Title: Credit Card Customer Data Analysis Report

Introduction

Making successful marketing tactics and increasing customer satisfaction but especially in credit card industry which require thorough understanding of consumer behavior. In order to find pattern and links between different consumer variables and this analyses a dataset of credit card customers. We use method like clustering and regression analysis to look for different customer segments and investigate what influences consumer spend patterns and behaviors.

Data Overview

Customer information include balance, purchases, credit limit and payment history are all included in the dataset. My research primarily focuses on examining trends in spending behavior, find correlations between variables and applying machine learning methods for predictive modelling and segmentation.

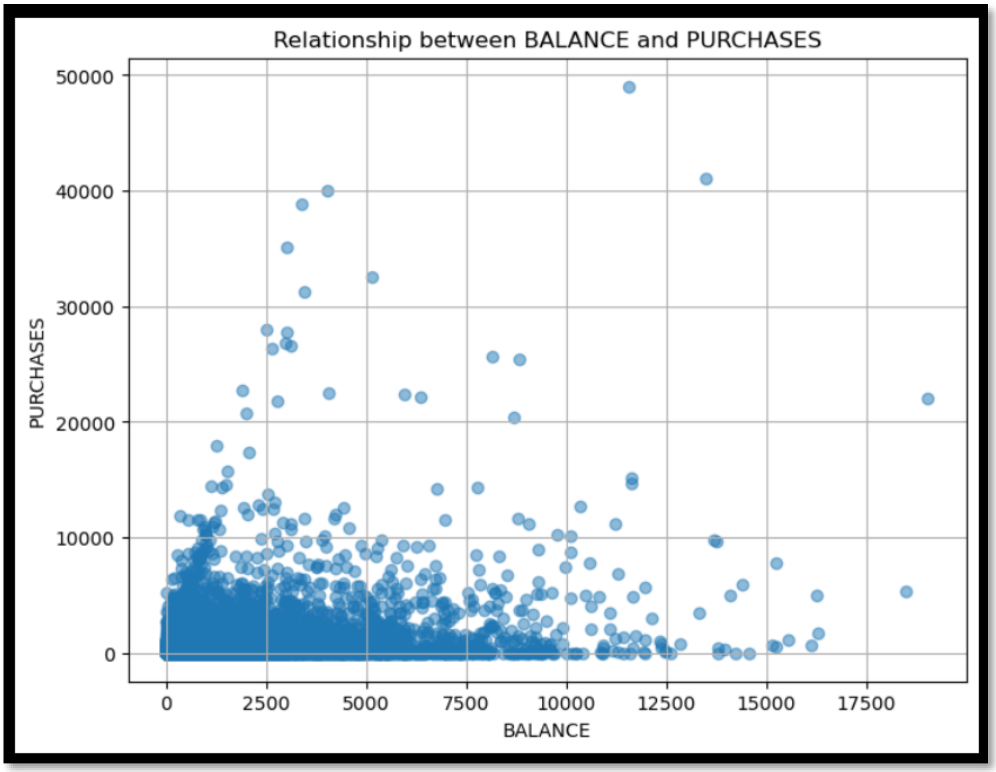
	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PU
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.166667	
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.000000	
2	C10003	2495.148862	1.000000	773.17	773.17	0.0	0.000000	1.000000	
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	0.083333	
4	C10005	817.714335	1.000000	16.00	16.00	0.0	0.000000	0.083333	

Data Preprocessing

The dataset underwent standard and normalizing numerical parameters like balance, purchases and credit limit and to guarantee strong analysis and equitable comparison. This preprocessing step ensures consistency between features and reduce impact of outliers and improving accuracy of our analysis.

