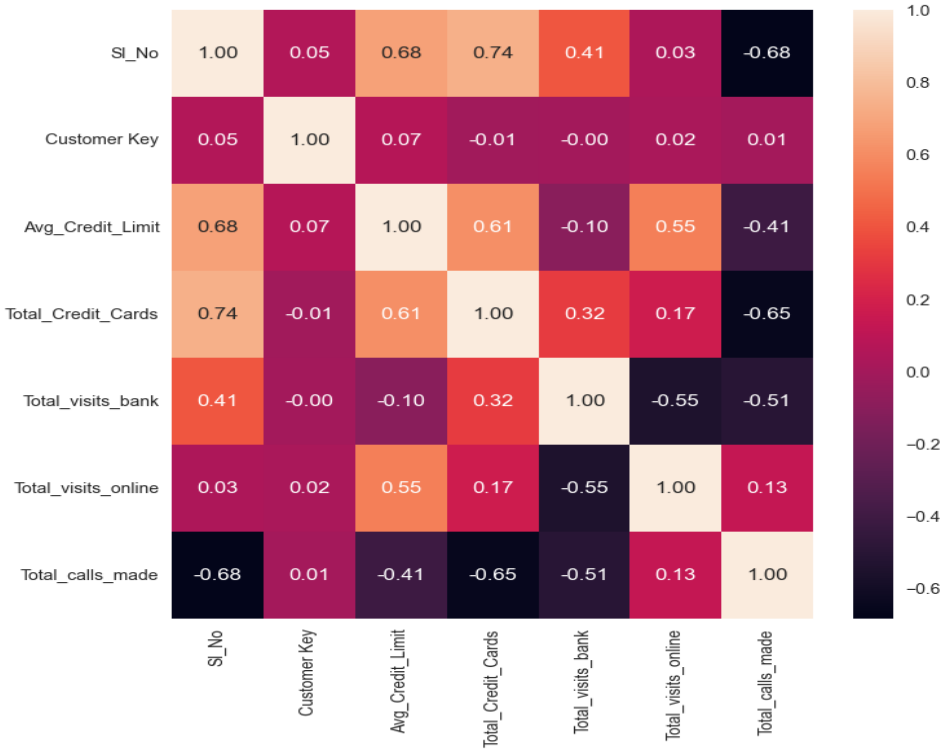
This plot provides insight into the concentration of customers with lower credit limits.



This correlation heatmap highlights:

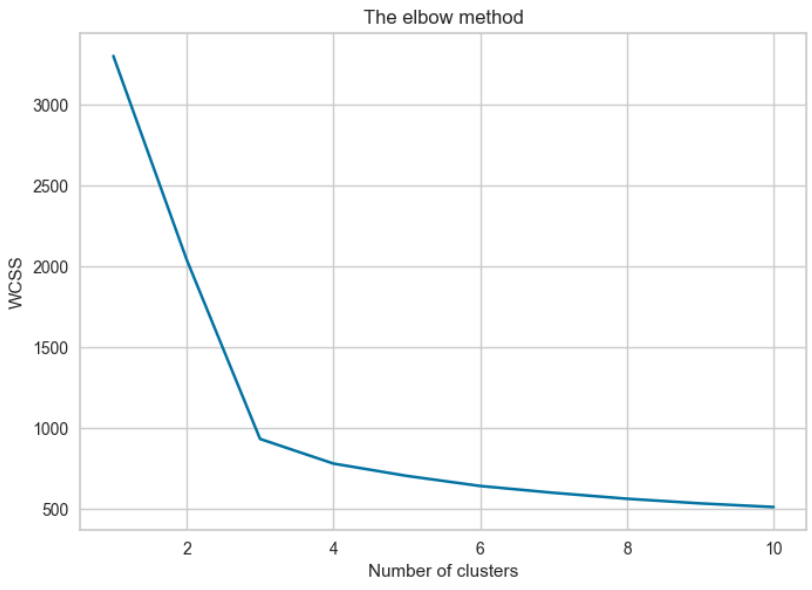
- **Avg_Credit_Limit** is positively correlated with **Total_Credit_Cards (0.61)** and **Total_visits_online (0.55)**, but negatively with **Total_calls_made (-0.41)**.
- **Total_calls_made** shows negative correlations with most features, including **Avg_Credit_Limit** and **Total_visits_bank**.

