

# DESCRIPTIVE ANALYSIS OF CREDIT SCORE CLASSIFICATION DATASET

GROUP 2



Prepared By :

s15553 Janith Ramanayake

s15677 Bhashitha Wijesinghe

## Contents

1. Introduction .....	2
2. Description of the Question .....	3
3. Description of the Dataset .....	3
4. Data Cleaning and Preprocessing.....	4
5. Results of Descriptive Analysis.....	6
5.1 Target variable – Credit Score .....	6
5.2 Correlations among the variables.....	6
5.3 Variation of the predictor variables with the response variable .....	8
5.4 Bivariate Analysis of Variables .....	9
5.5 Partial Least Squares Regression .....	11
6. Discussion and Analysis.....	12
7. Suggestions for Advanced Analysis.....	13
8. Appendix.....	13

## List of Figures

Figure 1The Credit Score Scales .....	2
Figure 2 Bar plot for the Credit Scores.....	6
Figure 3 Pearson Correlation Plot .....	7
Figure 4 Distribution of Delay from due date and num of delayed payment by Credit Score.....	8
Figure 5 Distribution of Outstanding debt and Credit utilization ration by Credit Score.....	8
Figure 6 Distribution of Credit history age and Num of credit card by Credit Score .....	9
Figure 7 Scatter plot of Annual income vs Monthly inhand salary .....	9
Figure 8 Scatter plots of Credit History Age with No. of loans, No. of credit inquiries and Outstanding Debt.....	10
Figure 9 Scatter plots of Outstanding Debt with Credit Inquiries and No. of loans.....	10
Figure 11 Loading Plot.....	11
Figure 10 Score Plot .....	11

## List of Tables

Table 1 Dataset Description .....	4
Table 2 Spearman Correlation values of variables .....	7

# 1. Introduction

Credit scores, which represent numerical values of risk associated with an individual on the basis of his/her credit history and financial behavior, are the real drivers of modern finance. Banks and credit card companies use credit scores to ascertain the amount of risk involved in lending money to a borrower. If one has a high credit score, he can expect better interest rates and terms of loans, and it will make obtaining a mortgage, an auto loan, or even a personal loan quite easy. Moreover, credit scores impact insurance premium rates, apartment rentals, and even job applications, since some employers view credit history as a yardstick for one's reliability and character. Understanding how to properly work with credit scoring will allow people more flexibility and stability in this financially driven age.

The credit scores are the results of applying complex mathematical algorithms to data from a person's credit report. Traditionally, lenders used to assess the credit risk manually in the early part of the 20th century because there were no standardized scoring systems then in use. This changed by the late 1950s and 1960s, with the development of mathematical models for predicting credit risk and the need to speed up the process of lending.

An even major breakthrough came about in 1989, introducing the FICO score. It is a three-digit number ranging from 300 to 850 and the first credit risk model developed using statistical analysis. Also, it would prove to be the most widely used credit scoring model in the United States. Its simplicity and reliability gained wide acceptance of the FICO score.

With the growing relevance of credit scores, firms began working on in-house credit scoring models to compete with FICO. One very good example is the VantageScore, introduced in 2006. Nowadays, with technological advancement, models of credit scoring are very sophisticated, able to efficiently process vast volumes of credit data and make very accurate risk assessments that will be better at distinguishing among consumers based on their creditworthiness.

Even after all these developments, most people still have a bad credit score simply because they are uninformed and ill aware of the rules of the game that govern credit scores. Therefore, educating people on the fundamentals that drive credit scores is very relevant to achieving better financial health. Thorough analyses of data on finance should be done to identify trends and key factors that largely control credit scores.

