INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT051-3-M-DM

DATA MANAGEMENT

APDMF2112DSBA(DE)(PR)

Assignment Part 2

Data Pre-Processing & Data Exploration

Lee Kean Lim

TP065778

Assoc. Prof. Dr. Raja Rejeswari

ABSTRACT

Loan approval is a labor and time intensive task which require the processing of voluminous information to sift through the applicants. Typically faced problems include inaccurate loan approval evaluation, leading to credit risks. Therefore, the development of machine learning models to assist the evaluation process achieving better evaluation rate and minimizes credit risks. However, the loan data typically contains missing values, outliers, and inconsistent data. Hence, require pre-processing prior to any use in the predictive models. This study investigates the methods in data pre-processing and performs exploratory data analysis (EDA) in a loan dataset from Kaggle. The data pre-processing would involve missing value imputation and removal of outliers while the EDA produces graphical analysis to identify trends and relationships in the data. In addition, EDA is utilized to validate six hypotheses formulated in which the graphics from EDA provides a quick and easy way of identifying the relationship among variables.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SMOTE.....Synthetic Minority Oversampling Technique

EDA .....Exploratory Data Analysis

SECTION 1

INTRODUCTION

1.1 INTRODUCTION

Banks provide money lending services as part of their business model to gain revenue. The banks would typically go through an evaluation process of the applicant to ensure the applicant is able to fulfill the repayment obligations. The evaluation process of approving a loan application involves the validation and verification of information submitted by applicants. This process can be labor and time intensive and subjected to bias of the staff handling the case (Fati, 2021). Therefore, a better method is required in collecting and handling the application data to achieve a more consistent and accurate approval results while minimizing the risk faced by the banks.

The use of machine learning is widely applied in the loan prediction tasks (Ambika & Biradar, 2021). Where the use of machine learning model would automate the process of validating the information and produces a real time approval result. Therefore, improving the work efficiency and reduces the need of manual labors. However, to achieve high prediction accuracy, the condition of the dataset plays an important role. Typically, loan dataset contains a lot of missing values and outliers which need to be treated for the dataset to be useful for the predictive model.

Data pre-processing is the manipulation and transformation of the data to enhance the value derived by the predictive models. It is a crucial process in any typical data science project life cycle. Typical steps involved in data pre-processing include data cleaning, data transformation, and data reduction. Data pre-processing ensures the data is properly representing the problem and formatted to suit the requirements of specific predictive models.

A loan dataset from Kaggle will be used in this study which the dataset was designed to predict the approval of housing loan of applicants. The dataset will be processed by imputing the missing values and removal of the outliers. After the data pre-processing, the data will be used in exploratory data analysis (EDA) to identify relationships among variables.

This study is structured as followed. Section 2 discusses the literature survey, section 3 discusses the initial data exploration, section 4 discusses the data pre-processing, section 5 discusses the EDA, and finally section 6 concludes the study.

1.2 AIM & OBJECTIVES

1.2.1 Aim

The aim of this study is to perform data pre-processing for a dataset to ensure the dataset usability for a machine learning model. In addition, exploration of the dataset is performed to identify relationship between variables to derive insights from the data.

1.2.2 Objectives

The objectives of the study are as followed:

1. To identify and apply the appropriate missing value imputation methods to the dataset.
2. To identify and apply outliers removal methods to the dataset.
3. To explore and identify relationship between variables using EDA.

1.3 HYPOTHESIS

The hypotheses of the study are as followed:

1. Applicants with higher income would apply for a higher loan amount.
2. Male applicants have a higher income than the female applicants.
3. Applicants with higher education level would have a higher income.
4. Applicants with good repayment history would have a higher loan approval rate.
5. Housing price in the urban area is higher than the semiurban and rural area.
6. Applicants who are married would have a higher loan approval rate.

SECTION 2

RELATED WORKS

2.1 LITERATURE SURVEY MATRIX

This section outlines the data pre-processing and EDA methods used in related works. The domain in the related works would cover the credit risk assessment for credit approval. Table 2.1 shows the work conducted by researchers in developing a prediction model to assist the loan approval process. However, only the data pre-processing and EDA process will be mentioned in the table in line with the objectives of this study. Following the table, will be a discussion on the findings based on the literature survey.

Table 2.1: List of pre-processing methods used in related works

Reference	Data Pre-Processing / Exploratory Data Analysis	Comments
Wang *et al.* (2022)	- Removal of irrelevant and duplicate variables as some of the variables are a complete duplicate - Variables with missing values greater than 55%, the whole variable will be removed from dataset - Complete removal of some observations that have many missing values, as it accounts for a very low proportion to the entire dataset - Outliers identified using box plot and removed from dataset as it accounts for very low proportion to the entire dataset - Dataset is seriously imbalanced and utilized Synthetic Minority Oversampling Technique (SMOTE) to balance the classes	- Did not mention the use of EDA to identify information from dataset - Did not mention missing value imputation method as observations with missing values are directly removed from dataset
Wen Zhang *et al.* (2022)	- Class imbalance treatment using SMOTE technique - Outliers are identified using boxplots and verified with domain expert	- Did not mention missing value treatment - Did not mention further treatment after verifying outliers with domain expert

		- Did not mention the use of EDA to identify information from dataset
Ashwini S. Kadam *et al.* (2021)	- Missing values for numerical variables imputed using mean, while for categorical variables imputed using mode - Outliers treatment is mentioned	- Mention of outlier treatment but did not outline procedure and identification for outliers
Gupta *et al.* (2021)	- Missing values imputed using K-Nearest Neighbors - Class imbalanced solved by using undersampling technique - EDA: Skewness, boxplot, QQ plot	- Did not mention of outlier treatment
H. Zhang *et al.* (2021)	- Missing value imputation using mean - Descriptive analysis of dataset and produces variance, average, minimum, maximum	- Did not mention missing value imputation for categorical variables - Did not mention treatment for outliers
Munoz *et al.* (2021)	- Class imbalance treatment using SMOTE technique to oversample dataset - Observations with missing values are removed as the total number of missing values is less than 1%	- Did not mention the use of EDA to identify information from dataset - Did not mention of outlier identification and treatment
L. Udaya Bhanu and Narayana (2021)	- Missing value imputation is mentioned - Outliers treatment is mentioned	- Mention of missing value imputation and outlier treatment but did not outline procedure and identification method
Fati (2021)	- EDA used to understand data and relationships between variables. Charts used in EDA: box plots, bar plots, histograms, heat maps, etc - Identified significant variables with high correlation using heat map - Missing values are identified using heat maps and imputed using mean for numerical variables and mode for categorical variables - Outliers are identified using box plot and removed from the dataset	
Xia *et al.* (2020)	- Missing values for numerical variables imputed using mean, while for categorical variables imputed using mode	- Did not mention treatment for outliers - Did not mention the use of EDA to identify information from dataset

Tripathi *et al.* (2020)	- Missing value imputation for categorical variables using a unique integer value - Missing values in numerical observations are removed from dataset	- Did not mention the use of EDA to identify information from dataset - Did not mention of outlier identification and treatment
Wei Zhang *et al.* (2020)	- Missing value imputation for numerical variables using mean and for categorical variables using mode - Outliers are removed from dataset	- Did not mention the use of EDA to identify information from dataset - Did not mention outlier identification method
Bao *et al.* (2019)	- Most of the missing values were imputed based on the experience of the author - Variables with missing values greater than 95%, the whole variable will be removed from dataset - Missing values treatment for categorical variables that are unable to be identified by experience of author will be replace by "1" as empty and by "0" as non-empty - Missing values treatment for numerical variables that are unable to be identified by experience of author will be replaced by mean value - Outliers are identified using boxplot and checked individually for rationality based on experience - If outliers are deemed rational, the data points will be kept as it is - Else the outliers will be replaced by the upper or lower bound of boxplot	- Did not mention the use of EDA to identify information from dataset
Blessie and Rekha (2019)	- Missing values for numerical variables imputed using mean, while for categorical variables imputed using mode - EDA is performed to identify the need of normalization, identifying missing values, identifying outliers, identifying significant variables, and identifying need of feature engineering	- Mention of outlier treatment but did not outline procedure for outliers - Mention of EDA but did not outline procedure of EDA
Chen *et al.* (2019)	- Variables with missing values greater than 60%, the whole variable will be removed from dataset - Missing values imputed using interpolation method, median, and mean	- Did not specify what criteria on variables to use the interpolation, median, or mean method for missing value imputation - Did not mention treatment for outliers

		- Did not mention the use of EDA to identify information from dataset

2.2 DISCUSSION

Based on Table 2.1, it is observed that data pre-processing is a crucial task in any data science project which can significantly affect the performance of the predictive models. The data pre-processing methods can be identified through EDA. Although many researchers mention the use of EDA. However, not all researchers outline the procedures and components for the EDA task. From the EDA task, few observations are typically identified, namely data outliers detection, dataset class imbalance detection, and missing values identification. Generally, graphics in EDA are performed using boxplots, histograms, and bar charts to identify the proportion and distribution of data. However, in some cases quantile-quantile plot and heat maps are used.

