[1.0 Introduction 1](#_Toc1273428858)

[1.1 Data Description: 1](#_Toc1809443104)

[1.2 Assumptions: 1](#_Toc1629342321)

[1.3 Hypothesis: 1](#_Toc1687559545)

[To investigate the impact of credit, mix on credit scores. (Shong Ming Xuan) 1](#_Toc837415263)

[To explore the effect of Credit history age on credit scores. (Chee Kai Jian) 1](#_Toc93047077)

[To examine the influence of annual income on credit scores. (Wong Wei Jun) 1](#_Toc1607621454)

[2.0 Data Preparation 2](#_Toc1160660259)

[2.1 Data Import: 3](#_Toc528769420)

[2.2 Data Cleaning 3](#_Toc1174260971)

[2.2.1 Cleaning Procedure 3](#_Toc875910805)

[Random Characters Removal 4](#_Toc1442118424)

[Data Type Correction 4](#_Toc195164409)

[NA Data Mutation 5](#_Toc995656970)

[Number Irregularity 5](#_Toc523217176)

[3.0 Data Analysis 13](#_Toc558507600)

[3.1 13](#_Toc2131458481)

[Analysis 1 13](#_Toc2144149156)

[Analysis 2 17](#_Toc187308318)

[Analysis 3 20](#_Toc2076396800)

[Analysis 4(Machine Learning) 20](#_Toc588304591)

[4.0 General Conclusion 23](#_Toc1463459152)

[5.0 Workload Matrix 24](#_Toc586147989)

[6.0 Reference 24](#_Toc950731424)

# ***1.0 Introduction***

## 1.1 Data Description:

The dataset consists of 100,000 objects and 28 variables, including "ID", "Customer\_ID", "Month", "Name", "Age", "SSN", "Occupation", "Annual\_Income", "Monthly\_Inhand\_Salary", "Num\_Bank\_Accounts", "Num\_Credit\_Card", "Interest\_Rate", "Num\_of\_Loan", "Type\_of\_Loan", "Delay\_from\_due\_date", "Num\_of\_Delayed\_Payment", "Changed\_Credit\_Limit", "Num\_Credit\_Inquiries", "Credit\_Mix", "Outstanding\_Debt", "Credit\_Utilization\_Ratio", "Credit\_History\_Age", "Payment\_of\_Min\_Amount", "Total\_EMI\_per\_month", "Amount\_invested\_monthly", "Payment\_Behaviour", "Monthly\_Balance", and "Credit\_Score". With 28 columns and 100,000 rows, each entry represents its respective value in the classification data.

## 1.2 Assumptions:

1. The currency use is Ringgit Malaysia.

2. The oldest age is 100

3. Annual income larger than RM100000 a year is considered high income individual

4. Assume every customer has a unique customer ID.

5. Assume \_10000\_ in Amount invested monthly is null

6. Max Num of loan is assumed to be 10

7. Highest Num Credit Card is 10

8. Lowest Num\_Bank\_Account and Num\_Credit\_Card is 1

9.Negative delay form due day is assumed to be early payments

10. Max Interest rate is 100 percent

## 1.3 Hypothesis:

### *To investigate the impact of credit, mix on credit scores. (Shong Ming Xuan)*

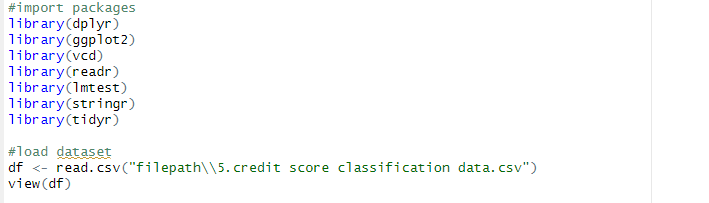
### *To explore the effect of Credit history age on credit scores. (Chee Kai Jian)*

### *To examine the influence of annual income on credit scores. (Wong Wei Jun)*

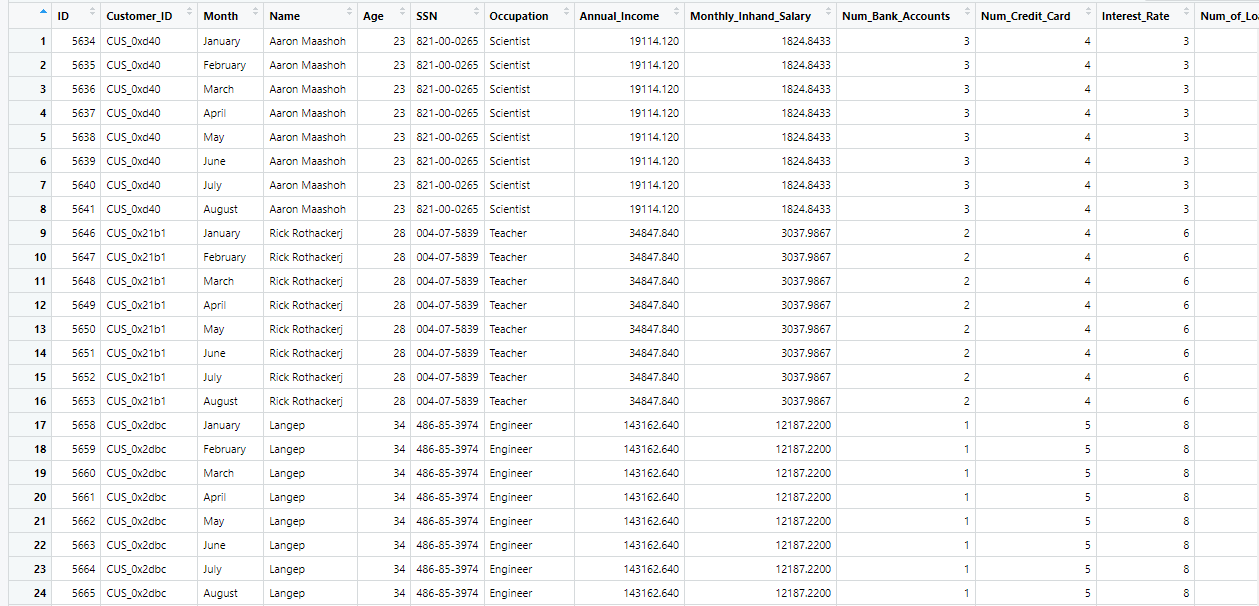
# **2.0 Data Preparation**

## 2.1 Data Import:

**Input:**

 *Figure input Dataset*

**Output:**

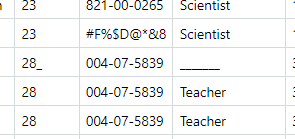
 *Figure Output Dataset*

## 2.2 Data Cleaning

Prior to analysing the data given, is it vital to follow the procedure and making sure the data is in the correct condition to be utilized. This process is completed to ensure a more thorough and accurate representation of the data.

