PART 2

7.0 Chapter 7 ( Literature Review )

7.1 Modern banking system

Bank act as intermediate between depositors and borrowers and also at the same time serve as a middleman between buyers and people who sell shares in the stockbroker’s industry. The figure below shows the banking firm intermediary, where the x axis shows the rate of interest and the y- axis shows the volume of deposits or loans.

A diagram of a graph

Description automatically generated

Figure 14

Based on the figure above the supply of deposits curve has upwards slope and there is supply of loans curve showing upwards slope where the bank will increase loan value as interest rates increasing. The demand for loans curve shows that it decreasing if the interest rate increasing. It shows that for the the market interest rate is indicating that the interest rate would succeed in a competitive market with no intermediation cost. The volume of business indicated by but there is exception where bank induce costs such as monitoring, verification, search, and enforcement costs. All of these are done to build credit worthiness of potential borrowers. Big corporations and companies use bank loans as chunk of their external financing due loan agreements indicated that in the financial markets that borrowers are creditworthy (Stiglitz and Weiss, 1988).

Bank organizational structure exists because the main goal of bank is to maximize potential of profit. The way of organizing economic activity is through a firm which provides alternatives method to market transaction as it is more efficient to command people than depend on market price (Coase, 1937). Coase theory states that payment services to customers exist due to intermediates that occurs between borrowers and lenders and thus create the traditional banking system. This system is much more efficient to be operate under organizational structure due to loans and deposit are internal to bank. Later on, this idea is refined where emphasis is made on monitoring function of firm and creation of incentives system (Alchian and Demsetz, 1972). Another approach argues that when there is scepticism happen, firm allowed to economise on cost outside of contracts (Williamson, 1981).

There is level of banking system in terms of their hierarchy as bank can be a financial conglomerate where a financial conglomerate usually a big group of companies that in the business of banking, insurance, and investment services. Financial conglomerate can be defined as firm that comprise at least two out of five financial activities such as insurance, payments, corporate finance, fund management and retail. (Briault,2000) An argument that has been made on financial conglomerate such as big banking corporations is that financial institutions usually expand financial functions with subsidiaries around the world as this approach allows them to survive economic downturns.

7.2 Determinant of profit and shareholder value creation in banking

There are some questions that need to be raised on how a bank making money or profit for its owners such as various factors that affecting the performance of the bank. Moreover, it is quite surprising to find that creating value for shareholders by creating return in excess of cost oof capital has been used in most banks in the past decades. It is found that in a study showing that there is a link or relationship with how bank productive efficiency correlated to stock returns generating positive relationship (Beccalli et al., 2005, Fernandez et al., 2002, Eisenbeis et al. 1999, Chu and Lim, 1998).

Bank performances are measured using shareholder value and profits, whereby profits are defined as to gain income superior to costs over period of time. On the other hand, the value of a shareholder is created when the return on investment for the capital is higher than the opportunity costs. There are factors that have impact on bank profits such as bank efficiency, risk management ability, bank competitive strength and bank financial structure.

In the banking world cost efficiency is defined as the capability of a firm to choose input or output levels and mixing of these both to minimize cost. Meanwhile, profit efficiency is the ability of a bank to produce the highest possible profit at given level of input and output prices. The relationship of higher efficiency shows that it will result in bad influence on free cash flow and later on the bank returns. Another ability to achieve return is influenced by risk taking bias whereby there are several main types of risk exist in banking such as credit risk, market risk and liquidity risk and operational risk. There are several studies carried out to deal with credit risk where the issues deal with things such as the sufficiency of new capital prerequisite in regard to credit risk or liability management implemented in banking industry (Jacobson et al., 2006), measurement technique (Duffie 2005, Lucas and Klaassen, 2006 and Galluccio and Roncoroni, 2006); and relationship with other risks (Zheng 2006 and Jobst et al., 2006).

The shareholder profit and value are influenced by how capable the bank run and operate the financial market activity and the market risk undertaken. The market risk undertaken can be defined as risk that investment value will decrease because of interest rates, inflation, market sentiment, economic conditions, and geopolitical events. The market risk undertaken also can be defined as risk of losses in on and off-balance sheet positions soaring from market prices movement. There is also operational risk exist in banking where it is the risk of loss due to unsuccessful internal process, people and system that occurs from external factors by looking at measurement issues (Scandizzo 2005, De Fontnouvelle et al., 2005).

On the other hand, liquidity risk is the risk that exist where there is a potential difficulty of buy and sell assets at profitable price because of poor market activity. It becomes less significant in a lesser liquid market or when in period where market stress occurs. It is also a risk that a bank holds inadequate liquid assets and it incapable to match conditions without damaging its financial capital. Another factor that influences a bank profit is the bank financial structure itself where it is one of determinant factor of profit and shareholder value creation. It is founded that companies that provide a bond rating above the S&P investment trade level usually have a higher price various on net income and lower pricing various on book value relative to a less healthy firms (Barth et al.,1998).

* 1. Factors that influence bank profitability

There are several studies conducted on bank profitability across several European countries, north America, African countries, emerging market economies and China. In several European countries there is a dependent variable that study the profitability of European banking such as net profit ratio before tax over capital and reserves, net profit ratio after tax over capital and reserves, net profit ratio before tax over total borrowings, reserves, and before tax over capital, net profit ratio before tax before over total assets, net profit ratio before tax and staff overhead over total assets, and net profit ratio before staff overhead, taxes and arrangement for loan before loan losses over overall assets.

In general, there are several factors to determine bank profitability such as stable and longstanding bond rate, increase in money supply, bank ownership, bank asset concentration ratio, net ratio of capital and reserves over the total assets, annual inflation rate, staff expenses ratio over total assets, ratio of cash, and bank investment and deposits securities over total assets. It can be deduced that there are factors that have positive outcome and impact on bank cashflow or profitability whereby the non-interest income, capital and concentration has the most impact. There are several case studies conducted on sample of Greek bank whereby specific case used where bank profitability is determined under Generalized Method of Moments (GMM) system estimator in which bank profitability is measured with profitability indicators such as Return on Equity (ROE) and Return on Assets (ROA). (Athanasoglou et al, 2008).

7.4 Bank loan collection from its customers

It is known that the way bank financing loan works is that bank make profit off from the interest rates. In a layman term the borrowers need to repay the borrowed funds, loan, or money at higher interest rate than what is paid to depositors. The profit is the difference between interest paid and interest received. Hence it can be called debt for someone that borrows money from a banking or any financial institution that is giving out loans.