As the loan dataset typically contains missing values, various methods of missing value imputation are used by the researchers. Generally, the frequently used method is by imputing mean value for numerical variables and imputing mode value for categorical variables. In addition, other not so frequently used methods of missing value imputation include interpolation of data points, imputation based on opinions from domain experts, machine learning algorithms, and using median value. Furthermore, missing value imputation is typically not conducted on variables containing more than 55% of missing values. Such variables are directly dropped from the dataset.

Outliers in the dataset is typically identified using box plots. Outliers if not treated can significantly skewed the results leading to false interpretation. Generally, data points deemed as outliers will be removed from the dataset. However, in some cases outliers treatments are applied to retain the data points which include verification and value adjustments based on experiences from domain experts and value replacement by the upper or lower bound of the box plot.

Wang *et al.* (2022) mentioned that typically, loans applied in commercial banks have higher quantity of customers with good credit rating as compared to those of bad credit rating. This would result in the imbalance of classes in the dataset. Hence, would cause the predictive results to be biased towards the majority class and affect the accuracy of the result. Generally, the SMOTE algorithm is applied to solve the imbalance class problem as this technique is widely applied and recognized by academic community.

SECTION 3

INITIAL DATA EXPLORATION

3.1 DATASET DESCRIPTION

A public dataset to predict housing loan eligibility from Kaggle website is used in this study. The link to the dataset is as followed, https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset?select=loan-train.csv. The dataset contains 981 observations and 13 features. Figure 3.1 shows the first five observations of the dataset which provides the initial view of the dataset.

Obs	Loan_ID	Gender	Married	Dependents	Education	ApplicantIncome	CoapplicantIncome
1	LP001002	Male	No	0	Graduate	5849	0
2	LP001003	Male	Yes	1	Graduate	4583	1508
3	LP001005	Male	Yes	0	Graduate	3000	0
4	LP001006	Male	Yes	0	Not Graduate	2583	2358
5	LP001008	Male	No	0	Graduate	6000	0

Obs	LoanAmount	Self_Employed	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	.	No	360	1	Urban	Y
2	128	No	360	1	Rural	N
3	66	Yes	360	1	Urban	Y
4	120	No	360	1	Urban	Y
5	141	No	360	1	Urban	Y

Figure 3.1: First five observations of the dataset

Based on Figure 3.1, it is observed that the dataset contains a mix of categorical and numerical variables. Some categorical variables are in numeric format while some are in character format. The “Loan_ID” feature can be removed, as it serves as a reference and does not provide useful information to the dataset.

Table 3.1: Feature description and data type

Feature Name	Feature Description	Data Type	Data Sample
Gender	Gender of applicant	Categorical - Nominal	Male, Fema
Married	Marital status of applicant	Categorical - Nominal	Yes, No
Dependents	Number of dependents	Categorical – Nominal	0, 1, 2, 3+
Education	Education or qualification of applicant	Categorical – Nominal	Graduate, Not Graduate
Self_Employed	Employment status of applicant	Categorical – Nominal	Yes, No
ApplicantIncome	Monthly income of applicant	Numerical – Ratio	0 – 81000
CoapplicantIncome	Monthly income of co-applicant	Numerical – Ratio	0 – 41667
LoanAmount	Loan amount in thousands	Numerical - Ratio	9 – 700
Loan_Amount_Term	Loan term in months	Categorical – Nominal	6, 12, 36, 60, 84, 120, 180, 240, 300, 350, 360, 480
Credit_History	Missed repayment where 0 indicates missed, 1 indicate no miss	Categorical – Nominal	0, 1
Property_Area	Location of property	Categorical – Nominal	Rural, semiurban, Urban
Loan_Status	Approval status of loan	Categorical - Nominal	N, Y

Table 3.1 describes each of the features and their respective data type while displaying some data samples from each feature. Based on the table, it is observed that nine out of the 12 features are categorical features with nominal data type. While three out of the 12 features are numerical features with ratio data type.

3.2 CATEGORICAL VARIABLES EXPLORATION

This section describes the descriptive statistic and missing value identification for categorical variables prior to any data pre-processing. Figure 3.2 shows the class frequencies of each feature along with the number of missing values in each feature highlighted in red.

```
15 ods noproctitle;
16 title "Frequencies for Categorical Variables";
17 proc freq data=WORK.IMPORT1;
18     tables Credit_History Gender Married Dependents Education Self_Employed
19         Property_Area Loan_Status;
20 run;
```

Credit_History	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	148	16.41	148	16.41
1	754	83.59	902	100.00
Frequency Missing = 79				

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Fema	182	19.02	182	19.02
Male	775	80.98	957	100.00
Frequency Missing = 24				

Married	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	347	35.48	347	35.48
Yes	631	64.52	978	100.00
Frequency Missing = 3				

Dependents	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	545	57.01	545	57.01
1	160	16.74	705	73.74
2	160	16.74	865	90.48
3+	91	9.52	956	100.00
Frequency Missing = 25				

Education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Graduate	763	77.78	763	77.78
Not Graduate	218	22.22	981	100.00

Self_Employed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	807	87.15	807	87.15
Yes	119	12.85	926	100.00
Frequency Missing = 55				

Property_Area	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Rural	290	29.56	290	29.56
Semiurban	349	35.58	639	65.14
Urban	342	34.86	981	100.00

Loan_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	192	31.27	192	31.27
Y	422	68.73	614	100.00
Frequency Missing = 367				

Figure 3.2: Measures of frequency for categorical variables with code snippet

Based on Figure 3.2, the following observations are identified and listed as followed:

- Missing values are present in six of the eight categorical variables namely “Credit_History”, “Gender”, “Married”, “Dependents”, “Self_Employed”, and “Loan_Status”. The “Education” and “Property_Area” variables do not have missing values observed.
- The “Loan_Status” has 367 missing values which is about 37% with respect to the “Loan_Status” total observations. Two classes are observed in “Loan_Status” namely “N” and “Y”. The proportion of class “Y” in “Loan_Status” is observed to be higher than class “N”, which is about 31% of class “N” and 69% of class “Y”. This indicates that a higher proportion of applicants got their loan approved as compared to applicants who got their loan rejected.

- The “Property_Area” does not have any missing values. Three classes are observed in “Property_Area” namely “Rural”, “Semiurban”, and “Urban”. The proportion of class “Semiurban” is similar to “Urban” which is about 35% for both classes and is both higher than class “Rural” which has about 30% proportion. This indicates majority of loan applicants arrive from the semi-urban and urban area.
- The “Self_Employed” has 55 missing values which is about 6% with respect to the “Self_Employed” total observations. Two classes are observed in “Self_Employed” namely “Yes” and “No”. The proportion of class “No” in “Self_Employed” is observed to be higher than class “Yes”, which is about 87% for class “No” and about 13% for class “Yes”. This indicates that majority of applicants are employees and might be earning a median income range.
- The “Education” does not have any missing values. Two classes are observed in “Education” namely “Graduate” and “Not Graduate”. The proportion of class “Graduate” is higher than class “Not Graduate”, which is about 78% for “Graduate” and about 22% for “Not Graduate”. This indicates majority of applicants are graduates and might be earning a higher income thus has more buying power.
- The “Dependents” has 25 missing values which is about 3% with respect to the “Dependents” total observations. Four classes are observed in “Dependents” namely “0”, “1”, “2”, “3+”. The “0” class is observed to have the highest proportion as compared to other classes which takes up about 57% of the total observations. The “1” and “2” classes have same proportion of about 17% of the total observations. While class “3+” has the least proportion of about 9% of the total observations. This indicates majority of the applicants does not have children and even if the applicants have children, most of them have one or two children.
- The “Married” has three missing values which is about less than 1% with respect to the “Married” total observations. Two classes are observed in “Married” namely “Yes” and “No”. The proportion of class “Yes” is higher than class “No”, which is about 65% for “Yes” and about 35% for “No”. This indicates majority of the applicants are married and probably applying for housing loan to buy a house for settling down and have a family.
- The “Gender” has 24 missing values which is about 3% with respect to the “Gender” total observations. Two classes are observed in “Gender” namely “Male” and “Fema”. The

proportion of class “Male” is higher than class “Fema”, which is about 81% for “Male” and about 19% for “Fema”. This indicates majority of the applicants are male which is probably due to the norm of society where the men have to provide for the family.

