



Individual Assignment

TECHNOLOGY PARK MALAYSIA

MODULE CODE-DAP

Data Analytical Programming

APUMP2205-DSBA(PR)

Student's TP: TP067696

Student's Name: Mr.Muhammad Arif Bin Jamaluddin

Lecturer's Name: Mr.DHASON PADMAKUMAR

ACKNOWLEDGMENT

On this acknowledgement part, I want to take this opportunity to express my gratitude and very special thanks to my lecturer Mr. Dhason Padmakumar who guided me and my classmates for this subject (DAP). His years and years of experience of teaching this subject really showed when he was conducting the online class as the pace of the class was really good. Moreover, I want to thank my classmates for their participation in the classroom.

In addition, I would like to express my special thanks to my family, especially my mother who has been supporting me throughout my master program. It has been quite a roller coaster journey for me, and I am glad that I am able to pull it off.

Also, I want to express my gratitude towards the staff of APU for helping me with the students matters and without them student life would not be easy. In conclusion, I want to thank God for helping me getting through all the challenges that I faced throughout this study.

TABLE OF CONTENT

Contents

ACKNOWLEDGMENT.....	2
TABLE OF CONTENT	3
1.0 Chapter 1 (Introduction)	12
2.0 Chapter 2 (Problem Statement).....	14
3.0 Chapter 3 (Background Of Lasiandra Finance Inc. (LFI), New York, USA)	16
4.0 Chapter 4 (ASSUMPTION, PROGRAM DEMONSTRATION, CODING AND JUSTIFICATION).	18
5.0 Chapter 5 (Methodology).....	20
6.0 Chapter 6 (Data Dictionary/ Metadata).....	22
6.1 Introduction	22
6.1 Location of the datasets on SAS.....	23
6.2 Project datasets found inside the SAS permanent library/folder.	23
6.2.1 Description.....	23
6.2.2 Display the structure of the datasets-DAP67696.TRAINING_DS	24
6.2.3 Data dictionary of the dataset- DAP67696.TRAINING_DS	24
6.2.4 SAS PROC SQL Codes	24
6.2.5 Screenshot(s) of the output	24
6.2.6 Description.....	25
6.2.7 SAS PROC SQL Codes	25
6.2.8 Screenshot(s) of the output	26
6.2.9 Description.....	27
7.0 Chapter 7 (Literature Review)	28
7.1 Modern banking system.....	28
7.2 Determinant of profit and shareholder value creation in banking	30
7.3 Factors that influence bank profitability	31
7.4 Bank loan collection from its customers	32
7.5 Machine learning in loan process application	34
7.6 Outcome of the research	34
8.0 Chapter 8 (Data Analysis and Data Cleansing)	35

8.1 Analysis of the Categorical variables found in DAP67696.TRAINING_DS	35
8.1.1 Univariate Analysis of the Categorical variable.....	35
8.1.2 Univariate Analysis of the Categorical variable - GENDER	35
8.1.3 SAS Source Codes.....	36
8.1.4 Screenshot(s) of the Output	36
8.1.5 Description.....	36
8.1.6 Univariate Analysis of the Categorical variable – MARITAL_STATUS	37
8.1.7 SAS Source Codes.....	37
8.1.8 Screenshot(s) of the Output	37
8.1.9 Description.....	37
8.2 Univariate Analysis of the Categorical variable – LOAN_LOCATION	38
8.2.1 SAS Source Codes.....	38
8.2.2 Screenshot(s) of the Output	38
8.2.3 Description.....	38
8.2.4 Univariate Analysis of the Continuous/Numeric variable.....	39
8.2.5 Univariate Analysis of the Continuous/Numeric variable – GUARANTEE_INCOME .	
8.2.6 SAS Source Codes.....	39
8.2.7 Screenshot(s) of the Output	40
8.2.8 Description.....	41
8.2.9 Univariate Analysis of the Continuous/Numeric variable – LOAN_DURATION.....	42
8.3 SAS Source Codes.....	42
8.3.1 Screenshot(s) of the Output	43
8.3.2 Description.....	43
8.3.3 Univariate Analysis of the Continuous/Numeric variable – LOAN_AMOUNT	44
8.3.4 SAS Source Codes.....	44
8.3.5 Screenshot(s) of the Output	45
8.3.6 Description.....	45
8.3.7 Bivariate analysis of the variables found in DAP67696.TRAINING_DS	46
8.3.8 Introduction	46
8.3.9 Bivariate Analysis of the variables (categorical vs categorical).....	46
8.4. Bivariate Analysis of the variable – (GENDER vs MARITAL_STATUS); (Categorical vs categorical variable).....	46

8.4.1 SAS Source Codes.....	46
8.4.2 Screenshot(s) of the Output	47
8.4.3 Description.....	48
8.4.5 SAS Source Codes.....	49
8.4.6 Screenshot(s) of the Output	49
8.4.7 Description.....	50
8.4.8 Bivariate Analysis of the variable – (LOAN_LOCATION vs LOAN_HISTORY); (Categorical vs categorical variable)	51
8.4.9 SAS Source Codes.....	51
8.5 Screenshot(s) of the Output	51
8.5.1 Description.....	52
8.5.2 Bivariate Analysis of the variables (categorical vs continuous).....	53
8.5.3 Bivariate Analysis of the variable – (GENDER vs LOAN_DURATION); (Categorical vs continuous variable)	53
8.5.4 SAS Source Codes.....	53
8.5.5 Screenshot(s) of the Output	53
8.5.6 Description.....	54
8.5.7 Bivariate Analysis of the variable – (LOAN_HISTORY vs CANDIDATE_INCOME); (Categorical vs continuous variable)	54
8.5.8 SAS Source Codes.....	55
8.5.9 Screenshot(s) of the Output	55
8.6 Description.....	55
8.6.1 Bivariate Analysis of the variable – (QUALIFICATION vs LOAN_AMOUNT); (Categorical vs continuous variable)	56
8.6.2 SAS Source Codes.....	56
8.6.2 Screenshot(s) of the Output	56
8.6.3 Description.....	57
8.6.4 Analysis of the Categorical variables found in DAP67696.TESTING_DS.....	57
8.6.5 Univariate Analysis of the Categorical variable using SAS MACRO	57
8.6.6 Introduction	57
8.6.7 SAS Source Codes.....	58
8.6.8 Screenshot(s) of the Output	59
8.6.9 Description.....	59

8.7 Univariate Analysis of the Continuous/Numeric variable – using the SAS MACRO	60
8.7.1 SAS Source Codes.....	60
8.7.2 Screenshot(s) of the Output	62
8.7.3 Description.....	65
8.7.4 Bivariate Analysis of the variables – (Categorical vs categorical variable) using SAS Macro.....	66
8.7.5 SAS Source Codes.....	66
8.7.6 Screenshot(s) of the Output	67
8.7.7 Description.....	70
8.7.8 Bivariate Analysis of the variables – (Categorical vs continuous variable) using SAS Macro.....	71
8.7.9 SAS Source Codes.....	71
8.8 Screenshot(s) of the Output	72
8.8.1 Description.....	73
8.8.2 Data Cleaning	74
8.8.3 Imputing the missing values found in the categorical variables in the datasets DAP67696.TRAINING_DS.....	74
8.8.4 Imputing the missing values found in the categorical variables - GENDER.	74
8.8.5 SAS Source Codes.....	74
8.8.6 Screenshot(s) of the Output	75
8.8.7 Description.....	75
8.8.8 SAS Source Codes.....	75
8.8.9 Screenshot(s) of the Output	76
8.9 Description.....	76
8.9.1 SAS Source Codes.....	76
8.9.2 Screenshot(s) of the Output	77
8.9.3 Description.....	77
8.9.4 SAS Source Codes.....	77
8.9.5 Screenshot of the Output	78
8.9.6 Description.....	78
8.9.7 SAS Source Codes.....	78
8.9.8 Screenshot(s) of the Output	79

8.9.9 Description.....	79
8.0.1.1 SAS Source Codes.....	80
8.0.1.2 Screenshot(s) of the Output	80
8.0.1.3 Description.....	80
8.0.1.4 SAS Source Codes.....	80
8.0.1.5 Screenshot(s) of the Output	81
8.0.1.6 Description.....	81
8.0.1.7 Imputing the missing values found in the categorical variables – MARITAL_STATUS.....	81
8.0.1.8 SAS Source Codes.....	81
8.0.1.9 Screenshot(s) of the Output	82
8.0.2 Description.....	82
8.0.2.1 SAS Source Codes.....	82
8.0.2.2 Screenshot(s) of the Output	83
8.0.2.3 Description.....	83
8.0.2.4 SAS Source Codes.....	83
8.0.2.5 Screenshot(s) of the Output	84
8.0.2.6 Description.....	84
8.0.2.7 SAS Source Codes.....	84
8.0.2.8 Screenshot(s) of the Output	84
8.0.2.9 Description.....	84
8.0.3 Imputing the missing values found in the categorical variables – FAMILY_MEMBERS.	85
8.0.3.1 SAS Source Codes.....	85
8.0.3.2 Screenshot(s) of the Output	85
8.0.3.3 Description.....	85
8.0.3.4 SAS Source Codes.....	86
8.0.3.5 Screenshot(s) of the Output	86
8.0.3.6 Description.....	86
8.0.3.7 SAS Source Codes.....	87
8.0.3.8 Screenshot(s) of the Output	87
8.0.3.9 Description.....	88

8.0.4 SAS Source Codes.....	88
8.0.4.1 Screenshot(s) of the Output	88
8.0.4.2 Description.....	88
8.0.4.3 SAS Source Codes.....	89
8.0.4.4 Screenshot(s) of the Output	89
8.0.4.5 Description.....	89
8.0.4.6 SAS Source Codes.....	90
8.0.4.7 Screenshot(s) of the Output	90
8.0.4.8 Description.....	90
8.0.4.9 SAS Source Codes.....	90
8.0.5 Screenshot(s) of the Output	91
8.0.5.1 Description.....	91
8.0.5.2 Imputing the missing values found in the continuous variables – LOAN_AMOUNT.	
8.0.5.3 SAS Source Codes.....	92
8.0.5.4 Screenshot(s) of the Output	92
8.0.5.5 Description.....	92
8.0.5.6 SAS Source Codes.....	93
8.0.5.7 Screenshot(s) of the Output	93
8.0.5.8 Description.....	93
8.0.5.9 SAS Source Codes.....	94
8.0.6 Screenshot(s) of the Output	94
8.0.7 Description.....	94
8.0.7.1 SAS Source Codes.....	95
8.0.7.2 Screenshot(s) of the Output	95
8.0.7.3 Description.....	95
8.0.7.4 Imputing the missing values found in the continuous variables – LOAN_DURATION.	95
8.0.7.5 SAS Source Codes.....	96
8.0.7.6 Screenshot(s) of the Output	96
8.0.7.7 Description.....	96
8.0.7.8 SAS Source Codes.....	97
8.0.7.9 Screenshot(s) of the Output	97

8.0.8 Description.....	97
8.0.8.1 SAS Source Codes.....	98
8.0.8.2 Screenshot(s) of the Output	98
8.0.8.3 Description.....	98
8.0.8.4 SAS Source Codes.....	99
8.0.8.5 Screenshot(s) of the Output	99
8.0.8.6 Description.....	99
8.0.8.7 Imputing the missing values found in the continuous variables – GUARANTEE_INCOME	99
8.0.8.8 SAS Source Codes.....	100
8.0.8.9 Screenshot(s) of the Output	100
8.0.9.1 Description.....	100
8.0.9.2 SAS Source Codes.....	101
8.0.9.3 Screenshot(s) of the Output	101
8.0.9.4 Description.....	101
8.0.9.5 SAS Source Codes.....	102
8.0.9.6 Screenshot(s) of the Output	102
8.0.9.7 Description.....	102
8.0.9.8 SAS Source Codes.....	103
8.0.9.9 Screenshot(s) of the Output	103
8.0.0.1 Description.....	103
9.0 Chapter 9 (Model Creation and Prediction).....	104
9.1.1 SAS Source Codes.....	104
9.1.2 Screenshot(s) of the Output	105
9.1.3 Description.....	105
9.1.4 Screenshot(s) of the Output	106
9.1.5 Description.....	106
9.1.6 Screenshot(s) of the Output	106
9.1.7 Description.....	106
9.1.8 Screenshot(s) of the Output	107
9.1.9 Description.....	107
9.2 Screenshot(s) of the Output	108

9.2.1 Description.....	108
9.2.2 SAS Source Codes\.....	108
9.2.3 Screenshot(s) of the Output	109
9.2.4 Description.....	109
9.2.5 List the details of the dataset carrying the loan approval status predicted- DAP67696.TESTING_PREDICTED_DS	109
9.2.6 SAS Source Codes.....	109
9.2.7 Screenshot(s) of the Output	110
9.2.8 Description.....	110
10.0 Chapter 10 (Data Visualization and Report Generation).....	111
10.1.1 Data visualization	111
10.1.2 Introduction	111
10.1.3 SAS Source Codes.....	111
10.1.4 Screenshot(s) of the Output	112
10.1.5 Description.....	112
10.1.6 SAS Source Codes.....	112
10.1.7 Screenshot(s) of the Output	113
10.1.8 Description.....	113
10.1.9 SAS Source Codes.....	114
10.2 Screenshot(s) of the Output	114
10.2.1 Description.....	114
10.2.2 SAS Source Codes.....	115
10.2.3 Screenshot(s) of the Output	115
10.2.4 Description.....	116
10.2.5 SAS Source Codes.....	116
10.2.6 Screenshot(s) of the Output	117
10.2.7 Description.....	117
10.2.8 SAS Source Codes.....	118
10.2.9 Screenshot(s) of the Output	118
10.3 Description.....	119
10.3.1 SAS Source Codes.....	120
10.3.2 Screenshot(s) of the Output	120

10.3.2 Description.....	121
10.3.3 SAS Source Codes.....	121
10.3.4 Screenshot(s) of the Output	121
10.3.5 Description.....	122
10.3.6 SAS Source Codes.....	122
10.3.7 Screenshot(s) of the Output	122
10.3.8 Description.....	123
10.3.9 Report Generation.....	123
10.4 Physical location of SAS library	123
10.4.1 SAS Source Codes.....	123
10.4.2 Screenshot(s) of the Output	123
10.4.2 Description.....	124
10.4.3 Introduction to ODS	125
10.4.4 SAS Source Codes.....	126
10.4.5 Screenshot.....	126
10.4.6 Description.....	127
10.4.7 SAS Source Codes.....	128
10.4.8 Screenshot(s) of the Output	128
10.4.9 Description.....	128
10.5 SAS Source Codes.....	129
10.5.1 Screenshot(s) of the Output	130
10.5.2 Description.....	134
10.5.3 Discussion.....	135
11.0 Chapter 11 (Conclusion).....	137
12.0 REFERENCES	138

Part 1

1.0 Chapter 1 (Introduction)

Banks have been the bedrock of modern economies in the modern world right now where it plays crucial role in financial infrastructure of many countries around the globe. Banks act as an intermediary between the depositors (institution who loan money to the bank) and borrowers (to whom the bank lends money). A banking system is a group or network of institutions that provide financial services and typically banking institutions operate a payment system, provide loans, take deposits, and help people, individuals, or companies with investments. There is a certain amount of money banks pay for deposits and interest is often the income the bank receives when giving out loans. So, there are two key people here, the depositors and the borrowers where depositors can be the common man such as individuals, financial and non-financial firms, and local governments.

For example, payments and loans from typical commercial banks allow people to use deposit funds and use checking and debit cards to pay bills or make any purchases. So, what is a bank? A bank like many others is a business but unlike any other business, bank don't manufacture a physical product like factory or industry but instead bank provide services. These services such as business loans which is related to this assignment, Lasiandra Finance Inc. (LFI), that provide loan to small startups. Other services bank provided are car loans, home mortgages loan, credit card services and retirement accounts.

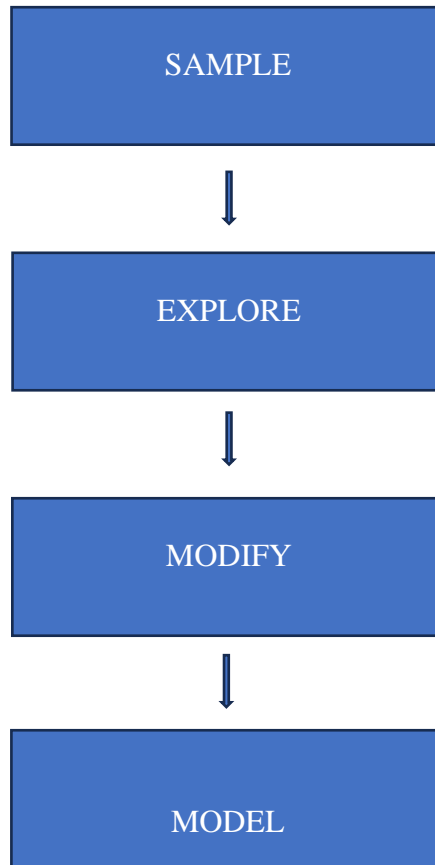
On the other hand, central banks such the federal reserve bank in USA and in Malaysia (Bank Negara Malaysia) main role is to distribute currency and establish money related policies. Another type and form of bank is investment bank where it conducts trades or deals with capital markets and most banks are profit-seeking entities where the major goal is to protect the interest of the shareholders. Bank make profit by charging more interest in loans and paying less interest on deposits. For example, a typical interest rate for a house mortgage with a 30-year loan would be around 3.29%. There are several key takeaways such as both banking systems operate differently and have different target clientele. Private banks focus on providing the best services and caters to

wealthy individuals, companies, and corporations where often the goal is to generate profit for shareholders.

The public banking system's main goal is to adhere to societal role, provide apprehensive and accessible financial services to the masses, contribute to economic growth and development as instructed by the government. So, in conclusion, the way banks and the banking industry work is that it manages and operates the flow or exchange of money between business and people where it offers deposit account where people can store their money in a secured place. This money is used by the banks to provide loans to businesses. In return the bank makes a profit from the interest payments they receive on those loans from the borrowers.

2.0 Chapter 2 (Problem Statement)

As a data scientist at this company in the Headquarters of LFI, Washington, D.C., United, it is required by the company to analyse the datasets obtained from the past customers and build an accurate model or in other word model building to predict the approval process of loan applications either it is being approved or being rejected.



Workflow of model proposed.

Above shows how the workflow of the model proposed where further detail explanation will be done to explain the model output of the analytical study of the Lasiandra Finance Inc.

Sl.No	Variable	Description
1	SME_LOAN_ID_NO	Reference No for the Loan
2	GENDER	Male / Female
3	MARITAL STATUS	Married / Not Married
4	FAMILY_MEMBERS	Total no of Family Members
5	QUALIFICATION	Graduate / Under Graduate
6	EMPLOYMENT	Yes / No
7	CANDIDATE INCOME	Monthly Income of the Applicant
8	GUARANTEE INCOME	Joint Applicant Income
9	LOAN-AMOUNT	Amount applied for in thousands
10	LOAN-DURATION	Repayment duration for the loan
11	LOAN-HISTORY	Past loan records positive / negative
12	LOAN_LOCATION	City / Town / Village
13	LOAN_APPROVAL	Approval of Loan Yes / No

Figure 1 : Datasets of the bank loan

Figure above shows the datasets used for the assignment and further details about the datasets is that it contained training and testing datasets in the csv format where it is imported into the SAS software.

3.0 Chapter 3 (Background Of Lasiandra Finance Inc. (LFI), New York, USA)

A leading private financing company called Lasiandra Finance Inc. (LFI) New York, USA which caters to small startup or small companies especially (SME) Small and Medium Enterprises that needed funding. This company clearly understands that some businesses need a dream of extra push or fund so that it can accelerate the company growth faster. This company requires to upgrade their loaning process or system to be tailor made and customer centric. Also, in the past years this company massively upgraded its wings to speed up and increase its process as it needs to automate loan eligibility process based on customer portfolio entered online.

Hence the main problem faced by the company is the process of approving the loans as the process of loan approval is very complicated. The procedure is very complex as it needed constant verification and validation such loan criteria and eligibility as there is no guarantee that the chosen or eligible applicant will be accepted. There are several key takeaways here such as companies like Lasiandra Finance Inc typically fall under private sector banking where it is a type of banking system that are generally held by private companies or very wealthy individuals. The table below shows the benefits of private banks and the counterpart if it is public bank although both of them served the served purpose to make money or profit.

Table 1

Private Bank	Public bank
Offers fast and quick services to its customers	Low interest charges on loans
Provide customized services according to customers preferences	High interest rate on deposits
Fast and quick financial decision making	Full security jobs for employees
Streamlined management system	Offers its services to large customer base

The way manual loan applications work is that it consists of 7 steps before the loan application is approved by the bank. The process consists of pre-qualification process, loan application,

application processing, underwriting process, credit decision, quality check and loan funding. Table below shows the detailed steps of each manual loan application process.

Table 2

Step	Details
pre-qualification process	In the original step of loan application process, this method works whereby the borrower needs to submit several lists of items to the lender to get loan.
Loan Application	At this stage the borrower completed the loan application and mostly nowadays it is done through web and mobile app.
Application Processing	The credit department then received the application where it reviewed for its completeness, accuracy and how genuine it is
Underwriting Process	At this stage there are several criteria that need to be checked by the lenders such as credit scores and risk scores
Credit decision	From the results of the underwriting process, the applicants will be notified by the banker whether the loan approval is approved or denied.
Quality Check	The lenders must ensure the quality of the manual loan application process since the rate of the lending is high and must be regulated most of the time.
Loan Funding	After the loan application process is completed when the documents are signed, it is hence loans fund shortly afterwards.

4.0 Chapter 4 (ASSUMPTION, PROGRAM DEMONSTRATION, CODING AND JUSTIFICATION).

For this assignment, the deliverables and fulfillment of the assignment is that it is required to conduct the data analysis on the datasets using SAS program and the report requires it to introduce the data, method/technique, and coding problems. One of ultimate goals is to discuss the objective of the analysis and from that the output of the code results are interpreted in datasets exploration.

Regards to the program demonstration and coding, there are a few steps that need to be taken such as ensuring that break down the complex procedure of the coding into the important steps. These essential steps include showing the ability to demonstrate subsets of PROC, SQL code, macro code or supplementary code. Hence, each step the code explanation must done the way it anticipated and identify the section where improvement needed to be done and repeat again. After that, discuss the obstacle when doing the SQL code programming and show the mistakes done. It is reasonable to show the programming problem and later show some modifications to make the project/assignment better. Often the times the modifications done when there are some changes occur on the goal and can be done by tweak the code and the output.

The end goal of the program is to have a complete code that demonstrate the datasets used which in this assignment the banking loan datasets, along with the product of code, tables and ODS output (Output Delivery System). Before that let dives deep into understand the PROC SQL, the relational database management system where relational database is an organized and structured collection of information where the data are arranged into a two-dimensional table. Each of the table contain usually one or more row and columns. Often the time the data are related, based upon their values and not according to other data structures.

On the other hand, for this data analytical assignment (DAP), SQL statements is used to read and update table. PROC SQL used structured query language that talks to relational database management system as SQL provide familiar action or task such as CREATE, SELECT, UPDATE, INSERT, DROP and DELETE, Below shows the task of SQL.

- Create and delete dataset.

- Retrieve and manipulate SAS dataset.
- Add or modify data values in a dataset.
- Create and delete indexes on columns in a dataset.
- Add, modify, or drop columns in a dataset.

It is known that PROC SQL can be used on SAS files, databases tables and combinations of these to execute query operations. The reason why SAS are chosen is that SAS as a programming language is very user friendly and it is easy to use compared to other programming language. This is due to the SAS used very simple syntax that uses abbreviated and direct commands and hence this makes it is an excellent language for people that has very little or no knowledge for programming. SAS programming language provides a simple user interface whereby this software includes charts, graph and plots and hence makes it easier to plot bars, graphs, and charts.

Moreover, benefits of using SAS are that it enhances data security as it provides high data security to many businesses and enterprises as SAS is one of the major key players of analytics tool used in most companies, business, and corporations. Hence due to this high security measures guaranteed by SAS, data manipulation is nearly impossible due data security provided by SAS.

5.0 Chapter 5 (Methodology)

The methodology used in this assignment is SEMMA method whereby it stands for "Sample, Explore, Modify, Model, and Assess.", whereby this method is a data mining and predictive analytics method, or technique developed and created by SAS institute. SAS is a software company widely known for data management tools and analytics solutions and below shows the detailed step of each of the SEMMA methods. Figure below shows the SEMMA method.