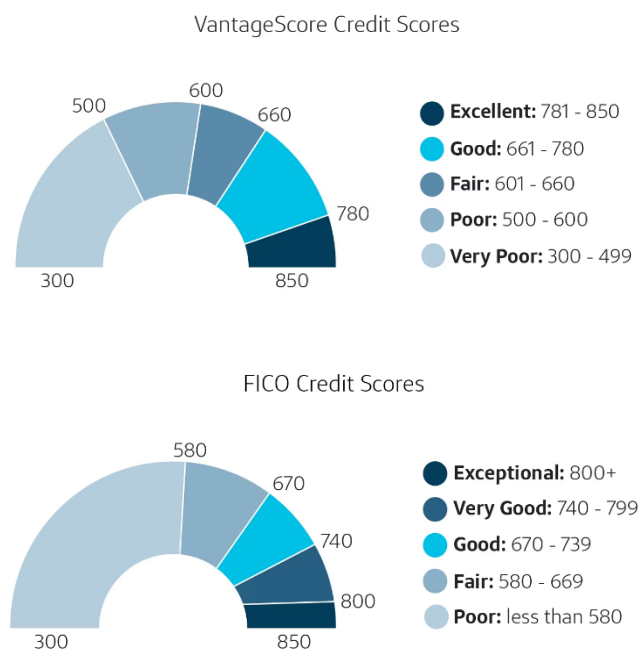


Figure 1The Credit Score Scales

This analysis will evaluate trends identification and factors associated with credit scores on the Credit Score Classification Dataset taken from Kaggle. We are going to look through this data while trying to answer such questions: what are the chief credit-related factors and which of the personal data elements are heavy on credit scores?.

## 2. Description of the Question

Our main variable of interest was "Credit Score," which is a qualitative variable users derived from their financial history and personal track record. Today, society is embracing credit in almost all purchases or monetary decisions. For example, if you want to buy a home, the mortgage lending institution will use your credit data to estimate the probability of your defaulting on the loan repayment. You may struggle to get a mortgage with bad credit, or you will with higher interest rates, directly impacting the monthly cost of your mortgage payment. Those with low credit scores may be rejected for the loan or charged higher rates.

Good credit offers a host of advantages that include, among others, the following:

- Lower auto insurance premiums
- Lower premiums on home insurance
- Lower interest rates on credit cards
- Higher credit limits on credit cards
- More likely to get approved for utility services, such as electricity or internet
- You are an attractive candidate for potential employers

Some of our key objectives are:

- Determination of credit-related properties and personal banking data that impact the credit score
- Determination of the optimal conditions of these impactful variables so as to keep a good credit score
- Development of a predictive model in order to predict credit scores based on users' financial history and personal data

Due to the fact that machine learning models can support large data sets, credit score computation is way more accurate as compared to traditional ways by identifying patterns for precise predictions.

## 3. Description of the Dataset

The dataset used here is the Credit Score Classification Dataset from Kaggle, which includes basic bank details and some credit-related information gathered from 12,500 people in a global finance company. There were, in all, 100,000 observations in this dataset, of monthly data collected from each individual during the January to August period. It contains 29 variables, including one response variable of interest: Credit Score, classifying credit scores into three groups.

Variable	Description of variable	Variable Type
ID	The unique identification of an entry	String
Customer ID	The unique identification of a customer	String
Month	Month of the Year	Categorical(Ordinal)
Name	Name of the person	String
Age	Age of the person	Numerical
SSN	The social security number of a person	String
Occupation	The occupation of the person	Categorical(Nominal)
Annual Income	The annual income of the person	Numerical
Monthly_Inhand_Salary	The monthly salary of a person	Numerical
Num_Bank_Accounts	The number of bank accounts a person holds	Numerical
Num_Credit_Card	The number of other credit cards held by a person	Numerical
Interest_Rate	The interest rate on credit card (percent)	Numerical
Num_of_Loan	The number of loans taken from the bank	Numerical
Type_of_Loan	The type of the loans taken from the bank	Categorical(Nominal)
Delay_from_due_date	The average number of days delayed from the payment date in days	Numerical
Num_of_Delayed_Payment	The average number of payments delayed by a person	Numerical
Changed_Credit_Limit	The percentage change in credit card limit (percent)	Numerical
Num_Credit_Inquiries	The number of credit card inquiries	Numerical
Credit_Mix	The classification of the mix of credits (Bad, Standard, Good)	Categorical(Ordinal)
Outstanding_Debt	The remaining debt to be paid	Numerical
Credit_Utilization_Ratio	The utilization ratio of credit card (percent)	Numerical
Credit_History_Age	The age of credit history of the person (days)	Numerical
Payment_of_Min_Amount	Represents whether only the minimum amount was paid by the person	Categorical(Ordinal)
Total_EMI_per_month	The monthly Equated Monthly Installment made by customer	Numerical
Amount_invested_monthly	The monthly amount invested by the customer	Numerical
Payment_Behaviour	The payment behavior of the customer	Categorical(Nominal)
Monthly_Balance	The monthly balance amount of the customer	Numerical
Credit_Score	Represents the bracket of credit score (Poor, Standard, Good)	Categorical(Ordinal)

Table 1 Dataset Description

## 4. Data Cleaning and Preprocessing

The original dataset was divided into test and train sets. However, it was noted that the test set was created specifically for competition purposes and did not include credit scores. As a result, this test set was not suitable for future modeling tasks. Hence training set consider as the primary data source. The train dataset contained **100,000 records** with **12,500 unique customers**.