- The “Credit_History” has 79 missing values which is about 8% with respect to the “Credit_History” total observations. Two classes are observed in “Credit_History” namely “0” and “1”. The proportion of class “1” is higher than class “0”, which is about 84% for “1” and about 16% for “0”. This indicates majority of the applicants has good history of repayments which can be used to judge future repayment behavior.

A special treatment is required for the “Loan_Amount_Term” variable. It is a categorical variable with 12 levels of classes. Figure 3.3 shows the frequency for each class in the variable along with the number of missing values highlighted in red.

```
15 proc freq data=WORK.IMPORT;
16     tables Loan_Amount_Term / plots=none;
17 run;
```

Loan_Amount_Term	Frequency	Percent	Cumulative Frequency	Cumulative Percent
6	1	0.10	1	0.10
12	2	0.21	3	0.31
36	3	0.31	6	0.62
60	3	0.31	9	0.94
84	7	0.73	16	1.66
120	4	0.42	20	2.08
180	66	6.87	86	8.95
240	8	0.83	94	9.78
300	20	2.08	114	11.86
350	1	0.10	115	11.97
360	823	85.64	938	97.61
480	23	2.39	961	100.00
Frequency Missing = 20				

Figure 3.3: Frequency table for “Loan_Amount_Term” with code snippet

Based on Figure 3.3 it observed that the “Loan_Amount_Term” has 20 missing values which is about 2% with respect to the “Loan_Amount_Term” total observations. 12 classes are observed in the table. The variable represents the periods of repayment in months which the applicant must commit to paying back to the bank. It is observed that most applicants applied for the 360 months term which is equivalent to 30 years. This is the standard term typically used for a housing loan as the price of a house is typically very large in amount and requires a long time to pay back. In addition, an inconsistent data was observed which one applicant has applied for 350 months term which might be an error and can be converted to the 360 months term. As all other terms are a multiple of six months.

3.3 NUMERICAL VARIABLES EXPLORATION

This section describes the descriptive statistic, missing value identification, and outlier identification for numerical variables prior to any data processing. Figure 3.4 shows the central tendency and dispersion along with the number of missing values in each feature highlighted in red. While Figure 3.5 shows the box plots for each continuous variable. In addition, the code snippet is attached above each figure.

```
15 ods noproctitle;
16
17 /*** Analyze numeric variables ***/
18 title "Descriptive Statistics for Numeric Variables";
19
20 proc means data=WORK.IMPORT n nmiss min mean median max std;
21     var ApplicantIncome CoapplicantIncome LoanAmount;
22 run;
```

Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
ApplicantIncome	981	0	0	5179.80	3800.00	81000.00	5695.10
CoapplicantIncome	981	0	0	1601.92	1110.00	41667.00	2718.77
LoanAmount	954	27	9.0000000	142.5115304	126.0000000	700.0000000	77.4217431

Figure 3.4: Measures of central tendency and dispersion for numerical variables with code snippet

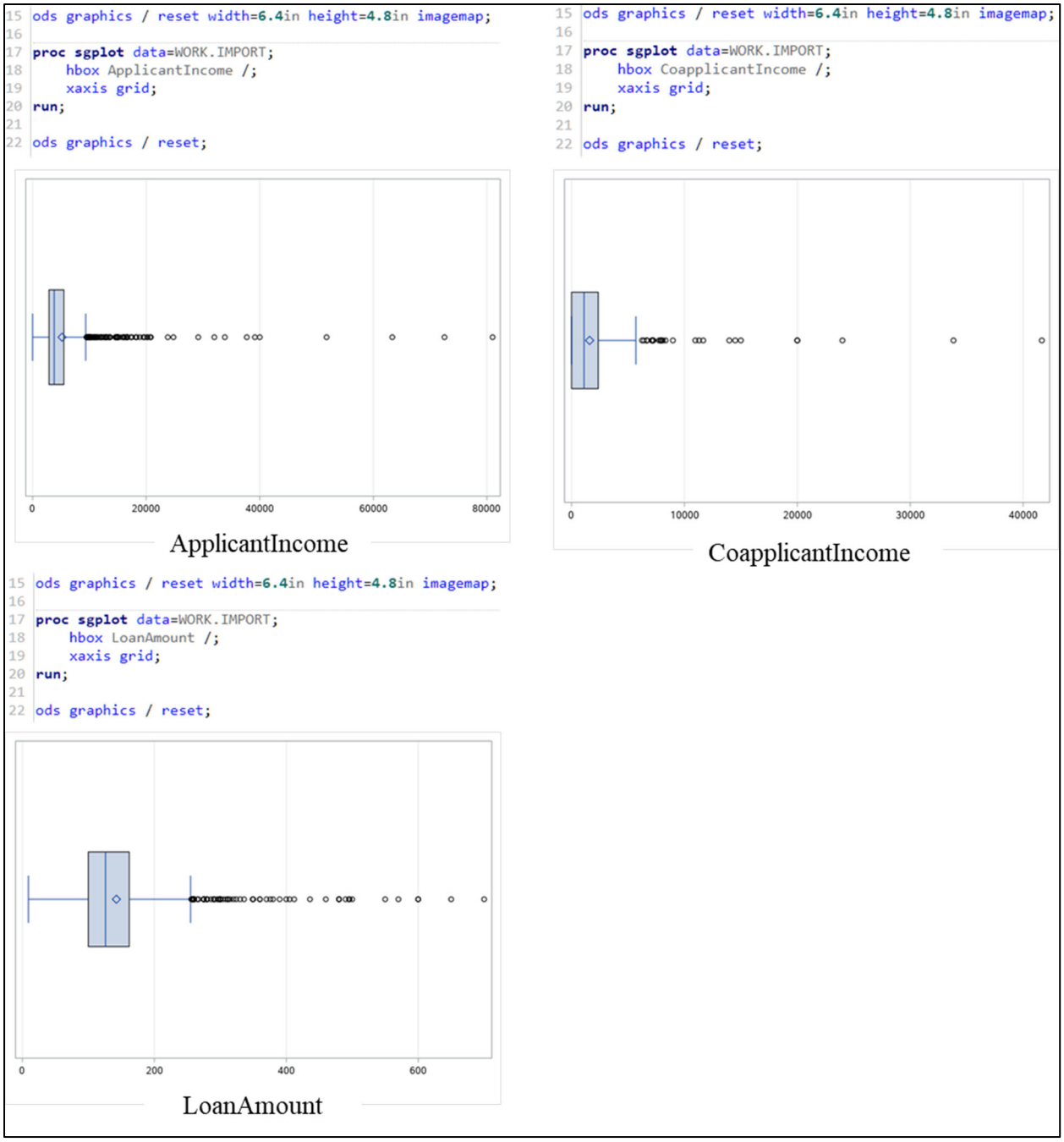


Figure 3.5: Box plots for numerical variables with code snippet

Based on Figure 3.4 and Figure 3.5, the following observations are identified and listed as followed:

- The “ApplicantIncome” does not have any missing values but based on the box plot a high number of outliers are observed in the higher value region. Furthermore, based on the box plot the variable is positively skewed. This indicates majority of applicant has income in the lower bracket range.