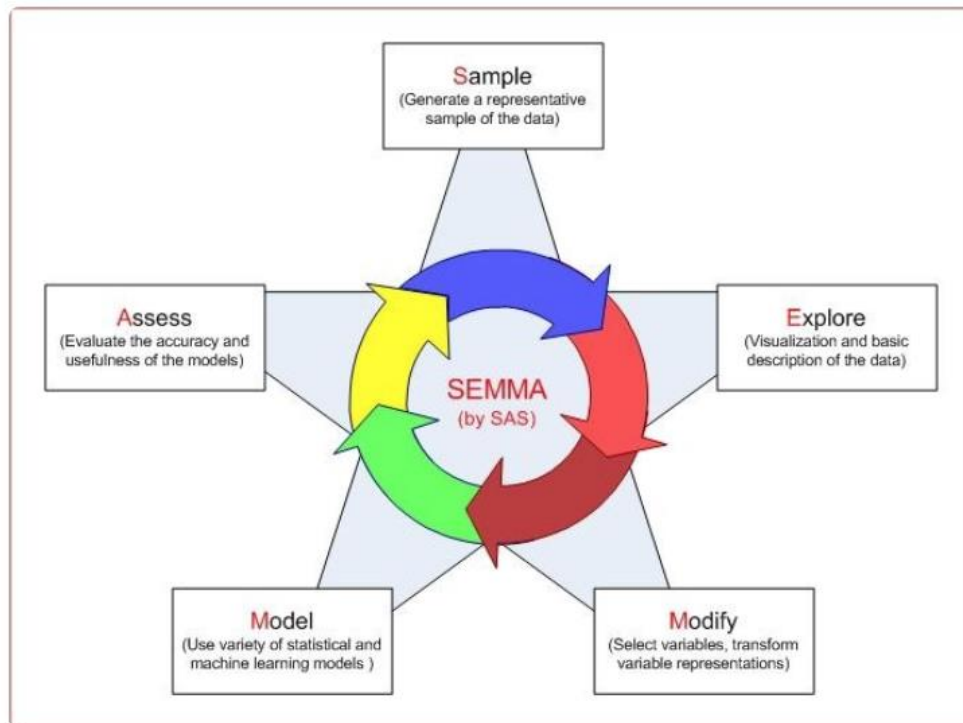


Figure 1

The first one is Sample whereby representative sample of the datasets is selected which in this case the TRAINING AND TESTING datasets. The sample datasets are very crucial to analyze a subset of data instead of the whole datasets. Next is Explore where data exploration is done to gain insights on the datasets chosen as for this part it is done to studied is there any observations, and anomalies between the variables. Modify is a method where the data undergoes pre-processing and it undergoes imputation, data transformation, cleaning data and handling missing values. Next is the Model where after the data preparation from Modify, this step participating in building a

predictive model where the data is prepared to train the machine learning model selected. Hence the proposed machine learning model used is for predictions or classification based on new data.

The last one is Assess where after model building was done the performance of the model is assessed using evaluation matrices such as performance matrices to determine their accuracy at making predictions. This method is also crucial to study the effectiveness of the proposed model and for these datasets some of the categorical variables are education level, gender, loan amount, loan approval, and loan type.

6.0 Chapter 6 (Data Dictionary/ Metadata)

6.1 Introduction

Exploration on the datasets is done where the datasets used contain train and test datasets which are imported into the SAS studio software. EDA stands for "Exploratory Data Analysis" where it is an approach for data analysis which solely focuses on studying the characteristics of the datasets using statistical measures. Some of key features of EDA include the data summarization, data visualization, data cleaning, pattern recognition, hypothesis study, feature selection and data transformation. Figure below shows the loan datasets used in this assignment.

Sl.No	Variable	Description
1	SME_LOAN_ID_NO	Reference No for the Loan
2	GENDER	Male / Female
3	MARITAL STATUS	Married / Not Married
4	FAMILY MEMBERS	Total no of Family Members
5	QUALIFICATION	Graduate / Under Graduate
6	EMPLOYMENT	Yes / No
7	CANDIDATE INCOME	Monthly Income of the Applicant
8	GUARANTEE INCOME	Joint Applicant Income
9	LOAN-AMOUNT	Amount applied for in thousands
10	LOAN-DURATION	Repayment duration for the loan
11	LOAN-HISTORY	Past loan records positive / negative
12	LOAN_LOCATION	City / Town / Village
13	LOAN_APPROVAL	Approval of Loan Yes / No

Figure 2

Data exploration is very crucial to find insights on the datasets by understanding the data structure or in this case metadata which means by data that describes other data.

6.1 Location of the datasets on SAS



Figure 3

6.2 Project datasets found inside the SAS permanent library/folder.

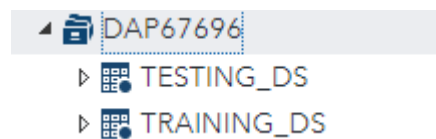


Figure 4

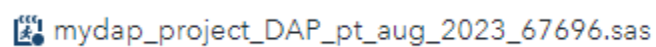


Figure 5

6.2.1 Description

The purpose of SAS library is simply a collection of SAS files that are stored in the same folder or directory of the computer. Figure above shows the datasets stored inside SAS permanent library whereby a permanent SAS library is stored inside external storage and will be not deleted when SAS session terminated. The maximum length of a library is typically 32 characters, and the length is maximum 8 for numeric letters. The permanent library name for this assignment is DAP67696 where it stored the TESTING_DS and TRAINING_DS files.

6.2.2 Display the structure of the datasets-DAP67696.TRAINING_DS

6.2.3 Data dictionary of the dataset- DAP67696.TRAINING_DS

```
1  /*****  
2  Developer name : Muhammad Arif Bin Jamaluddin  
3  Job Position: Data Scientist, Dazztech Solutions Sdn Bhd  
4  Program name: mydap_project_DAP_pt_aug_2023_67696.sas  
5  Description: Loan application status prediction - 1-2  
6  Date first written: Wed, 23-Sept-2023  
7  Date last updated: Mon, 16-Oct-2023  
8  Folder name: DAP_PT_AUG_2023_TP067696  
9  Library name: DAP67696
```

Figure 6

6.2.4 SAS PROC SQL Codes

```
.....  
PROC SQL;  
  
DESCRIBE TABLE DAP67696.TRAINING_DS;  
  
QUIT;
```

Figure 7

6.2.5 Screenshot(s) of the output


```

1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69          PROC SQL;
70
71          DESCRIBE TABLE DAP67696.TRAINING_DS;
NOTE: SQL table DAP67696.TRAINING_DS was created like:

create table DAP67696.TRAINING_DS( bufsize=131072 )
(
  SME_LOAN_ID_NO char(8) format=$8. informat=$8.,
  GENDER char(6) format=$6. informat=$6.,
  MARITAL_STATUS char(11) format=$11. informat=$11.,
  FAMILY_MEMBERS char(2) format=$2. informat=$2.,
  QUALIFICATION char(14) format=$14. informat=$14.,
  EMPLOYMENT char(3) format=$3. informat=$3.,
  CANDIDATE_INCOME num format=BEST12. informat=BEST32.,
  GUARANTEE_INCOME num format=BEST12. informat=BEST32.,
  LOAN_AMOUNT num format=BEST12. informat=BEST32.,
  LOAN_DURATION num format=BEST12. informat=BEST32.,
  LOAN_HISTORY num format=BEST12. informat=BEST32.,
  LOAN_LOCATION char(7) format=$7. informat=$7.,
  LOAN_APPROVAL_STATUS char(1) format=$1. informat=$1.
);

72          RUN;
NOTE: PROC SQL statements are executed immediately; The RUN statement has no effect
73
74          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84

```

Figure 8

6.2.6 Description

The diagram above shows the structure of the TRAINING_DS dataset where CREATE TABLE command or syntax is used to create a table shown in diagram above. The table above shows the variable name along with data type and length in the new table created and for this project the library created is DAP67696 and from the output of the SAS log window the table is created successfully.

6.2.7 SAS PROC SQL Codes

```

PROC CONTENTS DATA = DAP67696.TRAINING_DS;

RUN;

```

Figure 9

6.2.8 Screenshot(s) of the output

The CONTENTS Procedure			
Data Set Name	DAP67696.TRAINING_DS	Observations	614
Member Type	DATA	Variables	13
Engine	V9	Indexes	0
Created	23/08/2023 21:08:16	Observation Length	96
Last Modified	23/08/2023 21:08:16	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Figure 10

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1363
Obs in First Data Page	614
Number of Data Set Repairs	0
Filename	/home/u61522473/DAP_PT_AUG_2023_TP067696/training_ds.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	7118417964
Access Permission	rw-r--r--
Owner Name	u61522473
File Size	256KB
File Size (bytes)	262144

Figure 11

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
7	CANDIDATE_INCOME	Num	8	BEST12.	BEST32.
6	EMPLOYMENT	Char	3	\$3.	\$3.
4	FAMILY_MEMBERS	Char	2	\$2.	\$2.
2	GENDER	Char	6	\$6.	\$6.
8	GUARANTEE_INCOME	Num	8	BEST12.	BEST32.
9	LOAN_AMOUNT	Num	8	BEST12.	BEST32.
13	LOAN_APPROVAL_STATUS	Char	1	\$1.	\$1.
10	LOAN_DURATION	Num	8	BEST12.	BEST32.
11	LOAN_HISTORY	Num	8	BEST12.	BEST32.
12	LOAN_LOCATION	Char	7	\$7.	\$7.
3	MARITAL_STATUS	Char	11	\$11.	\$11.
5	QUALIFICATION	Char	14	\$14.	\$14.
1	SME_LOAN_ID_NO	Char	8	\$8.	\$8.

Figure 12

6.2.9 Description

Based on the figure above the it shows the data structure of the DAP6796.TRAINING_DS whereby the number of variables is 13, the number of observations is 614 and the observation length is 96. The purpose of this method is to store data in the SAS system memory and the data can easily be monitored and retrieved by organizations to understand the characteristics of the datasets better.

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.

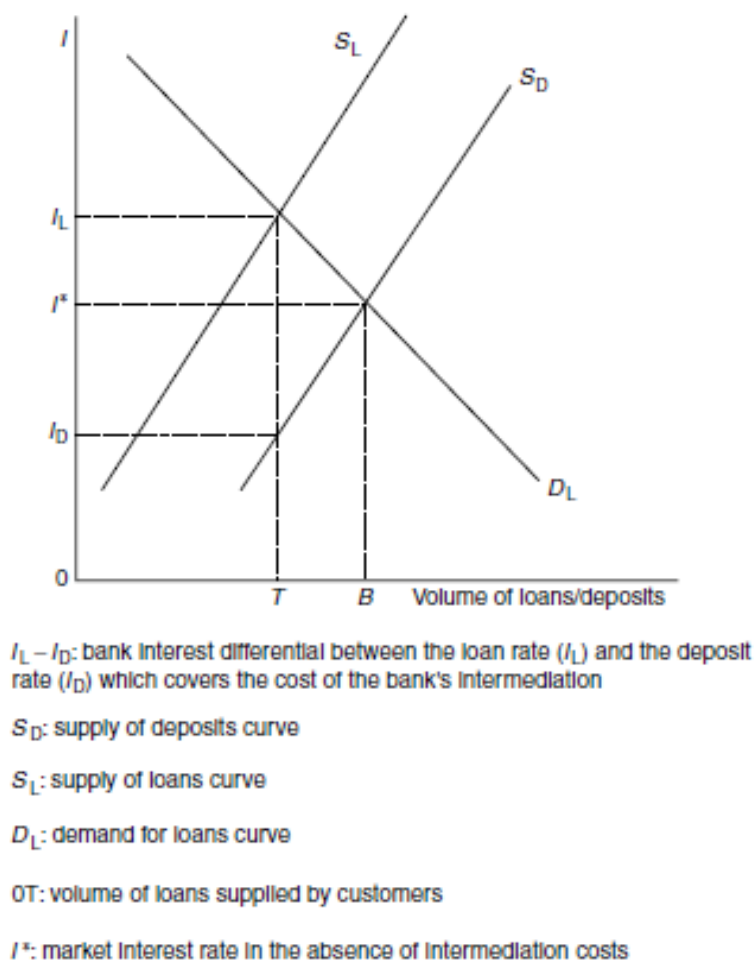


Figure 13

Based on the figure above the supply of deposits curve S_D has upwards slope and there is supply of loans curve S_L showing upwards slope where the bank will increase loan value as interest rates increasing. The demand for loans curve D_L shows that it decreasing if the interest rate increasing. It shows that for the i^* 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 OB 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 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 of 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 grade 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).

7.3 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. Hence 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.



Figure 14

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. Hence 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

```
24 TITLE 'Figure no - Univariate Analysis of the Categorical variable: Gender';  
25  
26 PROC FREQ DATA = DAP67696.TRAINING_DS;  
27  
28 TABLE GENDER;  
29  
30 RUN;
```

Figure 15

8.1.4 Screenshot(s) of the Output

Figure no - Univariate Analysis of the Categorical variable: Gender				
The FREQ Procedure				
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	112	18.64	112	18.64
Male	489	81.36	601	100.00
Frequency Missing = 13				

Figure 16

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 female, 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

```
32 TITLE 'Figure no - Univariate Analysis of the Categorical variable: Marital Status';  
33  
34 PROC FREQ DATA = DAP67696.TRAINING_DS;  
35  
36 TABLE MARITAL_STATUS;  
37  
38 RUN;
```

Figure 17

8.1.8 Screenshot(s) of the Output

Figure no - Univariate Analysis of the Categorical variable: Marital Status				
The FREQ Procedure				
MARITAL_STATUS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Married	398	65.14	398	65.14
Not Married	213	34.86	611	100.00
Frequency Missing = 3				

Figure 18

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

```
72 TITLE 'Figure no - Univariate Analysis of the Categorical variable: Loan location';  
73  
74 PROC FREQ DATA = DAP67696.TRAINING_DS;  
75  
76 TABLE LOAN_LOCATION;  
77  
78 RUN;
```

Figure 19

8.2.2 Screenshot(s) of the Output

Figure no - Univariate Analysis of the Categorical variable: Loan location

The FREQ Procedure

LOAN_LOCATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
City	202	32.90	202	32.90
Town	233	37.95	435	70.85
Village	179	29.15	614	100.00

Figure 20

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

```

87
88 TITLE 'Figure number - Univariate analysis of the continuous/numeric variable: GUARANTEE_INCOME';
89
90 PROC MEANS DATA = DAP67696.TRAINING_DS N NMISS MIN MAX MEAN MEDIAN STD;
91
92 VAR GUARANTEE_INCOME;
93
94 RUN;
95
96 ODS GRAPHICS / RESET WIDTH = 4.0 IN HEIGHT = 3.0 IN IMAGEMAP;
97
98 PROC SGPLOT DATA = DAP67696.TRAINING_DS;
99
100 HISTOGRAM GUARANTEE_INCOME;
101
102 TITLE 'Figure number - Univariate analysis of the continuous/numeric variable: GUARANTEE_INCOME';
103
104 RUN;
105
106
107

```

Figure 21

8.2.7 Screenshot(s) of the Output

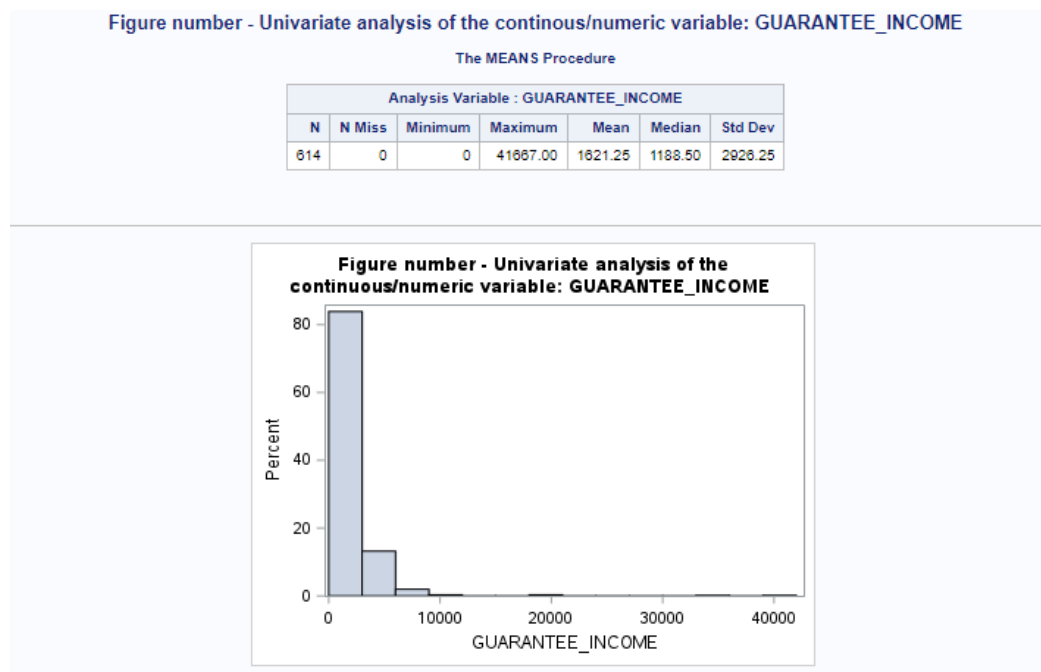


Figure 22

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

```
108 |
109 |
110 | TITLE 'Figure number - Univariate analysis of the continuous/numeric variable: LOAN_DURATION';
111 |
112 | PROC MEANS DATA = DAP67696.TRAINING_DS N NMISS MIN MAX MEAN MEDIAN STD;
113 |
114 | VAR LOAN_DURATION;
115 |
116 | RUN;
117 |
118 | ODS GRAPHICS / RESET WIDTH = 4.0 IN HEIGHT = 3.0 IN IMAGEMAP;
119 |
120 | PROC SGPLOT DATA = DAP67696.TRAINING_DS;
121 |
122 | HISTOGRAM LOAN_DURATION;
123 |
124 | TITLE 'Figure number - Univariate analysis of the continuous/numeric variable: LOAN_DURATION';
125 |
126 | RUN;
127 |
128 |
```

Figure 23

8.3.1 Screenshot(s) of the Output

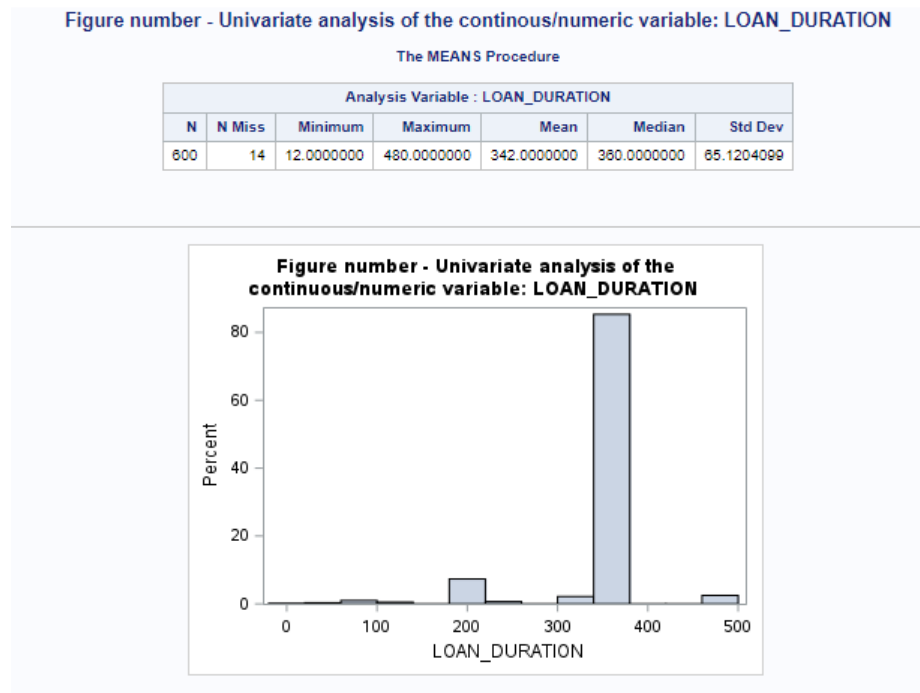


Figure 24

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

```
129  
130 TITLE 'Figure number - Univariate analysis of the continuous/numeric variable: LOAN_AMOUNT';  
131  
132 PROC MEANS DATA = DAP67696.TRAINING_DS N NMISS MIN MAX MEAN MEDIAN STD;  
133  
134 VAR LOAN_AMOUNT;  
135  
136 RUN;  
137  
138 ODS GRAPHICS / RESET WIDTH = 4.0 IN HEIGHT = 3.0 IN IMAGEMAP;  
139  
140 PROC SGPLOT DATA = DAP67696.TRAINING_DS;  
141  
142 HISTOGRAM LOAN_AMOUNT;  
143  
144 TITLE 'Figure number - Univariate analysis of the continuous/numeric variable: LOAN_AMOUNT';  
145  
146 RUN;
```

Figure 25

8.3.5 Screenshot(s) of the Output

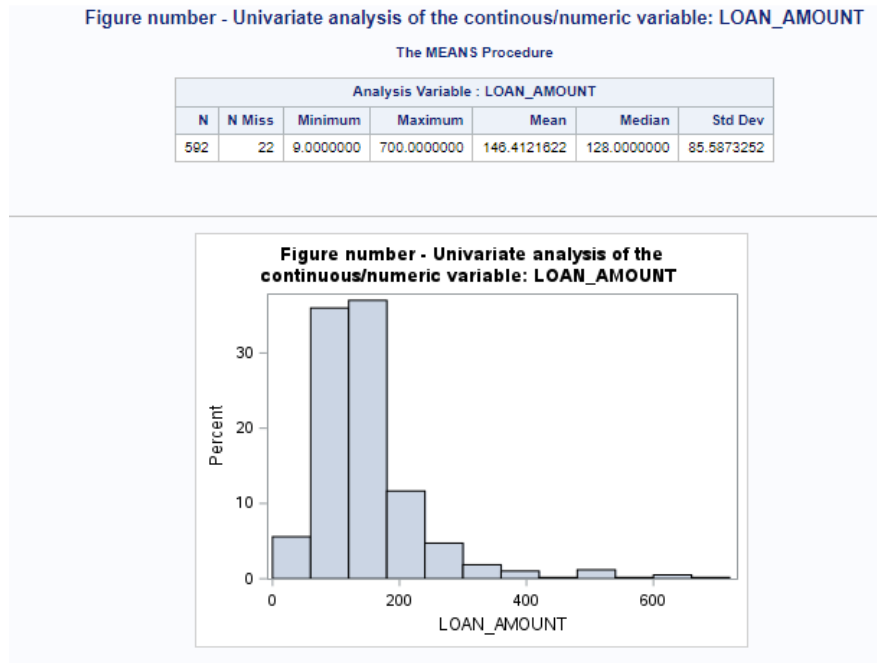


Figure 26

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 find 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

```

149 TITLE1 'Figure number Bivariate analysis of the variable: ';
150 TITLE2 'Categorical variable[GENDER] vs Categorical variable[MARITAL_STATUS]';
151 FOOTNOTE '-----END-----';
152
153 PROC FREQ DATA = DAP67696.TRAINING_DS;
154
155 TABLE gender * marital_status/
156 PLOTS = FREQPLOT( TWOWAY = STACKED SCALE = GROUPPCT );
157
158 RUN;
159

```

Figure 27

8.4.2 Screenshot(s) of the Output

**Figure number Bivariate analysis of the variable:
Categorical variable[GENDER] vs Categorical variable[MARITAL_STATUS]**

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of GENDER by MARITAL_STATUS			
	MARITAL_STATUS			
	GENDER	Married	Not Married	Total
Female		31	80	111
		5.18	13.38	18.56
		27.93	72.07	
		7.99	38.10	
Male		357	130	487
		59.70	21.74	81.44
		73.31	26.69	
		92.01	61.90	
Total		388	210	598
		64.88	35.12	100.00
Frequency Missing = 16				

Figure 28

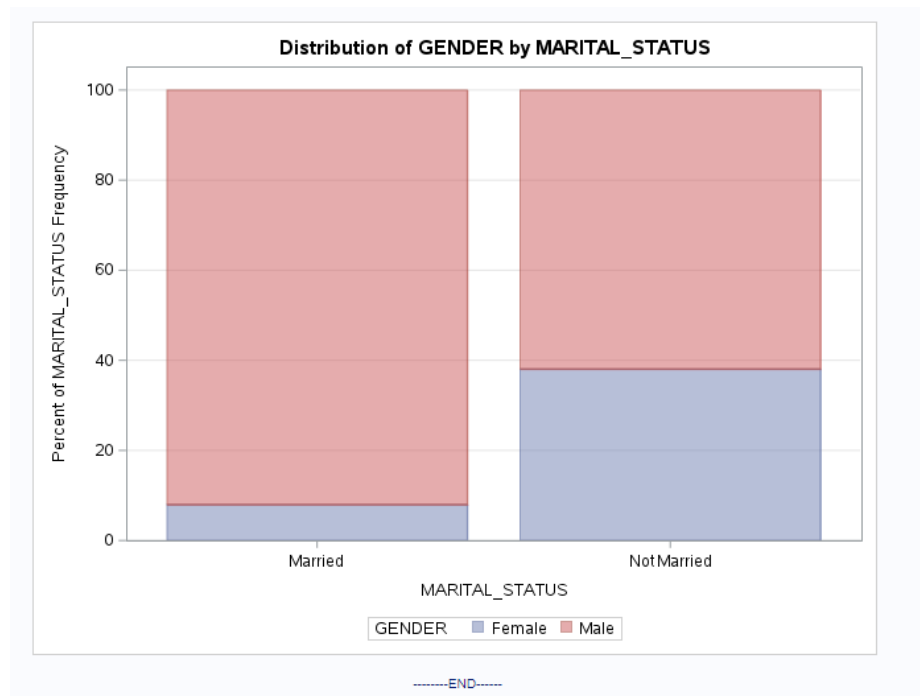


Figure 29

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

```

160
161 TITLE1 'Figure number Bivariate analysis of the variable: ';
162 TITLE2 'Categorical variable[GENDER] vs Categorical variable[EMPLOYMENT]';
163 FOOTNOTE '-----END-----';
164
165 PROC FREQ DATA = DAP67696.TRAINING_DS;
166
167 TABLE gender * employment/
168 PLOTS = FREQPLOT( TWOWAY = STACKED SCALE = GROUPPCT );
169
170 RUN;
171