Values of numerical variables such as *Age, Annual\_Income, Num\_of\_Loan, Num\_of\_Delayed\_Payment, Changed\_Credit\_Limit, Outstanding\_Debt, Amount\_invested\_monthly*, and *Monthly\_Balance* were found to be in object form, with some values starting or ending with string characters. These characters were identified and removed, converting the values into numerical form.

For a particular Customer\_ID, variables such as *Name, Age, SSN, Occupation, Annual\_Income, Interest\_Rate, Num\_of\_Loan, Type\_of\_Loan, Credit\_Mix*, and *Outstanding\_Debt* typically have unique values for every month. However, occasional discrepancies were observed in one or two months for certain Customer\_IDs. These discrepancies were considered human errors and were corrected by assigning the appropriate values.

For a particular Customer\_ID, variables such as *Monthly\_Inhand\_Salary, Num\_Bank\_Accounts, Num\_Credit\_Card, Interest\_Rate, Num\_of\_Loan, Num\_of\_Delayed\_Payment, Changed\_Credit\_Limit, Num\_Credit\_Inquiries, Credit\_Mix*, and *Total\_EMI\_per\_month* typically follow clear patterns over time. However, occasional unusual values disrupt these patterns. These anomalies were identified and corrected using forward and backward fill methods.

There were 1,426 customers without any Type\_of\_Loan value. Consequently, it was decided to drop these customers from the dataset. Missing values in *Amount\_invested\_monthly, Payment\_Behaviour*, and *Monthly\_Balance* were then addressed. For each Customer\_ID, these missing values were imputed using the **mean and mode** of the respective customer's data.

The Month variable was then converted to a numerical format. Initially, the Type\_of\_Loan variable had over 6,000 unique categories, making analysis challenging. To address this, the Type\_of\_Loan variable was transformed into a numeric format using NLP techniques. This approach leverages the potential relationships and sequences among different loan types. By applying NLP techniques, a numeric value was obtained, enabling the identification of similar loan combinations more effectively.

After completing all the cleaning and preprocessing steps, the final dataset contained **88,592 records** with **11,074 unique customers**.

When splitting the dataset into train and test sets, randomly dividing the entire dataset is not meaningful because all instances of a customer must be treated as a single unit. Therefore, the Customer\_IDs are first divided randomly, with 80% allocated for training and 20% for testing. Then, all instances belonging to each customer are placed in the correct set accordingly.

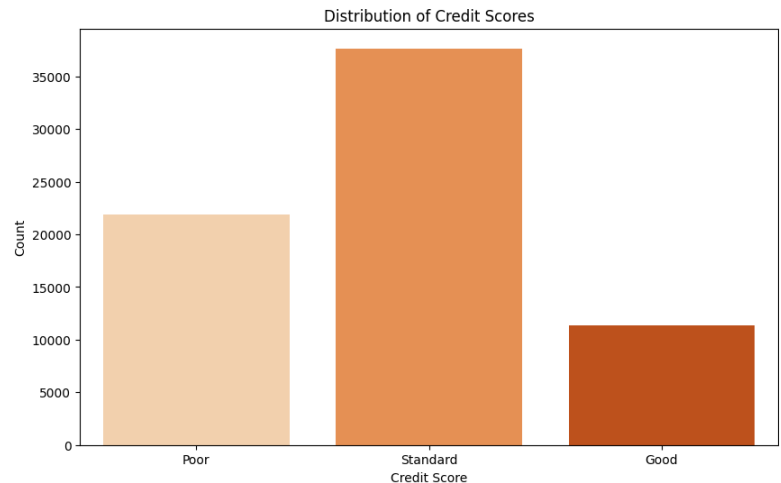
## 5. Results of Descriptive Analysis

### 5.1 Target variable – Credit Score

Assuming that the observations are independent of one another, a bar plot was created for the response variable, Credit Score. This variable consists of three categories: Poor, Standard, and Good.

Due to the lack of information on the recoding scheme used, the FICO score ranges (300-850) were employed to define these categories:

- **Good Category:** Very Good (740+)
- **Standard Category:** Average (670–739)
- **Poor Category:** Poor (less than 669)



*Figure 2 Bar plot for the Credit Scores*

The analysis revealed that the majority of observations fall within the Standard range, with relatively fewer observations in the Good and Poor categories. This indicates that most customers have a Standard Credit Score.