### 2.2.1 Cleaning Procedure

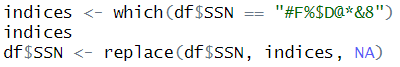
#### **Random Characters Removal**



*Figure of Random Characters*

In the following figure above, we can observe an example of random characters that does not belong there.

*Figure of underline removal*



*Figure of Character removal*

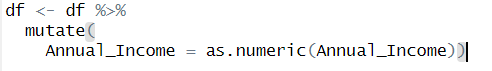
To solve these problems, we had changed the data to NA type to prepare the data for mutation later fill its NA value.

#### **Data Type Correction**

this procedure is to ensure that the data type matches the one that data is representing, for example: if we observe that Annual Income in this instance is a character data type, we need to change its data type to one of numeric or integer.

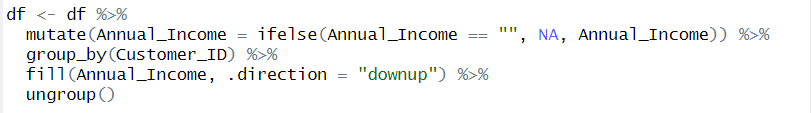
We know this due to the how the data in this column are display to us. If the values are all numbers (1,2,3,4,5) or the data required to represent the column must be a number (Age, Num\_Credit\_Card, Annual Income, Monthly Income Salary....)

We can immediately discard the possibility of its data type needing to be a character, in instances where the column is a numbers-based data, but it was given to us as character, a change of data type is needed.



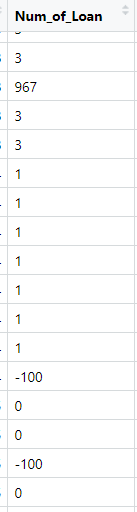
*Figure of Changing data type*

#### **NA Data Mutation**

*Figure : Mutation of data column*

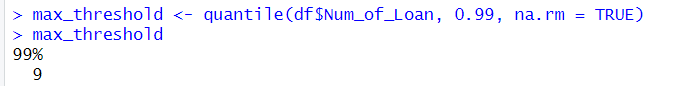
Once we have replaced all the data irregularities with NA, the code above will group the table by using customer\_ID . When each customer IDs are grouped into individual parts, it will then proceed to fill the NA (if NA exist within a group) with the data of the row above or below within the specified group .If the NA is the first value withing the group , it will fill it with the data of the row below.

#### **Number Irregularity**



*Figure of number irregularities*

In a few cases, we noticed number irregularity, these irregularities are often obvious in terms of being too large or below zero.

 *Figure :Code for 99 percentiles of Num\_of\_Loan*

To begin with limiting the data distance, we need to find out the maximum amount that we will accept within this data set when it comes to Num\_of\_Loan .

In this instance, utilization of percentile-based approach is the most appropriate due to its focus on the majority and ignoring the outliers.

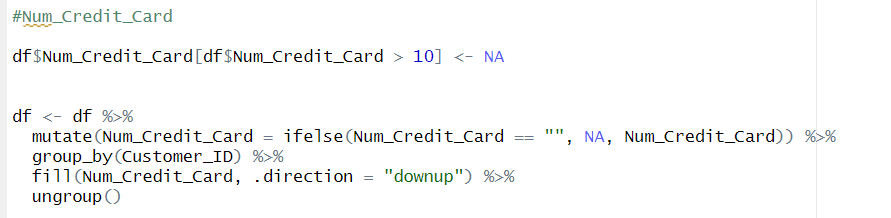


*Figure: Irregularities to NA*

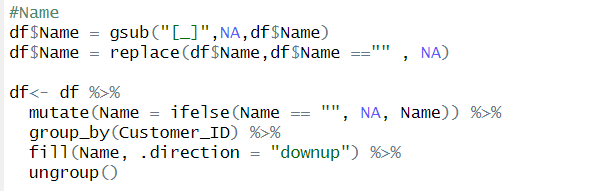
It is showed from the figures that 99 percentiles of the rows within Num\_of\_Loan has a maximum number of 9. After we have gained the threshold, we can use it to convert the rows that are larger than 9 to NA ensuring the extreme outliers are secluded.

In summary, values within Num of loans that are below 0 and above 9 are remove and replace with NA, after we handle the outliers, we can effectively mutate the data to fill in the NA gaps.