There are two types of debt, which is good debt and bad debt and in this case of Lasiandra Finance Inc the type of debt it helping its customers is good debt. This is due to the nature of this financing company that provides loans and funds to small companies; hence it falls under good debt. Figure below shows the comparison between a good debt and a bad debt.

A diagram of a comparison of credit cards

Description automatically generated with medium confidence

Figure 15

A good debt is money borrowed for appreciating assets such as real estate student loan, meanwhile a bad debt is for depreciating assets such as cars, credit card. There must be an effective way for any bank or financial institution to handle debt collection. The traditional debt collection process starts with credit assessment, customer loan approval, payment collection, database collection on late payment, customer follow up and debt recovery.

However, there is weakness and issues that exist in traditional debt recovery or collection such as manual credit assessment is very prone to human error and a very tedious job to do. Next is due to communication error between the lenders the bank itself with the defaulters or the borrower due to each individual has a different style, profile, and background. Hence, the approach by the bank management to assume every borrower to react and behave the same totally hinders the process of debt collection efficiently. Next is the debt collectors may harass or kept on calling the defaulters whereby effective solution needs to be implemented to ease the process of debt collection.

Hence an effective debt collection should be adopted by the banking or financial institution such implemented data driven solution using dashboard monitoring system to monitor customer behavior. This approach allows the banker to study the pattern of their debtors and identify the debtors with high default risk. Next approach is making use of alternative data such the economy, recession, market confidence with investors, and employment rate whereby these factors influence debtor ability to pay loan or debt. Dynamic model approach should be used whereby it incorporated AI and machine learning to identify and classify debtors into a category of high-risk debt or in the category of bankruptcy using both dynamic and static data to give early warning alert. After that, imposed an enhance recommendation system that act as backup plan whereby using AI recommendations system the lenders can provide alternative payments to the debtors such extended the duration of payment and reduce the monthly payment. Lastly, the financial institutions or banks should incorporate behavioral science to study their debtor’s pattern on paying their debt with the use of data as it will solve any risk or money loss from the bank.

7.5 Machine learning in loan process application

Nowadays many banking start to implement machine learning in their loan application process whereby in loan prediction based on the applicant background such as gender, age, income, and more dependent variables. There are several features that influence bank loan approval such as credit history, total amount of assets, career, and lifestyle by utilizing machine learning algorithm to predict the loan status of a new applicant. There is literature review conducted regarding machine learning on loan process applications whereby python programming languages is used using three algorithm such as random forest, decision tree and logistic regression (Kumar, Rajiv, et al. 2019). There is some improvement that can be made from previous study whereby data pre-processing can be implemented to get rid of any anomalies occurs in the banking datasets. (Supriya, Pidikiti, et al. ,2019).

7.6 Outcome of the research

It can be deduced that there are five subtopic of the literature review which are the modern banking system, determinant of profit and shareholder value creation in banking, factors that influence bank profitability, bank loan location from its customers, and machine learning in loan process application. It can be deduced that banks play a critical role in the economy as they serve as financial bridges between borrowers and lenders to stimulate economic growth. The banking industry also offers diverse services such as wealth management, investment, payment processing and more as banking as this diversity indicates that bank adaptability to evolving market. The banking industry also works within the regulatory framework where the central banking system must oversee banks to guaranteed consumers’ rights are protected, ensuring stability and compliance with laws and regulations.

8.0 Chapter 8 (Data Analysis and Data Cleansing)

8.1 Analysis of the Categorical variables found in DAP67696.TRAINING\_DS

In chapter 8, there will be data analysis and cleansing done on the DAP67696.TRAINING\_DS and DAP67696.TESTING\_DS where several analyses will be done on both categorical and numeric/continuous variables. Before that, brief explanations on categorical variable whereby categorical variables are a type of variable that used in data analysis and statistics to visualize and represent data that can be split and divided into certain distinct class and group.

8.1.1 Univariate Analysis of the Categorical variable

Univariate analysis is a type of statistical analysis where it involves studying single variables from a dataset at a time. In a laymen term it involves without relationship with other variables while examine their statistical patterns and distributions as the main goal is to investigate the data central tendency, dispersion and more.

Below shows the list of categorical variables found from the TRAINING\_DS but only three variables are selected for the data analysis and data cleaning which are the gender, marital status, and loan location for this univariate analysis.

* GENDER
* MARITAL\_STATUS
* FAMILY\_MEMBERS
* QUALIFICATION
* EMPLOYMENT
* LOAN\_HISTORY
* LOAN\_LOCATION
* LOAN\_APPROVAL\_STATUS

8.1.2 Univariate Analysis of the Categorical variable - GENDER

8.1.3 SAS Source Codes

A close-up of a computer screen

Description automatically generated

Figure 16

8.1.4 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 17

8.1.5 Description

For the Univariate analysis of categorical variables – GENDER, from figure above it shows that there are two categories of GENDER variables which are male and female. The distribution of male is higher than male, and it also shows that thirteen gender value are missing as shown by frequency missing = 13.

8.1.6 Univariate Analysis of the Categorical variable – MARITAL\_STATUS

8.1.7 SAS Source Codes

A close-up of a computer screen

Description automatically generated

Figure 18

8.1.8 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 19

8.1.9 Description

For the Univariate analysis of categorical variables – MARITAL\_STATUS, from figure above it shows that there are two categories of MARITAL\_STATUS variables which are married and not married. The distribution of number of people married is higher than people who are not married, and it also shows that three MARITAL\_STATUS variable value are missing as shown by frequency missing = 3.

8.2 Univariate Analysis of the Categorical variable – LOAN\_LOCATION

8.2.1 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 20

8.2.2 Screenshot(s) of the Output

A screenshot of a graph

Description automatically generated

Figure 21

8.2.3 Description

For the Univariate analysis of categorical variables – LOAN\_LOCATION, from the figure above it shows that for the LOAN\_LOCATION variables there are three categories which are city, town, and village. Based on frequency value, the town category occurs more often than other categories. There is a logical sense on the number is like this for loan location to be higher in town than city due to the geographical location. This is due to city maybe dense and populated but it falls under one municipality or county whereas for town it is bounded by different municipalities. Hence, the finance company should study and look at factors such as the borrower history of debt, credit score, location of living, income before extending the period of loan to maximize profit.

8.2.4 Univariate Analysis of the Continuous/Numeric variable

This part here is to do data analysis on continuous/numeric variables found in the TRAINING\_DS whereby unlike categorical variables, the continuous/numeric variables is a type of variable that has the infinite number of values for a given range. Continuous/ numeric variables possessed a characteristic of measurement that is continuously or repetitively and also include uncountable number of possible values.

