LASIANDRA FINANCE INC. (LFI) LOAN APPROVAL PROCESS USING SAS PROGRAM

Table of Contents

[CHAPTER 1: 1](#_Toc123910460)

[Introduction 1](#_Toc123910461)

[CHAPTER 5: 2](#_Toc123910462)

[Data Exploration 2](#_Toc123910463)

[CHAPTER 6: 4](#_Toc123910464)

[Methodology. 4](#_Toc123910465)

[CHAPTER 7: 6](#_Toc123910466)

[Experimentation 6](#_Toc123910467)

[7.1 Create a folder on SAS 6](#_Toc123910468)

[7.2 Uploading dataset to SAS 6](#_Toc123910469)

[7.3 Creating a Library for the project 7](#_Toc123910470)

[7.4.1 Importation of the datasets to the library. 8](#_Toc123910471)

[Chapter 8: 9](#_Toc123910472)

[Exploratory Data Analysis 9](#_Toc123910473)

[8.1 Data Dictionary structure of the dataset 9](#_Toc123910474)

[8.2 Analysis of variables in the dataset -LIB067275.TRAINING\_DS 10](#_Toc123910475)

[8.2.1 Univariate Analysis of categorical variables – GENDER 10](#_Toc123910476)

[8.2.2 Univariate Analysis of categorical variables – QUALIFICATION 12](#_Toc123910477)

[8.2.3 Univariate Analysis of categorical variables – FAMILY\_MEMBERS 13](#_Toc123910478)

[8.2.4 Univariate Analysis of categorical variables – EMPLOYMENT 14](#_Toc123910479)

[8.2.5 Univariate Analysis of categorical variables – LOAN\_HISTORY 16](#_Toc123910480)

[8.2.6 Univariate Analysis of categorical variables – LOAN LOCATION 17](#_Toc123910481)

[8.2.7 Univariate Analysis of categorical variables – LOAN APPROVAL STATUS 19](#_Toc123910482)

[8.3 Univariate Analysis of Numeric variables 20](#_Toc123910483)

[8.3.1 Univariate Analysis of Numeric Variables – CANDIDATE\_INCOME 20](#_Toc123910484)

[8.3.2 Univariate Analysis of Numeric Variables – GUARANTEE\_INCOME 22](#_Toc123910485)

[8.3.3 Univariate Analysis of Numeric Variables – LOAN\_AMOUNT 23](#_Toc123910486)

[8.3.4 Univariate Analysis of Numeric Variables – LOAN\_DURATION 24](#_Toc123910487)

[8.4 Bivariate analysis of variables 26](#_Toc123910488)

[8.4.1 Bivariate analysis of the variables (Gender – Categorical Variables Vs Qualification Categorical Variables 26](#_Toc123910489)

[8.4.2 Bivariate analysis of the variables (Marital status – Categorical Variables Vs Loan approval status Categorical Variables) 27](#_Toc123910490)

[8.4.3 Bivariate analysis of the variables (Loan location – Categorical Variables Vs Candidate Income - Continuous Variables) 29](#_Toc123910491)

[8.4.4 Bivariate analysis of the variables (Marital status – Categorical Variables Vs Guarantee income - Continuous Variables 30](#_Toc123910492)

[Chapter 9: 32](#_Toc123910493)

[Data preprocessing. 32](#_Toc123910494)

[9.1 Missing value imputation 32](#_Toc123910495)

[9.2 Missing value imputation in categorical variables – Marital status 32](#_Toc123910496)

[9.3 Missing value imputation in categorical variables – Family members 35](#_Toc123910497)

[9.4 Missing value imputation in categorical variables – Gender 40](#_Toc123910498)

[9.5 Missing value imputation in categorical variables – Employment 44](#_Toc123910499)

[9.6 Missing value imputation in categorical variables – Loan history 49](#_Toc123910500)

[9.7 Missing value imputation in continuous variables – Loan Amount 53](#_Toc123910501)

[9.8 Missing value imputation in continuous variables – Loan Duration 57](#_Toc123910502)

[Chapter 10. 60](#_Toc123910503)

[Data Exploration: TESTING\_DS 60](#_Toc123910504)

[10.1 Introduction of SAS MACRO 60](#_Toc123910505)

[10.2 Univariate Analysis of the categorical variables using SAS MACRO 60](#_Toc123910506)

[10.3 Univariate Analysis of Numeric Variables – CANDIDATE\_INCOME 63](#_Toc123910507)

[10.4 Univariate Analysis of Numeric Variables – GUARANTEE\_INCOME 64](#_Toc123910508)

[10.5 Univariate Analysis of Numeric Variables – LOAN\_AMOUNT 66](#_Toc123910509)

[10.5 Univariate Analysis of Numeric Variables – LOAN\_DURATION 67](#_Toc123910510)

[10.6 Bivariate analysis of categorical variables 69](#_Toc123910511)

[10.7 Bivariate analysis of categorical variables & numeric variables using SAS MACRO 72](#_Toc123910512)

[Chapter 11: 75](#_Toc123910513)

[Data Preprocessing: TESTING DS 75](#_Toc123910514)

[11.1 Missing value imputation of categorical variables 75](#_Toc123910515)

[11.1.1 Missing value imputation in the categorical variables – GENDER 75](#_Toc123910516)

[11.1.2 Missing value imputation in the categorical variables – FAMILY MEMBERS 79](#_Toc123910517)

[11.2 Missing value imputation of numeric variables 83](#_Toc123910518)

[11.2.1 Missing value imputation in the numeric variables – LOAN AMOUNT 83](#_Toc123910519)

[11.2.1 Missing value imputation in the numeric variables – LOAN DURATION 86](#_Toc123910520)

[Chapter 12: 90](#_Toc123910521)

[Logistic regression model. 90](#_Toc123910522)

[Chapter 13: 94](#_Toc123910523)

[SAS ODS – Output delivery system 94](#_Toc123910524)

[13.1 Introduction 94](#_Toc123910525)

[CHAPTER 14: 95](#_Toc123910526)

[Conclusion 95](#_Toc123910527)

[CHAPTER 15: 96](#_Toc123910528)

[References 96](#_Toc123910529)

# CHAPTER 1:

# Introduction

Lasiandra Finance Inc. (LFI) is a leading private financing company that provides funding to small and medium-sized businesses (SMEs). The number of loan applications at LFI has increased due to expansion, as a result, LFIs face a problem in the loan approval process. The current approval process is a complicated procedure in terms of loan validation and verification. A machine learning model is a proposed solution to automate the loan approval process.

The data scientist must carefully analyse the data set obtained from previous customers and construct the most accurate model to predict whether the approval process will be approved or rejected using SAS studio. This project is divided into sections that outline the process of developing a machine learning solution for LFI.

The chapters explain the background of the company, the assumptions and justifications of the project and finally an extensive review of literatures on loans, and the loan approval process. The datasets are analyzed and documented, followed by preprocessing and model implementation. The flow process is carefully documented accompanied by SAS codes, outputs and appropriate explanations.

# CHAPTER 5:

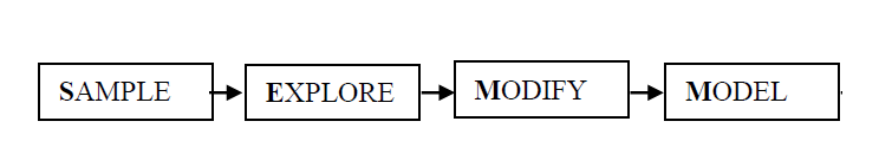
# Data Exploration

**5.1 Metadata of the dataset.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of variable | Description | Data Type | Length | Sample Data |
| SME\_LOAN\_ID\_NO | Loan application number | Char | 8 | LP001002,  LP001003,  LP001005 |
| GENDER | Gender of the applicant; Male or Female | Char | 6 | Female; Male |
| LOAN\_AMOUNT | Loan amount | Numeric | 5 | 128,66, |
| MARITAL\_STATUS | The relationship status of the applicant | Char | 6 | Not married,  Married |
| FAMILY\_MEMBERS | Number of the family members | Char | 6 | 0,1,2,3+ |
| QUALIFICATION | Education level  Graduate/U | Char | 6 | Graduate,  Undergraduate |
| EMPLOYMENT | Employment status; Yes/no | Char | 6 | No, yes |
| CANDIDATE\_INCOME | Monthly Income of the applicant | Numeric | 6 | 5849,  3000 |
| GUARANTEE\_INCOME | Joint income of the applicant | Numeric | 5 | 0,  1508 |
| LOAN\_DURATION | Repayment period of the loan | Numeric | 4 | 360 |
| LOAN\_HISTORY | Past loan records | Numeric | 6 | 0, 1 |
| LOAN\_LOCATION | City/Town of the applicant | Char | 8 | Town, City |
| LOAN\_APPROVAL\_STATUS | Approval of the loan yes/no | Char | 6 | Y, N |

# CHAPTER 6:

# Methodology.



**Sampling.**

This is the first stage of the data mining methodology. This stage involves the splitting of the data set for the machine model. The original dataset is split into two samples, the training set and the testing set. The training set is the dataset that will be used by the machine model to learn the interdependence and the weights of the independent variables to the target variables. Whereas the testing set will be used by the machine learning model to validate the model.

**Explore**

This stage involves the understanding of the training set and testing set. It comprises univariate and bivariate analyses of each variable involved in the dataset. In this analysis, the data type, the characteristics of the data, the distribution, and the presence of missing values and outliers will be explored. These patterns explored will assist in accurately developing the machine model for the prediction and assisting in explaining the core problem.

**Modify.**

This stage includes all data preparation before modelling. The accuracy of the prediction is dependent on how well the data was processed, so data preparation is critical. The dataset will be cleaned to remove any outliers and missing values (data cleaning), the number of irrelevant features will be reduced (feature selection), and the data will be normalized if necessary (data transformation)

**Modelling**

The data scientist will use a machine learning algorithm, specifically a classification technique, to predict whether or not a loan application will be approved. The proposed machine learning models will then be used to train the dataset. The logistic regression machine learning model will be used.

# CHAPTER 7:

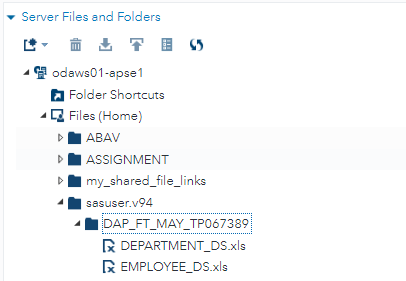
# Experimentation

## 7.1 Create a folder on SAS

7.1.1 Explanation

Within the SAS environment, the data scientist created a folder to save the programme code files and datasets, that will be used to build the machine learning tool for LFI. This folder is known as DAP\_FT\_TP067389.

7.1.2 Screenshot(s)

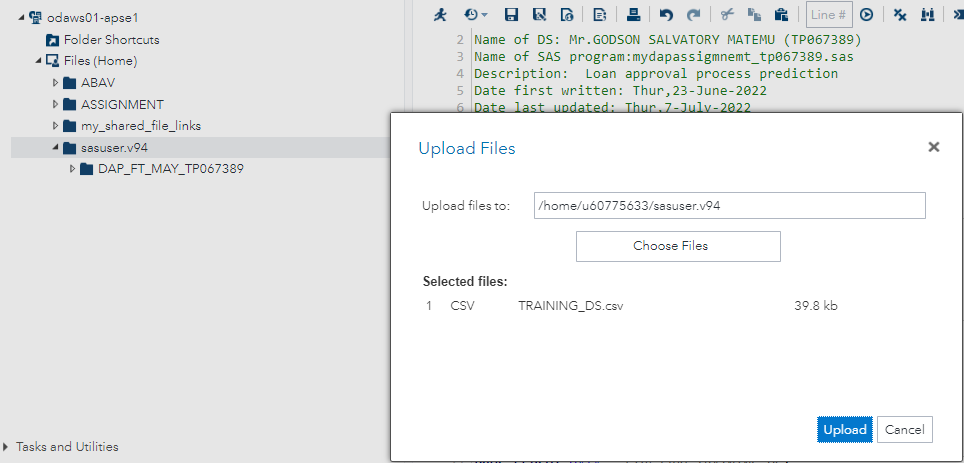


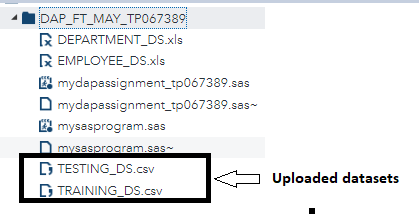
## 7.2 Uploading dataset to SAS

7.2.1 Explanation

The training and test datasets were uploaded to the folder DAP\_FT\_TP067275, as shown below. This makes it simple to refer to when performing data mining on the variables.

7.2.2 Screenshots



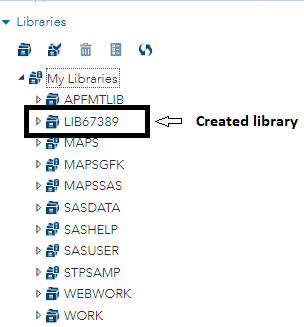


## 7.3 Creating a Library for the project

7.3.1 Explanation

A library was created to help with project organization and secondary storage. The datasets must be uploaded to the library folder for easy access when performing exploratory data analysis, imputation of missing values, and modelling using the library name LIB067389.

7.3.2 Screenshots

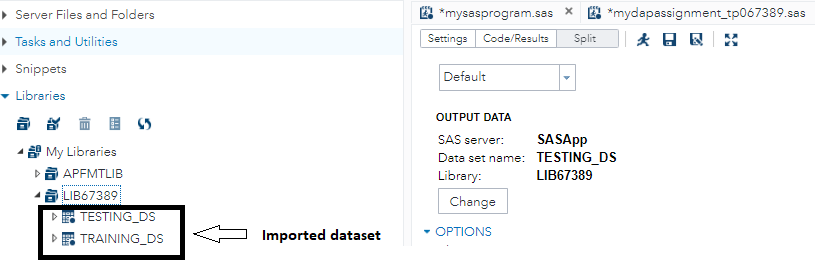


## 7.4.1 Importation of the datasets to the library.

7.4.2 Explanation

The datasets are imported to the library to act as the secondary storage of the dataset. Primarily the dataset was at the work library which the library is not permanent storage for the data, because the data is lost when SAS is closed. But the importation of the dataset to the library solves the problem.

7.4.3 Screenshots



# Chapter 8:

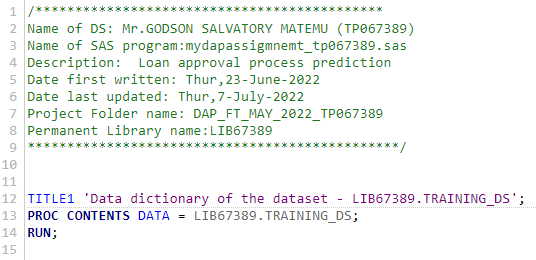
# Exploratory Data Analysis

## 8.1 Data Dictionary structure of the dataset

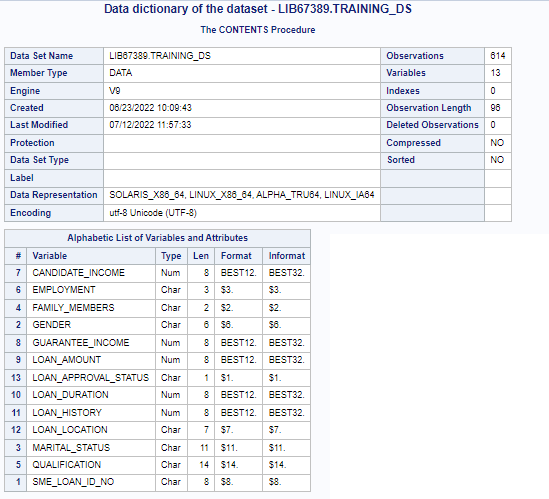
8.1.1 Explanation

The training data was investigated, and it was discovered that it contains 614 instances and 13 variables. The maximum variable length was 14 characters, and the minimum variable length was 1, which belongs to the LOAN STATUS variable with binary values. The qualification variable, on the other hand, is the longest variable, with each instance containing 14 characters.

8.1.2 SAS Codes



8.1.3 Output

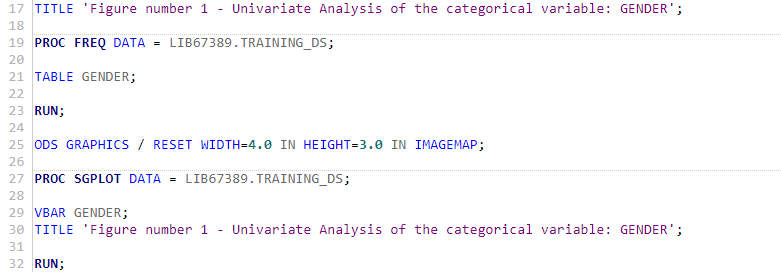


## 8.2 Analysis of variables in the dataset -LIB067275.TRAINING\_DS

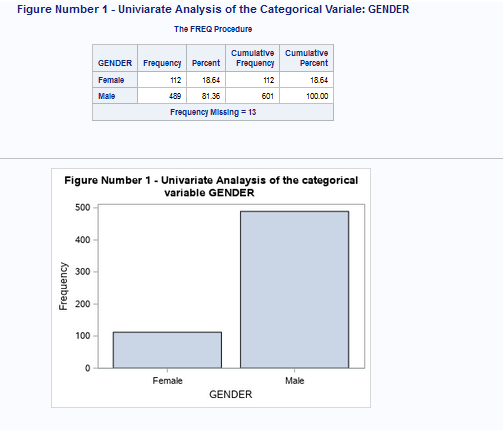
The dataset contains seven categorical variables, the univariate analysis of each categorical variable shall each be analyzed in isolation to determine if they have missing values and the mode of each category represented in the dataset. This analysis serves as the foundation for a thorough understanding of the LFI loan process and which customers are drawn to the LFI group.