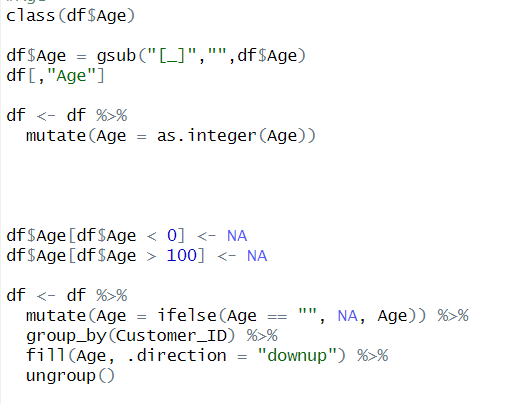
**Num\_Credit\_Card**



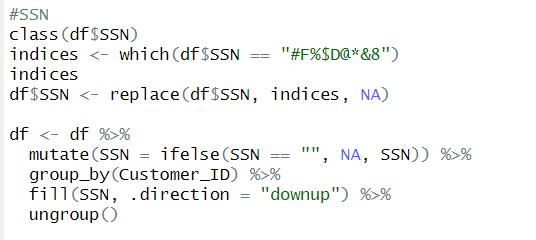
**Name**



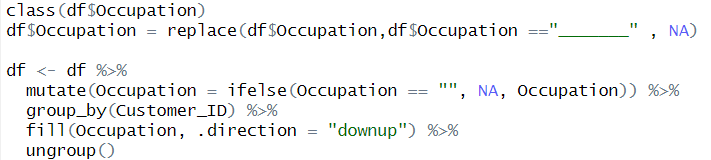
**Age**



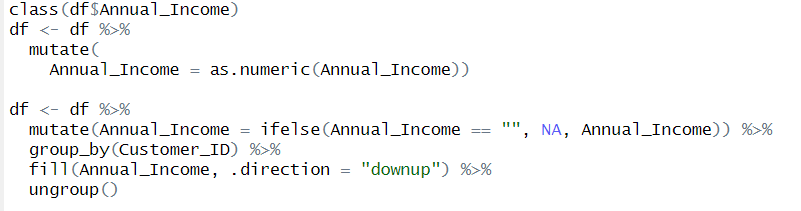
**SSN**

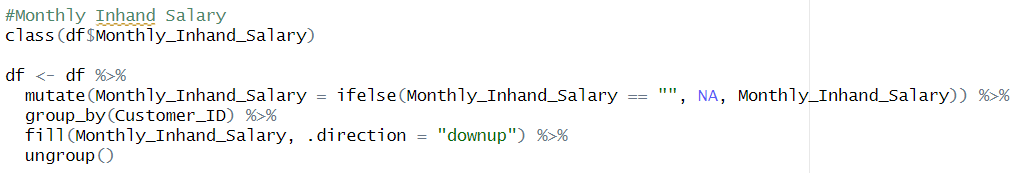


**Occupation**

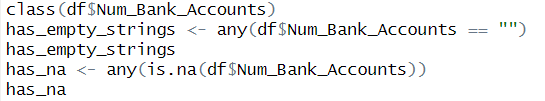


**Annual Income**

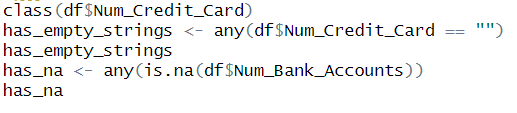
**Monthly Income Salary**



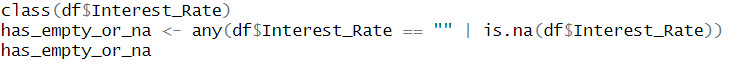
**Num\_Bank\_Account**



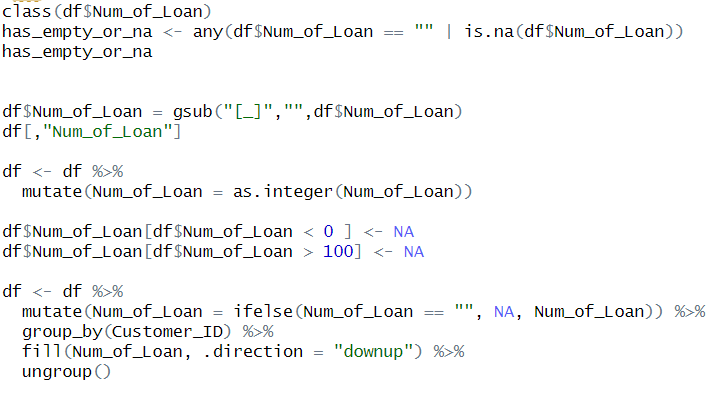
**Num\_Credit\_Card**



**Interest Rate**



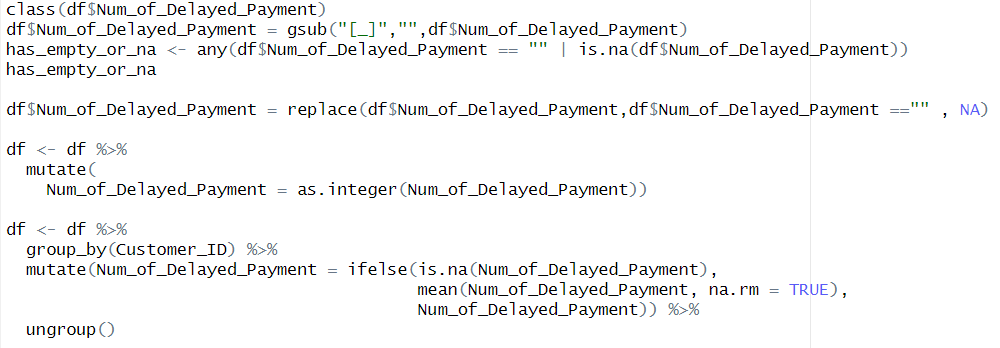
**Num\_of\_loan**



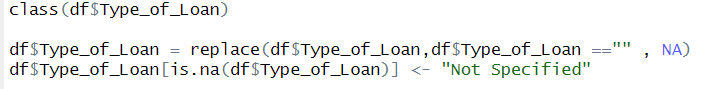
**Delay From Due Date**



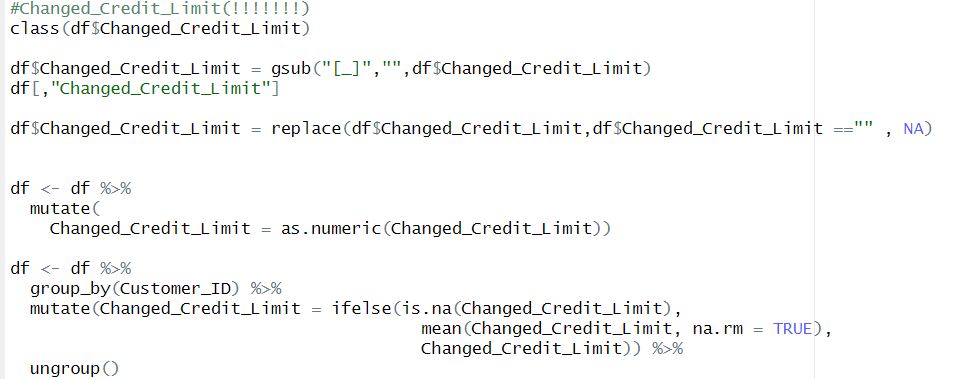
**Num\_of\_Delayed\_Payment**



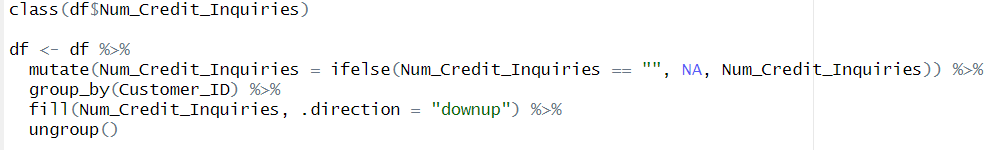