Below shows the list of continuous/ numeric variables for the TRAINING\_DS but there are only three variables selected for the data analysis which are the guarantee income, loan amount and loan duration.

* CANDIDATE\_INCOME
* GURANTEE\_INCOME
* LOAN\_AMOUNT
* LOAN\_DURATION

8.2.5 Univariate Analysis of the Continuous/Numeric variable – GUARANTEE\_INCOME

8.2.6 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 22

8.2.7 Screenshot(s) of the Output

A screenshot of a computer screen

Description automatically generated

Figure 23

8.2.8 Description

Based on the figure above it shows the univariate analysis of the Continuous/Numeric variable – GUARANTEE\_INCOME where the ‘N’ variable represents the number of observations of this dataset which is 614. The bar chart shows the distribution of the variable whereby it is right skewed as the right side extends further and more data concentrated on the left side of the bar chart. Other than that, the mean value is higher than the median value which is one of the characteristics of right-skewed distribution and also indicates that it has not equal central tendency.

Other than that, there is no frequency value missing for Continuous/Numeric variable – GUARANTEE\_INCOME and due to this right-skewed distribution of the data it indicates that most of the people or the population with guaranteed income are low-income earners than high income earners. This indicates that there is huge gap in income distribution among the population or the group itself.

8.2.9 Univariate Analysis of the Continuous/Numeric variable – LOAN\_DURATION

8.3 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 24

8.3.1 Screenshot(s) of the Output

A screenshot of a graph

Description automatically generated

Figure 25

8.3.2 Description

Figure above shows the univariate analysis of the Continuous/Numeric variable – LOAN\_DURATION, whereby for this time the graph of the LOAN\_DURATION indicates that it is mostly left- skewed. This is due to the distribution of the bar chart has it tails extend to the left side and most of the data concentrated on the right. From the table it shows that the median value is slightly higher than mean which indicates it is left skewed and it has 14 number of observations missing from the total 600 number of observations.

Loan duration has the same concept as loan period or loan term where the duration is set and agreed between the lender and borrowers. For small medium businesses, the loan duration is typically between lenders and borrowers. The period of loan duration significantly influences the monthly payment of the borrower and the total cost of borrowing. The loan duration ranges from 0 minutes to 500 minutes and most of the loan duration concentrated between 300 to 400 minutes.

8.3.3 Univariate Analysis of the Continuous/Numeric variable – LOAN\_AMOUNT

8.3.4 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 26

8.3.5 Screenshot(s) of the Output

A screenshot of a graph

Description automatically generated

Figure 27

8.3.6 Description

Figure above shows the univariate analysis of the Continuous/Numeric variable – LOAN\_AMOUNT, where from it can be noticed from the bar chart above the graph is right skewed as the right side extends further and more data concentrated on the left side of the bar chart. As usual the original number of observations is 614 but there are 22 number of observations that are missing, lost or displaced.

This can be represented by the N Miss 22 as shown in figure above and also one of the characteristics of right skewed graph distributions is that the value of mean is higher than the median. Based on the graph above the loan amount made ranges from 7 to 800 whereby the most loan amount approved is around 100 to 200. This graph distribution correlated to the variable of GUARANTEE\_INCOME whereby most of the population belong to lower income category, hence the reason the loan amount approved is high in the region of 100 to 200 and not in the region of 600 because of 100 to 200 belong to loan amount approved for low-income earner and 600 belong to high income earners.

The amount of loan guaranteed by lenders depends on the income and credit score of the borrowers as bank assumed that wealthy borrowers posed less risk compared to lower income one.

8.3.7 Bivariate analysis of the variables found in DAP67696.TRAINING\_DS

8.3.8 Introduction

Bivariate analysis is when study of two data is being conducted for example studying a group of college students to fine the average of SAT score and the age. For example, the variable can be categorical vs categorical or categorical vs continuous variable or continuous vs continuous. The keywords here is it focused on two variables instead of one like univariate analysis. The main purpose of bivariate analysis is to identify patterns and study the relationship between the two variables.

8.3.9 Bivariate Analysis of the variables (categorical vs categorical)

For this analysis, there are three bivariate analyses of the variables will be done which are GENDER vs MARITAL\_STATUS, GENDER vs EMPLOYMENT, and LOAN\_LOCATION vs LOAN\_HISTORY

8.4. Bivariate Analysis of the variable – (GENDER vs MARITAL\_STATUS ); (Categorical vs categorical variable)

8.4.1 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 28

8.4.2 Screenshot(s) of the Output

A screenshot of a data

Description automatically generated

Figure 29

A screenshot of a graph

Description automatically generated

Figure 30

8.4.3 Description

Based on the screenshot output of the GENDER vs MARITAL\_STATUS it shows the relationship between these two variables and the total number of observations is 598 out of the 614 data points. This indicates that 16 data points are missing, and the variable marital status are indicated with married and not married and variable gender is indicated by male and female. The number of male populations is significantly higher than female whereby the population of male is 487 meanwhile female is at 111.

The marital status of people who are married irrespective of their gender; male or female is at 388 meanwhile for people who are not married is at 210. This indicates that in terms of marital status there are more people who married compared to those who did not. By looking at the information given above, there are more not married woman compared to married women, meanwhile there are more married men than unmarried men. Assumptions can be made that most of the applicant who are eligible for loan is mostly man who are working and essentially, they are business owner and most likely men who are working are married and they have job to support kids and wife at home.

8.4.4 Bivariate Analysis of the variable – (GENDER vs EMPLOYMENT); (Categorical vs categorical variable)

8.4.5 SAS Source Codes

A computer code with text

Description automatically generated with medium confidence

Figure 31

8.4.6 Screenshot(s) of the Output

A screenshot of a computer screen

Description automatically generated

Figure 32

A graph with red and blue squares

Description automatically generated

Figure 33

8.4.7 Description

Based on figure above it shows the relationship between the GENDER vs EMPLOYMENT variable where again gender variables contain female and male. On the other hand, the employment variable contained either yes or no which indicates if the individual employed or not employed with a job. Employment is one of the most crucial factors in approving loans, especially for companies like lasiandra finance that lend money to small startups.

Based on the table it shows that there are 569 number of observations which means 45 number of observations are missing from the total of 614. Again, the male population is higher than female and for this time the number of people who are not employed for both genders are higher than people who are employed. Hence, it can be deduced that this trend correlated with guarantee income variables where this shows that lot of the population for both male and female don’t have any source of income. This maybe not be a good indicator as banks take this as risks to give out loan to people who are not employed as there is no guaranteed that the payment can be done on monthly basis.