```

Figure 30

8.4.6 Screenshot(s) of the Output

**Figure number Bivariate analysis of the variable:
Categorical variable[GENDER] vs Categorical variable[EMPLOYMENT]**

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of GENDER by EMPLOYMENT		
	EMPLOYMENT		
	No	Yes	Total
Female	89 15.64 85.58 18.13	15 2.64 14.42 19.23	104 18.28
Male	402 70.65 86.45 81.87	63 11.07 13.55 80.77	465 81.72
Total	491 86.29	78 13.71	569 100.00
Frequency Missing = 45			

Figure 31



Figure 32

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 bank 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

```

172 TITLE1 'Figure number Bivariate analysis of the variable: ';
173 TITLE2 'Categorical variable[LOAN_LOCATION] vs Categorical variable[LOAN_HISTORY]';
174 FOOTNOTE '-----END-----';
175
176 PROC FREQ DATA = DAP67696.TRAINING_DS;
177
178 TABLE loan_location * loan_history/
179 PLOTS = FREQPLOT( TWOWAY = STACKED SCALE = GROUPPCT );
180
181 RUN;
182
183

```

Figure 33

8.5 Screenshot(s) of the Output

Figure number Bivariate analysis of the variable:
Categorical variable[LOAN_LOCATION] vs Categorical variable[LOAN_HISTORY]

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of LOAN_LOCATION by LOAN_HISTORY			
	LOAN_LOCATION	LOAN_HISTORY		
		0	1	Total
City	31	151	182	
	5.50	26.77	32.27	
	17.03	82.97		
	34.83	31.79		
Town	30	187	217	
	5.32	33.16	38.48	
	13.82	86.18		
	33.71	39.37		
Village	28	137	165	
	4.96	24.29	29.26	
	16.97	83.03		
	31.46	28.84		
Total	89	475	564	
	15.78	84.22	100.00	
Frequency Missing = 50				

Figure 34

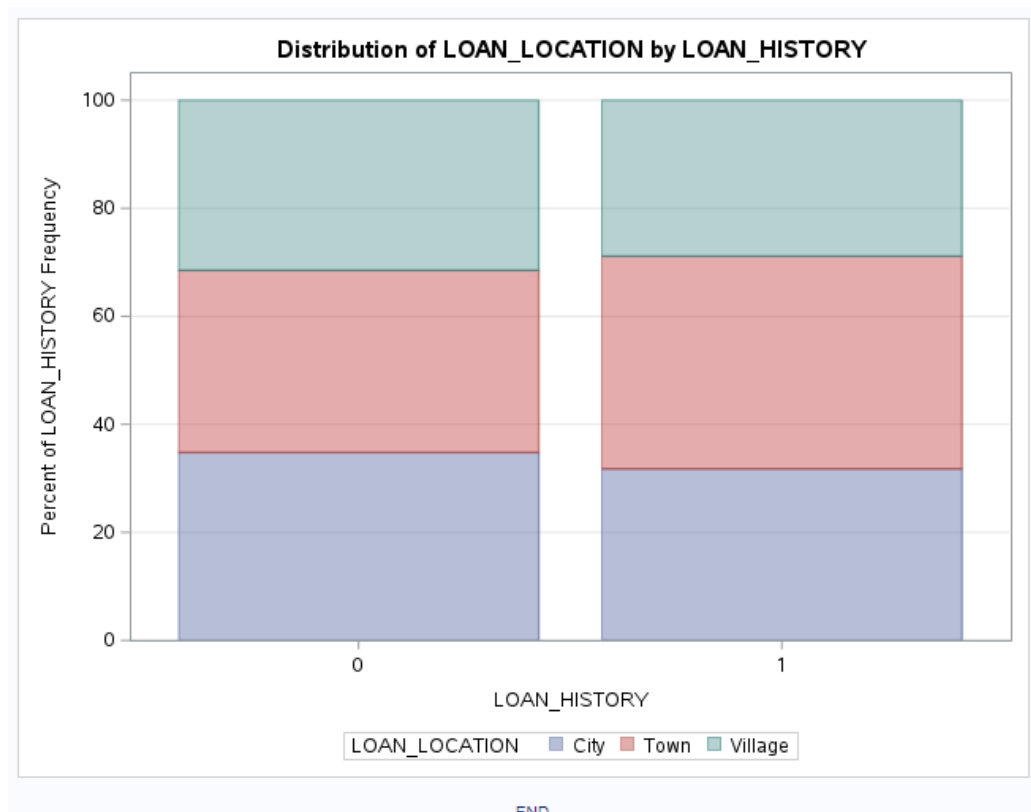


Figure 35

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 history of loans and these two labels in the context of data analysis refers 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 has 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 it 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

```

3
4 TITLE1 'Figure number: Bivariate analysis of variable: ';
5 TITLE2 'Categorical variable[GENDER] vs Numeric/Continous variable[LOAN_DURATION]';
6 FOOTNOTE '-----END-----';
7
8 PROC MEANS DATA = DAP67696.TRAINING_DS;
9
10 CLASS gender; /* It is a Categorical variable*/
11 VAR loan_duration; /* It is a Numeric/Continous variable */
12
13 RUN;

```

Figure 36

8.5.5 Screenshot(s) of the Output

Figure number: Bivariate analysis of variable:
Categorical variable[GENDER] vs Numeric/Continuous variable[LOAN_DURATION]

The MEANS Procedure

Analysis Variable : LOAN_DURATION						
GENDER	N Obs	N	Mean	Std Dev	Minimum	Maximum
Female	112	109	352.2935780	56.7220812	36.0000000	480.0000000
Male	489	478	339.6652720	67.0891400	12.0000000	480.0000000

-----END-----

Figure 37

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

```
16 TITLE1 'Figure number: Bivariate analysis of variable: ';
17 TITLE2 'Categorical variable[LOAN_HISTORY] vs Numeric/Continous variable[CANDIDATE_INCOME]';
18 FOOTNOTE '-----END-----';
19
0 PROC MEANS DATA = DAP67696.TRAINING_DS;
1
2 CLASS loan_history; /* It is a Categorical variable*/
3 VAR candidate_income; /* It is a Numeric/Continous variable */
4
```

Figure 38

8.5.9 Screenshot(s) of the Output

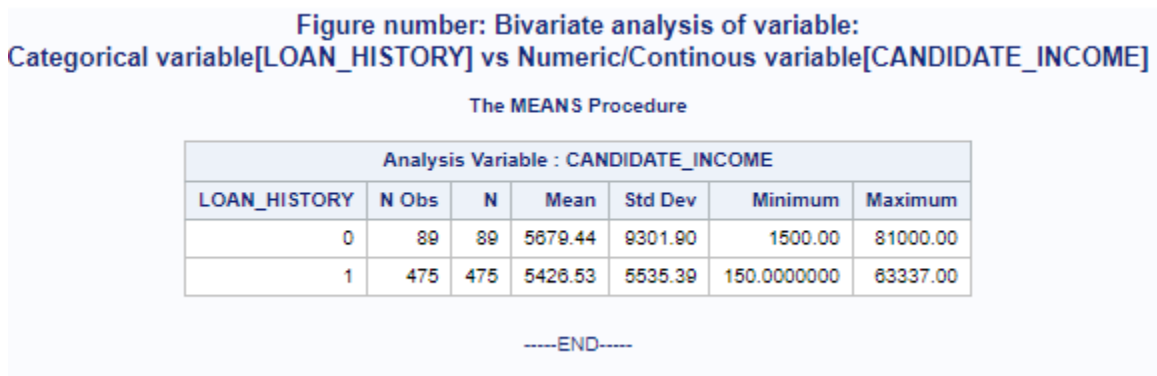


Figure 39

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

```
TITLE1 'Figure number: Bivariate analysis of variable: ';
TITLE2 'Categorical variable[QUALIFICATION] vs Numeric/Continuous variable[LOAN_AMOUNT]';
FOOTNOTE '-----END-----';

PROC MEANS DATA = DAP67696.TRAINING_DS;

CLASS qualification; /* It is a Categorical variable*/
VAR loan_amount; /* It is a Numeric/Continuous variable */

RUN;
```

Figure 40

8.6.2 Screenshot(s) of the Output

**Figure number: Bivariate analysis of variable:
Categorical variable[QUALIFICATION] vs Numeric/Continuous variable[LOAN_AMOUNT]**

The MEANS Procedure

Analysis Variable : LOAN_AMOUNT						
QUALIFICATION	N Obs	N	Mean	Std Dev	Minimum	Maximum
Graduate	480	465	154.0602151	92.8833658	9.0000000	700.0000000
Under Graduate	134	127	118.4094488	39.7736361	25.0000000	279.0000000

-----END-----

Figure 41

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

```
1  /* The SAS Macro begins */
2  OPTIONS MCOMPILENOTE=ALL;
3  %MACRO MACRO_UVACATEVARI(ptitle_name,pds_name,pcate_vari_name);
4  TITLE &ptitle_name;
5
6  PROC FREQ DATA = &pds_name;
7
8  TABLE &pcate_vari_name;
9
10 RUN;
11 %MEND MACRO_UVACATEVARI;
12
13 /* The SAS Macro ends */
```

Figure 42

```
79          %MEND MACRO_UVACATEVARI;
NOTE: The macro MACRO_UVACATEVARI completed compilation without errors.
      11 instructions 328 bytes.
80
81
82          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
92
```

Figure 43

```
4  /* To call the MACRO_UVACATEVARI for the categorical variable */
5
6  /* LOAN_LOCATION */
7  %MACRO_UVACATEVARI ('Univariate Analysis of the variable - LOAN_LOCATION', DAP67696.TESTING_DS, LOAN_LOCATION);
8
9  /* GENDER */
10 %MACRO_UVACATEVARI ('Univariate Analysis of the variable - GENDER', DAP67696.TESTING_DS, GENDER);
11
12 /* LOAN_LOCATION */
13 %MACRO_UVACATEVARI ('Univariate Analysis of the variable - MARITAL_STATUS', DAP67696.TESTING_DS, MARITAL_STATUS);
14
15 /* FAMILY_MEMBERS */
16 %MACRO_UVACATEVARI ('Univariate Analysis of the variable - FAMILY_MEMBERS', DAP67696.TESTING_DS, FAMILY_MEMBERS);
17
18 /* QUALIFICATION */
19 %MACRO_UVACATEVARI ('Univariate Analysis of the variable - QUALIFICATION', DAP67696.TESTING_DS, QUALIFICATION);
20
21 /* EMPLOYMENT */
22 %MACRO_UVACATEVARI ('Univariate Analysis of the variable - EMPLOYMENT', DAP67696.TESTING_DS, EMPLOYMENT);
23
24 /* LOAN_HISTORY */
25 %MACRO_UVACATEVARI ('Univariate Analysis of the variable - LOAN_HISTORY', DAP67696.TESTING_DS, LOAN_HISTORY);
26
```

Figure 44

8.6.8 Screenshot(s) of the Output

Univariate Analysis of the variable - LOAN_LOCATION

The FREQ Procedure

LOAN_LOCATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
City	140	38.15	140	38.15
Town	116	31.61	256	69.75
Village	111	30.25	367	100.00

Figure 45

Univariate Analysis of the variable - GENDER

The FREQ Procedure

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	70	19.66	70	19.66
Male	286	80.34	356	100.00

Frequency Missing = 11

Figure 46

Univariate Analysis of the variable - MARITAL_STATUS

The FREQ Procedure

MARITAL_STATUS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Married	233	63.49	233	63.49
Not Married	134	36.51	367	100.00

Figure 47

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 is 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

```
271
272 /*The SAS MACRO begin here*/
273
274 OPTIONS MCOMPILENOTE=ALL;
275 %MACRO MACRO_UVACONTI_VARI(ptitle,pds_name,pcontinue_vari_name);
276 TITLE &ptitle;
277
278 PROC MEANS DATA = &pds_name N NMISS MIN MAX MEAN MEDIAN STD;
279
280 VAR &pcontinue_vari_name;
281
282 RUN;
283
284 ODS GRAPHICS / RESET WIDTH = 4.0 IN HEIGHT = 3.0 IN IMAGEMAP;
285
286 PROC SGPLOT DATA = &pds_name;
287
288 HISTOGRAM &pcontinue_vari_name;
289
290 TITLE &ptitle;
291
292 RUN;
293 %MEND MACRO_UVACONTI_VARI;
294 /* The SAS MACRO ends here*/
295
```

Figure 48

```

90      RUN;
91      %MEND MACRO_UVACONTI_VARI;
NOTE: The macro MACRO_UVACONTI_VARI completed compilation without errors.
      13 instructions 544 bytes.
92      /* The SAS MACRO ends here*/
93
94      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
104

```

Figure 49

```

297 /* To call the SAS MACRO - MACRO_UVACONTI_VARI */
298
299 /* CANDIDATE_INCOME */
300 %MACRO_UVACONTI_VARI('Univariate Analysis of the continous variable - CANDIDATE_INCOME' , DAP67696.TESTING_DS, candidate_income);
301
302 /* GUARANTEE_INCOME */
303 %MACRO_UVACONTI_VARI('Univariate Analysis of the continous variable - GUARANTEE_INCOME' , DAP67696.TESTING_DS, guarantee_income);
304
305 /* LOAN_AMOUNT */
306 %MACRO_UVACONTI_VARI('Univariate Analysis of the continous variable - LOAN_AMOUNT' , DAP67696.TESTING_DS, loan_amount);
307
308 /* LOAN_DURATION */
309 %MACRO_UVACONTI_VARI('Univariate Analysis of the continous variable - LOAN_DURATION' , DAP67696.TESTING_DS, loan_duration);
310

```

Figure 50

8.7.2 Screenshot(s) of the Output

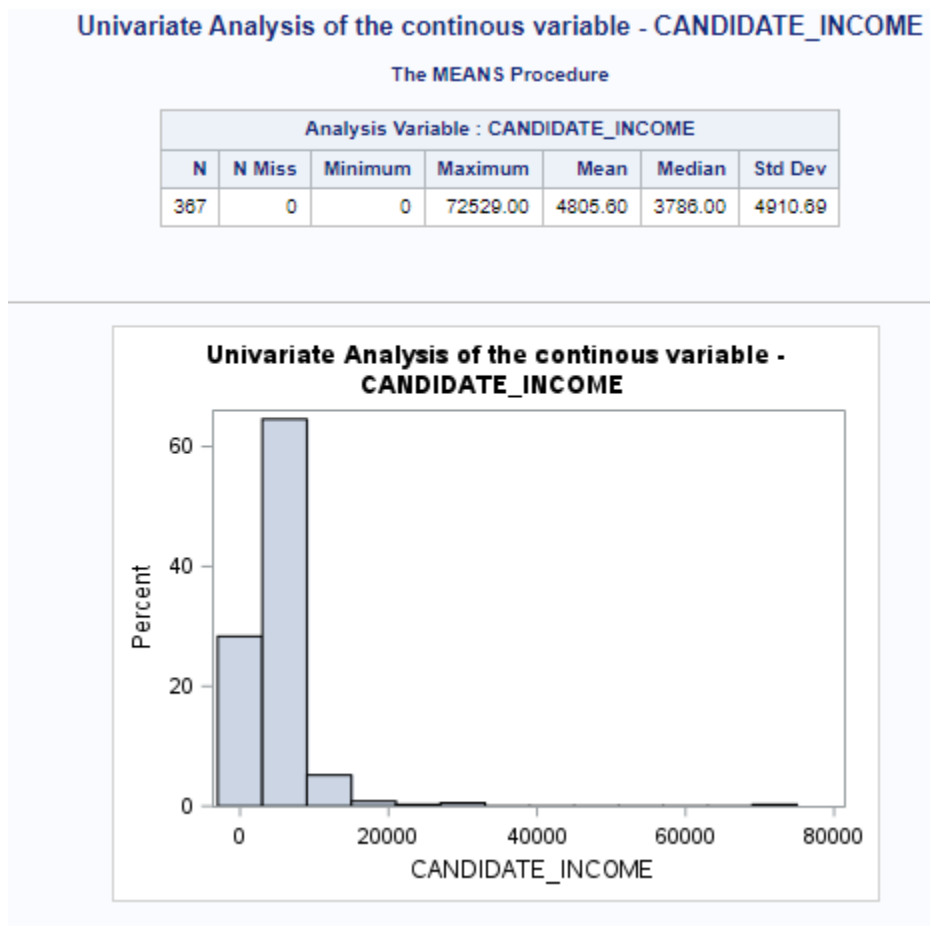


Figure 51

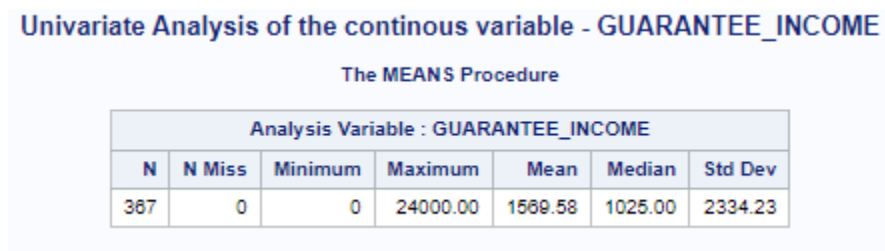


Figure 52

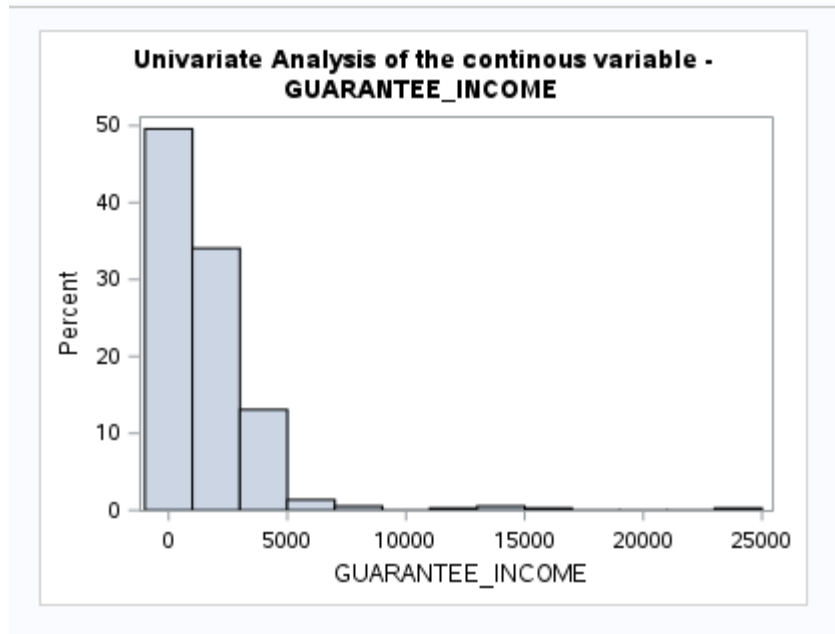


Figure 53

Univariate Analysis of the continuous variable - LOAN_AMOUNT

The MEANS Procedure

Analysis Variable : LOAN_AMOUNT						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
362	5	28.0000000	550.0000000	138.1325967	125.0000000	61.3666524

Figure 54

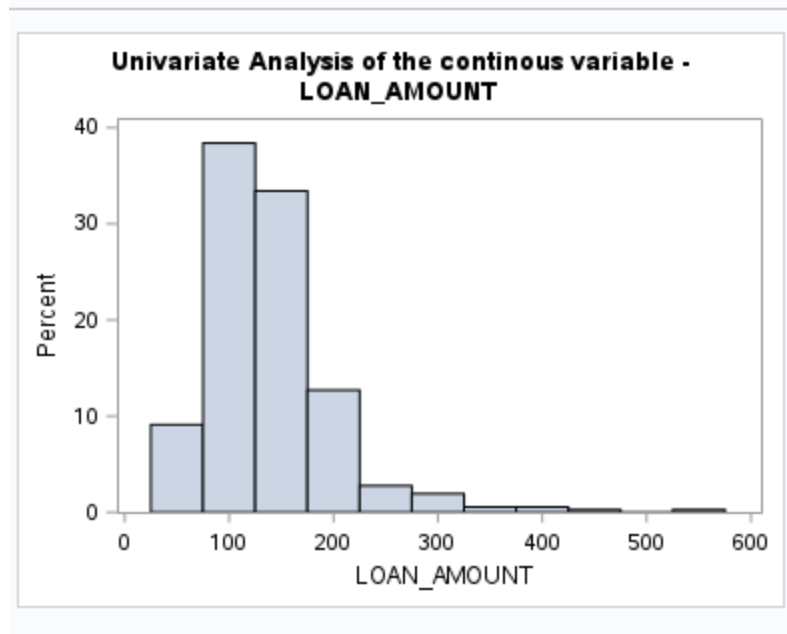


Figure 55

Univariate Analysis of the continous variable - LOAN_DURATION

The MEANS Procedure

Analysis Variable : LOAN_DURATION						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
361	6	6.0000000	480.0000000	342.5373961	360.0000000	65.1566434

Figure 56

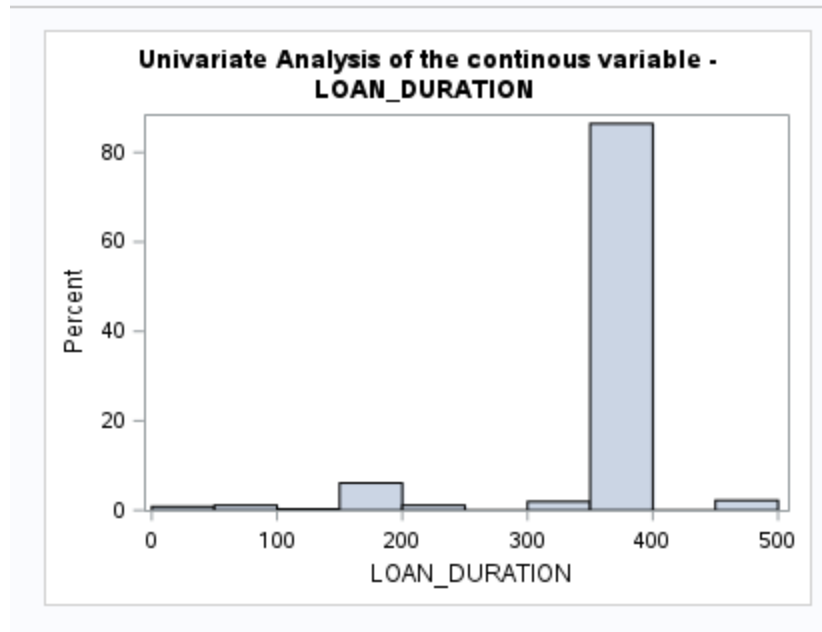


Figure 57

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

```
313 /* The SAS MACRO begins here */
314 OPTIONS MCOMPILENOTE=ALL;
315 %MACRO MACRO_BVA_CATE_CATE(ptitle1,ptitle2,pds_name,pcate_vari_name1,pcate_vari_name2);
316 TITLE1 &ptitle1;
317 TITLE2 &ptitle2;
318 FOOTNOTE '-----END-----';
319
320 PROC FREQ DATA = &pds_name;
321
322 TABLE &pcate_vari_name1 * &pcate_vari_name2/
323 PLOTS = FREQPLOT( TWOWAY = STACKED SCALE = GROUPPCT );
324
325 RUN;
326 %MEND MACRO_BVA_CATE_CATE;
327 /* The SAS MACRO ends here*/
328
329
```

Figure 58

```
81      non,
82      %MEND MACRO_BVA_CATE_CATE;
NOTE: The macro MACRO_BVA_CATE_CATE completed compilation without errors.
      15 instructions 576 bytes.
83      /* The SAS MACRO ends here*/
84
85      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
95
```

Figure 59

```

330 /* To call the SAS MACRO - MACRO_BVA_CATE_CATE */
331
332 /* Gender vs loan_location*/
333 %MACRO_BVA_CATE_CATE('Bivariate Analysis of the variables',
334 'Categorical variable vs Categorical variable',
335 DAP67696.TESTING_DS, GENDER, LOAN_LOCATION);
336
337 /* Gender vs loan_history*/
338 %MACRO_BVA_CATE_CATE('Bivariate Analysis of the variables',
339 'Categorical variable vs Categorical variable',
340 DAP67696.TESTING_DS, GENDER, LOAN_HISTORY);
341
342 /* Marital_status vs loan_location*/
343 %MACRO_BVA_CATE_CATE('Bivariate Analysis of the variables',
344 'Categorical variable vs Categorical variable',
345 DAP67696.TESTING_DS, MARITAL_STATUS, LOAN_LOCATION);
346