**Type Of Loan**



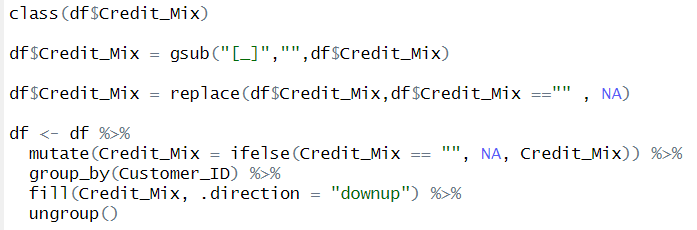
**Changed Credit Limit**



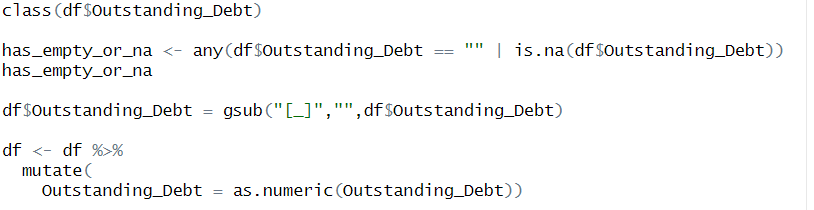
**Num Credit Inquiries**



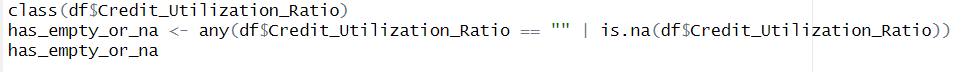
**Credit Mix**



**Outstanding Debt**



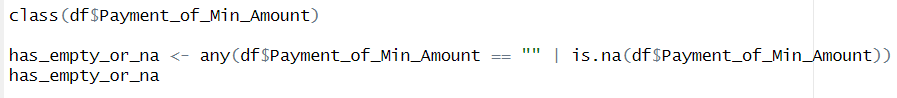
**Credit Utilization Ratio**



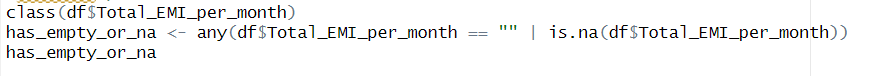
**Credit History Age**



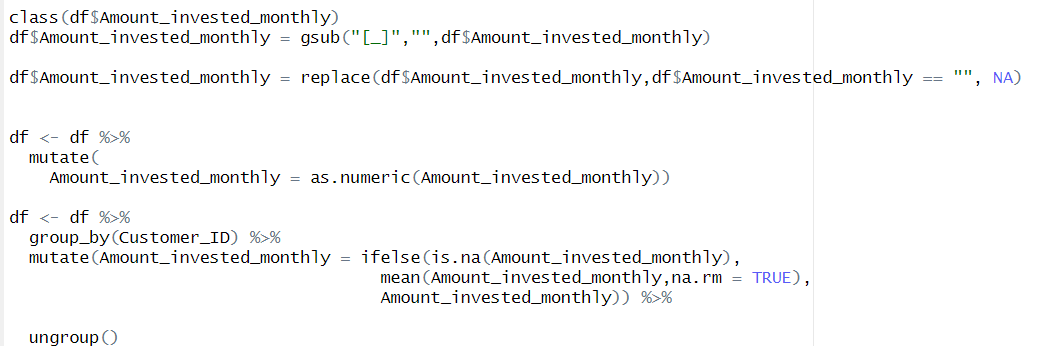
**Payment Of Min Amount**



**Total EMI per month**



**Amount Invested Monthly**



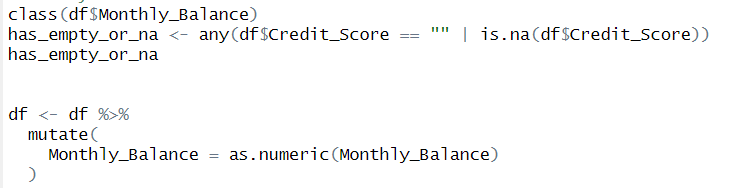
**Credit Score**



**Payment Behaviour**



**Monthly Balance**

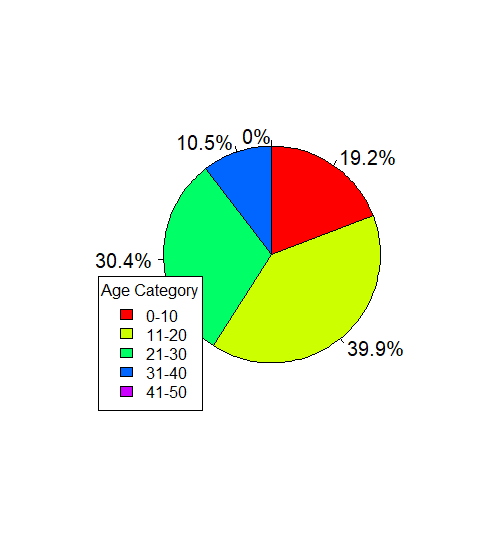


# **3.0 Data Analysis**

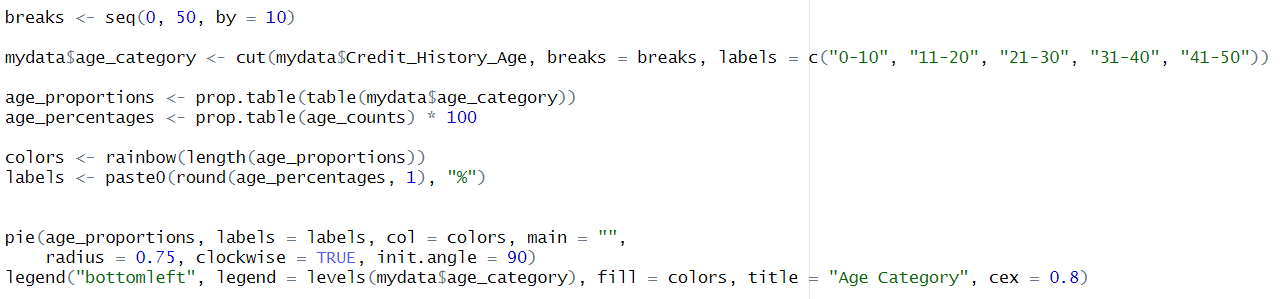
## 3.1

**Objective: To Investigate the relationship between Credit History Age and Credit Score**

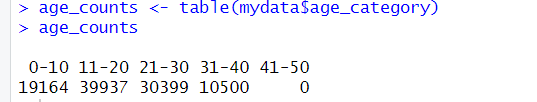