8.4.8 Bivariate Analysis of the variable – (LOAN\_LOCATION vs LOAN\_HISTORY); (Categorical vs categorical variable)

8.4.9 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 34

8.5 Screenshot(s) of the Output

A screenshot of a computer screen

Description automatically generated

Figure 35

A graph of a distribution of loan

Description automatically generated

Figure 36

8.5.1 Description

Based on the figure above it shows the relationship between LOAN\_LOCATION vs LOAN\_HISTORY where the variables of loan contained the likes of city, town, and village. On the other hand, the loan history has two categories which are the ‘0’ and ‘1’, whereby in data analysis, label ‘0’ represents that the borrower does not have any record history of loans. The label ‘1’ on the other hand indicates that the borrower has a history of loans and these two labels in the context of data analysis refer to binary variables of whether a borrower has history of taking loans from any bank or financial institution.

From the information given, data points have 564 total of data points out of the original 564 of the number of observations and 50 number of observations are missing. From the variable above, most of the population lives in town where the number of people lives there is 217 people. To make comparison between different loan location with the city, town and village with the loan history, it is found that that most of the population has loan history which means that most of them have history of borrowing money from the bank. It is found that the people from the town have the history of borrowing money the most and which makes sense due to it high population and thus more banking activity or money transaction happening compared to village and city.

8.5.2 Bivariate Analysis of the variables (categorical vs continuous)

8.5.3 Bivariate Analysis of the variable – (GENDER vs LOAN\_DURATION); (Categorical vs continuous variable)

8.5.4 SAS Source Codes

A screen shot of a computer code

Description automatically generated

Figure 37

8.5.5 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 38

8.5.6 Description

The figure above shows the relationship between the categorical variable (GENDER) vs numeric/continuous variable (LOAN\_DURATION) for the bivariate analysis where from the information above it is shown that there are 112 female and 489 males. From the information above the minimum loan duration or processing for female is 36 min and the maximum is 480 minutes. Meanwhile for male is 12 min and the maximum is 480 min same as female. There is such a huge disparity in terms of processing time for loan duration for male and female as male take faster time to process than female.

This is due to bias by the banking system as they favor male applicants as the male population has guaranteed income and job and hence it is easier for them to process the loan duration time compared to female. This is also due to high male population than female and hence it is better for the business to lower the processing time or the loan duration.

8.5.7 Bivariate Analysis of the variable – (LOAN\_HISTORY vs CANDIDATE\_INCOME); (Categorical vs continuous variable)

8.5.8 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 39

8.5.9 Screenshot(s) of the Output

A screenshot of a computer screen

Description automatically generated

Figure 40

8.6 Description

The figure above shows the relationship between the LOAN\_HISTORY vs CANDIDATE\_INCOME which the analysis is done for the bivariate analysis of both variable. The variable loan history has binary variable of 0 and 1 where 0 indicates that borrower has no loan history borrowing money from the bank meanwhile for 1 it has loan history borrowing money from the bank. The number of observations or people that has loan history ‘0’ or no loan history is lower than the people who has loan history ‘1’. There are 475 people who has history of making loans and 89 people who have not single loan in their life. The minimum amount and maximum amount of candidate income of people who have no loan history is 1500 and 81000 compared with people who has loan history at 150 and 63337. The binary value 1 shows that there is credit history and the range of income of people who has candidate history is between 150 and 63337 which make sense due to most of the population is low-income earner compared to only small population who makes more money but has no credit history of 0. It can be deduced that low-income earner has the most loan history compared to high income earner by looking at the number of observations and minimum and maximum candidate income.

8.6.1 Bivariate Analysis of the variable – (QUALIFICATION vs LOAN\_AMOUNT); (Categorical vs continuous variable)

8.6.2 SAS Source Codes

A screen shot of a computer

Description automatically generated

Figure 41

8.6.2 Screenshot(s) of the Output

A screenshot of a computer screen

Description automatically generated

Figure 42

8.6.3 Description

Based on figure above it shows bivariate analysis of the variable QUALIFICATION vs LOAN\_AMOUNT whereby the variable qualification indicates two categories which are the population that is graduate and undergraduate. Graduate indicated a person who pursue education beyond bachelor’s degree meanwhile undergraduate only with bachelor’s degree. From the information given above the number of graduates are higher than undergraduate at 480 compared to 134 and the loan amount graduate is higher at 700 compared to 279. Often the time, qualification has link with the level of income, hence lenders, bank or any financial institution will give the maximum amount of value for loan amount compared to undergraduate population. This is due graduate student who pursue master’s usually has job beforehand and several years of working experience, so the loan amount set by the lenders is the maximum due to the ability to repay the debt or the monthly installments.

8.6.4 Analysis of the Categorical variables found in DAP67696.TESTING\_DS

8.6.5 Univariate Analysis of the Categorical variable using SAS MACRO

8.6.6 Introduction

SAS macro is a unique programming feature that helps programmers to avoid writing repetitive code and reuse that code when necessary. Furthermore, SAS macro helps programmers in creating dynamic variables that can take on changing values that occur in the code. Hence, SAS Macro variables are SAS variables that stored or keep values in SAS program that can be utilized repeatedly. Univariate analysis will be done on the categorical variable found in the DAP67696.TESTING\_DS whereby three categorical variables are selected which are the loan location, marital status and gender from the testing dataset. Based on the SAS code below all the seven categorical variables are executed but only three variables selected for the analysis.

8.6.7 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 43

A white background with black text

Description automatically generated

Figure 44

A white background with text

Description automatically generated

Figure 45

8.6.8 Screenshot(s) of the Output

A screenshot of a graph

Description automatically generated

Figure 46

A screenshot of a graph

Description automatically generated

Figure 47

A screenshot of a graph

Description automatically generated

Figure 48

8.6.9 Description

Based on the three figures above it shows the LOAN\_LOCATION, GENDER and MARITAL\_STATUS are the categorical variables selected for the univariate analysis using SAS MACRO. From the information above the categorical variable GENDER has missing number of observations at 11. Because of its univariate analysis it only study single variable and hence the outcome or output of the SAS code is that it shows people who are married and not married, the gender of female and male and the loan location. In terms of loan location, people in the city have the most frequency of loan due to cost of living in the city is very high and hence taking loan and debt is quite the norm. The least frequent people taking out loan is in the village due to there is less business activity or money transaction compared to in the city or town. The same with marital status as people who are married more likely to get loan due to high commitment such as mortgage house payments to provide a roof for family compared to not married person.

