**TABLE OF CONTENTS**

[1. Introduction 4](#_Toc108706584)

[1.1 Overview 4](#_Toc108706585)

[1.2 Assumption 5](#_Toc108706586)

[1.3 Objectives 5](#_Toc108706587)

[1.4 Tools used 6](#_Toc108706588)

[2. Dataset 6](#_Toc108706589)

[2.1. Qualitative Data 7](#_Toc108706590)

[2.2. Quantitative Data 7](#_Toc108706591)

[3. Method 7](#_Toc108706592)

[4. Initial Data Exploration 8](#_Toc108706593)

[4.1. Data Import: 8](#_Toc108706594)

[4.2 Classification of scales 9](#_Toc108706595)

[4.3 Outliers, Missing Values, and Outliers Treatment 10](#_Toc108706596)

[4.3.1 Identify Missing Values 10](#_Toc108706597)

[4.3.2 Identify Unique values 11](#_Toc108706598)

[5. Data Pre-processing 12](#_Toc108706599)

[5.1 Handling incomplete, noisy and inconsistent data 12](#_Toc108706600)

[Data Cleaning and Transformation: 12](#_Toc108706601)

[6. Exploratory Data Analysis (EDA) Graph 14](#_Toc108706602)

[6.1 Descriptive Analysis: 15](#_Toc108706603)

[6.2 Correlation: 16](#_Toc108706604)

[6.3 Univariate Analysis: 16](#_Toc108706605)

[6.3.1 Using a Single variable 17](#_Toc108706606)

[6.3.2 Using Multiple Variable 18](#_Toc108706607)

[6.4 Identify Outliers in the data 19](#_Toc108706608)

[6.5 Scatter Plot 22](#_Toc108706609)

[7. Apache Hadoop 31](#_Toc108706610)

[8. Hypothesis 41](#_Toc108706611)

[9. Conclusion 48](#_Toc108706612)

[10. REFERENCES 49](#_Toc108706613)

**List of Figures**

[Figure 1: Import of Sample dataset into SAS Studio 8](#_Toc108704705)

[Figure 2: Sample dataset view in SAS Studio 9](#_Toc108704706)

[Figure 3: Code to check missing values in the dataset 11](#_Toc108704707)

[Figure 4: Number of missing values in different variables 11](#_Toc108704708)

[Figure 5: Code for Identifying Unique Values 12](#_Toc108704709)

[Figure 6: Unique values in the Given Dataset 12](#_Toc108704710)

[Figure 7: Code for data cleaning for the 'GENDER' variable 13](#_Toc108704711)

[Figure 8: Output 1 14](#_Toc108704712)

[Figure 9: Code 2 14](#_Toc108704713)

[Figure 10: Output 2 15](#_Toc108704714)

[Figure 11: Code for cleaning of data 15](#_Toc108704715)

[Figure 12: Output for cleaning of data 16](#_Toc108704716)

[Figure 13: Code for descriptive Analysis 16](#_Toc108704717)

[Figure 14: Descriptive Analysis of the given dataset 17](#_Toc108704718)

[Figure 15: Correlation between different variables 17](#_Toc108704719)

[Figure 16: Code for determining correlation among numeric variables 18](#_Toc108704720)

[Figure 17::Code for using a single variable in univariate analysis 19](#_Toc108704721)

[Figure 18: Single variable output for 'CANDIDATE\_INCOME' 19](#_Toc108704722)

[Figure 19: Single variable output for 'LOAN\_AMOUNT' 20](#_Toc108704723)

[Figure 20: Code for using multiple variables 20](#_Toc108704724)

[Figure 21: Output 1 for multiple variable 21](#_Toc108704725)

[Figure 22: Output 2 for multiple-variable 21](#_Toc108704726)

[Figure 23: Code to check outliers in the 'EMPLOYMENT' category 22](#_Toc108704727)

[Figure 24: Code to check outlier in the 'GENDER' category 22](#_Toc108704728)

[Figure 25: Box plot for the above analysis of ‘CANDIDATE\_INCOME’ in terms of the ‘EMPLOYMENT’ variable 23](#_Toc108704729)

[Figure 26:Box plot for the above analysis of ‘CANDIDATE\_INCOME’ in terms of the ‘GENDER' variable 24](#_Toc108704730)

[Figure 27: Code for scatter plot using the relationship between two variables 25](#_Toc108704731)

[Figure 28: Scatter plot derived from the code 26](#_Toc108704732)

[Figure 29: Code for scatter plot for a loan amount 26](#_Toc108704733)

[Figure 30: Scatter plot derived from the code 27](#_Toc108704734)

[Figure 31: Code for simple linear regression 28](#_Toc108704735)

[Figure 32: Observations, analysis of variance, and parameters estimations 29](#_Toc108704736)

[Figure 33: Output for Simple Linear Regression 29](#_Toc108704737)

[Figure 34: Residuals for CANDIDATE\_INCOME 30](#_Toc108704738)

[Figure 35: Fit Diagnostics for CANDIDATE\_INCOME 31](#_Toc108704739)

[Figure 36: The output derived from the Multiple linear regression 32](#_Toc108704740)

[Figure 37: Graph Output using Multiple Linear Regression 33](#_Toc108704741)

[Figure 38: Residual by Regressor for Candidate Income using Multiple Linear Regression 34](#_Toc108704742)

[Figure 39: Architecture of Hadoop using the AWS services 35](#_Toc108704743)

[Figure 40: File from local (csv file) is uploaded into S3 bucket using CLI command. 36](#_Toc108704744)

[Figure 41: Creation of EC2 Key Pair 36](#_Toc108704745)

[Figure 42: Creating EC2 Key Pair 37](#_Toc108704746)

[Figure 43: Configurations added 38](#_Toc108704747)

[Figure 44: Master and slave group extra ports added 38](#_Toc108704748)

[Figure 45: Checking if master DNS is not present 39](#_Toc108704749)

[Figure 46: Generation of master DNS 39](#_Toc108704750)

[Figure 47: Hostname in Putty to make connection 40](#_Toc108704751)

[Figure 48: Writing code after connection 41](#_Toc108704752)

[Figure 49: Sum of Loan amount by grnder and mariatal status in Quick Site 42](#_Toc108704753)

[Figure 50: Sum of loan amount by loan\_Location using AWS QuickSite 42](#_Toc108704754)

[Figure 51: Sum of loan\_amount by Location and Qualification 43](#_Toc108704755)

[Figure 52: Complete visualization using AWS Quick Sight 43](#_Toc108704756)

[Figure 53: Code for Chi-Square Test 45](#_Toc108704757)

[Figure 54: Statistics for a table of GENDER and use of Fisher's Exact Test 46](#_Toc108704758)

[Figure 55: Chi-Square test for LOAN\_APPROVAL\_STATUS for MARIATAL\_STATUS 46](#_Toc108704759)

[Figure 56: Statistics for LOAN\_APPROVAL\_STATUS by MARIATAL\_STATUS 47](#_Toc108704760)

List of Tables

[Table 1:Classification of scales of measurement 10](#_Toc108704761)

[Table 2: Types of hypotheses in SAS 44](#_Toc108704762)

# Introduction

Data without a doubt is the most vital asset for any company, and management of data is inevitable for better quality and future decision-making processes. Machine learning is a useful process in the development of predictive and descriptive models, which further improves the preparation and understanding of data. Performance on the other hand depends on various pre-processing techniques. Our dataset can be assumed as a dataset from a financial institution that provides loan services to its customers. This research focuses on the pre-processing of future engineering techniques and the transformation of the dataset. SAS studio is an important machine learning tool that will be used for performing these techniques and with the help of metadata, the data will be explored, and processed and the graphs will be constructed and the hypothesis will be analyzed accordingly. Apache Hadoop will also be performed using the Amazon Web Services tools such as EMR and Athena. QuickSight will also be used for visualization of transformed data in the Hadoop.

## Overview

In our previous assignment, we came across various terminologies of feature engineering techniques such as feature extraction, selection, label coding, one-hot coding, and feature transformation. The metadata for the dataset was also discussed, and various analysis such as single variable analysis and cross tab analysis was formed. Similarly, PROC CONTENT and PROC FREQ of the data were calculated to find the missing frequency number. Overall, we got an insight that the total loans approved were 133 where 69 people had been rejected from the loan. Likewise, the total female population was 18.84%, which is much less than the male population of 81.36%.

## 1.2 Assumption

“XYZ Loan” is a financial institution that provides loan services to their customers, depending on the various attributes such as candidate income and guaranteed income, qualification, employment, and gender. The loan amount and approval status will depend upon these particular attributes. Future engineering in machine learning will guide us to predict how the loans are distributed and how the different attributes on the dataset impact the loan approval status in the dataset.

## 1.3 Objectives

The main purpose of this report is to understand the various data types and data storage systems at XYZ Loan company, and perform indexing and retrieving of data through various tools and techniques of data management. We are given a raw dataset that will be used and analyzed with the help of feature engineering techniques, and will then be transformed into meaningful data that will be able to make predictions for decisions. We will perform the following objectives.

* To explore the dataset for analysis through its classification in terms of nominal, interval, ordinal, and ratio data.
* To identify the mean, median, frequency, percentile, and variances to further support the initial exploration.
* To identify the outliers, treatment of outliers, and the missing values in the dataset.
* To perform data preprocessing to clean and organizes the data.
* To perform the Exploratory Data Analysis (EDA) to explain the relationships and characteristics of the various data using a graph for better visualization.
* To perform the hypothesis to interpret the query and visualization, and decide the conclusion.

## 1.4 Tools used

We will be using the future engineering techniques of the Statistical Analysis System, popularly known as SAS. Through the SAS studio, we shall clean, process, and use graphs and hypotheses to decide which attributes impact the loan approval least and most. The future engineering process in SAS studio will involve feature creation and transformation; feature extraction, exploratory data services (EDA), and, the benchmark. Various graphs including the scatter plot, pie-chart, and histogram will be used.

# Dataset

The dataset used in this project is a sample Dataset related to the income and loan contributed based on various categories like Gender, Qualification, Loan Location, and Employment status.

The following fields are being used in the given sample dataset.

* 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

## 2.1. Qualitative Data

The type of data which are based on description and interpretation and which cannot be counted in numbers is called qualitative data. In the provided dataset, the following fields represent the qualitative data.

* GENDER
* QUALIFICATION
* EMPLOYMENT
* LOAN\_LOCATION
* LOAN\_APPROVAL\_STATUS

## Quantitative Data

The type of data that is countable and represented in the form of numbers is called quantitative data. In the provided dataset, the following fields represent the quantitative data.

* CANDIDATE\_INCOME
* GUARANTEE\_INCOME
* LOAN\_AMOUNT
* LOAN\_DURATION

# Method

Our method will include the initial data exploration, data preprocessing, Exploratory data analysis, and hypothesis. We will go through them one by one, and will use SAS studio.

# Initial Data Exploration

## 4.1. Data Import:

The Sample dataset was imported into SAS Studio using the following code.

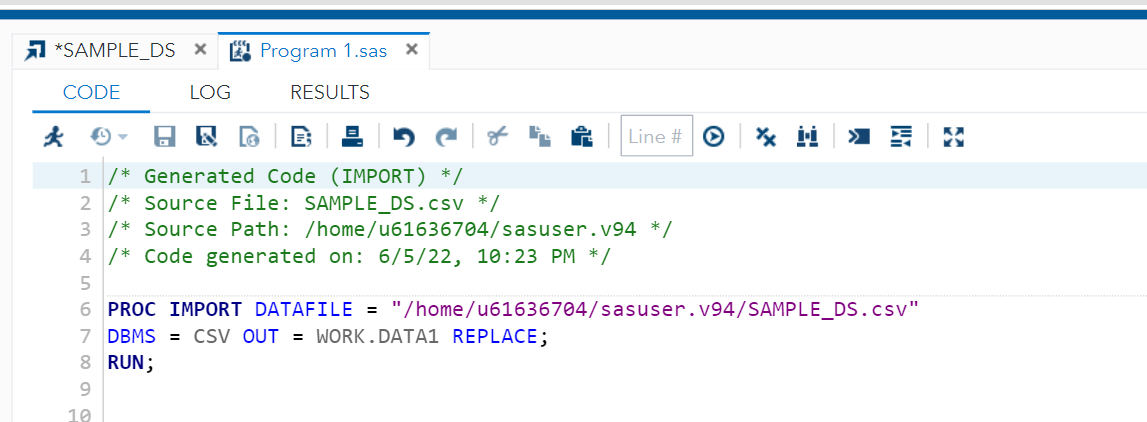


Figure 1: Import of Sample dataset into SAS Studio

The snapshot below shows the result of the import of data using the above code.

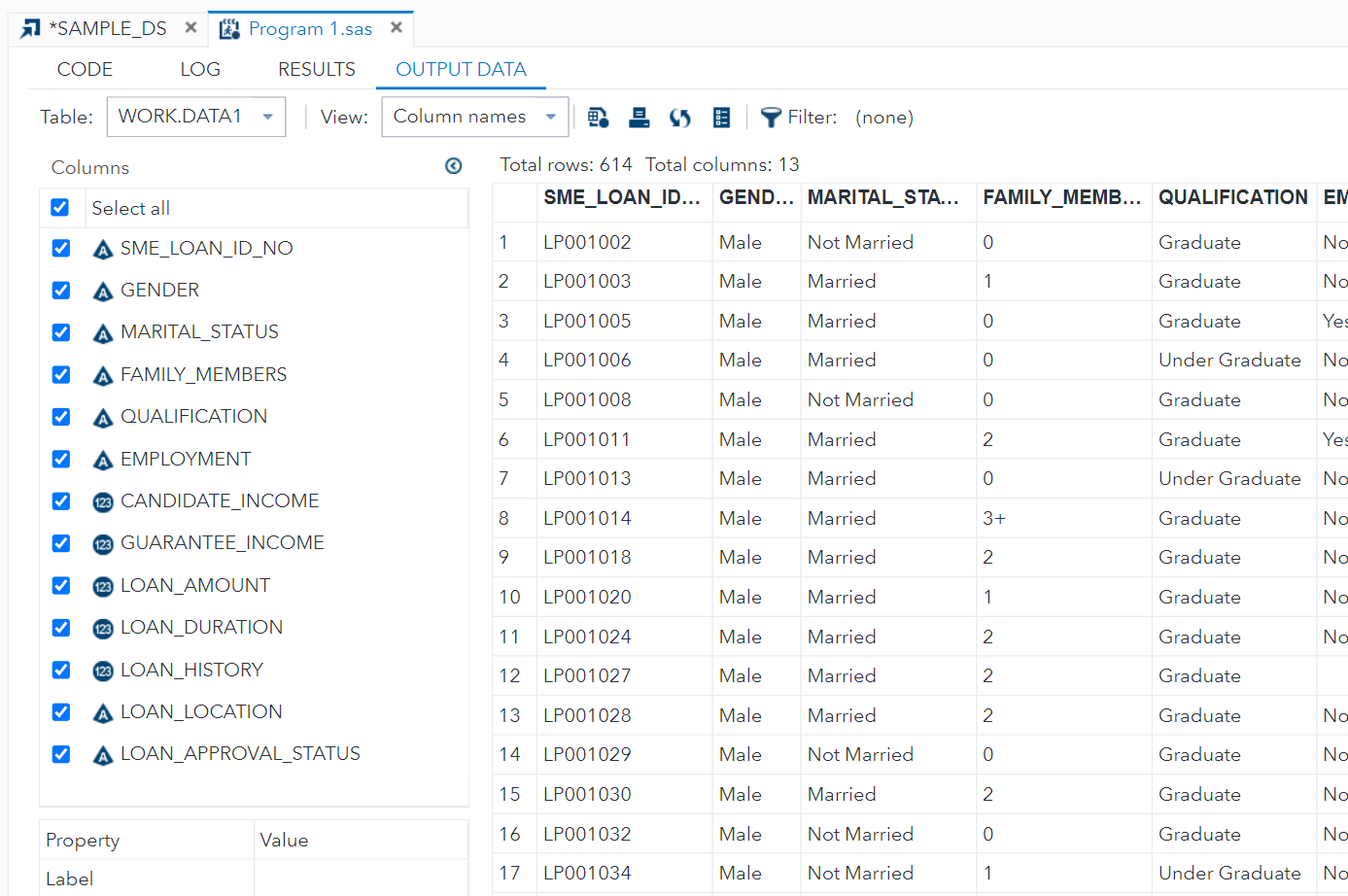


Figure 2: Sample dataset view in SAS Studio

## 4.2 Classification of scales

The fundamental scales of measurement that helps in the effective measurement of variables are usually in the form of questionnaires, surveys, and, question types. There are four types of scales which are nominal, ordinal, interval, and ratio scale.

A nominal scale is used in various classifications to label the required variables and is not quantitative They are the first level of measurement, and some of the examples of nominal scale include the gender and the view on politics.

The ordinal scale explains the frequency, happiness, and satisfaction data, and similar ideas but does not explain the numbers or any mathematical data. They are the second level of measurement of variables and may include examples such as grades, satisfaction, and the level of happiness.

Interval scales usually involve numeric values and may include the order and differences of variables. They are the third level of the measurement scale and examples include the temperature of a particular place or the income of the family.

The ratio scale not only measures the differences between variables and their order, but also the data revolving around zero. They are the fourth level of measurement scale and examples of such scale include the height and weight of a person.

(Land et al., 1990)

Now I will Classify the dataset of “XYZ” financial institution into the nominal, ordinal, interval, and ratio terms:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Nominal** | **Ordinal** | **Interval** | **Ratio** |
| **GENDER** | **Checkmark with solid fill** |  |  |  |
| **MARITAL\_STATUS** | **Checkmark with solid fill** |  |  |  |
| **FAMILY\_MEMBERS** | **Checkmark with solid fill** |  |  |  |
| **QUALIFICATION** |  | **Checkmark with solid fill** |  |  |
| **EMPLOYMENT** |  | **Checkmark with solid fill** |  |  |
| **CANDIDATE\_INCOME** |  |  |  | **Checkmark with solid fill** |
| **GUARANTEE\_INCOME** |  |  |  | **Checkmark with solid fill** |
| **LOAN\_ID** | **Checkmark with solid fill** |  |  |  |
| **LOAN\_DURATION** |  |  |  | **Checkmark with solid fill** |
| **LOAN\_HISTORY** |  |  |  |  |
| **LOAN\_LOCATION** | **Checkmark with solid fill** |  |  |  |
| **LOAN\_APPROVAL\_STATUS** | **Checkmark with solid fill** |  |  |  |

Table :Classification of scales of measurement

## 4.3 Outliers, Missing Values, and Outliers Treatment

### 4.3.1 Identify Missing Values

To identify the missing values in the quantitative data, the following code was used.