### **Analysis 1**



*Circle Graph Plot for Credit History Age*

*Code for Circle Graph*

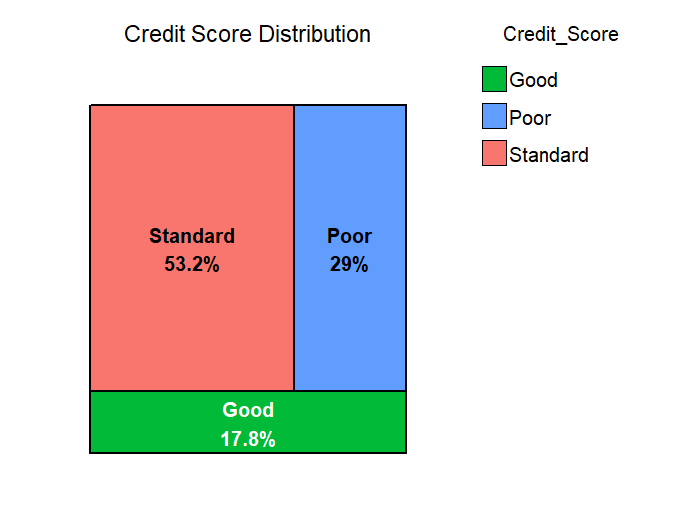
The analysis of this section helps us to understand and visualize the data spread of Credit History Age. To better understand to data, we have utilized a descriptive statistics and data visualization method, known as Pie Chart.



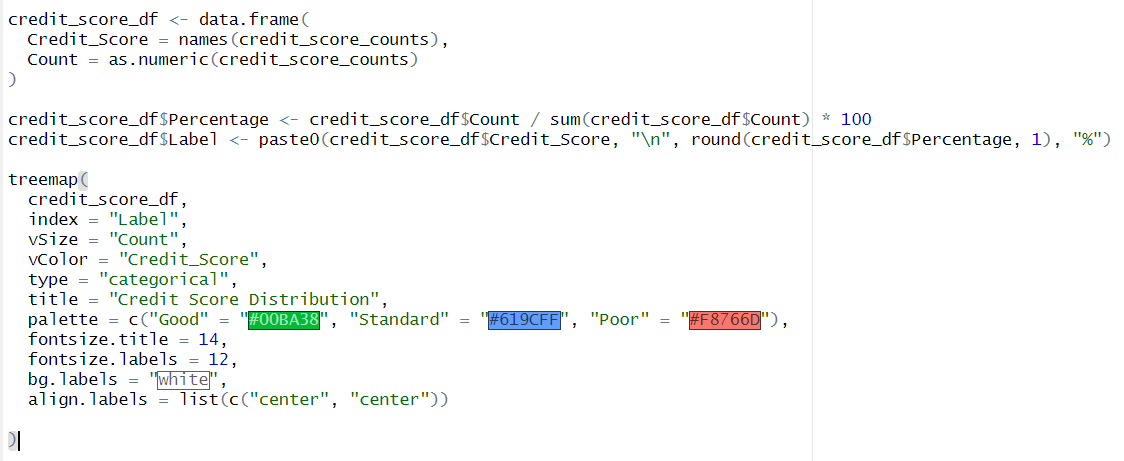
*Credit History Age frequency table*

Within this Pie Chart, distinct colours are used to separate and differentiate the different credit age groups. Regarding the data patterns, we can acknowledge the following:

Majority of the Credit age fall between the ages of 11 to 20 and 21 to 30, taking up a significant 70.3% of the data and totalling 70,336, whereas the remaining 29.7% amount to only amount to 29,664.

