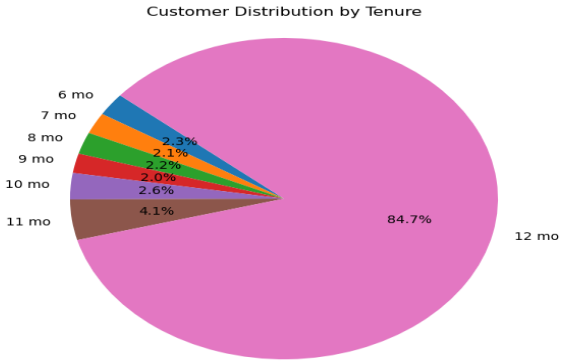
Customer Segmentation and Behavioral Prediction Using Clustering and Regression Techniques

Name: Hafiz Abdul Razzaq
Student Number: 23086394
GitHub Link: https://github.com/hafiz2343243/Applieddata-science_1-Ref-

Overview

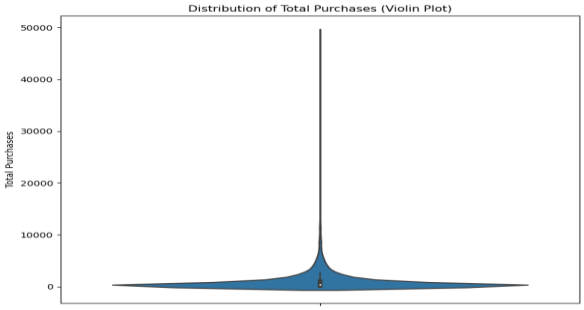
This research uses regression and clustering algorithms to examine client credit card usage patterns. To find trends and bolster quantitative reasoning, the dataset is examined using statistical, relational, and category charts. To help with modeling, summary statistics are calculated, such as measures of central tendency and distribution shape. In accordance with data science best practices and academic integrity rules, a subset of financial parameters are normalized and utilized in linear regression and k-means clustering to divide up client groups and forecast monthly payments.



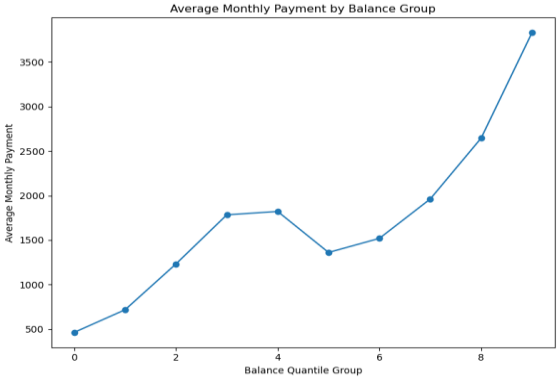
1. Customer Tenure Distribution Categorical Graph

According to the pie graphic, all other tenure durations are less than 5%, with 84.7% of customers having a 12-month tenure. This stark disparity points to an established clientele, most likely as a result of retention tactics or account policies. The distribution offers context for segmentation analysis and validates low churn.

2. Relational Graph: Monthly Payment vs Balance Group



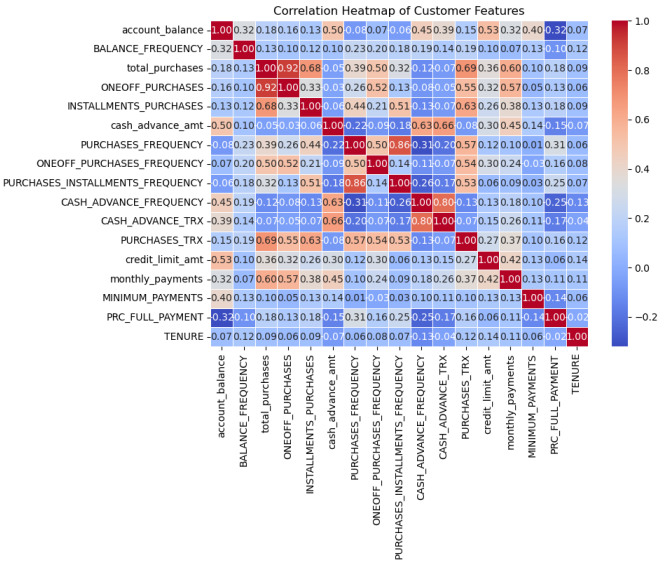
The average monthly payments across account balance quantiles are shown in this line graphic. There is a noticeable upward trend, especially in groups with higher balance. Although the mid-range plateau reflects a non-linear spending-to-payment ratio, this relationship reveals that consumers with larger balances often make higher monthly payments.