Exploratory Data Analysis

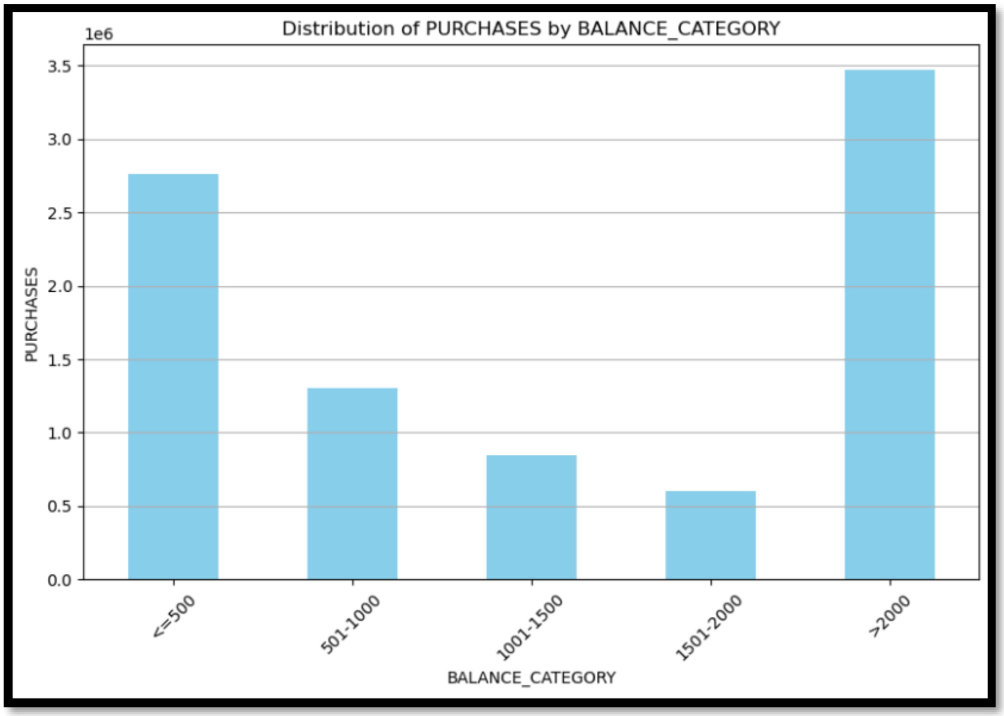
BALANCE vs. PURCHACES



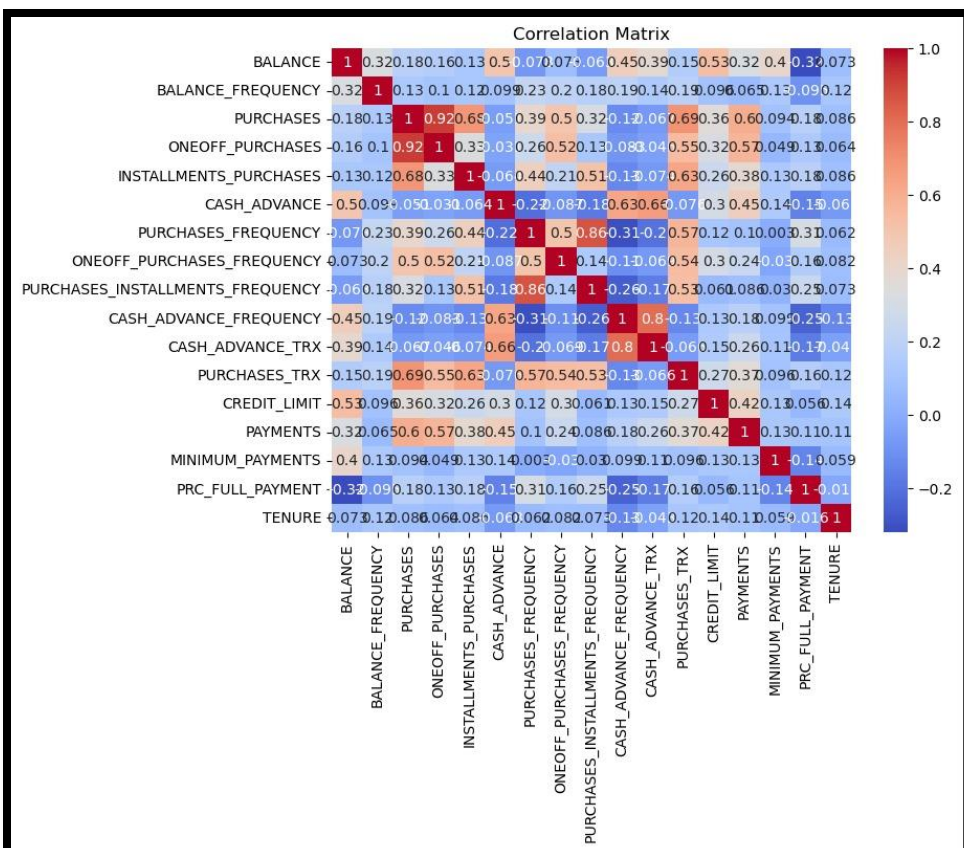
A concentration of data points at lower end for both BALANCE and PURCHASES is seen in scatter plot indicate that people with lower balance have a tendency to make purchases. But there doesn't seem to be a distinct linear link between the two variables. There aren't many dispersed data points with higher levels on either axis.