8.7 Univariate Analysis of the Continuous/Numeric variable – using the SAS MACRO

There are four continuous/numeric variable from the testing dataset which are the candidate income, guarantee income, loan amount and loan duration. All these four variables will undergo data analysis using univariate which mean examine individual variable using SAS MACRO method.

8.7.1 SAS Source Codes

A screenshot of a computer program

Description automatically generated

Figure 49

A computer code with text

Description automatically generated

Figure 50

A screenshot of a computer

Description automatically generated

Figure 51

8.7.2 Screenshot(s) of the Output

A screenshot of a computer screen

Description automatically generated

Figure 52

A screenshot of a computer screen

Description automatically generated

Figure 53

A graph of a bar graph

Description automatically generated with medium confidence

Figure 54

A screenshot of a computer screen

Description automatically generated

Figure 55

A graph of a graph

Description automatically generated

Figure 56

A screenshot of a computer screen

Description automatically generated

Figure 57

A graph with numbers and a bar

Description automatically generated

Figure 58

8.7.3 Description

Based on the figure above shows previously the in the SAS code and the screenshot of the output it shows the data analysis of four continuous/numeric variables where the SAS MACRO univariate analysis of those four variables from the TESTING datasets. For the univariate analysis of the TESTING dataset, it is found that that the original data points (N) or the number of observations is 367 compared with TESTING datasets at 614. By looking at the univariate analysis of the categorical variable candidate income it is found that the standard deviation is greater than the mean same with the guarantee income. This shows there is a dispersion in the data and the datapoints has skewed distributions and outliers exist in this graph distribution and on the other hand, the loan amount has missing values same with loan duration.

There are quite few methods to reduce monthly payments when taking out loans which are through the refinance loans by going for the one with lower interest rates, extending the loan period for example from 15 years to 30 years, and consolidate debt taken such credit card debt to a lower interest loan rate.

8.7.4 Bivariate Analysis of the variables – (Categorical vs categorical variable) using SAS Macro

8.7.5 SAS Source Codes

A computer screen shot of a program

Description automatically generated

Figure 59

A computer code with blue text

Description automatically generated with medium confidence

Figure 60

A screenshot of a computer code

Description automatically generated

Figure 61

8.7.6 Screenshot(s) of the Output

A screenshot of a data analysis

Description automatically generated

Figure 62

A graph of a distribution of gender

Description automatically generated

Figure 63

A screenshot of a computer

Description automatically generated

Figure 64

A graph with red and blue squares

Description automatically generated

Figure 65

A screenshot of a data

Description automatically generated

Figure 66

A graph of a number of people

Description automatically generated with medium confidence

Figure 67

8.7.7 Description

This part here is for the data analysis of the bivariate Analysis of the variables – (Categorical vs categorical variable) using SAS Macro where the first one is GENDER vs LOAN\_LOCATION, next is GENDER vs LOAN\_HISTORY and the last one is MARITAL\_STATUS vs LOAN\_LOCATION. Again, this data analysis used macro function on TESTING datasets whereby looking at the variables above GENDER vs LOAN\_LOCATION and GENDER vs LOAN\_HISTORY has frequency missing number of observations meanwhile MARITAL\_STATUS vs LOAN\_LOCATION is not.

It is found that loan location city and gender male is the most when it comes to applying for loan due to city is the place where most of business transaction takes place. The reason why men dominated is due to most of them are bread winner of the family and hence loan is frequent. By looking at GENDER vs LOAN\_HISTORY, it is found that each gender has history of taking loans and male are the highest of having taking loan previously compare to female. In terms of MARITAL\_STATUS vs LOAN\_LOCATION, its is found that married couple is more likely taking out loans compared to unmarried one and most of them living city as people in the city as people in the city has established credit with bank and more likely have success in convincing banks extend period of credit than people in town or village.

8.7.8 Bivariate Analysis of the variables – (Categorical vs continuous variable) using SAS Macro

8.7.9 SAS Source Codes

A computer screen shot of a program

Description automatically generated

Figure 68

A computer screen shot of a computer code

Description automatically generated

Figure 69

A screenshot of a computer

Description automatically generated

Figure 70

8.8 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 71

A table with numbers and letters

Description automatically generated

Figure 72

A screenshot of a data analysis

Description automatically generated

Figure 73

8.8.1 Description

This part here is for the data analysis of the bivariate Analysis of the variables – (Categorical vs continuous variable) using SAS Macro where the first one is LOAN\_LOCATION vs LOAN\_AMOUNT, next is FAMILY\_MEMBERS vs LOAN\_DURATION and the last one is EMPLOYMENT vs GUARANTEE\_INCOME. From the figure above it is found that people who live in the city compared to town and village do the most loan activity, but their maximum amount of loan is surprisingly lower than people in town. From the information given it is quite surprising to find that people with 0 family members take out or get approved loan the most compared to people with family members where their minimum loan duration is the lowest compared to rest. This is due to lower minimum loan duration, faster debt repayment, and lower interest cost. On the other hand, it is found that unemployed people have the maximum amount of guaranteed income due to several reasons such as unemployment benefits, social assistance programs by the government and income support initiatives and most of the population are unemployed.

8.8.2 Data Cleaning

Data cleaning is the second part of chapter 8 where handling missing values, removing duplicate data, remove errors and inaccuracies, standardizing data, handling consistent data are done to ensure the data used has quality. It is one of the most crucial step in data analysis as it is needed to be done to ensure that the data has consistency, maintaining the data integrity, removing bias in the data and a clean data is essential to train machine learning model to produce accurate predictions or classifications.

8.8.3 Imputing the missing values found in the categorical variables in the datasets DAP67696.TRAINING\_DS

Hence data imputation will impute missing values found in the categorical variables in the DAP67696.TRAINING\_DS where for this data cleaning the selected variables are GENDER, MARITAL\_STATUS and FAMILY\_MEMBERS.

8.8.4 Imputing the missing values found in the categorical variables - GENDER.

8.8.5 SAS Source Codes

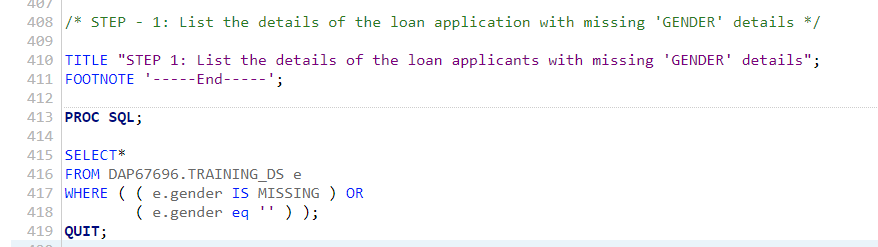


Figure 74

8.8.6 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 75

8.8.7 Description

Figure above shows the list of details of loan applicants with missing GENDER details where there around 13 applicants with missing GENDER details and around 10 of the applicants are married and 3 of them are single and not married.