### 8.2.1 Univariate Analysis of categorical variables – GENDER

**SAS codes**

****

**Output**

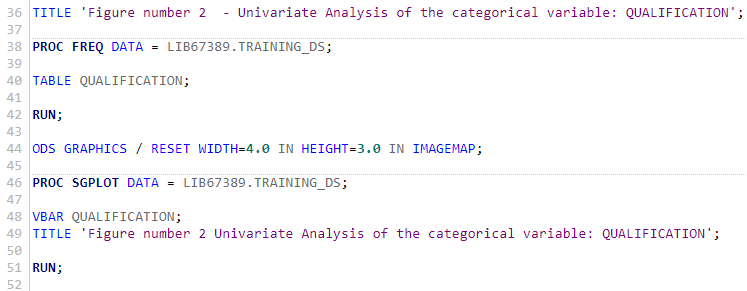
****

**Explanation**

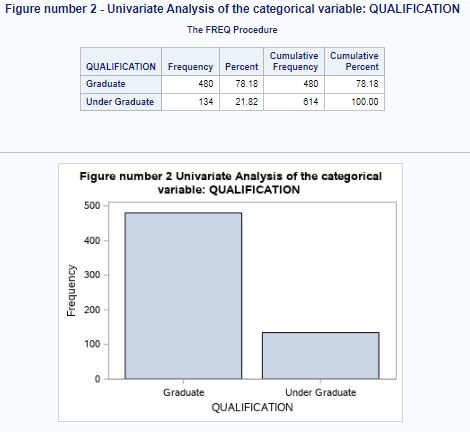
The analysis below shows that males exceed the number of females in the dataset. From a total of 614 instances, there are 13 missing values. Males represent 81.14 per cent of the population, while females represent 18.64 per cent. Missing values will be imputed later in this section.

### 8.2.2 Univariate Analysis of categorical variables – QUALIFICATION

**SAS codes**

****

**Output**

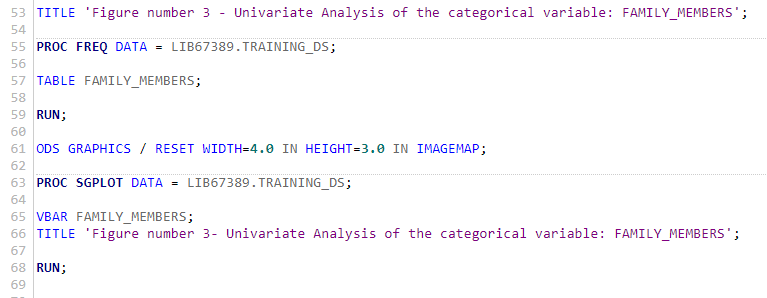
****

**Explanation**

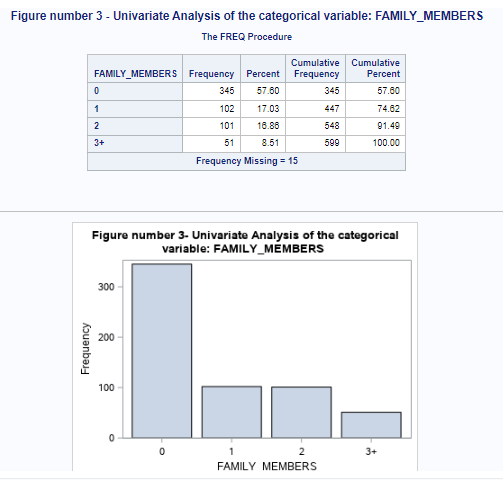
The vast majority of LFI customers in the LFI dataset have a graduate degree which accounts for 78.18 per cent of their total customers, while 21.62 per cent of customers are undergraduates. This implies that most graduate individuals are the ones who are involved in the loan application process. Lastly, the dataset has no missing values.

### 8.2.3 Univariate Analysis of categorical variables – FAMILY\_MEMBERS

**SAS codes**



**Output**

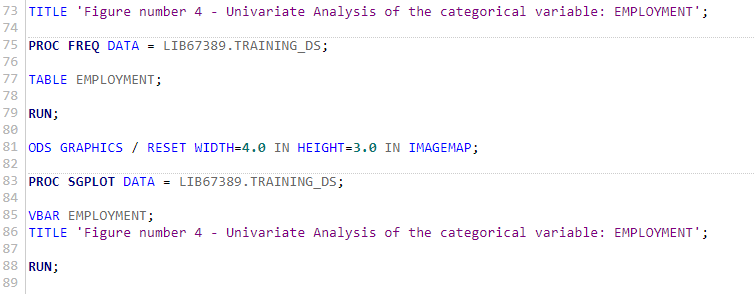
****

**Explanation**

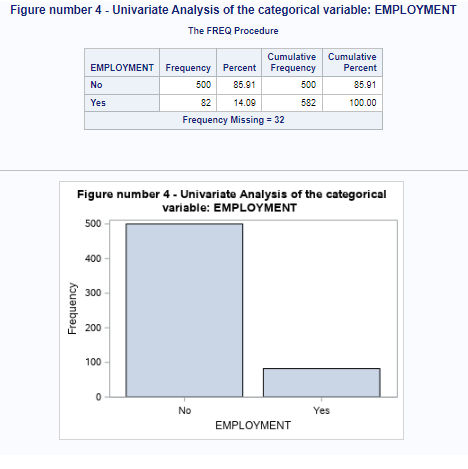
According to the analysis, there are 509 non-missing values and 13 missing values. The majority of people in the dataset do not have any family members, accounting for 57.60 per cent, while the remaining 43.40 per cent have one, two, or three or more family members, with proportions of 17.03 per cent, 16.86 per cent, and 8.51 per cent, respectively.

### 8.2.4 Univariate Analysis of categorical variables – EMPLOYMENT

**SAS codes**

****

**Output**

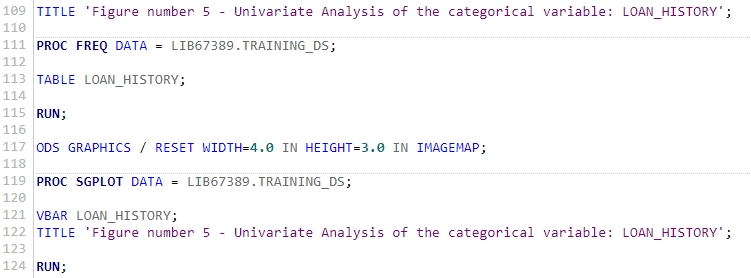
****

**Explanation**

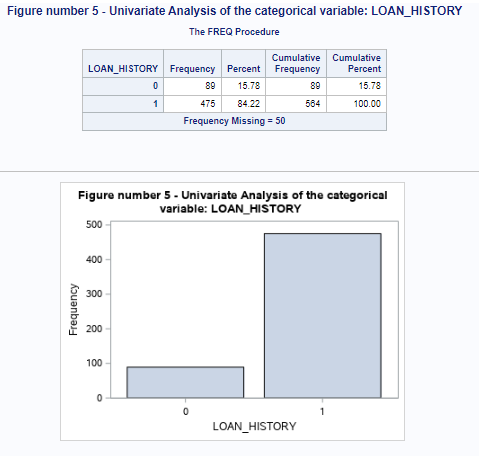
The majority of applicants in the LFI training dataset are unemployed, with unemployed people accounting for 78.18 per cent of customers and employed people accounting for 21.82 per cent of people. This increases LFI's risk of high credit defaults from their customers because their customers' lack of employment implies that they may have difficulty repaying their LFI loans. This could hurt LFI’s profitability and cash flows.

### 8.2.5 Univariate Analysis of categorical variables – LOAN\_HISTORY

**SAS codes**

****

**Output**

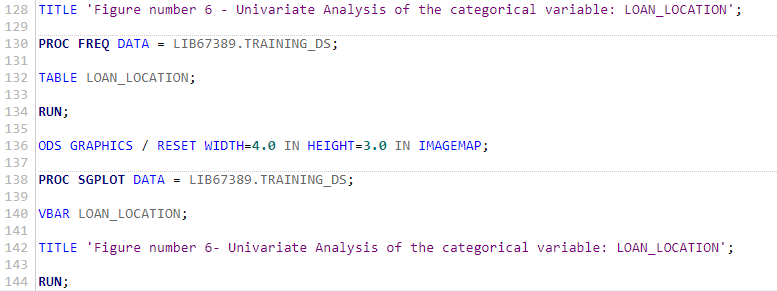
****

**Explanation**

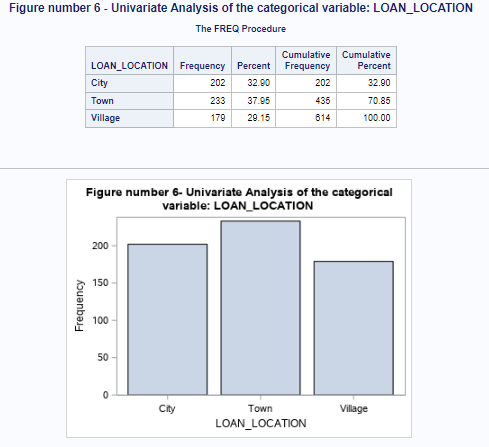
According to the analysis, 50 loan applicants had an unidentified history of loan applicants due to missing values. 57.6 per cent (345 applicants) have no family members, while 17.03 per cent (102 applicants) have one. Furthermore, 16.86 per cent (101 applicants) have two family members. 8.51 per cent of applicants (51 in total) have three or more family members.

### 8.2.6 Univariate Analysis of categorical variables – LOAN LOCATION

**SAS codes**

****

**Output**

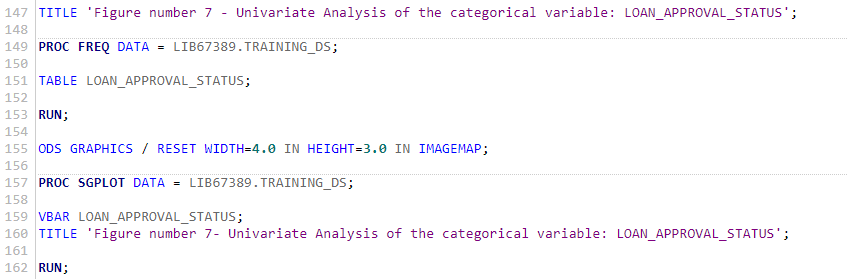
****

**Explanation**

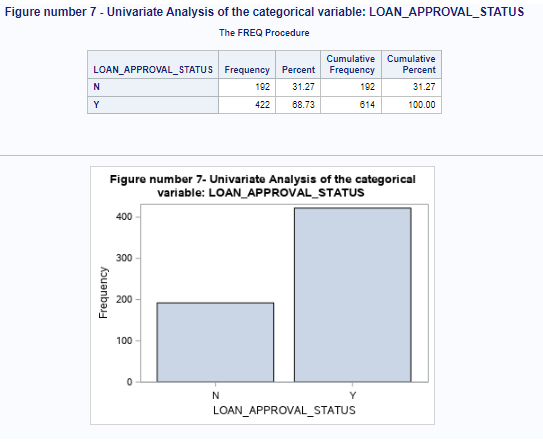
The loan location variable has no missing values and represents a significant majority of where the majority of LFI's customers are located. The majority of customers are almost evenly distributed across cities, towns, and villages, with slightly more people living within towns. The percentages of customers living in cities, towns, and villages are 32.90%, 37.95%, and 29.15%, respectively.

### 8.2.7 Univariate Analysis of categorical variables – LOAN APPROVAL STATUS

**SAS codes**

****

**Output**

****

**Explanation**

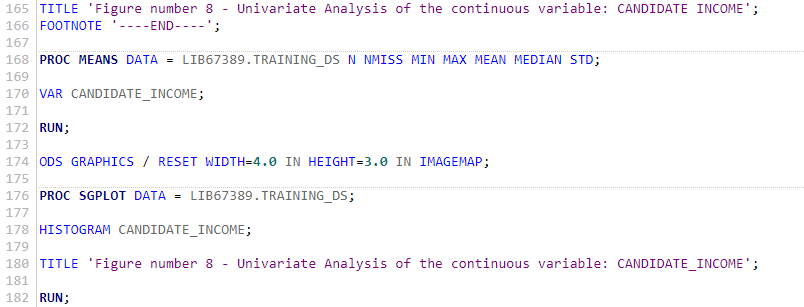
The dataset has an uneven distribution of approved (Y) and rejected (N) loans, with the percentage of approved loans (Y) being 68.73 per cent (422 applicants). The rejected loan percentage (N) is 31.27 per cent (192 applicants). There are no missing values in the dataset that have an unidentified loan approval status.

## 8.3 Univariate Analysis of Numeric variables

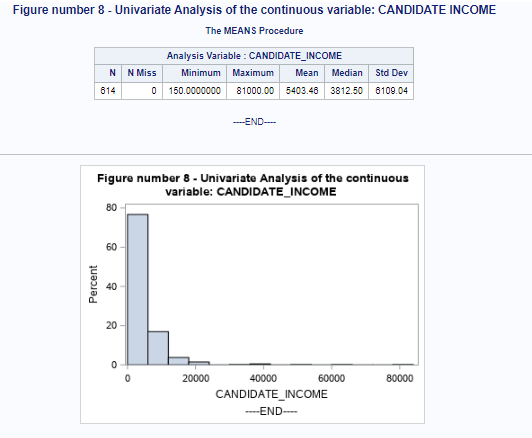
The dataset contains five continuous variables, the univariate analysis of each continuous variable shall each be analyzed in isolation to determine if they have missing values and the statistical measure of each variable represented in the dataset.

### 8.3.1 Univariate Analysis of Numeric Variables – CANDIDATE\_INCOME

**SAS Code**

****

**Output**

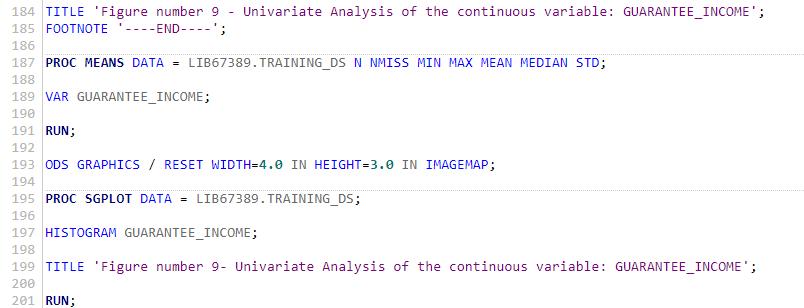
****

**Explanation**

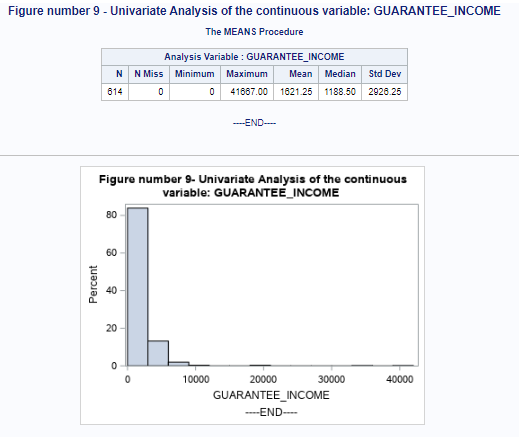
From the analysis, there are no missing values or applicants with unknown income in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the mean and median at 5,403.46 and 3,812.5 respectively. The maximum value is greater than the (mean + 3x standard deviation) value, indicating that this variable contains extreme outliers.

### 8.3.2 Univariate Analysis of Numeric Variables – GUARANTEE\_INCOME

**SAS Code**

****

**Output**

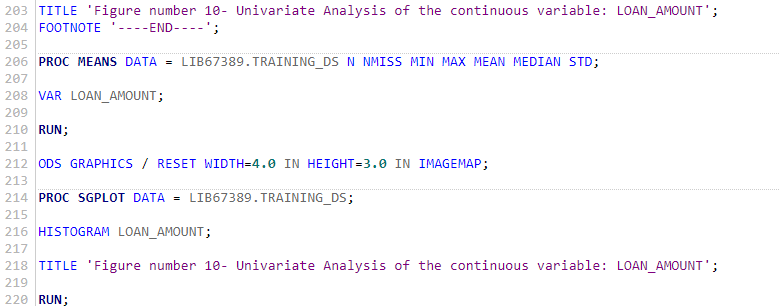
****

**Explanation**

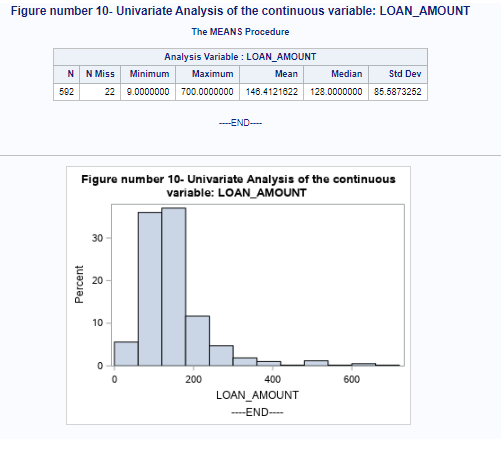
From the analysis, there are no missing values or applicants with unknown income in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the median and mean at 1188.50 and 1621.25 respectively. The maximum value is greater than the (mean + 3x standard deviation) value, indicating that this variable contains extreme outliers.

### 8.3.3 Univariate Analysis of Numeric Variables – LOAN\_AMOUNT

**SAS Code**

****

**Output**

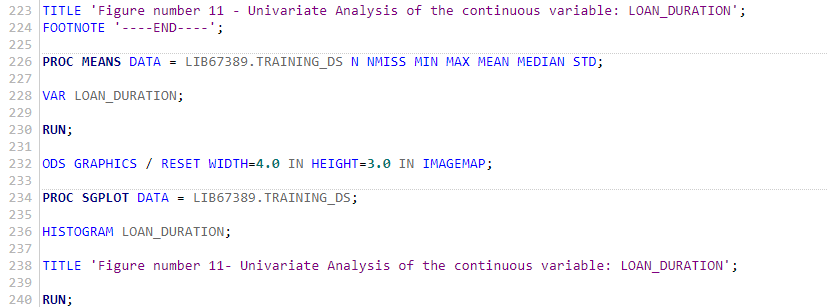
****

**Explanation**

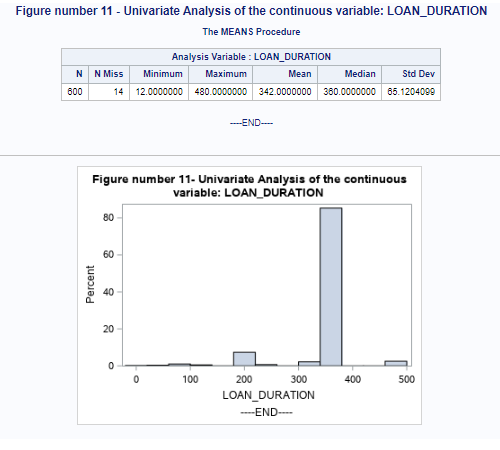
From the analysis, there are 22 missing values or applicants with unknown loan amounts in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the median and mean at 128 and 146.41 respectively.