```

Figure 60

8.7.6 Screenshot(s) of the Output

Bivariate Analysis of the variables

Categorical variable vs Categorical variable

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of GENDER by LOAN_LOCATION				
	LOAN_LOCATION				
	GENDER	City	Town	Village	Total
	Female	25 7.02 35.71 18.25	27 7.58 38.57 24.32	18 5.06 25.71 16.67	70 19.66
	Male	112 31.46 39.16 81.75	84 23.60 29.37 75.68	90 25.28 31.47 83.33	286 80.34
	Total	137 38.48	111 31.18	108 30.34	356 100.00
Frequency Missing = 11					

Figure 61

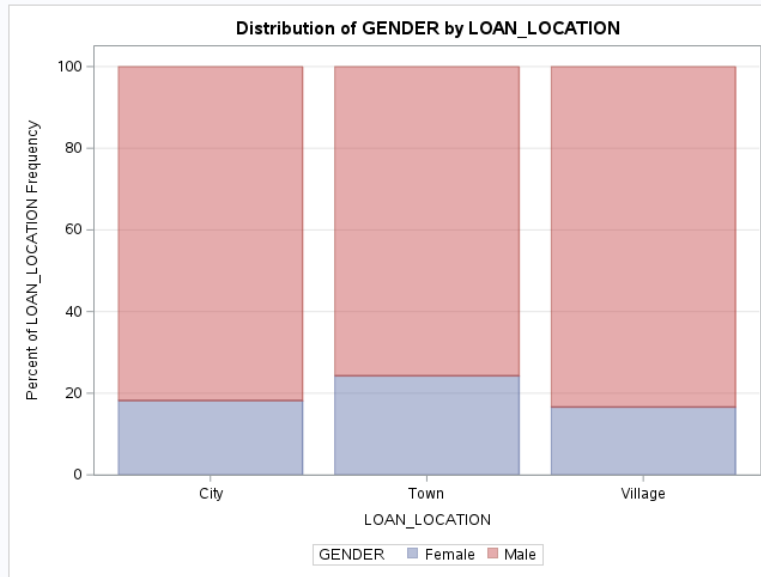


Figure 62

Bivariate Analysis of the variables
Categorical variable vs Categorical variable

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of GENDER by LOAN_HISTORY			
GENDER	LOAN_HISTORY		Total
	0	1	
Female	13	51	64
	3.98	15.60	19.57
	20.31	79.69	
	22.81	18.89	
Male	44	219	263
	13.46	66.97	80.43
	16.73	83.27	
	77.19	81.11	
Total	57	270	327
	17.43	82.57	100.00

Frequency Missing = 40

Figure 63

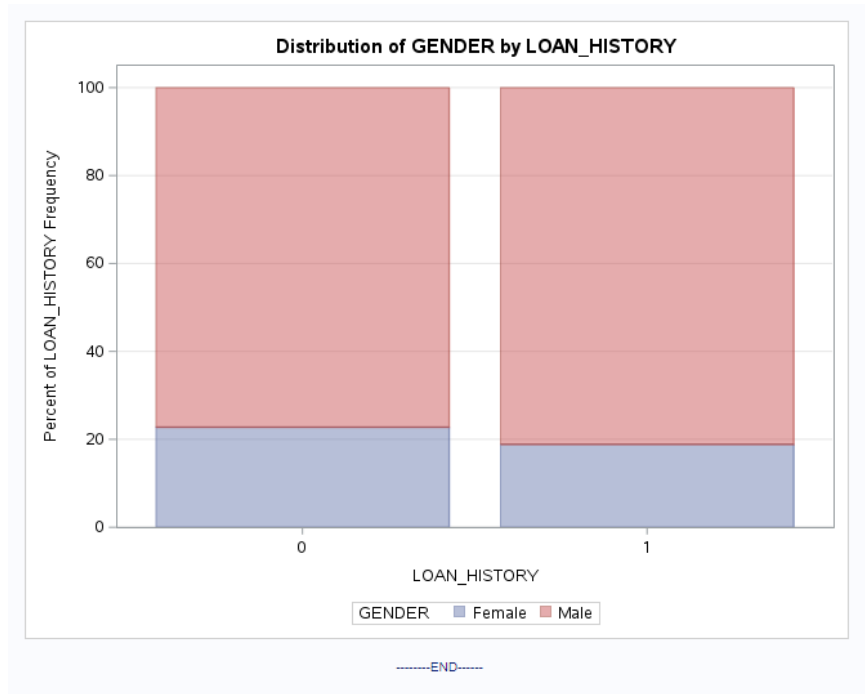


Figure 64

Bivariate Analysis of the variables
Categorical variable vs Categorical variable
The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of MARITAL_STATUS by LOAN_LOCATION			
	LOAN_LOCATION			
MARITAL_STATUS	City	Town	Village	Total
Married	91 24.80 39.06 65.00	71 19.35 30.47 61.21	71 19.35 30.47 63.96	233 63.49
Not Married	49 13.35 36.57 35.00	45 12.26 33.58 38.79	40 10.90 29.85 38.04	134 36.51
Total	140 38.15	116 31.61	111 30.25	367 100.00

Figure 65

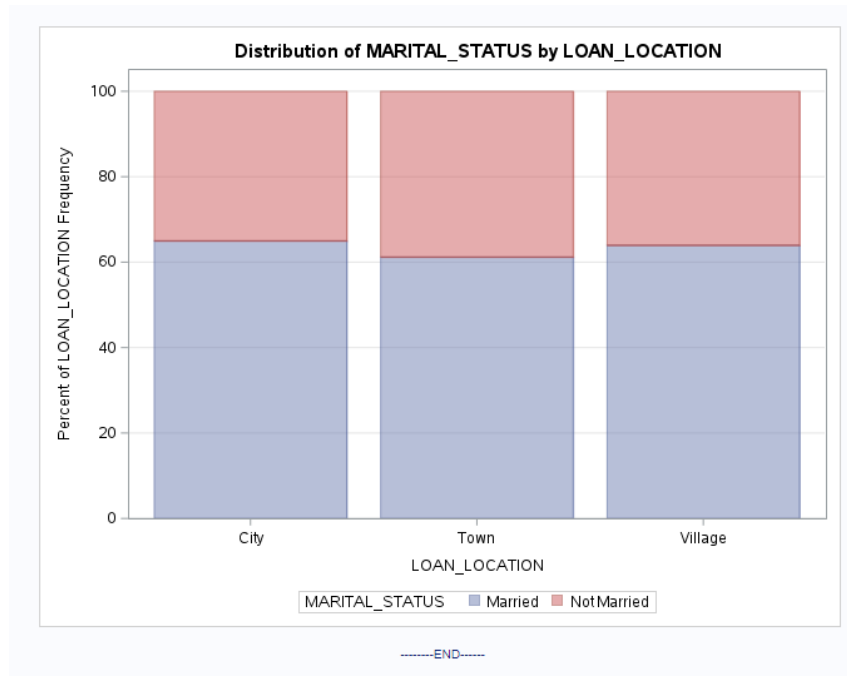


Figure 66

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

```

349 /* The SAS MACRO begins here*/
350 OPTIONS MCOMPILENOTE=ALL;
351 %MACRO MACRO_BVACATE_CONTI(ptitle1,ptitle2,pds_name,pcate_vari_name,pconti_vari_name);
352 TITLE1 &ptitle1;
353 TITLE2 &ptitle2;
354
355
356 PROC MEANS DATA = &pds_name;
357
358 CLASS &pcate_vari_name; /* It is a Categorical variable*/
359 VAR &pconti_vari_name; /* It is a Numeric/Continous variable */
360
361 RUN;
362 %MEND MACRO_BVACATE_CONTI;
363
364 /* The SAS MACRO ends here*/
365

```

Figure 67

```

81      NON,
82      %MEND MACRO_BVACATE_CONTI;
NOTE: The macro MACRO_BVACATE_CONTI completed compilation without errors.
      15 instructions 496 bytes.
83
84      /* The SAS MACRO ends here*/
85
86
87      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
97

```

Figure 68

```

366
367 /* To call the SAS MACRO - MACRO_BVACATE_CONTI */
368
369 /* loan location vs loan amount */
370 %MACRO_BVACATE_CONTI('Bivariate Analysis of the variables :', 'Categorical vs Continous', DAP67696.TESTING_DS,
371 LOAN_LOCATION, LOAN_AMOUNT);
372
373 /* family members vs loan duration */
374 %MACRO_BVACATE_CONTI('Bivariate Analysis of the variables :', 'Categorical vs Continous', DAP67696.TESTING_DS,
375 FAMILY_MEMBERS, LOAN_DURATION);
376
377 /* employment vs guarantee income */
378 %MACRO_BVACATE_CONTI('Bivariate Analysis of the variables :', 'Categorical vs Continous', DAP67696.TESTING_DS,
379 EMPLOYMENT, GUARANTEE_INCOME);
380
381
382
383

```

Figure 69

8.8 Screenshot(s) of the Output

**Bivariate Analysis of the variables :
Categorical vs Continous**

The MEANS Procedure

Analysis Variable : LOAN_AMOUNT						
LOAN_LOCATION	N Obs	N	Mean	Std Dev	Minimum	Maximum
City	140	138	136.2246377	65.0807492	28.0000000	480.0000000
Town	116	114	134.0438596	61.8013361	35.0000000	550.0000000
Village	111	110	138.1818182	56.3947720	28.0000000	390.0000000

Figure 70

**Bivariate Analysis of the variables :
Categorical vs Continous**

The MEANS Procedure

Analysis Variable : LOAN_DURATION						
FAMILY_MEMBERS	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	200	197	345.0558376	60.9019112	6.0000000	480.0000000
1	58	56	346.5000000	60.7184261	60.0000000	480.0000000
2	59	59	340.8813559	72.4852148	12.0000000	480.0000000
3+	40	39	330.7692308	75.2324334	120.0000000	480.0000000

Figure 71

Bivariate Analysis of the variables : Categorical vs Continuous						
The MEANS Procedure						
Analysis Variable : GUARANTEE_INCOME						
EMPLOYMENT	N Obs	N	Mean	Std Dev	Minimum	Maximum
No	307	307	1585.98	2446.09	0	24000.00
Yes	37	37	1381.16	1569.67	0	4831.00

Figure 72

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

```
407  
408 /* STEP - 1: List the details of the loan application with missing 'GENDER' details */  
409  
410 TITLE "STEP 1: List the details of the loan applicants with missing 'GENDER' details";  
411 FOOTNOTE '-----End-----';  
412  
413 PROC SQL;  
414  
415 SELECT*  
416 FROM DAP67696.TRAINING_DS e  
417 WHERE ( ( e.gender IS MISSING ) OR  
418         ( e.gender eq '' ) );  
419 QUIT;
```

Figure 73

8.8.6 Screenshot(s) of the Output

STEP 1: List the details of the loan applicants with missing 'GENDER' details

SME_LOAN_ID_NO	GENDER	MARITAL_STATUS	FAMILY_MEMBERS	QUALIFICATION	EMPLOYMENT	CANDIDATE_INCOME	GUARANTEE_INCOME	LOAN_AMOUNT	LOAN_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS
LP001050		Married	2	Under Graduate	No	3365	1917	112	360	0	Village	N
LP001448		Married	3+	Graduate	No	23803	0	370	360	1	Village	Y
LP001585		Married	3+	Graduate	No	51763	0	700	300	1	City	Y
LP001844		Married	0	Graduate	Yes	674	5295	168	360	1	Village	Y
LP002024		Married	0	Graduate	No	2473	1843	159	360	1	Village	N
LP002103		Married	1	Graduate	Yes	9833	1833	182	180	1	City	Y
LP002478		Married	0	Graduate	Yes	2083	4083	160	360	1	Town	Y
LP002501		Married	0	Graduate	No	16502	0	110	360	1	Town	Y
LP002530		Married	2	Graduate	No	2873	1872	132	360	0	Town	N
LP002825		Not Married	0	Graduate	No	3593	0	98	360	1	City	N
LP002872		Married	0	Graduate	No	3087	2210	136	360	0	Town	N
LP002925		Not Married	0	Graduate	No	4750	0	94	360	1	Town	Y
LP002933		Not Married	3+	Graduate	Yes	9357	0	282	360	1	Town	Y

-----End-----
Figure 74

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

```

423
424
425 /* STEP - 2: Find the total number of loan applicants who submitted their loan applications with GENDER details */
426
427 TITLE 'STEP 2 Find the total number of loan applicants who submitted their loan applications with GENDER details';
428 FOOTNOTE '-----End-----';
429
430 PROC SQL;
431
432 SELECT COUNT(*) Label = 'Number of Loan applicants'
433 FROM DAP67696.TRAINING_DS e
434 WHERE ( ( e.gender IS MISSING ) OR
435         ( e.gender eq '' ) );
436
437 QUIT;
438

```

Figure 75

8.8.9 Screenshot(s) of the Output

STEP 2 Find the total number of loan applicants who submitted their loan applications with GENDER details

Number of Loan applicants
13

-----End-----

Figure 76

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

```
442  
443 /* STEP - 3: Find the statistics and save the statistics in a temporary dataset */  
444  
445 PROC SQL;  
446  
447 CREATE TABLE DAP67696.TRAINING_GENDER_STAT_DS AS  
448 SELECT e.gender AS gender,  
449        COUNT(*) AS counts  
450 FROM DAP67696.TRAINING_DS e  
451 WHERE ( ( e.gender IS NOT MISSING ) OR  
452        ( e.gender ne '' ) )  
453 GROUP BY e.gender;  
454  
455 QUIT;  
456
```

Figure 77

8.9.2 Screenshot(s) of the Output

CODE
LOG
RESULTS
OUTPUT DATA

Table: DAP67696.TRAINING_GENDER_STAT_DS
View: Column names

Columns
Total rows: 2 Total columns: 2

☒ Select all
☒ gender
☒ counts

	gender
1	Female
2	Male

Figure 78

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

```
470 /* STEP 4: Find the Mod ... */
471
472 PROC SQL;
473
474 SELECT t.gender
475 FROM DAP67696.TRAINING_GENDER_STAT_DS t
476 WHERE t.counts eq ( SELECT MAX(t.counts) Label = 'highest_count'
477                     FROM DAP67696.TRAINING_GENDER_STAT_DS t );
478 /* Above is a sub-program to find the highest count */
479 QUIT;
480
481
482
```

Figure 79

8.9.5 Screenshot of the Output



gender
Male

Figure 80

8.9.6 Description

The output of the code display above is to show 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

```
484  
485 /* STEP 5: Make a copy of the dataset DAP67696.TRAINING_DS */  
486  
487 PROC SQL;  
488  
489 CREATE TABLE DAP67696.TRAINING_BK_DS AS  
490 SELECT*  
491 FROM DAP67696.TRAINING_DS;  
492  
493 QUIT;  
494
```

Figure 81

8.9.8 Screenshot(s) of the Output

Table: DAP67696.TRAINING_BK_DS

View: Column names

Filter: (none)

Columns

☒

Select all

☒

SME_LOAN_ID_NO

☒

GENDER

☒

MARITAL_STATUS

☒

FAMILY_MEMBERS

☒

QUALIFICATION

☒

EMPLOYMENT

☒

CANDIDATE_INCOME

☒

GUARANTEE_INCOME

☒

LOAN_AMOUNT

☒

LOAN_DURATION

☒

LOAN_HISTORY

☒

LOAN_LOCATION

☒

LOAN_APPROVAL_STATUS

Property

Value

Label

Name

Length

Type

Format

Informat

Total rows: 614

Total columns: 13

Rows 1-100

	SME_LOAN_ID...	GEND...	MARITAL_STA...	FAMILY_MEMB...	QUALIFICATION	EMPLOYM...	CANDIDATE_INCOME
1	LP001002	Male	Not Married	0	Graduate	No	5849
2	LP001003	Male	Married	1	Graduate	No	4583
3	LP001005	Male	Married	0	Graduate	Yes	3000
4	LP001006	Male	Married	0	Under Graduate	No	2583
5	LP001008	Male	Not Married	0	Graduate	No	6000
6	LP001011	Male	Married	2	Graduate	Yes	5417
7	LP001013	Male	Married	0	Under Graduate	No	2333
8	LP001014	Male	Married	3+	Graduate	No	3036
9	LP001018	Male	Married	2	Graduate	No	4006
10	LP001020	Male	Married	1	Graduate	No	12841
11	LP001024	Male	Married	2	Graduate	No	3200
12	LP001027	Male	Married	2	Graduate		2500
13	LP001028	Male	Married	2	Graduate	No	3073
14	LP001029	Male	Not Married	0	Graduate	No	1853
15	LP001030	Male	Married	2	Graduate	No	1299
16	LP001032	Male	Not Married	0	Graduate	No	4950
17	LP001034	Male	Not Married	1	Under Graduate	No	3596
18	LP001036	Female	Not Married	0	Graduate	No	3510
19	LP001038	Male	Married	0	Under Graduate	No	4887
20	LP001041	Male	Married	0	Graduate		2600
21	LP001043	Male	Married	0	Under Graduate	No	7660
22	LP001046	Male	Married	1	Graduate	No	5955
23	LP001047	Male	Married	0	Under Graduate	No	2600

Figure 82

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

```
499  
500 /* STEP 6: Impute the missing values found in the variable - GENDER */  
501  
502 PROC SQL;  
503  
504 UPDATE DAP67696.TRAINING_DS  
505 SET gender = ( SELECT t.gender  
506                FROM DAP67696.TRAINING_GENDER_STAT_DS t  
507                WHERE t.counts eq ( SELECT MAX(t.counts) Label = 'highest count'  
508                                FROM DAP67696.TRAINING_GENDER_STAT_DS t ) )  
509                                /* Above is a sub-program to find the highest count */  
510 WHERE ( ( gender IS MISSING ) OR  
511         ( gender eq '' ) );  
512  
513 QUIT;  
514
```

Figure 83

8.0.1.2 Screenshot(s) of the Output

```
80                                ( gender eq '' ) );  
NOTE: 13 rows were updated in DAP67696.TRAINING_DS.
```

Figure 84

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

```
518  
519 /* STEP - 7: (AI) List the details of the loan application with missing 'GENDER' details */  
520 TITLE "STEP 7 (AI): List the details of the loan applicants with missing 'GENDER' details";  
521 FOOTNOTE '-----End-----';  
522  
523 PROC SQL;  
524  
525 SELECT*  
526 FROM DAP67696.TRAINING_DS e  
527 WHERE ( ( e.gender IS MISSING ) OR  
528         ( e.gender eq '' ) );  
529 QUIT;
```

Figure 85

8.0.1.5 Screenshot(s) of the Output



STEP 7 (AI): List the details of the loan applicants with missing 'GENDER' details

-----End-----

Figure 86

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

```
537  
538 /* STEP - 1: List the details of the loan application with missing 'marital_status' details */  
539  
540 TITLE "STEP 1: List the details of the loan applicants with missing 'MARITAL_STATUS' details";  
541 FOOTNOTE '-----End-----';  
542  
543 PROC SQL;  
544  
545 SELECT*  
546 FROM DAP67696.TRAINING_DS e  
547 WHERE ( ( e.marital_status IS MISSING ) OR  
548         ( e.marital_status eq '' ) );  
549 QUIT;  
550
```

Figure 87

8.0.1.9 Screenshot(s) of the Output

STEP 1: List the details of the loan applicants with missing 'MARITAL_STATUS' details												
SME_LOAN_ID_NO	GENDER	MARITAL_STATUS	FAMILY_MEMBERS	QUALIFICATION	EMPLOYMENT	CANDIDATE_INCOME	GUARANTEE_INCOME	LOAN_AMOUNT	LOAN_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS
LP001357	Male			Graduate	No	3816	754	160	360	1	City	Y
LP001760	Male			Graduate	No	4758	0	158	480	1	Town	Y
LP002393	Female			Graduate	No	10047	0	.	240	1	Town	Y
-----End-----												

Figure 88

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

```
554 /* STEP - 2: Find the statistics and save the statistics in a temporary dataset */
555 PROC SQL;
556
557
558 CREATE TABLE DAP67696.TRAINING_MS_STAT_DS AS
559 SELECT e.marital_status AS marital_status,
560        COUNT(*) AS counts
561 FROM DAP67696.TRAINING_DS e
562 WHERE ( ( e.marital_status IS NOT MISSING ) OR
563        ( e.marital_status ne '' ) )
564 GROUP BY e.marital_status;
565
566 QUIT;
567
```

Figure 89

8.0.2.2 Screenshot(s) of the Output

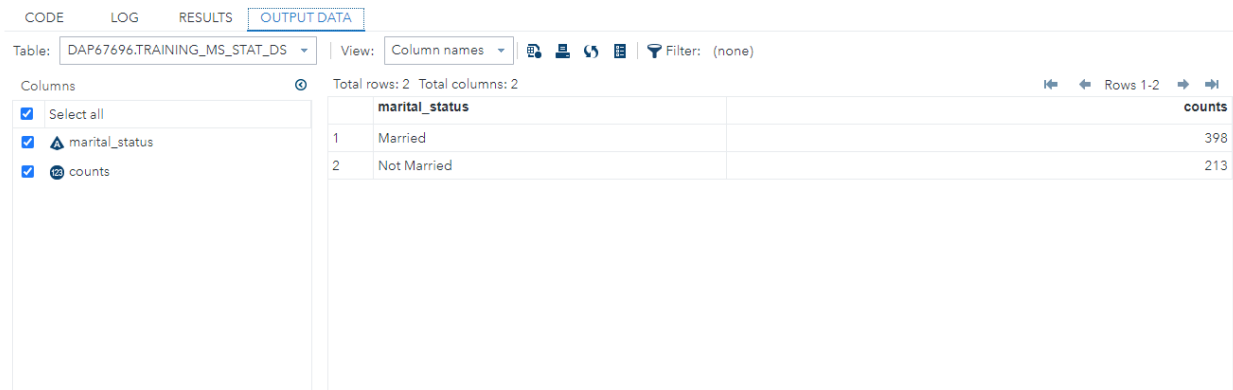


Table: DAP67696.TRAINING_MS_STAT_DS | View: Column names | Filter: (none)

Total rows: 2 Total columns: 2

	marital_status	counts
1	Married	398
2	Not Married	213

Figure 90

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

```
571 /* STEP 3: Impute the missing values found in the variable - MARITAL_STATUS */
572
573 PROC SQL;
574
575 UPDATE DAP67696.TRAINING_DS
576 SET marital_status = ( SELECT t.marital_status
577                       FROM DAP67696.TRAINING_MS_STAT_DS t
578                       WHERE t.counts eq ( SELECT MAX(t.counts) Label = 'highest count'
579                                           FROM DAP67696.TRAINING_MS_STAT_DS t ) )
580 /* Above is a sub-program to find the highest count */
581 WHERE ( ( marital_status IS MISSING ) OR
582         ( marital_status eq '' ) );
583
584 QUIT;
585
```

Figure 91

8.0.2.5 Screenshot(s) of the Output

NOTE: 3 rows were updated in DAP67696.TRAINING_DS.