8.8.8 SAS Source Codes

A close-up of a computer screen

Description automatically generated

Figure 76

8.8.9 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 77

8.9 Description

The results of the code displayed as shown in figure above where number of loans applicants who submitted loan applications with missing GENDER details are 13.

8.9.1 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 78

8.9.2 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 79

8.9.3 Description

The step number 3 is done to find the statistics and save the statistics in the temporary datasets called DAP67696.TRAINING DS where all the data including gender and total counts are stored there. The output data shows the gender female and male.

8.9.4 SAS Source Codes

A screen shot of a computer

Description automatically generated

Figure 80

8.9.5 Screenshot of the Output

A screen shot of a computer

Description automatically generated

Figure 81

8.9.6 Description

The output of the code display above is to shows the finding of the highest count in the DAP67696.TRAINING DS where the gender male has the highest count.

8.9.7 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 82

8.9.8 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 83

8.9.9 Description

The figure above is to create a copy of dataset DAP67696.TRAINING\_DS whereby it create a backup copy of the dataset. The backup dataset are stored in the table called DAP67696.TRAINING\_BK\_DS.

8.0.1.1 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 84

8.0.1.2 Screenshot(s) of the Output



Figure 85

8.0.1.3 Description

The information given previously shows the imputation is done to find any missing values found in the variable GENDER. The results show that there are 13 new rows updated in DAP67696.TRAINING\_DS indicates that GENDER variable are imputed.

8.0.1.4 SAS Source Codes

A screen shot of a computer

Description automatically generated

Figure 86

8.0.1.5 Screenshot(s) of the Output

A close-up of a sign

Description automatically generated

Figure 87

8.0.1.6 Description

Step 7 indicates that the imputation lists successfully done for the missing variables and this shows that it successfully lists all the details of loan applicants with missing GENDER.

8.0.1.7 Imputing the missing values found in the categorical variables – MARITAL\_STATUS.

8.0.1.8 SAS Source Codes

A computer screen shot of a computer code

Description automatically generated

Figure 88

8.0.1.9 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 89

8.0.2 Description

The step 1 above shows the list of details of loan applications with missing marital status details and it is found out that there two men and a women who have applied for this role.

8.0.2.1 SAS Source Codes

A screenshot of a computer program

Description automatically generated

Figure 90

8.0.2.2 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 91

8.0.2.3 Description

Figure above shows the outcome whereby the statistics are saved in a temporary dataset called DAP6796.TRAINING where all the data including marital status and counts are stored. It is found that married individuals are 398 and not married are 213.

8.0.2.4 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 92

8.0.2.5 Screenshot(s) of the Output



Figure 93

8.0.2.6 Description

Step 3 is to impute missing values found in the variable MARITAL\_STATUS where the missing data are imputed, and the result shows 3 rows were updated in DAP67696.TRAINING\_DS.

8.0.2.7 SAS Source Codes

A computer screen shot of a computer code

Description automatically generated

Figure 94

8.0.2.8 Screenshot(s) of the Output

A close-up of a text

Description automatically generated

Figure 95

8.0.2.9 Description

Step 4 shows that imputation is done, and the information related to list the details of the loan applicants with missing MARITAL\_STATUS is executed successfully.

8.0.3 Imputing the missing values found in the categorical variables – FAMILY\_MEMBERS.

8.0.3.1 SAS Source Codes

A computer screen shot of a computer code

Description automatically generated

Figure 96

8.0.3.2 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 97

8.0.3.3 Description

Step 1 again is the method of listing the details of loan applicants with missing family members detail where there are around 15 applicants with missing family members details where it can be seen in figure above.

8.0.3.4 SAS Source Codes

A computer screen shot of a computer code

Description automatically generated

Figure 98

8.0.3.5 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 99

8.0.3.6 Description

Step 2 above shows to find the statistics and save the statistics in a temporary dataset DAP67696.TRAINING dataset where family members and counts are stored. It is found that 345 people have 0 family members, 102 have 1 family members, 101 have 2 family members and 51 people have more than 3 family members.

8.0.3.7 SAS Source Codes

A computer screen shot of a program

Description automatically generated

Figure 100

8.0.3.8 Screenshot(s) of the Output

A screenshot of a document

Description automatically generated

Figure 101

8.0.3.9 Description

Figure above shows the list of details of the loan applicants with ‘3+’ family members where there are 51 applicants with more than three family members.

8.0.4 SAS Source Codes

A white screen with blue and green text

Description automatically generated

Figure 102

8.0.4.1 Screenshot(s) of the Output



Figure 103

8.0.4.2 Description

Step 4 is done remove the ‘+’ symbol found in the family members and update the dataset DAP67696.TRAINING\_DS whereby there are 51 rows were updated in the DAP67696.TRAINING\_DS. This step is to remove the ‘+’ symbol as it is crucial to data cleaning of the datasets.

8.0.4.3 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 104

8.0.4.4 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 105

8.0.4.5 Description

The figure above shows the results of the output data for family members variable after removing the ‘+’ symbol. From the figure above the ‘+’ symbol is removed.

8.0.4.6 SAS Source Codes

A computer code with text

Description automatically generated

Figure 106

8.0.4.7 Screenshot(s) of the Output



Figure 107

8.0.4.8 Description

The code above indicates that imputation of the missing values found in the variable is done for the family members where 15 new rows were updated in the DAP67696.TRAINING\_DS. Also, it is also done to find the highest count in the datasets.

8.0.4.9 SAS Source Codes

A computer screen shot of text

Description automatically generated

Figure 108

8.0.5 Screenshot(s) of the Output

A close-up of a message

Description automatically generated

Figure 109

8.0.5.1 Description

Step 7 shows after indexation are done on the list of details of the loan applicants with missing family members details and it indicates that it is executed successfully.

It can be deduced that for the categorical variables, 3 categorical variables are cleansed which are the gender, marital status, and family members. There are other categorical variables which are the qualification, employment, loan history, loan location and loan approval status where data cleansing and cleaning is done on this variable but not documented. It can be deduced that out of these 8 categorical variables, 5 of them have missing values which are the gender, marital status, family members, employment, and loan history meanwhile qualification, loan location and loan approval status don’t have missing values.

