THINES KUMAR NADARAJA
TP063097

CT050-3-M – DATA ANALYTICAL PROGRAMMING (DAP)

INDIVIDUAL ASSIGNMENT

INTAKE : APDMF2009DSBA

LECTURER : DHASON PADMAKUMAR

ASIA PACIFIC UNIVERSITY
DATA ANALYTICAL PROGRAMMING

TABLE OF CONTENTS

CHAPTER 1: Introduction

Applying for loans has proven to be a tedious process for various Small and Medium Enterprises (SMEs). SMEs require funding to improve financial performance of their organizations, as well as aiding certain areas in a company that requires improvement. In this current digital age, research, and development of products in an organization is pivotal, but costly. A company's financial report is always cross-checked against the current and non-current assets, as well as liabilities that the company incurred on an annual basis. Funding for an organization usually comes from its' stakeholders, in the form of equity. Equity, however, must be cross balanced against the company's assets and liabilities. An SME that acquires funding to be used for revenue generation, or product innovation, will certainly be able to grow and expand organically, as well as exponentially, over a certain period. However, it is vital that SMEs acquire funding with low and fixed interest rates, as this would ascertain their financial reports with viable repayment of loans, over the period of the approval of loans.

A financial organization, such as Lasiandra Finance Inc. (LFI) New York, USA, will play a pivotal role in arming SMEs with a strong financial aid, to help them sustain and grow. The main benefit of LFI would be automating and customizing loans based on applicant-centric necessities. To achieve this goal, LFI requires Data Scientists to analyze historical datasets of applicants and build a predictive model for said applicants in getting loans approved or rejected. Data analytical programming is handy in addressing this issue, which would benefit both LFI and SMEs applying for loans, as waiting time in getting loans approved or rejected is reduced with the usage of Data Science.

CHAPTER 2: Background of Lasiandra Finance Inc. (LFI), New York, USA

LFI, which is a leading private financing company, provides loans for SMEs to exacerbate the businesses of SMEs rapidly. Providing loans to SMEs would be the kick-start that these organizations need, to increase revenue and progress of said organizations in the SMEs. The current manual loan application at LFI is tailor-made based on applicant needs and requirements. This would mean that applicants applying for a business loan would have to manually register their loan applications, to release funding for SMEs. This can benefit applicants by providing a channel for funding, that would uplift the businesses of SMEs in need of funding. The manual loan application process in LFI is said to meet the requirements of the applicant, and the selection of loans along with its loan amount and interest, would be matched with the amount of loan that each applicant is asking for. The manual loan application process of LFI takes a significantly great amount of time. Hence, LFI feels the need to automate their loan application process to provide encouragement for SMEs, by granting loans based on the approval status of applicants. The automation process of loans to SMEs would consider the portfolio that the applicants have previously entered, and these loans would be vetted automatically.

The manual process of applying for loans has proven to be of inconvenience to applicants. The current manual loan application process of LFI would have to consider having geographically distributed branch offices, to offer an efficient loan process, that would regulate and regulate the loan process and procedures. Currently, at LFI, the manual application process of loans requires a few simple procedures:

1) Assessing if the amount requested by the applicant matches the business model and operational expenditure of their organizations.
2) Ensuring that the correct loan program is provided to the applicant, based on their needs and requirement.
3) Conducting background checks on applicant's credit score, along with previous financial reports of applicants.
4) Verifying and validating the loan process for applicant, by granting approval or rejected status.

The approval of a manual loan application would take 3 – 7 working days, which is a long time for applicants to wait. The manual loan application of LFI would require an interview to be conducted with the borrower, to assess the loan in its entirety along with additional inputs made during the interview process. Next, once the loan application has been entered into LFI's loan processing computer system, a credit score report would then be provided to applicants. Usually, these reports have been processed earlier, before processing the loan applications. Moreover, the source of income of which the SMEs built the foundations of their business on, will then be verified. This would require verification of the employment of the borrower, income of the borrower and finally, the assets listed by the borrower. SMEs are then subjected to appraisals, insurances, and inspections of their organizations, for verification purposes. Upon the completion of this verification and validation process would then grant an approval status for applicants through the manual loan application process of LFI. Hence, LFI requires an automated loan application process, to speed up the approval of loans for applicants.

Automation would benefit LFI and its applicants, in terms of reducing processing application time. Approval of loans have been challenging, in terms of verifying and validating each application. Initially, applicants were vetted manually, which is a difficult process in terms of time taken for approval. Manual application of loans would require time, as background checks on organizations applying for loans would have to be thoroughly conducted to assess the performance of said companies and ability to re-pay loans, over a fixed period. Automation of loans, whilst necessary, would be a tricky task to achieve. This is because each applicant's loan status would have to be verified and validated quickly. LFI requires a model that would predict the approval loan status of applicants in a quick manner, which reduces the time taken for verification and validation of the loan given. Based on the above procedures, the automation process of LFI would significantly reduce the process required for applicants, along with verifying and validating applicants through an automated vetting process. The automation process would consider the following criteria of the applicants:

1) Marital Status of applicants
2) Number of dependencies of applicants
3) Academic qualifications of applicants
4) Employment status of applicants

5) Income status of applicants
6) Amount of loan requested by applicants
7) Duration of loan requested by applicants
8) Credit score of applicants
9) Location of applicants

Based on this information, the loan process can now be automatically vetted, which would provide applicants with a quicker response to their loan applications via LFI.

The objective of this assignment is to ensure that the chosen applicant for loan is a deserving candidate, using Data Analytical Programming. The dataset for analysis purpose used in this assignment is analyzed with the intent of building an accurate model using predictive analytics. This would then grant approval to loan applicants, based on the criteria required for loans, as mentioned earlier.

CHAPTER 3: Assumption and Justification

For this assignment, Data Scientists were tasked with the usage of the SAS Studio software, as this is an important tool for Data Analytical Programming. SAS Studio is used due to its syntax, which is easy to program and debug. SAS Studio uses a MySQL syntax for programming, which is a popular syntax for data science and analytics. Moreover, another advantage of SAS Studio is its ability to handle large databases. This is pivotal for bank loan applications in processing large datasets or Big Data, that can benefit LFI both in the short-term and long-term.

However, SAS studio is not open-sourced, which means a license is required to use SAS, that can lead to high cost in purchasing the license by LFI. Next, SAS is more of a procedural language, and as such, has more lines of code along with packages that are costly to purchase.

SAS Studio is useful for Predictive Analytics, which, in the usage of bank loan applications, will be beneficial to LFI in its ability to forecast bank loan applications based on historical data of their applicants. Data can also be presented in the form of tables and graphs in SAS Studio, which is useful in Prescriptive Analytics for applicants and SMEs that apply for loans, for future loan application purposes.

CHAPTER 4: Literature Review

4.1 Loan Applications for SMEs

The large numbers of Small and Medium Enterprises (SMEs) make up 98.5% of business enterprises in the United States of America (Wang et al., 2020). Small and Medium Enterprises (SMEs) face obstacles in procuring funding, as SMEs are not public listed companies, which means that funding for organizations cannot come from a constant stream of stakeholders that trade on the share market (Abuka et al., 2019). Various investors are skeptical in investing with SMEs (Moreira et al., 2018), as SMEs do not necessarily have a proven track record over a long period (Zhao & Zou, 2021), and equity within SMEs are usually privately owned (Al Azzawi, 2019). Moreover, application for loans in financial institutions such as banks have proven to be tricky (Zhao & Zou, 2021). The interest rates imposed on such loans are high, which would greatly reduce the amount of loan applicants (Al-Blooshi & Nobanee, 2020). Moreover, financial investors provide a thorough background check on SMEs (Moreira et al., 2018), to assess their eligibility in getting loans granted (Abuka et al., 2019). These background checks usually take a substantial amount of time, which would delay the process of loan approvals (Al Azzawi, 2019). Hence, having a strong financial investment company would greatly benefit SMEs in procuring required funds (Abuka et al., 2019), to run their operations and to increase revenue of their enterprises (Hidayat et al., 2020). Hence, a financial organization that can provide funding would be necessary in aiding the SMEs in the United States of America (Moreira et al., 2018), to ensure that companies can keep afloat in difficult times, and to provide the necessary impetus for these organizations to generate revenue with a source of funding (Al-Blooshi & Nobanee, 2020). However, there are drawbacks in funding SMEs (Abuka et al., 2019). As previously mentioned, background checks on SMEs are thoroughly conducted (Melnychenko et al., 2020), to ensure that financial performance over the period of funding can match the expectations of the lender to the borrower (Al Azzawi, 2019). This would require a thorough auditing of SMEs requiring financial assistance (Ivashina et al., 2020), and this process is a challenging and complicated one (Al-Blooshi & Nobanee, 2020). SMEs would have to be thoroughly vetted, with historical financial statements being produced, for assessment to take place (Abuka et al., 2019).

The interest rate charged by lenders also provide a setback for SMEs seeking for loans (Abuka et al., 2019). High interest rates would require a longer payback period for organizations, and the interest rates charged by funding institutions would have to be measured against the rates charged by banking institutions (Al-Blooshi & Nobanee, 2020). Hence, LFI would have to consider revising their interest rates when lending money (Al Azzawi, 2019).

The main considerations for a funding organization in providing loans for SMEs are:

1. Security and Documentations

   SMEs are required to provide proof of type of guarantees that can be used, in the event of bankruptcy of failure in repaying loans (Al-Blooshi & Nobanee, 2020). Moreover, these documentations would have to be approved by both the funding organizations and the government (Shrestha & Paudel, 2019), in assessment of deceit or fraud (Hidayat et al., 2020). Negligence on any parties involved in providing these documentations would be detrimental in the financial performance of SMEs and impact the credit score of said companies (Abuka et al., 2019). Thus, documentations that are legitimate would ensure that financial organizations such as LFI, would minimize risks in lending money to SMEs (Moreira et al., 2018).

2. Type of Guarantees

   The guarantees provided by SMEs to financial institutions consider the following (Hidayat et al., 2020):

   - Loan mortgages and immovable assets
   - Cash security, based on Cash Flow of the SMEs
   - Pledging of current assets
   - Personal guarantee of partners in loan repayment
   - Imported assets shipping documents
   - Title deed signifying ownership of land

These guarantees are pivotal for lenders such as LFI to provide loans, as SMEs are then obligated legally in repayment of loans (Shrestha & Paudel, 2019), and thus, loans would not be defaulted (Abuka et al., 2019). This would ensure that financial organizations providing loans would be protected, in cases of deceit or bankruptcy of SMEs (Gupta et al., 2020). Thus, it is vital that SMEs

applying for loans have adequate current and non-current assets (Saha & Waheed, 2017), along with a positive cash flow to ensure sustainability of these companies' financial performance over the years (Hidayat et al., 2020). Defaulting in payment of loans would increase debt in SMEs, that over time (Ivashina et al., 2020), would lead to the company folding and stakeholders losing their money (Hidayat et al., 2020).

Once approval of loans is successful, a sanction letter would then be provided to SMEs (Melnychenko et al., 2020), that would display the terms and conditions of the funding, of which a contract can then be signed between the SMEs and the financial organizations that provide funding (Abuka et al., 2019).

4.2 Challenges Faced by SMEs

Loan applications for SMEs come with its own mitigating risks and challenges (Saha & Waheed, 2017). The figure below depicts the relationship between challenges, risks and sustainability of SMEs applying for funding:

Figure 1: Challenges, Risks and Sustainability of SMEs in Applying Loans (Al Azzawi, 2019)

Challenges:

- It is pivotal for SMEs to acquire accounting and business information (Zhao & Zou, 2021), to ensure operations of the organization run smoothly (Ivashina et al., 2020). SMEs should primarily target long-term achievable goals, which increases the necessity of qualified accountants and auditors in said organizations (Basten & Ongena, 2020).
- In the event of a financial crisis, the demand of products and services of the SMEs would be challenging. This would affect the longevity of SMEs applying for loans, in terms of financial performance (Gupta et al., 2020).
- SMEs that face obstacles in maintaining operations centered on cash flow difficulties can prove tricky when applying for loans (Melnychenko et al., 2020).

Risks:

- High interest rates provide a risk to SMEs in application of loans (Serrasqueiro et al., 2018), which tend to prefer fixed-rate loans compared to adjustable-rate loans that have high interest rates (Ivashina et al., 2020).
- Risk of liquidity of an organization would result in collateral losses for the stakeholders. High debts and net losses in a SMEs financial statement would affect the loan application processes (Gupta et al., 2020).
- Technological adoption in SMEs come with obstacles, along with opportunities for said organizations (Wang et al., 2020). Inability in technological adoption would run the risk of SMEs to be left behind, in terms of revenue generation and sustainability over the years (Melnychenko et al., 2020).

Sustainability:

- SMEs are expected to survive, expand, and grow organically over a period, with financial backing (Saha & Waheed, 2017). Due to the uncertainty of the current and perhaps future economic conditions, this would prove to be tricky for SMEs in sustaining their businesses for the long run (Kassem & Trenz, 2020).
- Innovation of products and processes to sustain the SMEs future business operations is key, which is challenging as funding is required for research and development, which would reflect in the accounts of an organization when applying for loans (Kassem &

Trenz, 2020). Moreover, SMEs must match competitors' products and innovation over the years, which can prove tricky (Gupta et al., 2020).

The challenges faced by SMEs in applying for loans are great, which can be mitigated with proper planning and smooth operations of business (Al Azzawi, 2019). SMEs would have to navigate these challenges, both short-term and long-term (Zhao & Zou, 2021), before applying for loans (Gupta et al., 2020). Funding from financial institutions would consider the growth and long-term sustainability of the SME before granting loans (Saha & Waheed, 2017). As mentioned earlier, the failure to repay loans would result in grave financial implications, that can lead to financial bankruptcy and folding of the organizations (Basten & Ongena, 2020). Hence, it is pivotal that SMEs have good bookkeeping and financial planning (Serrasqueiro et al., 2018), that can ensure sustainability over the loan period (Gupta et al., 2020).

4.3 Automation Process of Loan Applications

Currently, the most practiced loan application process can be summarized in Figure (2) below:

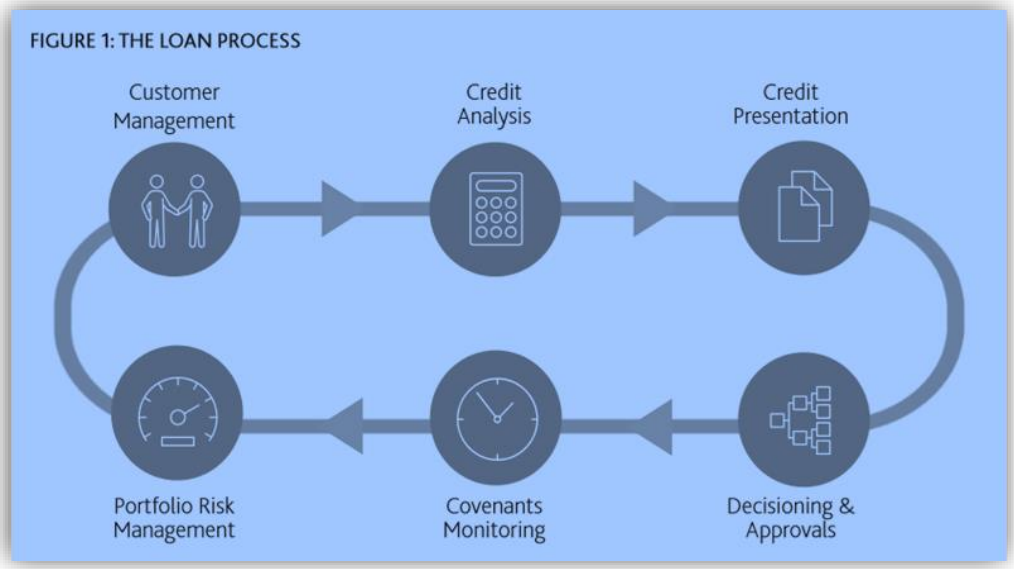


Figure 2: Current Loan Application Process of SMEs (Kassem & Trenz, 2020)

This process can be challenging for SMEs, as the time taken for loans to be granted can be lengthy (Zhao & Zou, 2021). It is pivotal that the process of loan application is conducted at a quicker rate (Gupta et al., 2020), hence the need for automation of these loan application processes (Al Azzawi, 2019). In the current digital era, automation of loan applications can be conducted at a higher rate (Basten & Ongena, 2020). There are various financial organizations already providing such services, such as Credit Online. The automation loan process of Credit Online is depicted in the figure below:

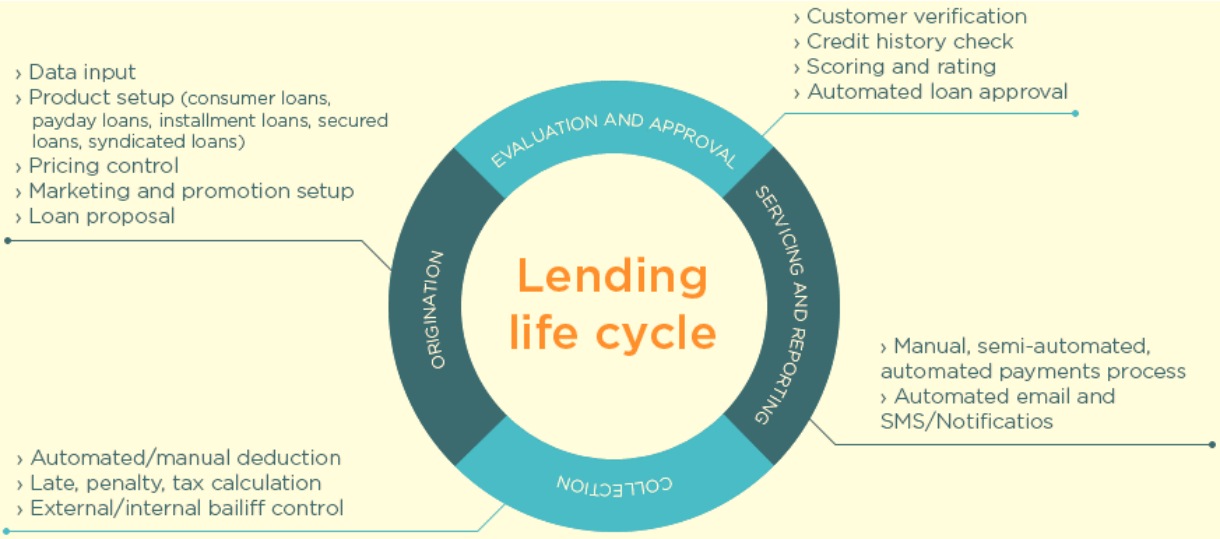


Figure 3: Automation Loan Process of Lenders (Al-Blooshi & Nobanee, 2020)

Automation process of loan applications would require data science and data analytics. Borrowers are required to submit necessary documents for loans (Serrasqueiro et al., 2018), and approvals can be granted or denied based on certain criteria that financial organizations impose (Al Azzawi, 2019). Data science that considers factors of a loanee is beneficial in the automation process of loan applications (Al-Blooshi & Nobanee, 2020). Since Big Data Analytics consist of both Predictive and Prescriptive Analytics (Wang et al., 2020), SMEs and funding organizations can benefit from the usage of data analytics in predicting the loan approvals (Basten & Ongena, 2020), along with providing solutions to SMEs in future loan applications and sustainability of their business operations (Al Azzawi, 2019).