### 5.2 Correlations among the variables

The dataset contained both categorical and numerical predictors. Therefore, three approaches were used to determine the correlations among the variables:

1. **Spearman's Rank Correlation** was employed to assess the association between the numerical predictors and the categorical target variable.
2. **Chi-Square Test** was used to determine the association between the categorical predictors and the categorical target variable.
3. **Pearson Correlation Coefficient** was utilized to calculate the correlations among the numerical variables.



Table 2 Spearman Correlation values of variables

Variable Name	Spearman's Rank Correlation
Age	-0.0299
Annual Income	-0.0368
Monthly_Inhand_Salary	-0.0356
Num_Bank_Accounts	0.0886
Num_Credit_Card	0.0250
Interest_Rate	0.0486
Num_of_Loan	0.0042
Delay_from_due_date	0.0596
Num_of_Delayed_Payment	0.1192
Changed_Credit_Limit	0.1830
Num_Credit_Inquiries	-0.0073
Outstanding_Debt	-0.0548
Credit_Utilization_Ratio	-0.0055
Credit_History_Age	-0.0445
Total_EMI_per_month	-0.0403
Amount_invested_monthly	-0.0272
Monthly_Balance	-0.0168

The analysis's most important factors were determined and looked at individually based on the results. It is significant to highlight that, for the sake of simplicity, the values were determined under the assumption that each observation is independent. But as was already mentioned, each customer's month-by-month data for the months of January through August was included in the dataset. With 11,074 customers in total, it was a laborious process to analyse the variables for each individual consumer. In order to solve this, the predictors were examined month by month, which made the research simpler while still taking each consumer into account.