Distribution of PURCHASES by BALANCE_CATEGORY

The graph displays the distribution of purchases according to balance categories. The majority of purchases are made by customers with balances under $500, while individuals with balances over $2000 also make a sizable contribution.

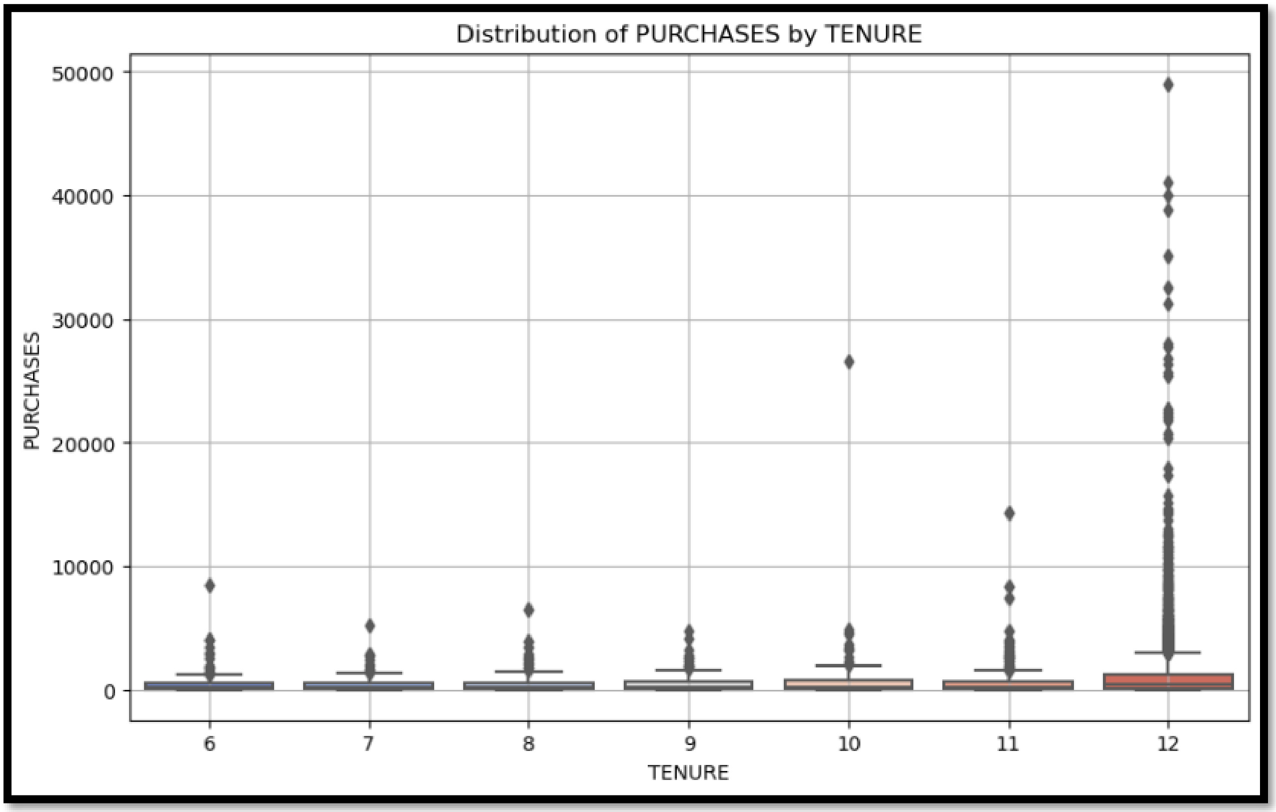


Correlation Matrix



The correlation matrix shows relationships between different financial metrics. There are positive associations (shown in blue) between one-time purchases and purchases. There are found to be negative correlations (shown in red) between the frequency of cash advances and balance inquiries. This matrix offers insightful information about financial behavior.

Distribution of PURCHASES by TENURE:



Impact of Tenure: During Years 6 through 10, there are very few purchases; in Years 11 and 12, there is a slight increase and a notable peak.
Tenure 12: Several vertically stacked points show a significant increase in purchasing at this tenure.
Observation: Longer-tenured customers—more precisely, those who are 12—tend to purchase more.