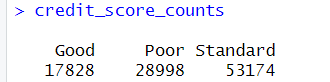


*Credit Score Tree Map*

*Code for Tree Map*



*Code for Credit Score frequency table*

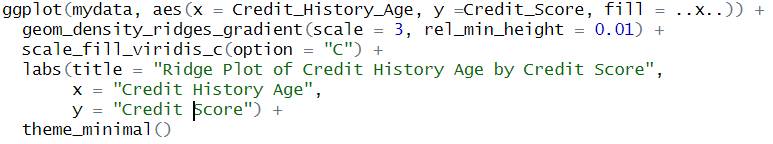


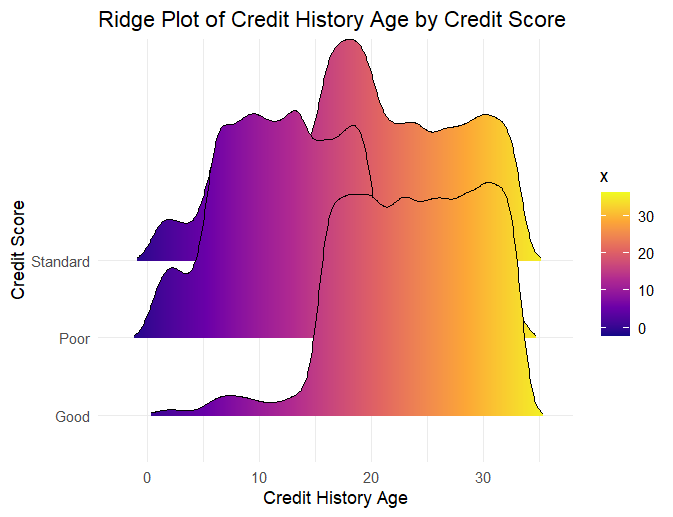
*Credit Score frequency table*

By using a tree map to fully visualize the different portions of Credit Score Category, as well as labelling the portions with percentages. Observing from the tree map we can notice three category and those are Standard, Poor, and Good. In terms of credit scores, we can learn a lot about a person's financial situation, debt, and credit management. The tree map is dominated by Standard portion with 53.2 percent amounting to, followed by Poor with 29 percent and the remaining belonging to Good with the least amount at 17.8 percent amounting to 17828. In summary, majority of people does not have good credit score but as we expected, the standard category encompasses the most people.

### **Analysis 2**

Seeks to determine the correlation between Credit Score History and Credit Score

*Code for Ridge plot*



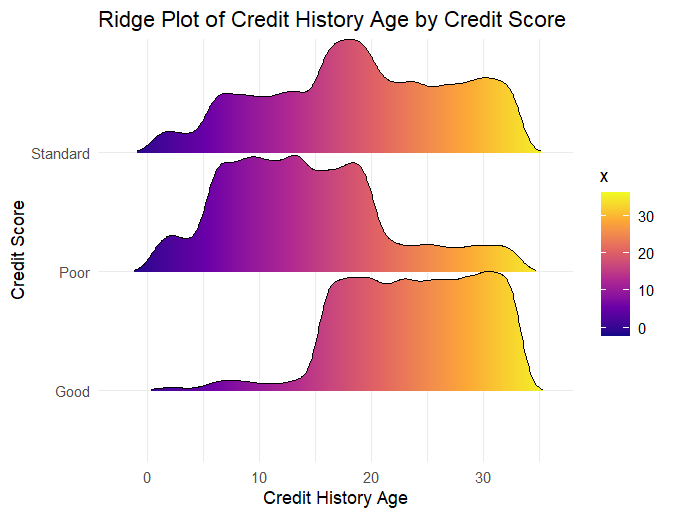
*Ridge plot between two variables*

To effectively understand how these two variables can be influential towards one another, utilization of the Ridge plot is deployed. Along the x axis represents Credit History Age, while the y axis represents Credit Score in three categories: Standard, Good and Poor. The height and shape of the ridges exhibit density of people. On the other hand, the colouration of the ridges indicates the density value of the colour bar shown on the right (Age of Credit History).

Immediately from the plot itself we can observed a significant impact on credit score when it comes to Credit History Age.

Credit Score Good on the Y-axis has a more concentrated Age range that leans a lot toward the older side of the X-axis. Which indicate a significant correlation that Credit History Age does come into account when it comes to the categorization of the good Credit score.

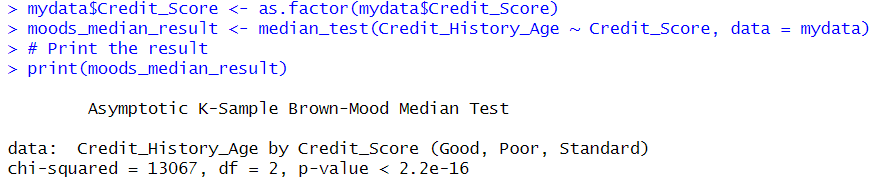
Moreover, the poor credit score ridge display the same trend and effect when it comes to Credit History Age influence on its categorization of people’s credit. We can notice on the poor’s ridge that it leans heavily to the younger side of Y-axis, that demonstrate most of the people in poor category when it comes to their credit classification is on the younger side.



*Ridge plot between two variables (scale =1)*

In regards of the Standard classification of credit score, even though compared to the previous two category it can be perceive as more even when it comes to certain ranges. However, there is still a noticeable increase in the density of the shape and sizes of the ridge. We can see that the middle part of the ridge has a noticeable increased density which falls in line with the previous two category’s pattern.

**Asymptotic K-Sample Brown-Mood Median Test**

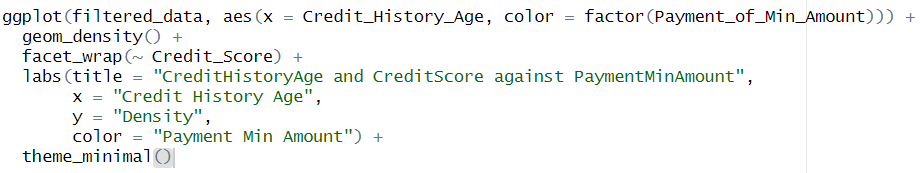


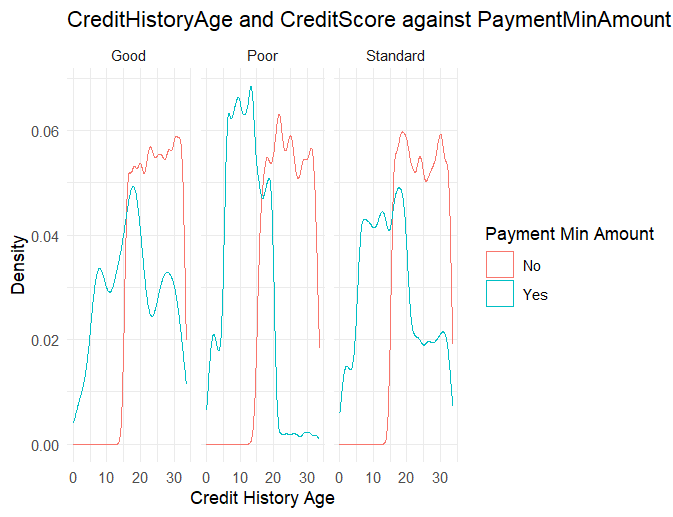
K-sample Median test is a non-parametric test that is applied when the dependent variable is ordinally scaled and is intended to investigate differences between more than two groups. From the Median test itself we can notice that the p value is smaller than 2.2e-16, this represents an exceedingly small number. Much smaller than the 0.05 significance level. This indicates the case of string evidence against the null hypothesis.