Figure 92

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

```
589 /* STEP - 4: List the details of the loan application with missing 'marital_status' details */
590
591 TITLE "STEP 4: List the details of the loan applicants with missing 'MARITAL_STATUS' details";
592 FOOTNOTE '-----End-----';
593
594 PROC SQL;
595
596 SELECT*
597 FROM DAP67696.TRAINING_DS e
598 WHERE ( ( e.marital_status IS MISSING ) OR
599        ( e.marital_status eq '' ) );
600 QUIT;
601
```

Figure 93

8.0.2.8 Screenshot(s) of the Output

STEP 4: List the details of the loan applicants with missing 'MARITAL_STATUS' details

-----End-----

Figure 94

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

```
~~~~~
604 /* FAMILY_MEMBERS*/
605
606
607 /* STEP - 1: List the details of the loan application with missing 'FAMILY_MEMBERS' details */
608
609 TITLE "STEP 1: List the details of the loan applicants with missing 'FAMILY_MEMBERS' details";
610 FOOTNOTE '-----End-----';
611
612 PROC SQL;
613
614 SELECT*
615 FROM DAP67696.TRAINING_DS e
616 WHERE ( ( e.family_members IS MISSING ) OR
617        ( e.family_members eq '' ) );
618 QUIT;
619
```

Figure 95

8.0.3.2 Screenshot(s) of the Output

STEP 1: List the details of the loan applicants with missing 'FAMILY_MEMBERS' details												
SME_LOAN_ID_NO	GENDER	MARITAL_STATUS	FAMILY_MEMBERS	QUALIFICATION	EMPLOYMENT	CANDIDATE_INCOME	GUARANTEE_INCOME	LOAN_AMOUNT	LOAN_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS
LP001350	Male	Married		Graduate	No	13650	0	.	360	1	City	Y
LP001357	Marrie	Married		Graduate	No	3816	754	160	360	1	City	Y
LP001426	Male	Married		Graduate	No	5667	2667	180	360	1	Village	Y
LP001754	Male	Married		Under Graduate	Yes	4735	0	138	360	1	City	N
LP001780	Marrie	Married		Graduate	No	4758	0	158	480	1	Town	Y
LP001945	Female	Not Married		Graduate	No	5417	0	143	480	0	City	N
LP001972	Male	Married		Under Graduate	No	2875	1750	105	360	1	Town	Y
LP002100	Male	Not Married		Graduate	No	2833	0	71	360	1	City	Y
LP002106	Male	Married		Graduate	Yes	5503	4490	70	.	1	Town	Y
LP002130	Male	Married		Under Graduate	No	3523	3230	152	360	0	Village	N
LP002144	Female	Not Married		Graduate	No	3813	0	116	180	1	City	Y
LP002393	Marrie	Married		Graduate	No	10047	0	.	240	1	Town	Y
LP002662	Male	Married		Under Graduate	No	3074	1600	123	360	0	Town	N
LP002847	Male	Married		Graduate	No	5116	1451	165	360	0	City	N
LP002943	Male	Not Married		Graduate	No	2967	0	88	360	0	Town	N

Figure 96

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

```
621  
622 /* STEP - 2: Find the statistics and save the statistics in a temporary dataset */  
623 PROC SQL;  
624  
625  
626 CREATE TABLE DAP67696.TRAINING_FM_STAT_DS AS  
627 SELECT t.family_members AS family_members ,  
628        COUNT(*) AS counts  
629 FROM DAP67696.TRAINING_DS t  
630 WHERE ( ( t.family_members IS NOT MISSING ) OR  
631        ( t.family_members ne '' ) )  
632 GROUP BY t.family_members;  
633  
634 QUIT;  
---
```

Figure 97

8.0.3.5 Screenshot(s) of the Output

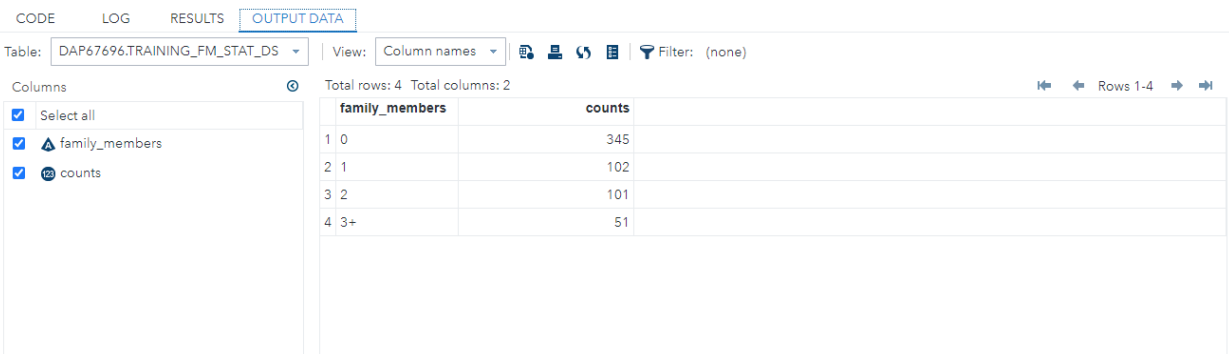


Table: DAP67696.TRAINING_FM_STAT_DS View: Column names Filter: (none)

Total rows: 4 Total columns: 2

	family_members	counts
1	0	345
2	1	102
3	2	101
4	3+	51

Figure 98

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

```
636  
637 /* STEP 3: List the details of the loan application with '3+' family members */  
638 TITLE "List the details of the loan application with '3+' family members";  
639 FOOTNOTE '-----End-----';  
640  
641 PROC SQL;  
642  
643 /*3+ SUBSTAR(family_members,1,1) UBSTR(family_members,2,1)*/  
644  
645 SELECT t.family_members,  
646 substr(t.family_members,1,1) Label = '1,1',  
647 substr(t.family_members,2,1) Label = '2,1'  
648 FROM DAP67696.TRAINING_DS t  
649 WHERE ( t.family_members eq '3+' );  
650  
651 QUIT;  
652
```

Figure 99

8.0.3.8 Screenshot(s) of the Output

List the details of the loan application with '3+' family members		
FAMILY_MEMBERS	1,1	2,1
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+
3+	3	+

Figure 100

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

```
656  
657 /* STEP 4: Remove the '+' found in the family members variable and update the dataset DAP67696.TRAINING_DS */  
658  
659  
660 PROC SQL;  
661  
662 UPDATE DAP67696.TRAINING_DS  
663 SET family_members = substr(family_members,1,1)  
664 WHERE ( family_members eq '3+' );  
665  
666 QUIT;
```

Figure 101

8.0.4.1 Screenshot(s) of the Output

NOTE: 51 rows were updated in DAP67696.TRAINING_DS.

Figure 102

8.0.4.2 Description

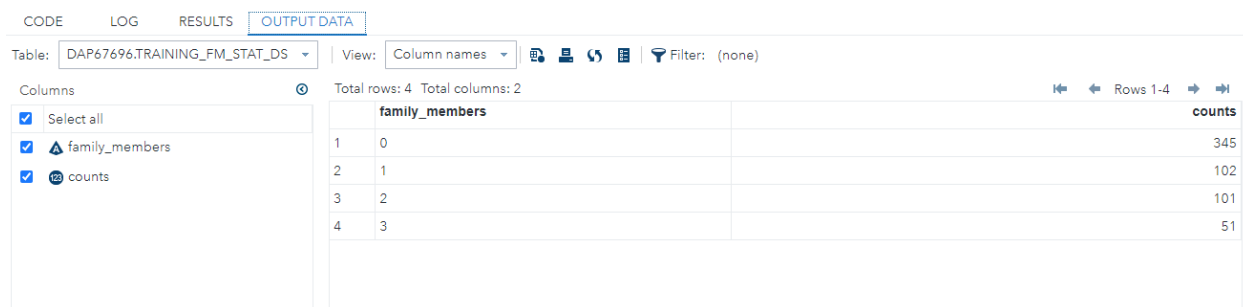
Step 4 is done remove the '+' symbol found in the family members and update the dataset DAP67696.TRAINING_DS whereby there are 51 rows were updated in the DAP67696.TRAINING_DS. This step is to remove the '+' symbol as it is crucial to data cleaning of the datasets.

8.0.4.3 SAS Source Codes

```
669  
670 /* STEP - 5: (After removing the + symbol found in the family_members variable ) */  
671 PROC SQL;  
672  
673  
674 CREATE TABLE DAP67696.TRAINING_FM_STAT_DS AS  
675 SELECT t.family_members AS family_members,  
676        COUNT(*) AS counts  
677 FROM DAP67696.TRAINING_DS t  
678 WHERE ( ( t.family_members IS NOT MISSING ) OR  
679        ( t.family_members ne '' ) )  
680 GROUP BY t.family_members;  
681  
682 QUIT;  
683
```

Figure 103

8.0.4.4 Screenshot(s) of the Output



CODE LOG RESULTS OUTPUT DATA

Table: DAP67696.TRAINING_FM_STAT_DS View: Column names Filter: (none)

Columns: Select all family_members counts

Total rows: 4 Total columns: 2

	family_members	counts
1	0	345
2	1	102
3	2	101
4	3	51

Figure 104

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

```
683  
684 /* STEP 6: Impute the missing values found in the variable - FAMILY_MEMBERS */  
685  
686 PROC SQL;  
687  
688 UPDATE DAP67696.TRAINING_DS  
689 SET family_members = ( SELECT t.family_members  
690                        FROM DAP67696.TRAINING_FM_STAT_DS t  
691                        WHERE t.counts eq ( SELECT MAX(t.counts) Label = 'highest count'  
692                                           FROM DAP67696.TRAINING_FM_STAT_DS t ) )  
693                        /* Above is a sub-program to find the highest count */  
694 WHERE ( ( family_members IS MISSING ) OR  
695         ( family_members eq '' ) );  
696  
697 QUIT;  
698
```

Figure 105

8.0.4.7 Screenshot(s) of the Output

NOTE: 15 rows were updated in DAP67696.TRAINING_DS.

Figure 106

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

```
700  
701  
702 /* STEP - 7: (After I) List the details of the loan applicants with missing 'FAMILY_MEMBERS' details */  
703  
704 TITLE "STEP 7: List the details of the loan applicants with missing 'FAMILY_MEMBERS' details";  
705 FOOTNOTE '-----End-----';  
706  
707 PROC SQL;  
708  
709 SELECT *  
710 FROM DAP67696.TRAINING_DS e  
711 WHERE ( ( e.family_members IS MISSING ) OR  
712         ( e.family_members eq '' ) );  
713  
714 QUIT;  
715
```

Figure 107

8.0.5 Screenshot(s) of the Output



```
STEP 7: List the details of the loan applicants with missing 'FAMILY_MEMBERS' details
----End----
```

Figure 108

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

```

722 /* LOAN_AMOUNT*/
723
724 /* STEP - 1: List the details of the loan applicants with missing 'LOAN_AMOUNT' details */
725
726 TITLE "STEP 1: List the details of the loan applicants with missing 'LOAN_AMOUNT' details";
727 FOOTNOTE '-----End-----';
728
729 PROC SQL;
730
731 SELECT *
732 FROM DAP67696.TRAINING_DS t
733 WHERE ( ( t.loan_amount IS MISSING ) OR
734         ( t.loan_amount eq . ) );
735 QUIT;
736

```

Figure 109

8.0.5.4 Screenshot(s) of the Output

STEP 1: List the details of the loan applicants with missing 'LOAN_AMOUNT' details												
SME_LOAN_ID_NO	GENDER	MARITAL_STATUS	FAMILY_MEMBERS	QUALIFICATION	EMPLOYMENT	CANDIDATE_INCOME	GUARANTEE_INCOME	LOAN_AMOUNT	LOAN_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS
LP001002	Male	Not Married	0	Graduate	No	5649	0	.	360	1	City	Y
LP001106	Male	Married	0	Graduate	No	2275	2067	.	360	1	City	Y
LP001213	Male	Married	1	Graduate	No	4945	0	.	360	0	Village	N
LP001266	Male	Married	1	Graduate	Yes	2395	0	.	360	1	Town	Y
LP001326	Male	Not Married	0	Graduate	No	6782	0	.	360	.	City	N
LP001350	0	Married	0	Graduate	No	13050	0	.	360	1	City	Y
LP001356	Male	Married	0	Graduate	No	4052	3583	.	360	1	Town	Y
LP001362	Female	Not Married	1	Graduate	Yes	7451	0	.	360	1	Town	Y
LP001449	Male	Not Married	0	Graduate	No	3885	1640	.	360	1	Village	Y
LP001662	Male	Married	3	Under Graduate	No	3992	0	.	180	1	City	N
LP001622	Male	Married	0	Graduate	No	20667	0	.	360	1	Village	N
LP001990	Male	Not Married	0	Under Graduate	No	2000	0	.	360	1	City	N
LP002054	Male	Married	2	Under Graduate	No	3601	1590	.	360	1	Village	Y
LP002113	Female	Not Married	3	Under Graduate	No	1630	0	.	360	0	City	N
LP002243	Male	Married	0	Under Graduate	No	3010	3136	.	360	0	City	N
LP002393	0	Married	0	Graduate	No	10047	0	.	240	1	Town	Y
LP002401	Male	Married	0	Graduate	No	2213	1125	.	360	1	City	Y
LP002533	Male	Married	2	Graduate	No	2647	1603	.	360	1	City	N
LP002667	Male	Not Married	0	Graduate	No	4680	2087	.	360	1	Town	N
LP002778	Male	Married	2	Graduate	Yes	6633	0	.	360	0	Village	N
LP002784	Male	Married	1	Under Graduate	No	2492	2375	.	360	1	Village	Y
LP002860	Male	Married	0	Under Graduate	No	2400	3800	.	180	1	City	N
-----End-----												

Figure 110

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

```
737  
738  
739 /* STEP - 2: Find the total number of the loan applicants with missing 'LOAN_AMOUNT' details */  
740  
741 TITLE "STEP 2: List the details of the loan applicants with missing 'LOAN_AMOUNT' details";  
742 FOOTNOTE '-----End-----';  
743  
744 PROC SQL;  
745  
746 SELECT COUNT(*) Label = 'Number of applicants'  
747 FROM DAP67696.TRAINING_DS t  
748 WHERE ( ( t.loan_amount IS MISSING ) OR  
749         ( t.loan_amount eq . ) );  
750 QUIT;  
751
```

Figure 111

8.0.5.7 Screenshot(s) of the Output



Figure 112

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

```

753
754 /* STEP 3: Impute the missing values found in the variable - LOAN_AMOUNT */;
755
756 PROC STDIZE DATA = DAP67696.TRAINING_DS REONLY
757
758 METHOD = MEAN OUT = DAP67696.TRAINING_DS;
759 VAR loan_amount;
760
761 QUIT;
---
```

Figure 113

8.0.6 Screenshot(s) of the Output

Table: DAP67696.TRAINING_DS View: Column names Filter: (none)

Columns: Select all

- ☒ SME_LOAN_ID_NO
- ☒ GENDER
- ☒ MARITAL_STATUS
- ☒ FAMILY_MEMBERS
- ☒ QUALIFICATION
- ☒ EMPLOYMENT
- ☒ CANDIDATE_INCOME
- ☒ GUARANTEE_INCOME
- ☒ LOAN_AMOUNT
- ☒ LOAN_DURATION
- ☒ LOAN_HISTORY
- ☒ LOAN_LOCATION
- ☒ LOAN_APPROVAL_STATUS

Total rows: 614 Total columns: 13 Rows 1-100

	SME_LOAN_ID...	GEND...	MARITAL_STA...	FAMILY_MEMB...	QUALIFICATION	EMPLOYM...	CANDIDATE_INCOME
1	LP001002	Male	Not Married	0	Graduate	No	5849
2	LP001003	Male	Married	1	Graduate	No	4583
3	LP001005	Male	Married	0	Graduate	Yes	3000
4	LP001006	Male	Married	0	Under Graduate	No	2583
5	LP001008	Male	Not Married	0	Graduate	No	6000
6	LP001011	Male	Married	2	Graduate	Yes	5417
7	LP001013	Male	Married	0	Under Graduate	No	2333
8	LP001014	Male	Married	3	Graduate	No	3036
9	LP001018	Male	Married	2	Graduate	No	4006
10	LP001020	Male	Married	1	Graduate	No	12841
11	LP001024	Male	Married	2	Graduate	No	3200
12	LP001027	Male	Married	2	Graduate		2500
13	LP001028	Male	Married	2	Graduate	No	3073
14	LP001029	Male	Not Married	0	Graduate	No	1853

Figure 114

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

```
764 /* STEP - 4: (AI) List the details of the loan applicants with missing 'LOAN_AMOUNT' details */
765
766
767 TITLE "STEP 4: List the details of the loan applicants with missing 'LOAN_AMOUNT' details";
768 FOOTNOTE '-----End-----';
769
770 PROC SQL;
771
772 SELECT *
773 FROM DAP67696.TRAINING_DS t
774 WHERE ( ( t.loan_amount IS MISSING ) OR
775         ( t.loan_amount eq . ) );
776 QUIT;
777
778
```

Figure 115

8.0.7.2 Screenshot(s) of the Output

STEP 4: List the details of the loan applicants with missing 'LOAN_AMOUNT' details

-----End-----

Figure 116

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

```

779
780 /* LOAN_DURATION*/
781
782 /* STEP - 1: List the details of the loan applicants with missing 'LOAN_DURATION' details */
783
784 TITLE "STEP 1: List the details of the loan applicants with missing 'LOAN_DURATION' details";
785 FOOTNOTE '-----End-----';
786
787 PROC SQL;
788
789 SELECT *
790 FROM DAP67696.TRAINING_DS t
791 WHERE ( ( t.loan_duration IS MISSING ) OR
792         ( t.loan_duration eq . ) );
793 QUIT;
794

```

Figure 117

8.0.7.6 Screenshot(s) of the Output

STEP 1: List the details of the loan applicants with missing 'LOAN_DURATION' details												
SME_LOAN_ID_NO	GENDER	MARITAL_STATUS	FAMILY_MEMBERS	QUALIFICATION	EMPLOYMENT	CANDIDATE_INCOME	GUARANTEE_INCOME	LOAN_AMOUNT	LOAN_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS
LP001041	Male	Married	0	Graduate	No	2600	3500	115	-	1	City	Y
LP001109	Male	Married	0	Graduate	No	1828	1330	100	-	0	City	N
LP001136	Male	Married	0	Under Graduate	Yes	4695	0	96	-	1	City	Y
LP001137	Female	Not Married	0	Graduate	No	3410	0	88	-	1	City	Y
LP001250	Male	Married	3	Under Graduate	No	4755	0	95	-	0	Town	N
LP001391	Male	Married	0	Under Graduate	No	3572	4114	152	-	0	Village	N
LP001574	Male	Married	0	Graduate	No	3707	3166	182	-	1	Village	Y
LP001669	Female	Not Married	0	Under Graduate	No	1907	2385	120	-	1	City	Y
LP001749	Male	Married	0	Graduate	No	7578	1010	175	-	1	Town	Y
LP001770	Male	Not Married	0	Under Graduate	No	3189	2598	120	-	1	Village	Y
LP002106	0	Married	0	Graduate	Yes	5503	4490	70	-	1	Town	Y
LP002188	Male	Not Married	0	Graduate	No	5124	0	124	-	0	Village	N
LP002357	Female	Not Married	0	Under Graduate	No	2720	0	80	-	0	City	N
LP002352	Male	Married	1	Graduate	No	7250	1687	110	-	0	City	N

-----End-----

Figure 118

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

```
795 |
796 /* STEP - 2: Find the total number of the loan applicants with missing 'LOAN_DURATION' details */
797 |
798 TITLE "STEP 2: List the details of the loan applicants with missing 'LOAN_DURATION' details";
799 FOOTNOTE '-----End-----';
800 |
801 PROC SQL;
802 |
803 SELECT COUNT(*) Label = 'Number of applicants'
804 FROM DAP67696.TRAINING_DS t
805 WHERE ( ( t.loan_duration IS MISSING ) OR
806         ( t.loan_duration eq . ) );
807 QUIT;
808 |
809 |
```

Figure 119

8.0.7.9 Screenshot(s) of the Output



Figure 120

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

```

810 /* STEP 3: Impute the missing values found in the variable - LOAN_DURATION */;
811
812 PROC STDIZE DATA = DAP67696.TRAINING_DS REONLY
813
814 METHOD = MEAN OUT = DAP67696.TRAINING_DS;
815 VAR loan_duration;
816
817 QUIT;
818
819

```

Figure 121

8.0.8.2 Screenshot(s) of the Output

CODE LOG RESULTS **OUTPUT DATA**

Table: DAP67696.TRAINING_DS View: Column names Filter: (none)

Columns: Select all

- ☒ SME_LOAN_ID_NO
- ☒ GENDER
- ☒ MARITAL_STATUS
- ☒ FAMILY_MEMBERS
- ☒ QUALIFICATION
- ☒ EMPLOYMENT
- ☒ CANDIDATE_INCOME
- ☒ GUARANTEE_INCOME
- ☒ LOAN_AMOUNT
- ☒ LOAN_DURATION
- ☒ LOAN_HISTORY
- ☒ LOAN_LOCATION
- ☒ LOAN_APPROVAL_STATUS

Total rows: 614 Total columns: 13 Rows 1-100

	SME_LOAN_ID...	GEND...	MARITAL_STA...	FAMILY_MEMB...	QUALIFICATION	EMPLOYM...	CANDIDATE_INCOME
1	LP001002	Male	Not Married	0	Graduate	No	5849
2	LP001003	Male	Married	1	Graduate	No	4583
3	LP001005	Male	Married	0	Graduate	Yes	3000
4	LP001006	Male	Married	0	Under Graduate	No	2583
5	LP001008	Male	Not Married	0	Graduate	No	6000
6	LP001011	Male	Married	2	Graduate	Yes	5417
7	LP001013	Male	Married	0	Under Graduate	No	2333
8	LP001014	Male	Married	3	Graduate	No	3036
9	LP001018	Male	Married	2	Graduate	No	4006
10	LP001020	Male	Married	1	Graduate	No	12841
11	LP001024	Male	Married	2	Graduate	No	3200
12	LP001027	Male	Married	2	Graduate		2500
13	LP001028	Male	Married	2	Graduate	No	3073
14	LP001029	Male	Not Married	0	Graduate	No	1853

Figure 122

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

```
819 |
820 | /* STEP - 4: (AI) List the details of the loan applicants with missing 'LOAN_DURATION' details */
821 |
822 | TITLE "STEP 4: List the details of the loan applicants with missing 'LOAN_DURATION' details";
823 | FOOTNOTE '-----End-----';
824 |
825 | PROC SQL;
826 |
827 | SELECT *
828 | FROM DAP67696.TRAINING_DS t
829 | WHERE ( ( t.loan_duration IS MISSING ) OR
830 |         ( t.loan_duration eq . ) );
831 | QUIT;
832 |
```

Figure 123

8.0.8.5 Screenshot(s) of the Output



The screenshot displays the output of a SAS program. It features a title bar at the top that reads "STEP 4: List the details of the loan applicants with missing 'LOAN_DURATION' details". Below the title bar, the text "-----End-----" is centered, indicating the end of the output for this step.

Figure 124

8.0.8.6 Description

Step 4 shows that the lists the details of the loan applicants with missing LOAN_DURATION 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

```
863 /* GUARANTEE_INCOME */
864
865 /* STEP - 1: List the details of the loan applicants with missing 'GUARANTEE_INCOME' details */
866
867 TITLE "STEP 1: List the details of the loan applicants with missing 'GUARANTEE_INCOME' details";
868 FOOTNOTE '-----End-----';
869
870
871 PROC SQL;
872
873 SELECT *
874 FROM DAP67696.TRAINING_DS t
875 WHERE ( ( t.guarantee_income IS MISSING ) OR
876        ( t.guarantee_income eq . ) );
877 QUIT;
```

Figure 125

8.0.8.9 Screenshot(s) of the Output

STEP 1: List the details of the loan applicants with missing 'GUARANTEE_INCOME' details

-----End-----

Figure 126

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

```
879  
880 /* STEP - 2: Find the total number of the loan applicants with missing 'GUARANTEE_INCOME' details */  
881  
882 TITLE "STEP 2: List the details of the loan applicants with missing 'GUARANTEE_INCOME' details";  
883 FOOTNOTE '-----End-----';  
884  
885 PROC SQL;  
886  
887 SELECT COUNT(*) Label = 'Number of applicants'  
888 FROM DAP67696.TRAINING_DS t  
889 WHERE ( ( t.guarantee_income IS MISSING ) OR  
890         ( t.guarantee_income eq . ) );  
891 QUIT;  
892
```

Figure 127

8.0.9.3 Screenshot(s) of the Output

STEP 2: List the details of the loan applicants with missing 'GUARANTEE_INCOME' details	
Number of applicants	0
-----End-----	

Figure 128

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

```

894 /* STEP 3: Impute the missing values found in the variable - GUARANTEE_INCOME */;
895
896 PROC STDIZE DATA = DAP67696.TRAINING_DS REONLY
897
898 METHOD = MEAN OUT = DAP67696.TRAINING_DS;
899 VAR guarantee_income;
900
901 QUIT;
902

```

Figure 129

8.0.9.6 Screenshot(s) of the Output

CODELOGRESULTS

OUTPUT DATA

Table: DAP67696.TRAINING_DS

View: Column names

Filter: (none)

Columns

Select all

SME_LOAN_ID_NO

GENDER

MARITAL_STATUS

FAMILY_MEMBERS

QUALIFICATION

EMPLOYMENT

CANDIDATE_INCOME

GUARANTEE_INCOME

LOAN_AMOUNT

LOAN_DURATION

LOAN_HISTORY

LOAN_LOCATION

LOAN_APPROVAL_STATUS

Property

Value

Label

Name

Length

Type

Format

Informat

Total rows: 614 Total columns: 13

Rows 1-100

	SME_LOAN_ID...	GEND...	MARITAL_STA...	FAMILY_MEMB...	QUALIFICATION	EMPLOYM...	CANDIDATE_INCOME
1	LP001002	Male	Not Married	0	Graduate	No	5849
2	LP001003	Male	Married	1	Graduate	No	4583
3	LP001005	Male	Married	0	Graduate	Yes	3000
4	LP001006	Male	Married	0	Under Graduate	No	2583
5	LP001008	Male	Not Married	0	Graduate	No	6000
6	LP001011	Male	Married	2	Graduate	Yes	5417
7	LP001013	Male	Married	0	Under Graduate	No	2333
8	LP001014	Male	Married	3	Graduate	No	3036
9	LP001018	Male	Married	2	Graduate	No	4006
10	LP001020	Male	Married	1	Graduate	No	12841
11	LP001024	Male	Married	2	Graduate	No	3200
12	LP001027	Male	Married	2	Graduate		2500
13	LP001028	Male	Married	2	Graduate	No	3073
14	LP001029	Male	Not Married	0	Graduate	No	1853
15	LP001030	Male	Married	2	Graduate	No	1299
16	LP001032	Male	Not Married	0	Graduate	No	4950
17	LP001034	Male	Not Married	1	Under Graduate	No	3596
18	LP001036	Female	Not Married	0	Graduate	No	3510
19	LP001038	Male	Married	0	Under Graduate	No	4887
20	LP001041	Male	Married	0	Graduate		2600
21	LP001043	Male	Married	0	Under Graduate	No	7660
22	LP001046	Male	Married	1	Graduate	No	5955
23	LP001047	Male	Married	0	Under Graduate	No	2600

Figure 130

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