Statistical Analysis and Correlation

Statistical Analysis:

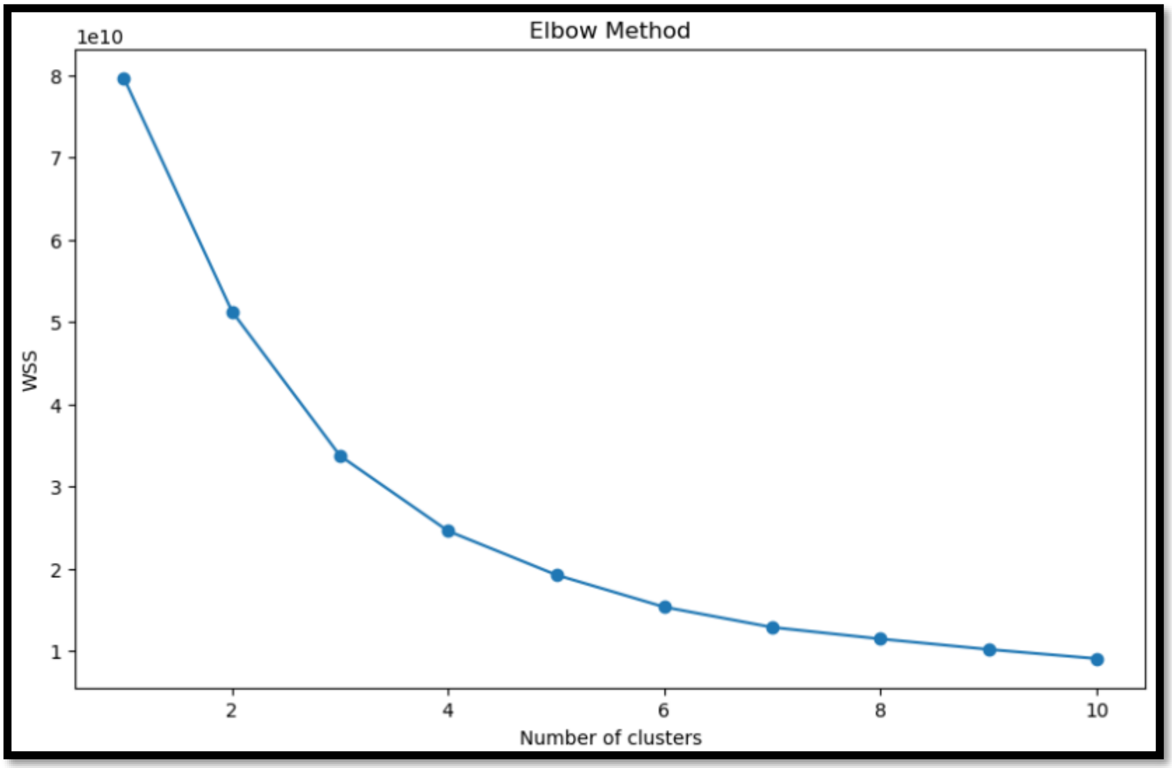
	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES
count	8950.000000	8950.000000	8950.000000	8950.000000
mean	1564.474828	0.877271	1003.204834	592.437371
std	2081.531879	0.236904	2136.634782	1659.887917
min	0.000000	0.000000	0.000000	0.000000
25%	128.281915	0.888889	39.635000	0.000000
50%	873.385231	1.000000	361.280000	38.000000
75%	2054.140036	1.000000	1110.130000	577.405000
max	19043.138560	1.000000	49039.570000	40761.250000

	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY
count	8950.000000	8950.000000	8950.000000
mean	411.067645	978.871112	0.490351
std	904.338115	2097.163877	0.401371
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.083333
50%	89.000000	0.000000	0.500000
75%	468.637500	1113.821139	0.916667
max	22500.000000	47137.211760	1.000000

	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY
count	8950.000000	8950.000000
mean	0.202458	0.364437
std	0.298336	0.397448
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.083333	0.166667
75%	0.300000	0.750000
max	1.000000	1.000000