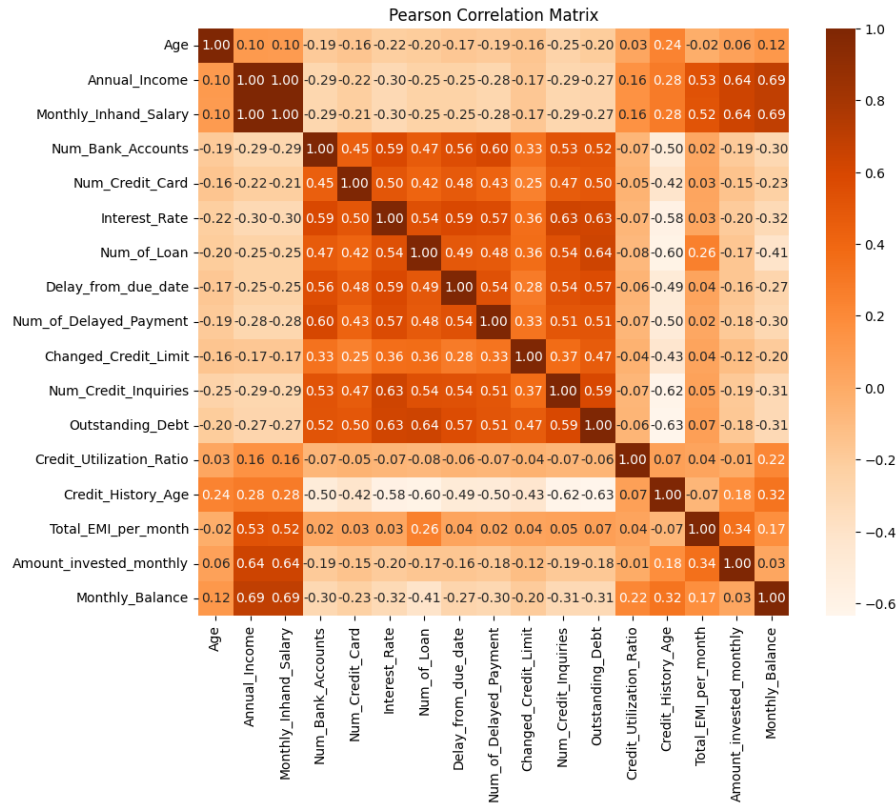


Figure 3 Pearson Correlation Plot



### 5.3 Variation of the predictor variables with the response variable

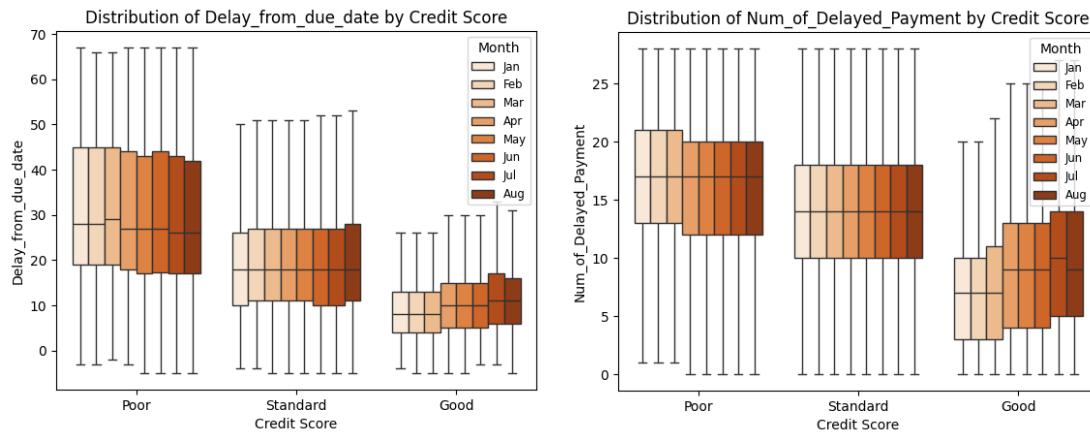


Figure 4 Distribution of Delay from due date and num of delayed payment by Credit Score

According to sources, the most important factor for credit scores is **payment history**, which accounts for 35% of the total importance. This category includes an individual's track record of paying bills on time, as well as any late payments, collections, and bankruptcies. In the dataset, the variables Delay\_from\_due\_date and Num\_of\_Delayed\_Payment are related to this factor. As observed in the plot, the credit score increases with the decreasing Delay\_from\_due\_date, and Num\_of\_Delayed\_Payment

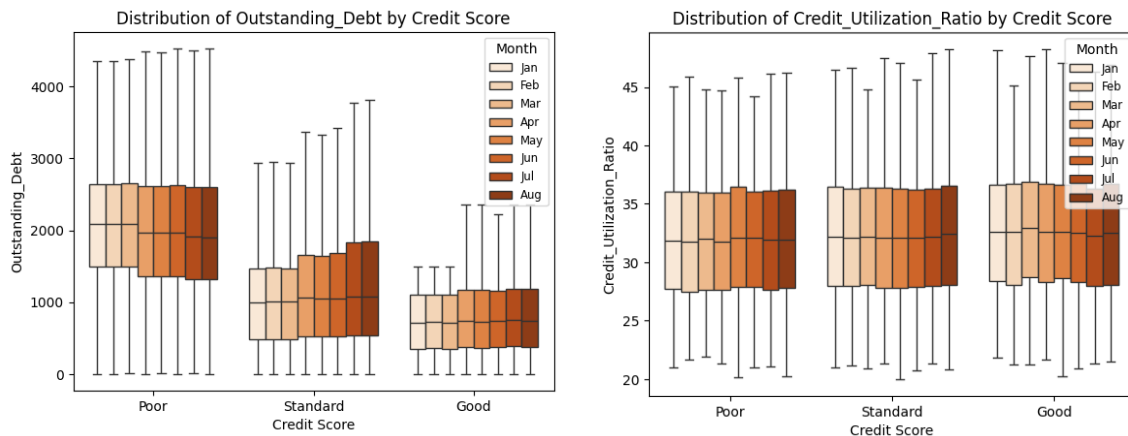


Figure 5 Distribution of Outstanding debt and Credit utilization ratio by Credit Score

**Amounts owed** have a 30% importance in the credit score according to sources. This factor considers the total amount of debt an individual has, the number of accounts with balances, and how much of their available credit they are using. Variables such as Outstanding\_Debt and Credit\_Utilization\_Ratio are related to this factor. As observed in the plot, the credit score increases with the decreasing Outstanding\_Debt, but there is no significant variation in the Credit\_Utilization\_Ratio.

According to sources, the **length of credit history** holds a 15% importance in determining a credit score. This factor considers how long an individual's credit accounts have been established, including the age of the oldest account, the average age of all accounts, and the age of specific types of accounts. Generally, a longer credit history can contribute to a higher score. The variable `Credit_History_Age` is related to this factor.