(Gupta et al., 2020) states the framework for automating loan applications in the figure below:

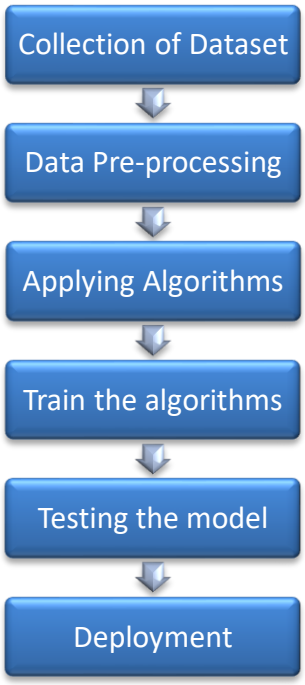


Figure 4: Framework for Automation of Loan Applications (Gupta et al., 2020)

Data analytics is pivotal for financial organizations such as LFI, in granting loan approvals for SMEs that would reduce waiting time for funds to be granted (Basten & Ongena, 2020). Automation of loan applications currently exist (Zhao & Zou, 2021), and with the further advancement of technologies such as Machine Learning and Artificial Intelligence (Gupta et al., 2020), this would further benefit both the lender and the borrower (Zhao & Zou, 2021).

CHAPTER 5: Exploration on the Given Dataset

5.1 Structure of the Training Dataset Given

Based on the training dataset, there are a total of 614 observations, with 13 variables, as depicted in the table below:

Name of variable	Description	Data Type	Length	Sample Data
SME_LOAN_ID_NO	Loan application number	Char	8	LP001002/LP001003
GENDER	Gender of the applicant; Male or Female	Char	6	Female; Male
MARITAL STATUS	Marital status of the applicant; Married or Not Married	Char	10	Married; Not Married
FAMILY MEMBERS	Number of family members of the applicant	Numeric	2	1;2;3+
QUALIFICATION	Academic qualification of the applicant	Char	14	Graduate; Under Graduate
EMPLOYMENT	Employment status of the applicant	Char	3	Yes; No
CANDIDATE_INCOME	Amount of monthly income of the applicant	Numeric	5	12481; 5000

GUARANTEE_INCOME	Amount of guaranteed income of the applicant	Numeric	5	10968; 1508
LOAN_AMOUNT	Loan amount	Numeric	5	128; 66
LOAN_DURATION	Duration of loan applied by applicant, in months	Numeric	3	360; 480
LOAN_HISTORY	History of previous loan undertaken by applicant	Numeric	1	0; 1
LOAN_LOCATION	Location of applicant when applying for loan	Char	7	Village; Town; City
LOAN_APPROVAL_STATUS	Status of loan approval	Char	1	Y; N

Table 1: Table of Training Dataset Given

CHAPTER 6: Methodology

The methodology for this project follows the workflow chart depicted in the figure below:

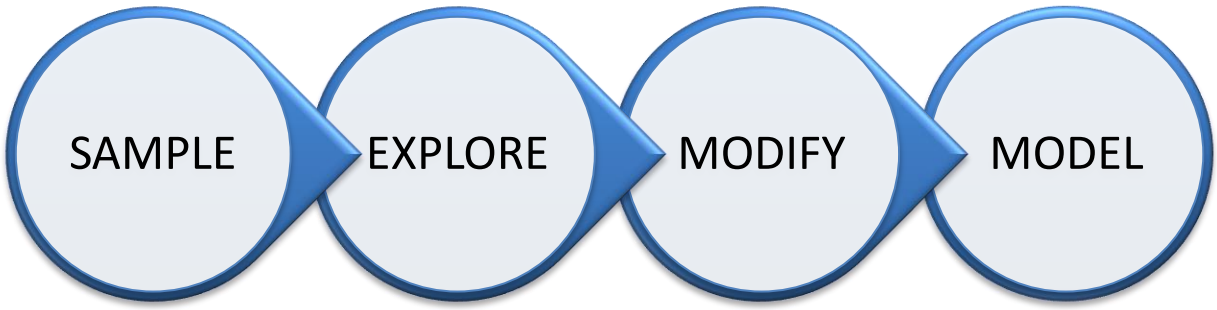


Figure 5: Assignment Workflow Chart

Based on Figure (5) above, the sample data which is the training dataset shown in Table (1) is extracted onto SAS Studio, by renaming a database file for the dataset to be loaded.

Next, data exploration is performed onto the dataset. In this process, data imputation for missing values is calculated, and performed to replace missing values with data imputation techniques. Dataset is also explored in terms of all the variables present in the dataset.

Based on this training dataset, a model is then built that would correspond to the outputs that can be provided based on the requirements of the assignment. Analysis is performed, using uni-variate or bi-variate analysis methods, to procure the accurate outputs required for this assignment. Data

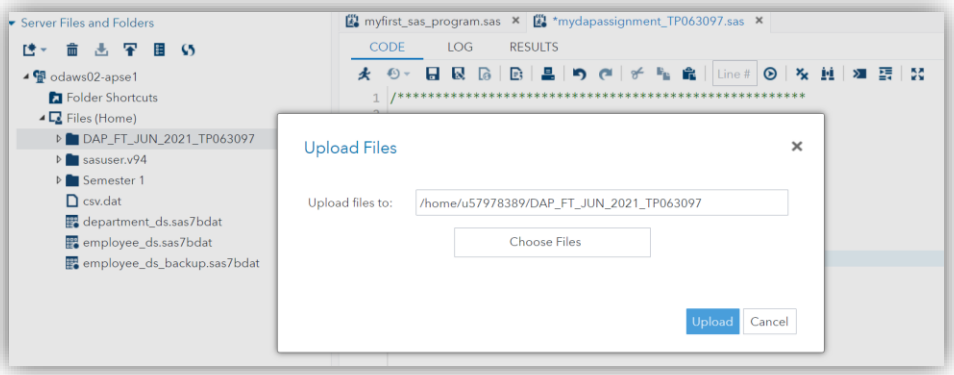
scientists are tasked with performing Predictive and Prescriptive Analytics based on Data Analytical Programming using SAS Studio.

Finally, the model of loan application process is ready to be used, and this can then be performed on all the observations in the dataset, to find the optimum result and to cross-check the number of applicants successful in their loan approval status. Graphs and tables can be produced as part of Exploratory Data Analysis to further boost the results that can be produced in the output section of this assignment.

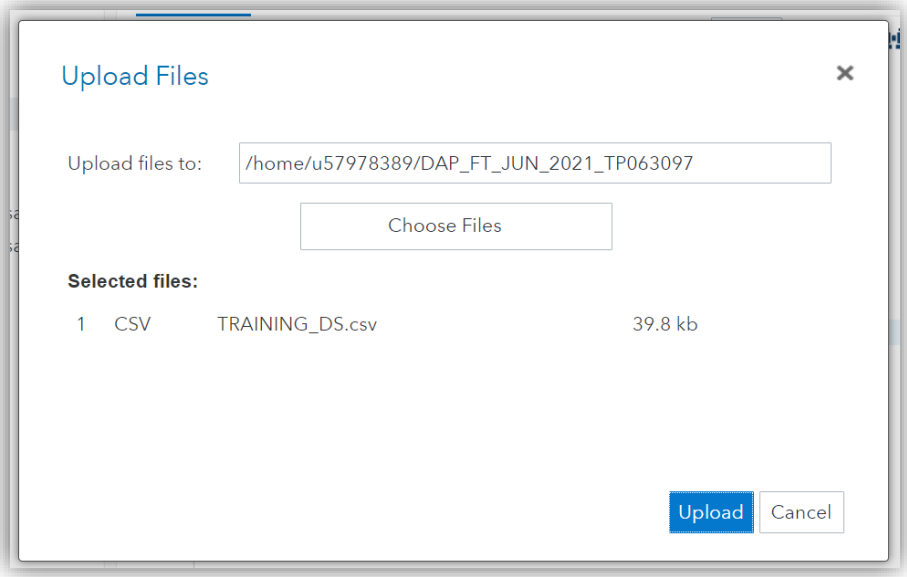
CHAPTER 7: Experimentation

7.1 Upload the dataset of TRAINING_DS and TESTING_DS to the newly created folder

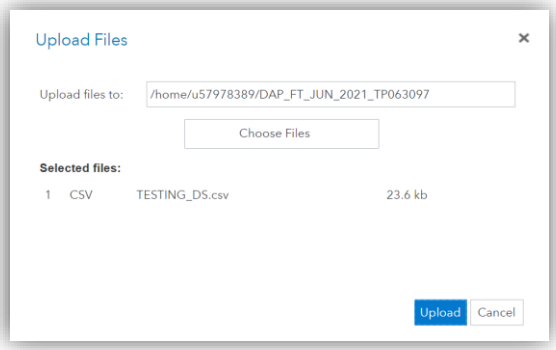
Step 1: Choose destination folder



Step 2: Upload TRAINING_DS file to destination folder



Step 3: Upload TESTING_DS file to destination folder



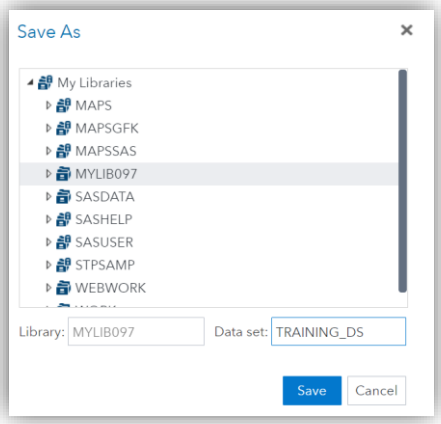
Both datasets have now been uploaded to the folder:

7.2 Create a permanent library on SAS

7.2.1 Explanation

Creating a permanent library would ensure that both the files would remain in the library on SAS, to ensure that files can be revisited and updated when necessary.

7.2.2 Screenshots



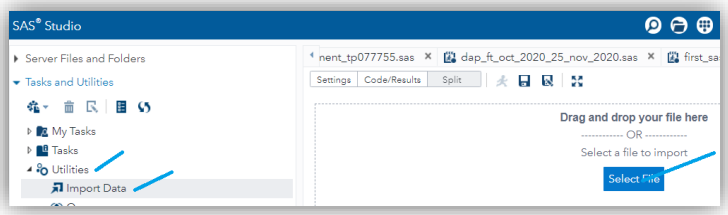
MYLIB097 is the destination for the library of the SAS file, for this assignment.

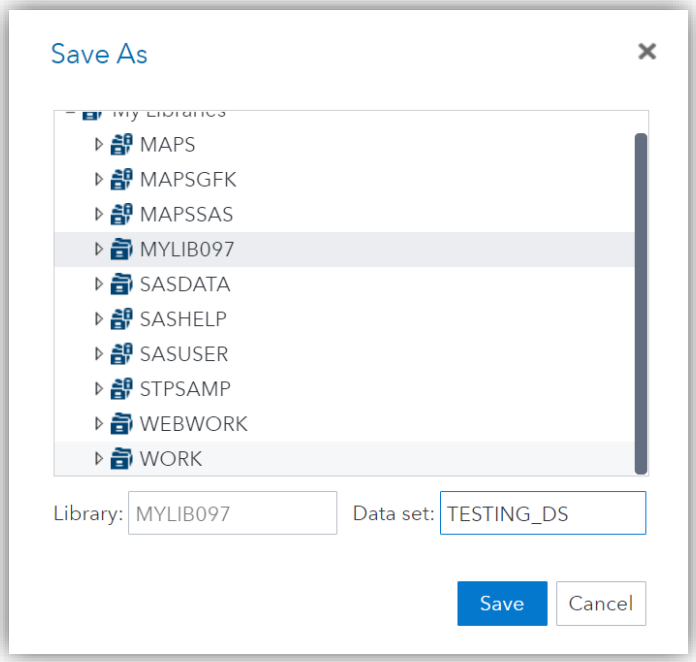
7.3 Import the dataset TRAINING_DS

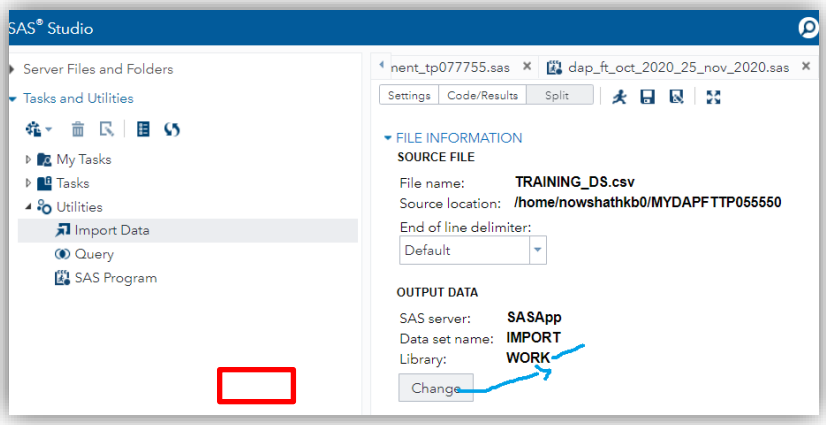
7.1.1 Explanation

Dataset of TRAINING_DS is imported into the library, to perform Exploratory Data Analysis (EDA), pre-processing, imputation, and analysis.

7.1.2 Screenshots





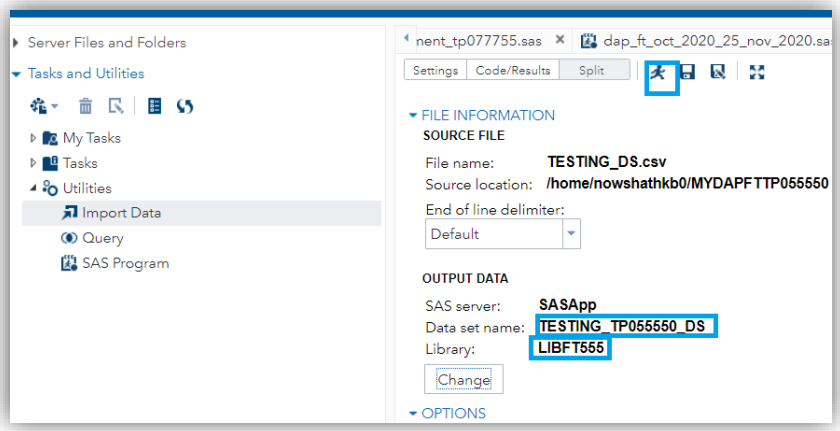


The library is changed to MYLIB097

7.2 Import the dataset TESTING_DS

7.2.1 Explanation

The same steps are repeated for the TESTING_DS dataset, as both TRAINING_DS and TESTING_DS datasets are used for this assignment.



7.3 Create a copy of the TRAINING_DS dataset

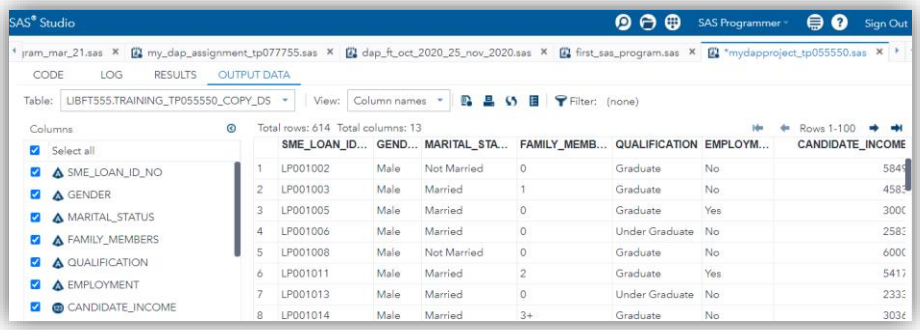
7.3.1 Explanation

A copy of the TRAINING_DS dataset is created, to ensure that any changes that occur will not damage the original dataset, for fault-tolerance and replication of the dataset. The new dataset is named as MYLIB097.TRAINING_DS_TP063097_BK

7.3.2 SAS Codes

```
/*****************************************************

Name of DS: Mr. Thines Kumar Nadaraja - (TP063097)
Name of SAS program: mydapassignment_TP063097.sas
Description:
Date first written: Monday 19 July 2021
Date last updated: Monday 19 July 2021
Folder name: DAP_FT_JUN_2021_TP063097
Library name: MYLIB097

*****************************************************/

TITLE 'DAP Assignment';
/* Title of DAP Assignment */
PROC SQL;

CREATE TABLE MYLIB097.TRAINING_DS_TP063097_BK AS
SELECT * FROM MYLIB097.TRAINING_DS;

QUIT;
```

7.3.3 Outputs/Results



This table depicts the dataset of the assignment, whereby all the variables that affect the loan of a applicant is shown.

```
PROC SQL;

DESCRIBE TABLE MYLIB097.TRAINING_DS_TP063097_BK;

QUIT;
```

```
create table MYLIB097.TRAINING_DS_TP063097_BK( 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.
  );
```

Each variable has a length and format, as shown in the image above.

7.4 Label each variable in the dataset

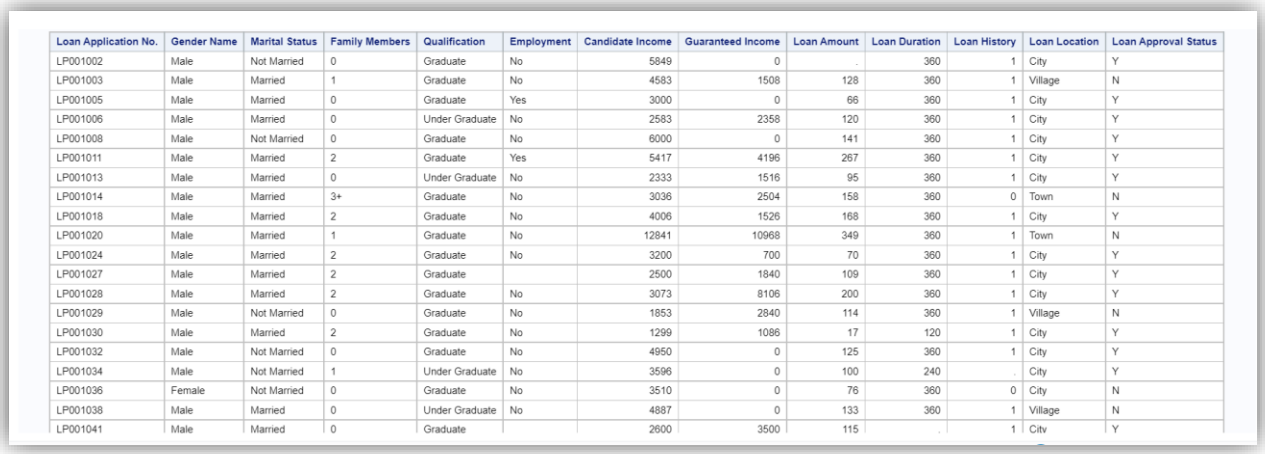
7.4.1. Explanation

Each variable is labelled, as the original label has many foreign characters (such as _), which would be difficult to code. The purpose of labelling variables is to enable easier coding, and better understanding of the dataset.