- The “CoApplicantIncome” does not have any missing values but based on the box plot a high number of outliers are observed in the higher value region. Furthermore, based on the box plot the variable is positively skewed. This indicates majority of co-applicant has income in the lower bracket range.
- The “LoanAmount” has 27 missing values in addition based on the box plot, a high number of outliers are observed in the higher value region. Furthermore, based on the box plot the variable is positively skewed. This indicates majority of applicant is applying for a small amount of loan.

SECTION 4

DATA PRE-PROCESSING

This section describes the data pre-processing methods applied to the dataset which includes missing value imputation, outliers treatment, and inconsistent data treatment.

4.1 MISSING DATA IMPUTATION

As described in the previous section, missing values are observed in multiple variables. The imputation process for the missing values will be described in this section and partitioned into categorical variables and numerical variables subsections.

4.1.1 Categorical variables missing value imputation

The missing values in categorical variables were identified in the previous section. Seven out of the nine categorical variables contain missing values. Based on the literature survey, missing values treatment for categorical variables are typically treated using the class mode to replace the missing value thus similar procedure will be applied here. Figure 4.1 shows the snippet for applying mode imputation to missing values in categorical variables.

```
data categorical_imputation;
    set work.import;
    if Credit_History ='' then Credit_History = 1;
    if Gender = '' then Gender = 'Male';
    if Married = '' then Married = 'Yes';
    if Dependents = '' then Dependents ='0';
    if Self_Employed = '' then Self_Employed = 'No';
    if Loan_Amount_Term = '' then Loan_Amount_Term = 360;
    if Loan_Amount_Term = 350 then Loan_Amount_Term = 360;
run;
```

Figure 4.1: Snippet for missing value imputation of categorical variables

Based on Figure 4.1, the following observations are identified and listed as followed:

- The missing values in “Credit_History” are replaced by “1” as it is the class with highest frequency.

- The missing values in “Gender” are replaced by “Male” as it is the class with highest frequency.
- The missing values in “Married” are replaced by “Yes” as it is the class with highest frequency.
- The missing values in “Dependents” are replaced by “0” as it is the class with highest frequency.
- The missing values in “Self_Employed” are replaced by “No” as it is the class with highest frequency.
- The missing values in “Loan_Amount_Term” are replaced by “360” as it is the class with highest frequency. In addition, one of the observations identified as inconsistent in the previous section is converted from “350” to “360”.

Since the “Loan_Status” has a high number of missing values, imputing using the mode would skew the data and results would not be reliable. Therefore, a different method is applied for imputing missing values for “Loan_Status” which is using logistic regression predictive model to compute the probability of the class for each missing value. Figure 4.2 shows the code snippet of using logistic regression to compute the class probability for “Loan_Status”. Backward elimination is performed to ensure only significant variables are used in the predictive model.

```
15 ods noproctitle;
16 ods graphics / imagemap=on;
17
18 proc logistic data=WORK.NUMERICAL_IMPUTATION;
19     class Gender Married Dependents Education Self_Employed Loan_Amount_Term
20         Credit_History Property_Area / param=glm;
21     model Loan_Status(event='Y')=Gender Married Dependents Education Self_Employed
22         Loan_Amount_Term Credit_History Property_Area ApplicantIncome
23         CoapplicantIncome LoanAmount / link=logit selection=backward slstay=0.05
24         hierarchy=single details technique=fisher;
25     output out=work.Logistic_stats0001 predicted=pred_;
26 run;
```

Figure 4.2: Logistic regression snippet for “Loan_Status”

Figure 4.3 shows the results from the predictive model provided in probability in the last column as “pred_”. Further processing is required to convert the probability into class “Y” and “N”. A probability greater than 0.5 would indicate class “Y”, and a probability lesser than 0.5 would indicate class “N”.

Obs	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status	pred_
9	LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban		0.79970
12	LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban		0.79970
18	LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban		0.79970
21	LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360	1	Urban		0.79970
29	LP001051	Male	No	0	Not Graduate	No	3276	0	78	360	1	Urban		0.70083

Figure 4.3: Probability output for class for “Loan_Status”

```
data complete_imputation;
    set work.import1;
    if Loan_Status = '' AND pred_ > 0.5 THEN Loan_Status = 'Y';
    if Loan_Status = '' AND pred_ < 0.5 THEN Loan_Status = 'N';
run;
```

Figure 4.4: Snippet for converting probability to class for “Loan_Status”

Figure 4.4 shows the conversion of probability into classes for the missing values in “Loan_Status” based on the logistic regression predictive result. Therefore, missing values in the categorical variables have completed the imputation process. Figure 4.5 shows that after imputing missing values, there are no more missing values in the categorical variables. Code snippet for identifying missing values in categorical variables will be attached under Appendix A.

Gender	Frequency	Percent
Non-missing	981	100.00

Married	Frequency	Percent
Non-missing	981	100.00

Dependents	Frequency	Percent
Non-missing	981	100.00

Education	Frequency	Percent
Non-missing	981	100.00

Self_Employed	Frequency	Percent
Non-missing	981	100.00

Loan_Amount_Term	Frequency	Percent
Non-missing	981	100.00

Credit_History	Frequency	Percent
Non-missing	981	100.00

Property_Area	Frequency	Percent
Non-missing	981	100.00

Loan_Status	Frequency	Percent
Non-missing	981	100.00

Figure 4.5: Missing values after imputation for categorical variables

4.1.2 Numerical variables missing value imputation

The missing values in numerical variables were identified in the previous section. One out of the three numerical variables contain missing values. Based on the literature survey, missing values treatment for numerical variables are typically treated using the mean value to replace the missing value thus similar procedure will be applied here. Figure 4.6 shows the snippet for applying mean value imputation to missing values in numerical variable.

```
1 data NUMERICAL_IMPUTATION;
2     SET WORK.CATEGORICAL_IMPUTATION;
3     if LoanAmount = '' then LoanAmount = 142.5;
4 run;
```

Figure 4.6: Snippet for missing value imputation of numerical variable

Based on Figure 4.6, the snippet shows the missing values in "LoanAmount" are replaced by "142.5" as it is the mean value of the variable. Therefore, missing values in the numerical variables have completed the imputation process. Figure 4.7 shows that after imputing missing values, there are no more missing values in the numerical variables. Code snippet for identifying missing values in numerical variables will be attached under Appendix A.

ApplicantIncome	Frequency	Percent
Non-missing	981	100.00

CoapplicantIncome	Frequency	Percent
Non-missing	981	100.00

LoanAmount	Frequency	Percent
Non-missing	981	100.00

Figure 4.7: Missing values after imputation for numerical variables

4.2 OUTLIERS TREATMENT

The previous section has identified the present of outliers in the numerical variables namely "ApplicantIncome", "CoapplicantIncome", and "LoanAmount" while Figure 4.8 to Figure 4.10 shows the box plots for each numerical variable respectively. In addition, Table 4.1 shows the measures from the box plots which will be used to identify outliers. Based on the literature survey, outliers treatment are typically by removing them from the dataset.