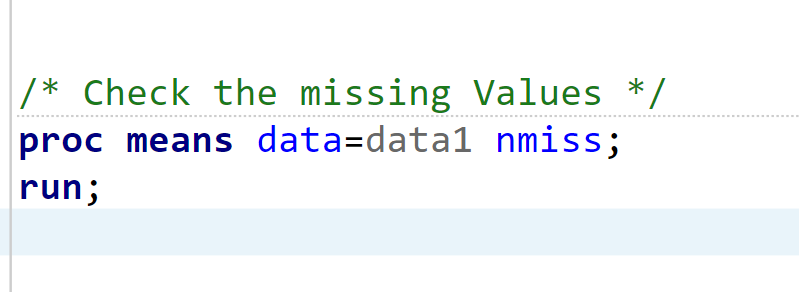


Figure 3: Code to check missing values in the dataset

The result from the above missing value code can be seen in the below screenshot.

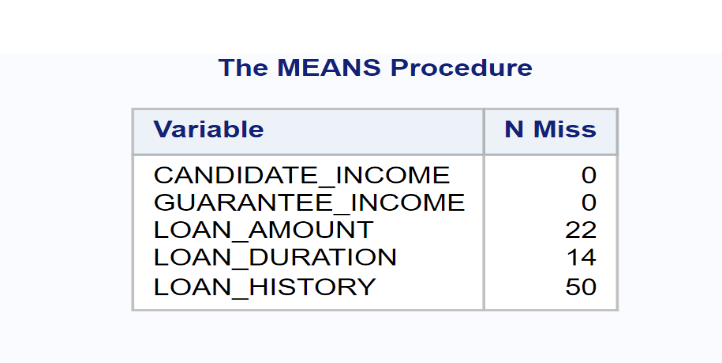


Figure 4: Number of missing values in different variables

## 4.3.2 Identify Unique values

The count of unique values in the given dataset in different qualitative variables was identified using the code below.

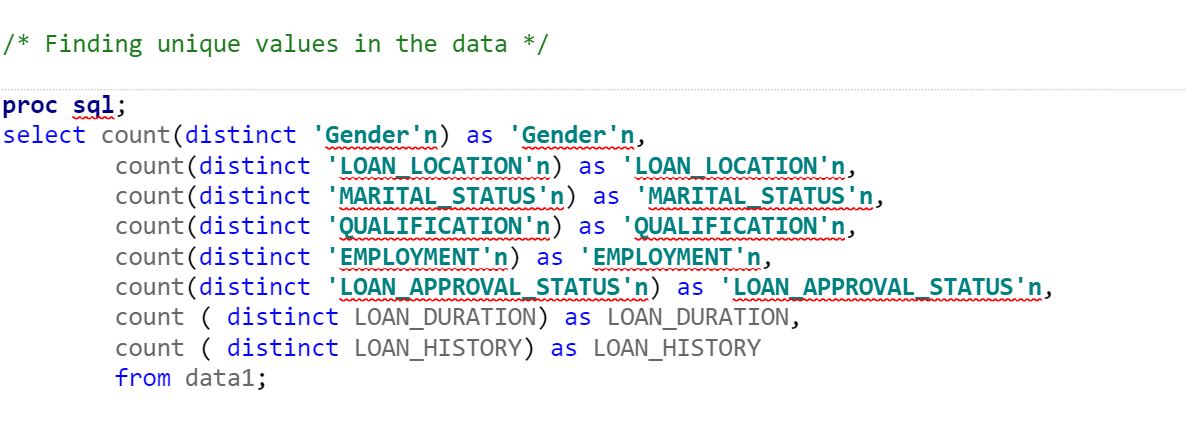


Figure 5: Code for Identifying Unique Values

Similarly, the output from the above code can be shown below.

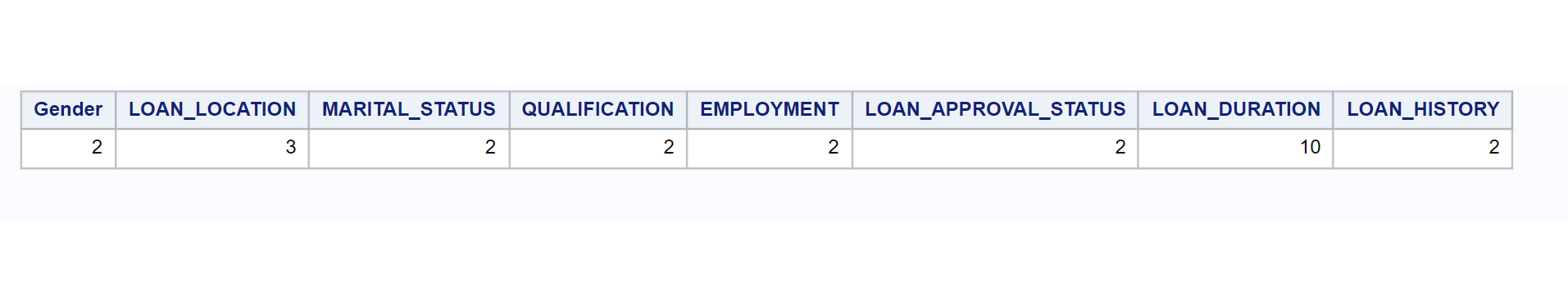


Figure 6: Unique values in the Given Dataset

# Data Pre-processing

Data pre-processing is the technique used before the data mining process that solves knowledge discovery issues. This preprocessing method makes the best use of the data mining algorithm and makes the data processing feasible. They usually involve preprocessing of the complex data and consume processing time. These methods use discipline processes such as data reduction techniques and come before the transformation, cleaning, and normalization method. And only after the completion of the data preprocessing stage, the data collected can be called reliable.

Data preprocessing not only involves classification and regression tasks to mine the data but also improves the models. In our dataset, the preprocessing will involve dealing with the imperfect data and finding the missing values, feature selection, and handling incomplete, inconsistent, and noisy data to clean, integrate and transform data.

(Garcia et al., 2016)

## 5.1 Handling incomplete, noisy and inconsistent data

## Data Cleaning and Transformation:

First of all, the NULL values in gender were replaced with ‘UNKNOWN’ values as a part of data cleaning.

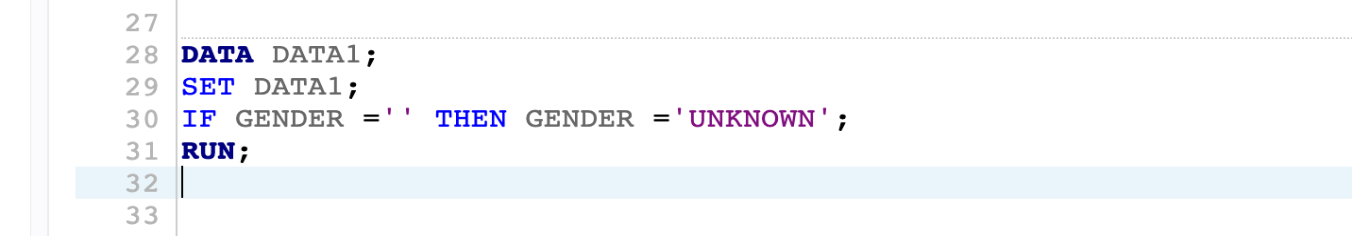
****

Figure : Code for data cleaning for the 'GENDER' variable

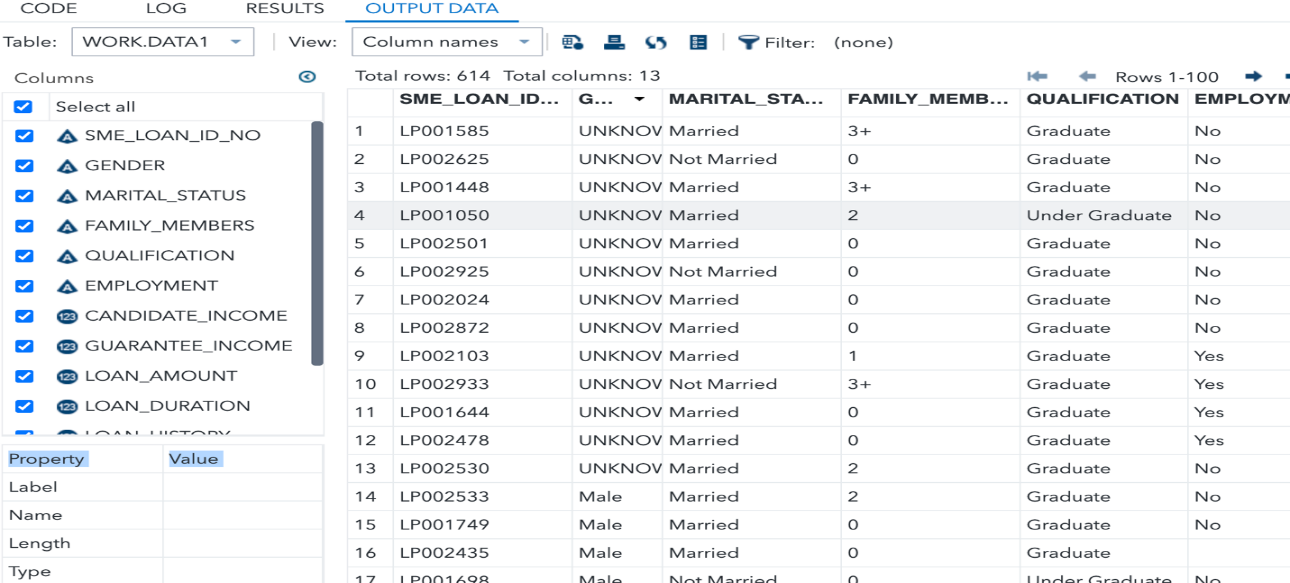
****

Figure : Output 1

Similarly, the NULL values in marital status were deleted since only a few records were contributing to NULL values. The code below shows the cleaning of Marital\_status.

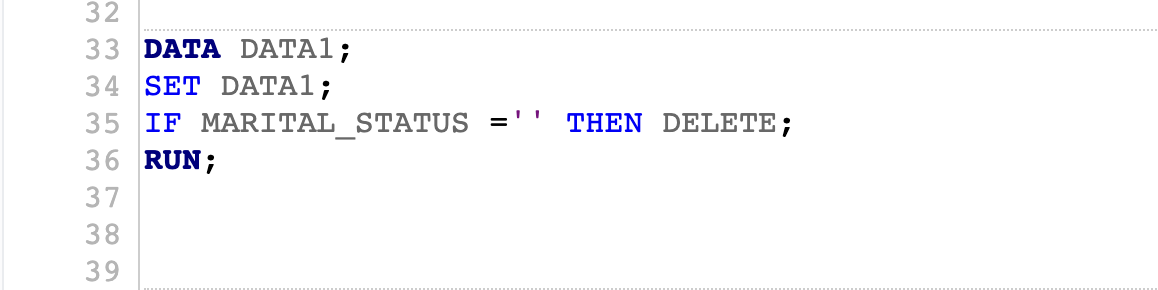
****

Figure : Code 2

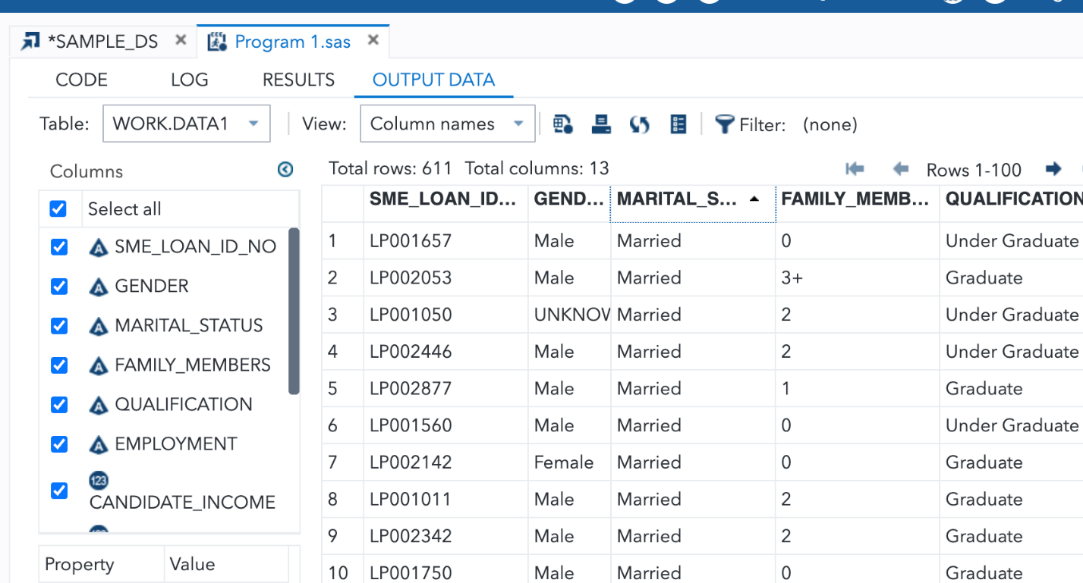
****

Figure : Output 2

From the above output, it is seen that the Gender value has inconsistent values i.e., some are in upper case and some are in regular case. The snapshot below shows the cleaning of inconsistent values in the data.

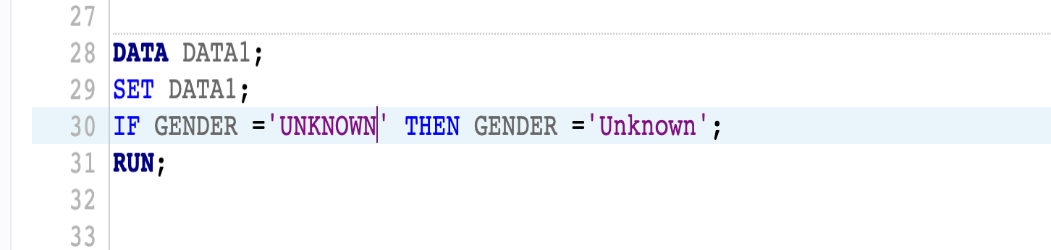


Figure : Code for cleaning of data

Now the consistency is maintained in the Gender variable.

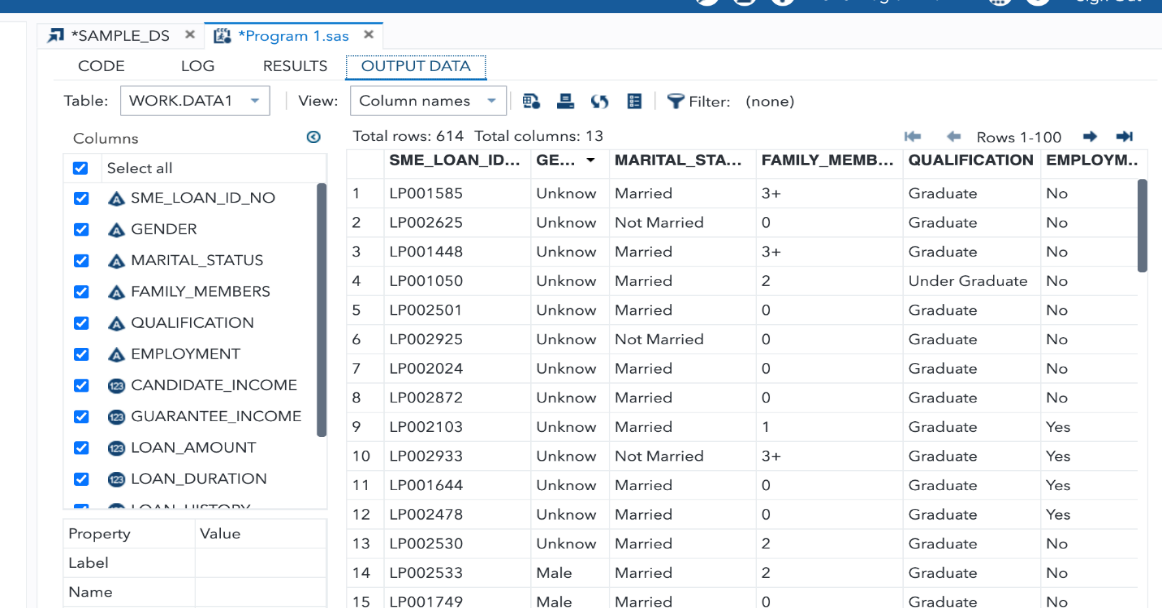


Figure : Output for cleaning of data

# Exploratory Data Analysis (EDA) Graph

In any research analysis of complex data, the EDA has been proven effective for distribution, outliers, and anomalies. These also include the graphs that will act as a tool for hypothesis purposes. The various techniques of the feature section also come under the EDA. EDA is widely regarded as the fundamental step after the collection of data and the pre-processing steps and involves the visualization of plotted and manipulated data to assist the development of quality in the models. Therefore, most of the EDA are graphical and qualitative and may also be univariate or multivariate. (Komorowski et al., 2016)

In our project, we will use the EDA to visualize the relationships between various variables, and will also detect our outliers and anomalies accordingly. Through the help of EDA, we shall understand the structure of our database, and we can create relevant assumptions and hypotheses based on our EDA.

## 6.1 Descriptive Analysis

The descriptive analysis including mean, median, mode, standard deviation, variance, minimum and maximum in the dataset was measured using the following code.