### 8.3.4 Univariate Analysis of Numeric Variables – LOAN\_DURATION

**SAS Code**

****

**Output**

****

**Explanation**

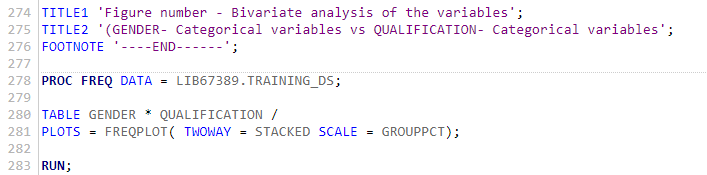
From the analysis, there are 14 missing values or applicants with unknown loan amounts in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the median and mean at 360 and 342 respectively.

## 8.4 Bivariate analysis of variables

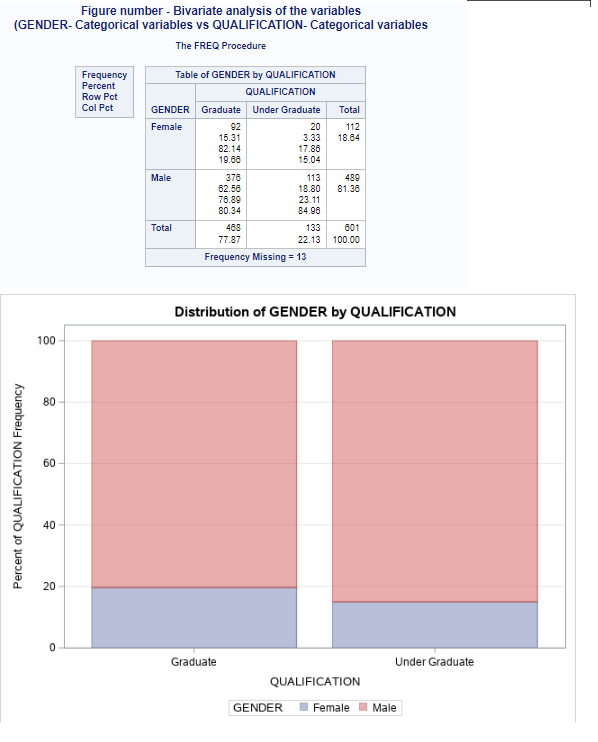
Bivariate analysis is the analysis of bivariate data. It is one of the most basic types of statistical analysis, used to determine whether or not two sets of values have a relationship.

### 8.4.1 Bivariate analysis of the variables (Gender – Categorical Variables Vs Qualification Categorical Variables

**SAS codes**

****

**Output**

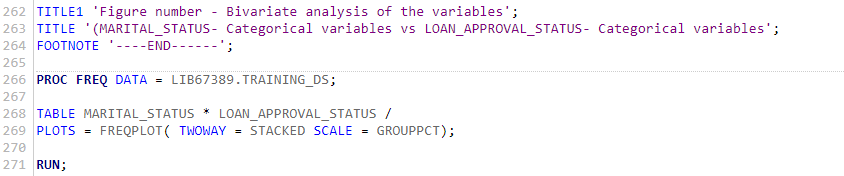
****

**Explanation**

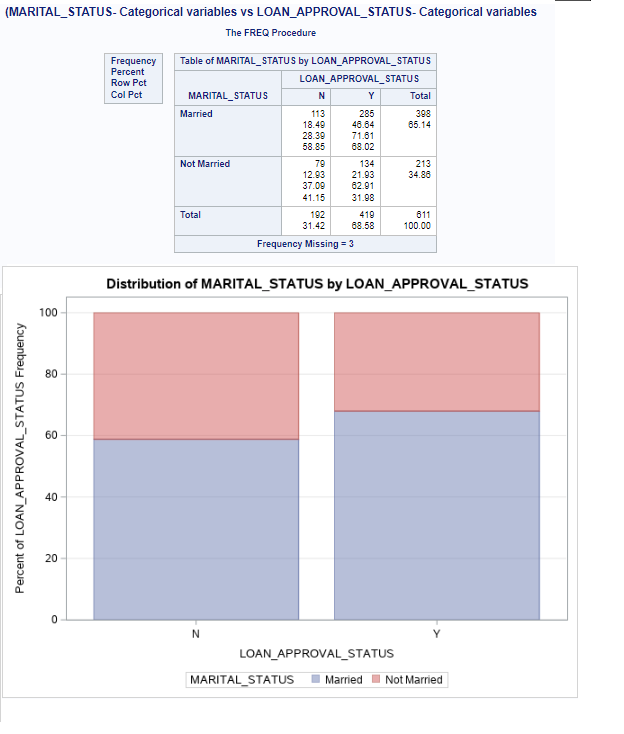
From the bivariate analysis, there is a high proportion of the number of men applicants compared to females. This affects the distribution of the proportion of qualification criteria in terms of gender. Out of the 468 graduate applicants, 19.66% of the applicants are female while 80.34% are male. On the other hand, out of 133 undergraduate applicants, 15.04% of applicants are female while 84.96% are male.

### 8.4.2 Bivariate analysis of the variables (Marital status – Categorical Variables Vs Loan approval status Categorical Variables)

**SAS codes**

****

**Output**

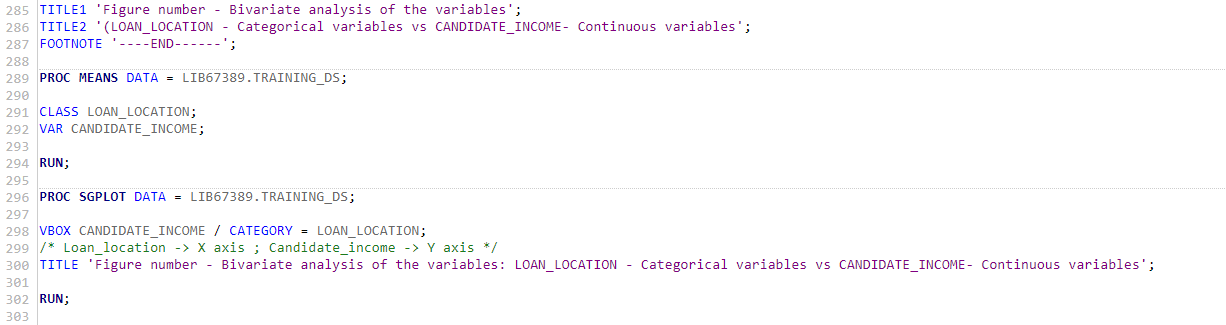
****

**Explanation**

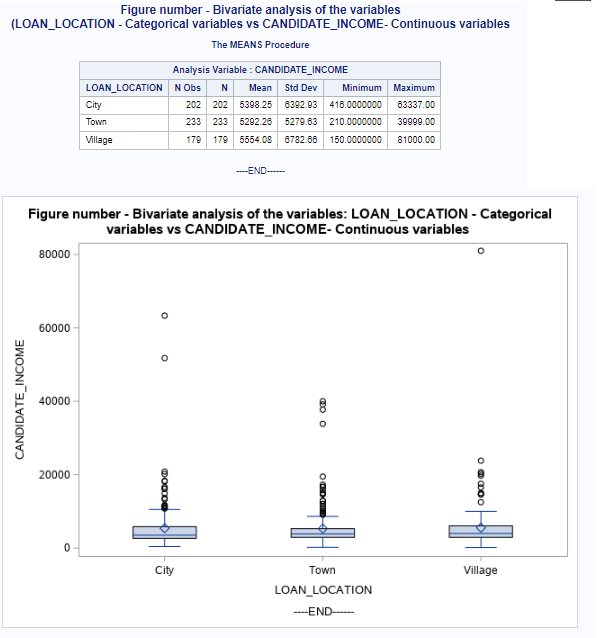
From the above analysis, the proportion of loan approval varies on the marital status of the applicants. 68.58% of the loan applications were approved while 32.42% were rejected. The analysis extracts further the proportion of the marital status according to the approval status, from the 419 approved loans 68.02% are married while the remaining are not married and also from the 192 rejected loans 58.85% are married and 41.15% are not married. Generally, there is a high proportion of married applicants compared to the not married applicants.

### 8.4.3 Bivariate analysis of the variables (Loan location – Categorical Variables Vs Candidate Income - Continuous Variables)

**SAS codes**

****

**Output**

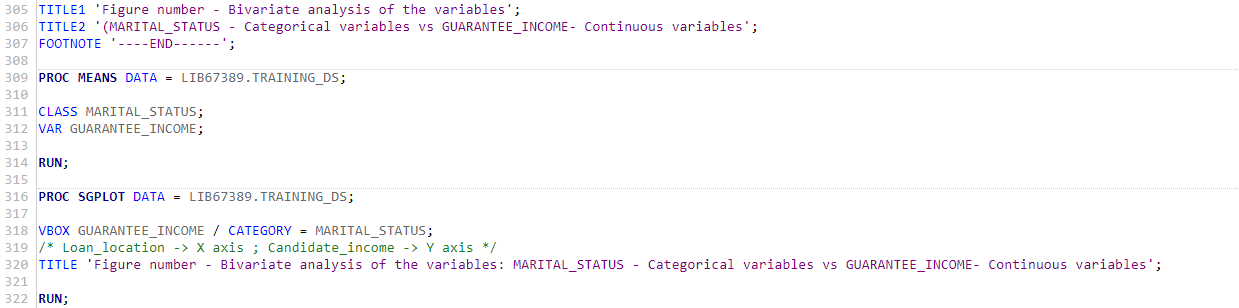
****

**Explanation**

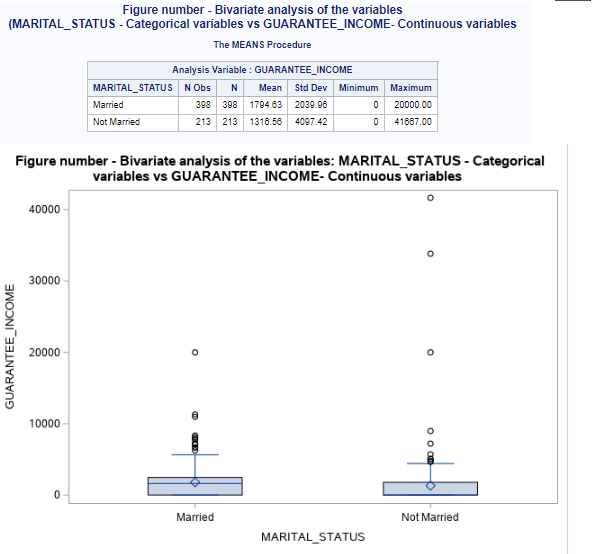
From the analysis, the candidate income of the applicants in the village is slightly higher compared to the other locations this is proved by the mean candidate income of different locations with the village candidate income average being 5554.08. But also, there is high dispersion of candidate income in the villages compared to the other locations. Due to variability of the minimum and maximum value of the income. As seen in the output above the maximum and minimum values are 81000 and 160 respectively which are all from the villages.

### 8.4.4 Bivariate analysis of the variables (Marital status – Categorical Variables Vs Guarantee income - Continuous Variables

**SAS codes**

****

**Output**

****

**Explanation**

From the analysis, the mean guarantee income of the married applicants is higher than that of the unmarried applicants, which is quite opposite from the SD perspective whereas the SD of the unmarried participants is higher. This means the unmarried applicants’ guaranteed income is more spread out than that of the married. This is further proved by the boxplot which shows there are more outliers in the unmarried applicants compared to the married.

# Chapter 9:

# Data preprocessing.

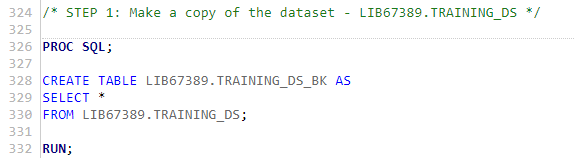
## 9.1 Missing value imputation

Categorical and numerical variables with missing values will be imputed using appropriate imputation methods using SAS studio.

## 9.2 Missing value imputation in categorical variables – Marital status

**9.2.1: Make a copy of the dataset**

**SAS codes.**

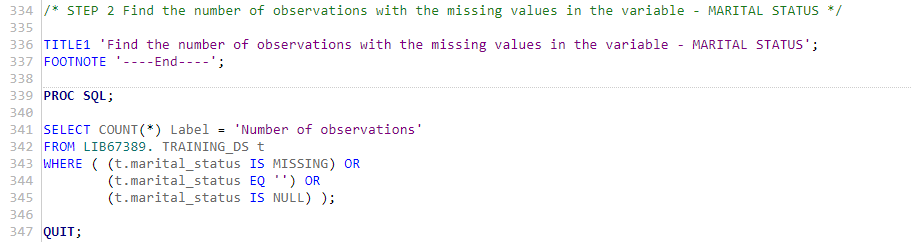
****

**Explanations**

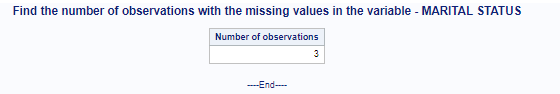
A backup of the training data set was created. The backup training set was created to protect against any unintentional changes made to the main training set. If any errors occurred on the main training set file, they could be recovered by using the backup file.

**9.2.2: The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output.**

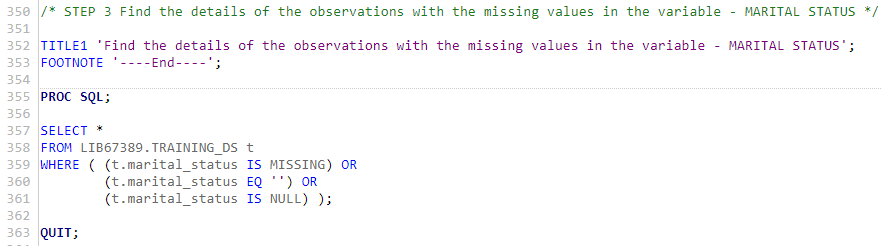
****

**Explanations**

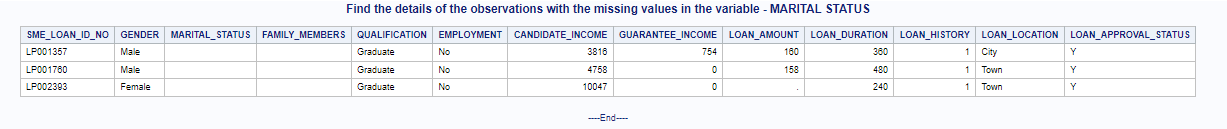
From the output above it can be observed that the marital status has 3 missing observations. An appropriate analysis is to be conducted to impute the missing values

**9.2.3:  Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output.**

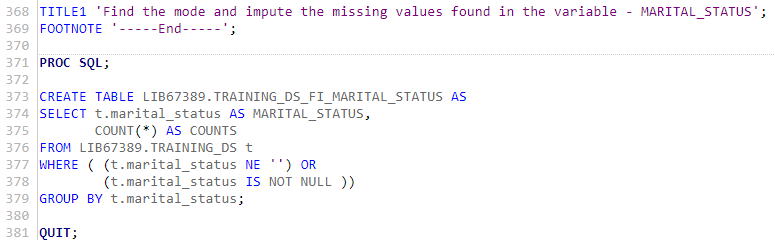
****

**Explanations**

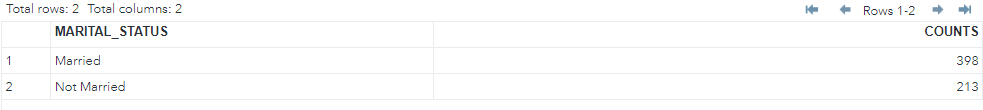
The details of the observations of the missing values are further analyzed. It can be observed that not only is the marital status observations missing but also the family members are also missing.

**9.2.4:** **Create a temporary dataset to hold marital status and the number of applicants to find the mode.**

**SAS codes.**

****

**Output.**

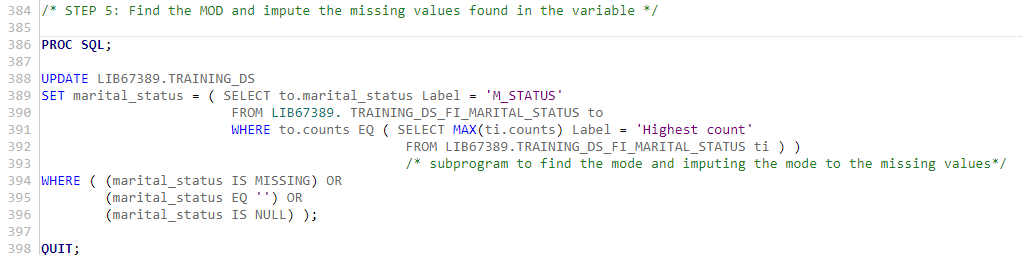
****

**Explanation**

**A** temporary dataset is created so that it can be easily used to calculate the number of observations on each category. The imputation method of categorical variables lies in imputing the missing values with the category with the highest mode.

**9.2.5: Find the Mode and impute the missing values in the variable**

**SAS codes.**

****

**Output.**

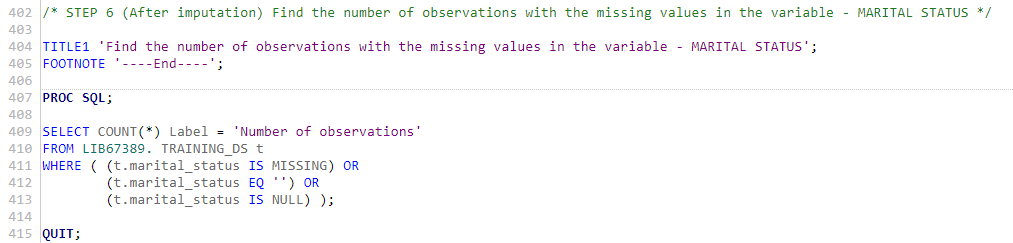
****

**Explanation**

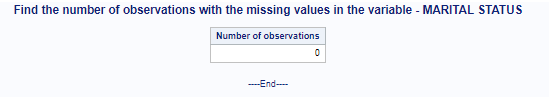
From the previous analysis, the mode was calculated, and the married category had the highest count. In this analysis, the mode calculated was used to be imputed to the missing values.