```
904 /* STEP - 4: (AI) List the details of the loan applicants with missing 'GUARANTEE_INCOME' details */
905
906 TITLE "STEP 4: List the details of the loan applicants with missing 'GUARANTEE_INCOME' details";
907 FOOTNOTE '-----End-----';
908
909 PROC SQL;
910
911 SELECT *
912 FROM DAP67696.TRAINING_DS t
913 WHERE ( ( t.guarantee_income IS MISSING ) OR
914        ( t.guarantee_income eq . ) );
915 QUIT;
```

Figure 131

8.0.9.9 Screenshot(s) of the Output



The screenshot displays the SAS output for Step 4. It features a title bar that reads "STEP 4: List the details of the loan applicants with missing 'GUARANTEE_INCOME' details" and a footnote below it that reads "-----End-----".

Figure 132

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

```
1399 /* Model Creation */
1400
1401 PROC LOGISTIC DATA = DAP67696.TRAINING_DS OUTMODEL = DAP67696.TRAINING_LR_MODEL;
1402 CLASS
1403 GENDER
1404 FAMILY_MEMBERS
1405 LOAN_LOCATION
1406 MARITAL_STATUS
1407 QUALIFICATION
1408 EMPLOYMENT
1409 LOAN_APPROVAL_STATUS
1410 LOAN_HISTORY;
1411
1412
1413 /* Above are categorical variables */
1414
1415 MODEL LOAN_APPROVAL_STATUS = /*place here all independent variables */
1416 /* LOAN_APPLICATION_STATUS is a dependent variable */
1417 GENDER
1418 FAMILY_MEMBERS
1419 LOAN_LOCATION
1420 LOAN_HISTORY
1421 MARITAL_STATUS
1422 QUALIFICATION
1423 EMPLOYMENT
1424 CANDIDATE_INCOME
1425 GUARANTEE_INCOME
1426 LOAN_AMOUNT
1427 LOAN_DURATION;
1428
1429 OUTPUT OUT = DAP67696.TRAINING_OUT_DS P = PPRED_PROB;
1430 /*PRED_PROB ->Predicted probability - variable to hold predicted probability
1431 OUT -> the output will be stored in the dataset
1432 Akaike Information criterion must ( AIC ) < SC (Schwarz Criterion) */
1433 /* If Pr > ChiSq is <= 0.05, it means that independent variable is an
1434 important variable and is truly contributing to predict dependent variable */
1435 RUN;
```

Figure 133

9.1.2 Screenshot(s) of the Output

The LOGISTIC Procedure	
Model Information	
Data Set	DAP87696.TRAINING_DS
Response Variable	LOAN_APPROVAL_STATUS
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Figure 134

Number of Observations Read	614
Number of Observations Used	614

Figure 135

The analysis is accepted.

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.

9.1.4 Screenshot(s) of the Output



A screenshot of a software output window titled "Model Convergence Status". It contains a single line of text: "Convergence criterion (GCONV=1E-8) satisfied." A blue arrow points from this text to a yellow-bordered box on the right.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

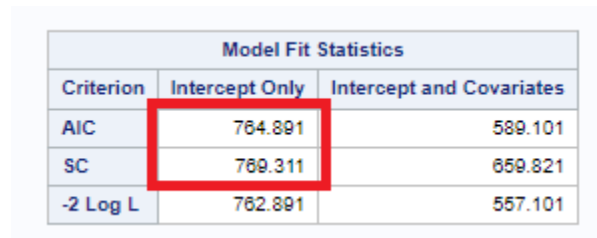
The model convergence status is acceptable.

Figure 136

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 software output window titled "Model Fit Statistics". It contains a table with three columns: "Criterion", "Intercept Only", and "Intercept and Covariates". The rows are "AIC", "SC", and "-2 Log L". The values for "Intercept Only" are 764.891, 769.311, and 762.891 respectively. The values for "Intercept and Covariates" are 589.101, 659.821, and 557.101 respectively. A red box highlights the "Intercept Only" column.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	764.891	589.101
SC	769.311	659.821
-2 Log L	762.891	557.101

Figure 137

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

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.00504	0.7383	0.0000	0.9946
GENDER	0	1	-0.1291	0.5449	0.0561	0.8128
GENDER	Female	1	0.0581	0.3268	0.0316	0.8590
FAMILY_MEMBERS	0	1	-0.0329	0.1881	0.0307	0.8609
FAMILY_MEMBERS	1	1	0.4290	0.2260	3.6024	0.0577
FAMILY_MEMBERS	2	1	-0.3334	0.2542	1.7201	0.1897
LOAN_LOCATION	City	1	0.1587	0.1522	1.0876	0.2970
LOAN_LOCATION	Town	1	-0.5319	0.1575	11.4034	0.0007
LOAN_HISTORY	0	1	1.9737	0.2115	87.0894	<.0001
MARITAL_STATUS	Married	1	-0.2865	0.1265	5.1324	0.0235
QUALIFICATION	Graduate	1	-0.2061	0.1299	2.5180	0.1126
EMPLOYMENT	No	1	-0.0131	0.1587	0.0068	0.9340
CANDIDATE_INCOME		1	-0.00001	0.000024	0.2246	0.6356
GUARANTEE_INCOME		1	0.000053	0.000035	2.2578	0.1329
LOAN_AMOUNT		1	0.00191	0.00180	1.4245	0.2327
LOAN_DURATION		1	0.00134	0.00184	0.5310	0.4662

Figure 138

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 P_r value of 0.007, loan history – 0 with P_r value of 0.001 and marital status – married with P_r value of 0.0235.

9.2 Screenshot(s) of the Output

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
GENDER	2	0.0813	0.9898
FAMILY MEMBERS	3	4.3321	0.2278
LOAN_LOCATION	2	12.0760	0.0024
LOAN_HISTORY	1	87.0894	<.0001
MARITAL_STATUS	1	5.1324	0.0235
QUALIFICATION	1	2.5180	0.1126
EMPLOYMENT	1	0.0068	0.9340
CANDIDATE_INCOME	1	0.2246	0.6356
GUARANTEE_INCOME	1	2.2578	0.1329
LOAN_AMOUNT	1	1.4245	0.2327
LOAN_DURATION	1	0.5310	0.4662

Figure 139

9.2.1 Description

Figure above shows the output of the type 3 analysis of effects where if $P_r > ChiSq$ is lesser than 0.05 it shows that these variables that meet the criteria of variable importance. The variable importance means that if the P_r 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\

```
958 /*****
959 Predict the loan approval status using the model created
960 erewerer
961 *****/
962
963 PROC LOGISTIC INMODEL = DAP67696.TRAINING_LR_MODEL; /* It is the model the model you created */
964
965 SCORE DATA = DAP67696.TRAINING_DS /* Test ds*/
966 OUT = DAP67696.TESTING_PREDICTED_DS; /*Location of output */
967
968 QUIT;
969
```

Figure 140

9.2.3 Screenshot(s) of the Output

mydap_project_DAP_pt_aug_2023_67696.sas

CODE

LOG

RESULTS

OUTPUT DATA

Table: DAP67696.TESTING_PREDICTED_DS

Views: Column names

Filter: (none)

Columns

Select all

SME_LOAN_ID_NO

GENDER

MARITAL_STATUS

FAMILY_MEMBERS

QUALIFICATION

EMPLOYMENT

CANDIDATE_INCOME

GUARANTEE_INCOME

LOAN_AMOUNT

LOAN_DURATION

LOAN_HISTORY

LOAN_LOCATION

LOAN_APPROVAL_STATUS

F_LOAN_APPROVAL_STATUS

Property

<

Figure 141

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

```

974 TITLE 'List the details of the dataset carrying the loan approval status predicted - DAP67696.TESTING_PREDICTED_DS';
975 FOOTNOTE '-----End-----';
976
977 PROC SQL;
978
979 SELECT*
980 FROM DAP67696.TESTING_PREDICTED_DS;
981
982 QUIT;
983

```

Figure 142

9.2.7 Screenshot(s) of the Output

Table of Contents

List the details of the dataset carrying the loan approval status predicted - DAP67696.TESTING_PREDICTED_DS

ML_LOAN_ID_NO	GENDER	MARITAL_STATUS	FAMILY_MEMBERS	QUALIFICATION	EMPLOYMENT	CANDIDATE_INCOME	GUARANTEE_INCOME	LOAN_AMOUNT	LOAN_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	Info: LOAN_APPROVAL_STATUS	Predicted Probability: LOAN_APPROVAL_STATUS	Predicted Probab: LOAN_APPROVAL_STATUS
P001002	Male	Not Married	0	Graduate	No	5546	0	146.41216216	360	1	City	Y	Y	Y	0.254652	0.73
P001003	Male	Married	1	Graduate	No	4553	1938	126	360	1	Village	N	N	Y	0.267124	0.70
P001008	Male	Married	0	Graduate	Yes	3000	0	66	360	1	City	Y	Y	Y	0.156087	0.84
P001009	Male	Married	0	Under Graduate	No	2563	2358	120	360	1	City	Y	Y	Y	0.255982	0.74
P001008	Male	Not Married	0	Graduate	No	6030	0	141	360	1	City	Y	Y	Y	0.262524	0.73
P001011	Male	Married	2	Graduate	Yes	5417	4195	287	360	1	City	Y	Y	Y	0.165205	0.83
P001013	Male	Married	0	Under Graduate	No	2333	1818	65	360	1	City	Y	Y	Y	0.235928	0.76
P001014	Male	Married	3	Graduate	No	3035	2504	158	360	0	Town	N	N	N	0.885981	0.13
P001018	Male	Married	2	Graduate	No	4008	1925	185	360	1	City	Y	Y	Y	0.14759	0.8
P001020	Male	Married	1	Graduate	No	12341	10995	349	360	1	Town	N	N	Y	0.25177	0.7
P001024	Male	Married	2	Graduate	No	3200	100	70	360	1	City	Y	Y	Y	0.12216	0.8
P001027	Male	Married	2	Graduate	No	2500	1540	109	360	1	City	Y	Y	Y	0.133234	0.86
P001028	Male	Married	2	Graduate	No	3073	8105	200	360	1	City	Y	Y	Y	0.23909	0.7
P001029	Male	Not Married	0	Graduate	No	1253	2540	114	360	1	Village	N	N	Y	0.335214	0.66
P001030	Male	Married	2	Graduate	No	1339	1036	17	120	1	City	Y	Y	Y	0.268565	0.61
P001032	Male	Not Married	0	Graduate	No	4950	0	125	360	1	City	Y	Y	Y	0.255086	0.74
P001034	Male	Not Married	1	Under Graduate	No	3566	0	100	240	1	City	Y	Y	Y	0.408834	0.59
P001038	Female	Not Married	0	Graduate	No	3510	0	78	360	0	City	N	N	N	0.943038	0.05
P001039	Male	Married	0	Under Graduate	No	4887	0	133	360	1	Village	N	N	Y	0.272534	0.72
P001041	Male	Married	0	Graduate	No	2850	3900	115	342	1	City	Y	Y	Y	0.188274	0.81
P001043	Male	Married	0	Under Graduate	No	7850	0	104	360	0	City	N	N	N	0.854877	0.08
P001048	Male	Married	1	Graduate	No	5955	5525	315	360	1	City	Y	Y	Y	0.374903	0.62
P001047	Male	Married	0	Under Graduate	No	2800	1811	115	360	0	Town	N	N	N	0.895239	0.10
P001050	Male	Married	2	Under Graduate	No	3365	1917	112	360	0	Village	N	N	N	0.839511	0.08
P001052	Male	Married	1	Graduate	No	3717	2628	151	360	1	Town	N	N	Y	0.162399	0.83
P001058	Male	Married	0	Graduate	Yes	6560	0	191	360	1	Town	Y	Y	Y	0.260355	0.69
P001058	Male	Married	0	Graduate	No	2799	2253	122	360	1	Town	Y	Y	Y	0.101908	0.89
P001073	Male	Married	2	Under Graduate	No	4225	1040	110	360	1	City	Y	Y	Y	0.155329	0.81
P001086	Male	Not Married	0	Under Graduate	No	1442	0	35	360	1	City	N	N	Y	0.315455	0.68
P001087	Female	Not Married	2	Graduate	No	3750	2563	120	360	1	Town	Y	Y	Y	0.125511	0.87
P001091	Male	Married	1	Graduate	No	4155	3398	201	360	1	City	N	N	Y	0.303773	0.69
P001095	Male	Not Married	0	Graduate	No	3157	0	74	360	1	City	N	N	Y	0.244952	0.75
P001097	Male	Not Married	1	Graduate	Yes	4952	0	105	360	1	Village	N	N	Y	0.405777	0.59

Figure 143

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.

PART 3

10.0 Chapter 10 (Data Visualization and Report Generation)

10.1.1 Data visualization

10.1.2 Introduction

Data visualization is one of the important parts of any statistical data science project where it is important to tell the stories about the dataset and visualize them in charts, bars, and graphs. It is one of the most important skills to have as a data scientist to be able to do data storytelling and SAS offers many kinds of techniques and tools for data visualization. Some of them are using SAS Graphic Procedures where it provides various functions such as PROC SGPLOT, PROC SGPANEL, PROC GCHART that allow the data scientists to create scatter plots, charts, bar graphs and histograms. Another one is ODS graphics where it can create graphics in formats such PDF, HTML and SAS Visual Analytics where it provides web-based tools to visualize data and also provide platform for advanced analytics and predictive modelling.

10.1.3 SAS Source Codes

```
1553  
1554 /* Data Visualization */  
1555  
1556 /* SAS Simple Bar Chart */  
1557  
1558 /* Loan Location*/  
1559 TITLE 'Loan Applicants VS Loan Location';  
1560 PROC SGPLOT DATA = DAP67696.TESTING_PREDICTED_DS;  
1561 VBAR loan_location;  
1562 Label loan_location = 'Loan Location-----';  
1563 RUN;  
1564 QUIT;
```

Figure 144

10.1.4 Screenshot(s) of the Output

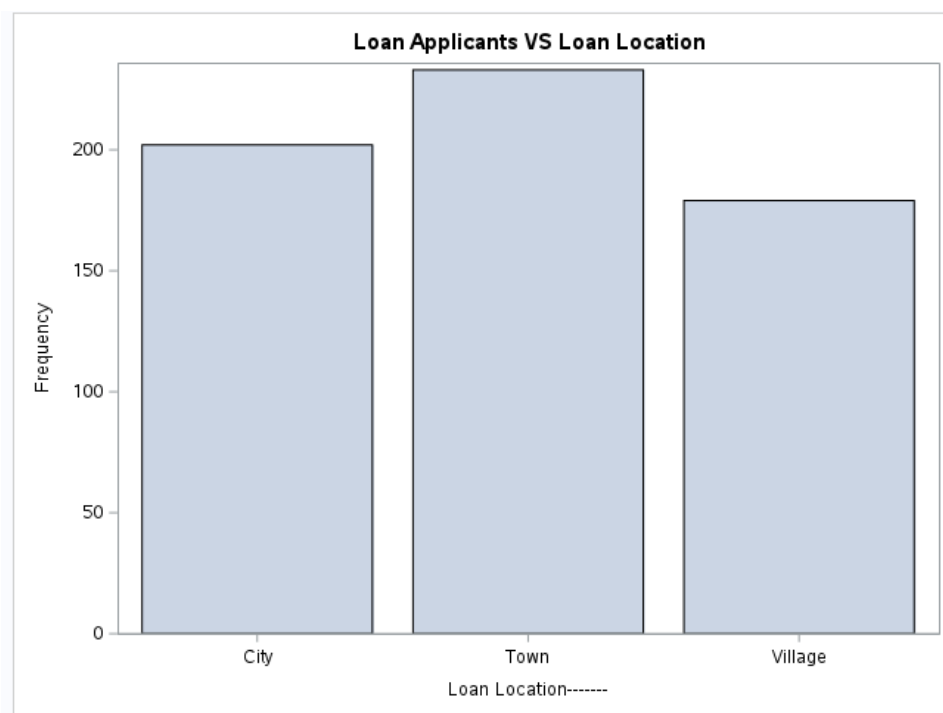


Figure 145

10.1.5 Description

Figure above shows the simple bar chart to visualize the relationship between the loan application and the loan location. It is found that from the figure above the variables town has the highest number of loan applications followed by city and the least number of applicants is village.

10.1.6 SAS Source Codes

```
1566 /* Marital Status*/  
1567 TITLE 'Loan Applicants VS Marital Status';  
1568 PROC SGPLOT DATA = DAP67696.TESTING_PREDICTED_DS;  
1569 VBAR marital_status;  
1570 Label marital_status = 'Marital Status-----';  
1571 RUN;  
1572 QUIT;  
1573
```

Figure 146

10.1.7 Screenshot(s) of the Output

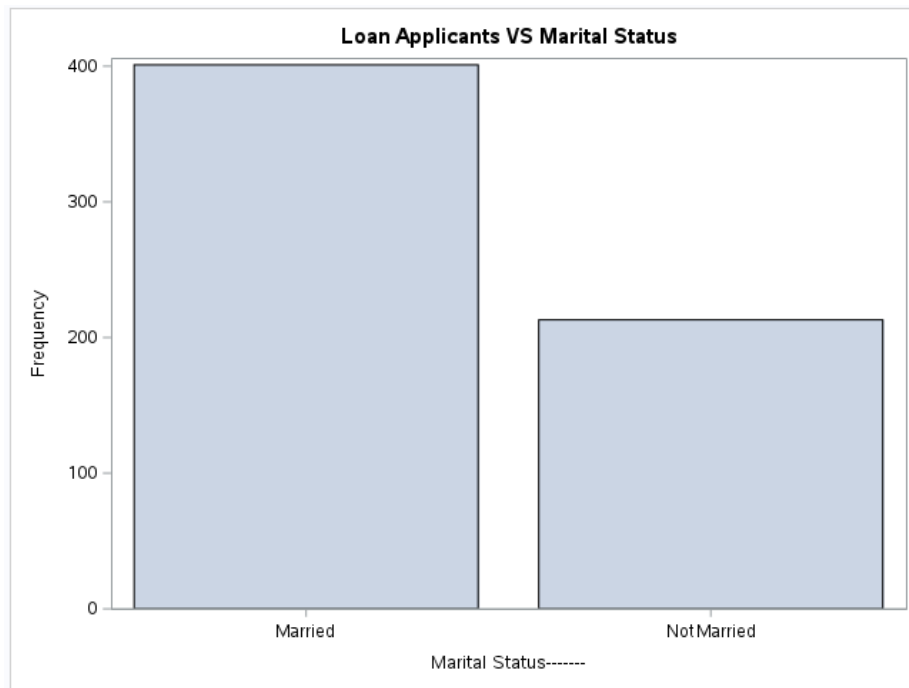


Figure 147

10.1.8 Description

Figure above shows the simple bar chart that describes the relationship between the variables of marital status against the loan applicants. For the marital status it consisted of married and not married where from the information above the married loan applicant is higher than not married applicants.

10.1.9 SAS Source Codes

```
1577 Stacked Bar Chart
1578 The groups were stacked one above the other
1579 *****/
1580
1581 /* Family Members */
1582 TITLE 'Number of family members by loan location';
1583 PROC SGPLOT data = DAP67696.TESTING_PREDICTED_DS;
1584 vbar family_members /group = loan_location groupdisplay = cluster;
1585 label family_members = 'Number of family members';
1586
1587 RUN;
1588 QUIT;
1589
```

Figure 148

10.2 Screenshot(s) of the Output

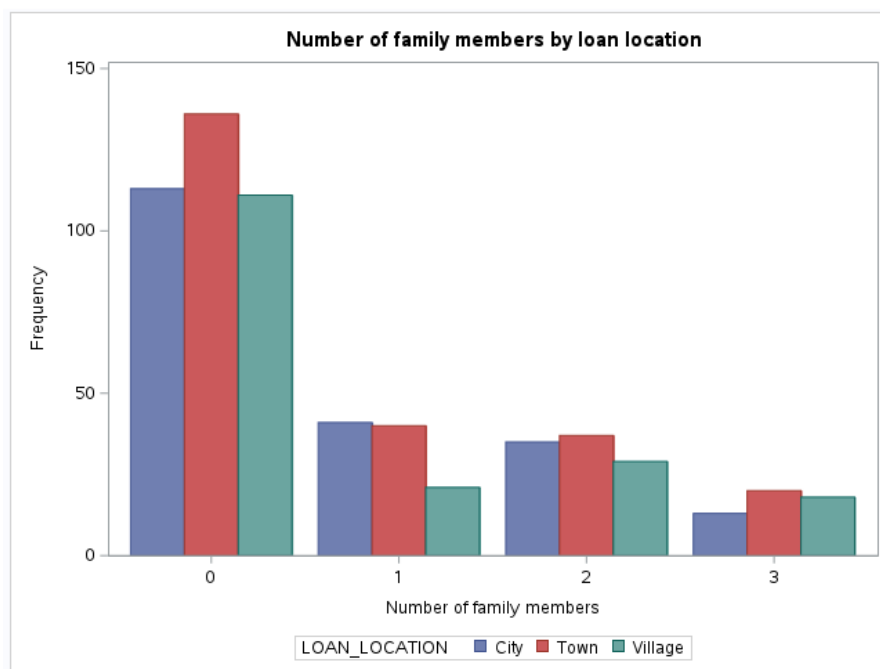


Figure 149

10.2.1 Description

Figure above shows the stacked bar chart of the relationship between the number of family members by loan location where for this data visualization, the groups were stacked side by side. Hence as shown above, the loan location has 3 categories which are city, town, and village which indicated by the color blue, red and green. The number of family members are categorized into 0

family members, 1,2 and 3 whereby loan applicants 0 family members have the highest number of applicants compared to 1,2 and 3. Moreover, by looking at figure above in the applicant with 0 family members the loan location town is the highest compared to city and village. For applicant with number of family members 1 has loan location city the highest and for applicant with number of family members 2 has loan location town the highest and for applicant with number of family members 3 has loan location town the highest as well.

10.2.2 SAS Source Codes

```

1590 /* Genders */
1591 TITLE 'Number of family members by Gender';
1592 PROC SGPLOT data = DAP67696.TESTING_PREDICTED_DS;
1593 vbar family_members /group = gender groupdisplay = cluster;
1594 Label family_members = 'Number of family members';
1595
1596 RUN;
1597 QUIT;
1598

```

Figure 150

10.2.3 Screenshot(s) of the Output

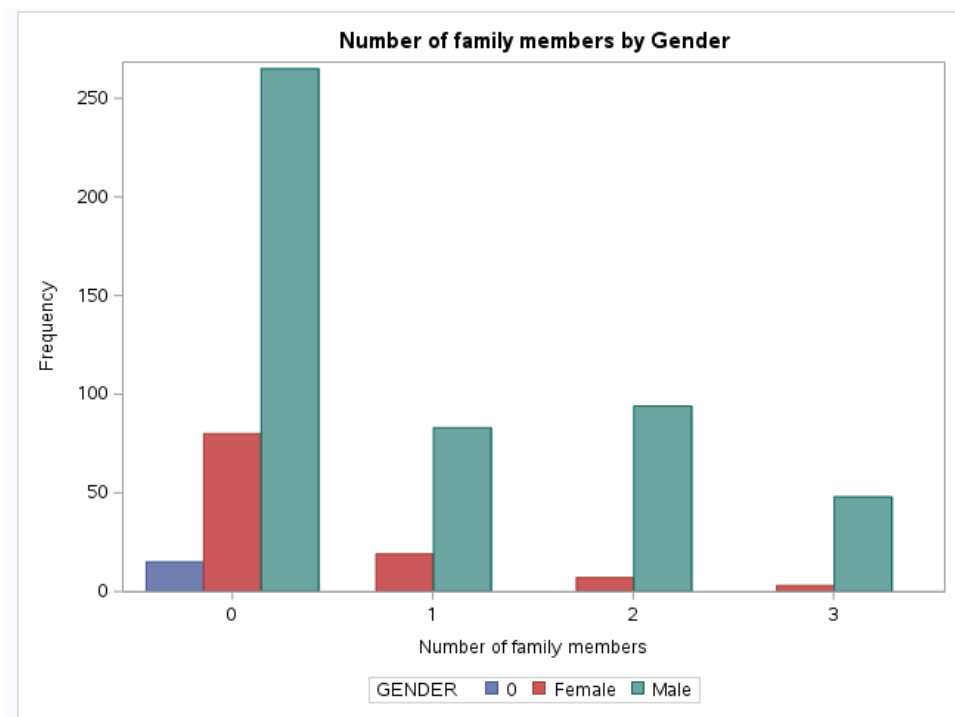


Figure 151

10.2.4 Description

Figure above shows the stacked bar chart that shows the relationship between number of family members and gender, and it is known that the gender variable has 2 categories which are the female and male. From the information above the gender female is represented by red color meanwhile for male it is green. The number of family members are categorized into 0 number of family members, 1,2, and 3 where loan applicant with 0 family members is the highest compared to number of family members 1,2, and 3. From the information above, in general male applicant are the highest to apply for loan compared to female.

10.2.5 SAS Source Codes