**Credit mix** accounts for 10% of a credit score's importance according to sources. FICO scores evaluate the variety of credit accounts an individual has, such as credit cards, retail accounts, installment loans, finance company accounts, and mortgage loans. A diverse mix of different types of credit can positively impact the score. The variable `Num_Credit_Card` is related to this factor. But as observed in the plot, the credit score decreases with the increasing `Num_Credit_Card`.

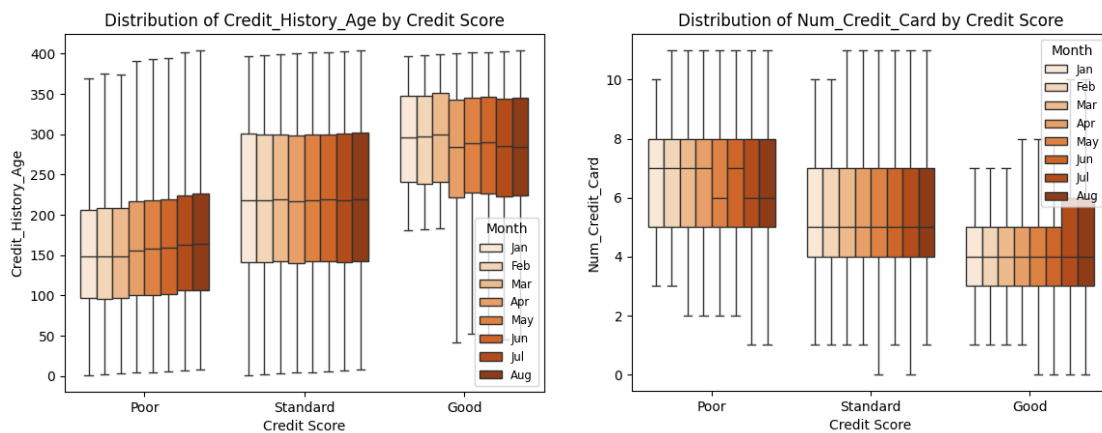


Figure 6 Distribution of Credit history age and Num of credit card by Credit Score

## 5.4 Bivariate Analysis of Variables

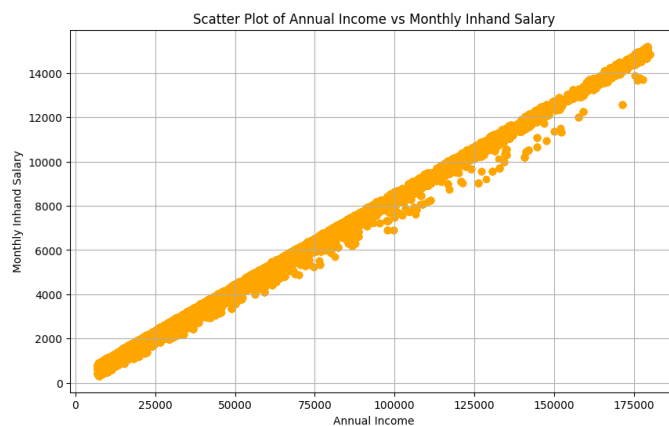


Figure 7 Scatter plot of Annual income vs Monthly inhand salary

There was a perfect correlation found between the year income and the monthly in-hand salary. The link arises from the fact that an increase in a person's monthly wage directly correlates with an increase in their income. While a perfect link was found, there is considerable dispersal in some areas. This could be because those who earn more each month have a tendency to invest, which raises their income.

There were weak negative correlations found between the number of loans, credit enquiries, and outstanding debt and the credit history age, measured in months.

This suggests that the `Num_of_Loan` reduces or vice versa as `Credit_History_Age` grows. A number of things could be the cause of this negative correlation, such as the fact that people with longer

Num\_of\_Loan may have developed a strong credit history and are hence less dependent on taking out multiple loans.