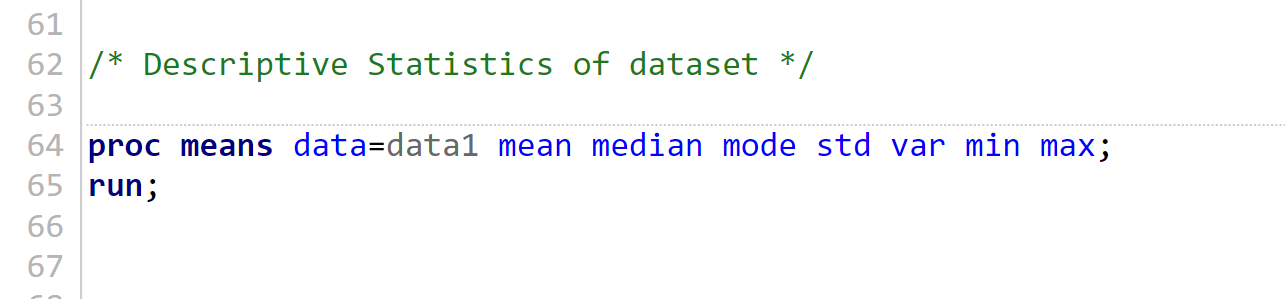


Figure 13: Code for descriptive Analysis

The results from the above code can be shown in the screenshot below.

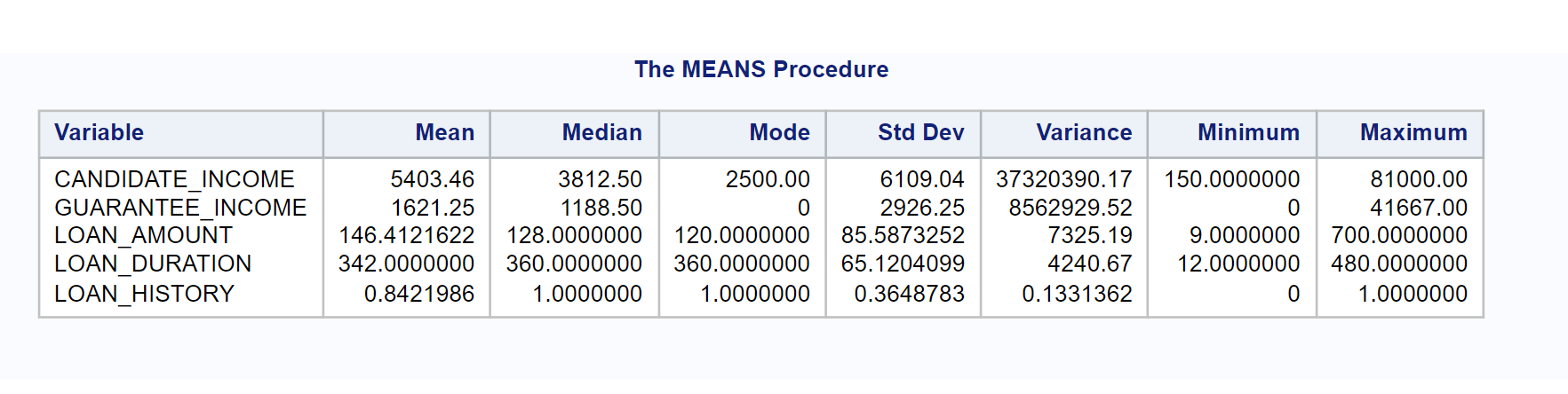


Figure 14: Descriptive Analysis of the given dataset

## 6.2 Correlation

The correlation between different variables can be shown in the snapshot below.

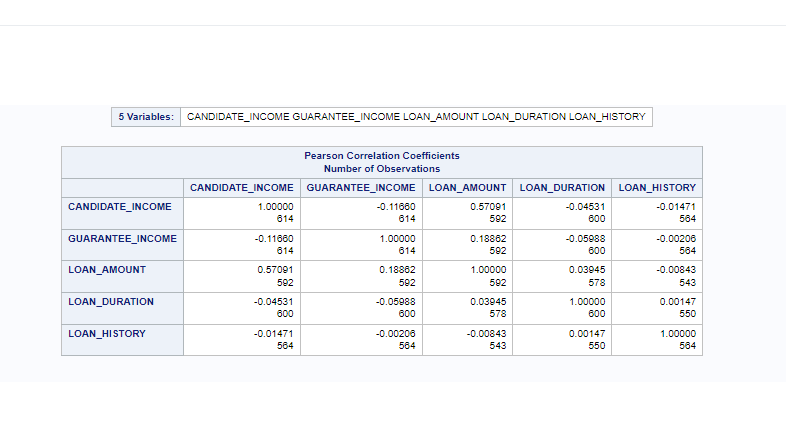


Figure 15: Correlation between different variables

The code for identifying the correlation between different variables in the dataset is shown below.

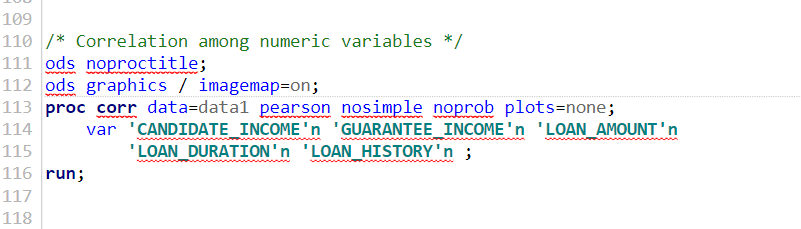


Figure 16: Code for determining correlation among numeric variables

## 6.3 Univariate Analysis

The univariate analysis in SAS is done through the proc univariate statement, which generates statistics to summarize the distribution of the data. The proc univariate analysis process is usually helpful in testing location, normality of data, extreme observations, the frequency of the observations, measuring the variability, and calculating the intervals and quantiles. (O Rourke et al., 2005)

In this project, we will perform the proc univariate analysis using a single variable as well as multiple variables.

### 6.3.1 Using a Single variable

Let us take the variable ‘CANDIDATE\_INCOME” from the dataset. We will use the ‘proc sgplot’ command to generate the histogram graph below.

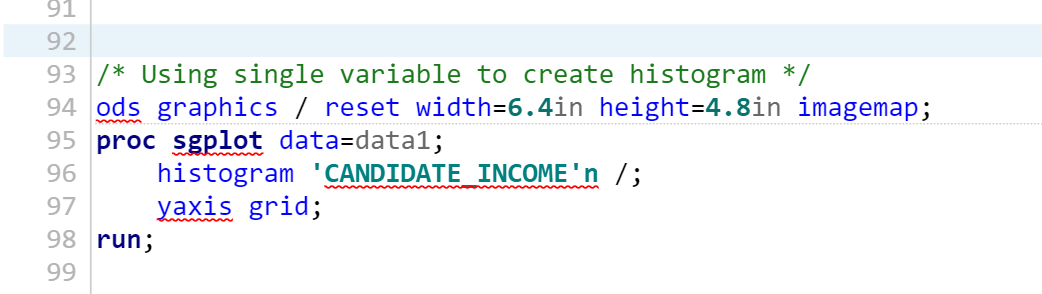
****

Figure ::Code for using a single variable in univariate analysis

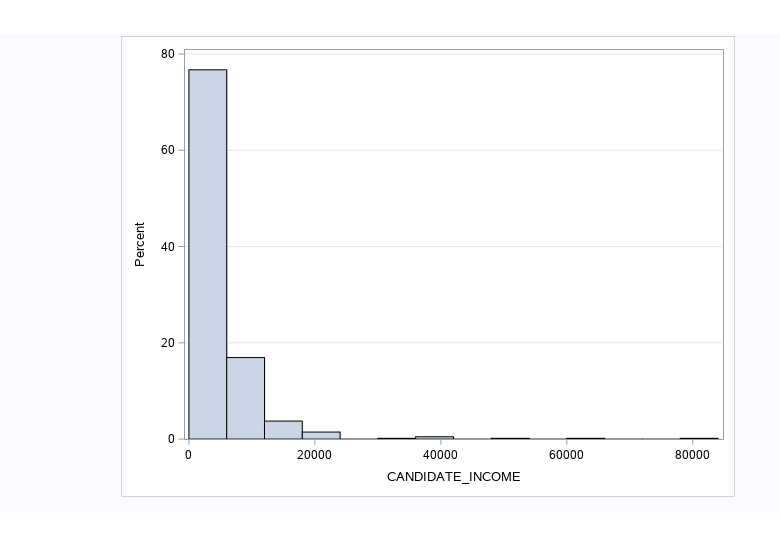


Figure : Single variable output for 'CANDIDATE\_INCOME'

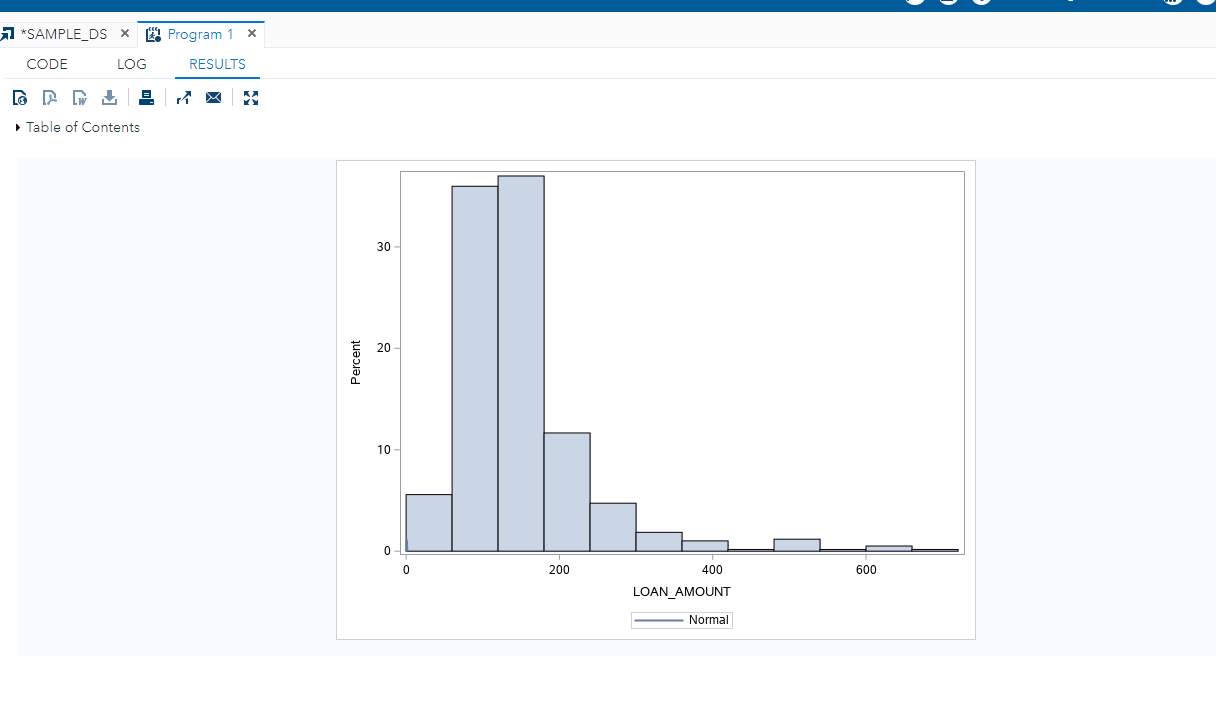
****

Figure : Single variable output for 'LOAN\_AMOUNT'

### 6.3.2 Using Multiple Variable

We will now use two variables i.e., ‘CANDIDATE\_INCOME’ and ‘LOAN\_AMOUNT’ from our dataset, and use the ‘proc univariate’ command to create the histogram from the analysis.

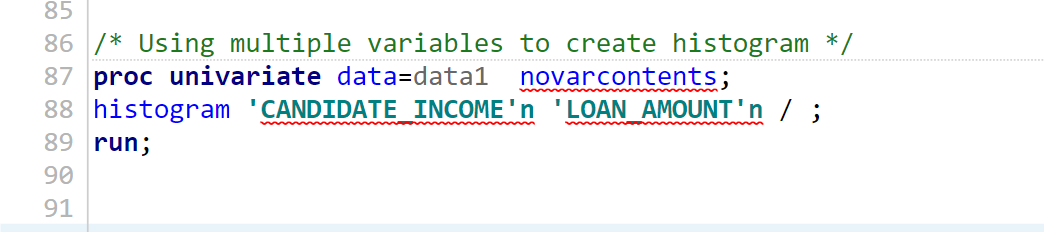
****

Figure : Code for using multiple variables

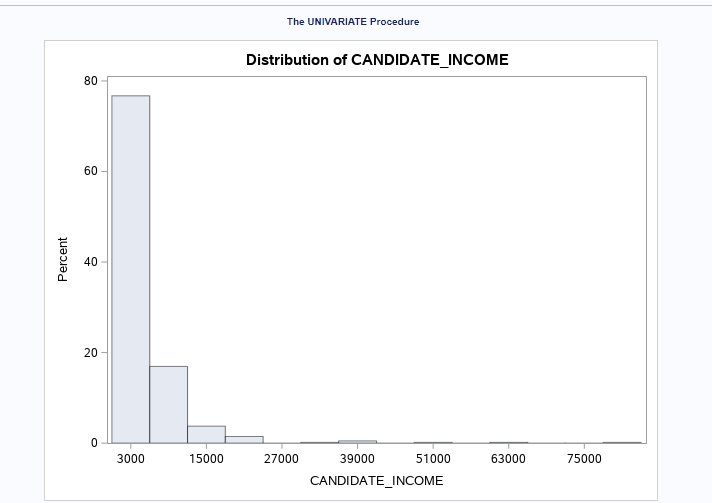


Figure : Output 1 for multiple variable

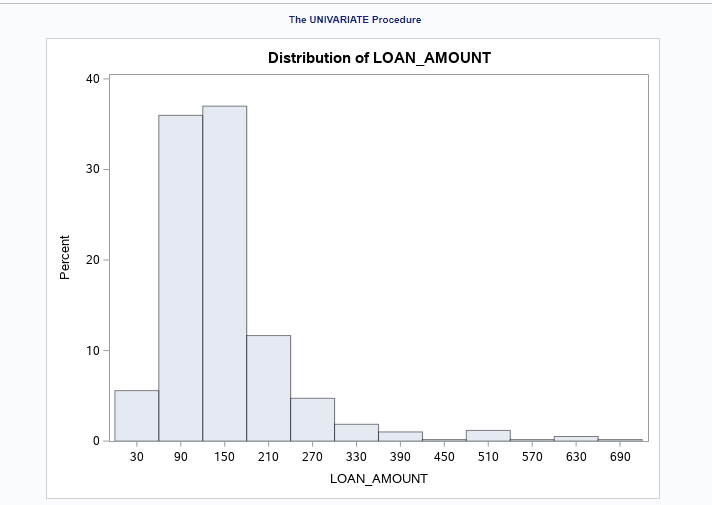


Figure : Output 2 for multiple-variable

## 6.4 Identify Outliers in the data

We will now identify the outliers in the univariate analysis using the single variable as well as the multiple variables. Outliers can be identified by using the following code. We will be using the ‘CANDIDATE\_INCOME’ variable in the box and ‘EMPLOYMENT’ as the category.

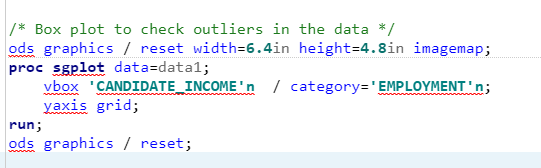


Figure : Code to check outliers in the 'EMPLOYMENT' category

Similarly, the ‘GENDER’ category is used for identifying the outliers below.

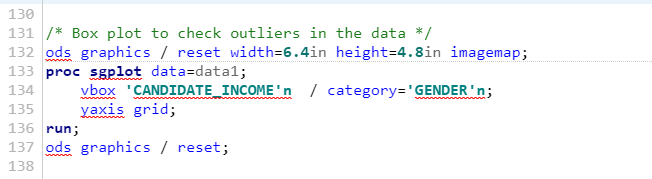


Figure : Code to check outlier in the 'GENDER' category

The box plot for the above analysis of CANDIDATE\_INCOME in terms of the ‘EMPLOYMENT’ variable is given below.

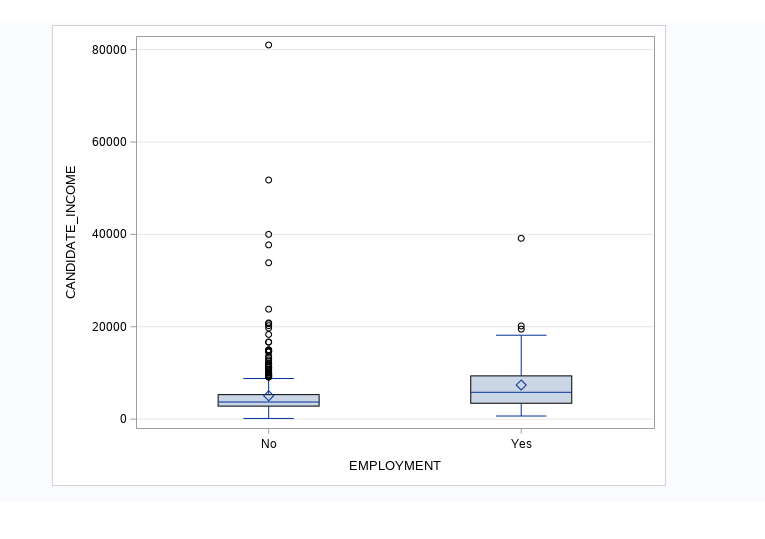


Figure 25: Box plot for the above analysis of ‘CANDIDATE\_INCOME’ in terms of the ‘EMPLOYMENT’ variable

The box plot for the above analysis of CANDIDATE\_INCOME in terms of the ‘GENDER’ variable is given below.