Strong correlations suggest key predictors of credit limit.



This elbow method plot shows WCSS vs. number of clusters.

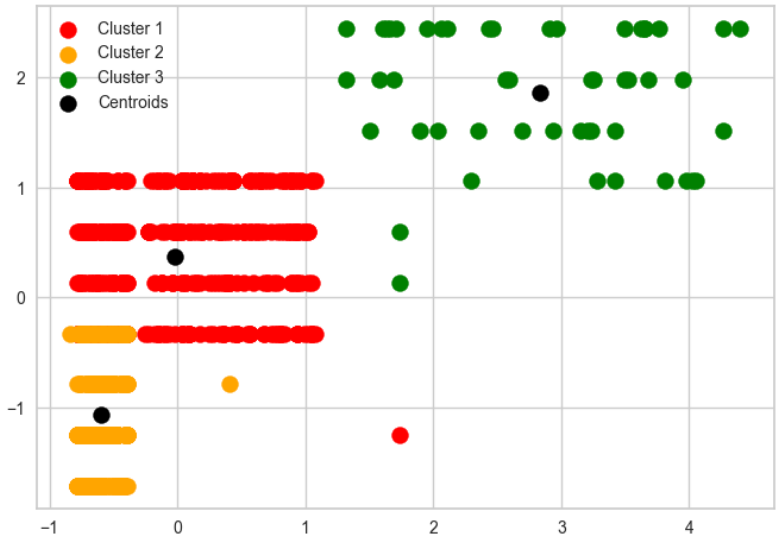
The "elbow" at **3 clusters** indicates the optimal number of clusters, where adding more clusters yields diminishing returns.



K-means clustering with three clusters:

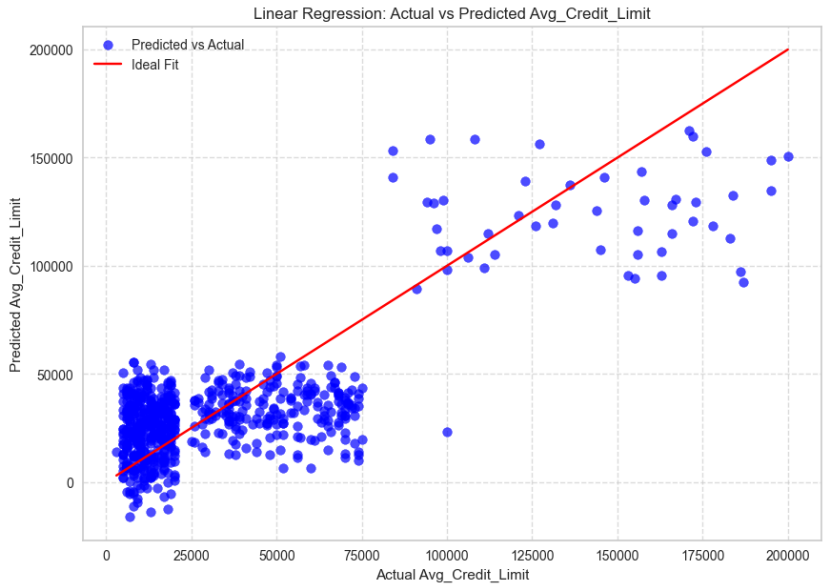
- **Cluster 1 (Red)**, **Cluster 2 (Orange)**, and **Cluster 3 (Green)** are groups of data points.
- **Black dots** represent the centroids (cluster centers).
- The x and y axes are likely scaled feature values.

It shows how the data is grouped, with Cluster 2 being compact and Cluster 3 more dispersed.



Linear Regression: Actual vs Predicted

The scatter plot below shows the predicted average credit limit against the actual values. The red line represents an ideal fit.



Model Evaluation

Mean Squared Error (MSE): 544315166.60

R^2 Score: 0.6149