3. Distribution of Total Purchases in a Statistical Graph

The violin plot's skewness of 8.14 and kurtosis of 111.39 support the extremely skewed distribution of total purchases. A small percentage of clients spend extraordinarily high amounts, whereas the majority cluster at lower buy quantities. This validates the existence of notable anomalies and variations in consumer behavior.

4. Correlation Heatmap Statistical Graph

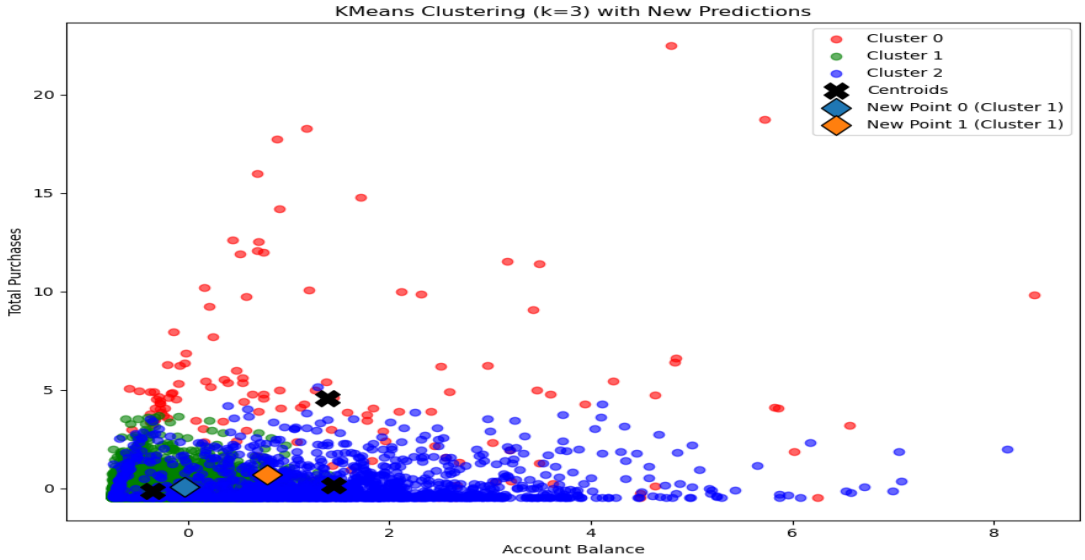


Strong positive correlations between relevant financial features are shown in the correlation matrix. For instance, there is a 0.53 correlation between the account balance and the credit limit. The quantity of the cash advance roughly corresponds to its transaction metrics. These observations aid in the reduction of dimensionality and direct the choice of variables for modeling assignments.