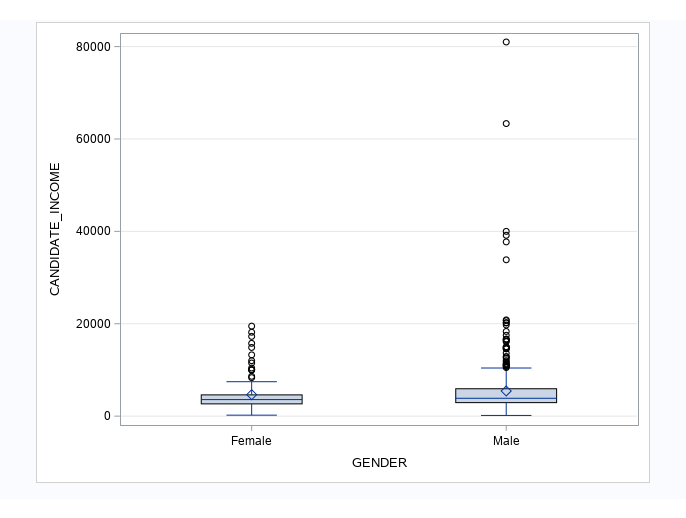


Figure :Box plot for the above analysis of ‘CANDIDATE\_INCOME’ in terms of the ‘GENDER' variable

## 6.5 Scatter Plot

A scatter plot analysis is a scatter graph that helps to explain values for a single variable as well as multiple variables. Each dot on the vertical, as well as horizontal axis, is formed on the graph, which will be helpful to find the relationships between the numeric variables.

Correlational relationships and identification of patterns within data are easier with the use of scattering lot. Furthermore, unexpected gaps and outliers can be easily observed, while segmentation of data is also possible with scatter plots. However, issues such as overplotting as well as interpreting correlation should be kept in mind while performing this analysis.

(Petruccelli et al., 1999)

A Scatter plot using two variables can be shown in the code below.

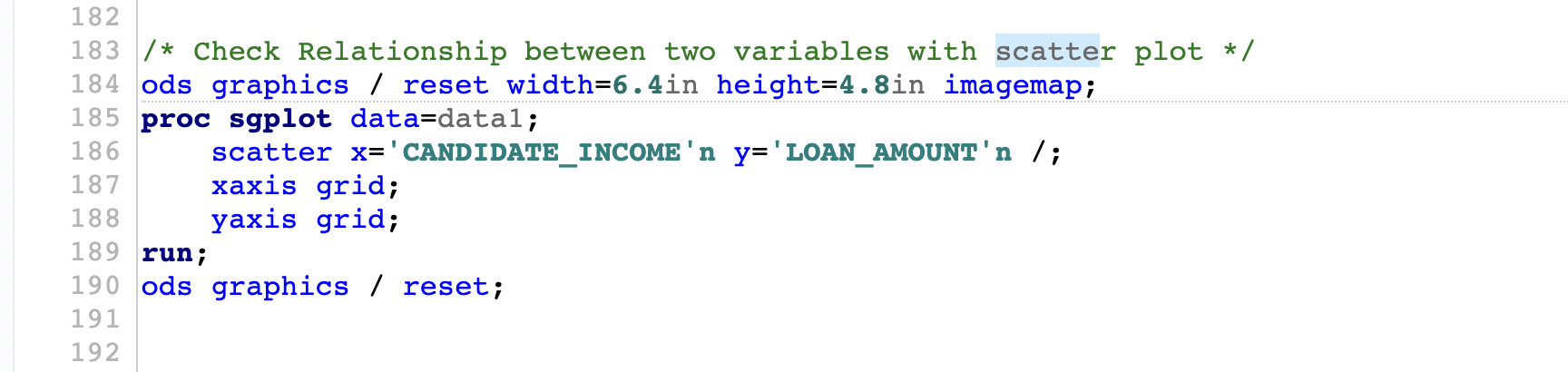
****

Figure : Code for scatter plot using the relationship between two variables

The output derived from the above code is shown below.

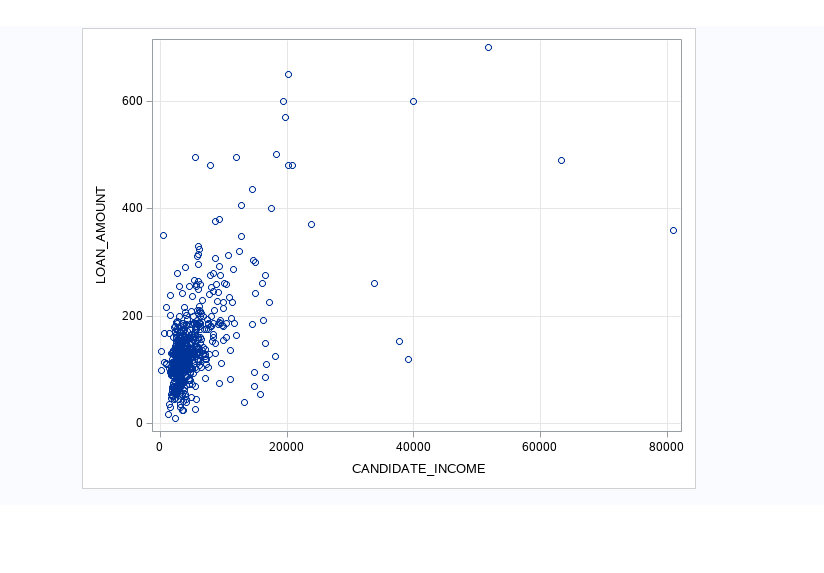


Figure : Scatter plot derived from the code

Similarly, the scatterplot for multiple variables can be drawn from the code below using variable Gender.

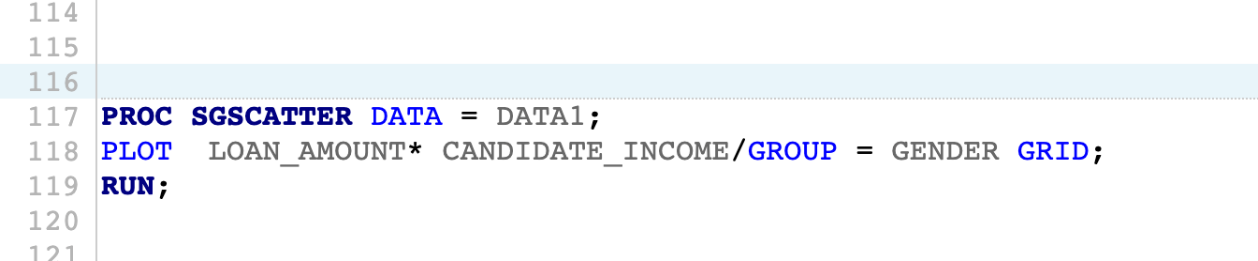


Figure : Code for scatter plot for a loan amount

The figure below shows the scatter plot using two variables – loan amount and candidate income in terms of variable Gender – male and female.

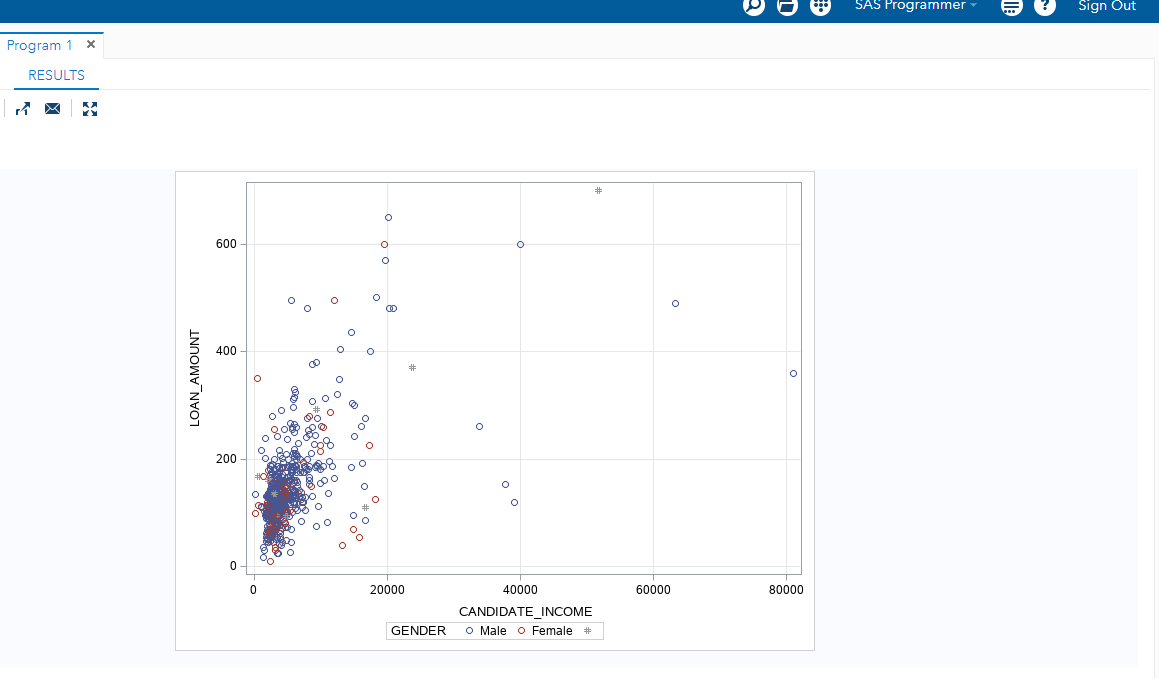


Figure : Scatter plot derived from the code

**6.3 Simple Linear Regression:**

Simple linear regression is a model that uses a straight line and predicts the relationship between the one dependent and one independent variable. Usually, in the SAS studio, the dependent and independent variable is selected from the dataset and then the simple linear regression is run.

Multiple linear regression is a technique to predict the outcome using two or more independent variables.

(Klein et al., 2008)

The code below shows the Simple Linear Regression as well as Multiple Linear Regression in the given dataset.

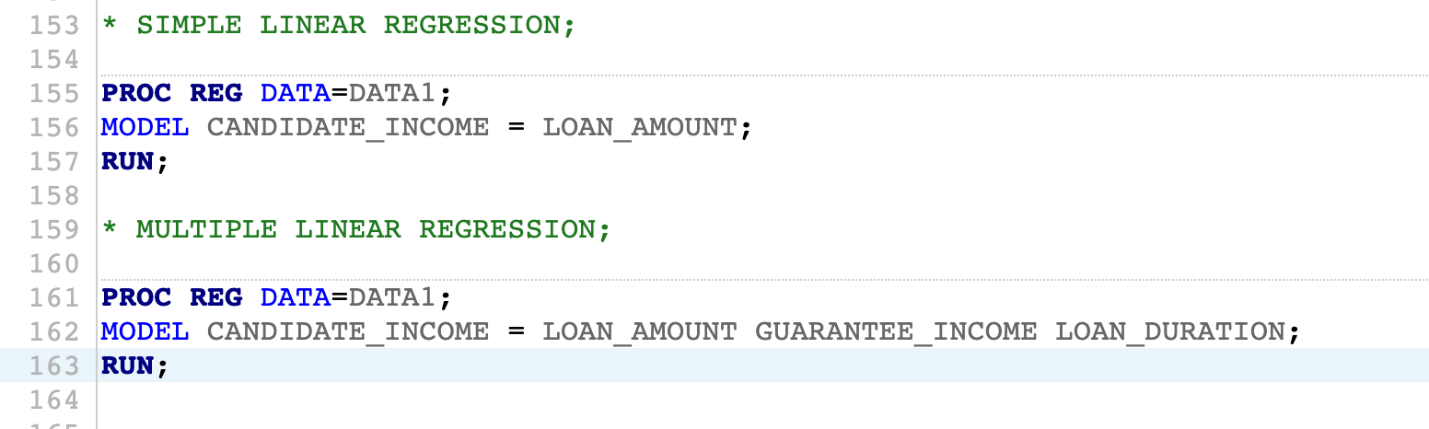
****

Figure : Code for simple linear regression

The output derived from the simple linear regression is shown below.

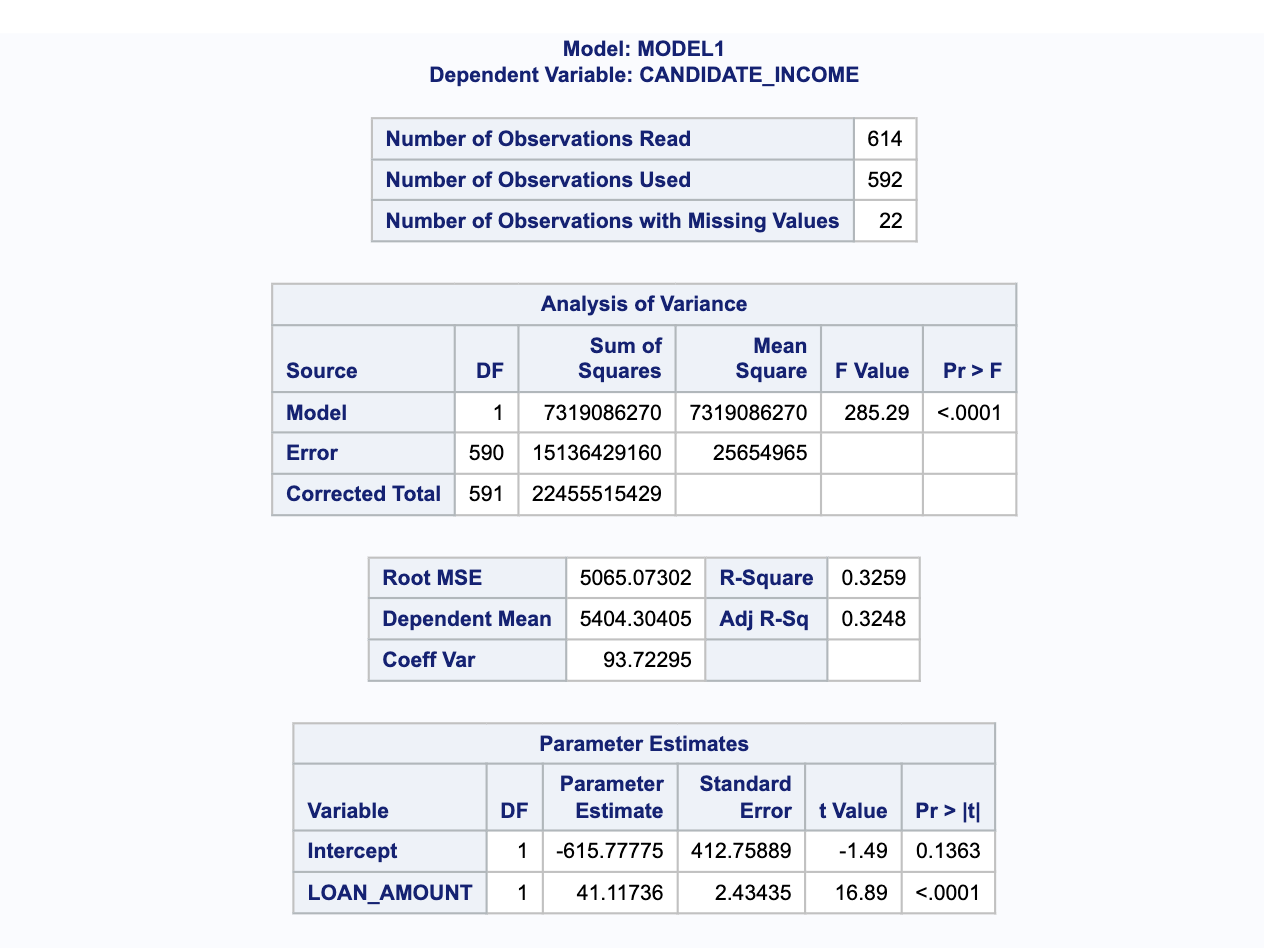
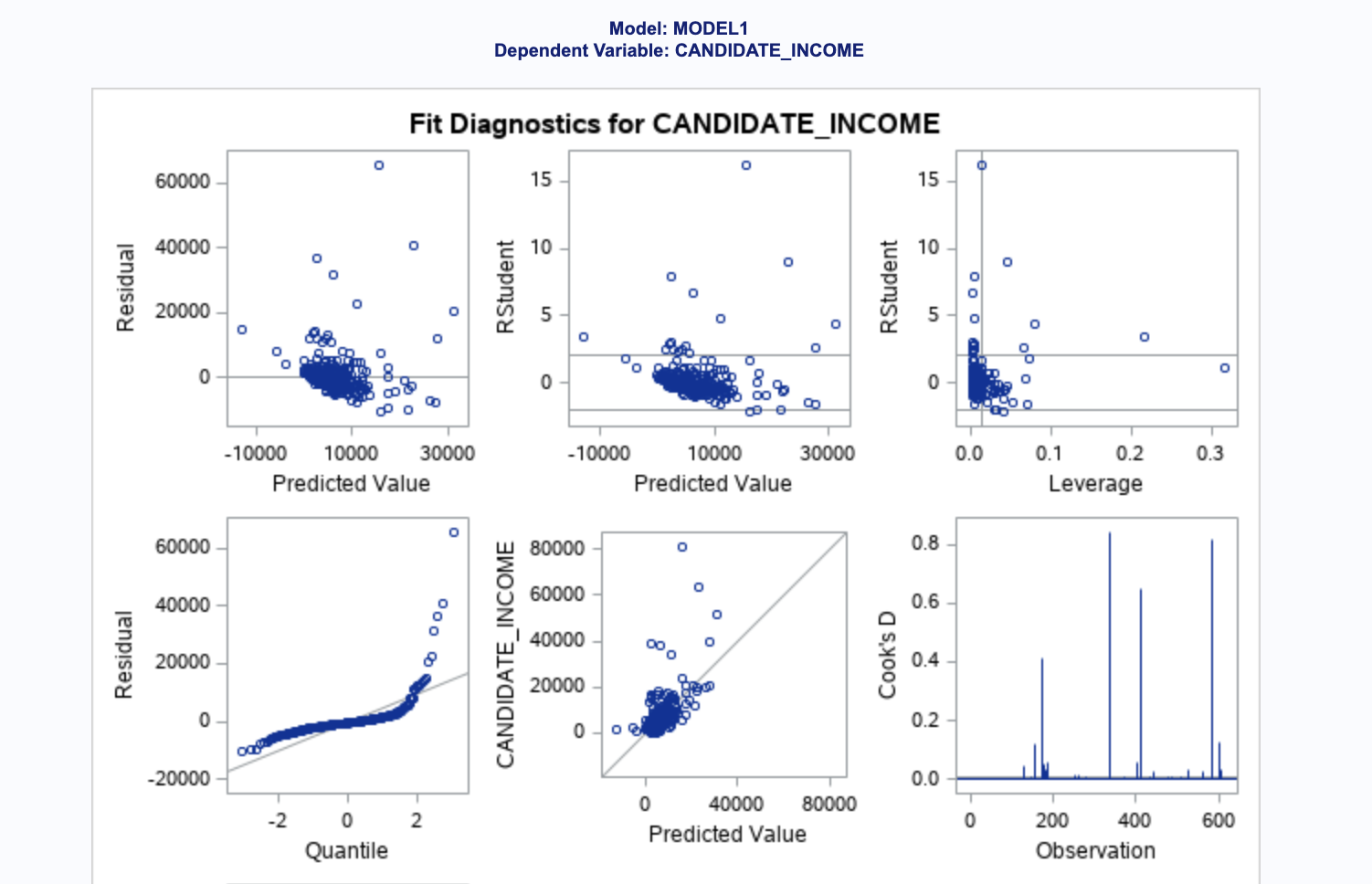
****

Figure : Observations, analysis of variance, and parameters estimations

****

****

Figure : Output for Simple Linear Regression

From the above figure, it is seen that the value of R-square is 0.3 which indicates that the model is not a good fit and does not have strong relationships among variables.