7.4.2 SAS Codes

```
DATA MYLIB097.TRAINING_DS_TP063097_BK;
SET MYLIB097.TRAINING_DS_TP063097_BK;
LABEL
    SME_LOAN_ID_NO = 'Loan Application No.'
    GENDER = 'Gender Name'
    MARITAL_STATUS = 'Marital Status'
    FAMILY_MEMBERS = 'Family Members'
    QUALIFICATION = 'Qualification'
    EMPLOYMENT = 'Employment'
    CANDIDATE_INCOME = 'Candidate Income'
    GUARANTEE_INCOME = 'Guaranteed Income'
    LOAN_AMOUNT = 'Loan Amount'
    LOAN_DURATION = 'Loan Duration'
    LOAN_HISTORY = 'Loan History'
    LOAN_LOCATION = 'Loan Location'
    LOAN_APPROVAL_STATUS = 'Loan Approval Status';
RUN;
```

7.4.3 Outputs/Results



Loan Application No.	Gender Name	Marital Status	Family Members	Qualification	Employment	Candidate Income	Guaranteed Income	Loan Amount	Loan Duration	Loan History	Loan Location	Loan Approval Status
LP001002	Male	Not Married	0	Graduate	No	5849	0	.	360	1	City	Y
LP001003	Male	Married	1	Graduate	No	4583	1508	128	360	1	Village	N
LP001005	Male	Married	0	Graduate	Yes	3000	0	66	360	1	City	Y
LP001006	Male	Married	0	Under Graduate	No	2583	2358	120	360	1	City	Y
LP001008	Male	Not Married	0	Graduate	No	6000	0	141	360	1	City	Y
LP001011	Male	Married	2	Graduate	Yes	5417	4196	267	360	1	City	Y
LP001013	Male	Married	0	Under Graduate	No	2333	1516	95	360	1	City	Y
LP001014	Male	Married	3+	Graduate	No	3036	2504	158	360	0	Town	N
LP001018	Male	Married	2	Graduate	No	4006	1526	168	360	1	City	Y
LP001020	Male	Married	1	Graduate	No	12841	10968	349	360	1	Town	N
LP001024	Male	Married	2	Graduate	No	3200	700	70	360	1	City	Y
LP001027	Male	Married	2	Graduate		2500	1840	109	360	1	City	Y
LP001028	Male	Married	2	Graduate	No	3073	8106	200	360	1	City	Y
LP001029	Male	Not Married	0	Graduate	No	1853	2840	114	360	1	Village	N
LP001030	Male	Married	2	Graduate	No	1299	1086	17	120	1	City	Y
LP001032	Male	Not Married	0	Graduate	No	4950	0	125	360	1	City	Y
LP001034	Male	Not Married	1	Under Graduate	No	3596	0	100	240	.	City	Y
LP001036	Female	Not Married	0	Graduate	No	3510	0	76	360	0	City	N
LP001038	Male	Married	0	Under Graduate	No	4887	0	133	360	1	Village	N
LP001041	Male	Married	0	Graduate		2600	3500	115	.	1	City	Y

The output shows that all variables have been labelled accordingly, for better understanding and clarity when coding.

7.5 Univariate Analysis on the variables found in the MYLIB097.TRAINING_DS_TP063097_BK

7.5.1 Explanation

Univariate analysis ins conducted to analyse the dataset in the most uncomplicated manner. Each data that is analysed only contains one variable. Univariate analysis enables data description, and for patterns to be recognised.

7.5.2 Univariate Analysis on GENDER – Categorical variable

7.5.2.1 SAS Codes

```
/* UNIVARIATE ANALYSIS FOR GENDER */

TITLE 'UNIVARIATE ANALYSIS GENDER';
PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE GENDER;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBAR GENDER;
TITLE 'GENDER';
RUN;
```

7.5.2.2 Outputs/Results

UNIVARIATE ANALYSIS GENDER

The FREQ Procedure

Gender				
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	112	18.64	112	18.64
Male	489	81.36	601	100.00
Frequency Missing = 13				

The count for gender of Male is 489, while Female is 112. There are 13 missing values for the GENDER variable, which is a Continuous variable.

7.5.3 Univariate Analysis on MARITAL_STATUS – Categorical variable

7.5.3.1 Explanation and Analysis

The univariate analysis is conducted on the MARITAL_STATUS variable, which is a categorical variable. This variable states the marital status of the applicant applying for loan, whether Married or Not Married.

7.5.3.2 SAS Codes

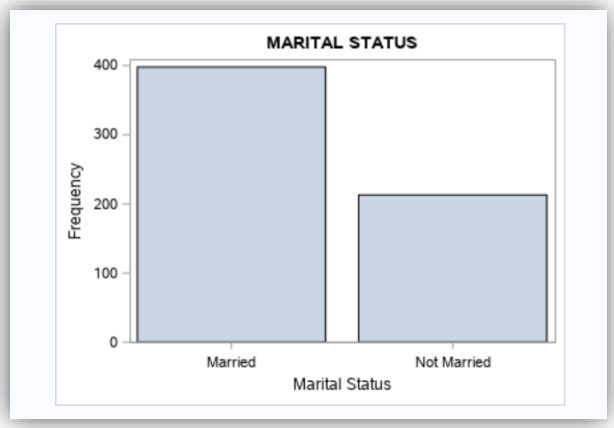
```
TITLE 'UNIVARIATE ANALYSIS';
PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE MARITAL_STATUS;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBAR MARITAL_STATUS;
TITLE 'FIGURE 7.8.2 TITLE';
RUN;
```

7.5.3.3 Outputs/Results

UNIVARIATE ANALYSIS FOR MARITAL STATUS

The FREQ Procedure

Marital Status				
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				



7.5.4 Univariate Analysis on LOAN_LOCATION– Categorical variable

7.5.4.1 Explanation and Analysis

LOAN_LOCATION is a categorical variable, that describes the location of each applicant applying for loan, whether City, Town or Village.

7.5.4.2 SAS Codes

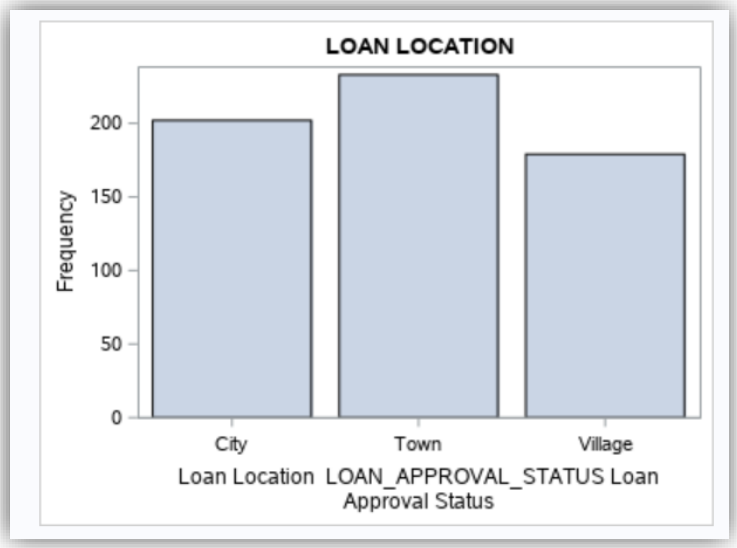
```
/* UNIVARIATE ANALYSIS FOR LOAN LOCATION */

TITLE 'UNIVARIATE ANALYSIS FOR LOAN LOCATION';
PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE LOAN_LOCATION;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBAR LOAN_LOCATION;
TITLE 'LOAN LOCATION';
RUN;
```

7.5.4.3 Outputs/Results

UNIVARIATE ANALYSIS FOR LOAN LOCATION

The FREQ Procedure

Loan Location LOAN_APPROVAL_STATUS Loan Approval Status				
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



7.5.5 Univariate Analysis on EMPLOYMENT– Categorical variable

7.5.5.1 Explanation and Analysis

EMPLOYMENT is a categorical variable, with outputs being either YES or NO.

7.5.5.2 SAS Codes

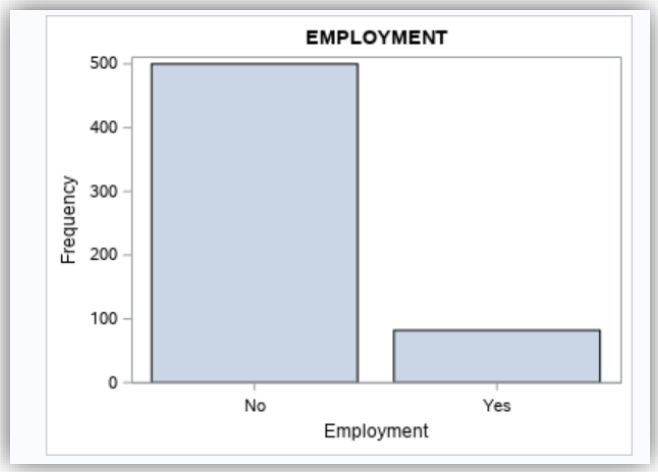
```
/* UNIVARIATE ANALYSIS FOR EMPLOYMENT */

TITLE 'UNIVARIATE ANALYSIS FOR EMPLOYMENT';
PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE EMPLOYMENT;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBAR EMPLOYMENT;
TITLE 'EMPLOYMENT';
RUN;
```

7.5.5.3 Outputs/Results

UNIVARIATE ANALYSIS FOR EMPLOYMENT

The FREQ Procedure

Employment				
EMPLOYMENT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	500	85.91	500	85.91
Yes	82	14.09	582	100.00
Frequency Missing = 32				



7.5.6 Univariate Analysis on QUALIFICATION– Categorical variable

7.5.6.1 Explanation and Analysis

QUALIFICATION is a categorical variable, to describe the academic qualification of the applicant, whether Graduate or Under Graduate.

7.5.6.2 SAS Codes

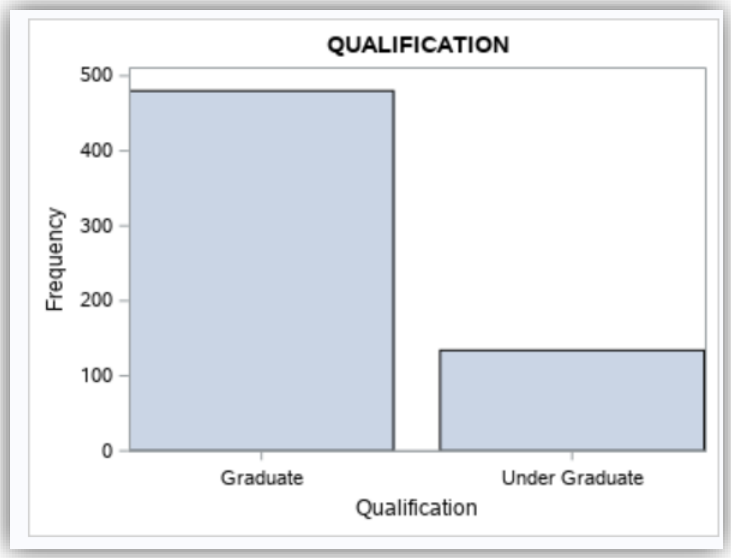
```
/* UNIVARIATE ANALYSIS FOR QUALIFICATION */

TITLE 'UNIVARIATE ANALYSIS FOR QUALIFICATION';
PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE QUALIFICATION;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBAR QUALIFICATION;
TITLE 'QUALIFICATION';
RUN;
```

7.5.6.3 Outputs/Results:

UNIVARIATE ANALYSIS FOR QUALIFICATION

The FREQ Procedure

Qualification				
QUALIFICATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Graduate	480	78.18	480	78.18
Under Graduate	134	21.82	614	100.00



7.5.7 Univariate Analysis on FAMILY_MEMBERS– Categorical variable

7.5.7.1 Explanation and Analysis

FAMILY_MEMBERS is a Categorical variable, describing the number of family members the applicant has. The output has either 0,1,2 or 3^{+}.

7.5.7.2 SAS Codes

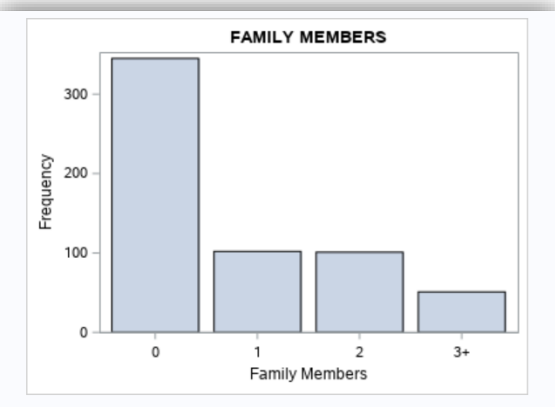
```
/* UNIVARIATE ANALYSIS FOR FAMILY MEMBERS */

TITLE 'UNIVARIATE ANALYSIS FOR FAMILY MEMBERS';
PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE FAMILY_MEMBERS;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBAR FAMILY_MEMBERS;
TITLE 'FAMILY MEMBERS';
RUN;
```

7.5.7.3 Outputs/Results:

UNIVARIATE ANALYSIS FOR FAMILY MEMBERS

The FREQ Procedure

Family Members				
FAMILY_MEMBERS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	345	57.60	345	57.60
1	102	17.03	447	74.62
2	101	16.86	548	91.49
3+	51	8.51	599	100.00
Frequency Missing = 15				



No. of observations after imputation

NO. OF OBSERVATIONS
0

Next, the removal of the + symbol from the metadata of the FAMILY_MEMBERS is done.

```
/*Step 1 */

TITLE 'Step 1';

PROC SQL;
SELECT COUNT (*) LABEL = 'NO. OF OBSERVATIONS'
FROM MYLIB097.TRAINING_DS_TP063097_FM t
WHERE ( SUBSTR (t.family_members,2,1) EQ '+'); /*No. of observations with + symbol */
QUIT;

/*
Before removal of the + symbol from 0,1,2,3+
*/

TITLE 'Step 2';

PROC SQL;
UPDATE MYLIB097.TRAINING_DS_TP063097_FM
SET family_members = SUBSTR(family_members,1,1)
WHERE ( SUBSTR (family_members,2,1) EQ '+'); /*No. of observations with + symbol */
QUIT;

/*
Step 3: After removal of the + sign, list the observation with missing values */

TITLE 'REMOVAL OF THE (+) SIGN';
PROC SQL;
SELECT t.FAMILY_MEMBERS LABEL = 'FAMILY CATEGORY',
        COUNT(*) LABEL = 'NO. OF APPLICANTS'
FROM MYLIB097.TRAINING_DS_TP063097_FM t
GROUP BY t.FAMILY_MEMBERS;
QUIT;
```

REMOVAL OF THE (+) SIGN

FAMILY CATEGORY	NO. OF APPLICANTS
0	360
1	102
2	101
3	51

7.5.8 Univariate Analysis on LOAN_HISTORY– Categorical variable

7.5.8.1 Explanation and Analysis

LOAN_HISTORY is a Categorical variable that describes the loan history of the applicant, whether Yes (Output as 1) or No (Output as 0).

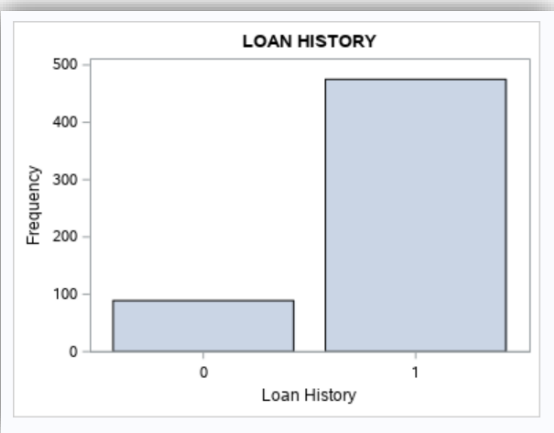
7.5.8.2 SAS Codes

```
/* UNIVARIATE ANALYSIS FOR LOAN HISTORY */

TITLE 'UNIVARIATE ANALYSIS FOR LOAN HISTORY';
PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE LOAN_HISTORY;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBAR LOAN_HISTORY;
TITLE 'LOAN HISTORY';
RUN;
```

UNIVARIATE ANALYSIS FOR LOAN HISTORY

The FREQ Procedure

Loan History				
LOAN_HISTORY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	89	15.78	89	15.78
1	475	84.22	564	100.00
Frequency Missing = 50				



7.5.9 Univariate Analysis on CANDIDATE_INCOME – Continuous variable

7.5.9.1 Explanation and Analysis

CANDIDATE_INCOME is a Continuous variable that describes the income amount of the applicant.

7.5.9.2 SAS Codes

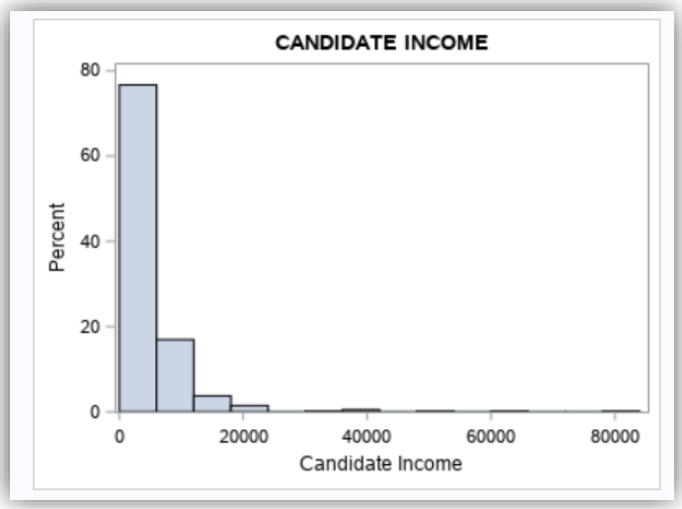
```
/* UNIVARIATE ANALYSIS FOR CANDIDATE INCOME */

TITLE 'UNIVARIATE ANALYSIS FOR CANDIDATE INCOME';
PROC MEANS DATA = MYLIB097.TRAINING_DS_TP063097_BK N NMISS MIN MAX MEAN MEDIAN STD;
VAR CANDIDATE_INCOME;
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
HISTOGRAM CANDIDATE_INCOME;
TITLE 'CANDIDATE INCOME';
RUN;
```

7.5.9.3 Outputs/Results

UNIVARIATE ANALYSIS FOR CANDIDATE INCOME

The MEANS Procedure

Analysis Variable : CANDIDATE_INCOME Candidate Income						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
614	0	150.0000000	81000.00	5403.46	3812.50	6109.04



7.5.10 Univariate Analysis on LOAN_AMOUNT – Continuous variable

7.5.10.1 Explanation and Analysis

LOAN_AMOUNT is a Continuous variable that describes the loan amount that the applicant is applying for.