**9.2.6: After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output.**

****

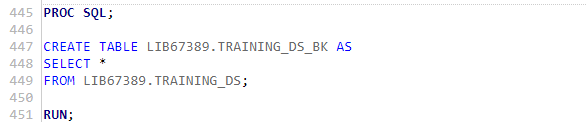
**Explanation**

**The** output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the marital status variable

## 9.3 Missing value imputation in categorical variables – Family members

**9.3.1: Make a copy of the dataset**

**SAS codes.**

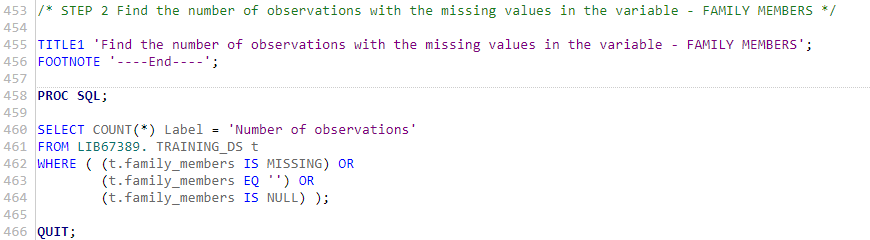
****

**Output**

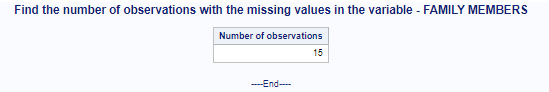
A backup of the training data set was created. The backup training set was created to protect against any unintentional changes made to the main training set. If any errors occurred on the main training set file, they could be recovered by using the backup file.

**9.3.2: The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output.**

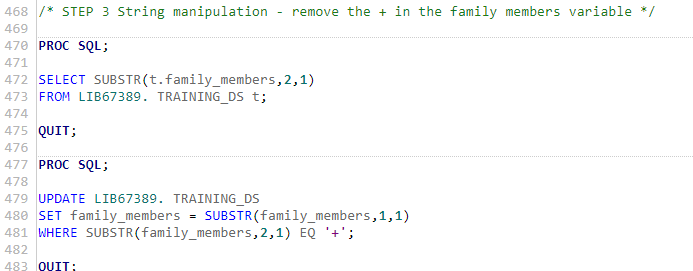
****

**Explanation**

From the output above it can be observed that the marital status has 15 missing observations. An appropriate analysis is to be conducted to impute the missing values

**9.3.3:  String manipulation remove the ‘+’ found in the family\_members variable**

**SAS codes.**

****

**Output.**

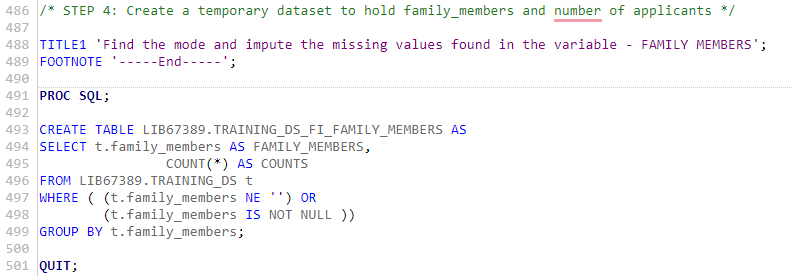
****

**Explanation**

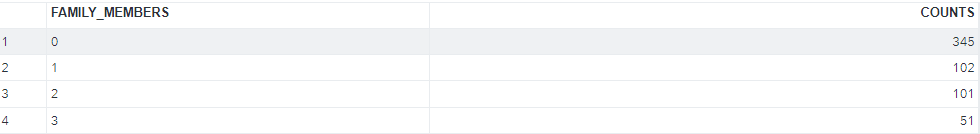
The family\_members variable had a string observation of 3+, With the help of SAS appropriate codes were used to separate the string ffrom3+ to three so that the datatype of all the variables be the same. The string was manipulated appropriately.

**9.3.4:** **Create a temporary dataset to hold marital status and the number of applicants to find the mode.**

**SAS codes.**

****

**Output.**

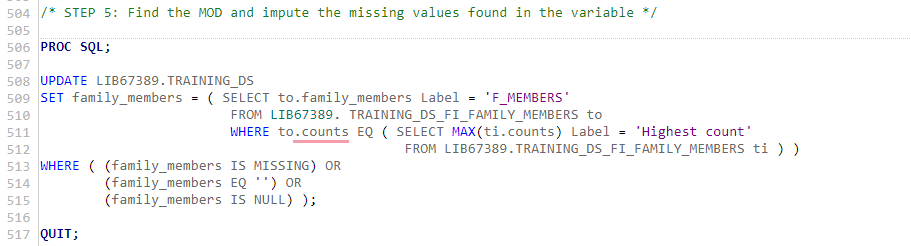
****

**Explanations**

**A** temporary dataset is created so that it can be easily used to calculate the number of observations on each category. The imputation method of categorical variables lies in imputing the missing values with the category with the highest mode. It can be observed that the family with 0o members has the highest mode.

**9.3.5: Find the Mode and impute the missing values in the variable**

**SAS codes.**

****

**Output.**

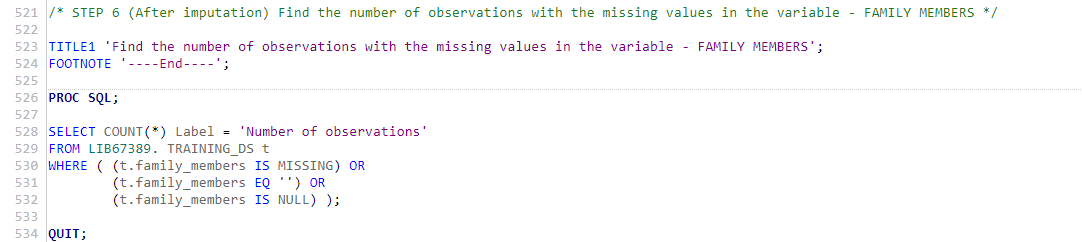
****

**Explanations**

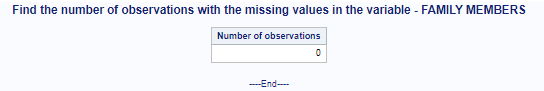
From the previous analysis, the mode was calculated, and the family members with 0 members had the highest count. In this analysis, the mode calculated was used to be imputed to the missing values.

**9.3.6: After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output.**

****

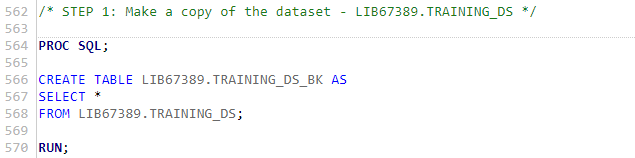
**Explanations**

The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the family\_members variable

## 9.4 Missing value imputation in categorical variables – Gender

**9.4.1: Make a copy of the dataset**

**SAS codes.**

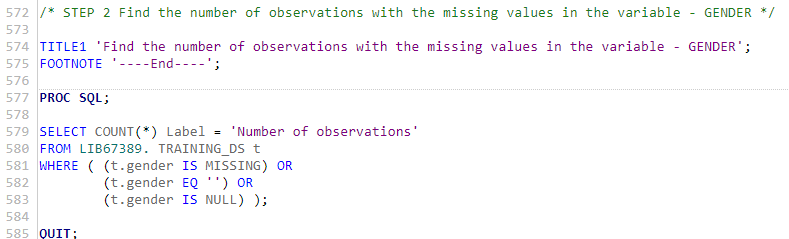
****

**Explanation**

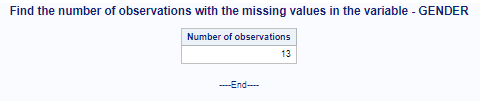
A backup of the training data set was created. The backup training set was created to protect against any unintentional changes made to the main training set. If any errors occurred on the main training set file, they could be recovered by using the backup file.

**9.4.2: The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output.**

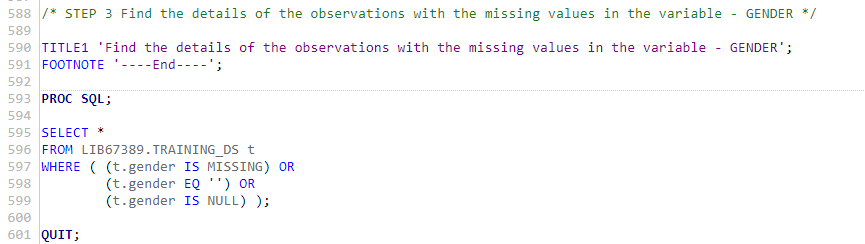
****

**Explanation**

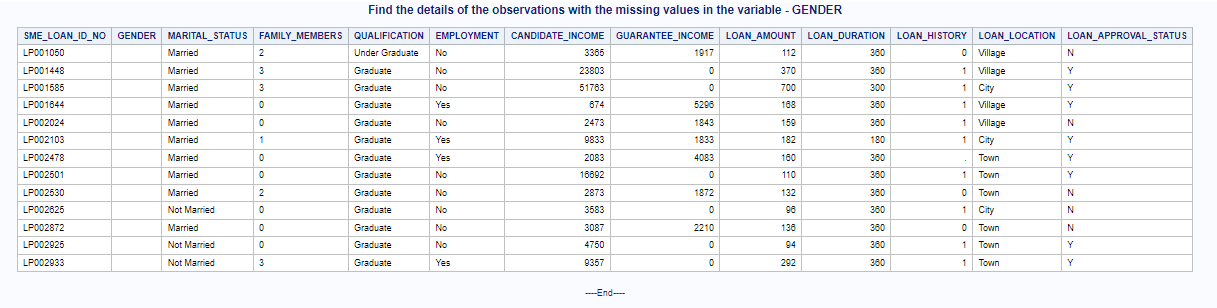
From the output above it can be observed that the gender has 13 missing observations. An appropriate analysis is to be conducted to impute the missing values

**9.4.3:  Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output.**

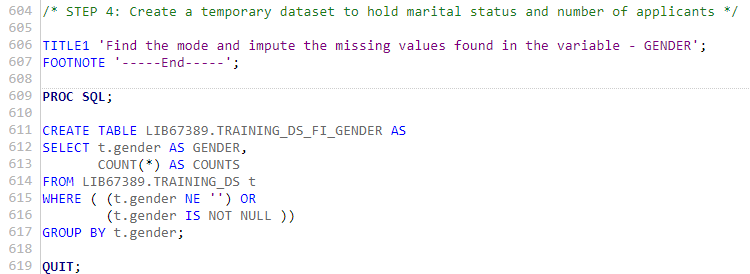
****

**Explanations**

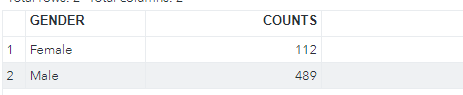
The details of the observations of the missing values are further analyzed. It can be observed that only the gender observations are missing.

**9.4.4:** **Create a temporary dataset to hold marital status and the number of applicants to find the mode.**

**SAS codes.**

****

**Output.**

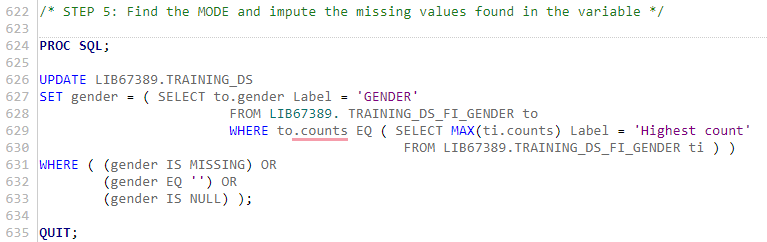
****

**Explanations**

**A** temporary dataset is created so that it can be easily used to calculate the number of observations on each category. The imputation method of categorical variables lies in imputing the missing values with the category with the highest mode.

**9.4.5: Find the Mode and impute the missing values in the variable –**

**SAS codes.**

****

**Output.**

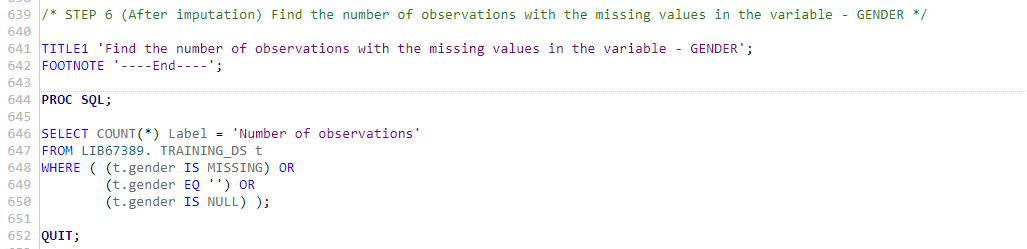
****

**Explanation**

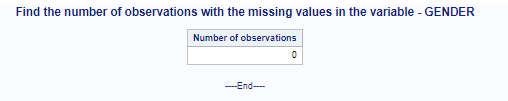
From the previous analysis, the mode was calculated, and the male gender has the highest count. In this analysis, the mode calculated was used to be imputed to the missing values.

**9.4.6: After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output.**

****

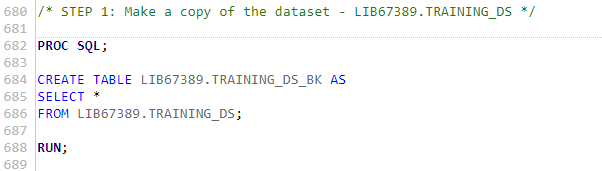
**Explanations**

The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the marital status variable

## 9.5 Missing value imputation in categorical variables – Employment

**9.5.1: Make a copy of the dataset**

**SAS codes.**

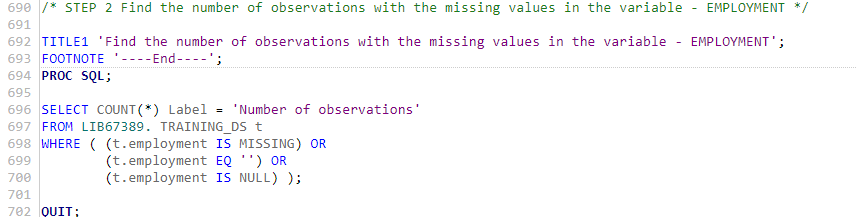
****

**Explanations**

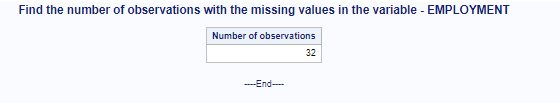
A backup of the training data set was created. The backup training set was created to protect against any unintentional changes made to the main training set. If any errors occurred on the main training set file, they could be recovered by using the backup file.

**9.5.2: The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output.**

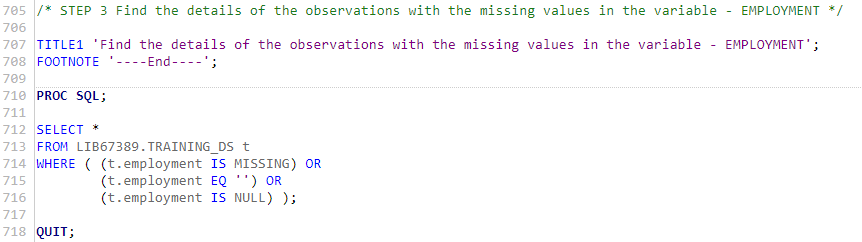
****

**Explanations**

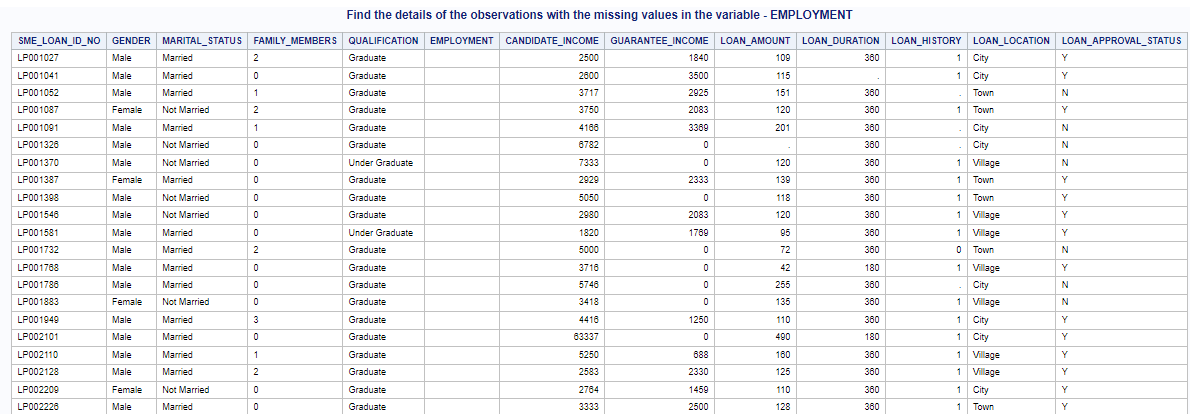
From the output above it can be observed that the employment status has 3 missing observations. An appropriate analysis is to be conducted to impute the missing values

**9.5.3:  Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output.**

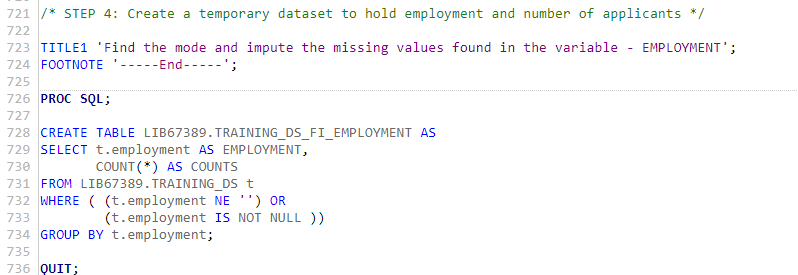
****

**Explanation**

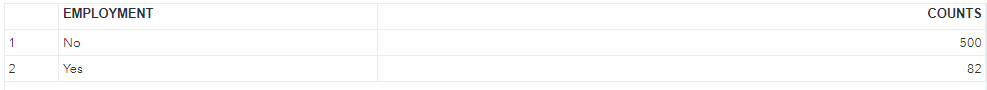
The details of the observations of the missing values are further analyzed. It can be observed that only the employment observations are missing.