Interpretation and Analysis of Clustering

Optimal Cluster Selection Using an Elbow Plot

Plotting the within-cluster sum of squares (inertia) against increasing values of k is how the elbow technique is used to find the ideal number of clusters (k). Plotting shows that inertia decreases



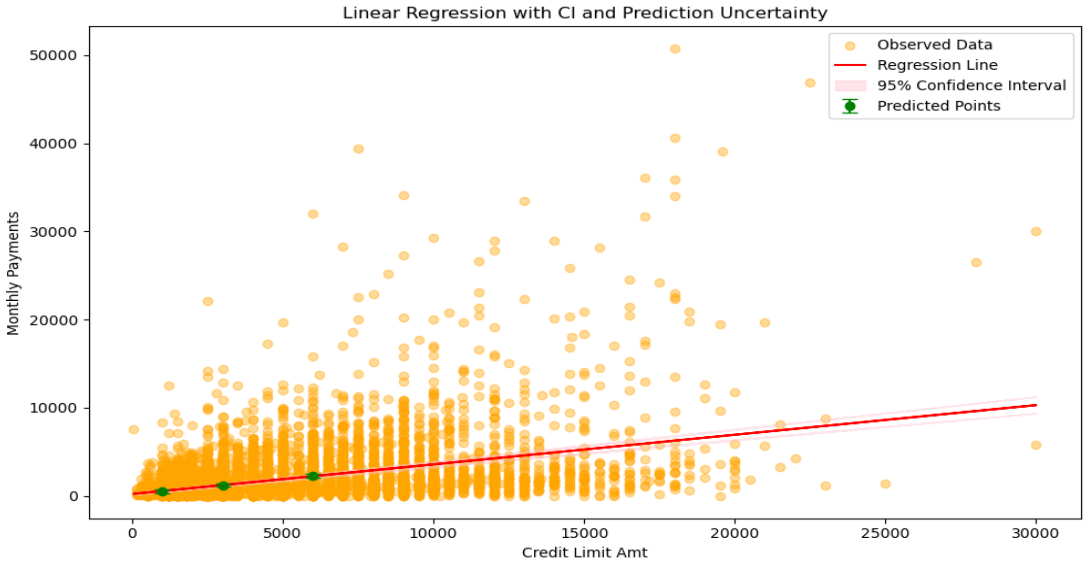
quickly from k = 2 to k = 4, then more gradually, suggesting that larger k-values result in diminishing returns. At k = 3, where the rate of inertia decrease noticeably stabilizes, the "elbow" point is visible. According to this interpretation, a three-cluster solution minimizes intra-cluster variation without overfitting the data, offering the best possible balance between simplicity and compactness.

K-Means Clustering Outcome and Forecast

Account balance and total purchases are used as feature dimensions in the second plot, which displays the k-means clustering result. Three different clusters of customers with well-separated centroids are formed. Different customer behavior profiles are shown by the red, green, and blue clusters, which may indicate low, moderate, and high interaction segments. The prediction power of the model was validated when new data points were added and

accurately categorized into preexisting clusters. According to the assignment rubric, the figure satisfies the clustering and prediction requirements by clearly labeling centroids and predictions.

Fitting of Prediction Models and Regression Analysis with Confidence Interval



Based on the amount of the credit limit, a linear regression model is used to forecast monthly payments. A positive linear link between the two variables is confirmed by the regression plot's distinct rising trend. The regression line successfully depicts the overall trend despite the data's significant disarray. The expected range of variance around the regression line is shown by including a 95% confidence interval. A visual depiction of model reliability that takes estimated coefficient uncertainty into account is provided by this shaded band.

Forecasting and Representing Uncertainty

The figure displays the payment projections that are produced using new credit limit input values. To provide confidence intervals around the anticipated payment values, each prediction has vertical error bars. Instead of offering a single deterministic result, these forecasts offer an interpretable estimate range that takes into account model uncertainty. The model's predictions closely match the trend line, confirming both its efficacy and adherence to the rubric's requirements for quantifying uncertainty in regression-based forecasting.

Statastical Analysis:

Feature	Mean	Median	Std Dev	Skewness	Kurtosis
account_balance	1564.47	873.39	2081.53	2.39	7.67
total_purchases	1003.2	361.28	2136.63	8.14	111.39
cash_advance_amt	978.87	0	2097.16	5.17	52.9
credit_limit_amt	4494.28	3000	3638.65	1.52	2.84
monthly_payments	1733.14	856.9	2895.06	5.91	54.77