```
1599 /*****
1600
1601 Pie Chart
1602 A pie-chart is a representation of values as slices of a circle with different colours
1603
1604 *****/
1605
1606 TITLE 'Loan approval status by loan location';
1607
1608 PROC GCHART data = DAP67696.TESTING_PREDICTED_DS;
1609 pie3d I_LOAN_APPROVAL_STATUS;
1610 RUN;
1611 QUIT;
```

Figure 152

10.2.6 Screenshot(s) of the Output

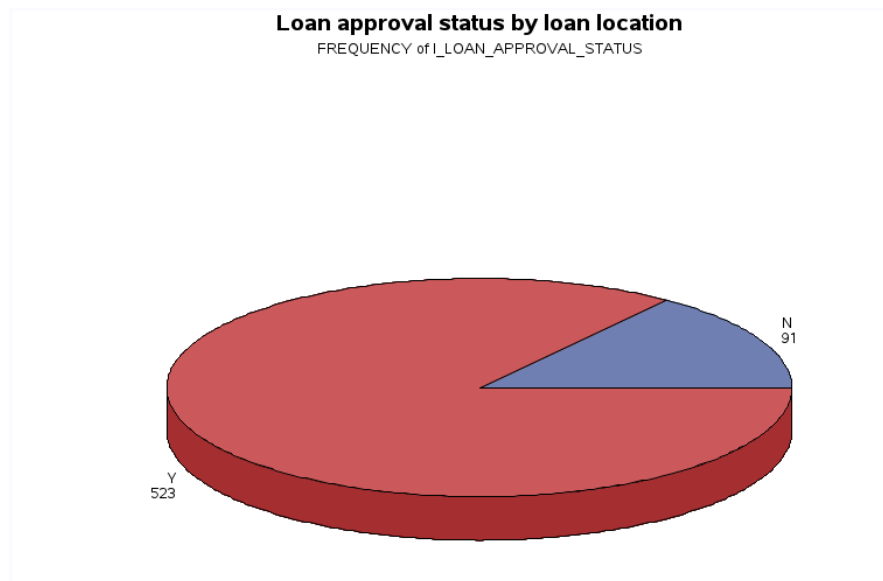


Figure 153

10.2.7 Description

Figure above shows the data visualization using pie chart approach where a pie chart is a representation of values as slices of circle with different colors. The output for this SAS code shows the pie chart above where the title is about loan approval status by location. It is known that Y indicates the number of loan applicants with loan approval status accepted meanwhile for N indicates number of loan applicants with loan approval status not accepted. From pie chart above, loan approval status accepted is 523 applicants which is higher than loan approval status not accepted at 91.

10.2.8 SAS Source Codes

```
1613  
1614 /* 3D Visualization */  
1615 TITLE 'Loan approval status by loan location';  
1616 goptions cback=black;  
1617 pattern1 c=red;  
1618 pattern1 c=green; pattern1 c=green;  
1619 PROC GCHART DATA = DAP67696.TESTING_PREDICTED_DS;  
1620 pie3d I_LOAN_APPROVAL_STATUS/woutline=2 coutline=white  
1621 ctext=white explore='M6' group=loan_location;  
1622 RUN;  
1623 QUIT;  
1624
```

Figure 154

10.2.9 Screenshot(s) of the Output

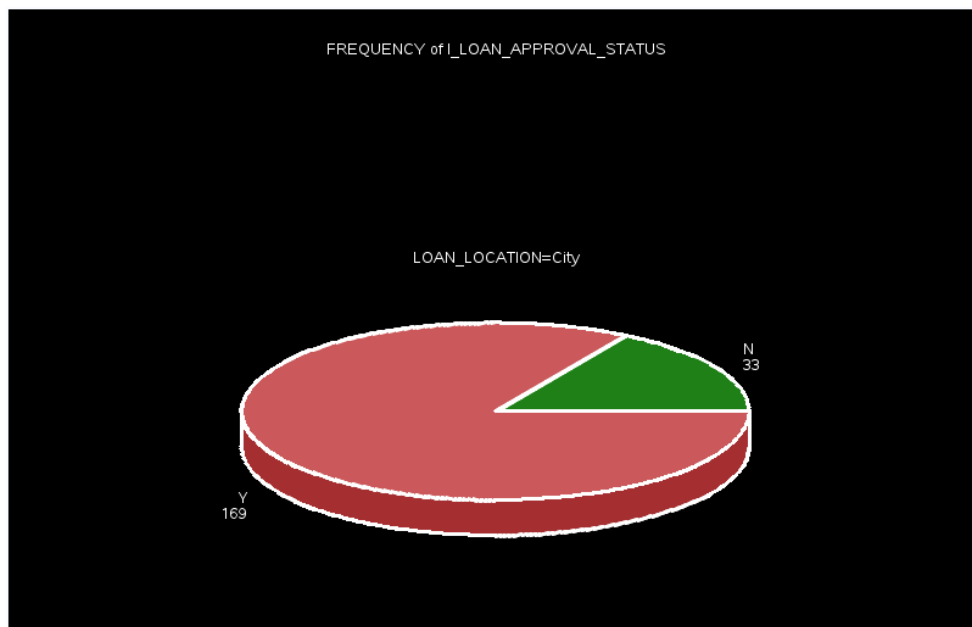


Figure 155

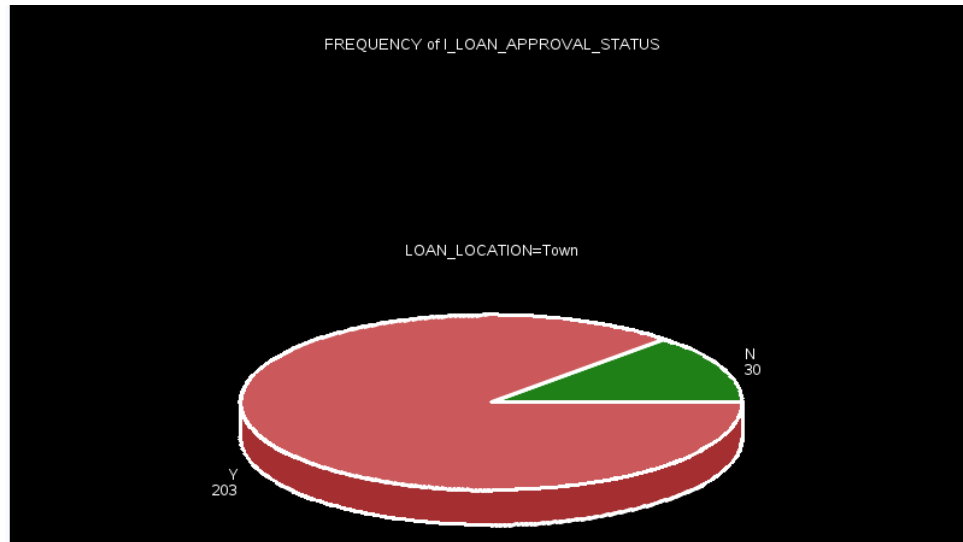


Figure 156

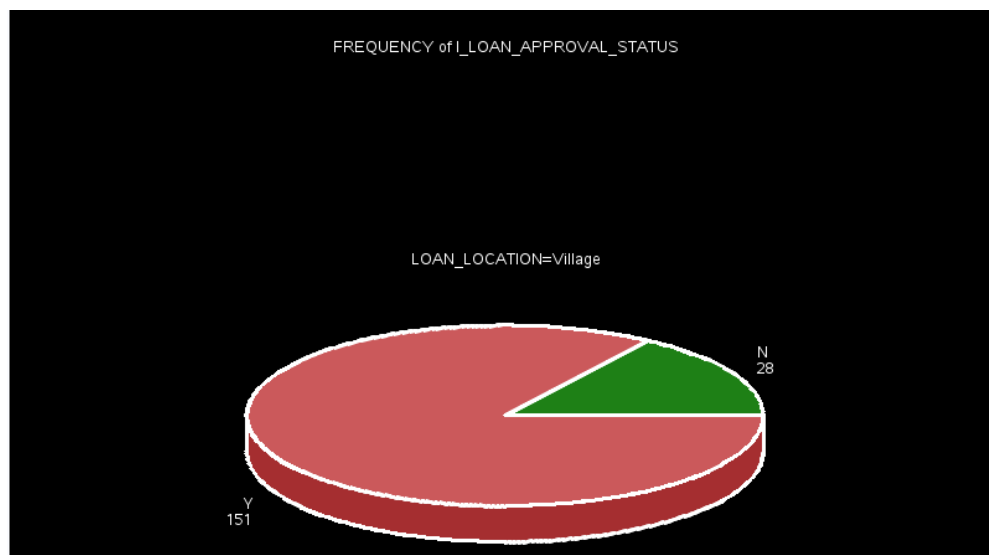


Figure 157

10.3 Description

The figure above shows an advanced version of data visualization of pie chart to show the relationship between loan approval status by loan location where the loan locations are divided into city, town, and village. For the loan location city, the number of loan applicants with loan approval status accepted is 169 compared to not accepted at 33. Meanwhile for loan location town, the number of loan applicants with loan approval status accepted is 203 compared to not accepted at 30 and for loan location village, the number of loan applicants with loan approval status accepted

is 151 compared to not accepted at 28. It can be deduced that loan approval status accepted is higher than not accepted and most of the accepted loan approval status lives in town.

10.3.1 SAS Source Codes

```

1627 /* Advanced pie chart */
1628 GOPTIONS RESET=ALL BORDER;
1629 TITLE "Family members vs Loan Location";
1630 PROC GCHART DATA=DAP67696.TESTING_PREDICTED_DS;
1631 pie family_members / detail=loan_location
1632 detail_percent=best
1633 detail_value=none
1634 detail_slice=best
1635 detail_threshold=2
1636 legend;
1637 RUN;
1638 QUIT;

```

Figure 158

10.3.2 Screenshot(s) of the Output

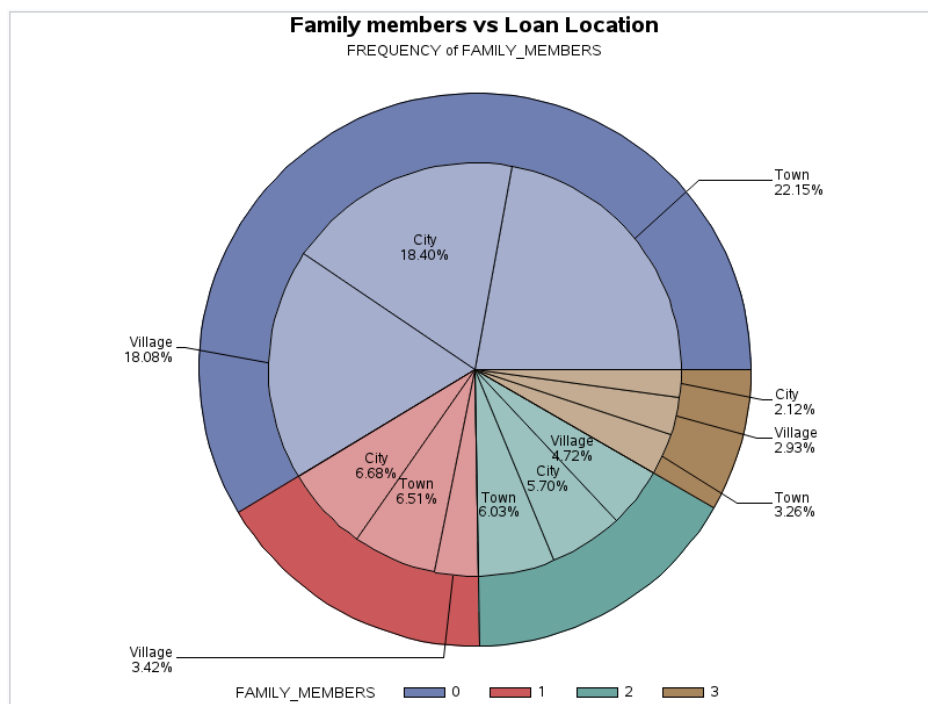


Figure 159

10.3.2 Description

The figure above shows an advanced pie chart that shows the relationship between family members vs loan location where it shows much detail explanation of each compositions in percentage of the loan location for town, city, and village. At the same time, it visualizes the number of family members of 0,1,2,and 3 with the color of blue, red, green, and brown representing each of the number of family members. From the pie chart above most of the loan applications have 0 family members which comprised at 58.68% of the applicants, meanwhile for 1 family members at 16.58%, 2 family members at 16.45% and 3 family members at 8.31%.

10.3.3 SAS Source Codes

```
1641 /*****  
1642 Scatter plot  
1643 Scatterplot is a type of graph which uses values from two variables.  
1644 It is usually used to find out the relationship between two variables.  
1645 *****/  
1646  
1647 PROC SGPLOT DATA = DAP67696.TESTING_PREDICTED_DS;  
1648 scatter x=candidate_income y=loan_amount / group=gender;  
1649 RUN;  
1650 QUIT;
```

Figure 160

10.3.4 Screenshot(s) of the Output

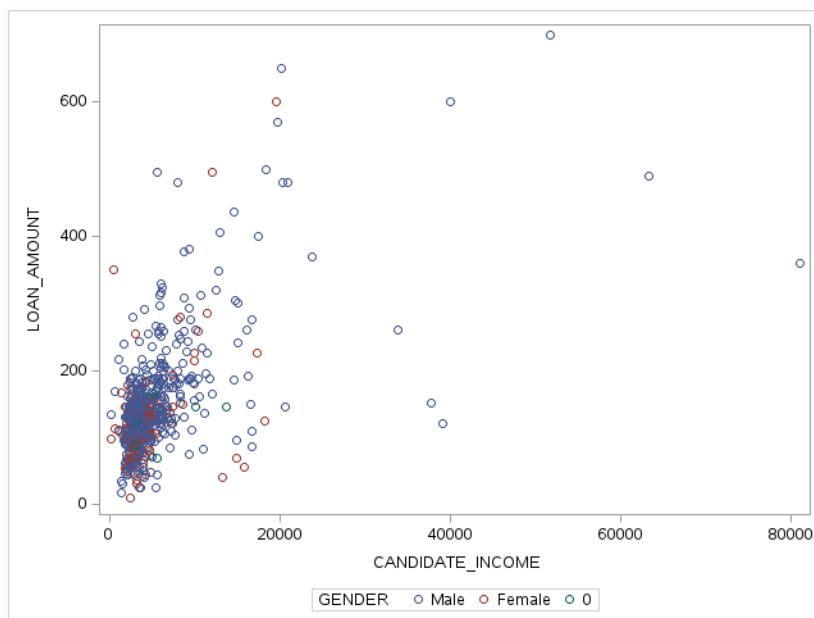


Figure 161

10.3.5 Description

Figure above shows the scatterplot to find out relationship between candidate income and loan amount by grouping both of it using gender where for this data visualization, it is found that most of the candidates income or salary is lower than 20000 for both male and female. Also, most of the loan amount approved by the bank is lower than 200 but another trend can be seen is that there are few outliers in the data points where male has the highest candidate income and loan amount.

10.3.6 SAS Source Codes

```
1652 /*****  
1653 Box Plot  
1654 *****/  
1655 PROC SGPLOT data = DAP67696.TESTING_PREDICTED_DS;  
1656 vbox candidate_income / category = gender;  
1657 RUN;  
1658 QUIT;  
1659
```

Figure 162

10.3.7 Screenshot(s) of the Output

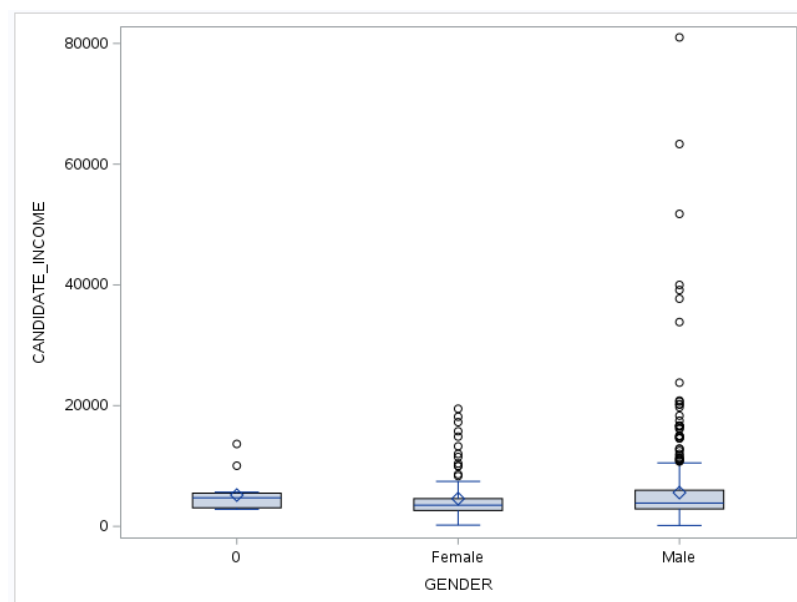


Figure 163

10.3.8 Description

Figure above shows another data visualization that shows the box plot of candidate income versus gender whereby looking at the box plot for female all of the candidate income is under 20000. Meanwhile for male candidate income majority of it is under 20000 but there are few outliers where the income is at 40000, 60000, and 80000.

10.3.9 Report Generation

Report generation in SAS

10.4 Physical location of SAS library

10.4.1 SAS Source Codes

```
1014  
1015 /* List the datasets found in the SAS Library */  
1016  
1017 PROC DATASETS LIBRARY = DAP67696 memtype = DATA;  
1018 RUN;
```

Figure 164

Location of the file

10.4.2 Screenshot(s) of the Output

Directory	
Libref	DAP67696
Engine	V9
Physical Name	/home/u61522473/DAP_PT_AUG_2023_TP067696
Filename	/home/u61522473/DAP_PT_AUG_2023_TP067696
Inode Number	7118327079
Access Permission	rwxf-xf-x
Owner Name	u61522473
File Size	4KB
File Size (bytes)	4096

Figure 165

#	Name	Member Type	File Size	Last Modified
1	QUALIFICATION_FM_STAT_DS	DATA	256KB	01/10/2023 07:35:07
2	QUALIFICATION_Q_STAT_DS	DATA	256KB	01/10/2023 07:42:28
3	TESTING_DS	DATA	256KB	23/08/2023 13:07:46
4	TESTING_PREDICTED_DS	DATA	256KB	03/10/2023 06:19:34
5	TRAINING_BK_DS	DATA	256KB	20/09/2023 16:20:28
6	TRAINING_DS	DATA	256KB	02/10/2023 04:11:28
7	TRAINING_EMPLOYMENT_STAT_DS	DATA	256KB	01/10/2023 08:50:34
8	TRAINING_FM_STAT_DS	DATA	256KB	20/09/2023 06:37:58
9	TRAINING_GENDER_STAT_DS	DATA	256KB	19/09/2023 16:28:59
10	TRAINING_LOAN_APPROVAL_STAT_DS	DATA	256KB	01/10/2023 13:44:55
11	TRAINING_LOAN_HISTORY_STAT_DS	DATA	256KB	01/10/2023 09:56:49
12	TRAINING_LOAN_LOCATION_STAT_DS	DATA	256KB	01/10/2023 13:09:32
13	TRAINING_LR_MODEL	DATA	256KB	03/10/2023 05:58:23
14	TRAINING_MS_STAT_DS	DATA	256KB	19/09/2023 17:11:10
15	TRAINING_OUT_DS	DATA	256KB	03/10/2023 05:58:23
16	TRAINING_QUALIFICATION_STAT_DS	DATA	256KB	01/10/2023 08:23:01

Figure 166

10.4.2 Description

The SAS code used is PROC function on the TESTING_DS whereby it is used to locate the file location in the SAS library. The physical file location is indicated by /home/u61522473/DAP_PT_AUG_2023_TP067696 which is a cloud based and Libref shows the name of the library which indicated by DAP67696. The other figure shows the other intermediate and temporary dataset created during the assignment where to know it is data is by looking at the Member Type and the file size can be also known and last modified shown when the last time the specific dataset is modified in the SAS studio.

10.4.3 Introduction to ODS

This part here is to discuss the report generation using SAS ODS – Output Delivery System where ODS where in the field of data processing, it refers to software component used in SAS software. ODS or Output Delivery System provide advanced control and the flexibility to generate and manage output from SAS procedures and program.

Using ODS the data scientist is allowed to customize the type, format, and appearance of the SAS output, for example create graphs, chart, tables and generate reports and deliver the output in multiple and various formats. These formats include the PDF,HTML and RTF(Rich Text Format) and main features of SAS ODS are output customization, various output formats, selective output, data manipulation and destination control.

For output customization the features allow the data scientist to modified the layout, style, and the formatting of the output to meet the data scientist preferences. On the other hand, for the various output format shows that ODS allow same analysis done on the report to be delivered in various format at same time. Thus, this allows easy access to information and makes it very easy to share information with other users or data scientists.

For selective output, ODS allow the data scientist or users to choose which parts of the output to be included and excluded meanwhile data manipulation allow the data scientist to manipulate the output results of the SAS programming. Destination control allows the data scientist to specify the output directory whether it's in a file, email, or printer format.

10.4.4 SAS Source Codes