**9.5.4:** **Create a temporary dataset to hold marital status and the number of applicants to find the mode.**

**SAS codes.**

****

**Output.**

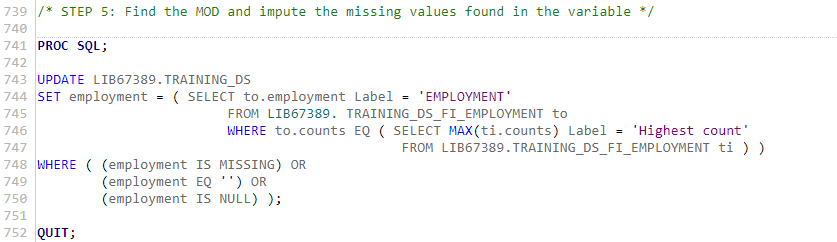
****

**Explanation**

**A** temporary dataset is created so that it can be easily used to calculate the number of observations on each category. The imputation method of categorical variables lies in imputing the missing values with the category with the highest mode.

**9.5.5: Find the Mode and impute the missing values in the variable –**

**SAS codes.**

****

**Output.**

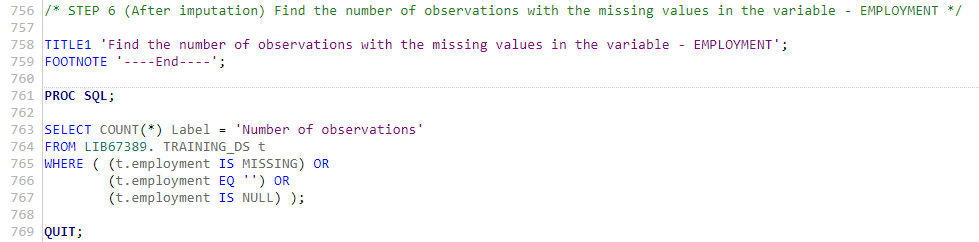
****

**Explanation**

From the previous analysis, the mode was calculated, and the applicants with no employment had the highest count. In this analysis, the mode calculated was used to be imputed to the missing values.

**9.5.6: After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output.**

****

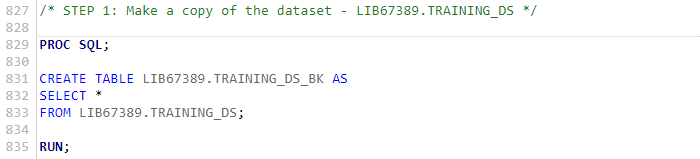
**Explanation**

The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the marital status variable

## 9.6 Missing value imputation in categorical variables – Loan history

**9.6.1: Make a copy of the dataset**

**SAS codes.**

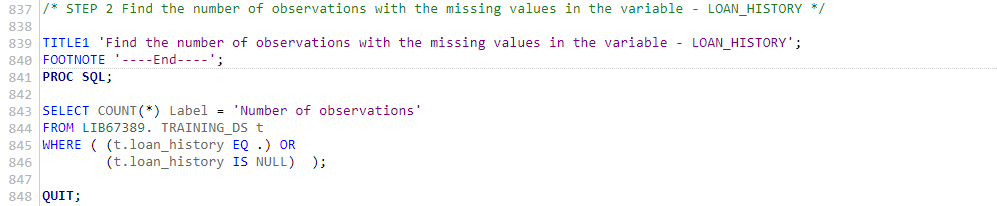
****

**Explanation**

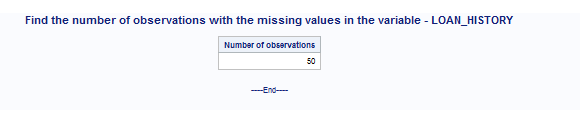
A backup of the training data set was created. The backup training set was created to protect against any unintentional changes made to the main training set. If any errors occurred on the main training set file, they could be recovered by using the backup file.

**9.6.2: The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output.**

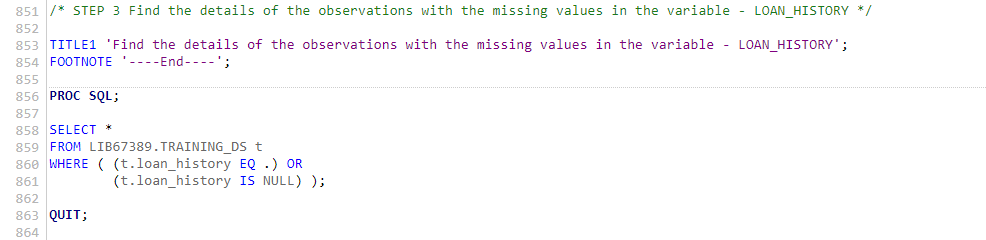
****

**Explanation**

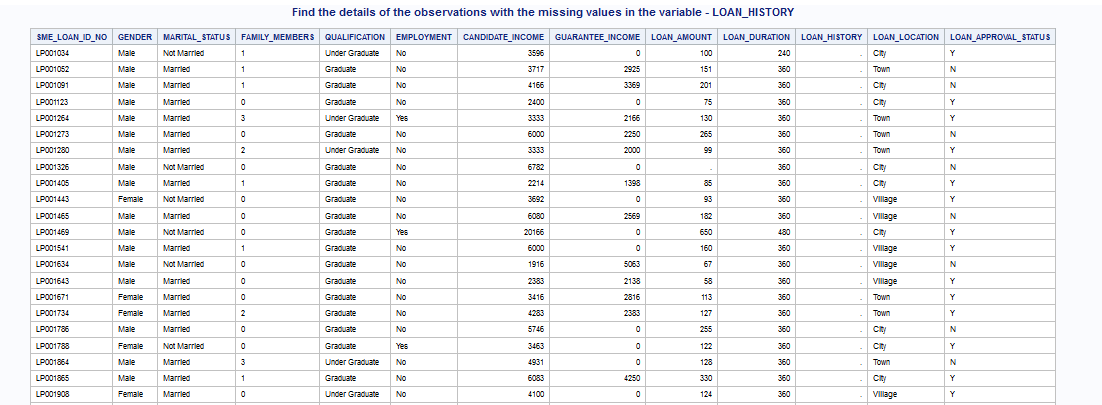
From the output above it can be observed that the loan history has 50 missing observations. An appropriate analysis is to be conducted to impute the missing values.

**9.6.3:  Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output.**

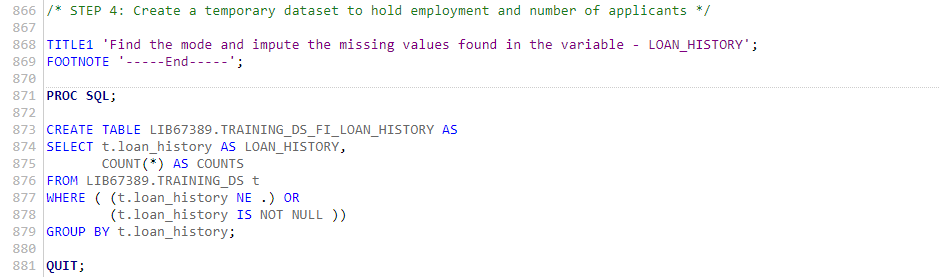
****

**Explanation**

The details of the observations of the missing values are further analyzed. It can be observed that only is the loan history observations missing.

**9.6.4:** **Create a temporary dataset to hold marital status and the number of applicants to find the mode.**

**SAS codes.**

****

**Output.**

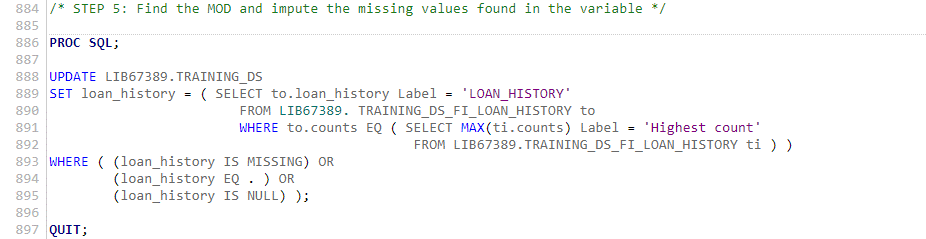
****

**Explanations**

**A** temporary dataset is created so that it can be easily used to calculate the number of observations on each category. The imputation method of categorical variables lies in imputing the missing values with the category with the highest mode.

**9.6.5: Find the Mode and impute the missing values in the variable –- LOAN HISTORY**

**SAS codes.**



**Output.**

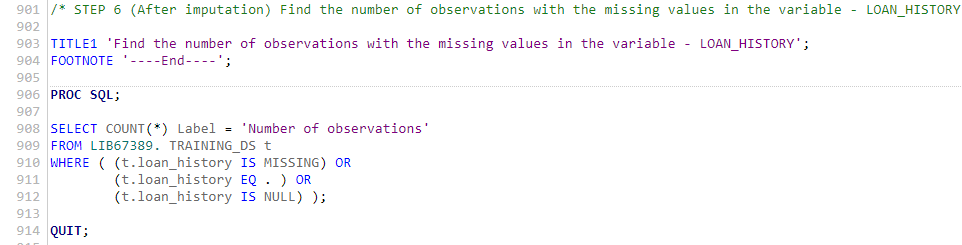
****

**Explanations**

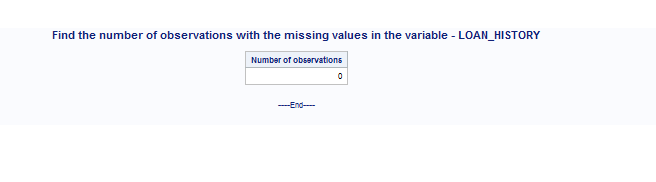
From the previous analysis, the mode was calculated, and applicants with a loan history had the highest count. In this analysis, the mode calculated was used to be imputed to the missing values.

**9.6.6: After Imputation finds the number of observations with missing values in the variable LOAN HISTORY.**

**SAS codes.**

****

**Output.**

****

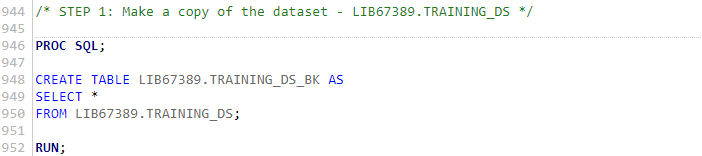
**Explanation**

The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the marital status variable.

## 9.7 Missing value imputation in continuous variables – Loan Amount

**9.7.1: Make a copy of the dataset**

**SAS codes.**

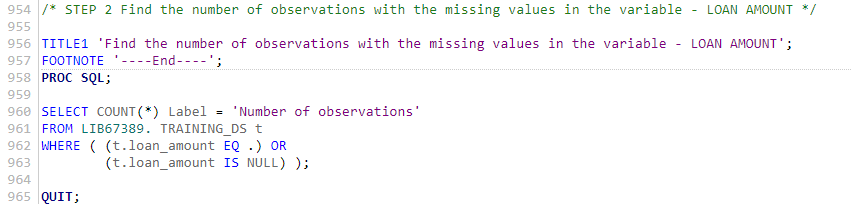
****

**Explanations**

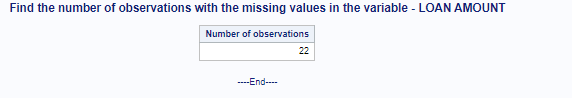
A backup of the training data set was created. The backup training set was created to protect against any unintentional changes made to the main training set. If any errors occurred on the main training set file, they could be recovered by using the backup file.

**9.7.2: The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output.**

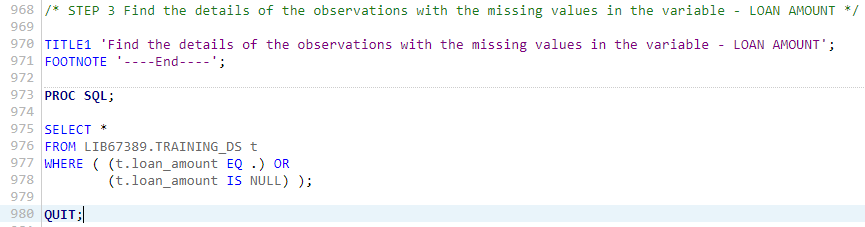
****

**Explanation**

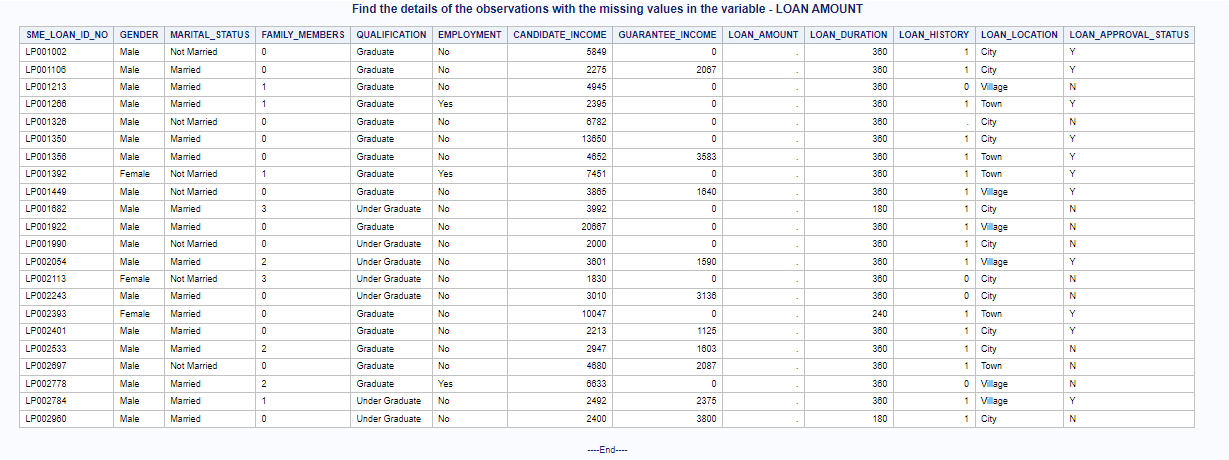
From the output above it can be observed that the Loan amount has 3 missing observations. An appropriate analysis is to be conducted to impute the missing values

**9.7.3:  Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output.**

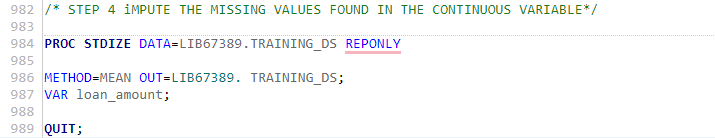
****

**Explanation**

The details of the observations of the missing values are further analyzed. It can be observed that only the loan amount observations are missing.

**9.7.4:** **Imputing the missing value found in the continuous variable**

**SAS codes.**

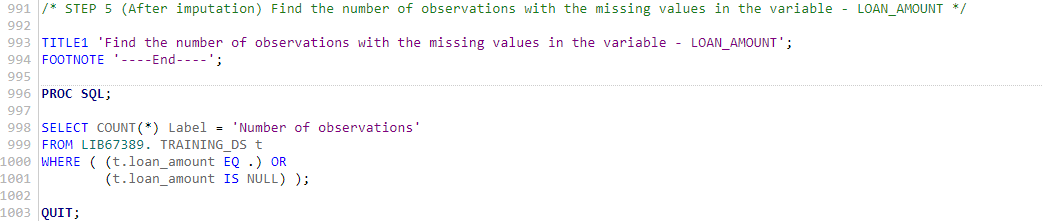
****

**Explanation.**

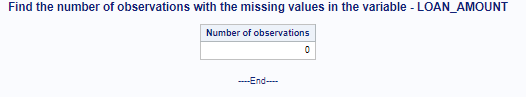
The imputation of continuous variables is different to that of categorical variables. From the above analysis, the mean of the loan amount was calculated and the resultant mean was imputed to replace the missing values

**9.7.5: After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output.**

****

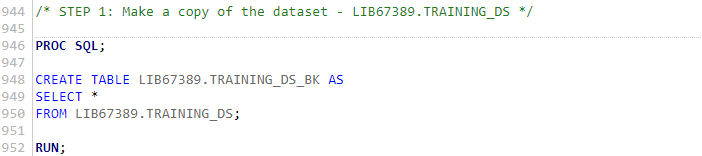
**Explanation**

The output above verifies the missing values are accurately imputed and that no more observations n null or missing values for the loan amount.

## 9.8 Missing value imputation in continuous variables – Loan Duration

**9.8.1: Make a copy of the dataset**

**SAS codes.**

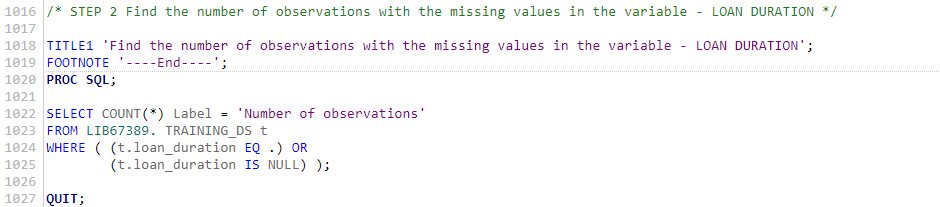
****

**Explanations**

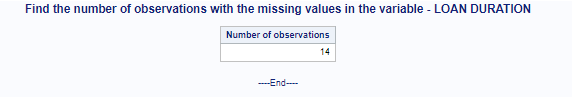
A backup of the training data set was created. The backup training set was created to protect against any unintentional changes made to the main training set. If any errors occurred on the main training set file, they could be recovered by using the backup file.