7.5.10.2 SAS Codes

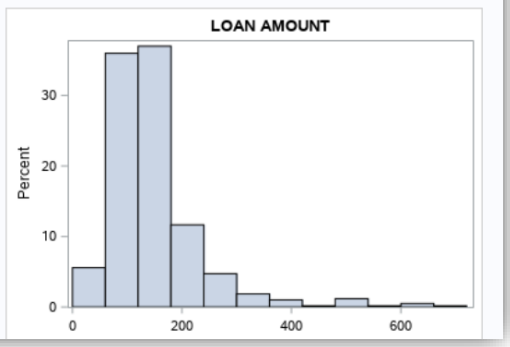
```
/* UNIVARIATE ANALYSIS FOR LOAN AMOUNT */

PROC MEANS DATA = MYLIB097.TRAINING_DS_TP063097_BK ;
VAR LOAN_AMOUNT;
TITLE 'Fig 7.11 Univariate Analysis on Candidate Income';
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
HISTOGRAM LOAN_AMOUNT;
TITLE 'LOAN AMOUNT';
RUN;
```

7.5.10.3 Outputs/Results

Fig 7.11 Univariate Analysis on Candidate Income

The MEANS Procedure

Analysis Variable : LOAN_AMOUNT Loan Amount				
N	Mean	Std Dev	Minimum	Maximum
592	146.4121622	85.5873252	9.0000000	700.0000000



Next, missing values of the numeric variable LOAN_AMOUNT is displayed.

```
TITLE 'DISPLAY MISSING VALUES IN LOAN_AMOUNT';
PROC SQL;

SELECT *
FROM MYLIB097.TRAINING_ds_TP063097_FM t
WHERE ( t.LOAN_AMOUNT EQ . );
QUIT;
```

DISPLAY MISSING VALUES IN LOAN_AMOUNT

Loan Application No.	Gender	Marital Status	Family Members	Qualification	Employment	Candidate Income	Guaranteed Income	Loan Amount	Loan Duration	Loan History	Loan Location LOAN_APPROVAL_STATUS Loan Approval Status	LOAN_APF
LP001002	Male	Not Married	0	Graduate	No	5849	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		6782	0	.	360	.	City	N
LP001350	Male	Married	0	Graduate	No	13650	0	.	360	1	City	Y
LP001356	Male	Married	0	Graduate	No	4652	3583	.	360	1	Town	Y
LP001392	Female	Not Married	1	Graduate	Yes	7451	0	.	360	1	Town	Y
LP001449	Male	Not Married	0	Graduate	No	3865	1640	.	360	1	Village	Y
LP001682	Male	Married	3	Under Graduate	No	3992	0	.	180	1	City	N
LP001922	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	1830	0	.	360	0	City	N
LP002243	Male	Married	0	Under Graduate	No	3010	3136	.	360	0	City	N
LP002393	Female		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	2947	1603	.	360	1	City	N
LP002697	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
LP002960	Male	Married	0	Under Graduate	No	2400	3800	.	180	1	City	N

```
TITLE 'FIND NUMBER OF OBSERVATIONS OF MISSING VALUES FOR LOAN_AMOUNT';

PROC SQL;
SELECT COUNT (*) LABEL = 'NUMBER OF OBS'
FROM MYLIB097.TRAINING_DS_TP063097_FM t
WHERE (t.LOAN_AMOUNT EQ . );
QUIT;
```

FIND NUMBER OF OBSERVATIONS OF MISSING VALUES FOR LOAN_AMOUNT

NUMBER OF OBS
22

Number of missing values for LOAN_AMOUNT is 22

```
/*STEP 3*/

TITLE 'COPY OF THE DATASET';

PROC SQL;
CREATE TABLE MYLIB097.TRAINING_DS_TP063097_LA AS /* Before imputation, a copy of the dataset
SELECT * FROM MYLIB097.TRAINING_DS_TP063097_FM;
QUIT;
```

7.5.11 Univariate Analysis on LOAN_DURATION– Continuous variable

7.5.11.1 Explanation and Analysis

LOAN_DURATION is the duration of time that the applicant is applying loan for. This is a continuous variable.

7.5.11.2 SAS Codes

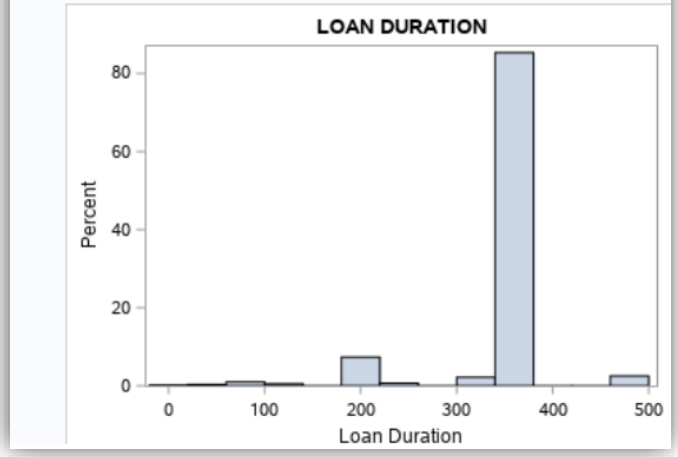
```
/* UNIVARIATE ANALYSIS FOR LOAN DURATION */

PROC MEANS DATA = MYLIB097.TRAINING_DS_TP063097_BK ;
VAR LOAN_AMOUNT;
TITLE 'Fig 7.11 Univariate Analysis on Candidate Income';
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
HISTOGRAM LOAN_DURATION;
TITLE 'LOAN DURATION';
RUN;
```

7.5.11.3 Outputs/Results

Fig 7.11 Univariate Analysis on Candidate Income

The MEANS Procedure

Analysis Variable : LOAN_AMOUNT Loan Amount				
N	Mean	Std Dev	Minimum	Maximum
592	146.4121622	85.5873252	9.0000000	700.0000000



7.5.12 Univariate Analysis on GUARANTEE_INCOME– Continuous variable

7.5.12.1 Explanation and Analysis

GUARANTEE_INCOME is a continuous variable, that describes the amount of income that the applicant is guaranteed on an annual basis. The reason this is important is to ensure viability of the applicant in repayment of the loan.

7.5.12.2 SAS Codes

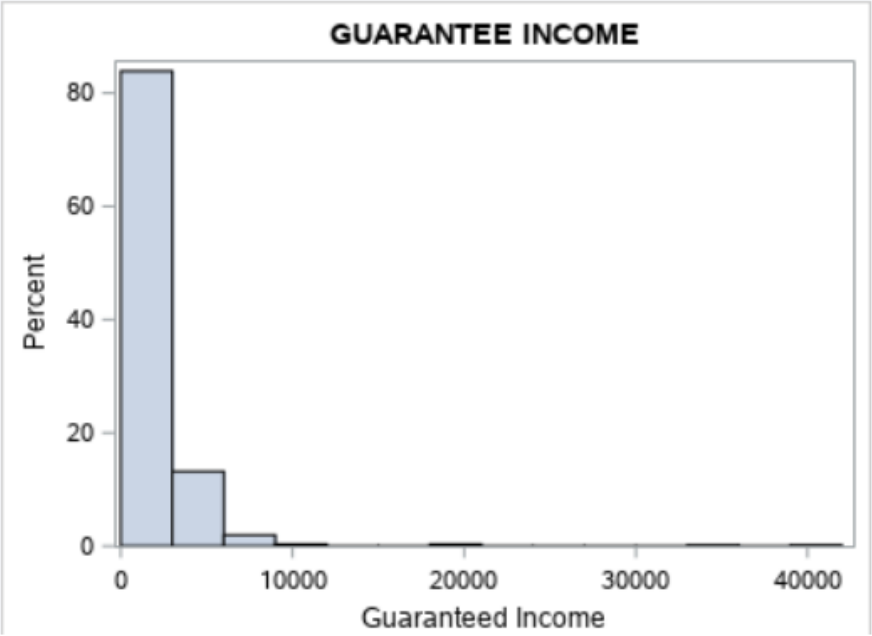
```
/* UNIVARIATE ANALYSIS FOR LOAN DURATION */

PROC MEANS DATA = MYLIB097.TRAINING_DS_TP063097_BK ;
VAR LOAN_AMOUNT;
TITLE 'Fig 7.11 Univariate Analysis on Guarantee Income';
RUN;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
HISTOGRAM GUARANTEE_INCOME;
TITLE 'GUARANTEE INCOME';
RUN;
```

7.5.12.3 Outputs/Results:

Fig 7.11 Univariate Analysis on Guarantee Income

The MEANS Procedure

Analysis Variable : LOAN_AMOUNT Loan Amount				
N	Mean	Std Dev	Minimum	Maximum
592	146.4121622	85.5873252	9.0000000	700.0000000



7.6 Bivariate Analysis on the variables found in the MYLIB097.TRAINING_DS_TP063097_BK dataset

7.6.1 Explanation

Bivariate analysis is important to ascertain the magnitude of predicting values for one variable, that could be a dependent variable, if the independent variable is known. Bivariate analysis is normally used to explain correlation or a linear regression of variables, to understand relationship between them.

7.6.2 Bivariate Analysis on Categorical variable (GENDER) VS Categorical variable (MARITAL_STATUS)

7.6.2.1 SAS Codes

```
/* Bivariate Analysis on GENDER (CATEGORICAL VARIABLE) & MARITAL_STATUS (CATEGORICAL VARIABLE) */

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE GENDER * MARITAL_STATUS /
PLOTS = FREQPLOT (TWOWAY = STACKED SCALE = GROUPPCT );
TITLE 'BIVARIATE ANALYSIS ON GENDER VS MARITAL STATUS';
RUN;
```

7.6.2.2 Outputs/Results

BIVARIATE ANALYSIS ON GENDER VS MARITAL STATUS

The FREQ Procedure

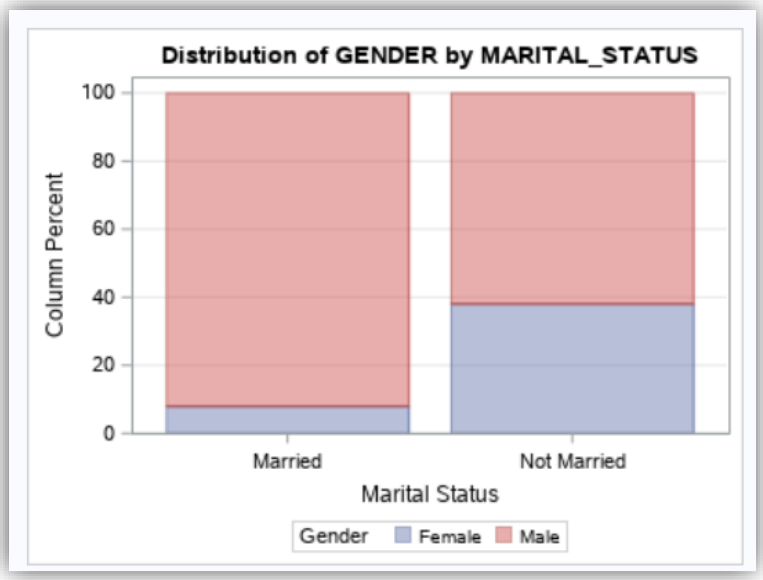
Frequency
Percent
Row Pct
Col Pct

Table of GENDER by MARITAL_STATUS

GENDER(Gender)	MARITAL_STATUS(Marital Status) Married	Not Married	Total
Female	31 5.18 27.93 7.99	80 13.38 72.07 38.10	111 18.56
Male	357 59.70 73.31 92.01	130 21.74 26.69 61.90	487 81.44
Total	388 64.88	210 35.12	598 100.00

Frequency Missing = 16

Missing values are 16 for this variable



7.6.3 Bivariate Analysis on Categorical variable (GENDER) VS Categorical variable (LOAN_APPROVAL_STATUS)

7.6.3.1 Explanation and Analysis

Bivariate analysis is performed on the Gender and LOAN_APPROVAL_STATUS variable. From this analysis, the frequency of approval of loans for both genders are analysed

7.6.3.2 SAS Codes

```
/* Bivariate Analysis on GENDER (CATEGORICAL VARIABLE) & LOAN_APPROVAL_STATUS (CATEGORICAL VARIA

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE GENDER * LOAN_APPROVAL_STATUS /
PLOTS = FREQPLOT (TWOWAY = STACKED SCALE = GROUPPCT );
TITLE 'BIVARIATE ANALYSIS ON GENDER VS LOAN APPROVAL STATUS';
RUN;
```

7.6.3.3 Outputs/Results

BIVARIATE ANALYSIS ON GENDER VS LOAN APPROVAL STATUS

The FREQ Procedure

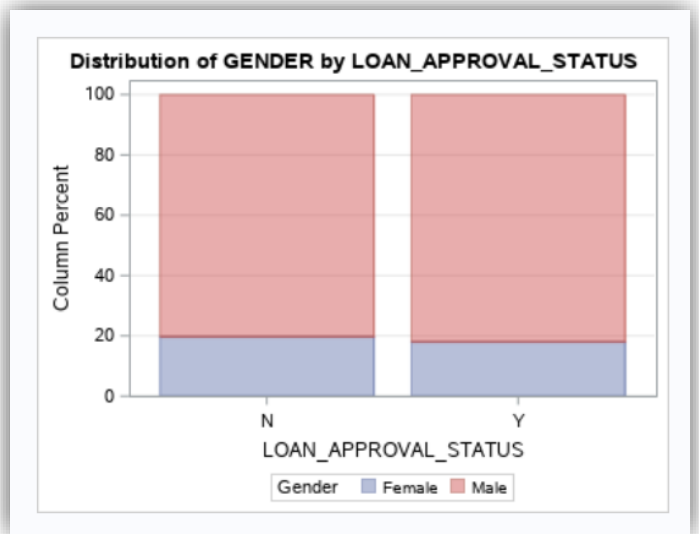
Frequency
Percent
Row Pct
Col Pct

Table of GENDER by LOAN_APPROVAL_STATUS

GENDER(Gender)	LOAN_APPROVAL_STATUS N	Y	Total
Female	37 6.16 33.04 19.79	75 12.48 66.96 18.12	112 18.64
Male	150 24.96 30.67 80.21	339 56.41 69.33 81.88	489 81.36
Total	187 31.11	414 68.89	601 100.00

Frequency Missing = 13

Missing values for this variable is 13



7.6.4 Bivariate Analysis on Categorical variable (MARITAL_STATUS) VS Categorical variable (FAMILY_MEMBERS)

7.6.4.1 Explanation

Bivariate analysis is performed between the Marital Status and Family Members variable, to analyse the number of family members each applicant that applies for loan has, according to their marital status.

7.6.4.2 SAS Codes

```
/* Bivariate Analysis on MARITAL STATUS (CATEGORICAL VARIABLE) & FAMILY MEMBERS (CATEGORICAL VAR

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE MARITAL_STATUS * FAMILY_MEMBERS /
PLOTS = FREQPLOT (TWOWAY = STACKED SCALE = GROUPPCT );
TITLE 'BIVARIATE ANALYSIS ON FAMILY MEMBERS VS MARITAL STATUS';
RUN;
```

7.6.4.3 Outputs/Results

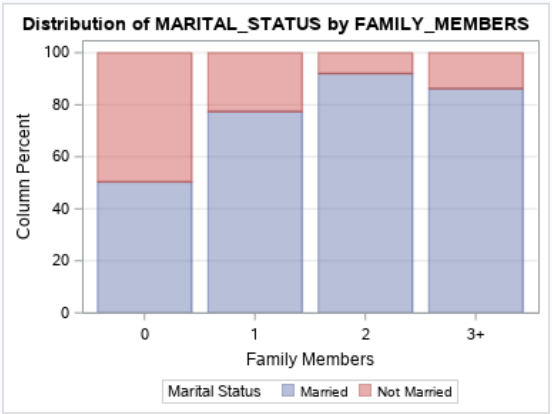
BIVARIATE ANALYSIS ON FAMILY MEMBERS VS MARITAL STATUS

The FREQ Procedure

Table of MARITAL_STATUS by FAMILY_MEMBERS

Frequency Percent Row Pct Col Pct	FAMILY_MEMBERS(Family Members)				
MARITAL_STATUS(Marital Status)	0	1	2	3+	Total
Married	174 29.05 44.62 50.43	79 13.19 20.26 77.45	93 15.53 23.85 92.08	44 7.35 11.28 86.27	390 65.11
Not Married	171 28.55 81.82 49.57	23 3.84 11.00 22.55	8 1.34 3.83 7.92	7 1.17 3.35 13.73	209 34.89
Total	345 57.60	102 17.03	101 16.86	51 8.51	599 100.00

Frequency Missing = 15



7.6.5 Bivariate Analysis on Categorical variable (GENDER) VS Categorical variable (QUALIFICATION)

7.6.5.1 Explanation

Bivariate analysis is performed on both Gender and Qualification variables, to analyse the qualification of each applicant via gender.

7.6.5.2 SAS Codes

```
/* Bivariate Analysis on GENDER (CATEGORICAL VARIABLE) & QUALIFICATION (CATEGORICAL VARIABLE) */

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE GENDER * QUALIFICATION /
PLOTS = FREQPLOT (TWOWAY = STACKED SCALE = GROUPPCT );
TITLE 'BIVARIATE ANALYSIS ON GENDER VS QUALIFICATION';
RUN;
```

7.6.5.3 Outputs/Results

BIVARIATE ANALYSIS ON GENDER VS QUALIFICATION

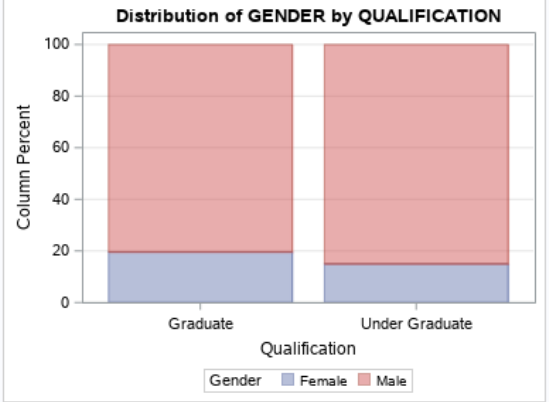
The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of GENDER by QUALIFICATION

GENDER(Gender)	QUALIFICATION(Qualification) Graduate	Under Graduate	Total
Female	92 15.31 82.14 19.66	20 3.33 17.86 15.04	112 18.64
Male	376 62.56 76.89 80.34	113 18.80 23.11 84.96	489 81.36
Total	468 77.87	133 22.13	601 100.00

Frequency Missing = 13



7.6.6 Bivariate Analysis on Categorical variable (GENDER) VS Categorical variable (LOAN_LOCATION)

7.6.6.1 Explanation

Bivariate analysis is performed between gender and loan_location variables, to investigate the location of each applicant by gender.