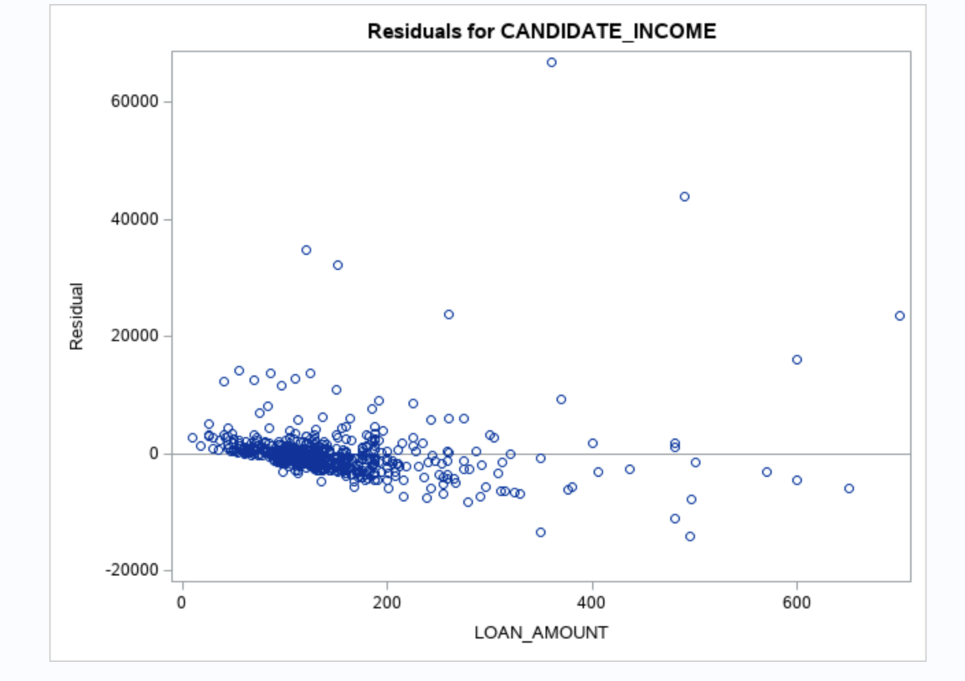
****

Figure : Residuals for CANDIDATE\_INCOME

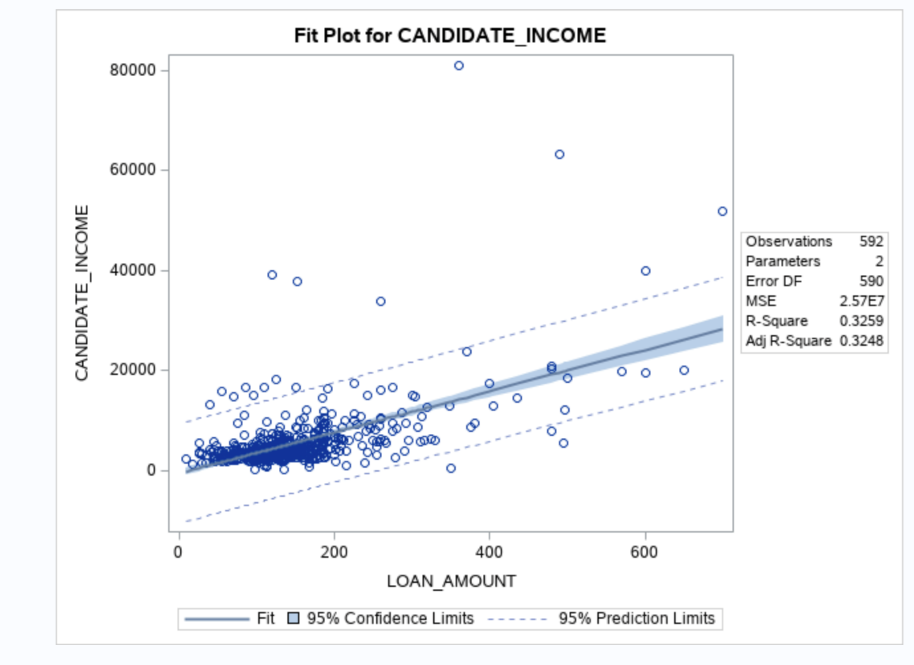
****

Figure : Fit Diagnostics for CANDIDATE\_INCOME

Multiple Linear Regression:

The output derived from the code using Multiple linear regression is shown in the snapshot below.

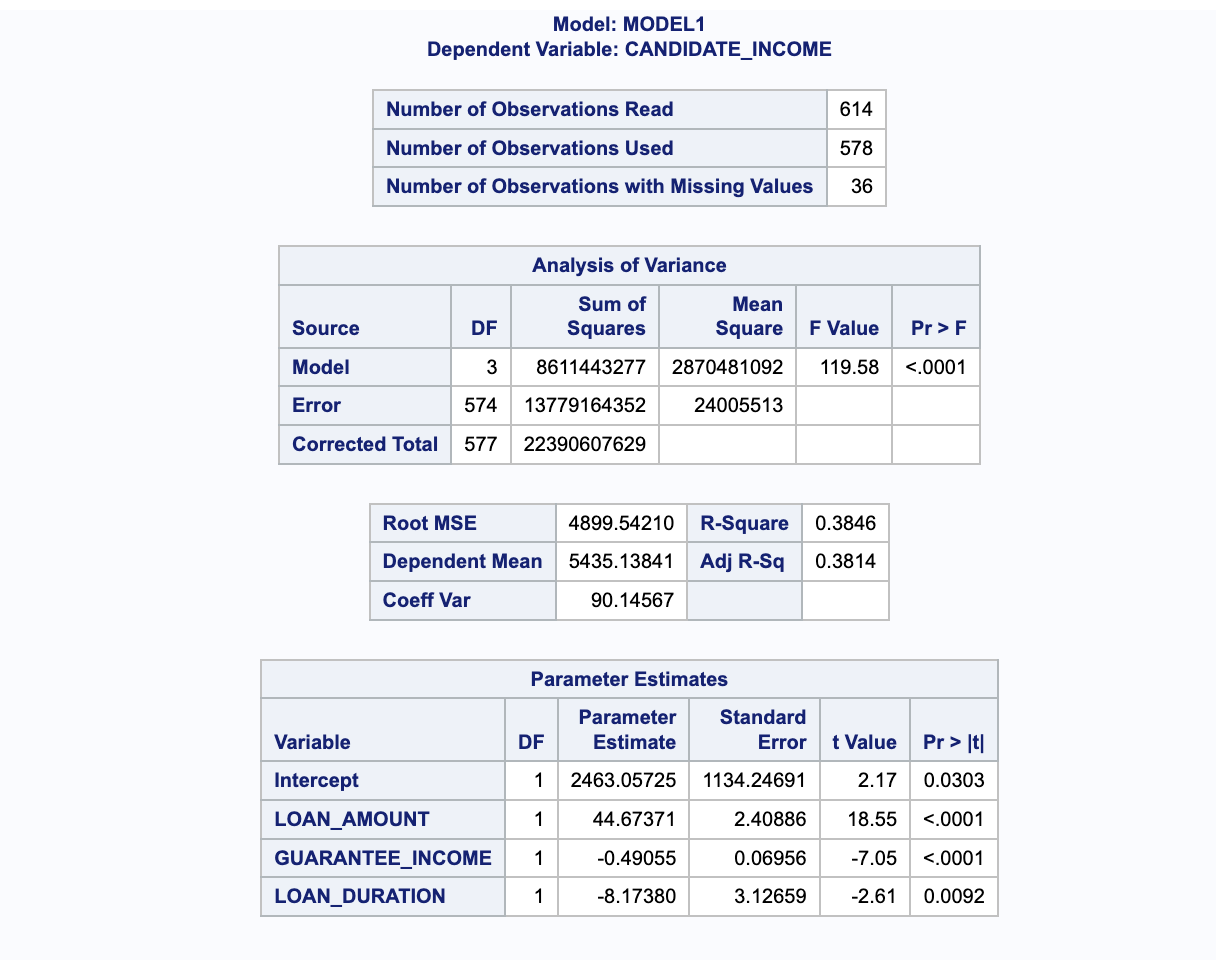
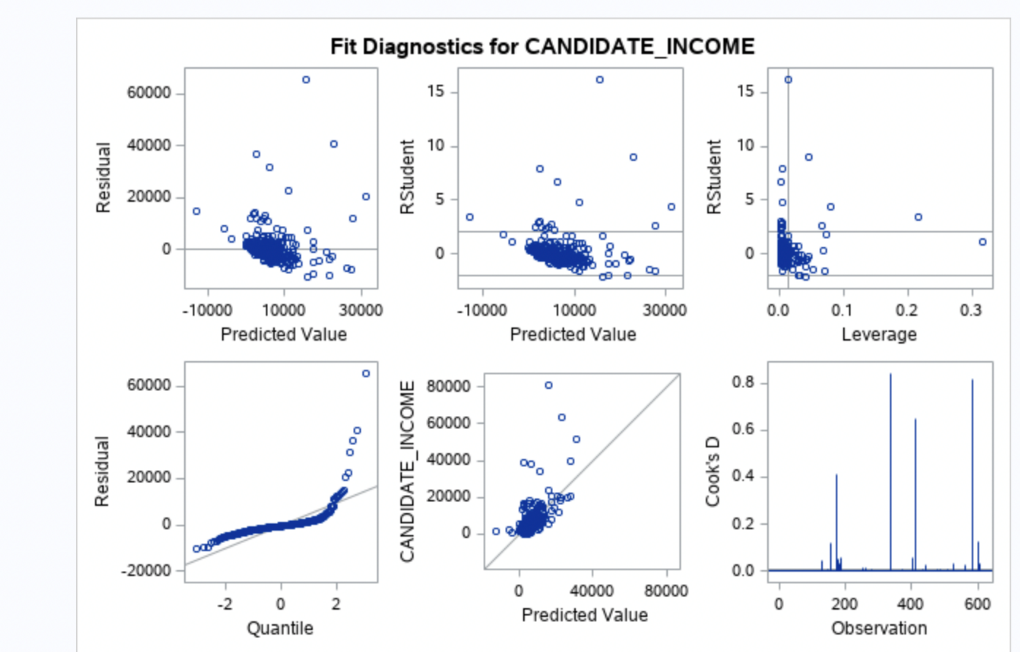
****

Figure : The output derived from the Multiple linear regression



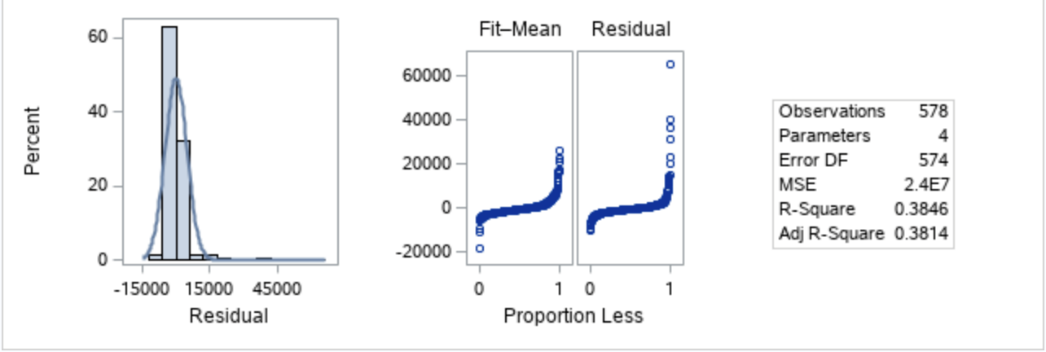


Figure : Graph Output using Multiple Linear Regression

From the above figure, it is seen that the value of R-square is 0.38 which indicates that the model is not a good fit and does not have strong relationships among variables.

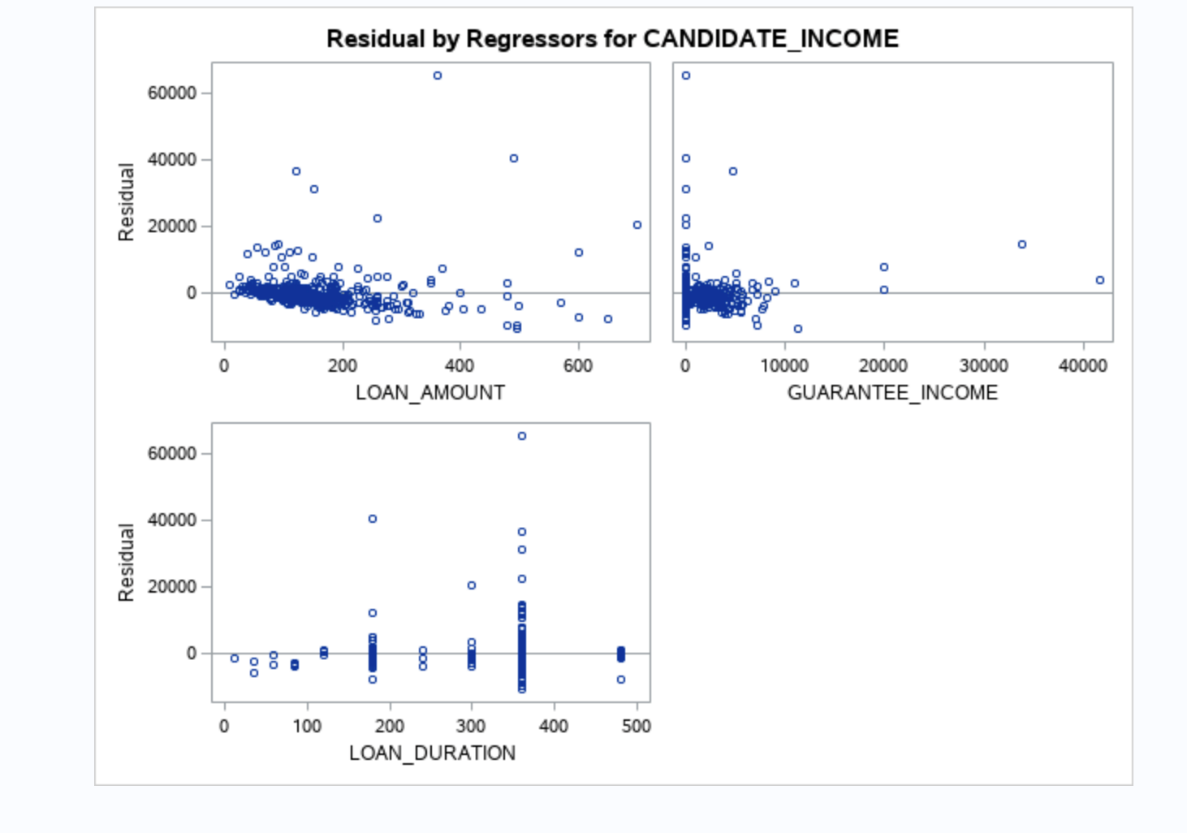
****

Figure : Residual by Regressor for Candidate Income using Multiple Linear Regression

# Apache Hadoop

We will now load the cleaned and transformed dataset into Hive and will use the AWS services as a choice of our Apache Hadoop distribution.

Hadoop is used for the management of big data, which have been distributed across many servers. Hadoop uses HDFS aka Hadoop distributed file system to store data and map-reduce method for the data processing. Hive, on the other hand, uses the Hadoop framework to provide an interface as similar to SQL but instead used HQL aka Hive query language. Hive and Hadoop help in the warehousing of the data and information, and making their best uses. (Thusoo et al., 2009)

I will use AWS EMR for Hadoop in this project. Amazon web services is a cloud-based application whereas EMR stands for Amazon Elastic MapReduce, which is an important tool of AWS used in big data analysis and processing. To perform Hadoop on AWS using EMR, the following steps and architecture have to be performed.

I will be using the following architecture.