**9.8.2: The number of observations with the missing values in the variable**

**SAS codes.**

****

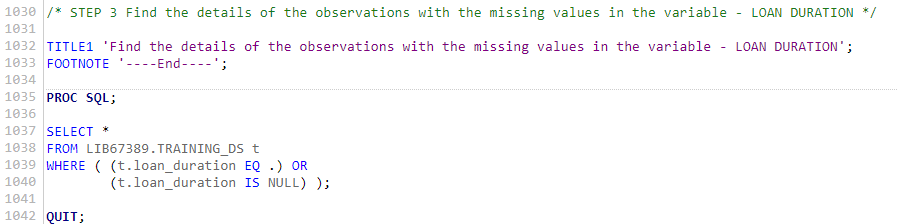
**Output.**

****

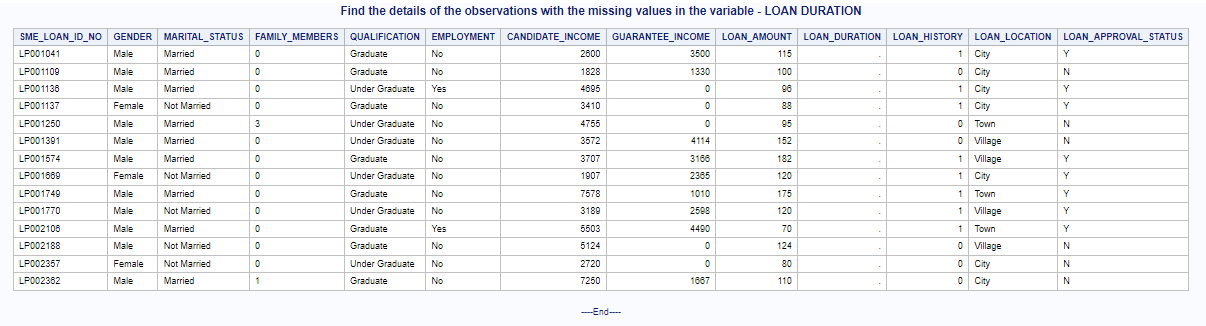
From the output above it can be observed that the loan duration has 14 missing observations. An appropriate analysis is to be conducted to impute the missing values

**9.8.3:  Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output**

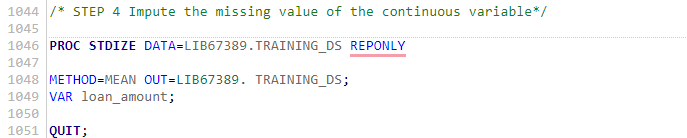
****

**Explanations**

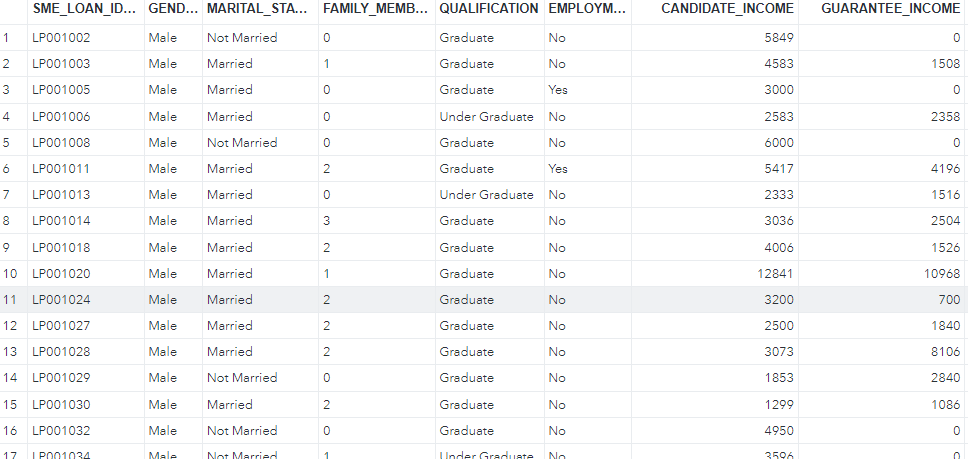
The details of the observations of the missing values are further analyzed. It can be observed that only the loan duration observations are missing.

**9.8.4:** **Imputing the missing value found in the continuous variable**

**SAS codes.**

****

**Output**

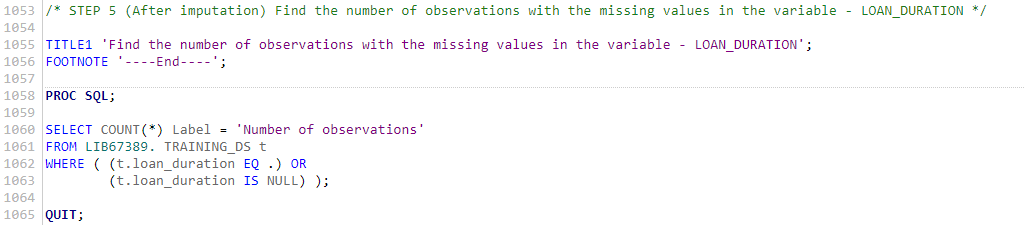
****

**Explanations**

From the above output, the mean of the loan duration was calculated and the resultant mean was imputed to replace the missing values

**9.8.5: After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output**

****

**Explanations**

The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the marital status variable

# Chapter 10.

# Data Exploration: TESTING\_DS

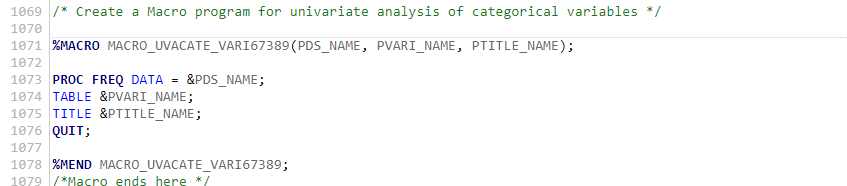
## 10.1 Introduction of SAS MACRO

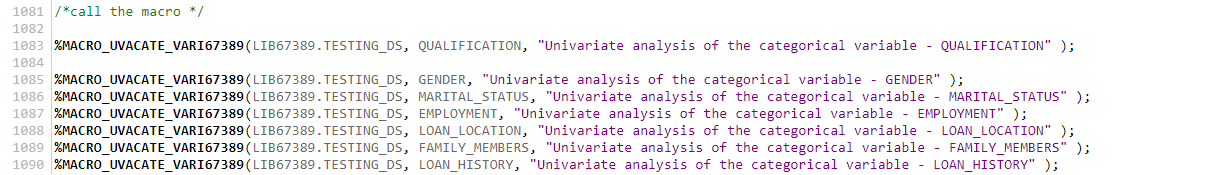
The SAS macro is a powerful and versatile tool. It is frequently used to reduce the amount of normal SAS code and to facilitate the transfer of information from one procedure to another. It can also be used to create SAS programmes that are "dynamic" and flexible.

Macros can be useful in a variety of ways. To begin, macros can be used to make a small change in a programme and have SAS echo that change throughout the programme. Second, macros enable to write a piece of code once and then reuse it in the same or different programmes. Third, it can make programmes data-driven, allowing SAS to make decisions based on actual data values.

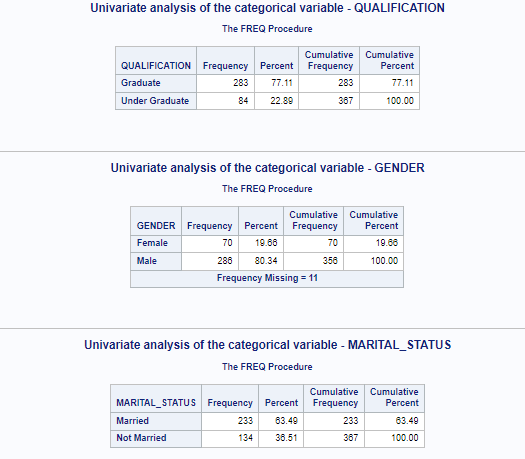
## 10.2 Univariate Analysis of the categorical variables using SAS MACRO

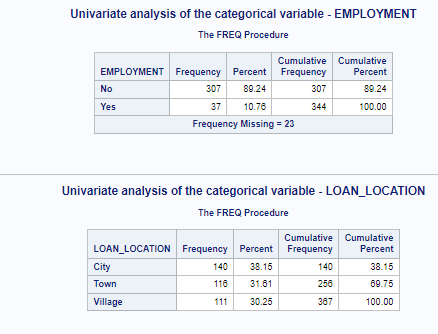
**SAS codes**

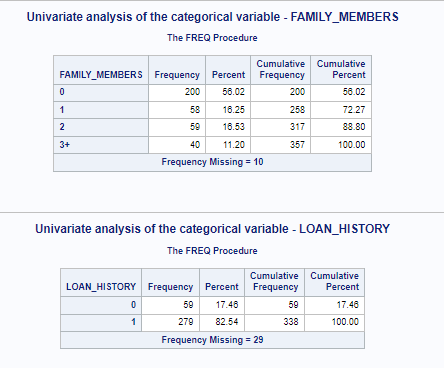




**Output**





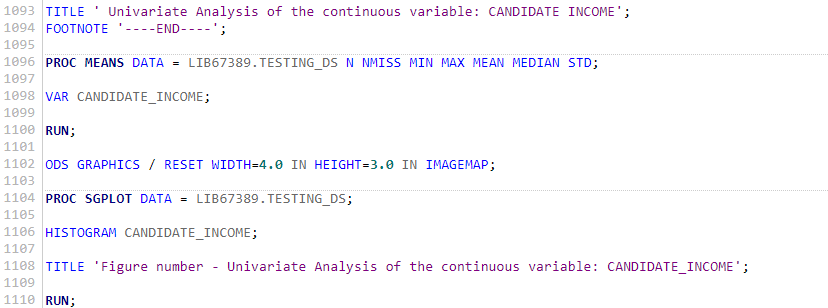


**Explanations**

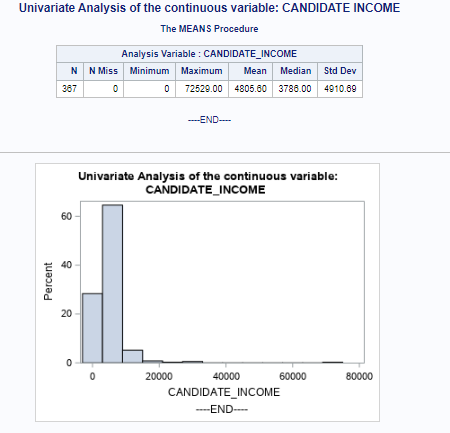
From the analysis, different observations are observed from the 7 categorical variables from the testing set. 4 of them contain missing values. Gender contains 11 missing values; loan history contains 29 missing values family members contain 10 missing values and finally, employment contains 23 missing values. Generally, all the variables have an adequate number of levels and only the data processing required is dealing with the missing values

## 10.3 Univariate Analysis of Numeric Variables – CANDIDATE\_INCOME

**SAS Code**

****

**Output**

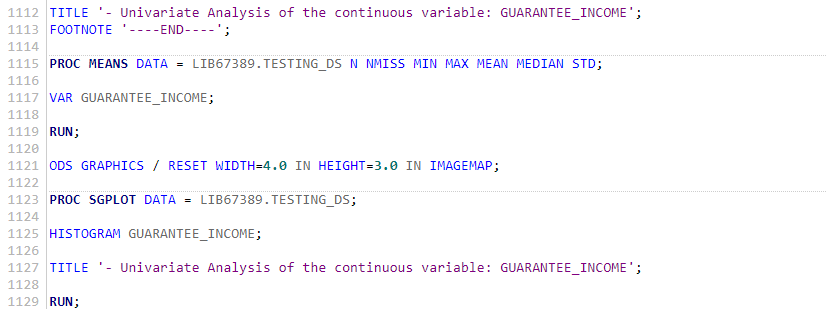
****

**Explanation**

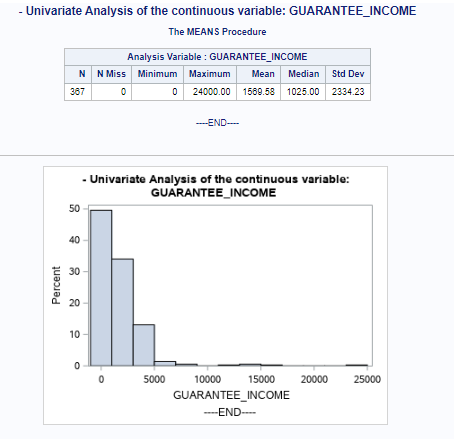
From the analysis, there are no missing values or applicants with unknown income in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the mean and median at 4805.60 and 3,786.00 respectively. The maximum value is greater than the (mean + 3x standard deviation) value, indicating that this variable contains extreme outliers.

## 10.4 Univariate Analysis of Numeric Variables – GUARANTEE\_INCOME

**SAS Code**

****

**Output**

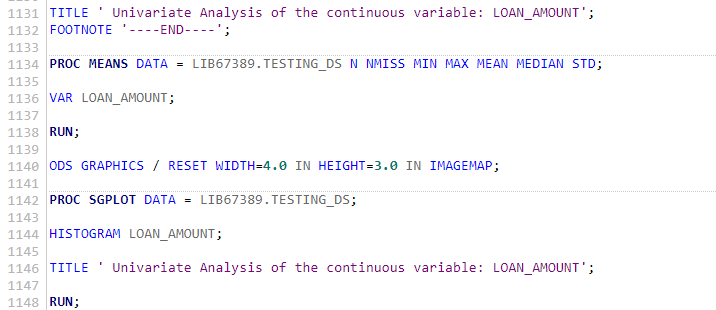
****

**Explanation**

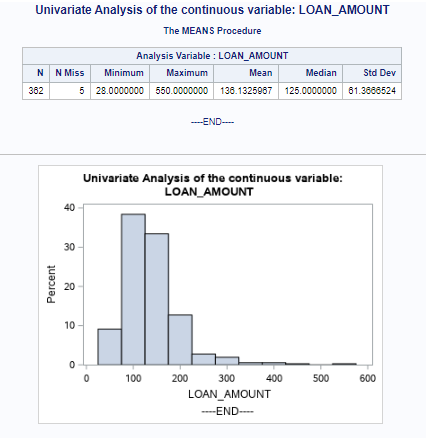
From the analysis, there are no missing values or applicants with unknown income in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the median and mean at 1025.00 and 1569.58 respectively. The maximum value is greater than the (mean + 3x standard deviation) value, indicating that this variable contains extreme outliers.

## 10.5 Univariate Analysis of Numeric Variables – LOAN\_AMOUNT

**SAS Code**

****

**Output**

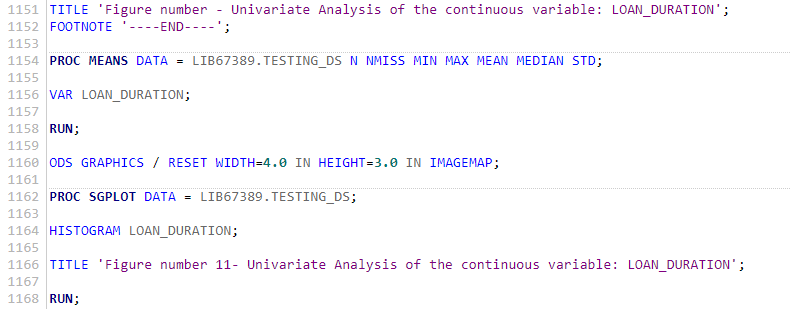
****

**Explanation**

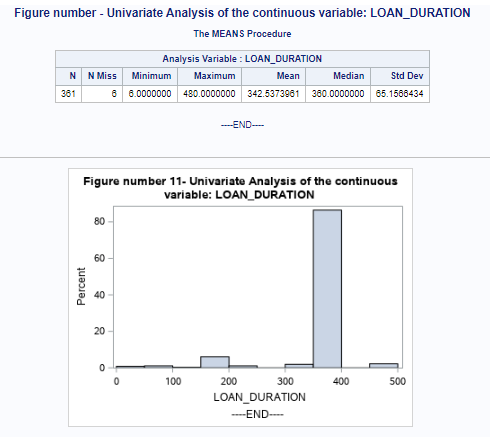
From the analysis, there are 5 missing values or applicants with unknown loan amounts in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the median and mean at 125 and 136.41 respectively.

## 10.5 Univariate Analysis of Numeric Variables – LOAN\_DURATION

**SAS Code**

****

**Output**

****

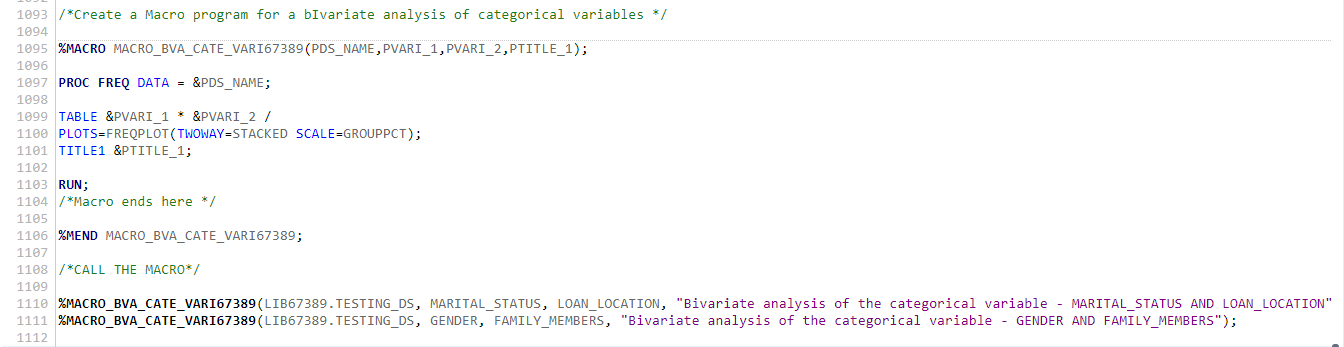
**Explanation**

From the analysis, there are 6 missing values or applicants with unknown loan amounts in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the median and mean at 360 and 342.53 respectively.

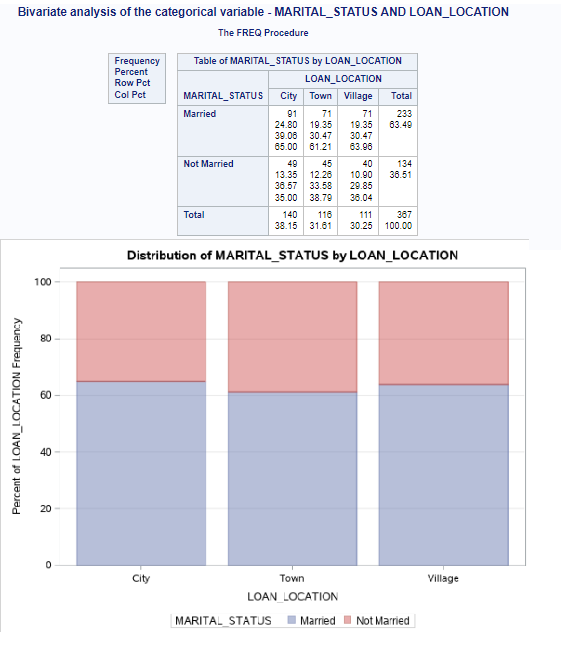
## 10.6 Bivariate analysis of categorical variables

**10.6.1 Bivariate analysis of categorical variables MARITAL\_STATUS vs LOAN\_LOCATION**

**SAS codes**



**Output**

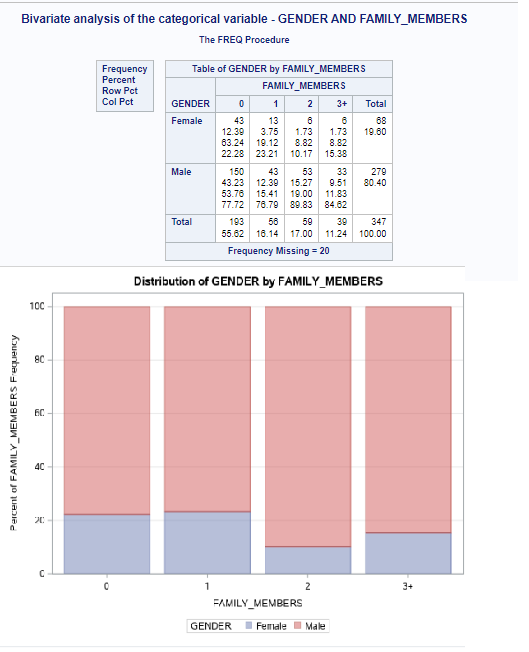


**Explanation**

From the above analysis, the proportion of loan locations varies according to the marital status of the applicants. 38.15% of the loan applicants are from the city, in which 65% of the applicants from the city are married while the rest are single. The analysis further shows that there is an equal distribution of the applicants in different loan locations. Generally, there is a high proportion of married applicants compared to the not married applicants.