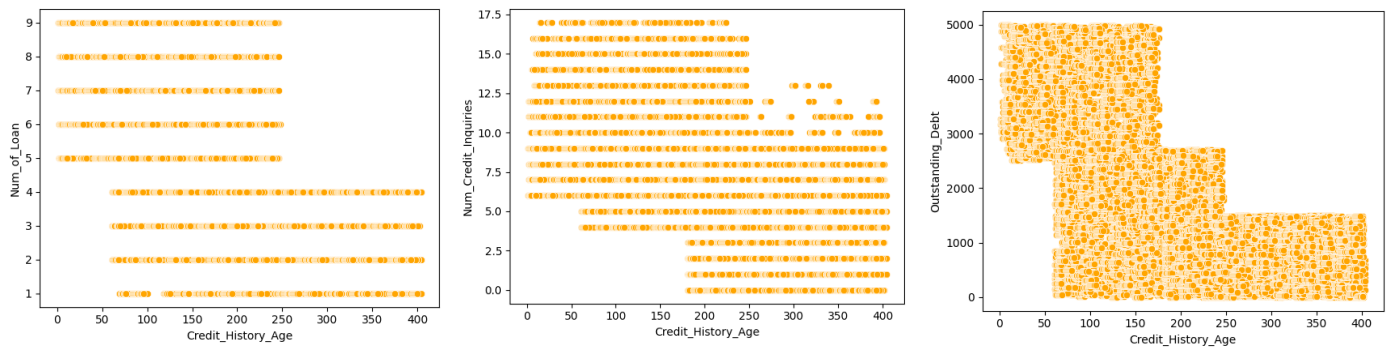


Figure 8 Scatter plots of Credit History Age with No. of loans, No. of credit inquiries and Outstanding Debt

Longer credit histories are generally seen by lenders as more creditworthy, which facilitates the acquisition of larger loans or the consolidation of preexisting debt.

Longer credit histories may indicate that a person has paid off or substantially decreased prior debt, which reduces the amount of current outstanding debt.

Individuals may experience changes in their credit demands and borrowing behaviours as they move through different periods of life. Elderly people may not require as many fresh loans, but younger people may be more inclined to take out several loans for different uses.

Likewise, those with a longer credit history tend to be more established, which lessens the necessity for credit enquiries. Additionally, because they have more expertise handling credit, they would be more circumspect about any potential repercussions while doing credit enquiries. This clarifies the unfavourable correlation.

Additionally, people with longer credit histories may have had more time to develop sound financial practices, which has led to lower levels of outstanding debt. This elucidates the inverse relationship between the age of credit history and outstanding debt.

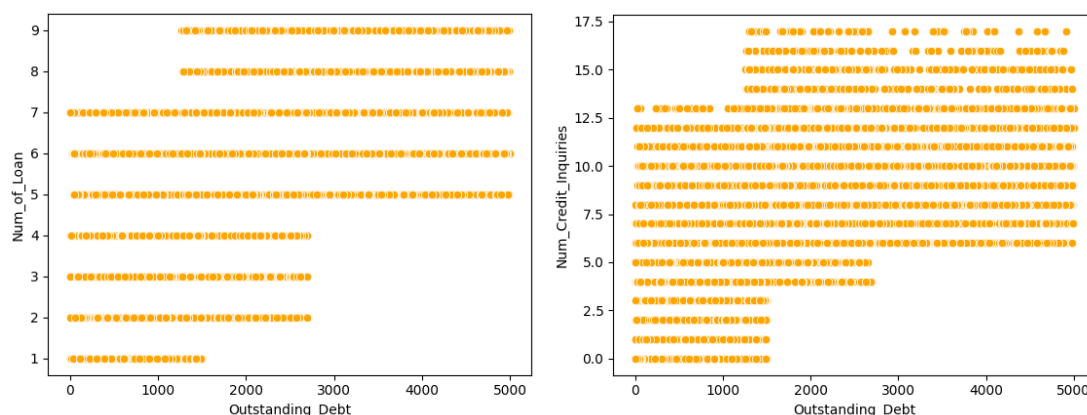


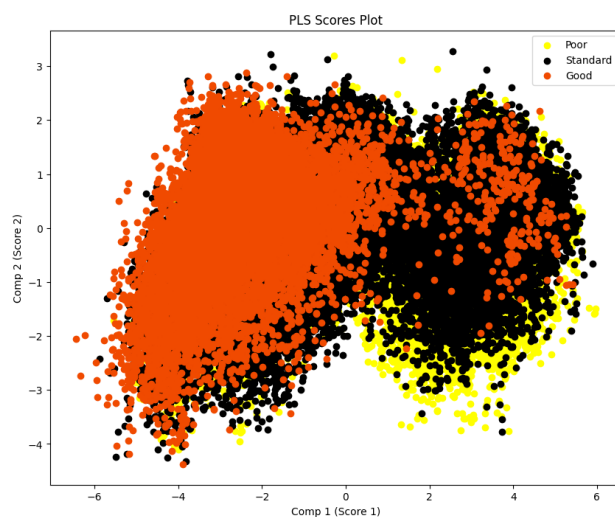
Figure 9 Scatter plots of Outstanding Debt with Credit Inquiries and No. of loans

The amount of outstanding debt positively correlates with both the number of credit enquiries and loan denials.