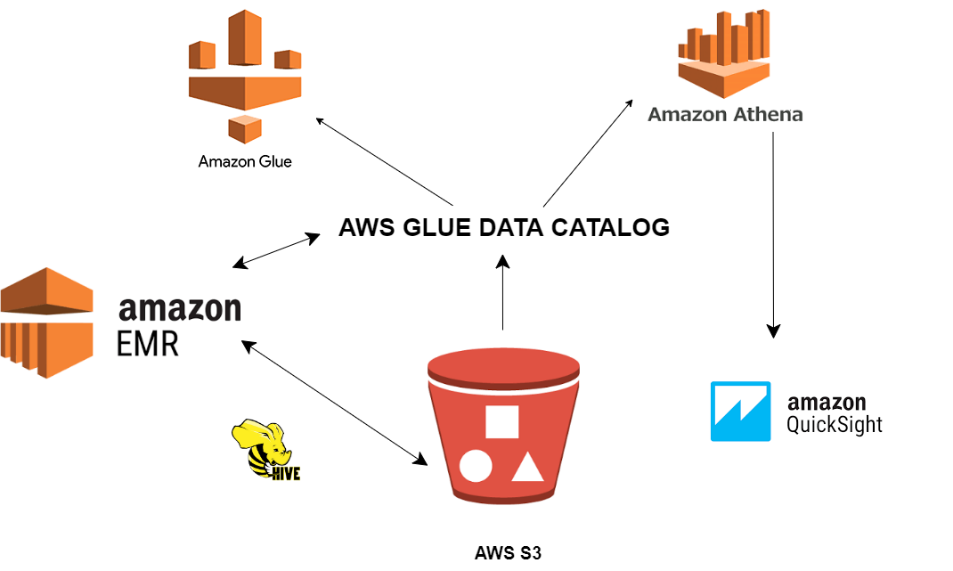


Figure : Architecture of Hadoop using the AWS services

First of all, the S3 bucket will be created, then our cleaned data will be uploaded to the S3 bucket using the AWS CLI. Now, a 1 m5\* large EMR instance will be created. The next step is to create an aggregate table using the HiveQL in EMR, and the result is now stored back to S3. Athena is now used to read the result from the aggregate table. Finally, AWS Quick Sight is used to generate the chart from Aggregated table.

The services used are the S3 bucket, EMR, Glue, Athena and Quick Sight. S3 is used for storage purposes, EMR to read and process the data, Glue to store hive metadata, Athena to access S3 and Quick sight for analytics.

The screenshots for the processes are as follows:

File from local (csv file) is uploaded into S3 bucket using CLI command. CLI stands for Command Line Interface, which needs to be first downloaded.

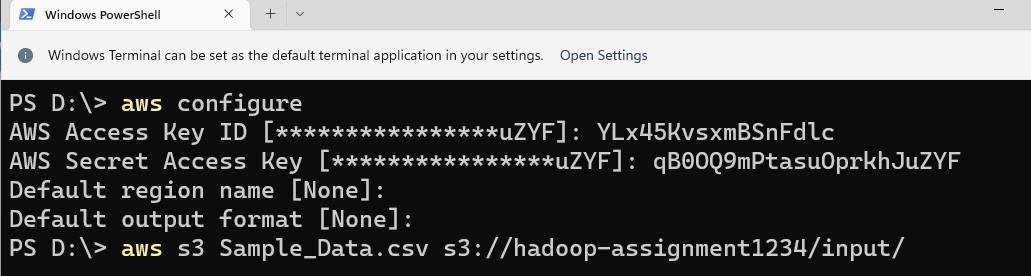


Figure : File from local (csv file) is uploaded into S3 bucket using CLI command.

Now, cluster is created using advanced options. Before creating cluster, we need to create EC2 Key Pair as shown in the figure below.

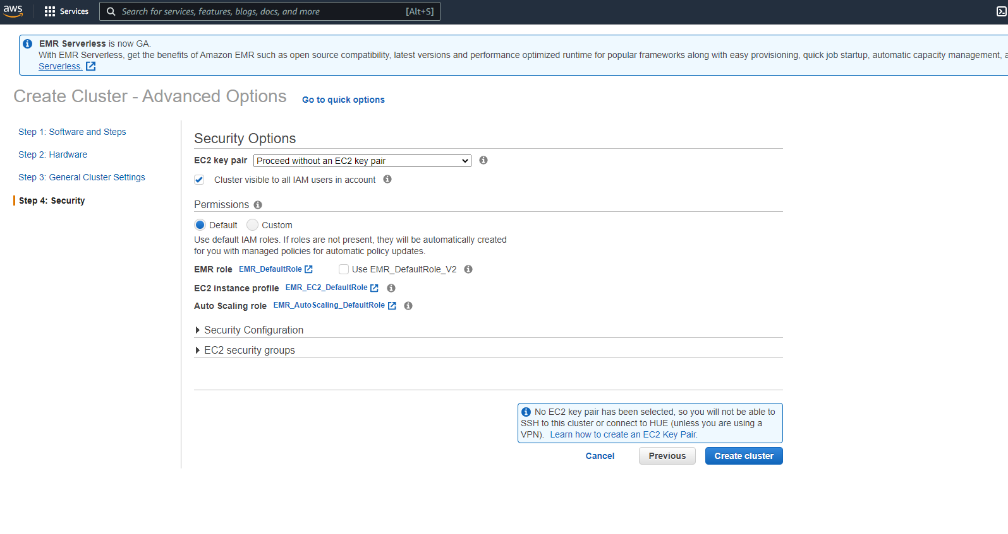
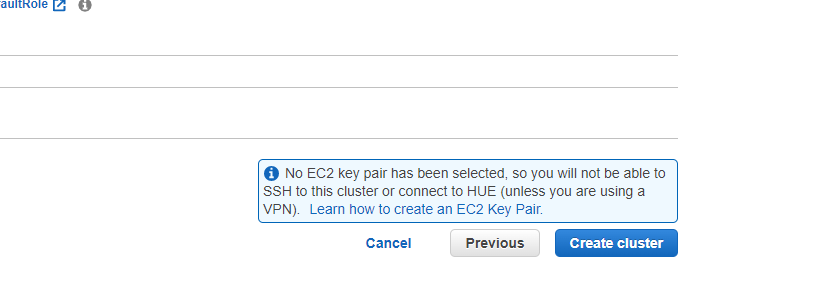
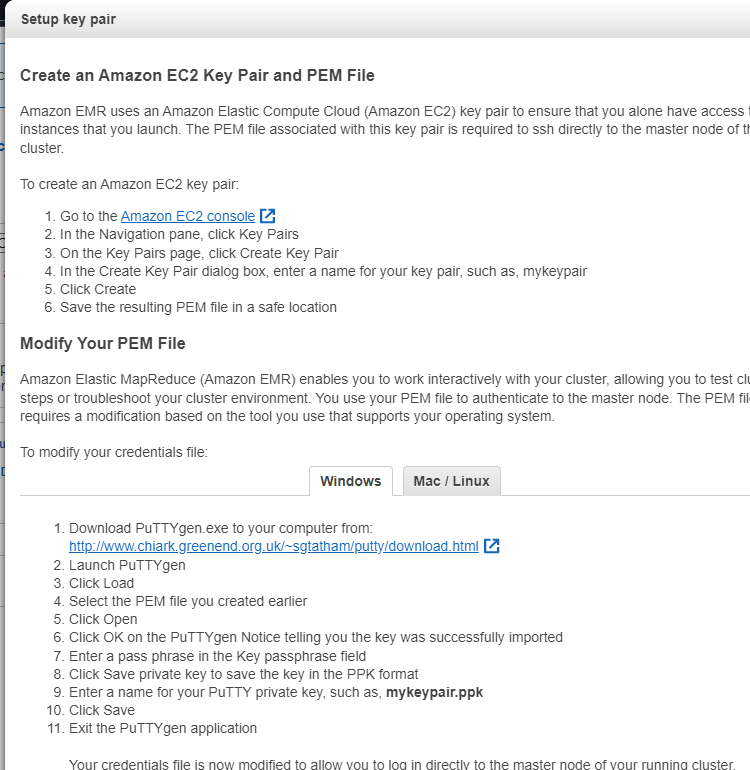


Figure : Creation of EC2 Key Pair





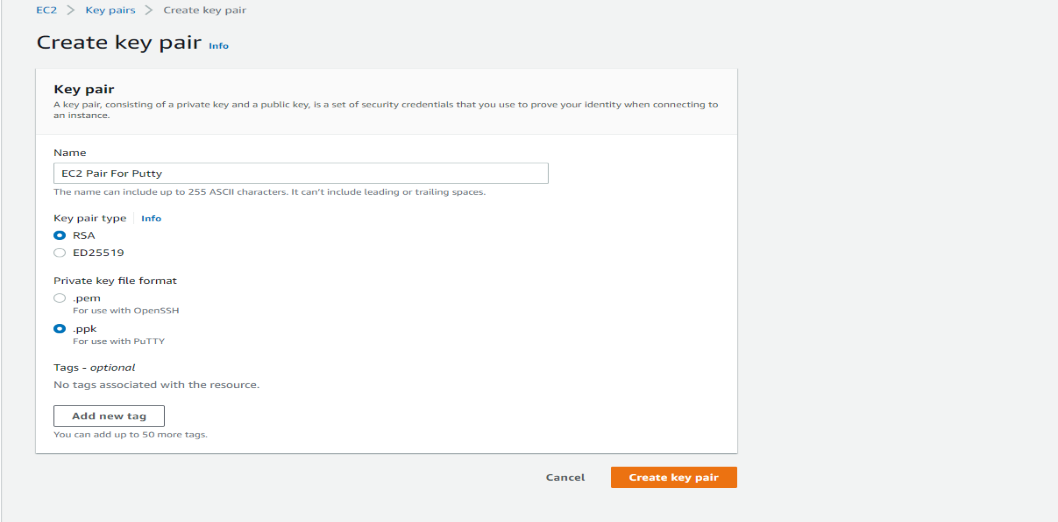


Figure : Creating EC2 Key Pair

Key pair is now created. Here .ppk is selected as we are doing the process from putty. If we need to carry out the process from SSH in mac, the we use .pem.

Now, before the cluster is created, we need to choose the type of software configuration.

Here, we have used Tez 0.92 for Hive and Hue 4.10.0 for Hue. Likewise, other configurations can be added as per the requirement as shown below.

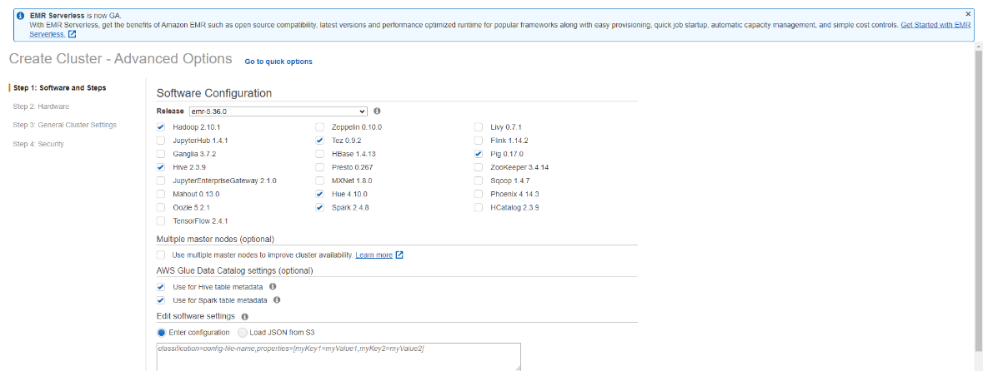


Figure : Configurations added

Now the two groups are created Master and Slave group for inbound security purpose.. Port 8888 (Type TCP) and Port 22 ( Type SSH) need to be enabled in the Master group. We do not need to add extra ports in the server node.

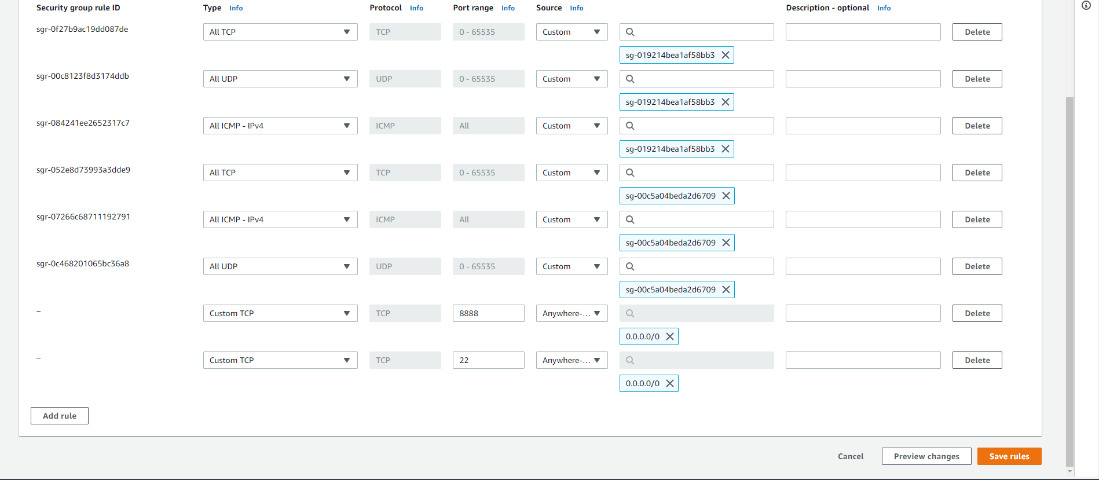


Figure : Master and slave group extra ports added

Now the connection is starting. We can see that master public DNS is not present at first.

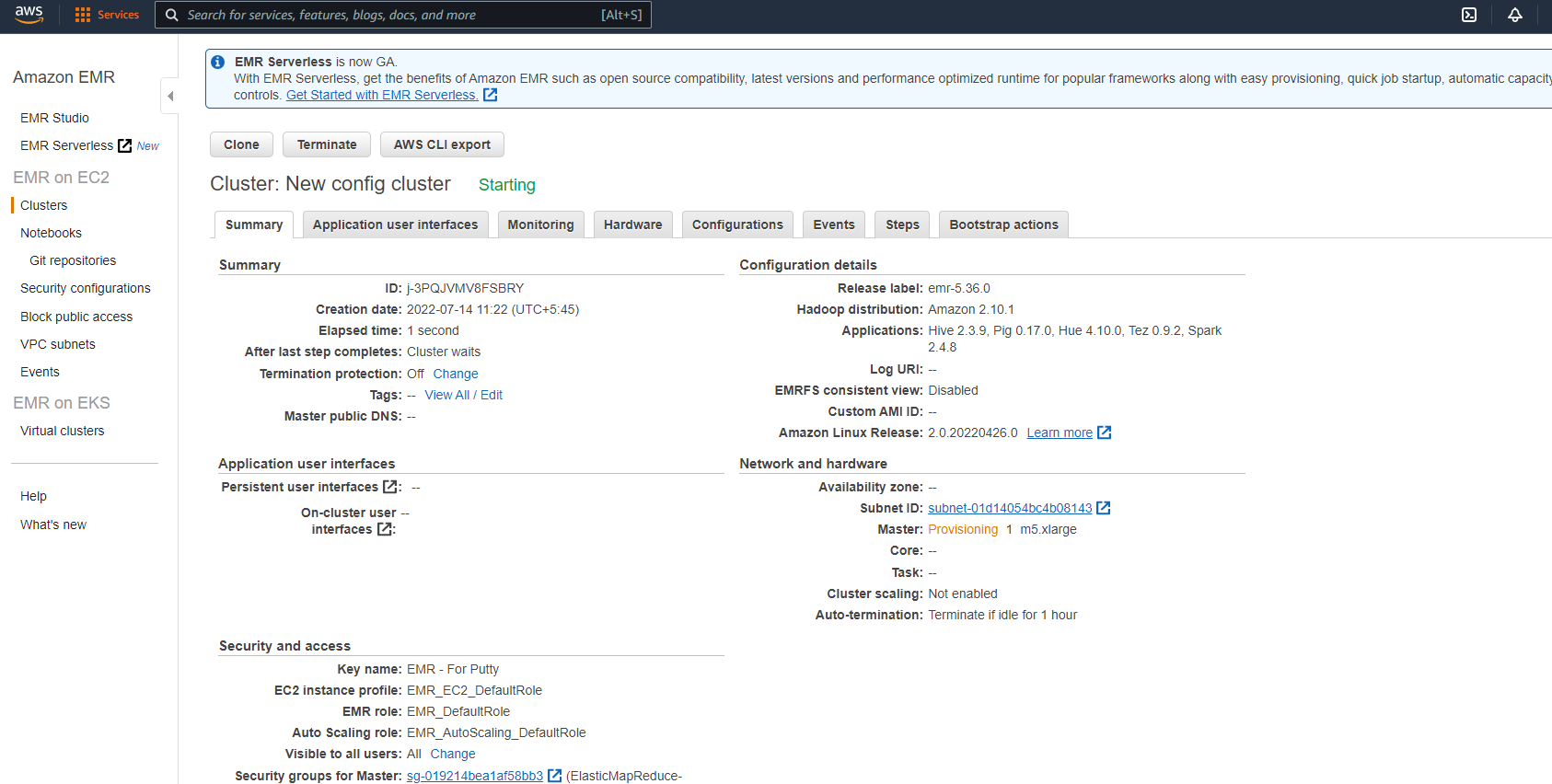


Figure : Checking if master DNS is not present

After sometime, the master public DNS is generated as shown in the figure below.