8.0.5.2 Imputing the missing values found in the continuous variables – LOAN\_AMOUNT.

8.0.5.3 SAS Source Codes

A screen shot of a computer

Description automatically generated

Figure 110

8.0.5.4 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 111

8.0.5.5 Description

Step 1 above shows the list of details of loan applicants with missing LOAN\_AMOUNT details where there are around 22 loan applicants with missing LOAN\_AMOUNT details as the figure above shows the value of unknown loan amount variable. The data scientist should know that loan amount refers to the amount of money the borrower takes loan from the bank or any of the financial institutions. It can also be called a mortgage and the whole amount of loan includes the loan principal, any recurring interest, and interest on late payments. Mortgages are influenced by overnight policy rate (OPR) where if the OPR rate goes up so did the mortgage payment.

8.0.5.6 SAS Source Codes

A computer screen shot of a computer code

Description automatically generated

Figure 112

8.0.5.7 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 113

8.0.5.8 Description

Step 2 above shows the output or outcome of listing the details or the number of applicants with missing LOAN\_AMOUNT details where it is around 22 number of applicants. This shows that there are 22 applicants who applied for a loan but failed to provide the details of the loan amount needed.

8.0.5.9 SAS Source Codes

A computer screen shot of a training program

Description automatically generated

Figure 114

8.0.6 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 115

8.0.7 Description

Figure above shows step 3 which is the imputation step where imputation is done to find the missing values found in the variable LOAN\_AMOUNT where a copy of the dataset is created.

8.0.7.1 SAS Source Codes

A computer screen shot of a computer program

Description automatically generated

Figure 116

8.0.7.2 Screenshot(s) of the Output

A close-up of a white background

Description automatically generated

Figure 117

8.0.7.3 Description

The figure above shows that output for step 4 is empty which means the step is successfully executed. Step 4 shows the after-indexation step to list the details of the loan applicants with missing LOAN\_AMOUNT details.

8.0.7.4 Imputing the missing values found in the continuous variables – LOAN\_DURATION.

8.0.7.5 SAS Source Codes

A computer screen shot of a computer program

Description automatically generated

Figure 118

8.0.7.6 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 119

8.0.7.7 Description

Figure above shows step 1 of list the details of the loan applicants with missing LOAN\_DURATION details where it displayed the applicant information. This information shows people who submit loan applications without giving any details about the duration of the loan taken. In this analysis the TRAINING\_DS is used, and loan duration can be defined as repayment term where it refers to certain length of time of which the borrower obliged to repay the loan. There are three common loan durations which are short term loans, medium term loans and long-term loans.

8.0.7.8 SAS Source Codes

A computer screen shot of text

Description automatically generated

Figure 120

8.0.7.9 Screenshot(s) of the Output

A screenshot of a computer error

Description automatically generated

Figure 121

8.0.8 Description

Step 2 is done to find the number of loan applicants with missing LOAN\_DURATION details whereby there are 14 number of applicants who did not include loan duration in loan submission to the Lasiandra Finance company. Based on the information given, there are 9 married people and 4 people who are not married and typically loan can be considered as financial agreement between the borrower and lender, typically the bank. Three key features of loan are principal which is the amount of money must be paid to the lender over period of time, interest, the calculated percentage of loan principal or in laymen terms the cost of borrowing money and the last one is loan term which is the repayment term. Loan is very common for small startup to start a business and typically a small company that is still growing and expanding business prefer loan terms as it is a quick way to gain cash flow to run and operate the company.

8.0.8.1 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 122

8.0.8.2 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 123

8.0.8.3 Description

Step 3 again is to impute the missing values found in the variable LOAN\_DURATION where figure above shows the output of the data.

8.0.8.4 SAS Source Codes

A computer screen shot of text

Description automatically generated

Figure 124

8.0.8.5 Screenshot(s) of the Output

A close-up of a text

Description automatically generated

Figure 125

8.0.8.6 Description

Step 4 shows that the lists the details of the loan applicants with missing LOAN\_DURATIN details are successfully created as step 4 shows the output is empty after the indexation method.

8.0.8.7 Imputing the missing values found in the continuous variables – GUARANTEE\_INCOME.

8.0.8.8 SAS Source Codes

A computer error message

Description automatically generated

Figure 126

8.0.8.9 Screenshot(s) of the Output

A close-up of a white background

Description automatically generated

Figure 127

8.0.9.1 Description

Figure above shows step 1 of listing the details of loan applicants with missing GUARANTEE\_INCOME where it shows that there is no output. By comparison with step 1 for LOAN\_AMOUNT and LOAN\_DURATION, step 1 has the outputs the missing details found in that said variables. This is due to the GUARANTEE\_INCOME has no missing details found which indicated by no output generated when running the code.

8.0.9.2 SAS Source Codes

A computer screen shot of text

Description automatically generated

Figure 128

8.0.9.3 Screenshot(s) of the Output

A screenshot of a computer error

Description automatically generated

Figure 129

8.0.9.4 Description

Figure above shows step 2 of the analysis where step 2 is done to find the total number of loan applicants with missing GUARANTEE\_INCOME details where based on the output get the number of applicants is 0. This means that there is 0 number of applicants with missing GUARANTEE\_INCOME details which indicate that the data is clean and without any missing values.

8.0.9.5 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 130

8.0.9.6 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 131

8.0.9.7 Description

Step 3 is done to impute the missing values found in the variable GUARANTEE\_INCOME in the TRAINING\_DS where in this method the dataset was copied and stored in another location. From the figure above the output is recorded.

8.0.9.8 SAS Source Codes

A computer screen with text

Description automatically generated

Figure 132

8.0.9.9 Screenshot(s) of the Output

A white background with black text

Description automatically generated

Figure 133

8.0.0.1 Description

Step 4 above shows after the indexation method of listing the details of loan applicants with missing GUARANTEE\_INCOME details where it shows the step is executed successfully.

It can be deduced that for continuous/numeric variables there are 4 list of variables that need to be clean which are the loan amount, loan duration, guarantee income and candidate income. From these 4 variables only 3 are used for documentation purpose and it is found out that from 4 of these variables the variables with missing values are loan amount and loan duration, meanwhile guarantee income and candidate income don’t have any missing values.

9.0 Chapter 9 ( Model Creation and Prediction)

This part here is model creation and prediction where for model creation categorical variable are listed and for the prediction the response variable which is the dependent variable are the loan approval status. The rest of the variable is set as independent variables and below shows the code for the model creation and prediction.

9.1.1 SAS Source Codes

A screenshot of a computer program

Description automatically generated

Figure 134

9.1.2 Screenshot(s) of the Output

A screenshot of a logistic procedure

Description automatically generated

Figure 135

The analysis is accepted.