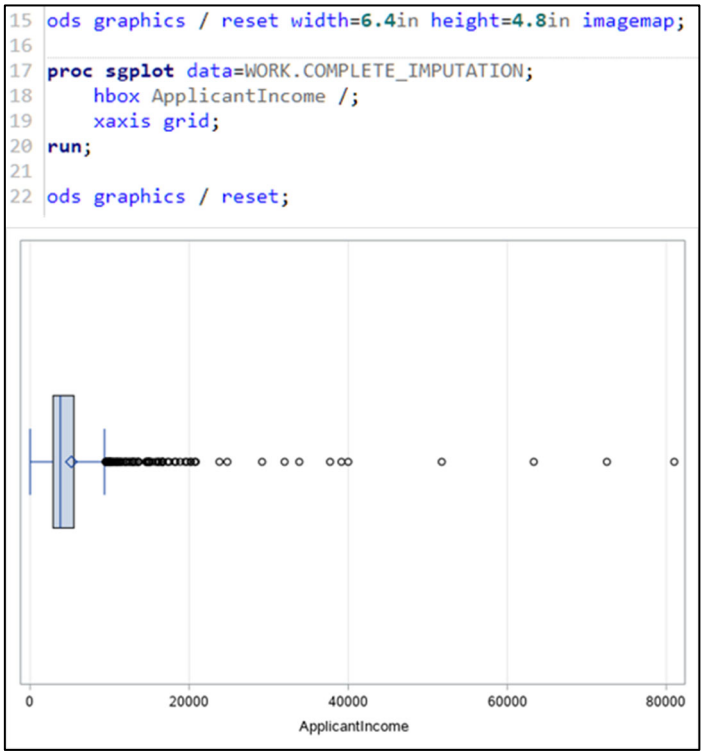


Figure 4.8: Box plot for "ApplicantIncome" with code snippet

```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sgplot data=WORK.COMPLETE_IMPUTATION;
18     hbox CoapplicantIncome /;
19     xaxis grid;
20 run;
21
22 ods graphics / reset;
```

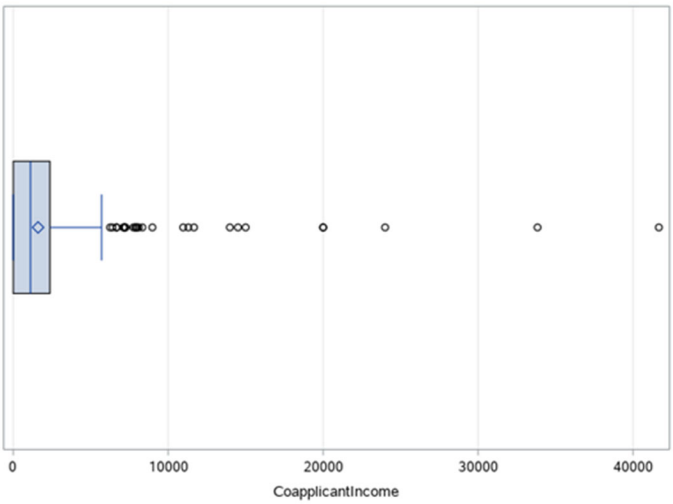


Figure 4.9: Box plot for “CoapplicantIncome” with code snippet

```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sgplot data=WORK.COMPLETE_IMPUTATION;
18     hbox LoanAmount /;
19     xaxis grid;
20 run;
21
22 ods graphics / reset;
```

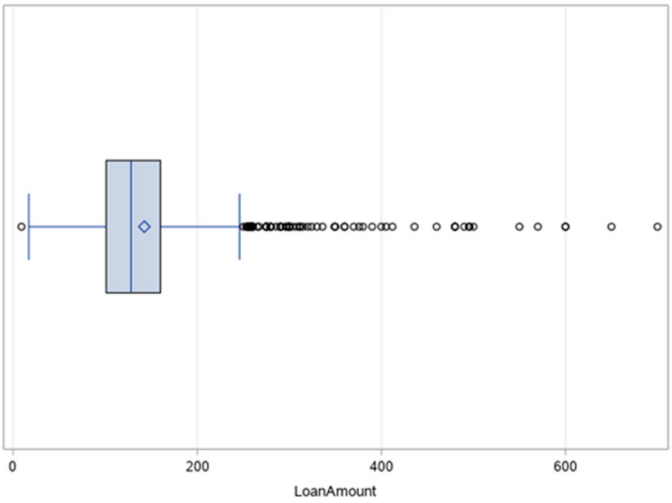


Figure 4.10: Box plot for “LoanAmount” with code snippet

Table 4.1: Box plots measures for each numerical variable

Measures	ApplicantIncome	CoapplicantIncome	LoanAmount
Data Minimum	0	0	9
Minimum Whisker	0	0	17
First Quartile	2875	0	101
Median	3800	1110	128
Third Quartile	5516	2365	160
Maximum Whisker	9357	5701	246
Data Maximum	81000	41667	700

Based on Figure 4.8 to Figure 4.10, all numerical variables contain outliers. Therefore, the identified outliers will be removed from the dataset. The following list the criteria based on Table 4.1 where data points are considered as outliers in each feature:

- In "ApplicantIncome", data points greater than 9357 are considered as outliers. There are no outliers below the minimum whisker.
- In "CoapplicantIncome", data points greater than 5701 are considered as outliers. There are no outliers below the minimum whisker.
- In "LoanAmount", data points less than 17 or greater than 246 are considered as outliers.

Figure 4.11 shows the code snippet for performing outliers removal from the dataset based on the criteria as mentioned. A total of 135 observations are removed from the dataset which are considered as outliers.

```
1 data outliers_treatment;
2     set work.complete_imputation;
3     if ApplicantIncome > 9357 OR
4     CoapplicantIncome > 5701 OR
5     LoanAmount > 246 OR
6     LoanAmount < 17 THEN delete;
7 run;
```

Figure 4.11: Code snippet for removing outliers

SECTION 5

EXPLORATORY DATA ANALYSIS

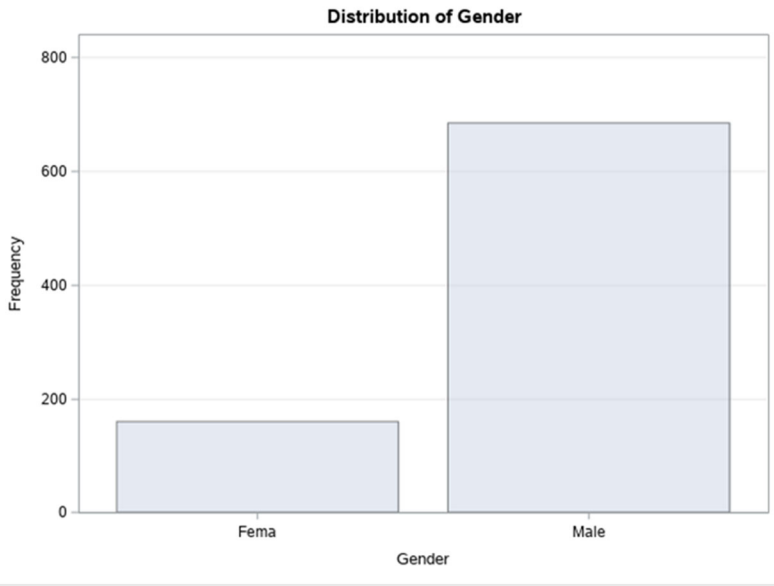
This section describes the EDA on the datasets by employing numerical and graphical examination to identify the data characteristics and relationships. The EDA will be performed on imputed dataset with no missing values. In addition, at the last section would be the interpretation of hypothesis formulated.

5.1 CATEGORICAL VARIABLES EDA

This section describes the EDA performed specifically on categorical variables. The first categorical variable "Loan_ID" will not be discussed as it is the row reference in the dataset. Figure 5.1 shows the bar chart along with the class frequencies for the "Gender" variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Fema	161	19.03	161	19.03
Male	685	80.97	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Gender / plots=(freqplot);
22 run;
```

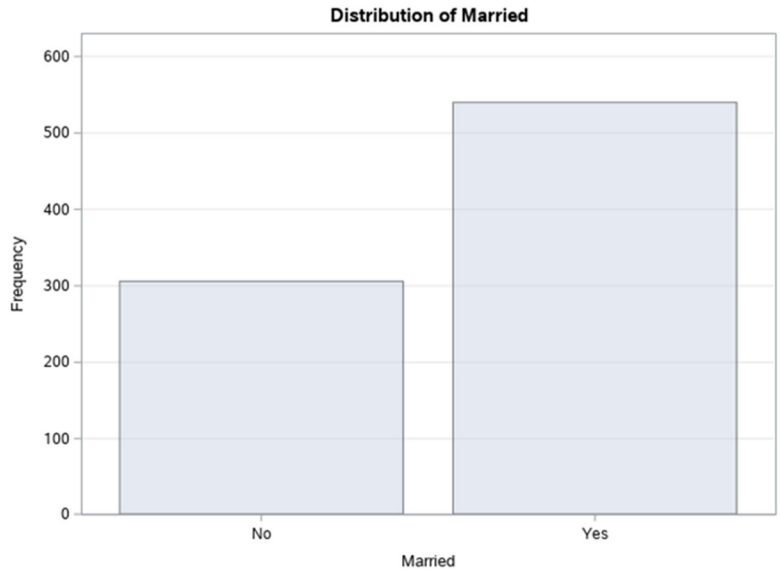
Figure 5.1: Bar chart for the “Gender” variable

Based on Figure 5.1, it is observed that the male gender is significantly greater than the female gender which the female is about 19% and the male is about 81%. Which in numbers, 161 for females and 685 for males. This indicates that there are more males applying for housing loan which can be due to the social norm where males are responsible to provide for the family by providing shelter.

Figure 5.2 shows the bar chart along with the class frequencies for the “Married” variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Married	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	306	36.17	306	36.17
Yes	540	63.83	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Married / plots=(freqplot);
22 run;
```