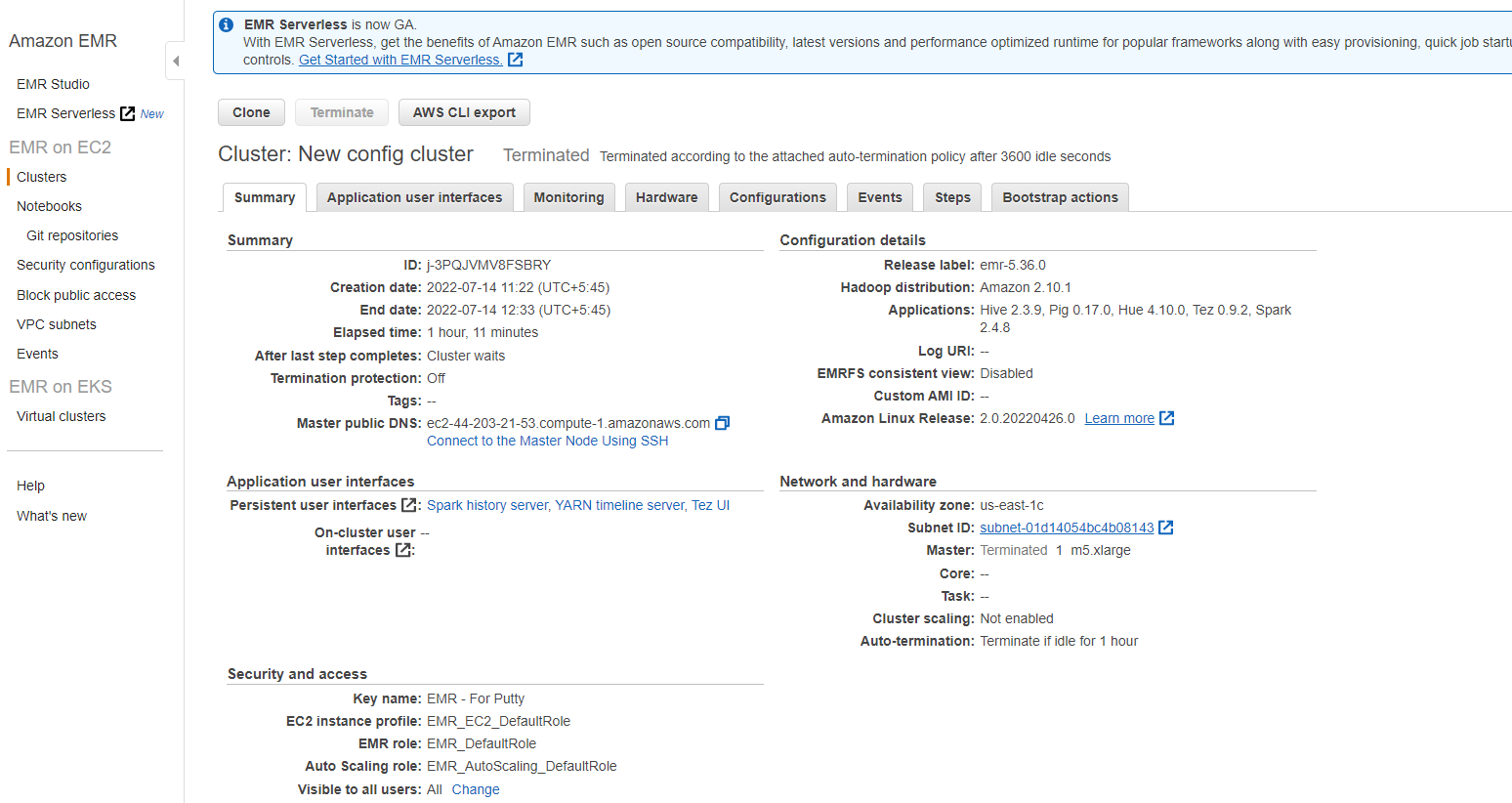


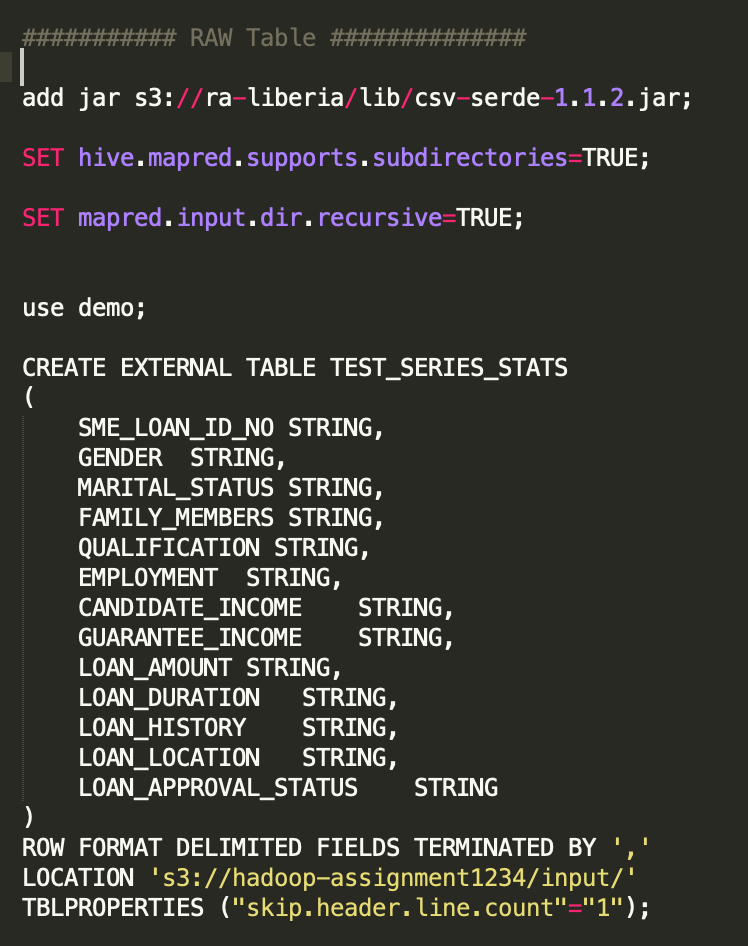
Figure : Generation of master DNS using EMR tool

We copy the master public DNS and open it using it as Hostname in Putty to make connection to the file. After we open putty, separate external table (for which sql file is already built using all the fields from the raw data file) is opened using Putty.



Figure : Hostname in Putty to make connection using EMR tool

Now, the Hue connection is enabled slowly and we using the following code.



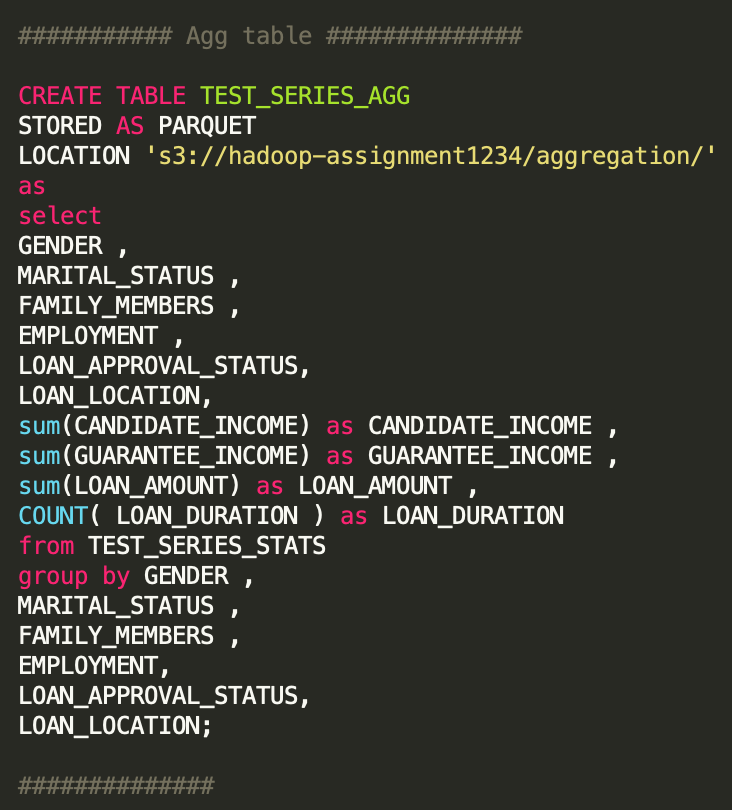


Figure : Writing code after connection

We have now finished up writing the query. We will now use AWS Quick Sight using Athena as a backend table. We must make sure as we are in the Quick Sight we are in the same region where the database was created. Datasets from the Hue are now uploaded and following visualizations are created.

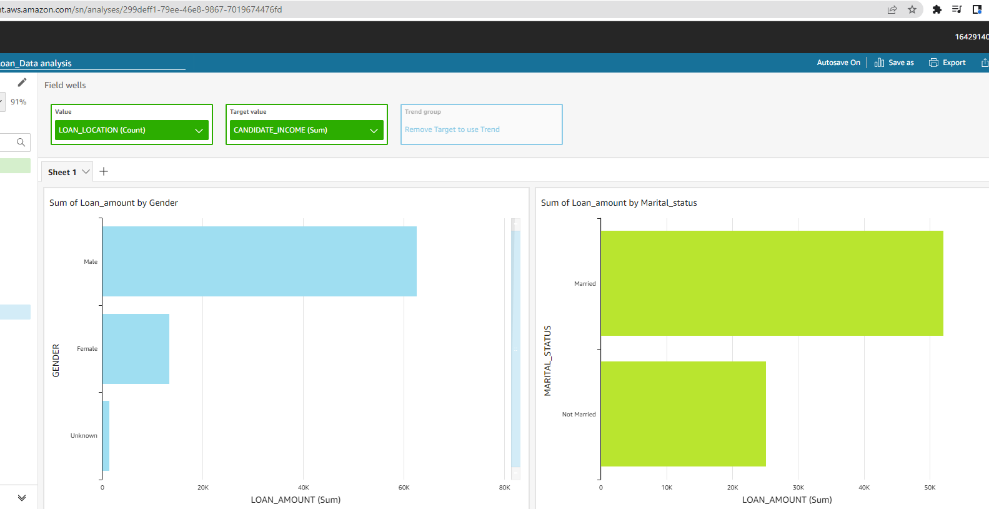


Figure : Sum of Loan amount by grnder and mariatal status in Quick Sight

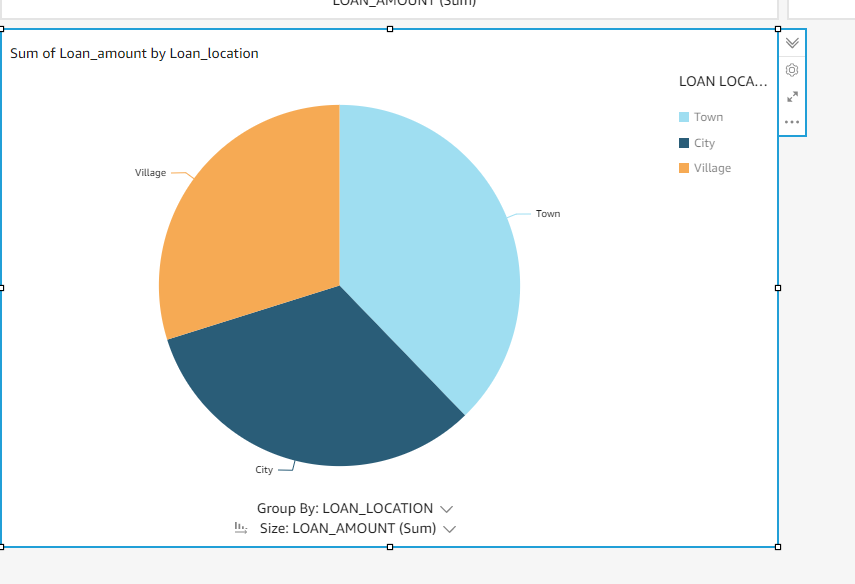


Figure : Sum of loan amount by loan\_Location using AWS QuickSight

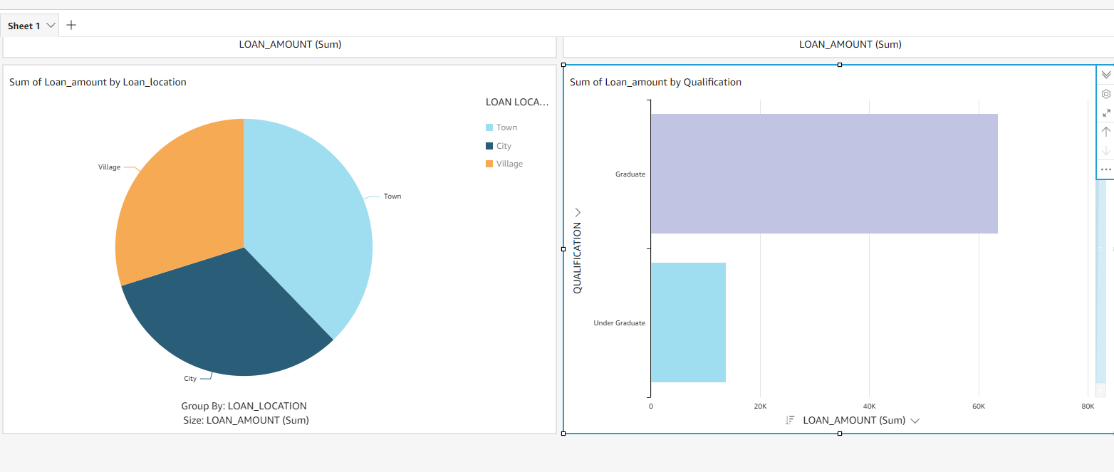


Figure : Sum of loan\_amount by Location and Qualification

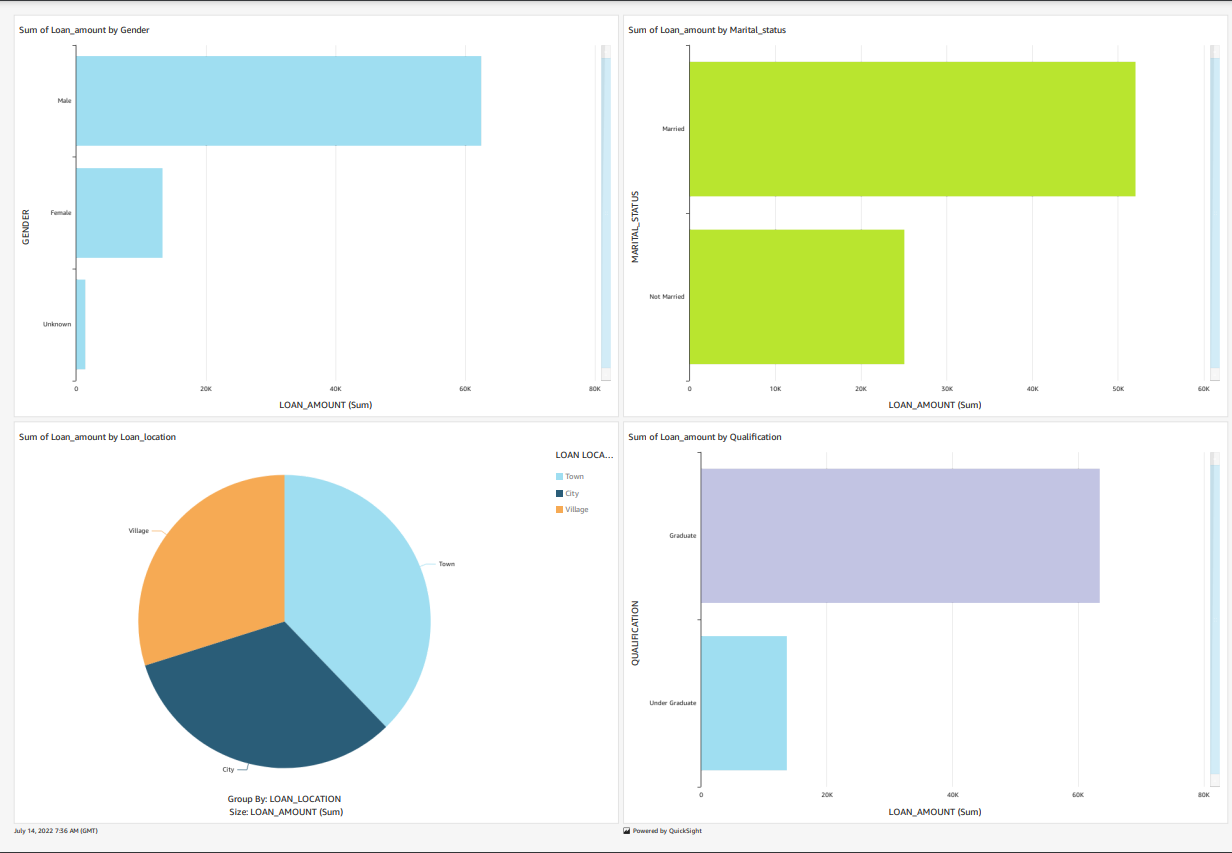


Figure : Complete visualization using AWS Quick Sight

# Hypothesis

Hypothesis uses statistics and finds out the probability of whether a hypothesis is true or not. To formulate the hypothesis, usually observations are done, and to find the truth of the null hypothesis, identification is used. Most of the time, the value is calculated, the if the p-value is small, we can assume more chances of the null hypothesis being false. The hypothesis may also include comparing p-value to alpha value for better predictions.

(Bzdok & Lindenberg, 2018)

The following types of hypotheses can be performed with the help of SAS studio.

|  |  |  |
| --- | --- | --- |
| SAS PROC | Test | Application |
| PROC TEST | T-Test | Checks if the mean value is different from the hypothesis value. The mean value includes both the paired and independent groups. |
| PROC FREQ | Chi-Square | To find the probability of frequency occurring by chance. |
| PROC ANOVA | ANOVA | They compare and evaluate if the means of an independent and categorical variable are different from the interval-dependent variable. |
| PROC REG | Linear Regression | It tells how one variable predicts another. |

Table : Types of hypotheses in SAS

Source: (Zhang & Hou, 2019)

We will now go through our hypothesis, which will include five of them and will mostly use the Chi square test and Anova. SAS studio will be used again for all hypothesis.

**Hypothesis 1:**

Null Hypothesis: Male population has a low chance of loan approval than the female population.

Alternate Hypothesis: Male population has a high chance of loan approval than the female population.