A screenshot of a computer

Description automatically generated

Figure 136

9.1.3 Description

Figure above shows the output of the model creation where it shows the LOGISTIC Procedure where it shows the response variable which is the LOAN\_APPROVAL\_STATUS, and the model is binary logit, and the optimization technique is Fisher’s scoring. Figure above shows that number of observations read is equal to number of observations which is 614 which indicates the analysis is accepted. One of the criteria of model prediction is that cleansing needs to be done on both train and test dataset in order to get the number of observations read equal to number of observations used. If the number of observations read and used are not equal, it is due to the data cleansing not done properly.

The model convergence status is acceptable.

9.1.4 Screenshot(s) of the Output



Figure 137

9.1.5 Description

The model convergence status shows that it is acceptable as shown in the figure above as the output indicates that the model convergence status is satisfied.

9.1.6 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 138

9.1.7 Description

Figure above shows the output of the predicted probability where the output of the predicted probability will be stored in the dataset. The Akaike Information Criterion must be less than the Schwarz Criterion as it holds the predicted probability information. Based on information above the AIC < SC as AIC= 764.891 and SC=769.311 which indicates the analysis is accepted as SC > AIC.

9.1.8 Screenshot(s) of the Output

A table of numbers and numbers

Description automatically generated with medium confidence

Figure 139

9.1.9 Description

Figure above shows the analysis of Maximum Likelihood Estimates where If Pr > ChiSq is <= 0.05, it means that independent variable is an important variable and is truly contributing to predict dependent variable which is shown by loan location – town with value of 0.007, loan history – 0 with value of 0.001 and marital status – married with value of 0.0235.

9.2 Screenshot(s) of the Output

A screenshot of a data sheet

Description automatically generated

Figure 140

9.2.1 Description

Figure above shows the output of the type 3 analysis of effects where if is lesser than 0.05 it shows that these variables that meet the criteria of variable importance. The variable importance means that if the value is less than 0.05 it indicates that the variables is the most contributing factor to the analysis of effects while others are not. This shows that loan location, loan history and marital status are the most contributing variables while others are not to predict the loan approval status.

9.2.2 SAS Source Codes\

A white background with text

Description automatically generated

Figure 141

9.2.3 Screenshot(s) of the Output

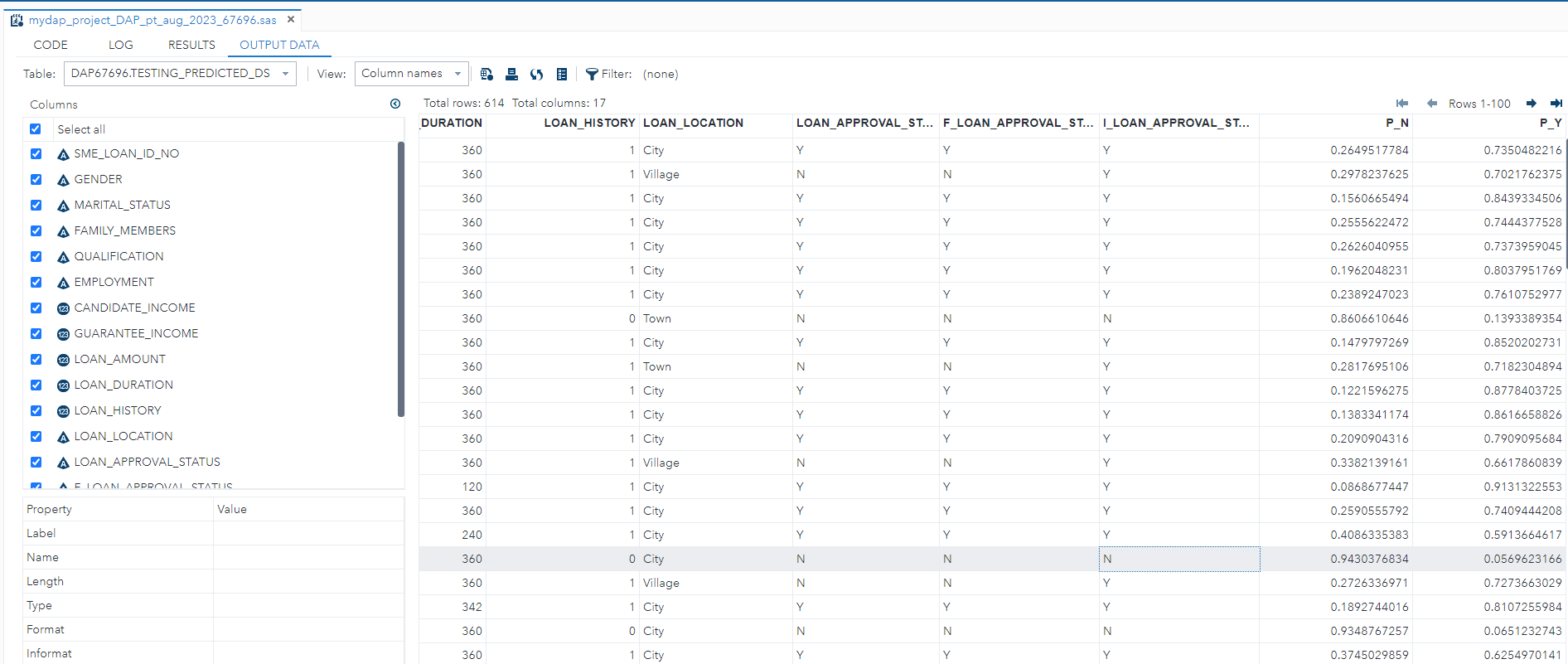


Figure 142

9.2.4 Description

Figure above shows the output of the DAP67696.TESTING\_PREDICTED\_DS where it shows the output of the data. The output of the data shows the P\_N and P\_Y value which shows the predicted value entry and predicted value exit. The score data is where the cleansed dataset is stored as cleaned dataset indicated by number of observations read equivalent to number of observations used.

9.2.5 List the details of the dataset carrying the loan approval status predicted- DAP67696.TESTING\_PREDICTED\_DS

9.2.6 SAS Source Codes

A close-up of a computer screen

Description automatically generated

Figure 143

9.2.7 Screenshot(s) of the Output

A screen shot of a computer

Description automatically generated

Figure 144

9.2.8 Description

Figure above shows the output of listing the details of the dataset carrying the loan approval status predicted for DAP67696.TESTING\_PREDICTED\_DS. From the information obtained above it can be deduced that both of the train and test dataset are being cleansed properly as there are no missing values of loan approval status as most of it are filled with N and Y.