**10.6.2 Bivariate analysis of categorical variables GENDER vs FAMILY\_MEMBERS**

**Output**

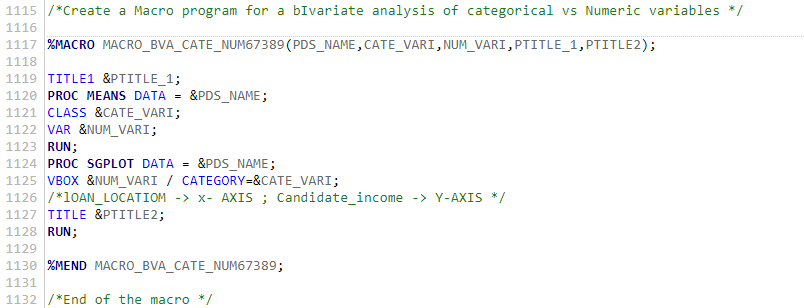


**Explanations**

From the bivariate analysis, there is a high proportion of the number of men applicants compared to females. This affects the distribution of the proportion of family members’ criteria in terms of gender. Out of the families with 0 and 1 family members, 77.72% and 76.69% are male applicants respectively. From the analysis, it can be observed that the proportion of males is higher in families with 2 and 3+ members this proves the fact that generally, the proportion of males is very higher compared to females.

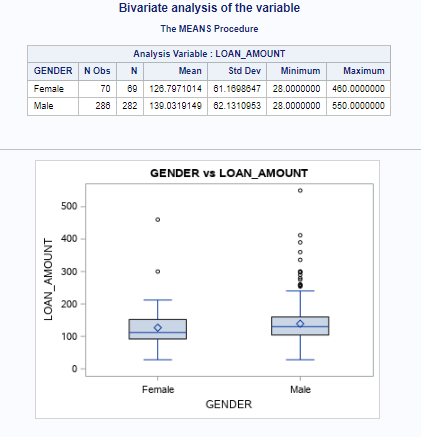
## 10.7 Bivariate analysis of categorical variables & numeric variables using SAS MACRO

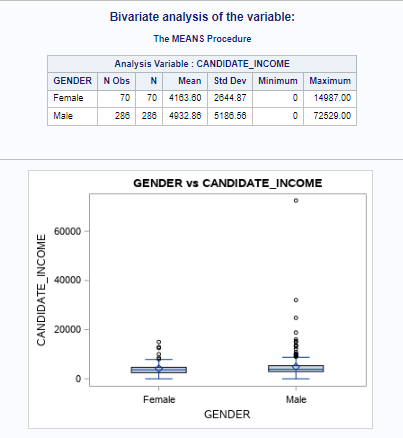
**SAS codes**





**Output**





**Explanation**

From the analysis, the mean loan amount of the male applicants is higher than that of the female applicants, which is the same as the SD perspective whereas the SD of the male participants is higher. This means the male applicants’ loan amount is more spread out than that of the female. This is further proved by the boxplot which shows there are more outliers in the male applicants compared to the female.

Similarly, to the previous analysis. The distribution of the candidate amount income and gender is the same with the male having an average higher candidate income compared to the females. This can be caused by the high number of males compared to females.

# Chapter 11:

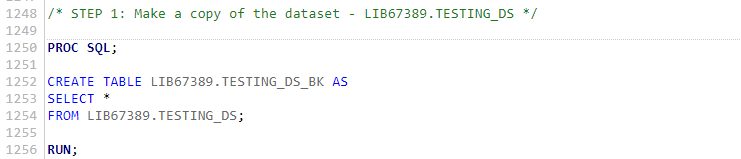
# Data Preprocessing: TESTING DS

## 11.1 Missing value imputation of categorical variables

### 11.1.1 Missing value imputation in the categorical variables – GENDER

**Make a copy of the dataset**

**SAS codes.**

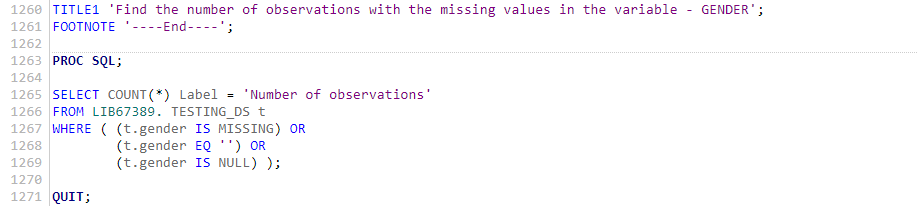
****

**Explanations**

A backup of the testing data set was created. The backup testing set was created to protect against any unintentional changes made to the main testing set. If any errors occurred on the main testing set file, they could be recovered by using the backup file.

**The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output**

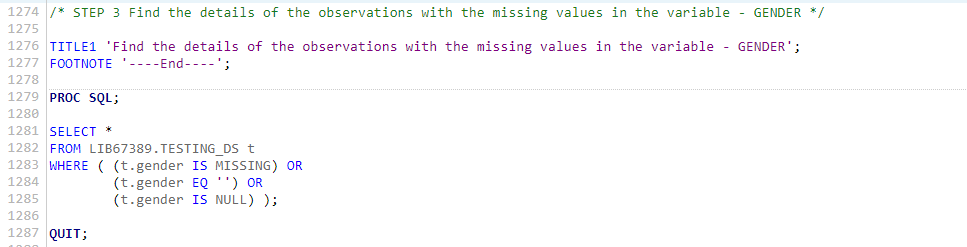
****

**Explanation**

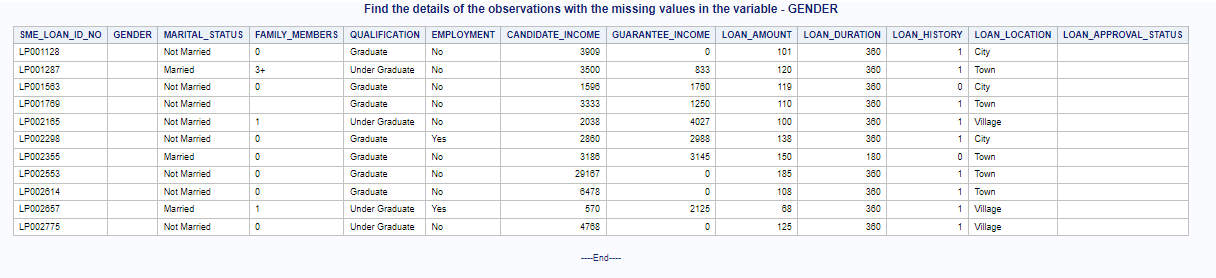
From the output above it can be observed that the gender has 11 missing observations. An appropriate analysis is to be conducted to impute the missing values

**Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output**

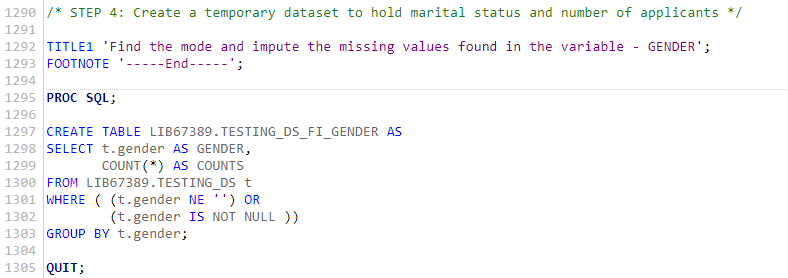
****

**Explanation**

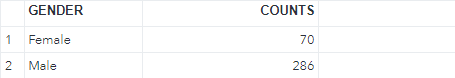
The details of the observations of the missing values are further analyzed. It can be observed that only the gender observations are missing.

**Creating a temporary dataset to hold marital status and the number of applicants to find the mode.**

**SAS codes.**

****

**Output**

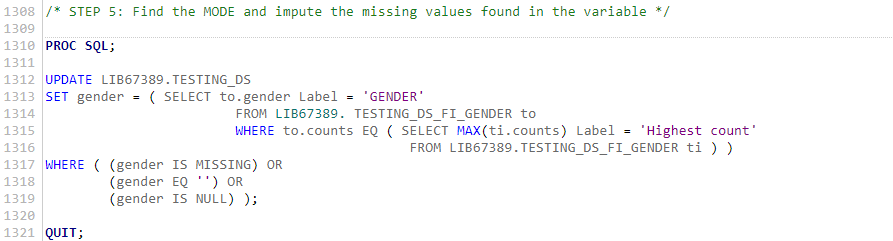
****

**Explanations**

**A** temporary dataset is created so that it can be easily used to calculate the number of observations on each category. The imputation method of categorical variables lies in imputing the missing values with the category with the highest mode.

**Find the Mode and impute the missing values in the variable –**

**SAS codes.**

****

**Output**

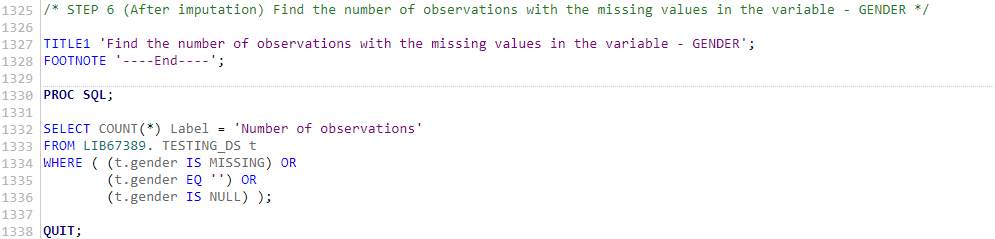
****

**Explanations**

From the previous analysis, the mode was calculated, and the male applicant’s category had the highest count. In this analysis, the mode calculated was used to be imputed to the missing values.

**After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output**

****

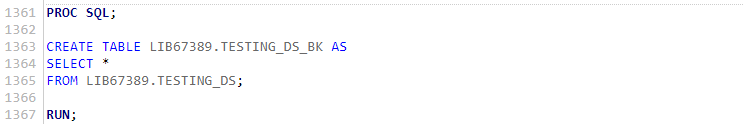
**Explanations**

The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the gender variable

### 11.1.2 Missing value imputation in the categorical variables – FAMILY MEMBERS

**Make a copy of the dataset**

**SAS codes.**

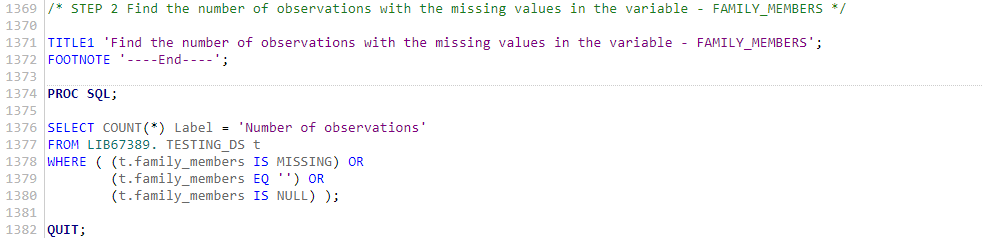
****

**Explanations**

A backup of the testing data set was created. The backup testing set was created to protect against any unintentional changes made to the main testing set. If any errors occurred on the main testing set file, they could be recovered by using the backup file.

**The number of observations with the missing values in the variable**

**SAS codes**

****

**Output**

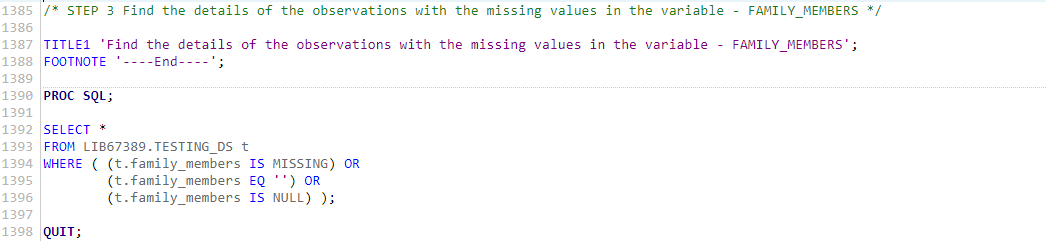
****

**Explanations**

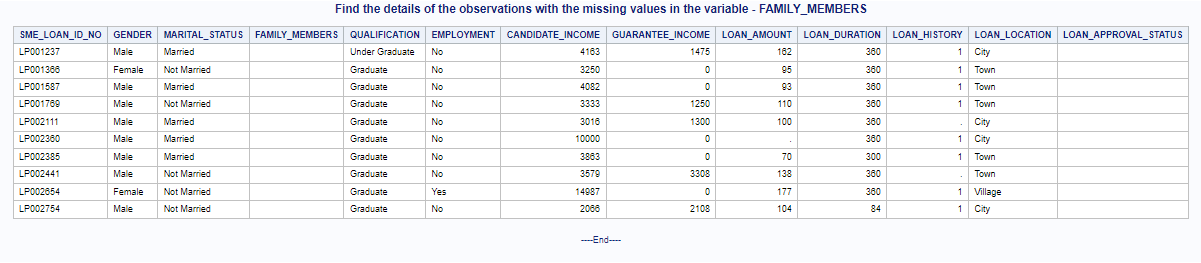
From the output above it can be observed that the family members have 10 missing observations. An appropriate analysis is to be conducted to impute the missing values

**Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output**

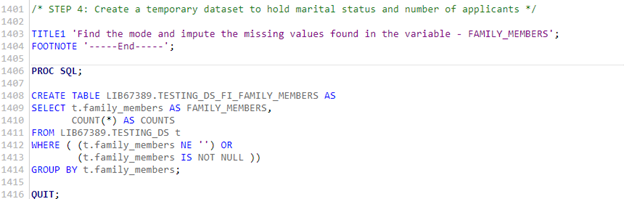
****

**Explanations**

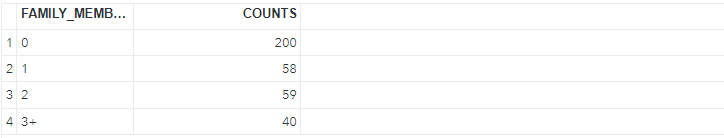
The details of the observations of the missing values are further analyzed. It can be observed that not only is the family members’ observations missing but also the loan amount and loan history are missing from some observations.

**Creating a temporary dataset to hold marital status and the number of applicants to find the mode.**

**SAS codes.**



**Output**

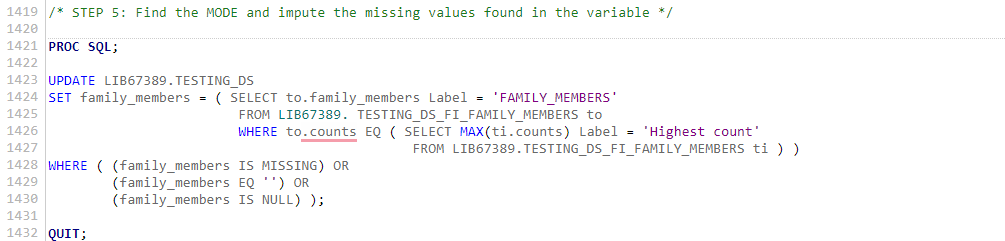
****

**Explanations**

**A** temporary dataset is created so that it can be easily used to calculate the number of observations on each category. The imputation method of categorical variables lies in imputing the missing values with the category with the highest mode.

**Find the Mode and impute the missing values in the variable –**

**SAS codes.**

****

**Output**

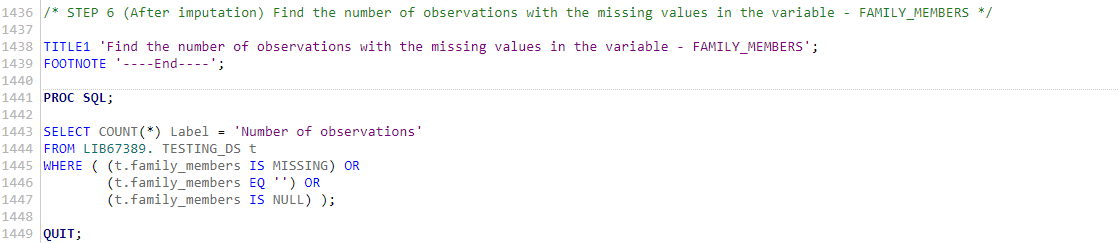
****

**Explanation**

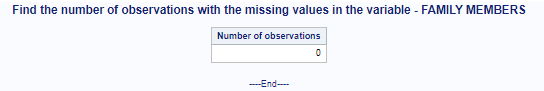
The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the family members’ variable

**After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Outputs**

****

**Explanation**

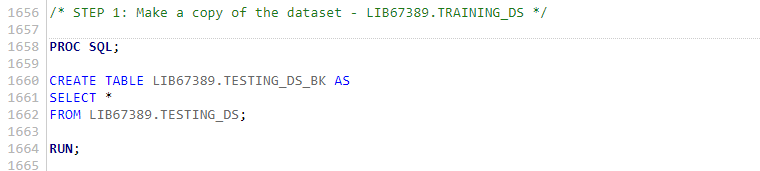
**The** output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the family members variable.