Figure : Code for Chi-Square Test

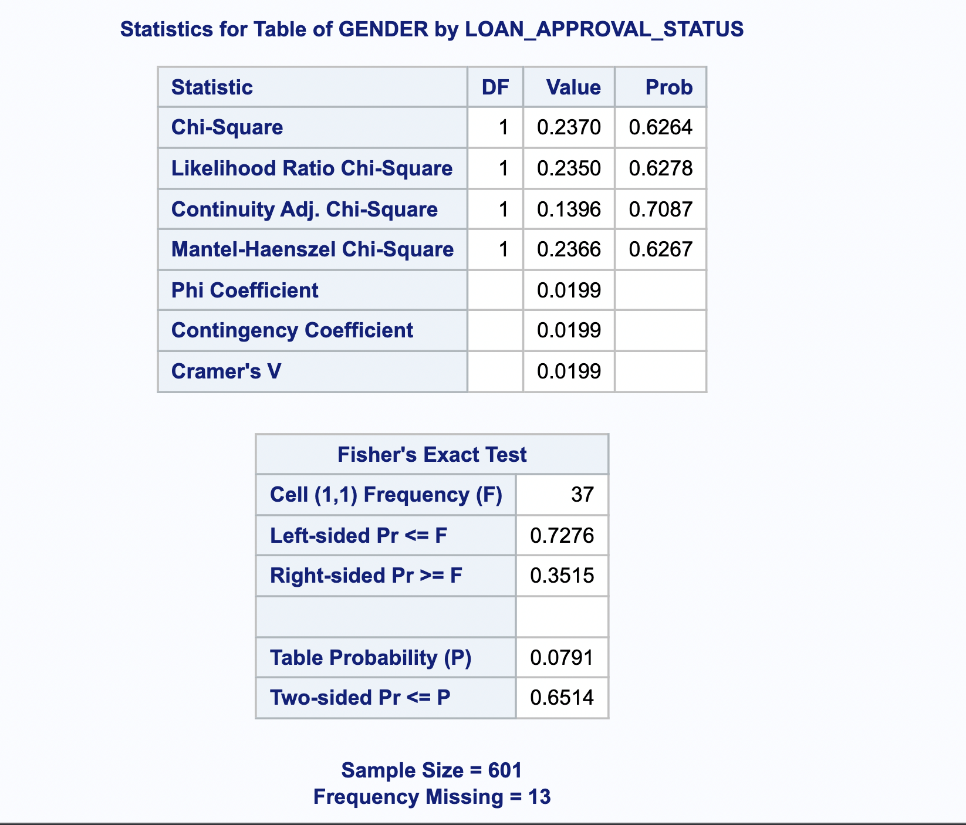


Figure : Statistics for a table of GENDER and use of Fisher's Exact Test

Chi-Square Test = 0.6264

Alpha = 0.6514

P<Alpha, therefore, the Null Hypothesis is rejected.

**Hypothesis 2:**

Null Hypothesis: The married population has a high chance of loan approval.

Alternate Hypothesis: The married population has a low chance of loan approval.

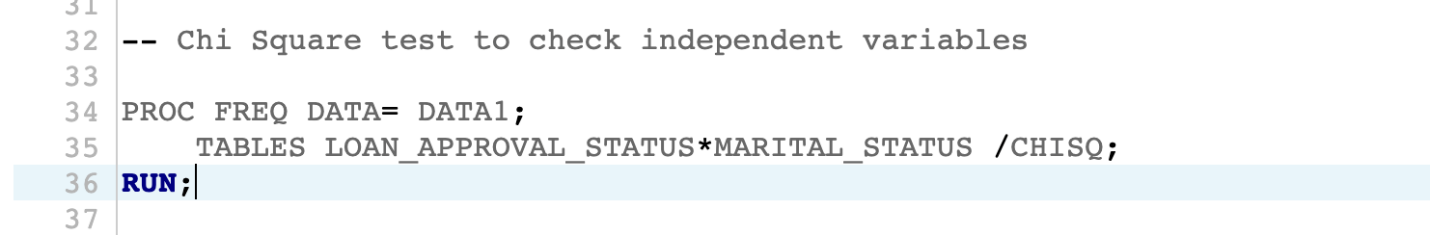


Figure : Chi-Square test for LOAN\_APPROVAL\_STATUS for MARIATAL\_STATUS

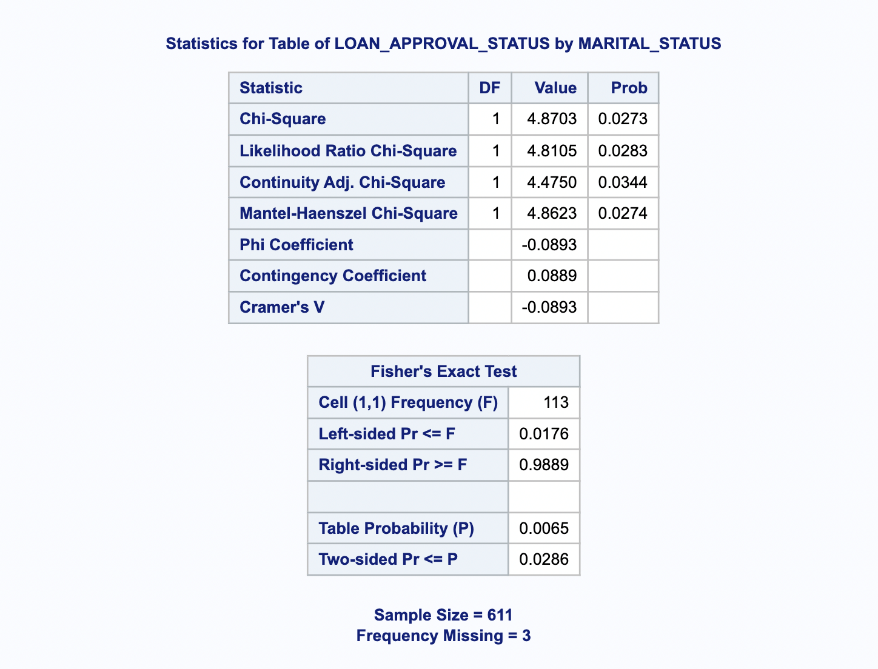


Figure : Statistics for LOAN\_APPROVAL\_STATUS by MARIATAL\_STATUS

Chi-Square Test = 0.0273

Alpha = -0.0893

P>Alpha therefore, NULL Hypothesis is accepted.

Similarly, the other hypothesis set in the data is shown below.

**Hypothesis 3:**

Null Hypothesis: Population with High Incomes have more chances of approval for loans.

Alternate Hypothesis: Population with high incomes have low chances of approval for the loan.

**Hypothesis 4:**

Null Hypothesis: A population with less loan amount has a high chance of loan approval.

Alternate Hypothesis: A population with less loan amount has a low chance of loan approval.

**Hypothesis 5:**

Null Hypothesis: A population with a high guarantee income has a high chance of loan approval.

Alternate Hypothesis: A population with a high guarantee income has a low chance of loan approval. In this hypothesis, we shall be using the anova.

One Way Anova:

Null Hypothesis: Married and unmarried population has equal loan\_amount.

Alternate Hypothesis: Married Population and unmarried population do not have equal loan\_amount.

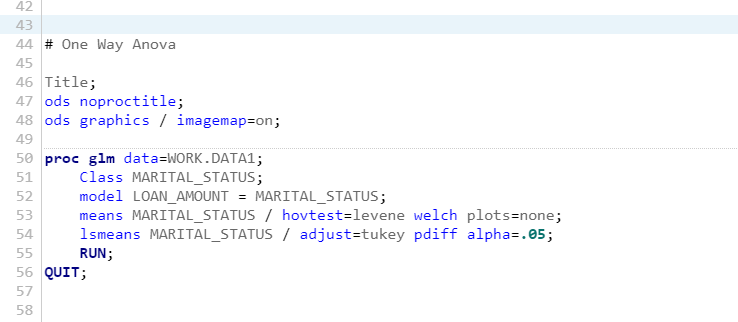
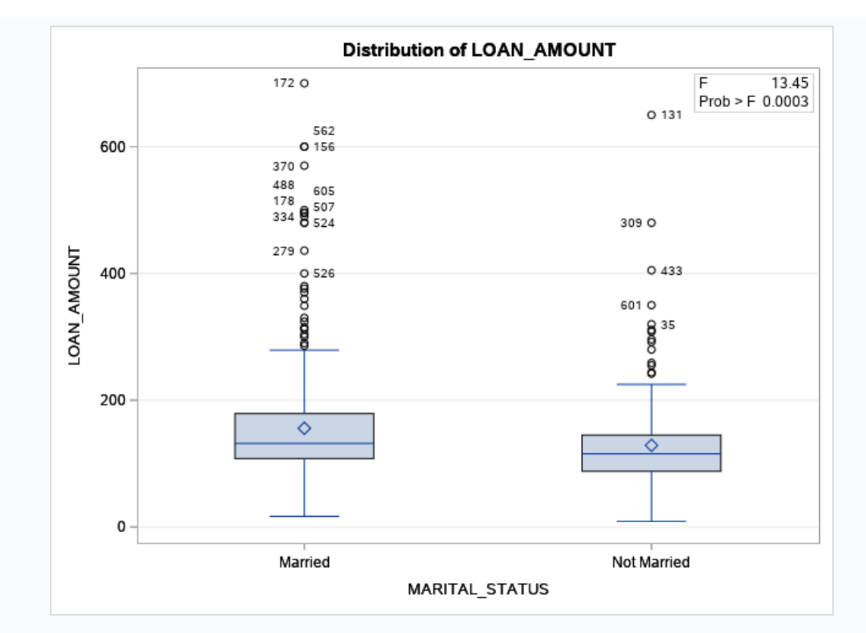
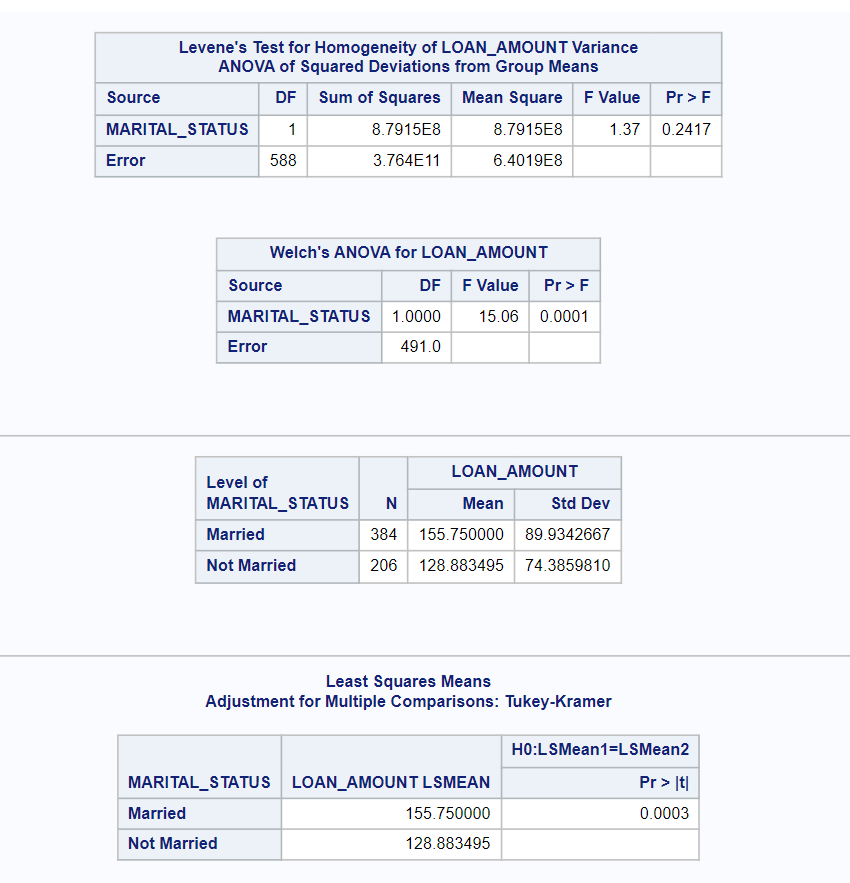


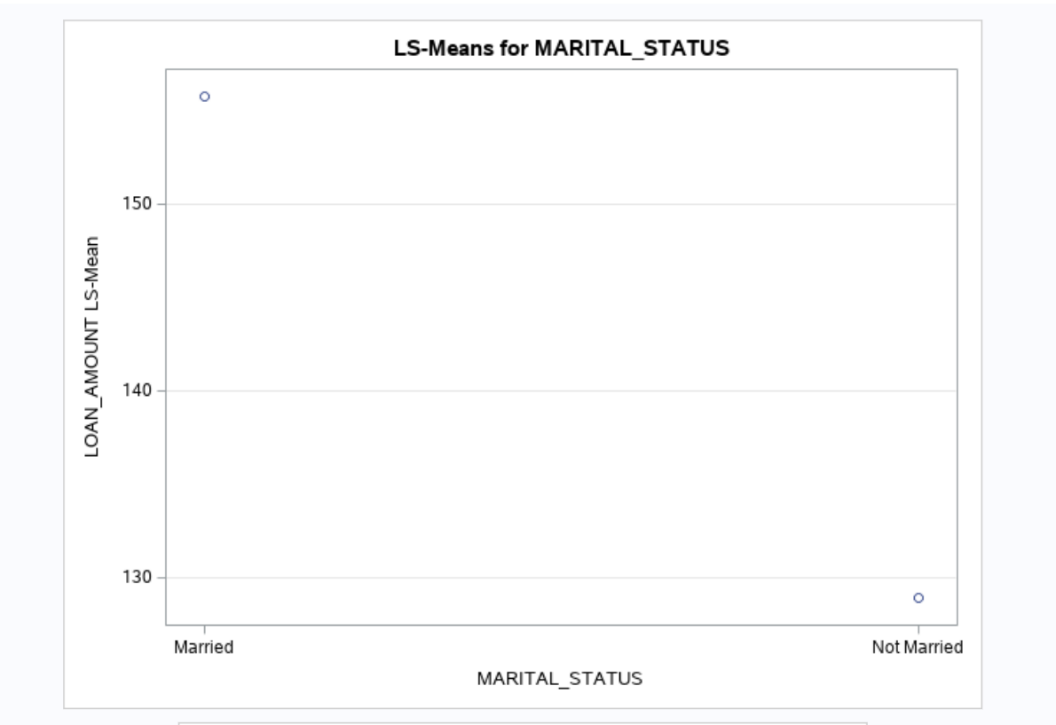
Figure : Code for hypothesis 5 using anova test

The result derived from the above code using One Way ANOVA is shown below:









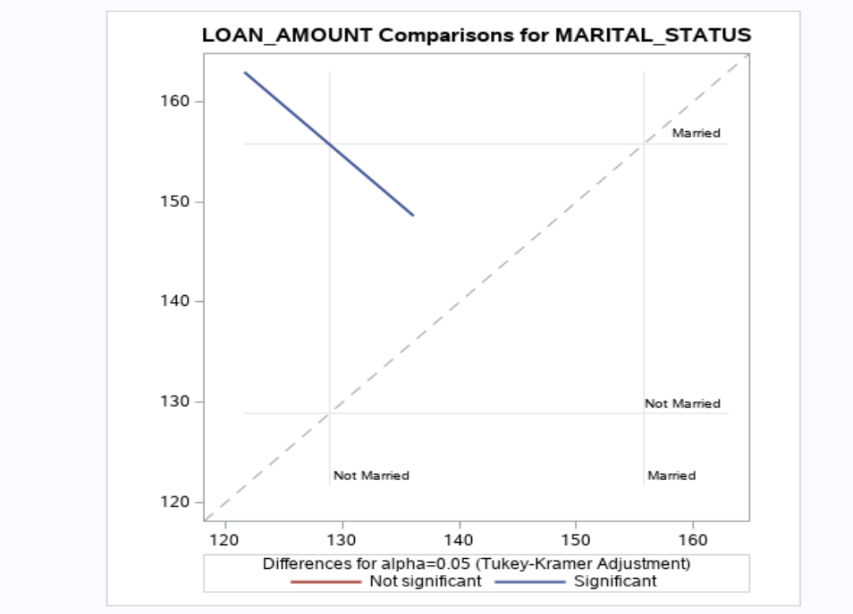


Figure : Output of hypothesis 5 using anova test

Here, in this case p value is 0.0003. When the value of P is less than 0.5, then we reject NULL Hypothesis.

This means that the NULL hypothesis is rejected. Therefore, the married and unmarried groups do not have same amount of loan.

# Conclusion

To conclude, machine learning and future engineering play an inevitable role in the classification, cleaning, and analysis of data for future predictions and decision-making processes. With the help of SAS studio, we were able to understand each variable inside the dataset in the initial data exploration stage and that these variables were indeed vital for decision-making. However, we also realized that not all of the variables were important, so they needed to be cleaned, handled, and fix the noisy and inconsistent data in the pre-processing processes. Through the help of single and multiple variable analysis, and the aggregation of variables, we were able to check the combination and calculate aggregation and transformations.

The exploratory data analysis was then possible and with the help of EDA graphs such as scatter charts, histogram, descriptive analysis, univariate analysis, and correlation, the outliers within the data were identified. For Apache Hadoop, Amazon web services tools including the EMR, AWS Glue, Athena, S3 bucket and Amazon QuickSight have been used.

Furthermore, the hypothesis, involving the Chi-Square Test and Anova tests has been proved to be effective in predicting that high loan approval was for those with high income, less debt, more likely a female population, and married population. Hence, these future engineering techniques in machine learning, and with the use of SASs has been a great prediction tool for the “XYZ Loan” company.

# REFERENCES

Bzdok, D., & Lindenberg, A. M. (2018). *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Garcia, S., Gallego, S. R., Luengo, J., Benitez, J. M., & Herrera, F. (2016). *Big data preprocessing: Methods and prospects*. https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0

Klein, J., Gerster, M., Andersen, P. K., & Tarima, S. (2008). *SAS and R functions to compute pseudo-values for censored data regression*. *89*(3), 289–300. https://doi.org/10.1016/j.cmpb.2007.11.017

Komorowski, M., Marshall, D., Salciccioli, J., & Crutain, Y. (2016). *Exploratory Data Analysis*. https://doi.org/10.1007/978-3-319-43742-2\_15

Land, D., N, P., & Wm, W. (1990). *Statistics Notebook Entry I.J, Types of Data Nominal, Ordinal, Interval, and Ratio Scales*. *67*(2), 155.

O Rourke, N., Hatcher, L., & Stepanski, E. (2005). *A Step-by-Step Approach to Using SAS for Univariate & Multivariate Statistics*.

Petruccelli, J. D., Nandram, B., & Chen, M. (1999). *Doing it with SAS: A Supplement to Applied Statistics for Engineers and Scientists*. http://www.math.wpi.edu/Course\_Materials/MA3627C04/diws12-14.pdf

Thusoo, A., Sharma, J. S., Jain, N., Chakka, P., Anthony, S., Liu, H., & Wyckoff, P. (2009). *Hive: A warehousing solution over a map-reduce framework*. *2*(2). https://doi.org/10.14778/1687553.1687609

Zhang, X., & Hou, W. (2019). *Introduction to SAS*. https://osa.stonybrookmedicine.edu/sites/default/files/SAS-workshop-2020%20Jan.pdf