7.6.6.2 SAS Codes

```
/* Bivariate Analysis on GENDER (CATEGORICAL VARIABLE) & LOAN LOCATION (CATEGORICAL VARIABLE) */

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE GENDER * LOAN_LOCATION /
PLOTS = FREQPLOT (TWOWAY = STACKED SCALE = GROUPPCT );
TITLE 'BIVARIATE ANALYSIS ON GENDER VS LOAN_LOCATION';
RUN;
```

7.6.6.3 Outputs/Results

BIVARIATE ANALYSIS ON GENDER VS LOAN_LOCATION

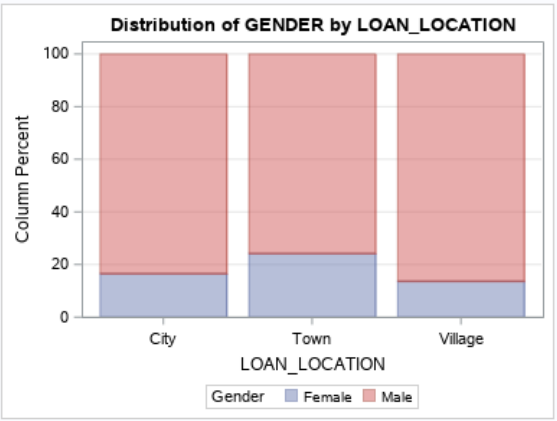
The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of GENDER by LOAN_LOCATION

GENDER(Gender)	LOAN_LOCATION(Loan Location LOAN_APPROVAL_STATUS Loan Approval Status) City	Town	Village	Total
Female	33 5.49 29.46 16.58	55 9.15 49.11 24.23	24 3.99 21.43 13.71	112 18.64
Male	166 27.62 33.95 83.42	172 28.62 35.17 75.77	151 25.12 30.88 86.29	489 81.36
Total	199 33.11	227 37.77	175 29.12	601 100.00

Frequency Missing = 13



7.6.7 Bivariate Analysis on Categorical variable (QUALIFICATION) VS Categorical variable (LOAN_APPROVAL_STATUS)

7.6.7.1 Explanation and Analysis

Bivariate analysis is performed between the Qualification and Loan Approval Status variables, to investigate the relationship of applications with graduate or undergraduate qualifications, and whether loan has been approved for these applicants.

7.6.7.2 SAS Codes

```
/* Bivariate Analysis on QUALIFICATION (CATEGORICAL VARIABLE) & LOAN APPROVAL STATUS (CATEGORICAL VARIABL

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE QUALIFICATION * LOAN_APPROVAL_STATUS /
PLOTS = FREQPLOT (TWOWAY = STACKED SCALE = GROUPPCT );
TITLE 'BIVARIATE ANALYSIS ON QUALIFICATION VS LOAN APPROVAL STATUS';
RUN;
```

7.6.7.3 Outputs/Results

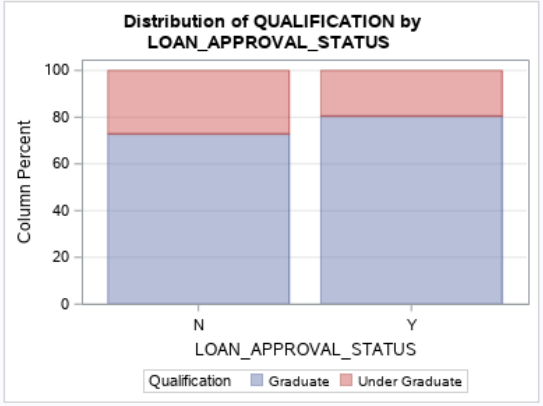
BIVARIATE ANALYSIS ON QUALIFICATION VS LOAN APPROVAL STATUS

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of QUALIFICATION by LOAN_APPROVAL_STATUS

QUALIFICATION(Qualification)	LOAN_APPROVAL_STATUS N	Y	Total
Graduate	140 22.80 29.17 72.92	340 55.37 70.83 80.57	480 78.18
Under Graduate	52 8.47 38.81 27.08	82 13.36 61.19 19.43	134 21.82
Total	192 31.27	422 68.73	614 100.00



7.6.8 Bivariate Analysis on Categorical variable (QUALIFICATION) VS Categorical variable (LOAN_LOCATION)

7.6.8.1 Explanation and Analysis

Bivariate analysis is performed between the qualification and loan location variables, to investigate the location of applicants with graduate or undergraduate qualifications.

7.6.8.2 SAS Codes

```
/* Bivariate Analysis on QUALIFICATION (CATEGORICAL VARIABLE) & LOAN LOAN LOCATION (CATEGORICAL VARIABLE)

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_BK;
TABLE QUALIFICATION * LOAN_LOCATION /
PLOTS = FREQPLOT (TWOWAY = STACKED SCALE = GROUPPCT );
TITLE 'BIVARIATE ANALYSIS ON QUALIFICATION VS LOAN LOCATION';
RUN;
```

7.6.8.3 Outputs/Results

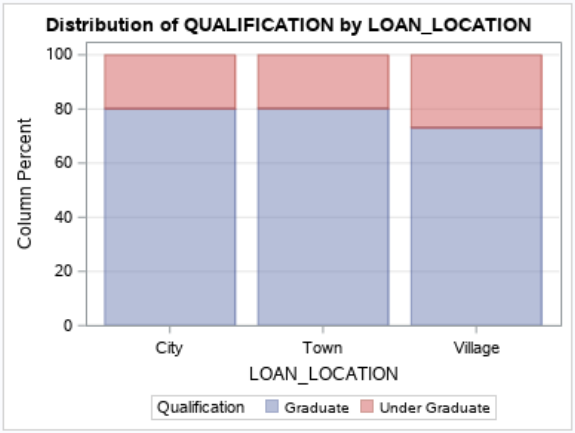
BIVARIATE ANALYSIS ON QUALIFICATION VS LOAN LOCATION

The FREQ Procedure

Table of QUALIFICATION by LOAN_LOCATION

LOAN_LOCATION(Loan Location LOAN_APPROVAL_STATUS Loan Approval Status)

Frequency Percent Row Pct Col Pct	City	Town	Village	Total
Graduate	162 26.38 33.75 80.20	187 30.46 38.96 80.26	131 21.34 27.29 73.18	480 78.18
Under Graduate	40 6.51 29.85 19.80	46 7.49 34.33 19.74	48 7.82 35.82 26.82	134 21.82
Total	202 32.90	233 37.95	179 29.15	614 100.00



7.6.9 Bivariate Analysis on FAMILY_MEMBERS (Categorical variable) Versus CANDIDATE_INCOME (Continuous variable)

7.6.9.1 Explanation and Analysis

Bivariate analysis is performed to investigate the relationship between the family members and candidate income variable. This is to find a causal link between the amount of family members an applicant has, along with their income.

7.6.9.2 SAS Codes

```
PROC MEANS DATA = MYLIB097.TRAINING_DS_TP063097_BK;
CLASS FAMILY_MEMBERS; /* CHAR */
VAR CANDIDATE_INCOME; /* NUMERIC */
TITLE 'BIVARIATE ANALYSIS ON FAMILY_MEMBERS (CATEGORICAL VARIABLE) VS. CANDIDATE_INCOME (CONTI
RUN;

PROC SGPLOT DATA = MYLIB097.TRAINING_DS_TP063097_BK;
VBOX CANDIDATE_INCOME / CATEGORY = FAMILY_MEMBERS;
/* FM X-AXIS CI Y-AXIS */
TITLE 'BIVARIATE ANALYSIS ON FAMILY_MEMBERS (CATEGORICAL VARIABLE) VS. CANDIDATE_INCOME (CONTI
RUN;
```

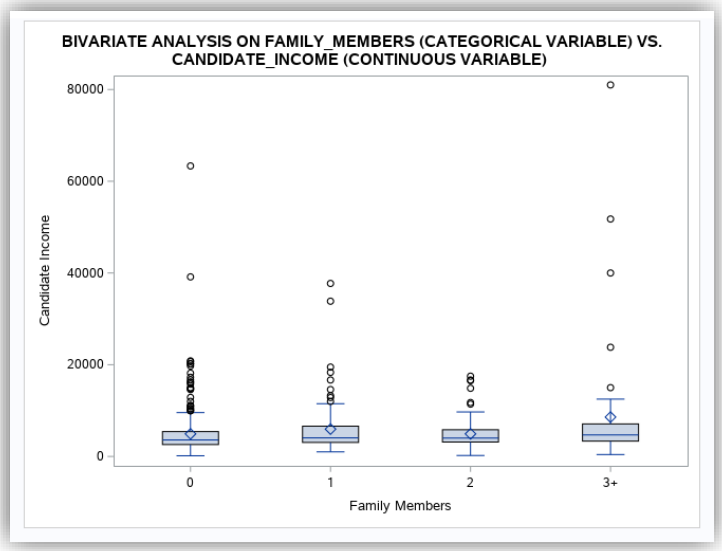
7.6.9.3 Outputs/Results

BIVARIATE ANALYSIS ON FAMILY_MEMBERS (CATEGORICAL VARIABLE) VS. CANDIDATE_INCOME (CONTINUOUS VARIABLE)

The MEANS Procedure

Analysis Variable : CANDIDATE_INCOME Candidate Income

Family Members	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	345	345	4917.42	5029.42	150.0000000	63337.00
1	102	102	5962.27	5587.40	1000.00	37719.00
2	101	101	4926.78	3153.83	210.0000000	17500.00
3+	51	51	8581.22	13603.94	416.0000000	81000.00



7.6.10 Bivariate Analysis of Loan Location vs Gender

7.6.10.1 SAS Code:

```
%MACRO_BIVATDS_TP97(MYLIB097.TESTING_DS_TP063097_BK, GENDER, LOAN_LOCATION, 'Figure of GE
```

7.6.10.2 Output/Results

Figure of GENDER VS. LOAN_LOCATION

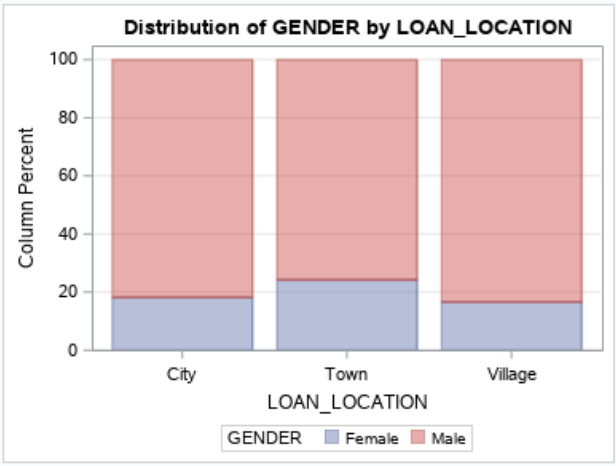
The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of GENDER by LOAN_LOCATION

GENDER	LOAN_LOCATION			
	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



7.7 Bivariate Analysis for the combination of variables (Categorical Variable vs Continuous Variable)

7.7.1 SAS MACRO Codes

```
/*************************************************************************************************
Bivariate Analysis on LOAN_APPROVAL_STATUS ( Categorical vatiable ) Versus CANDIDATE_INCOME  ( Continuous variable )
*************************************************************************************************/
%MACRO MACRO_BVA_CATCON_V(DATASET_NAME, CATE_VARIABLE,CONTI_VARIABLE,TITLE_1, TITLE_2);
PROC MEANS DATA =  &DATASET_NAME;
CLASS &CATE_VARIABLE;  /* CHAR*/
VAR &CONTI_VARIABLE; /* NUMERIC*/
TITLE &TITLE_1;
RUN;
PROC SGPLOT DATA =  &DATASET_NAME;
VBOX &CONTI_VARIABLE / CATEGORY=&CATE_VARIABLE;
/*FM X-AXIS CI Y-AXIS */
TITLE &TITLE_2;
RUN;
%MEND MACRO_BVA_CATCON_V;
```

7.7.1.1 SAS Codes

```
/* Bivariate analysis of combination of variables (categorical vs. numeric) */
/* The macro MACRO_BIVATDS_TP97 has parameters ()
pds_name - > Name of Data Set
pcate_variable_1 -> 1st variable name to be involved
pcate_variable_2 -> 2nd variable name to be involved
ptitle -> title to be displayed in the output
*************************************************************************/
%MACRO MACRO_BIVATDS_TP97(pds_name, pcate_variable_1, pconti_variable_2, ptitle);
PROC MEANS DATA = &pds_name;
CLASS &pcate_variable_1; /* CHAR */
VAR &pconti_variable_2; /* NUMERIC */
TITLE &ptitle;
RUN;

PROC SGPLOT DATA = &pds_name;
VBOX &pconti_variable_2 / CATEGORY = &pcate_variable_1;
TITLE &ptitle;
RUN;

%MEND MACRO_BIVATDS_TP97;

/*Call SAS Macro to do BIVARIATE ANALYSIS for the CV Gender vs. Candidate Income */

%MACRO_BIVATDS_TP97(MYLIB097.TESTING_DS_TP063097_BK, GENDER, CANDIDATE_INCOME, 'Figure of Bivat Analysis');
```

7.7.1.2 Explanation and Output/Results

SAS Macro can perform coding in a concise and simpler manner, to perform bivariate analysis between the selection of categorical and continuous variables.

Figure of Bivat Analysis

The MEANS Procedure

Analysis Variable : CANDIDATE_INCOME						
GENDER	N Obs	N	Mean	Std Dev	Minimum	Maximum
Female	70	70	4163.60	2644.87	0	14987.00
Male	286	286	4932.86	5186.56	0	72529.00

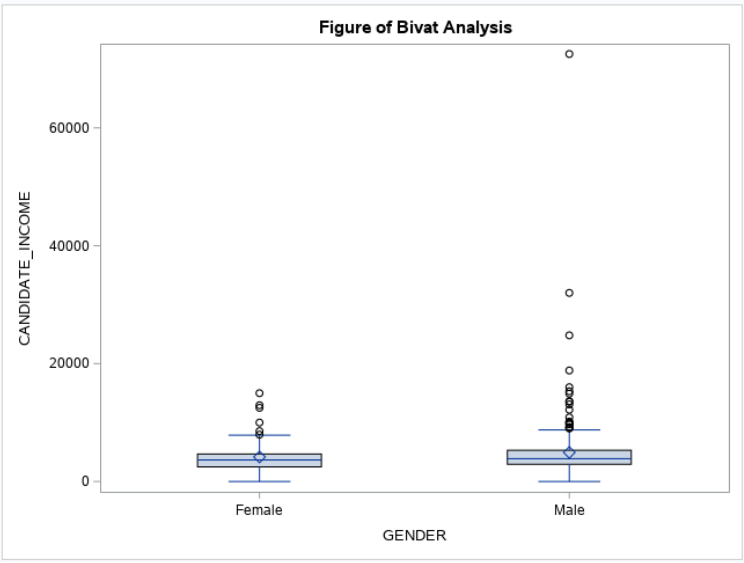
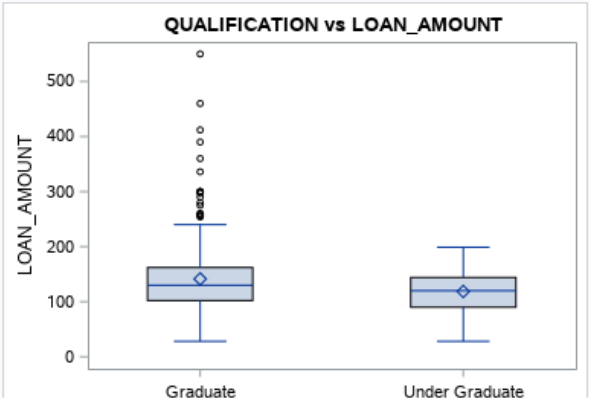


Figure 5.5

The MEANS Procedure

Analysis Variable : LOAN_AMOUNT						
QUALIFICATION	N Obs	N	Mean	Std Dev	Minimum	Maximum
Graduate	283	279	141.3584229	66.1702665	28.0000000	550.0000000
Under Graduate	84	83	118.5662651	36.4628755	28.0000000	199.0000000



7.8 Bivariate Analysis for the combination of continuous variable vs continuous variable (Candidate Income vs Guarantee Income)

7.8.1 SAS MACRO Codes

```
/*************************************************************************************************
Bivariate Analysis for the combination of Continuous vs Countinuous variable
*************************************************************************************************/

%MACRO BIVARIATE_CONTI_CONTI(DATASETNAME, VARIABLE_1, VARIABLE_2,TITLE_1);
PROC CORR DATA =  &DATASETNAME PLOTS = SCATTER;
VAR  &VARIABLE_1 &VARIABLE_2;
ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
TITLE1 &TITLE_1;
TITLE2 &VARIABLE_1 Vs &VARIABLE_2;
QUIT;
%MEND BIVARIATE_CONTI_CONTI;
```

```
/*Call MACRO to do Bivariate Analysis for the combination of Continuous variable (CANDIDATE_INCOME ) vs Countinuous variable(GUARANTEE_INCOME) '

%BIVARIATE_CONTI_CONTI(LIB77755.TRAINING_COPY_DS, CANDIDATE_INCOME, GUARANTEE_INCOME, "Bivariate Analysis for the combination of Continuous vs (
```

7.8.2 Outputs/Results

Bivariate Analysis for the combination of Continuous vs Countinuous variables
CANDIDATE_INCOME Vs GUARANTEE_INCOME

The CORR Procedure

2 Variables:	CANDIDATE_INCOME GUARANTEE_INCOME

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CANDIDATE_INCOME	614	5403	6109	3317724	150.00000	81000
GUARANTEE_INCOME	614	1621	2926	995445	0	41667

Pearson Correlation Coefficients, N = 614
Prob > |r| under H0: Rho=0

	CANDIDATE_INCOME	GUARANTEE_INCOME
CANDIDATE_INCOME	1.00000	-0.11660 0.0038
GUARANTEE_INCOME	-0.11660 0.0038	1.00000

Bivariate Analysis for the combination of Continuous vs Countinuous variables
CANDIDATE_INCOME Vs GUARANTEE_INCOME