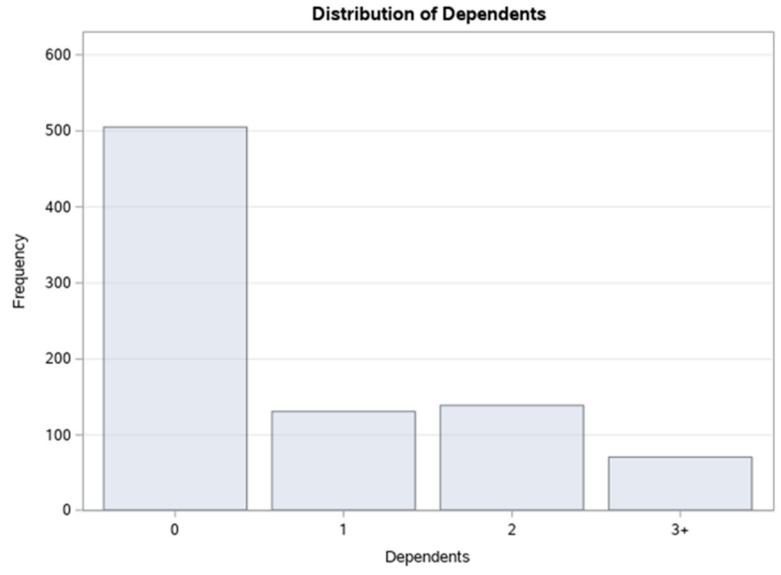
Figure 5.2: Bar chart for the “Married” variable

Based on Figure 5.2, it is observed that more applicants are married than applicants who are not married which is about 36% for not yet married and about 64% for married. Which in numbers, 306 for not yet married and 540 for married. This indicates that there are more married applicants applying for housing loan which can be due to the need of a house for accommodating bigger number of family members.

Figure 5.3 shows the bar chart along with the class frequencies for the “Married” variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Dependents	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	505	59.69	505	59.69
1	131	15.48	636	75.18
2	139	16.43	775	91.61
3+	71	8.39	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Dependents / plots=(freqplot);
22 run;
```

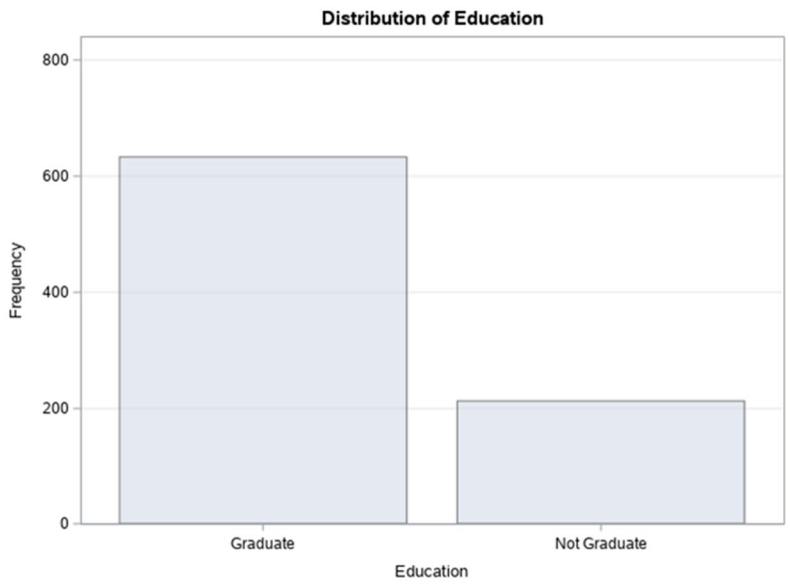
Figure 5.3: Bar chart for the “Dependents” variable

Based on Figure 5.3, it is observed that majority applicants do not have dependents. Even if the applicants have dependents, they would have about one or two dependents. Applicants without dependents makes up about 60% of the proportion, while applicants with one and two dependents makes up about 16% respectively, and applicants with three or more dependents makes up about 8% of the proportion. Which in numbers, 505 applicants without dependents, 131 applicants with one dependent, 139 applicants with two dependents, and 71 applicants with three or more dependents. This can indicate that majority applicants may be getting a house for the purpose of expecting an additional member to their family.

Figure 5.4 shows the bar chart along with the class frequencies for the “Education” variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Graduate	633	74.82	633	74.82
Not Graduate	213	25.18	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Education / plots=(freqplot);
22 run;
```

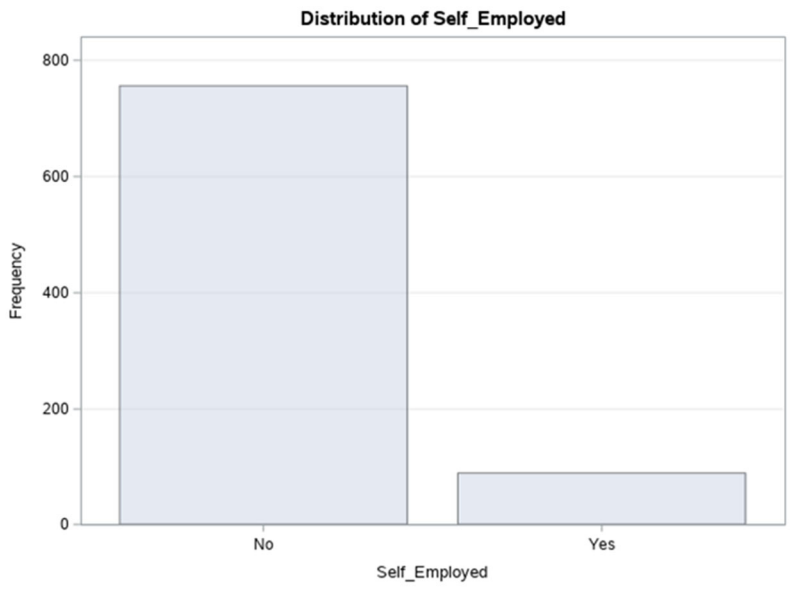
Figure 5.4: Bar chart for the “Education” variable

Based on Figure 5.4, it is observed that more applicants are graduates than applicants who are not graduates which is about 75% for graduates and about 25% for not graduates. Which in numbers, 633 for graduates and 213 for not graduates. This indicates that there are more graduate applicants applying for housing loan which can be due to the higher buying power as they are employed in higher salary jobs.

Figure 5.5 shows the bar chart along with the class frequencies for the “Education” variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Self_Employed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	756	89.36	756	89.36
Yes	90	10.64	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Self_Employed / plots=(freqplot);
22 run;
```

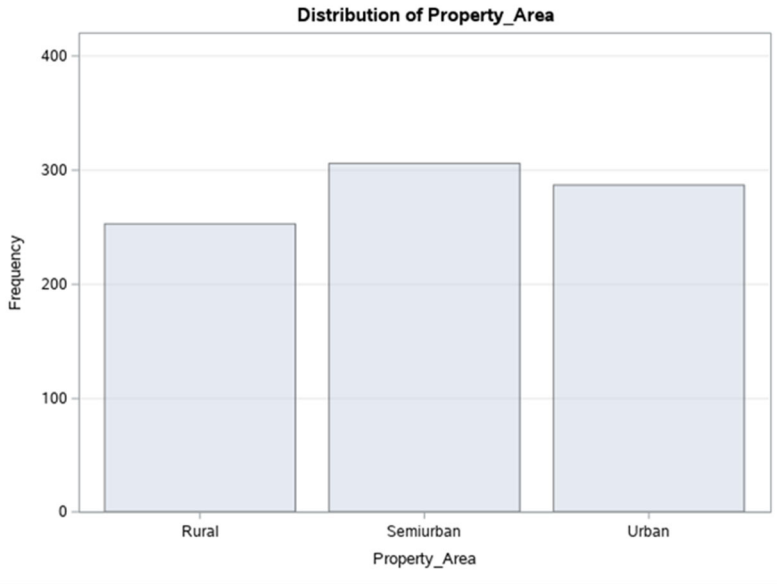
Figure 5.5: Bar chart for the “Self_Employed” variable

Based on Figure 5.5, it is observed that more applicants are not self-employed than applicants who are self-employed which is about 89% for not self-employed and about 11% for self-employed. Which in numbers, 756 for not self-employed and 90 for self-employed. This indicates that there are more applicants who are not self-employed and require the housing loan to assist them in owning a house as their salary might be around the median income range in their country.

Figure 5.6 shows the bar chart along with the class frequencies for the “Property_Area” variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Property_Area	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Rural	253	29.91	253	29.91
Semiurban	306	36.17	559	66.08
Urban	287	33.92	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Property_Area / plots=(freqplot);
22 run;
```