In basic terms, the test displays a variation of Credit History Age with different credit scores. Suggesting different credit score category groups are likely to have varying average credit history.

In summary, we can establish the following: The higher the Credit History Age, the better his/her Credit score classification Categories can fall into. From this ridge plot and the Mood Median test we have successfully proven the Credit History Age has an impact on credit Score.

### **Analysis 3**

*Code for faceted density plot*

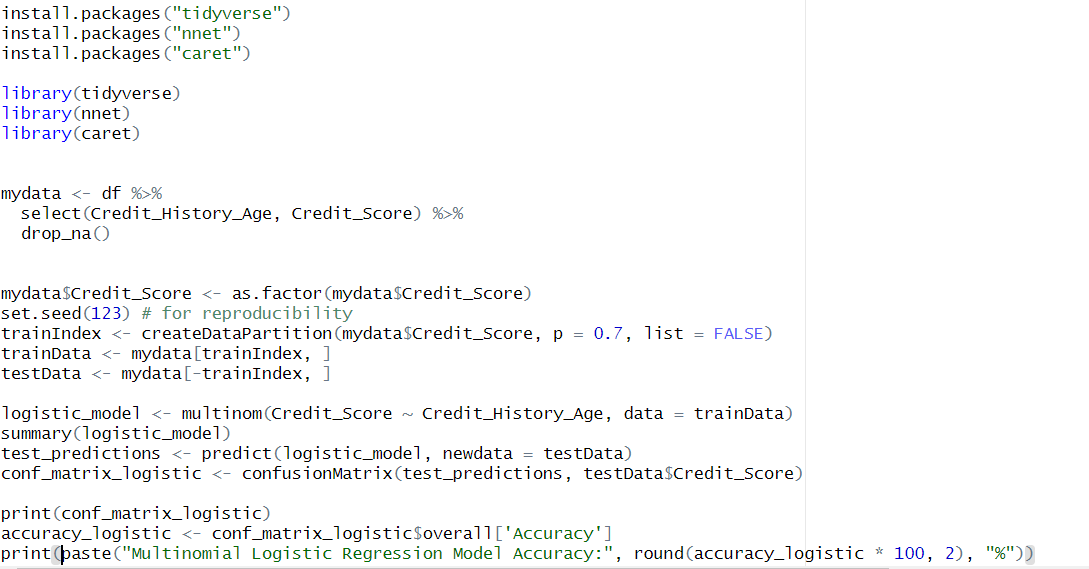


*Faceted density plot*

In this faceted density plot, along the X axis are the Credit History Age and the Y axis exist the Density and the top of the graph exist the credit score categories. The purpose of this graph is to display how only paying the minimum amount every month can affect the credit score.

The plost paints the picture clearly to us. When it comes to the Standard and Good Category, the majority does not only Pay the Minimum amount every month to sustain the credit account. Whereas within the poor category exist a more even spread of paying/not paying only minimum amount, however it is still noticeable when it comes to poor category having more density regarding only Paying Minimum amount.

## **Analysis 4(Machine Learning)**

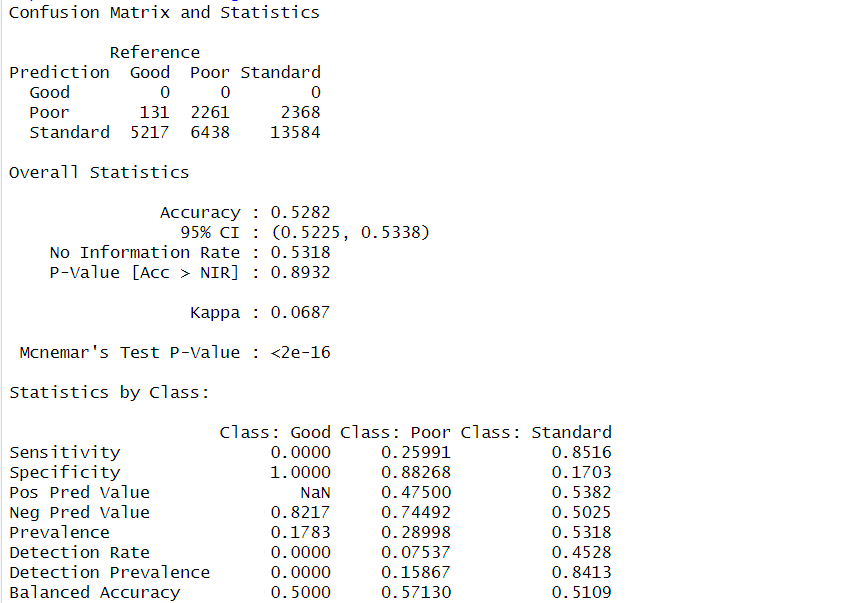


*Code for multinomial logistic regression*

For the fourth analysis, I have decided to utilize machine learning technique to dive deeper into the relationship between Credit score history and credit score. The ml model I have utilize is logistic regression model.

The multinomial logistic regression is an extending model of the logistic regression model, and it allows more than two categories within the dependent variable, in this care credit score has three categories (Good, Bad and Standard).

Tp prepare the data for training, I have split the data into two sets, those are training set with 70 percent of the data set and test set with 30 percent.