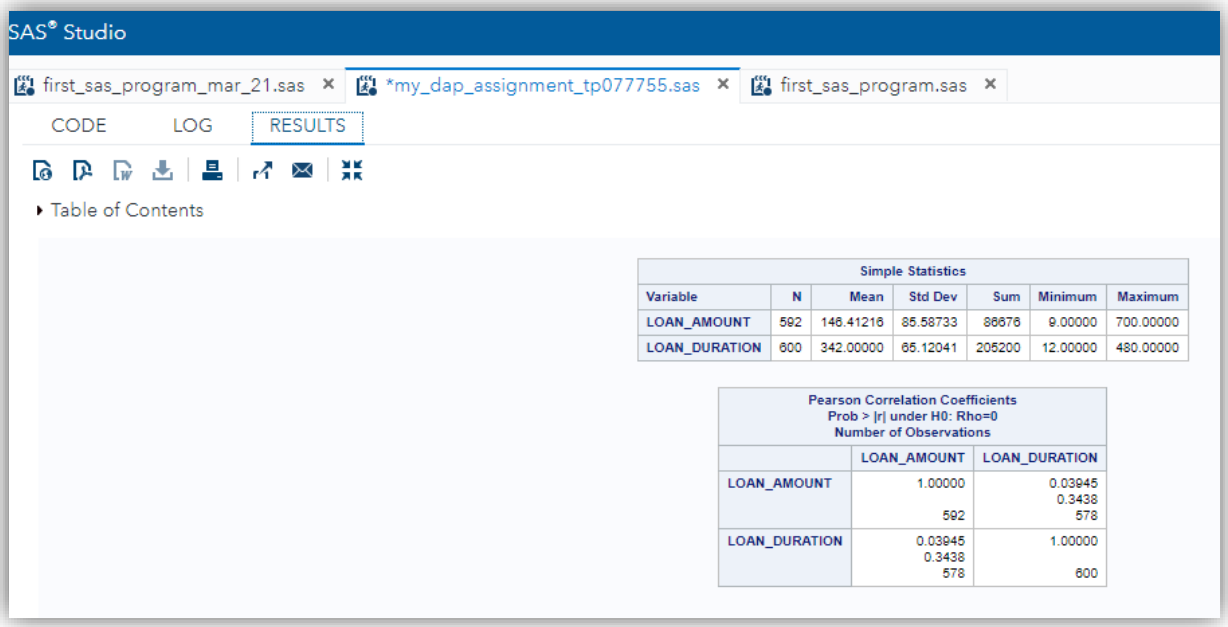
The CORR Procedure

7.9 Bivariate Analysis for the combination of continuous variable vs continuous variable (Loan Amount vs Loan Duration)

7.9.1 SAS Codes

```
/*Call MACRO to do Bivariate Analysis for the combination of Continuous variable (LOAN_AMOUNT ) vs Countinuous variable(LOAN_DURATION) */

%BIVARIATE_CONTI_CONTI(LIB77755.TRAINING_COPY_DS, LOAN_AMOUNT, LOAN_DURATION, "Bivariate Analysis for the combination of Continuous vs Countinuc
```

7.9.2 Outputs/Results



Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
LOAN_AMOUNT	592	146.41216	85.58733	86676	9.00000	700.00000
LOAN_DURATION	600	342.00000	65.12041	205200	12.00000	480.00000

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	LOAN_AMOUNT	LOAN_DURATION
LOAN_AMOUNT	1.00000 592	0.03945 0.3438 578
LOAN_DURATION	0.03945 0.3438 578	1.00000 600

CHAPTER 8. Imputing missing values found

8.1 Imputing missing values found in the variable GENDER

8.1.1 Explanation

Missing values are present in the dataset, as shown in the images and description earlier. It is pivotal that missing values are replaced (imputed) using SQL coding, as missing values will tamper with the final output. Missing values must be replaced and imputed with a viable number/name in the dataset. The process below depicts the process of imputation of missing values.

8.1.2 Make a copy of the TRAINING_DS_BK1

A new dataset (TRAINING_DS_TP063097_BK1) which is a copy of the previous dataset, is created.

8.1.3 SAS Codes and Outputs/Results

```
/* IMPUTE MISSING VALUES IN THE TRAINING DATASET *

STEP 1: MAKE A COPY OF THE TRAINING DATASET TO BE RENAMED AS BK1 FROM BK */

PROC SQL;
CREATE TABLE MYLIB097.TRAINING_DS_TP063097_BK1 AS
SELECT *FROM MYLIB097.TRAINING_DS_TP063097_BK;
QUIT;
```

Table: MYLIB097.TRAINING_DS_TP063097_BK1 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

Property	Value
Label	
Name	

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

	SME_LOAN_ID...	GEND...	MARITAL_STA...	FAMILY_MEMB...	QUALIFICATION	EMF
1	LP001002	Male	Not Married	0	Graduate	No
2	LP001003	Male	Married	1	Graduate	No
3	LP001005	Male	Married	0	Graduate	Yes
4	LP001006	Male	Married	0	Under Graduate	No
5	LP001008	Male	Not Married	0	Graduate	No
6	LP001011	Male	Married	2	Graduate	Yes
7	LP001013	Male	Married	0	Under Graduate	No
8	LP001014	Male	Married	3+	Graduate	No
9	LP001018	Male	Married	2	Graduate	No
10	LP001020	Male	Married	1	Graduate	No
11	LP001024	Male	Married	2	Graduate	No
12	LP001027	Male	Married	2	Graduate	
13	LP001028	Male	Married	2	Graduate	No
14	LP001029	Male	Not Married	0	Graduate	No
15	LP001030	Male	Married	2	Graduate	No
16	LP001032	Male	Not Married	0	Graduate	No

BEFORE IMPUTATION
ENTER

The FREQ Procedure

Gender				
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	112	18.64	112	18.64
Male	489	81.36	601	100.00
Frequency Missing = 13				

END OF ANALYSIS

DETAILS OF THE MISSING VALUES FOUND...

Loan Application No.	Gender	Marital Status	Family Members	Qualification	Employment	Candidate Income	Guaranteed Income	Loan Amount	Loan Duration	Loan History	Loan Location LOAN_APPROVAL_STATUS Loan Approval Status	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
LP001644		Married	0	Graduate	Yes	674	5296	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	.	Town	Y
LP002501		Married	0	Graduate	No	16692	0	110	360	1	Town	Y
LP002530		Married	2	Graduate	No	2873	1872	132	360	0	Town	N
LP002625		Not Married	0	Graduate	No	3583	0	96	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	292	360	1	Town	Y

```
/* STEP 4: TO FIND THE MOD G...*/

PROC SQL;

CREATE TABLE MYLIB097.TRAINING_DS_TP063097_BK2 AS
SELECT e.GENDER , COUNT (*) AS COUNTS FROM MYLIB097.TRAINING_DS_TP063097_BK1 e
WHERE (( e.gender IS NOT NULL) OR
        ( e.gender NE '' ) )
GROUP BY e.GENDER;

QUIT;
```

Columns

- Select all
- GENDER
- COUNTS

Total rows: 2 Total columns: 2 Rows 1-2

	GEND...	COUNTS
1	Female	112
2	Male	489

```
/* STEP 5: Display the details found in the dataset MYLIB097.TRAINING_DS_TP063097_BK2

PROC SQL;

TITLE 'TITLE 1';
TITLE2 ' TITLE 2';
FOOTNOTE 'END OF REPORT';

SELECT *
FROM MYLIB097.TRAINING_DS_TP063097_BK2;

QUIT;
```

TITLE 1
TITLE 2

Gender	COUNTS
Female	112
Male	489

END OF REPORT

```
QUIT;

SELECT * FROM MYLIB097.TRAINING_DS_TP063097_BK2;
CREATE TABLE MYLIB097.TRAINING_DS_COPY_GENDER AS

PROC SQL;

\* STEP 5: Copy of the dataset MYLIB097.TRAINING_DS_TP063097_BK2 is made *\
```

CODE LOG RESULTS OUTPUT DATA

Table: MYLIB097.TRAINING_DS_COPY_GENDER 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

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	EMP
1	LP001002	Male	Not Married	0	Graduate	No
2	LP001003	Male	Married	1	Graduate	No
3	LP001005	Male	Married	0	Graduate	Yes
4	LP001006	Male	Married	0	Under Graduate	No
5	LP001008	Male	Not Married	0	Graduate	No
6	LP001011	Male	Married	2	Graduate	Yes
7	LP001013	Male	Married	0	Under Graduate	No
8	LP001014	Male	Married	3+	Graduate	No
9	LP001018	Male	Married	2	Graduate	No
10	LP001020	Male	Married	1	Graduate	No
11	LP001024	Male	Married	2	Graduate	No
12	LP001027	Male	Married	2	Graduate	
13	LP001028	Male	Married	2	Graduate	No
14	LP001029	Male	Not Married	0	Graduate	No
15	LP001030	Male	Married	2	Graduate	No
16	LP001032	Male	Not Married	0	Graduate	No
17	LP001034	Male	Not Married	1	Under Graduate	No
18	LP001036	Female	Not Married	0	Graduate	No
19	LP001038	Male	Married	0	Under Graduate	No
20	LP001041	Male	Married	0	Graduate	
21	LP001043	Male	Married	0	Under Graduate	No

```
PROC SQL;

UPDATE MYLIB097.TRAINING_DS_COPY_GENDER
SET GENDER = ( SELECT eo.gender label = 'mod of gender'
                FROM MYLIB097.TRAINING_DS_TP063097_BK2 eo
                WHERE eo.counts EQ ( SELECT MAX (e.counts)
                                     FROM MYLIB097.TRAINING_DS_TP063097_BK2 e ) )
                WHERE ( ( gender IS NULL ) OR
                        ( gender EQ '') );

QUIT;
```

```
PROC FREQ DATA = MYLIB097.TRAINING_DS_COPY_GENDER ;

TABLE GENDER;
TITLE 'AFTER IMPUTATION';
TITLE2 'ENTER';
FOOTNOTE 'END OF ANALYSIS';

RUN;
```

AFTER IMPUTATION
ENTER

The FREQ Procedure

Gender				
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	112	18.24	112	18.24
Male	502	81.76	614	100.00

END OF ANALYSIS

8.2 Imputing missing values found in the variable FAMILY_MEMBERS

8.2.1 Explanation

Family Members variable contain missing values, as depicted below. Imputation is performed on this variable to replace missing values with the Mode (Frequently appeared value).

8.2.2 SAS Codes and Outputs/Results

```
/************************************
*STEP 1
************************************/

PROC SQL;
CREATE TABLE MYLIB097.TRAINING_DS_TP063097_FM AS
SELECT *FROM MYLIB097.TRAINING_DS_COPY_GENDER;
QUIT;

/* BEFORE IMPUTATION/CLEANSING OF THE MISSING VALUES FOUND IN THE CV FAMILY_MEMBERS

STEP 2: Perform univariate analysis on the FAMILY_MEMBERS variable before imputation */

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_FM ;

TABLE FAMILY_MEMBERS;
TITLE 'BEFORE IMPUTATION';
TITLE2 'ENTER';
FOOTNOTE 'END OF ANALYSIS';

RUN;

/* STEP 3: DETAILS OF THE MISSING VALUES FOUND... */

PROC SQL;

TITLE 'DETAILS OF THE MISSING VALUES FOUND...';
SELECT *
FROM MYLIB097.TRAINING_DS_TP063097_FM t
WHERE ( ( t.family_members IS NULL ) OR
        ( t.family_members EQ '') ) ;

QUIT;
```

```
/* STEP 4: TO FIND THE MOD G.... */

PROC SQL;

CREATE TABLE MYLIB097.TRAINING_DS_TP063097_FM2 AS
SELECT t.FAMILY_MEMBERS , COUNT (*) AS COUNTS FROM MYLIB097.TRAINING_DS_TP063097_FM t
WHERE (( t.family_members IS NOT NULL) OR
      ( t.family_members NE '' ) )
GROUP BY t.FAMILY_MEMBERS;

QUIT;

/* STEP 5: Display the details found in the dataset MYLIB097.TRAINING_DS_TP063097_FM2 (

PROC SQL;

TITLE 'TITLE 1';
TITLE2 ' TITLE 2';
FOOTNOTE 'END OF REPORT';

SELECT *
FROM MYLIB097.TRAINING_DS_TP063097_FM2;

QUIT;

/* STEP 7: Copy of the dataset MYLIB097.TRAINING_DS_TP063097_FM2 is made */

PROC SQL;

CREATE TABLE MYLIB097.TRAINING_DS_COPY_FM AS
SELECT * FROM MYLIB097.TRAINING_DS_TP063097_FM2;

QUIT;
```

```
/* STEP 8: IMPUTE MISSING VALUES FROM THE NEW COPY OF THE DATASET */

PROC SQL;

UPDATE MYLIB097.TRAINING_DS_TP063097_FM2
SET FAMILY_MEMBERS = ( SELECT to.family_members label = 'mod of family members'
                FROM MYLIB097.TRAINING_DS_COPY_FM to
                WHERE to.counts EQ ( SELECT MAX (t.counts)
                                     FROM MYLIB097.TRAINING_DS_COPY_FM t ) )
                WHERE ( ( family_members IS NULL ) OR
                        ( family_members EQ '') );

QUIT;

/* STEP 9: AFTER IMPUTATION */

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_FM2 ;

TABLE FAMILY_MEMBERS;
TITLE 'AFTER IMPUTATION';
TITLE2 'ENTER';
FOOTNOTE 'END OF ANALYSIS';

RUN;

/*
```

AFTER IMPUTATION
ENTER

The FREQ Procedure

Family Members				
FAMILY_MEMBERS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1	25.00	1	25.00
1	1	25.00	2	50.00
2	1	25.00	3	75.00
3+	1	25.00	4	100.00

END OF ANALYSIS

8.2 Imputing missing values found in the variable LOCATION

8.2.1 Explanation

Missing values for the variable LOCATION are imputed

8.2.2 SAS Codes

```
/*SAS Macro for Loan_Location */
%MYMACRO_UVA_TP97(MYLIB097.TESTING_DS_TP063097_BK, LOAN_LOCATION, 'Univariate Analysis', 'Figure 5: Loan Location');
```

8.3 Imputing missing values found in the Continuous variable LOAN_DURATION

8.3.1 Explanation

Missing values for the continuous variable LOAN_DURATION is imputed

8.3.2. SAS Codes

```
1058 /*STEP 1: Before imputing missing values found in the continuous variable - LOAN_DURATION */
1059
1060 PROC MEANS DATA =  LIB77755.TRAINING_COPY_DS_FI_GENDER_BI N NMISS MIN MAX MEAN MEDIAN STD;
1061 VAR LOAN_DURATION;
1062 TITLE "Fig 7.11 Univariate Analysis on ";
1063 TITLE2 'LOAN DURATION - Continuous variable';
1064 FOOTNOTE '--- End of Analyis ---';
1065 RUN;
```

Fig 7.11 Univariate Analysis on
LOAN DURATION - Continuous variable

The MEANS Procedure

Analysis Variable : LOAN_DURATION						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
600	14	12.0000000	480.0000000	342.0000000	360.0000000	65.1204099

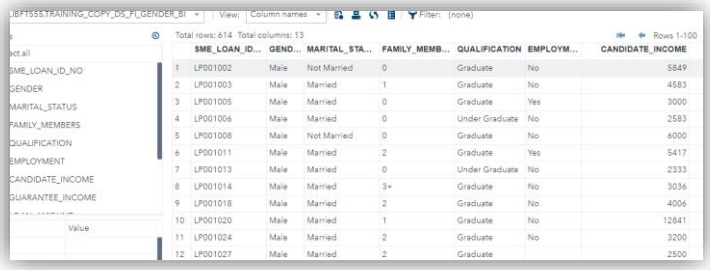
--- End of Analyis ---

Fig 7.11 Univariate Analysis on
LOAN DURATION - Continuous variable

The MEANS Procedure

Analysis Variable : LOAN_DURATION						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
614	0	12.0000000	480.0000000	342.0000000	360.0000000	64.3724886

--- End of Analyis ---



	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

8.4 Checking of any other missing values

```
1374 /* STEP 1: THU,6-MAY-2021 Check w
1375 Check whether the cleansed TRAINING dataset still has any missing values*/
1376
1377 PROC FREQ DATA=LIB77755.TRAINING_COPY_DS;
1378 TABLE
1379
1380 LOAN_AMOUNT
1381 FAMILY_MEMBERS
1382 CANDIDATE_INCOME
1383 GUARANTEE_INCOME
1384 LOAN_DURATION
1385 LOAN_HISTORY
1386 LOAN_LOCATION
1387 MARITAL_STATUS
1388 QUALIFICATION
1389 LOAN_LOCATION
1390 GENDER;
1391
1392 RUN;
```

FAMILY_MEMBERS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	345	57.60	345	57.60
1	102	17.03	447	74.62
2	101	16.86	548	91.49
3+	51	8.51	599	100.00
Frequency Missing = 15				

FAMILY_MEMBERS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	200	56.02	200	56.02
1	58	16.25	258	72.27
2	59	16.53	317	88.80
3+	40	11.20	357	100.00
Frequency Missing = 10				

Family_Members variable still has missing values.

8.5 Imputation for missing values of FAMILY_MEMBERS

8.5.1 Explanation

Based on the earlier analysis, Family_Members variable still has missing values, for which imputation is needed to remove missing values.

8.5.2 SAS Codes

```
/* STEP 9: AFTER IMPUTATION */

PROC FREQ DATA = MYLIB097.TRAINING_DS_TP063097_FM ;

TABLE FAMILY_MEMBERS;
TITLE 'AFTER IMPUTATION';
TITLE2 'ENTER';
FOOTNOTE 'END OF ANALYSIS';

RUN;

/*Number of observations after imputation */

PROC SQL;

TITLE 'No. of observations after imputation';
SELECT COUNT (*) LABEL = 'NO. OF OBSERVATIONS'
FROM MYLIB097.TRAINING_DS_TP063097_FM t
WHERE (( t.family_members EQ '') OR
            (t.family_members IS NULL ) );
```

This process imputes the continuous variable of Family_Members missing values.