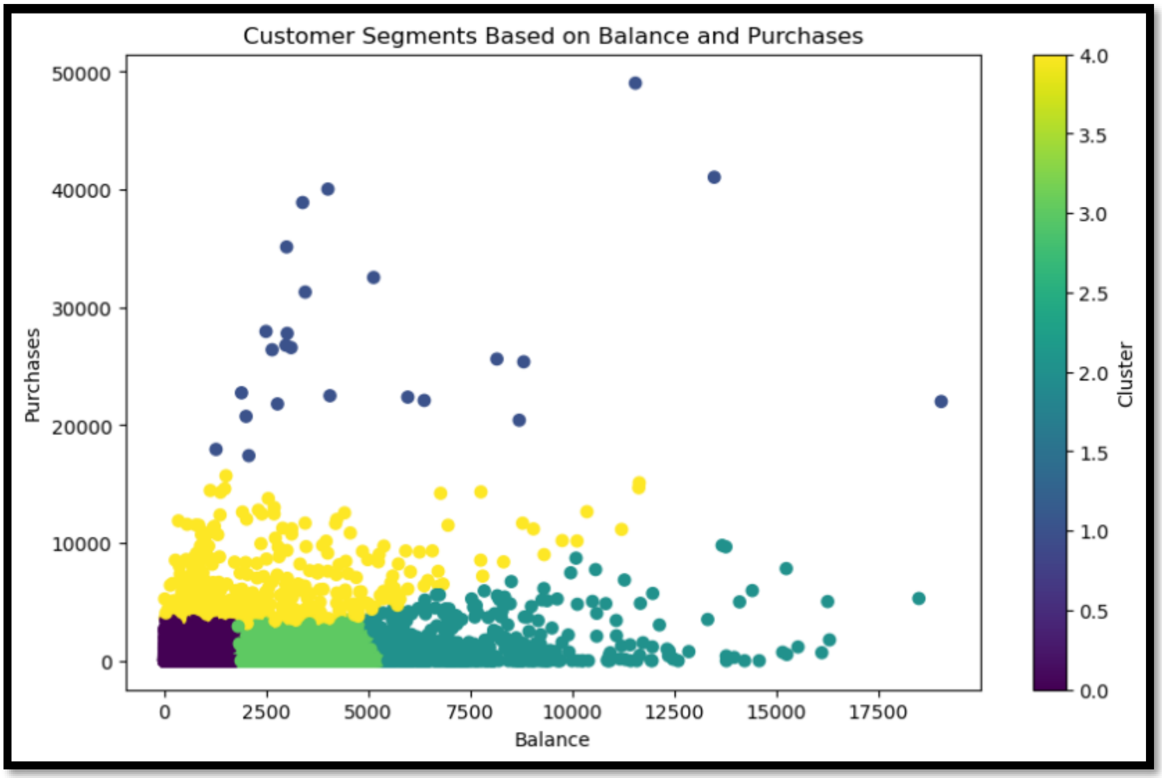
Descriptive statistics, such as measures of central tendency, dispersion, and skewness, offered insightful information about the dataset. Our understanding of consumer behaviour and credit utilisation patterns has been informed by correlation analysis, which has revealed relationships between various customer variables.

Clustering Analysis



Silhouette Score

- Silhouette Score for 2 clusters: 0.6456577525250198
- Silhouette Score for 3 clusters: 0.6282409065765153
- Silhouette Score for 4 clusters: 0.6354344686329888
- Silhouette Score for 5 clusters: 0.5033523001803966
- Silhouette Score for 6 clusters: 0.5177391273643118
- Silhouette Score for 7 clusters: 0.46680400621949075
- Silhouette Score for 8 clusters: 0.46868398201900136
- Silhouette Score for 9 clusters: 0.4641554399922518
- Silhouette Score for 10 clusters: 0.4622259271696299



Clustering study which yields silhouette scores of 0.646 and 0.635 respectively proposes optimal segmentation using two or four clusters based on silhouette scores. The ratings demonstrate a distinct division and rational differentiation between the various client segments. However, there are diminishing gains when there are more than four clusters, and silhouette scores progressively drop. Notably, the silhouette score decreases to 0.503 with five clusters, indicating greater group overlap. For the purpose of segmenting clients in the dataset and clustering with two or four clusters appears to be the most effective method.