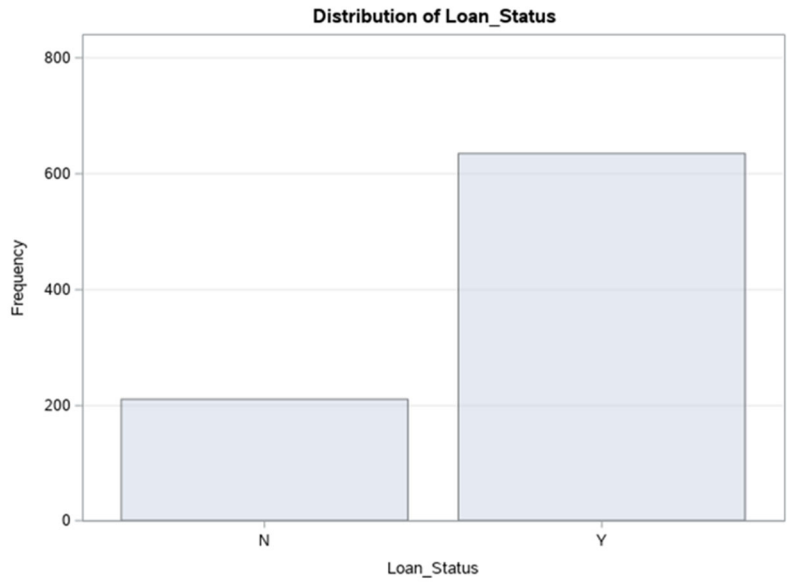
Figure 5.6: Bar chart for the “Property_Area” variable

Based on Figure 5.6, it is observed that the semiurban area has the highest frequency as compared to the urban area and rural area which is about 30% for rural area, about 36% for semiurban area, and about 34% for urban area. Which in numbers, 253 for rural area, 306 for semiurban area, and 287 for urban area. This indicates that more applicants are buying houses in the semiurban area which can be due to the housing prices are lower in semiurban area while situated near to the city.

Figure 5.7 shows the bar chart along with the class frequencies for the “Loan_Status” variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Loan_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	211	24.94	211	24.94
Y	635	75.06	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Loan_Status / plots=(freqplot);
22 run;
```

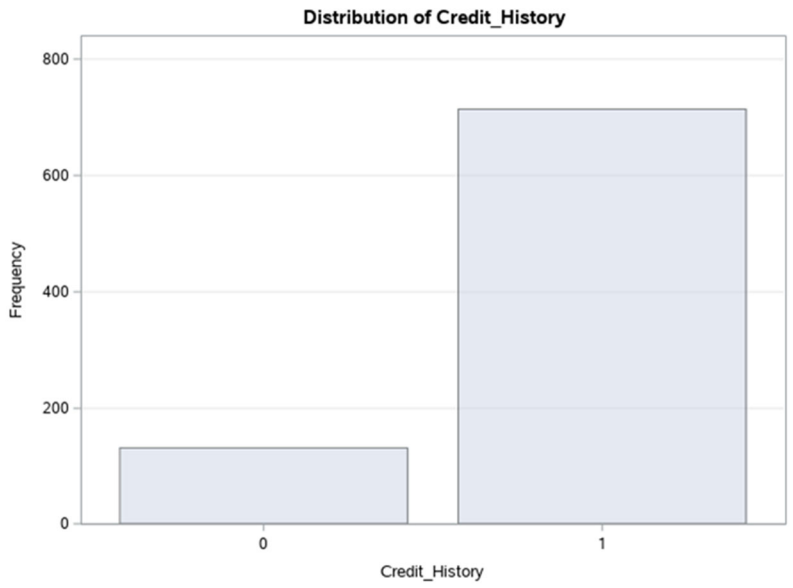
Figure 5.7: Bar chart for the “Loan_Status” variable

Based on Figure 5.7, it is observed that more applicants got their loan approved than applicants who got their loan rejected which is about 25% for rejected loans and about 75% for approved loans. Which in numbers, 211 for rejected loans and 635 for approved loans. This indicates that there are more applicants who got their loan approved which can be due to the lenient policy of the bank in providing loan services to customers.

Figure 5.8 shows the bar chart along with the class frequencies for the “Credit_History” variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Credit_History	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	132	15.60	132	15.60
1	714	84.40	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Credit_History / plots=(freqplot);
22 run;
```

Figure 5.8: Bar chart for the “Credit_History” variable

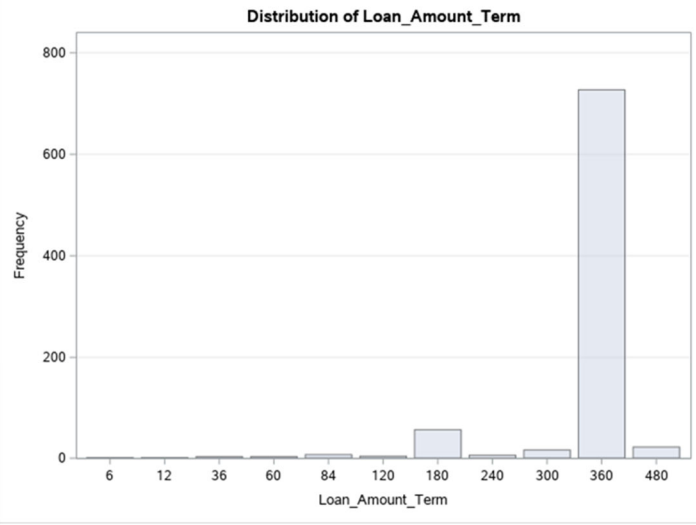
Based on Figure 5.8, it is observed that more applicants have a good performance in repayment history as compared to those who got bad performance in repayment history which is about 16% for bad performance and about 84% for good performance. Which in numbers, 132 for bad performance and 714 for good performance. This indicates that there are more applicants with

good performance in repayment history which can be used as a judgement to the future repayment performance of the applicants which will improve the loan approval rate.

Figure 5.9 shows the bar chart along with the class frequencies for the "Loan_Amount_Term" variable. In addition, code snippet for generating the figure is attached below the bar chart.