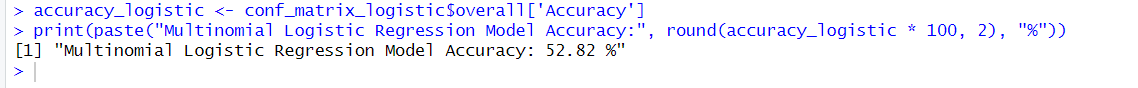


*Confusion Matrix*

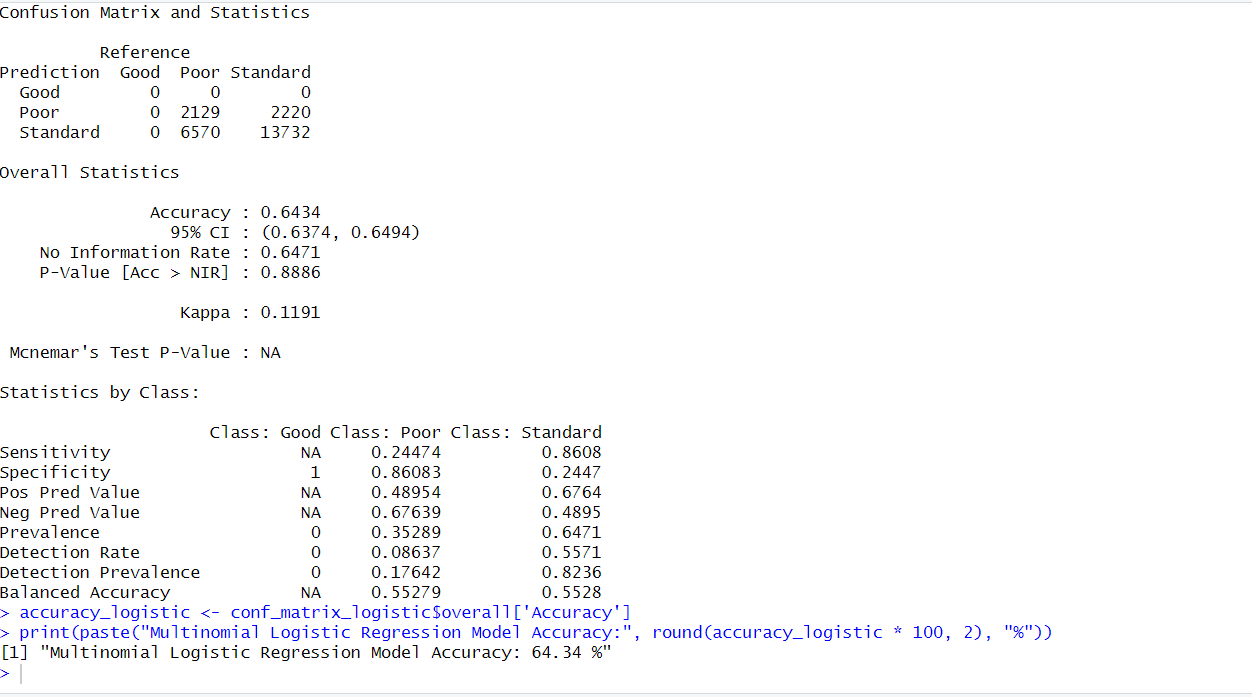
*Referring* to the figures above, I use the model to predict Credit score using credit score history and the predictions it made were not accurate enough. From the reference, as you can see the model is terrible at predicting the good credit score category with none correct predictions but where it really excels is at the standard class.

The other thing note is that the model’s specificity value regarding Poor class is 0.88, this is an indication that the model successfully identifies the cases where it is not poor 88 percent of the time, and this can also be said about good class.

Regarding sensitivity results it was exceptionally good at identifying the cases where it belongs to the standard classification and as for the other two, it performed poorly.

*Model accuracy*

In conclusion, the model did not perform, when the results were compared with the test data, it only resulted in 52.82 percent accurate.

*Confusion Matrix 2*

On the other hand, if the model were to only train only on both poor and standard classification it would perform much greater, as you can see form the figure above, when compare with testing data it resulted with 64.34 accuracy.

To conclude, machine learning did not accurately depict the accurate picture of the relationship between credit history score and credit score as seen in both the visualization and the analytical test that were used.

# 4.0 General Conclusion

From this analysis we can see that the influence of credit mix and credit history age on credit score is quite big. However, based on the correlation, the influence of annual income on credit score is surprisingly low.

Credit History Age and Credit Score: Longer credit histories positively impact credit scores, suggesting that the duration of credit use is a critical factor in creditworthiness.

In conclusion, the analysis underscores the multifaceted nature of credit scores, emphasizing the need for comprehensive approaches in assessment and education. Future research and policy efforts should refine credit scoring methodologies and support diverse consumer needs for equitable financial outcomes.

Last, Financial institutions should adopt a holistic approach to credit assessment, incorporating multiple factors beyond yearly income, such as credit history age and credit mix, for a more accurate evaluation of creditworthiness. Additionally, educational programs should be developed to inform consumers about the importance of maintaining a diverse credit mix and long-term credit histories to improve their scores.

# 5.0 Workload Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Work | Shong Ming Xuan TP074250 | Chee Kai Jian  TP074250 | Wong Wei Jun  TP068217 |
| Introduction | 40% | 40% | 20% |
| Data Preparation | 20% | 60% | 20% |
| Data Analysis | 33% | 33% | 33% |
| General Conclusion | 33% | 33% | 33% |

# 6.0 Reference

Articles. What is a Credit Mix? - Benefits of Credit Diversity | Equifax®. (n.d.). <https://www.equifax.com/personal/education/credit/score/articles/-/learn/what-is-a-credit-mix/#:~:text=Simply%20put%2C%20a%20credit%20mix,of%20calculating%20credit%20scores)%20used>.

*Cramer’s V*. R. (n.d.-b). <https://search.r-project.org/CRAN/refmans/confintr/html/cramersv.html>

Singh, G. (2024, May 21). *ANOVA: Complete guide to statistical analysis & applications (updated 2024)*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/>

(4932) total words