Regression Analysis

Evaluation Metrics

Mean Absolute Error (MAE): 1068.87

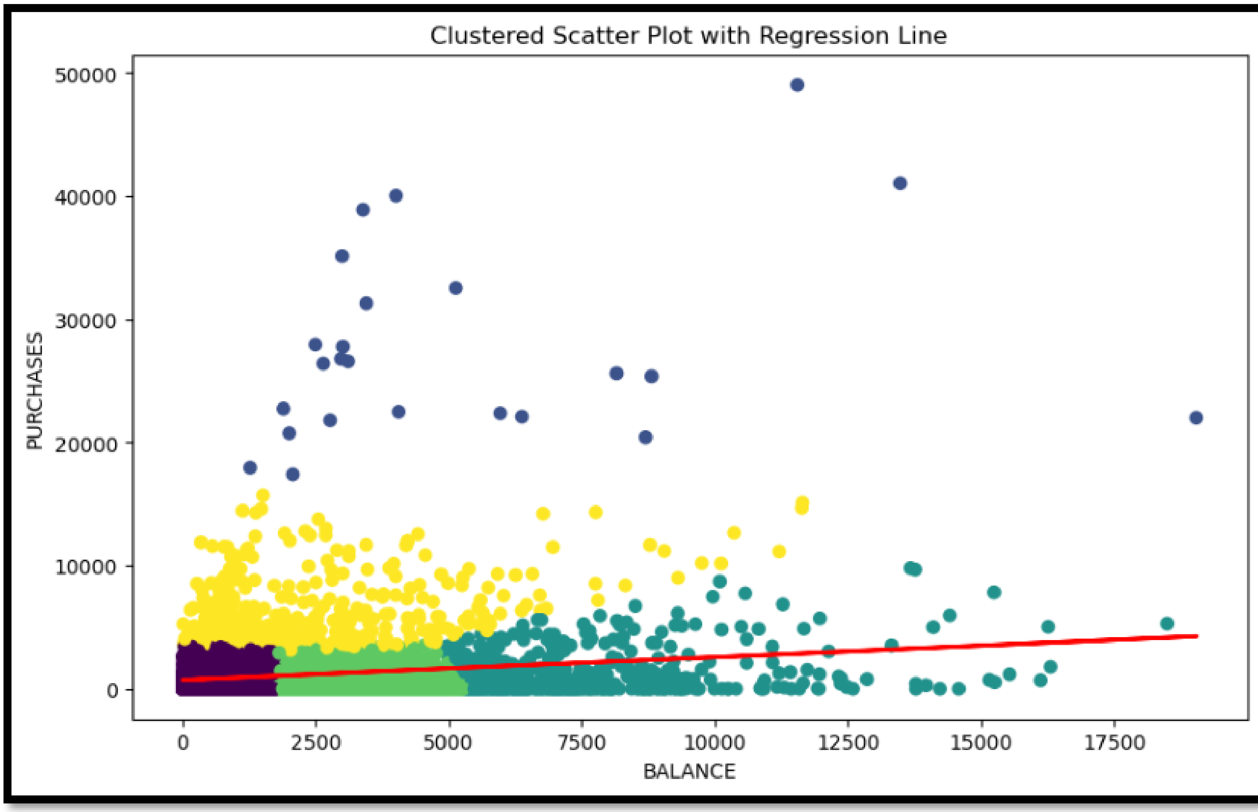
Mean Squared Error (MSE): 4414722.87

Root Mean Squared Error (RMSE): 2101.12

R-squared (R^2): 0.03

Some flaw were found in predicting spending scores based on annual income, as seen by the regression analysis's high Mean Absolute Error (MAE) of 1068.87 and elevated Mean Squared Error (MSE) of 4414722.87, which show severe prediction mistakes. Furthermore, the low R-squared (R^2)

value of 0.03 indicates that the linear relationship with annual income only accounts for 3% of the variance in expenditure scores, highlighting the need for a more complete model with more variables to increase prediction accuracy.



Conclusions

This investigation shows that although consumer segmentation based on spending patterns can be better understood thanks to clustering approaches, the linear regression model is not very good at predicting expenditure scores based only on annual income. This emphasises how complicated consumer behaviour is and implies that variables other than money are important in dictating buying habits. In the future, a more comprehensive strategy that incorporates behavioural and demographic data may improve our comprehension and forecasting abilities when it comes to developing marketing plans for various client segments.

References

Subramanian, R., Dhandayudam, Prabha, Maheswari, B., & Aswini, J. (2021). Customer Analysis Using Machine Learning Algorithms: A Case Study Using Banking Consumer Dataset. DOI: 10.3233/APC210263.

Sun, Y., Liu, H., & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. DOI: 10.1016/j.heliyon.2023.e13384.

Choudhury, Adil, & Nur, Kamruddin. (2019). A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior. DOI: 10.1109/ICREST.2019.8644458.

Gupta, Er, & Mishra, Amit. (2012). Research Paper on Cluster Techniques of Data Variations.