CHAPTER 9. TESTING_DS Dataset Analysis

9.1 Create a copy of the TESTING_DS Dataset

9.1.1 SAS Codes

```
/* 9.1 Create a copy of the TESTING_DS dataset */

/*************************************************************************************************/

TITLE 'COPY OF THE TESTING DATASET'
TITLE2 'TESTING DATASET';
PROC SQL;
CREATE TABLE MYLIB097.TESTING_DS_TP063097_BK AS
SELECT *FROM MYLIB097.TESTING_DS;

QUIT;
```

9.1.2 Outputs/Results

Table: MYLIB097.TESTING_DS_TP063097_BK | 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: 367 Total columns: 13 — Rows 1-100

	SME_LOAN_ID...	GEND...	MARITAL_STA...	FAMILY_MEMB...	QUALIFICATION	EMPLOYM...	CANDI
1	LP001015	Male	Married	0	Graduate	No	
2	LP001022	Male	Married	1	Graduate	No	
3	LP001031	Male	Married	2	Graduate	No	
4	LP001035	Male	Married	2	Graduate	No	
5	LP001051	Male	Not Married	0	Under Graduate	No	
6	LP001054	Male	Married	0	Under Graduate	Yes	
7	LP001055	Female	Not Married	1	Under Graduate	No	
8	LP001056	Male	Married	2	Under Graduate	No	
9	LP001059	Male	Married	2	Graduate		
10	LP001067	Male	Not Married	0	Under Graduate	No	
11	LP001078	Male	Not Married	0	Under Graduate	No	
12	LP001082	Male	Married	1	Graduate		
13	LP001083	Male	Not Married	3+	Graduate	No	
14	LP001094	Male	Married	2	Graduate		
15	LP001096	Female	Not Married	0	Graduate	No	
16	LP001099	Male	Not Married	1	Graduate	No	
17	LP001105	Male	Married	2	Graduate	No	
18	LP001107	Male	Married	3+	Graduate	No	
19	LP001108	Male	Married	0	Graduate	No	
20	LP001115	Male	Not Married	0	Graduate	No	
21	LP001121	Male	Married	1	Under Graduate	No	
22	LP001124	Female	Not Married	3+	Under Graduate	No	
23	LP001128		Not Married	0	Graduate	No	

9.2 Univariate Analysis on the variables found in the MYLIB097.TESTING_DS_TP063097 _BK dataset

9.2.1 SAS Codes/Macro

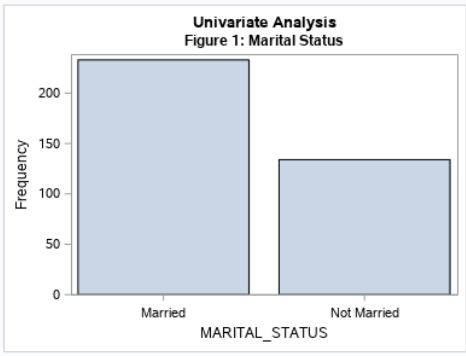
```
/* Univariate Analysis on the variables found in the MYLIB097.TESTING_DS_TP060397_BK. */

%MACRO MYMACRO_UVA_TP97(pds_name, pcate_variable_1, ptitle_1,ptitle_2);
TITLE &ptitle_1;
TITLE2 &ptitle_2;
PROC FREQ DATA = &pds_name;
TABLE &pcate_variable_1;
RUN;
ODS GRAPHICS / RESET WIDTH = 4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = &pds_name;
VBAR &pcate_variable_1;
RUN;
%MEND MYMACRO_UVA_TP97;

/*Call SAS Macro to perform Univariate Analysis for the cat. var. marital_status */

%MYMACRO_UVA_TP97(MYLIB097.TESTING_DS_TP063097_BK, marital_status,'Univariate Analysis', 'Figure 1: Marital Status')

/*SAS Macro for gender */
%MYMACRO_UVA_TP97(MYLIB097.TESTING_DS_TP063097_BK, gender, 'Univariate Analysis', 'Figure 2: Gender');

/*SAS Macro for qualification */
%MYMACRO_UVA_TP97(MYLIB097.TESTING_DS_TP063097_BK, QUALIFICATION, 'Univariate Analysis', 'Figure 3: Qualification');

/*SAS MACRO for employment */
%MYMACRO_UVA_TP97(MYLIB097.TESTING_DS_TP063097_BK, EMPLOYMENT, 'Univariate Analysis', 'Figure 4: Employment');

/*SAS Macro for Loan_Location */
%MYMACRO_UVA_TP97(MYLIB097.TESTING_DS_TP063097_BK, LOAN_LOCATION, 'Univariate Analysis', 'Figure 5: Loan Location');

/*SAS Macro for FAMILY_MEMBERS */

%MYMACRO_UVA_TP97(MYLIB097.TESTING_DS_TP063097_BK, FAMILY_MEMBERS, 'Univariate Analysis', 'Figure 6: Family Members'
```

9.2.2 Outputs/Results

Univariate Analysis
Figure 1: Marital Status

The FREQ Procedure

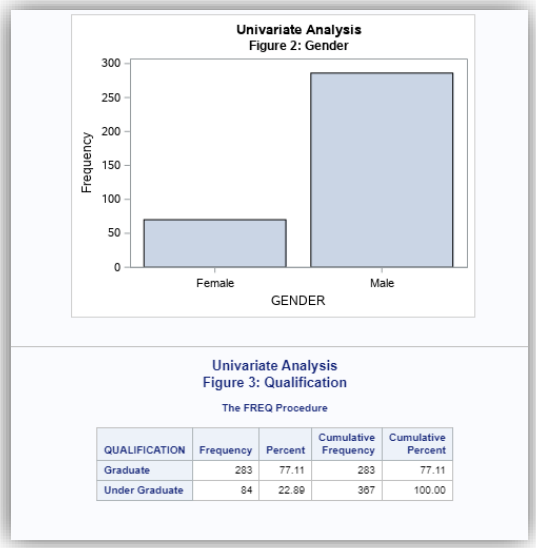
MARITAL_STATUS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Married	233	63.49	233	63.49
Not Married	134	36.51	367	100.00



Univariate Analysis
Figure 2: Gender

The FREQ Procedure

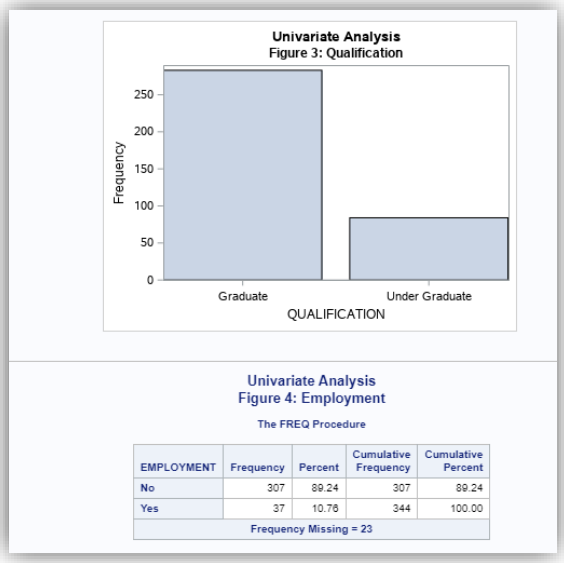
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	70	19.66	70	19.66
Male	286	80.34	356	100.00
Frequency Missing = 11				



Univariate Analysis
Figure 3: Qualification

The FREQ Procedure

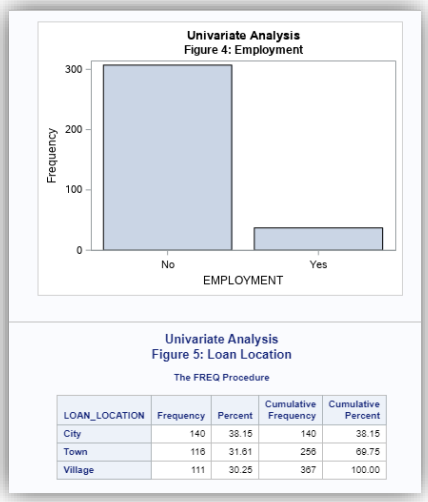
QUALIFICATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Graduate	283	77.11	283	77.11
Under Graduate	84	22.89	367	100.00



Univariate Analysis
Figure 4: Employment

The FREQ Procedure

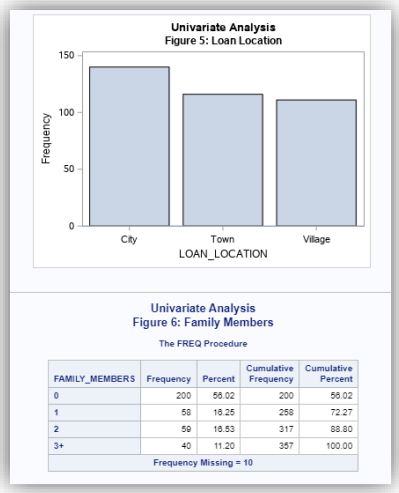
EMPLOYMENT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	307	89.24	307	89.24
Yes	37	10.76	344	100.00
Frequency Missing = 23				



Univariate Analysis
Figure 5: Loan Location

The FREQ Procedure

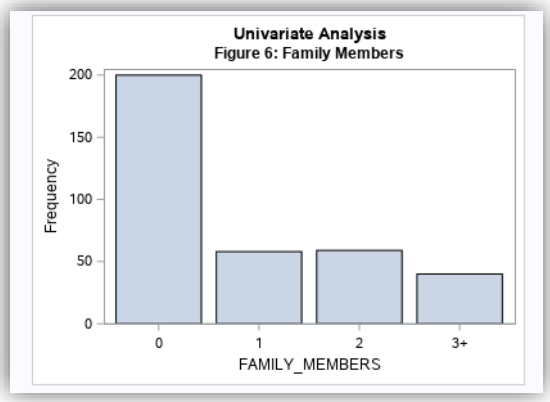
LOAN_LOCATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
City	140	38.15	140	38.15
Town	116	31.61	256	69.75
Village	111	30.25	367	100.00



Univariate Analysis
Figure 6: Family Members

The FREQ Procedure

FAMILY_MEMBERS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	200	56.02	200	56.02
1	58	16.25	258	72.27
2	59	16.53	317	88.80
3+	40	11.20	357	100.00
Frequency Missing = 10				



9.3 SAS Macro to perform Univariate Analysis on the Continuous Variable found in the TESTING_DS_TP063097_BK dataset

9.3.1 Explanation

Univariate Analysis is performed on all the continuous variables in the TESTING_DS dataset, before performing the same univariate analysis on the categorical variables. Since univariate analysis was already performed on the TRAINING_DS datasets earlier, using SAS Macro, now the process of performing univariate analysis on the TESTING_DS dataset can be expedited and automated. SAS Macro is beneficial in automating the process of univariate analysis, using the variable names that is the same for the TRAINING_DS dataset.

9.3.2 SAS Macro

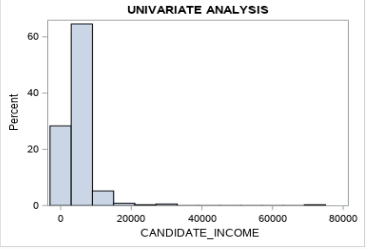
```
/*9.2 Univariate Analysis on the continuous variables found in the Testing dataset */

%MACRO MYMACRO_UVACONTIV_TP97(pds_name, pconti_variable_1, ptitle_1,ptitle_2);
PROC MEANS DATA = &pds_name N NMISS MIN MAX MEAN MEDIAN STD;
VAR &pconti_variable_1;
TITLE &ptitle_1;
RUN;

ODS GRAPHICS / RESET WIDTH=4.0 IN HEIGHT=3.0 IN IMAGEMAP;
PROC SGPLOT DATA = &pds_name;
HISTOGRAM &pconti_variable_1;
TITLE &ptitle_1;
RUN;
%MEND MYMACRO_UVACONTIV_TP97;

/*Call SAS Macro to do the Univariate Analysis on Candidate Income */
%MYMACRO_UVACONTIV_TP97 (MYLIB097.TESTING_DS_TP063097_BK, CANDIDATE_INCOME, 'UNIVARIATE ANALYSIS', 'FIGURE 1: CANDIATE INCOME');

/*CALL SAS MACRO TO DO UNIV. ANALYSIS FOR GUARANTEE INCOME */
%MYMACRO_UVACONTIV_TP97 (MYLIB097.TESTING_DS_TP063097_BK, GUARANTEE_INCOME, 'UNIVARIATE ANALSYSIS', 'FIGURE 2: GUARANTEE INCOME');

/*CALL SAS MACRO TO DO UNIV. ANALAYSIS FOR LOAN AMOUNT */
%MYMACRO_UVACONTIV_TP97 (MYLIB097.TESTING_DS_TP063097_BK, LOAN_AMOUNT, 'UNIVARIATE ANALYSIS', 'FIGURE 3: LOAN AMOUNT');

/*CALL SAS MACRO TO DO UNIV. ANALYSIS FOR LOAN DURATION */

%MYMACRO_UVACONTIV_TP97 (MYLIB097.TESTING_DS_TP063097_BK, LOAN_DURATION, 'UNIVARIATE ANALYSIS', 'FIGURE 4: LOAN DURATION');
```

9.3.3 Outputs/Results

UNIVARIATE ANALYSIS

The MEANS Procedure

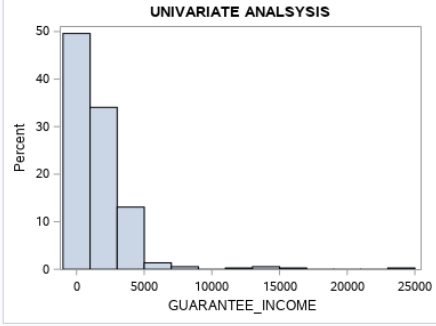
Analysis Variable : CANDIDATE_INCOME						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
367	0	0	72529.00	4805.60	3786.00	4910.69



UNIVARIATE ANALSYSIS

The MEANS Procedure

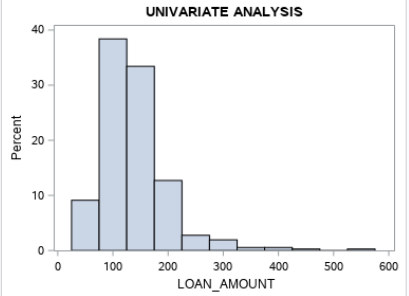
Analysis Variable : GUARANTEE_INCOME						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
367	0	0	24000.00	1569.58	1025.00	2334.23



UNIVARIATE ANALYSIS

The MEANS Procedure

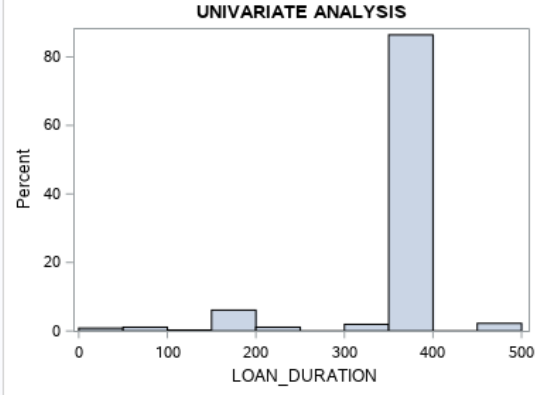
Analysis Variable : LOAN_AMOUNT						
N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
362	5	28.0000000	550.0000000	136.1325967	125.0000000	61.3666524



UNIVARIATE ANALYSIS

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



9.4 Impute missing values in the TESTING_DS_TP063097 dataset

9.4.1 SAS Codes

```
/*STEP 1: CREATE A COPY OF THE MYLIB097.TESTING_DS_TP063097_BK AS MYLIB097.TESTING_DS_TP063097_BK1 */

PROC SQL;

CREATE TABLE MYLIB097.TESTING_DS_TP063097_GENDER AS
SELECT * FROM MYLIB097.TESTING_DS_TP063097_BK;

QUIT;

/* STEP 2: BEFORE IMPUTATION: LIST THE OBSERVATIONS WITH MISSING VALUES */

TITLE 'BEFORE IMPUTATION';
TITLE2 'MISSING VALUES';
PROC SQL;
SELECT *
FROM MYLIB097.TESTING_DS_TP063097_GENDER t
WHERE ((t.gender EQ '') OR
       ( t.gender IS NULL ) ) ;

QUIT;

TITLE 'LIST THE MISSING VALUES';
PROC SQL;
SELECT COUNT (*) LABEL = 'NO OF OBSERVATIONS'
FROM MYLIB097.TESTING_DS_TP063097_GENDER t
WHERE (( t.gender EQ '') OR
       ( t.gender IS NULL ) );

QUIT;
```

```
618 /* STEP 3: LIST OBSERVATIONS WITH MISSING VALUES */
619
620 PROC SQL;
621
622 CREATE TABLE MYLIB097.TESTING_DS_TP063097_GENDER_N AS
623 SELECT t.gender AS GENDER_NAME, COUNT(*) AS TOTAL_COUNTS
624 FROM MYLIB097.TESTING_DS_TP063097_GENDER t
625 WHERE (( t.gender NE '') OR
626        (t.gender IS NOT NULL ) )
627 GROUP BY t.gender;
628 QUIT;
629
630 /* STEP 4: LIST GENDER COUNTS */
631
632 PROC SQL;
633 SELECT *
634 FROM MYLIB097.TESTING_DS_TP063097_GENDER_N t;
635 QUIT;
636
637 /* STEP 5: MAXIMUM TOTAL COUNTS OF GENDER */
638
639 PROC SQL;
640 SELECT MAX (t.total_counts)
641 FROM MYLIB097.TESTING_DS_TP063097_GENDER_N t;
642 QUIT;
643
644 /* STEP 6: To check if earlier step shows GENDER as FEMALE or MALE*/
645
646 PROC SQL;
647 SELECT to.gender_name
648 FROM MYLIB097.TESTING_DS_TP063097_GENDER_N to
649 WHERE to.total_counts EQ ( SELECT MAX (ti.total_counts)
650                            FROM MYLIB097.TESTING_DS_TP063097_GENDER_N ti);
651 QUIT;
```

```
653 /*STEP 7: impute missing values in GENDER*/
654
655 PROC SQL;
656 UPDATE MYLIB097.TESTING_DS_TP063097_GENDER
657 SET GENDER = ( SELECT to.gender_name
658                FROM MYLIB097.TESTING_DS_TP063097_GENDER_N to
659                WHERE to.total_counts EQ ( SELECT MAX (ti.total_counts)
660                                           FROM MYLIB097.TESTING_DS_TP063097_GENDER_N ti ) )
661 WHERE ( ( gender EQ '') OR
662         ( gender IS NULL ) );
663 QUIT;
664
665 /* STEP 8: AFTER IMPUTATION: LIST THE OBSERVATIONS */
666
667 TITLE 'AFTER IMPUTATION';
668 PROC SQL;
669 SELECT *
670 FROM MYLIB097.TESTING_DS_TP063097_GENDER t
671 WHERE ((t.gender EQ '') OR
672        ( t.gender IS NULL ) ) ;
673
674 QUIT;
675
676 PROC SQL;
677 SELECT COUNT (*) LABEL = 'NO OF OBSERVATIONS'
678 FROM MYLIB097.TESTING_DS_TP063097_GENDER t
679 WHERE (( t.gender EQ '') OR
680        ( t.gender IS NULL ) );
681
682 QUIT;
```

9.4.2 Outputs/Results

AFTER IMPUTATION

AFTER IMPUTATION

NO OF OBSERVATIONS
0

This step was repeated for the remaining variables. The data scientist has completed the univariate analysis for the numerical variables, of which the missing values has been imputed.