## 11.2 Missing value imputation of numeric variables

### 11.2.1 Missing value imputation in the numeric variables – LOAN AMOUNT

**Make a copy of the dataset**

**SAS codes.**

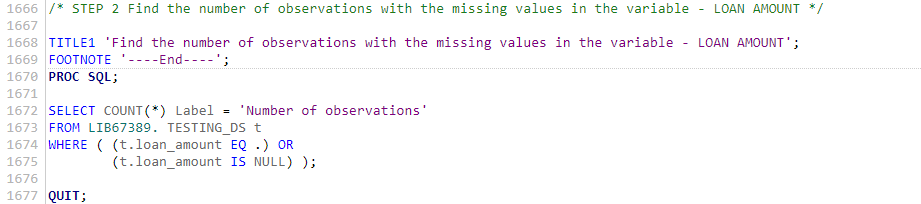
****

**Explanations**

A backup of the testing data set was created. The backup testing set was created to protect against any unintentional changes made to the main testing set. If any errors occurred on the main testing set file, they could be recovered by using the backup file.

**The number of observations with the missing values in the variable**

**SAS codes.**



**Output**

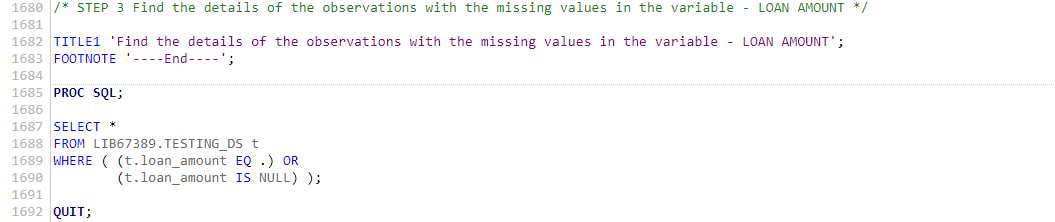
****

**Explanations**

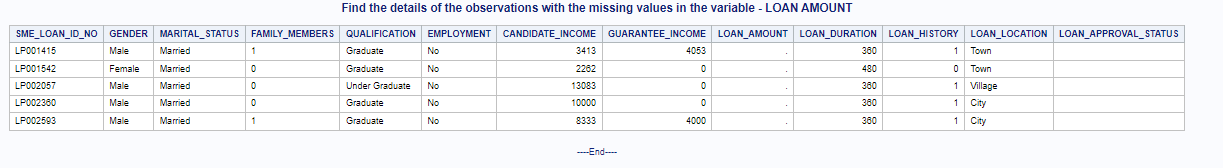
From the output above it can be observed that the Loan amount has 5 missing observations. An appropriate analysis is to be conducted to impute the missing values

**Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output**

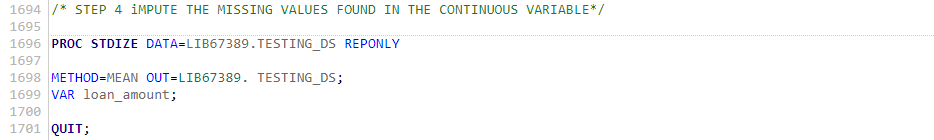
****

**Explanations**

The details of the observations of the missing values are further analyzed. It can be observed that only the loan amount observations are missing.

**Imputing the missing value found in the continuous variable**

**SAS codes.**

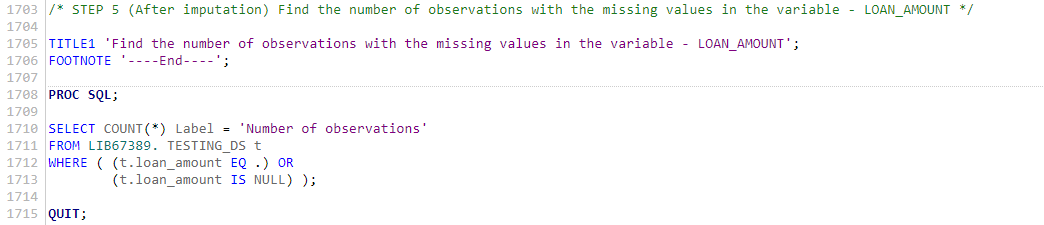
****

**Explanations**

The imputation of continuous variables is different to that of categorical variables. From the above analysis, the mean of the loan amount was calculated and the resultant mean was imputed to replace the missing values

**After Imputation find the number of observations with missing values in the variable**

**SAS codes.**

****

**Output**

****

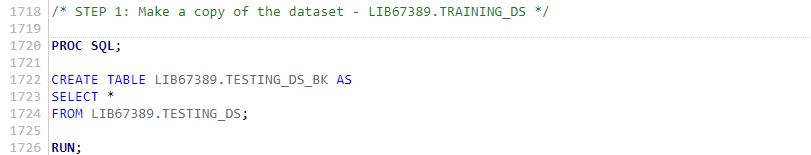
**Explanation**

The output above verifies the missing values are accurately imputed and that no more observations n null or missing values for the loan amount.

### 11.2.1 Missing value imputation in the numeric variables – LOAN DURATION

**Make a copy of the dataset**

**SAS codes.**

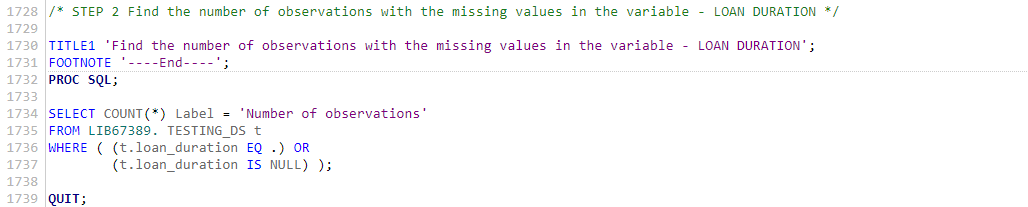
****

**Explanations**

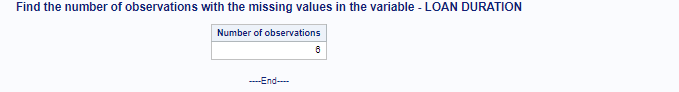
A backup of the testing data set was created. The backup testing set was created to protect against any unintentional changes made to the main testing set. If any errors occurred on the main testing set file, they could be recovered by using the backup file.

**The number of observations with the missing values in the variable**

**SAS codes.**

****

**Output**

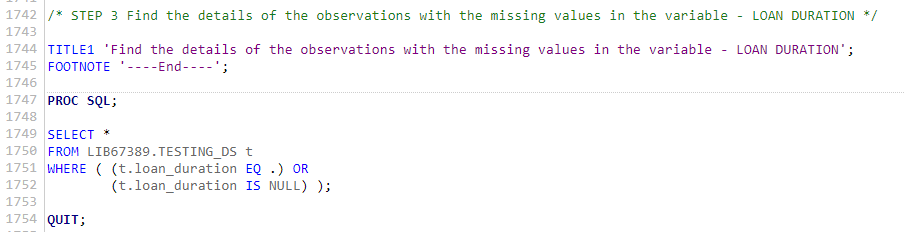
****

**Explanations**

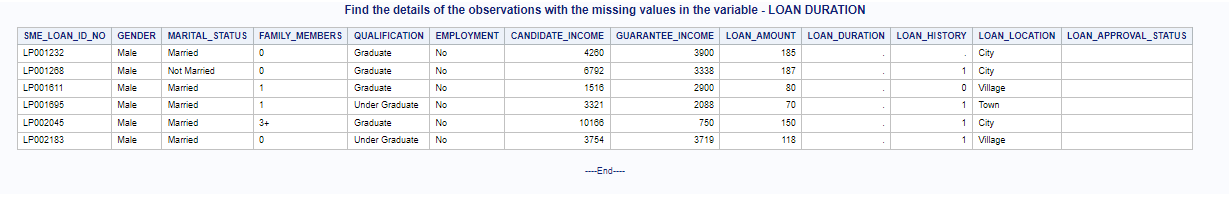
From the output above it can be observed that the loan duration has 6 missing observations. An appropriate analysis is to be conducted to impute the missing values

**Find the details of the observations with missing values in the variable**

**SAS codes.**

****

**Output**

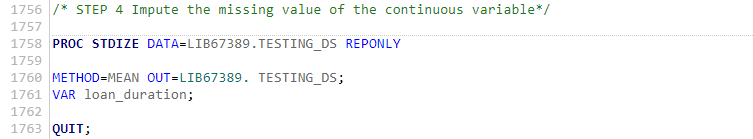
****

**Explanations**

The details of the observations of the missing values are further analyzed. It can be observed that only the loan duration observations are missing.

**Imputing the missing value found in the continuous variable**

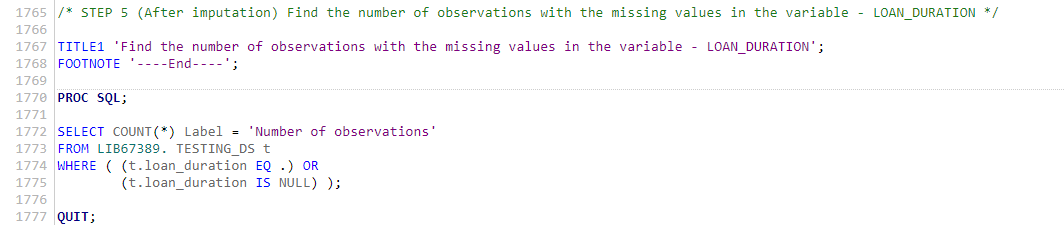
**SAS codes.**

****

**Explanations**

From the above code, the mean of the loan duration was calculated and the resultant mean was imputed to replace the missing values

**After Imputation find the number of observations with missing values in the variable**

**SAS codes** ****

**Outputs**

****

**Explanation**

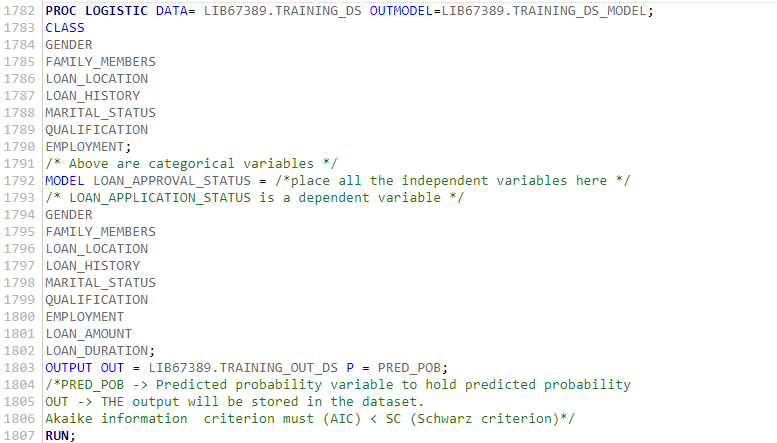
The output above verifies the missing values are accurately imputed and no more observations which contain null or missing values for the marital status variable

# Chapter 12:

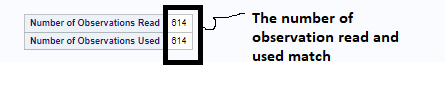
# Logistic regression model.

**Creating an L.R.M**

**SAS codes**

****

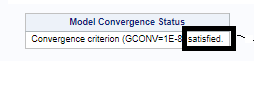
**Output**

****

**Explanations**

The dataset from which the model was built was properly cleaned. There are no longer any missing values as the number of observations read and used is the same enabling the model to use every observation in the dataset.

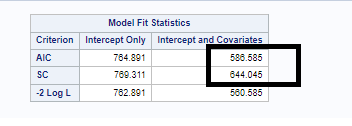
**Output**

****

**Explanations**

The Convergence Status table displays the iterative estimation process's status at the end of the optimization. It is satisfied by the logistic model.

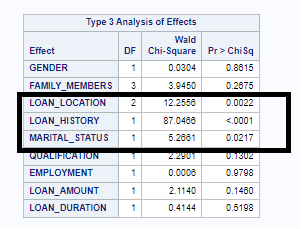
**Output**

****

**Explanations**

In the model fit statistic, it can be observed that the AIC is less than the SC. This demonstrates that the logistic regression model is an effective predictor.

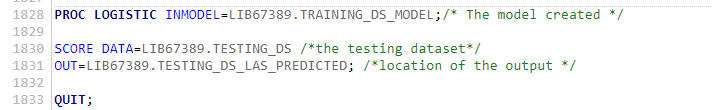
**Output**

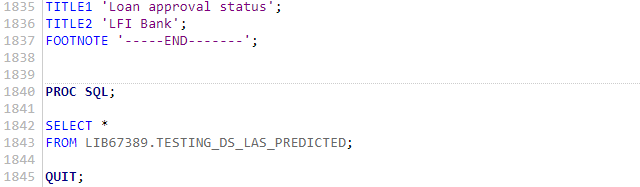
****

**Explanations**

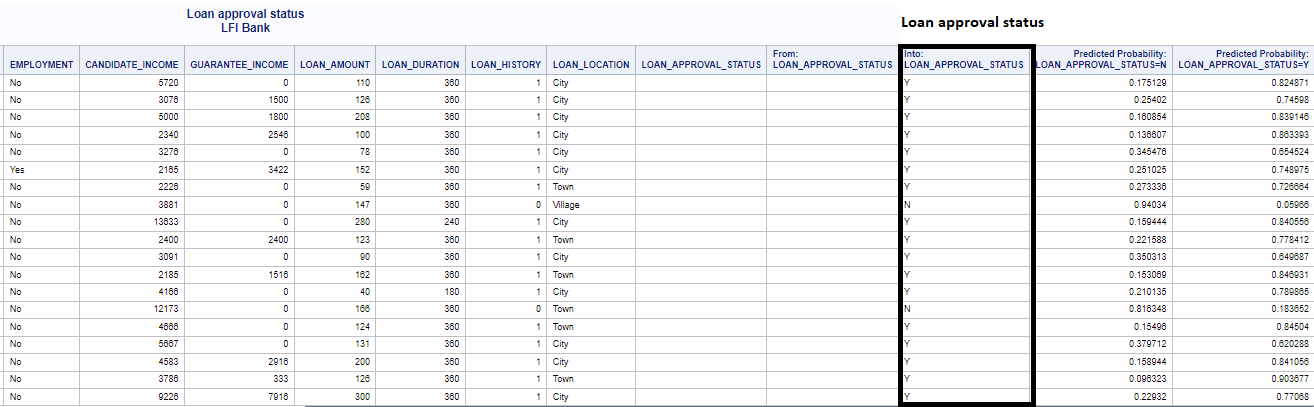
Because the p-value is less than 0.05, it is clear that marital status, loan history, and loan location have a significant correlation to loan approval status.

**SAS codes**

****

****

**Output**

****

**Explanation**

The above output shows the predicted loan approval status after the model prediction. The column comprises the estimated result from the model and the estimated probability.

# Chapter 13:

# SAS ODS – Output delivery system

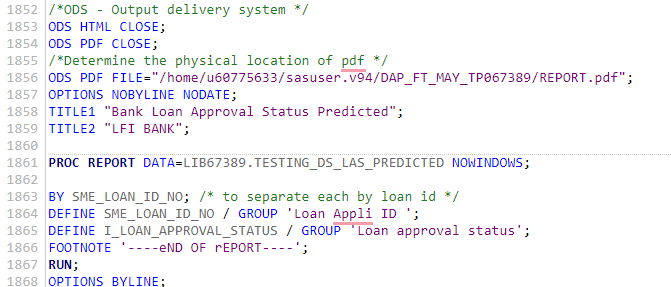
## 13.1 Introduction

The traditional SAS output is intended for use with a traditional line printer. The limitations of this type of output hinder the user from obtaining the most value from the results. ODS was created to address the limitations of traditional SAS output. It enables the delivery of output in a variety of formats and makes the formatted output easily accessible to SAS codes.

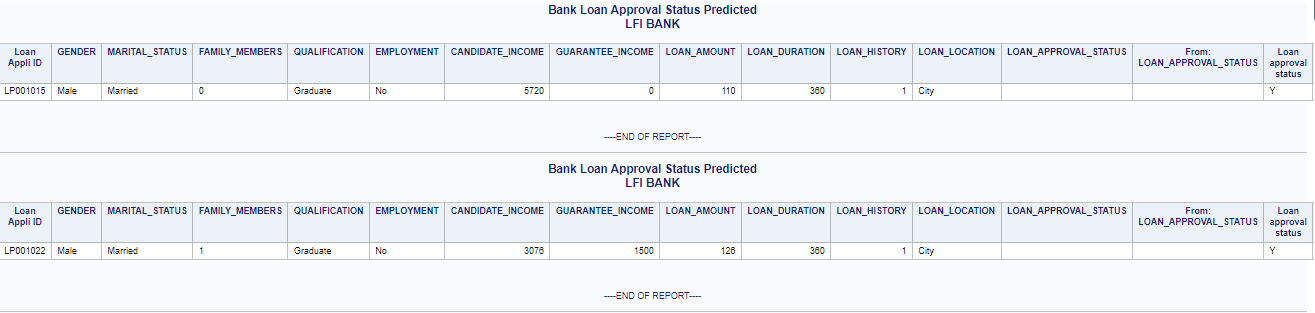
ODS opens up a whole new world of possibilities for generating high-quality, detailed SAS presentation output. ODS allows you to create files in a variety of formats, including HTML, Rich Text Format (RTF), PostScript (PS), Portable Document Format (PDF), and SAS data sets.

The SAS ODS outputs have been useful in various applications such as electricity and water bills, exam reports, etc.

**SAS Codes**

****

**Outputs**

****

**Explanations**

The SAS ODS is a report that explains the loan approval status for each applicant. In the report, the unique ID of the applicant is addressed, and all the other factors determined the loan status of the individual.

# CHAPTER 14:

# Conclusion

The data scientist has successfully created a model that can automatically approve loans in record time. This enables LFIs to process loans quickly, reducing their waiting list for credit approval and increasing customer satisfaction. Although, the LFI faces a default risk from lenders the model would help to accurately predict the clients who are eligible for the loans. Moreover, as the quality of the historical data from the clients increases also the accuracy in predicting the loan will be increasing there by reducing the misclassification rate of the approval status.

Personally, the project assisted me in learning how to code effectively in proc SQL. The lecturer understands the challenges that most students face when learning practical and technical skills such as coding. And thanks to his lessons, programming in SQL or SAS has become a lot easier. Moreover, the project has prepared me to work in the job environment as the lecturer consistently provided tips and tricks on how to thrive in the job place.

‌