Higher debt levels may indicate that a person is actively using credit to meet their demands and fund different expenses. There may be a rise in credit enquiries as a result of their increased credit usage and subsequent need to apply for more credit. Additionally, some people may seek for new loans in order to lower their interest rates or consolidate their current obligations. Their goal is to better manage their outstanding debt, thus they do this, which raises the number of credit enquiries.

In the case of loans, people may take out more loans due to their increasing financial needs, which could raise their total amount of outstanding debt. Additionally, company owners and entrepreneurs may take out several loans in order to finance the growth and operation of their enterprises. Their outstanding debt may rise as they make more investments in their company.

## 5.5 Partial Least Squares Regression



Under the assumption that the observations are independent, Partial Least Squares Regression was performed on the dataset to identify any clusters among the observations and to identify significantly correlated predictors. In the score plot provided, nearly 31% of the variation is explained by the first two components. The plot indicates that there are no significant clusters in the observation set.

Figure 11 Score Plot



Figure 10 Loading Plot

Moving on to the loadings plot, it was observed that there are strong correlations among some predictors and the response, whereas some predictors are orthogonal to the response. Additionally, some variable clusters were identified.

- Variables such as Age and Credit\_History\_Age exhibit a positive association with the response. Conversely, variables including Num\_of\_Loan, Outstanding\_Debt, Num\_of\_Delayed\_Payment, Num\_Bank\_Accounts, Num\_Credit\_Inquiries, Interest\_Rate, Delay\_from\_due\_date, and Num\_Credit\_Card show a negative association with the response.
- Variables like Changed\_Credit\_Limit, Credit\_Utilization\_Ratio, Total\_EMI\_per\_month, Amount\_invested\_monthly, Monthly\_Balance, Annual\_Income, and Monthly\_Inhand\_Salary are almost orthogonal to the response, indicating that these variables might not impact the credit score.
- Most notably, Monthly\_Inhand\_Salary and Annual\_Income seem to overlap, confirming the strong association between the two variables discovered earlier.

## 6. Discussion and Analysis

- Gaining a thorough knowledge of the variables included in the dataset was one of the most difficult jobs to complete. The data card that was connected to the dataset does not provide a description for any of the variables. Additionally, in order to determine the proper order for the answer variable, we had to go through the various scalers that were accessible for credit score calculation.
- Many banking journals state that "Credit History Age" is the average amount of time a customer has had credit accounts opened. Vantage Score and FICO, two of the most popular credit scoring models, determine credit history age by adding the customer's oldest and newest account ages (in months) and dividing the result by the total number of accounts the customer has. However, it should be noted that the "Credit History Age" variable in our dataset only relates to the duration of each account's credit history, not the entirety of them.
- The scatterplots did not make the moderate relationships between predictors, as determined by Pearson's Rank correlation, readily visible due to the huge number of observations in the dataset
- One of the main difficulties encountered during the investigation was the quantity of temporal structure in the data. The mixed effect model technique, which would be taken into consideration during advanced analysis, was compelled to be used as a result of the independent observation assumption being broken. It was crucial to preserve the data's structure throughout the descriptive analysis by taking observational groups into account. Therefore, rather than considering individual observations as is customary, groups of observations were taken into consideration during pre-processing through the use of data splitting and exploratory data analysis techniques.
- It can be concluded from the comprehensive descriptive analysis, which included partial least squares regression, bivariate analysis, and univariate analysis, that the variables Outstanding\_Debt, Interest\_Rate, Num\_of\_Credit\_Inquiries, and Credit\_Mix have a significant

influence on the response variable of Credit\_Score. To support the preceding assertion, however, a comprehensive Advanced Analysis is necessary because of the intricacy of the data and the conflicts that exist between correlation and causation.

## 7. Suggestions for Advanced Analysis

- The traditional classification algorithms cannot be used because our dataset includes multilayer categorical answer data with linked variables from January to August, pertaining to consumers of a financial company.
- It is necessary to employ proper evaluation criteria for model evaluation that take into account the data's clustered nature.
- Because of the quantity of the dataset, it will be crucial to consider the time efficiency of the models used. Therefore, while fitting models, it is advised to use a better GPU, particularly when employing grid search for hyperparameter tuning.

## 8. Appendix

All codes and reports at: [https://github.com/JanithRavinduRashmika/Credit\\_Score\\_Classification](https://github.com/JanithRavinduRashmika/Credit_Score_Classification)