CHAPTER 10. BUILDING A LOGISTIC REGRESSION MODEL

10.1 Explanation

Logistic regression modelling can be used for predictive modelling, and to investigate the relationship between variables. Since the aim of this assignment is to predict the loan approval status of an applicant that applies for a bank loan based on certain criteria, regression model is the technique used for prediction. Logistic regression is used for the categorical variables present in the dataset (Linear regression can only be used for continuous variables).

10.2 SAS Codes

```
/*BUILD A LOGISTIC REGRESSION MODEL */

PROC LOGISTIC DATA=MYLIB097.TRAINING_DS_TP063097_FM OUTMODEL=MYLIB097.TRAINING_DS_TP063097_FM_MODEL;
CLASS

LOAN_LOCATION
MARITAL_STATUS
QUALIFICATION
LOAN_LOCATION
;

/*ABOVE ARE CATEGORICAL VARIABLES */
MODEL LOAN_APPROVAL_STATUS =

/* LOAN_APPLICATION_STATUS IS A DEPENDENT VARIABLE */

LOAN_AMOUNT
LOAN_DURATION
LOAN_HISTORY
MARITAL_STATUS
QUALIFICATION
GUARANTEE_INCOME
LOAN_LOCATION
;

OUTPUT OUT = MYLIB097.TRAINING_DS_TP063097_FM P = PRED_PROB;
/*PRED_PROB -> PREDICTED PROBABILITY - VARIABLE TO HOLD PREDICTED PROBABILITY
OUT -> THE OUTPUT WILL BE STORED IN THE DATASET
AKA INFORMATION CRITERIA MUST ( AIX ) < SC (SCHWARZ CRITERION)
*/
RUN;
```

10.3 Outputs/Results:

The LOGISTIC Procedure

Model Information		
Data Set	MYLIB097.TRAINING_DS_TP063097_FM	Predicted Values and Diagnostic Statistics
Response Variable	LOAN_APPROVAL_STATUS	
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	614
Number of Observations Used	547

Response Profile		
Ordered Value	LOAN_APPROVAL_STATUS	Total Frequency
1	N	173
2	Y	374

Probability modeled is LOAN_APPROVAL_STATUS='N'.

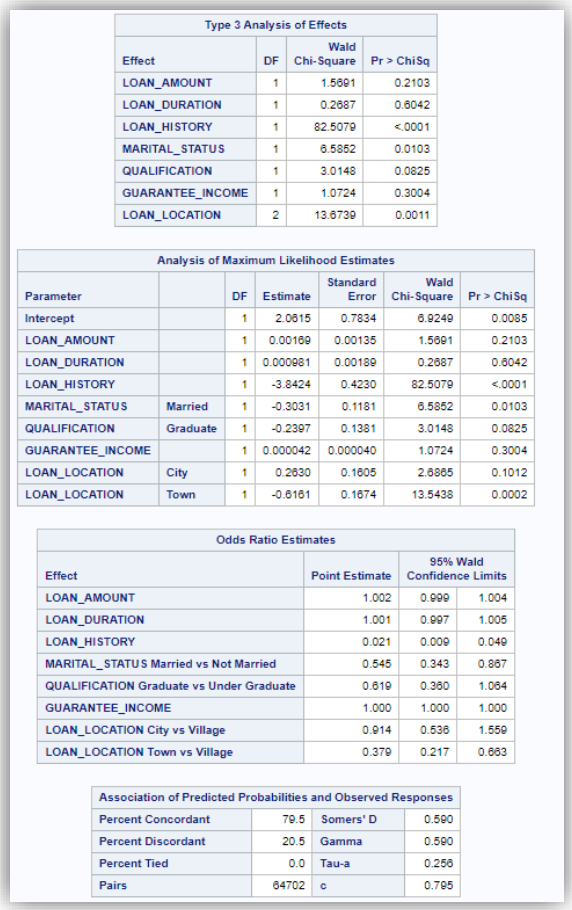
Note: 67 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
LOAN_LOCATION	City	1	0
	Town	0	1
	Village	-1	-1
MARITAL_STATUS	Married	1	
	Not Married	-1	
QUALIFICATION	Graduate	1	
	Under Graduate	-1	

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	684.685	515.709
SC	688.989	554.449
-2 Log L	682.685	497.709

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	184.9760	8	<.0001
Score	180.6855	8	<.0001
Wald	94.6481	8	<.0001



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
LOAN_AMOUNT	1	1.5691	0.2103
LOAN_DURATION	1	0.2687	0.6042
LOAN_HISTORY	1	82.5079	<.0001
MARITAL_STATUS	1	6.5852	0.0103
QUALIFICATION	1	3.0148	0.0825
GUARANTEE_INCOME	1	1.0724	0.3004
LOAN_LOCATION	2	13.6739	0.0011

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.0615	0.7834	6.9249	0.0085
LOAN_AMOUNT		1	0.00169	0.00135	1.5691	0.2103
LOAN_DURATION		1	0.000981	0.00189	0.2687	0.6042
LOAN_HISTORY		1	-3.8424	0.4230	82.5079	<.0001
MARITAL_STATUS	Married	1	-0.3031	0.1181	6.5852	0.0103
QUALIFICATION	Graduate	1	-0.2397	0.1381	3.0148	0.0825
GUARANTEE_INCOME		1	0.000042	0.000040	1.0724	0.3004
LOAN_LOCATION	City	1	0.2630	0.1605	2.6865	0.1012
LOAN_LOCATION	Town	1	-0.6161	0.1674	13.5438	0.0002

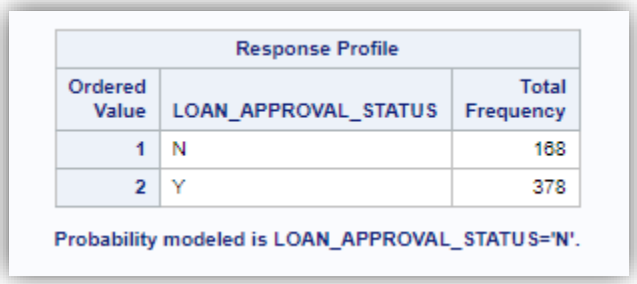
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
LOAN_AMOUNT	1.002	0.999	1.004
LOAN_DURATION	1.001	0.997	1.005
LOAN_HISTORY	0.021	0.009	0.049
MARITAL_STATUS Married vs Not Married	0.545	0.343	0.867
QUALIFICATION Graduate vs Under Graduate	0.619	0.360	1.064
GUARANTEE_INCOME	1.000	1.000	1.000
LOAN_LOCATION City vs Village	0.914	0.536	1.559
LOAN_LOCATION Town vs Village	0.379	0.217	0.663

Association of Predicted Probabilities and Observed Responses

Percent Concordant	79.5	Somers' D	0.590
Percent Discordant	20.5	Gamma	0.590
Percent Tied	0.0	Tau-a	0.256
Pairs	64702	c	0.795

The results show the Wald Confidence limits of the categorical variables in the dataset. The association of predicted probabilities and observed responses are of accepted levels. The odds ratio measures the probability of YES over NO for Loan Approval Status, as shown below:



Response Profile

Ordered Value	LOAN_APPROVAL_STATUS	Total Frequency
1	N	168
2	Y	378

Probability modeled is LOAN_APPROVAL_STATUS='N'.

The probability modeled is for LOAN_APPROVAL_STATUS of NO. NO (N) as a total frequency of 168 (168 loan approvals for applicants got rejected) while YES (Y) has a total frequency of 378 (378 loan approvals for applicants got approved) in the logistic regression model.

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

The model convergence status, using the convergence criterion, is satisfied. This means that the prediction model is at an acceptable level and can be used for prediction of loan approval for applicants, for this dataset.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
LOAN_LOCATION	2	9.4038	0.0091
GENDER	1	0.2720	0.6020
MARITAL_STATUS	1	5.9147	0.0150
QUALIFICATION	1	4.8387	0.0278
LOAN_AMOUNT	1	1.6401	0.2003
CANDIDATE_INCOME	1	0.2119	0.6453
EMPLOYMENT	1	0.0312	0.8599
GUARANTEE_INCOME	1	1.2157	0.2702

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.9111	0.2350	15.0331	0.0001
LOAN_LOCATION	City	1	0.1209	0.1343	0.8108	0.3679
LOAN_LOCATION	Town	1	-0.4048	0.1352	8.9613	0.0028
GENDER	Female	1	0.0686	0.1315	0.2720	0.6020
MARITAL_STATUS	Married	1	-0.2587	0.1064	5.9147	0.0150
QUALIFICATION	Graduate	1	-0.2505	0.1139	4.8387	0.0278
LOAN_AMOUNT		1	0.00179	0.00140	1.6401	0.2003
CANDIDATE_INCOME		1	9.318E-6	0.000020	0.2119	0.6453
EMPLOYMENT	No	1	-0.0243	0.1379	0.0312	0.8599
GUARANTEE_INCOME		1	0.000042	0.000038	1.2157	0.2702

The equation for logistic regression is:

Loan Approval Status = $-0.9111 + 0.1209_{\text{LOAN\_LOCATION(CITY)}} - 0.4048_{\text{LOAN\_LOCATION(TOWN)}} + 0.0686_{\text{GENDER(FEMALE)}} - 0.2587_{\text{MARITAL\_STATUS(MARRIED)}} - 0.2505_{\text{QUALIFICATION(GRADUATE)}} + 0.00179_{\text{LOAN\_AMOUNT}} + 9.318e^{-6}{}_{\text{CANDIDATE\_INCOME}} - 0.0243_{\text{EMPLOYMENT(NO)}} + 4.2e^{-5}$

GUARANTEE_INCOME

10.4 Final Outputs/Results:

```
PROC LOGISTIC INMODEL=MYLIB097.TRAINING_DS_TP063097_FM_MODEL; /*THE MODEL CREATED */
SCORE DATA=MYLIB097.TESTING_DS_TP063097_LH /*THE TESTING DATASSET */
OUT=MYLIB097.TESTING_DS_PREDICTION; /*LOCATION OF OUTPUT */
QUIT;
```

```
1472 /* To display the approval status of the loan applications found in the given TESTING Dataset */
1473
1474 PROC SQL;
1475
1476 SELECT SME_LOAN_ID_NO LABEL = 'LOAN ID',
1477 I_LOAN_APPROVAL_STATUS LABEL = 'LOAN APPROVAL STATUS',
1478 P_N,P_N LABEL = 'Probability_Entry',
1479 P_Y,P_N LABEL = 'Probability_Exit'
1480 FROM LIB77755.TESTING_COPY_DS_NEW_PREDICTIONS;
1481
1482 RUN;
```

LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	Into: LOAN_APPROVAL_STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
1	City			Y	0.176998	0.823002
1	City			Y	0.195789	0.804211
1	City			Y	0.224311	0.775689
1	City			Y	0.193893	0.806107
1	City			Y	0.379244	0.620756
1	City			Y	0.309813	0.690187
1	Town			Y	0.196465	0.803535
0	Village			N	0.951365	0.048635
1	City			Y	0.204714	0.795286
1	Town			Y	0.235126	0.764874
1	City			Y	0.386239	0.613761
1	Town			Y	0.099279	0.900721
1	City			Y	0.22806	0.77194
0	Town			N	0.812965	0.187035
1	Town			Y	0.146206	0.853794
1	City			Y	0.294111	0.705889
1	City			Y	0.229074	0.770926
1	Town			Y	0.087075	0.912925
1	City			Y	0.295834	0.704166
1	Town			Y	0.140856	0.859144
1	City			Y	0.25012	0.74988
1	City			Y	0.321373	0.678627
1	City			Y	0.284755	0.715245
1	City			Y	0.403143	0.596857
1	City			Y	0.275402	0.724598
0	Village			N	0.982465	0.017535
1	City			Y	0.19678	0.80322
1	City			Y	0.25904	0.74096
1	Town			Y	0.076677	0.923323
1	City			Y	0.283863	0.716137
1	Town			Y	0.201738	0.798262
1	City			Y	0.20946	0.79054

The image above shows the loan_approval_status of all the applicants, based on the logistic regression model. Y and N indicate the prediction of approval status for all the applicants in the dataset.

```
PROC REPORT DATA=MYLIB097.TESTING_DS_PREDICTION NOWINDOWS;
BY SME_LOAN_ID_NO;
DEFINE SME_LOAN_ID_NO / GROUP 'LOAN ID';
DEFINE I_LOAN_APPROVAL_STATUS / GROUP 'APPROVAL STATUS';
FOOTNOTE '-----END OF REPORT----';
RUN;
```

N_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	APPROVAL STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
360	1	City			Y	0.1769982	0.8230018

N_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	APPROVAL STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
360	1	City			Y	0.1957891	0.8042109

N_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	APPROVAL STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
360	1	City			Y	0.2243109	0.7756891

N_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	APPROVAL STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
360	1	City			Y	0.193893	0.806107

N_DURATION	LOAN_HISTORY	LOAN_LOCATION	LOAN_APPROVAL_STATUS	From: LOAN_APPROVAL_STATUS	APPROVAL STATUS	Predicted Probability: LOAN_APPROVAL_STATUS=N	Predicted Probability: LOAN_APPROVAL_STATUS=Y
360	1	City			Y	0.3792439	0.6207561

The image above shows the predicted probability in value, to predict the Y or N for approval status. Logistic regression uses probability of 0.5, with more than 0.5 being the chosen value. If Loan_Approval_Status for YES (Y) is higher than 0.5, then Y will be in the Approval_Status column, and vice versa. Hence, for the loan the be approved, the predicted probability for Y must be higher than 0.5, otherwise, the loan will be rejected (Loan_Approval_Status will be N).

CHAPTER 11: Conclusion

This assignment is beneficial for me, as an aspiring data scientist. Firstly, being able to understand coding better comes from this assignment. Using SQL for coding is beneficial, to produce queries and sub-queries to pre-process and analyse data. The dataset used for this assignment contains structured data with missing values, and with the usage of SAS Studio and SQL coding, I was able to impute missing values with Mode (Mean and Median can also be used), of which this assignment was beneficial in teaching me these methods. As a data scientist, I will encounter such issues with datasets. Hence, imputation techniques are pivotal in addressing these issues.

Moreover, this assignment taught me the benefits of keeping copies of datasets for fault-tolerance. Replicating the dataset will enable a backup to be kept, and once the final coding is ready, this can be used as the final output.

Logistic regression is a beneficial predictive modelling to be used, as the dataset contained mainly categorical variables. Predictive analytics is a vital part of data science, and using logistic regression is key to predict the loan approval status of the applicants. The aim of this assignment is to investigate the relationships between all variables (univariate and bivariate analysis), and to find the right model for prediction.

Data Analytical Programming is a beneficial course, with SQL coding being the key component of this course. Using SQL coding, the banking industry can certainly benefit, by expediting and automating their loan approval process based on an applicant’s requirements and existing conditions. This would help both the bank and applicant, as predicting future loan approval status can help both users and consumers to understand the requirements for acquiring a bank loan.

CHAPTER 12: References

1. Basten, C., & Ongena, S. (2020). A FinTech matching mortgage lenders with borrowers online and bank competition, diversification, and automation opportunities. *University Of Zurich*. Retrieved 15 July 2021, from https://www.eba.europa.eu/
2. Melnychenko, S., Volosovych, S., & Baraniuk, Y. (2020). DOMINANT IDEAS OF FINANCIAL TECHNOLOGIES IN DIGITAL BANKING. *Baltic Journal of Economic Studies*, *6*(1), 92. https://doi.org/10.30525/2256-0742/2020-6-1-92-99
3. Al Azzawi, F. (2019). Data Mining in Credit Insurance Information System for Bank Loans Risk Management in Developing Countries. *International Journal of Business Intelligence And Data Mining*, *1*(1), 1. https://doi.org/10.1504/ijbidm.2019.10016599
4. Zhao, S., & Zou, J. (2021). Predicting Loan Defaults Using Logistic Regression. *Journal Of Student Research*, *10*(1). https://doi.org/10.47611/jsrhs.v10i1.1326
5. Saha, S., & Waheed, S. (2017). Credit Risk of Bank Applicants can be Predicted from Applicant's Attribute using Neural Network. *International Journal of Computer Applications*, *161*(3), 39-43. https://doi.org/10.5120/ijca2017913170
6. Kassem, E., & Trenz, O. (2020). Automated Sustainability Assessment System for Small and Medium Enterprises Reporting. *Sustainability*, *12*(14), 5687. https://doi.org/10.3390/su12145687
7. Hidayat, A., Alam, F., & Helmi, M. (2020). Consumer Protection for Financial Technology Peer to Peer Lending Applicants in Indonesia. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY*, *9*(01). https://doi.org/10.7176/jlpg/102-08
8. Al-Blooshi, L., & Nobanee, H. (2020). Applications of Artificial Intelligence in Financial Management Decisions: A Mini-Review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3540140
9. Ivashina, V., Laeven, L., & Moral-Benito, E. (2020). Loan Types and the Bank Lending Channel. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3646021
10. Abuka, C., Alinda, R., Minoiu, C., Peydró, J., & Presbitero, A. (2019). Monetary policy and bank lending in developing countries: Loan applications, rates, and real effects. *Journal Of Development Economics*, *139*, 185-202. https://doi.org/10.1016/j.jdeveco.2019.03.004
11. Moreira, C., Haven, E., Sozzo, S., & Wichert, A. (2018). Process mining with real world financial loan applications: Improving inference on incomplete event logs. *PLOS ONE*, *13*(12), e0207806. https://doi.org/10.1371/journal.pone.0207806
12. Shrestha, S., & Paudel, L. (2019). Classification of Loan Applications of Garima Bikas Bank Ltd Using Decision Tree Classification Method. *Journal Of Advanced College of Engineering and Management*, *5*, 147-152. https://doi.org/10.3126/jacem.v5i0.26763
13. Wang, X., Han, L., & Huang, X. (2020). Bank competition, concentration, and EU SME cost of debt. *International Review of Financial Analysis*, *71*, 101534. https://doi.org/10.1016/j.irfa.2020.101534
14. Serrasqueiro, Z., Leitão, J., & Smallbone, D. (2018). Small- and medium-sized enterprises (SME) growth and financing sources: Before and after the financial crisis. Journal Of Management & Organization, 27(1), 6-21. https://doi.org/10.1017/jmo.2018.14
15. Gupta, A., Pant, V., Kumar, S., & Bansal, P. (2020). Bank Loan Prediction System using Machine Learning. 2020 9Th International Conference System Modeling and Advancement In Research Trends (SMART). https://doi.org/10.1109/smart50582.2020.9336801