```
1022
1023 /*****
1024 Generate report using SAS ODS - Output Delivery System
1025 *****/
1026
1027 ODS HTML CLOSE;
1028 ODS PDF CLOSE;
1029
1030 /* Determine the physical location of pdf */
1031
1032 ODS PDF FILE = "/home/u61522473/DAP_PT_AUG_2023_TP067696/LFI_LAS.pdf";
1033 OPTIONS NODATE;
1034 TITLE1 'Lasiandara Finance Loan Approval Status Predicted';
1035 TITLE2 'LFS,TPM';
1036
1037 PROC REPORT DATA = DAP67696.TESTING_PREDICTED_DS NOWINDOWS;
1038
1039 BY SME_LOAN_ID_NO; /* To separate each by SME_LOAN_ID_NO */
1040
1041 /* COLUMN SME_LOAN_ID_NO I_LOAN_APPROVAL_STATUS;*/
1042
1043 DEFINE SME_LOAN_ID_NO / GROUP 'LOAN ID';
1044 DEFINE GENDER / GROUP 'GENDER NAME';
1045 DEFINE MARITAL_STATUS / GROUP 'MARITAL STATUS';
1046 DEFINE FAMILY_MEMBERS / GROUP 'FAMILY MEMBERS';
1047 DEFINE CANDIDATE_INCOME / GROUP 'MONTHLY INCOME';
1048 DEFINE GUARANTEE_INCOME / GROUP 'CO-APPLICANT INCOME';
1049 DEFINE LOAN_AMOUNT / GROUP 'LOAN AMOUNT';
1050 DEFINE LOAN_DURATION / GROUP 'LOAN DURATION';
1051 DEFINE LOAN_HISTORY / GROUP 'LOAN HISTORY';
1052 DEFINE LOAN_LOCATION / GROUP 'LOAN LOCATION';
1053
1054 FOOTNOTE '-----End of Report-----';
1055 RUN;
1056
```

Figure 167

10.4.5 Screenshot

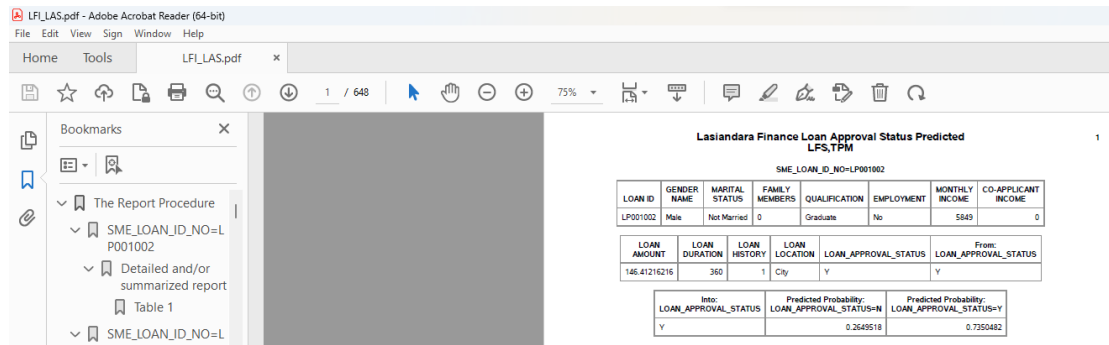
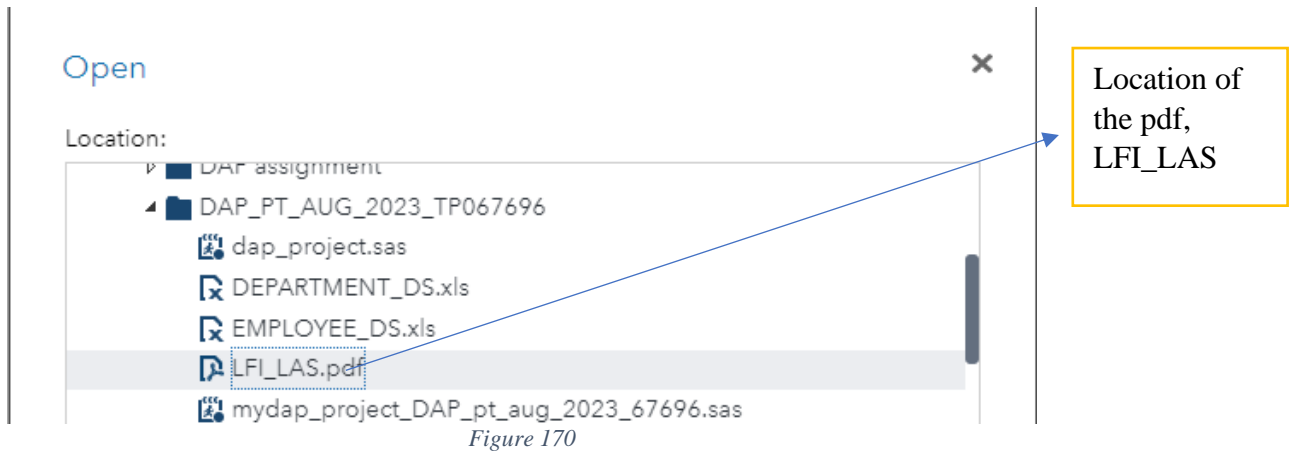
Table of Contents

Lasiandara Finance Loan Approval Status Predicted LFS,TPM																
LOAN ID	GENDER NAME	MARITAL STATUS	FAMILY MEMBERS	QUALIFICATION	EMPLOYMENT	MONTHLY INCOME	CO-APPLICANT INCOME	LOAN AMOUNT	LOAN DURATION	SME_LOAN_ID_NH_LP01002			From: LOAN_APPROVAL_STATUS	Info: LOAN_APPROVAL_STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
										LOAN HISTORY	LOAN LOCATION	LOAN_APPROVAL_STATUS				
LPD01002	Male	Not Married	0	Graduate	No	5540	0	145.412.15210	350	1	City	Y	Y	Y	0.2546518	0.7350482
-----End of Report-----																

Figure 168

Laslandara Finance Loan Approval Status Predicted LFS,TPM																
SME_LOAN_ID=NHLP00109																
LOAN ID	GENDER NAME	MARITAL STATUS	FAMILY MEMBERS	QUALIFICATION	EMPLOYMENT	MONTHLY INCOME	CO-APPLICANT INCOME	LOAN AMOUNT	LOAN DURATION	LOAN HISTORY	LOAN LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	Info: LOAN_APPROVAL_STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
LPD01100	Male	Married	0	Graduate	No	1520	1330	100	342	0	City	N	N	N	0.9135817	0.0864183
-----End of Report-----																

Figure 169



10.4.6 Description

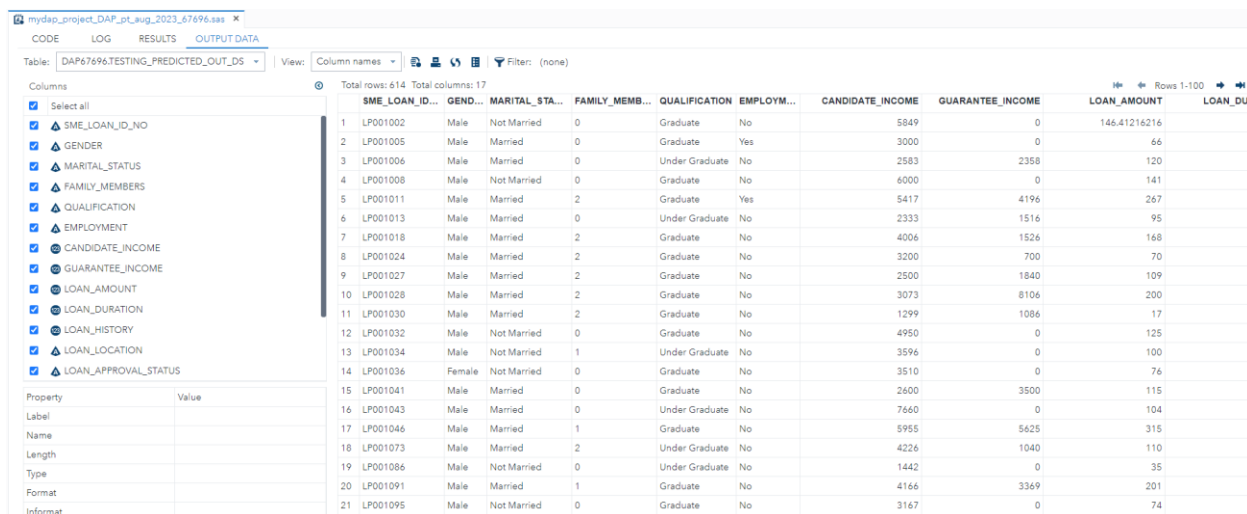
Figure above shows the SAS code and the output to generate report using SAS ODS – Output Delivery System where it omits the output Lasiandara Finance Loan Approval Status Predicted. Figure above shows that result of the output Lasiandara Finance have both train and test dataset cleaned which can be indicated no table generated is empty. In the SAS code above both categorical and continuous/numeric variables are used and for the SAS pdf report function of ODS HTML CLOSE; and ODS PDF CLOSE is used. The HTML CLOSE; and ODS PDF CLOSE used is to generate the output of the PDF where the location of the pdf can be seen in figure above and saved as LFI_LAS.pdf.

10.4.7 SAS Source Codes

```
1534 /*****
1535 Generate report carrying the loan approval status (without using SAS ODS)
1536 *****/
1537
1538 OPTIONS NODATE;
1539 PROC SORT DATA = DAP67696.TESTING_PREDICTED_DS OUT = DAP67696.TESTING_PREDICTED_OUT_DS;
1540
1541 BY loan_location
1542     sme_loan_id_no;
1543 RUN;
```

Figure 172

10.4.8 Screenshot(s) of the Output



The screenshot shows the SAS Output Data window for the table DAP67696.TESTING_PREDICTED_OUT_DS. The table contains 21 rows of data with 17 columns. The columns are: SME_LOAN_ID..., GEND..., MARITAL_STA..., FAMILY_MEMB..., QUALIFICATION, EMPLOYM..., CANDIDATE_INCOME, GUARANTEE_INCOME, LOAN_AMOUNT, and LOAN_DU. The data is sorted by loan_location and sme_loan_id_no.

SME_LOAN_ID...	GEND...	MARITAL_STA...	FAMILY_MEMB...	QUALIFICATION	EMPLOYM...	CANDIDATE_INCOME	GUARANTEE_INCOME	LOAN_AMOUNT	LOAN_DU
1 LP001002	Male	Not Married	0	Graduate	No	5849	0	146.41216216	
2 LP001005	Male	Married	0	Graduate	Yes	3000	0	66	
3 LP001006	Male	Married	0	Under Graduate	No	2583	2358	120	
4 LP001008	Male	Not Married	0	Graduate	No	6000	0	141	
5 LP001011	Male	Married	2	Graduate	Yes	5417	4196	267	
6 LP001013	Male	Married	0	Under Graduate	No	2333	1516	95	
7 LP001018	Male	Married	2	Graduate	No	4006	1526	168	
8 LP001024	Male	Married	2	Graduate	No	3200	700	70	
9 LP001027	Male	Married	2	Graduate	No	2500	1840	109	
10 LP001028	Male	Married	2	Graduate	No	3073	8106	200	
11 LP001030	Male	Married	2	Graduate	No	1299	1086	17	
12 LP001032	Male	Not Married	0	Graduate	No	4950	0	125	
13 LP001034	Male	Not Married	1	Under Graduate	No	3596	0	100	
14 LP001036	Female	Not Married	0	Graduate	No	3510	0	76	
15 LP001041	Male	Married	0	Graduate	No	2600	3500	115	
16 LP001043	Male	Married	0	Under Graduate	No	7660	0	104	
17 LP001046	Male	Married	1	Graduate	No	5955	5625	315	
18 LP001073	Male	Married	2	Under Graduate	No	4226	1040	110	
19 LP001086	Male	Not Married	0	Under Graduate	No	1442	0	35	
20 LP001091	Male	Married	1	Graduate	No	4166	3369	201	
21 LP001095	Male	Not Married	0	Graduate	No	3167	0	74	

Figure 173

10.4.9 Description

Figure above shows the SAS code to generate report carrying the loan approval status without using SAS ODS where the data used is DAP67696.TESTING_PREDICTED_DS where it is different from the previous one that used SAS ODS.

10.5 SAS Source Codes

```
1547 /*Generate the report */
1548
1549 PROC PRINT DATA = DAP67696.TESTING_PREDICTED_OUT_DS SPLIT = '*';
1550
1551 id loan_location;
1552 by loan_location;
1553 var sme_loan_id_no
1554     candidate_income
1555     loan_amount
1556     loan_duration
1557     i_loan_approval_status;
1558 sum candidate_income loan_amount;
1559
1560 label loan_location = 'LOAN LOCATION*====='
1561     sme_loan_id_no = 'LOAN ID*====='
1562     candidate_income = 'CANDIDATE INCOME*===== '
1563     loan_amount = 'LOAN AMOUNT*====='
1564     loan_duration = 'LOAN DURATION*====='
1565     i_loan_approval_status = 'LOAN APPROVAL STATUS*=====';
1566
1567 TITLE1 'Lasiandara Finance Loan Approval Status Predicted';
1568 TITLE2 'LFS,TPM';
1569
1570 RUN;
```

Figure 174

10.5.1 Screenshot(s) of the Output

Lasiandara Finance Loan Approval Status Predicted LFS,TPM					
LOAN LOCATION =====	LOAN ID =====	CANDIDATE INCOME =====	LOAN AMOUNT =====	LOAN DURATION =====	LOAN APPROVAL STATUS =====
City	LP001002	5849	146.41216216	360	Y
	LP001005	3000	66	360	Y
	LP001006	2583	120	360	Y
	LP001008	6000	141	360	Y
	LP001011	5417	267	360	Y
	LP001013	2333	95	360	Y
	LP001018	4006	168	360	Y
	LP001024	3200	70	360	Y
	LP001027	2500	109	360	Y
	LP001028	3073	200	360	Y
	LP001030	1299	17	120	Y
	LP001032	4950	125	360	Y
	LP001034	3596	100	240	Y
	LP001036	3510	76	360	N
	LP001041	2600	115	342	Y
	LP001043	7660	104	360	N
	LP001046	5955	315	360	Y
	LP001073	4226	110	360	Y
	LP001086	1442	35	360	Y
	LP001091	4166	201	360	Y
	LP001095	3167	74	360	Y
	LP001106	2275	146.41216216	360	Y
	LP001109	1828	100	342	N
	LP001114	4166	184	360	Y
	LP001119	3600	80	360	Y
	LP001120	1800	47	360	Y
	LP001123	2400	75	360	Y
	LP001136	4695	96	342	Y
	LP001137	3410	88	342	Y
	LP001138	5649	44	360	Y
	LP001144	5821	144	360	Y
	LP001146	2645	120	360	N

Figure 175

	LP002842	3417	188	360	Y
	LP002847	5116	165	360	N
	LP002855	16666	275	360	Y
	LP002868	3159	108	84	Y
	LP002874	3229	110	360	Y
	LP002888	3182	161	360	Y
	LP002893	1836	90	360	N
	LP002916	2297	104	360	Y
	LP002938	16120	260	360	Y
	LP002948	5780	192	360	Y
	LP002949	416	350	180	N
	LP002953	5703	128	360	Y
	LP002960	2400	146.41216216	180	Y
	LP002983	8072	253	360	Y
	LP002984	7583	187	360	Y
City		1090446	28770.533784		

Figure 176

LOAN LOCATION =====	LOAN ID =====	CANDIDATE INCOME =====	LOAN AMOUNT =====	LOAN DURATION =====	LOAN APPROVAL STATUS =====
Town	LP001014	3038	158	360	N
	LP001020	12841	349	360	Y
	LP001047	2600	116	360	N
	LP001052	3717	151	360	Y
	LP001066	9560	191	360	Y
	LP001068	2799	122	360	Y
	LP001087	3750	120	360	Y
	LP001098	3500	114	360	Y
	LP001112	3667	144	360	Y
	LP001116	3748	110	360	Y
	LP001131	3941	134	360	Y
	LP001151	4000	144	360	Y
	LP001155	1928	100	360	Y
	LP001157	3086	120	360	Y
	LP001164	4230	112	360	Y
	LP001194	2708	97	360	Y
	LP001195	2132	96	360	Y
	LP001222	4166	116	360	N
	LP001225	5726	258	360	Y
	LP001241	4300	136	360	N
	LP001245	1875	97	360	Y
	LP001248	3500	81	300	Y
	LP001250	4755	95	342	N
	LP001253	5266	187	360	Y
	LP001263	3167	180	300	N
	LP001264	3333	130	360	Y
	LP001265	3846	111	360	Y
	LP001266	2395	146.41216216	360	Y
	LP001273	6000	265	360	Y
	LP001279	2366	136	360	Y
	LP001280	3333	99	360	Y
	LP001282	2500	104	360	Y
	LP001310	5695	175	360	Y

Figure 177

	LP002872	3087	136	360	N
	LP002892	6540	205	360	Y
	LP002894	3166	36	360	Y
	LP002917	2165	70	360	Y
	LP002925	4750	94	360	Y
	LP002926	2726	106	360	N
	LP002928	3000	56	180	Y
	LP002931	6000	205	240	Y
	LP002933	9357	292	360	Y
	LP002943	2987	88	360	N
	LP002959	12000	496	360	Y
	LP002961	3400	173	360	Y
	LP002990	4583	133	360	N
Town		1233097	33907.060811		

Figure 178

LOAN LOCATION =====	LOAN ID =====	CANDIDATE INCOME =====	LOAN AMOUNT =====	LOAN DURATION =====	LOAN APPROVAL STATUS =====
Village	LP001003	4583	128	360	Y
	LP001029	1853	114	360	Y
	LP001038	4887	133	360	Y
	LP001050	3385	112	360	N
	LP001097	4692	106	360	Y
	LP001100	12500	320	360	Y
	LP001197	3366	135	360	Y
	LP001207	2609	165	180	N
	LP001213	4945	146.41216216	360	N
	LP001370	7333	120	360	Y
	LP001391	3572	152	342	N
	LP001401	14583	185	180	Y
	LP001421	5588	175	360	Y
	LP001426	5667	180	360	Y
	LP001439	4300	194	360	Y
	LP001443	3692	93	360	Y
	LP001448	23803	370	360	Y
	LP001449	3865	146.41216216	360	Y
	LP001465	6080	182	360	Y
	LP001489	4583	84	360	Y
	LP001493	4200	129	360	Y
	LP001497	5042	185	360	Y
	LP001519	10000	225	360	Y
	LP001528	6277	118	360	N
	LP001529	2577	152	360	Y
	LP001532	2281	113	360	Y
	LP001541	6000	160	360	Y
	LP001546	2980	120	360	Y
	LP001570	4167	158	360	Y
	LP001574	3707	182	342	Y
	LP001577	4583	112	360	Y
	LP001578	2439	129	360	Y
	LP001581	1820	95	360	Y

Figure 179

	LP002772	2526	145	360	Y
	LP002777	2785	110	360	Y
	LP002778	6633	146.41216216	360	N
	LP002784	2492	146.41216216	360	Y
	LP002789	3593	132	180	N
	LP002820	5923	211	360	Y
	LP002833	4467	120	360	Y
	LP002837	3400	123	360	N
	LP002877	1782	107	360	Y
	LP002898	1880	61	360	Y
	LP002911	2787	146	360	N
	LP002912	4283	172	84	Y
	LP002936	3859	142	180	Y
	LP002940	3833	110	360	Y
	LP002941	6383	187	360	Y
	LP002945	9963	180	360	Y
	LP002950	2894	155	360	Y
	LP002958	3676	172	360	Y
	LP002964	3987	157	360	Y
	LP002974	3232	108	360	Y
	LP002978	2900	71	360	Y
	LP002979	4106	40	180	Y
Village		994181	27219.472973		
		3317724	89897.067568		

Figure 180

10.5.2 Description

Figure above shows the output of SAS code to shows and visualize the dataset used is DAP67696.TESTING_PREDICTED_DS and based on the SAS code above the loan location and loan location id are used to generate the report of the variables. The variables shown in the figure above are candidate income, loan amount, loan duration, and loan approval status where i_loan_approval_status is what the machine learning algorithm read, and loan approval status is the variable name defined by the data scientist. From the table above, it can be deduced that the sum for candidate income and loan amount for loan location city is 1090446 and 28770.53, for loan location town its 1233097 and 33907.06 and for loan location village its 994181 and 279219.47. It can be deduced that the sum of candidate income for all 3-loan locations is 3317724 and for loan amount is 89897.06. For the loan approval status can be seen the outcome is either Yes or No indicates by Y and N for the loan applicants.

10.5.3 Discussion

Hence, it is known by the data scientists beforehand that the train and test dataset which is TRAINING.DS and TESTING.DS. It is known that from the SAS library for TRAINING DS has 614 total rows and 13 total number of columns. Meanwhile for TESTING DS it has 367 total rows and 13 total columns. Both of the dataset has the same variables which are the SME Loan ID, Gender, Marital Status, Family Member, Qualification, Employment, Candidate Income, Guarantee Income, Loan Amount, Loan History, LOAN Duration, Loan Location and Loan Approval Status.

As a data scientist it is important to study all of these variables as these variables influencing the decision making of the data scientist to approve the loans to the applicant. For this analysis, the dependent variable is the loan approval status meanwhile the rest of the variables are the independent variables. The loan approval status of the loan applicants will have outcome of either Yes or No depending on the independent variables.

For the model creation, logistic regression is used to predict the outcome of the loan approval status which is the response variable. It is known that the dependent variable is loan approval status is a categorical variable and the independent variable is the mixture of categorical and continuous/numeric variables. The data is cleansed when the number of observations read is equal to number of observations used, hence the model convergence status is accepted. Model convergence status is defined as convergence of the machine learning model during the training of the data, and which indicates it reach satisfactory state. The output for model convergence status used precision of 10^{-8} as it is the most common value convergence threshold in iterative algorithm optimization.

As a data scientist, it is important to know what the definition of Akaike Information Criterion (AIC) and Schwarz Criterion (SC) is where both of it are the statistical techniques used in model selection especially in linear regression machine learning model. AIC is used in model selection particularly in maximum likelihood estimation where $AIC = 2k - 2\ln(L)$, where k is the number of parameters in the model and $\ln(L)$ is the natural algorithm of the maximum likelihood of model with an input of data. On the other hand, Schwarz Criterion (SC) is the same as AIC but put more

penalty on model with a greater number of parameters. It is defined as $SC = -2 \ln(L) + k * \ln(n)$, where there is addition to previous criterion which n, the sample size. In general AIC and SC value must be lower which indicating the model is better fit which in this case the value of SC is 769.311 which is higher than AIC at 764.891

There is another criterion which is the $-2\log L$ where it is under the model fit statistics the term “ $-2\log L$ ” refers to one of the components used in model selection where L representing the likelihood of the data given in the model and $-2\log L$ indicates the measure of model goodness of fit. On the other hand, in the model fit statistics, there is intercept and covariates values of AIC and SC respectively at 589.101 and 659.821 where both of intercept and covariates play important role in calculation of the criterions to help determine the best fitting models.

Intercept usually often included in linear regression model where it shows the value of dependent variable which in this case loan approval status meanwhile all of the values of the independent variables (covariates) are equal to zero. On the other hand, the covariates are the independent variables in which it causes and effect on the dependent variables. In general, the relationship between covariates and dependent variables are quantified using regression coefficients.

For this analysis, there are 3 variables which act as the most contributing factors which are the loan location, loan history, and marital status. This is due to the value of $Pr > \text{Chisq}$ for these variables is less than 0.05 compared to other independent variables. The term “ $Pr > \text{Chisq}$ ” represents for the probability greater than chi-square, where it's a p-value that is related to chi-square statistic in statistical analysis. A chi-square test indicates that whether there is a significant relationship between two categorical variables and in general $Pr > \text{Chisq}$ is used to test and see is there any statistical significance between the variables.

If the P value is less than 0.05 in which in this case for loan location, loan history, and marital status where each of these p values are lesser than 0.05. Hence the data scientist should conclude that these variables are the contributing factors to approve the loan status whether Yes or No.

11.0 Chapter 11 (Conclusion)

It can be deduced that for this assignment, there are 3 parts which are divided into part 1, part 2, and part 3. For part 1, it comprises of chapter 1: introduction, chapter 2: problem statement, chapter 3: background of Lasiandra Finance, chapter 4: assumption, program demonstration, coding and justification, chapter 5: methodology and chapter 6: data dictionary/metadata. On the other hand, for part 2, it comprises of chapter 7: literature review, chapter 8: data analysis/data cleansing and chapter 9: model creation and prediction. The last section is part 3 where it consists of chapter 10: data visualization and report generation.

In general, this assignment has lot of knowledge that can be gained especially, understanding data management system using SQL programming in SAS studio. As a data scientist, a lot of knowledge and skills can be gained during this assignment where some of the outcomes gained are assess and study variety type and forms of datasets by reading, combined and categorizing the datasets used which is the training and testing datasets using data analytical programming method. Other outcomes from this assignment are that it taught the data scientist to produce analytical data models by creating reports and enhanced listings. It also gives new skills for the data scientist to learn about data visualization using SAS software.

A lot of knowledge had been taught by the course instructor Dr Dhason Padmakumar where for the past 2 months it has been quite a journey learning about data analytical programming. The course is very crucial to data scientists as it teaches one of the pillars of programming language that need to be mastered by data scientists in their profession. The domain of the datasets used is for banking and finance industry but after learning the skills of understanding the sql coding it can be applied to other domain of datasets in the field of engineering, healthcare, insurance and more. Dr Dhason lay out the structures of teaching the fundamentals of the sql programming techniques first before going through the assignment together which has been great way to understand about this assignment.

12.0 REFERENCES

Stiglitz, J.E. and A. Weiss (1988), “Banks as Social Accountants and Screening Devices for the Allocation of Credit”, Working Paper No. 2710, Washington: National Bureau of Economic Research.

Coase, R.H. (1937), “The Nature of the Firm”, *Economica*, **4**, 386–405.

Alchian, A. and H. Demsetz (1972), “Production, Information Costs and Economic Organisation”, *American Economic Review*, **62**, 777–795.

Williamson, O. (1981), “The Modern Corporation: Origins, Evolution, Attributes”, *Journal of Economic Literature*, **19**, 1537–1568.

Briault, C. (2000), Draft Paper, “FSA Revisited and Some Issues for European Securities Markets Regulation”, London: Financial Services Authority, December.

Beccalli, E., Casu, B., Girardone, C., 2005. Efficiency and Stock Performance in European Banking. *Journal of Business, Accounting and Finance* (forthcoming)

Fernández, P., 2002. EVA, Economic Profit and Cash Value Added do not measure shareholder value creation. University of Navarra, IESE, Research paper no. 453

Eisenbeis, R.A., Ferrier, G.D., Kwan, S.H., 1999. The Informativeness of Stochastic Frontier and Programming Frontier Efficiency Scores: Cost Efficiency and Other Measures of Bank Holding Company Performance. Federal Reserve Bank of Atlanta, Working Paper no. 99-23.

Chu, S.F., Lim, G.H., 1998. Share Performance and Profit Efficiency of Banks in an Oligopolistic Market: Evidence from Singapore. *Journal of Multinational Financial Management* **8**, 155--168.

Jacobson, T., Lindé, J., Roszbach, K., 2006. Credit risk versus capital requirements under Basel II: are SME loans and retail credit really different?. *Journal of Financial Services Research*, forthcoming

Duffie D., 2005. Credit risk modeling with affine processes. *Journal of Banking & Finance*, **29**, 11, pp. 2751-2802

- Lucas, A., Klaassen, P., 2006. Discrete versus continuous state switching models for portfolio credit risk. *Journal of Banking & Finance*, 30, 1, pp. 23-35
- Galluccio, S., Roncoroni, A., (2006). A new measure of cross-sectional risk and its empirical implications for portfolio risk management, *Journal of Banking & Finance*, (forthcoming)
- Zheng H., 2006. Interaction of credit and liquidity risks: Modelling and valuation. *Journal of Banking & Finance*, 30, 2, pp. 391-407
- Jobst, N.J., Mitra G., Zenios S.A., 2006. Integrating market and credit risk: A simulation and optimisation perspective. *Journal of Banking & Finance*, 30, 2, pp. 717-742
- Scandizzo, S., 2005. Risk Mapping and Key Risk Indicators in Operational Risk Management. *Economic-Notes*, 34, 2, pp. 231-56
- De Fontnouvelle, P., Jordan, J., Rosengren, E., 2005. Implications of Alternative Operational Risk Modeling Techniques. National Bureau of Economic Research, Inc, NBER, Working Papers: no. 11103
- Barth, M.E., Beaver, W.H., Landsman, W., 1998. Relative valuation roles of equity book value and net income as a function of financial health. *Journal of Accounting and Economics* 25, 1--34.
- Athanasoglou, P.P., Brissimis, S. N., Delis, M. (2008). Bank specific, industry specific and macroeconomic determinants of bank profitability. *Journal of international financial markets, institutions and money*, 18(2), 121-136.
- Kumar, Rajiv, et al. (2019). Prediction of loan approval using machine learning. *International Journal of Advanced Science and Technology*, 28(7), 455-460.