Frequencies for Categorical Variables

Loan_Amount_Term	Frequency	Percent	Cumulative Frequency	Cumulative Percent
6	1	0.12	1	0.12
12	1	0.12	2	0.24
36	3	0.35	5	0.59
60	3	0.35	8	0.95
84	7	0.83	15	1.77
120	4	0.47	19	2.25
180	56	6.62	75	8.87
240	6	0.71	81	9.57
300	16	1.89	97	11.47
360	727	85.93	824	97.40
480	22	2.60	846	100.00




```
15 ods noproctitle;
16
17 /*** Analyze categorical variables ***/
18 title "Frequencies for Categorical Variables";
19
20 proc freq data=WORK.IMPORT2;
21     tables Loan_Amount_Term / plots=(freqplot);
22 run;
```

Figure 5.9: Bar chart for the "Loan_Amount_Term" variable

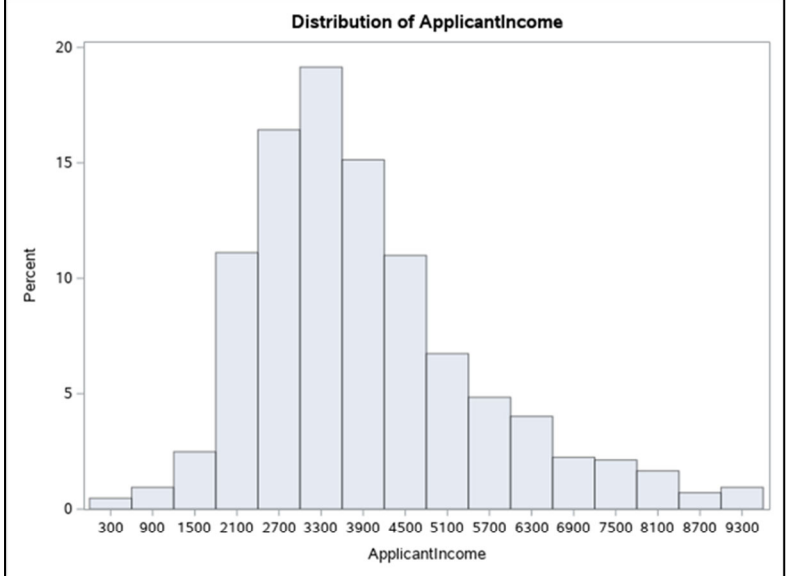
Based on Figure 5.9, it is observed that most applicants applied for the 360 months loan term as compared to other loan term periods. This can indicate that a housing loan is generally a large sum of money which requires a longer period of repayment. In which, 360 months is the typical loan term for a housing loan.

5.2 NUMERICAL VARIABLES EDA

This section describes the EDA performed specifically on numerical variables. Figure 5.10 shows the histogram along with the descriptive statistics for the "ApplicantIncome" variable. In addition, code snippet for generating the figure is attached below the histogram.

Descriptive Statistics for Numeric Variables

Analysis Variable : ApplicantIncome						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
846	0	0	3904.72	3585.50	9328.00	1638.70




```
15 ods noproctitle;
16
17 /*** Analyze numeric variables ***/
18 title "Descriptive Statistics for Numeric Variables";
19
20 proc means data=WORK.IMPORT2 n nmiss min mean median max std;
21     var ApplicantIncome;
22 run;
23
24 title;
25
26 proc univariate data=WORK.IMPORT2 noprint;
27     histogram ApplicantIncome;
28 run;
```

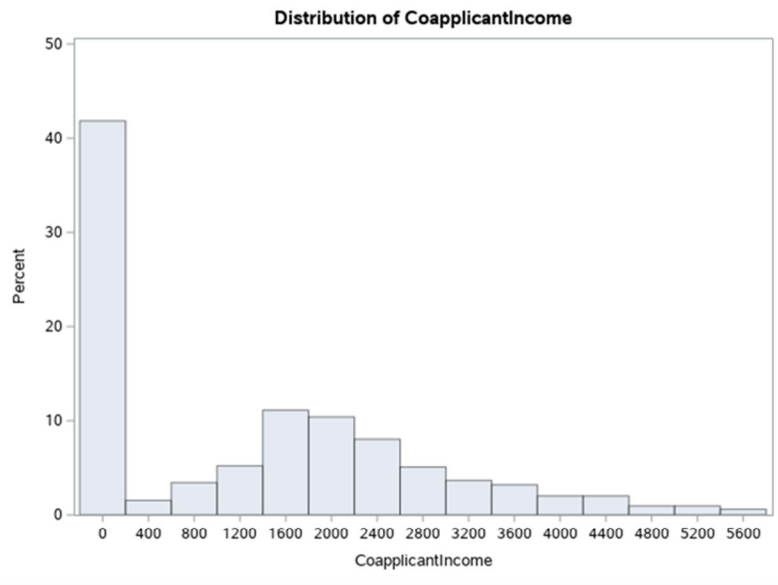
Figure 5.10: Histogram for the "ApplicantIncome" variable

Based on Figure 5.10, it is observed that the "ApplicantIncome" has a positively skewed distribution. This indicates that majority applicants have income on the lower bracket which is about the median value.

Figure 5.11 shows the histogram along with the descriptive statistics for the "CoApplicantIncome" variable. In addition, code snippet for generating the figure is attached below the histogram.

Descriptive Statistics for Numeric Variables

Analysis Variable : CoapplicantIncome						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
846	0	0	1343.92	1250.00	5701.00	1414.55




```
15 ods noproctitle;
16
17 /*** Analyze numeric variables ***/
18 title "Descriptive Statistics for Numeric Variables";
19
20 proc means data=WORK.IMPORT2 n nmiss min mean median max std;
21     var CoapplicantIncome;
22 run;
23
24 title;
25
26 proc univariate data=WORK.IMPORT2 noprint;
27     histogram CoapplicantIncome;
28 run;
```

Figure 5.11: Histogram for the "CoApplicantIncome" variable

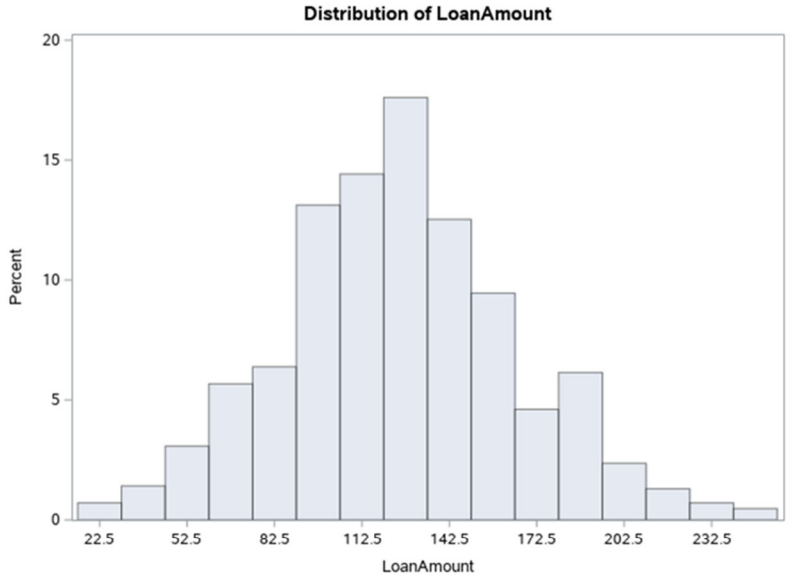
Based on Figure 5.11, it is observed that majority of co-applicants do not have an income. However, if the zero income is ignored, the "CoApplicantIncome" has a slight positive skewed distribution.

This indicates majority of co-applicants have income on the lower bracket or do not have any income at all.

Figure 5.12 shows the histogram along with the descriptive statistics for the "LoanAmount" variable. In addition, code snippet for generating the figure is attached below the histogram.

Descriptive Statistics for Numeric Variables

Analysis Variable : LoanAmount						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
846	0	17.0000000	124.4787234	123.0000000	246.0000000	40.3911587




```
15 ods noproctitle;
16
17 /*** Analyze numeric variables ***/
18 title "Descriptive Statistics for Numeric Variables";
19
20 proc means data=WORK.IMPORT2 n nmiss min mean median max std;
21     var LoanAmount;
22 run;
23
24 title;
25
26 proc univariate data=WORK.IMPORT2 noprint;
27     histogram LoanAmount;
28 run;
```

Figure 5.12: Histogram for the "LoanAmount" variable

Based on Figure 5.12, it is observed that the "LoanAmount" has a normal distribution. This indicates that majority applicants have a loan amount about the mean which is about 124,000. In which this can indicates majority of house prices are about this price range.

Pearson Correlation Coefficients, N = 846			
	ApplicantIncome	CoapplicantIncome	LoanAmount
ApplicantIncome	1.00000	-0.28670	0.38434
CoapplicantIncome	-0.28670	1.00000	0.30843
LoanAmount	0.38434	0.30843	1.00000

```
15 ods noproctitle;
16 ods graphics / imagemap=on;
17
18 proc corr data=WORK.IMPORT2 pearson nosimple noprob plots=none;
19     var ApplicantIncome CoapplicantIncome LoanAmount;
20 run;
```

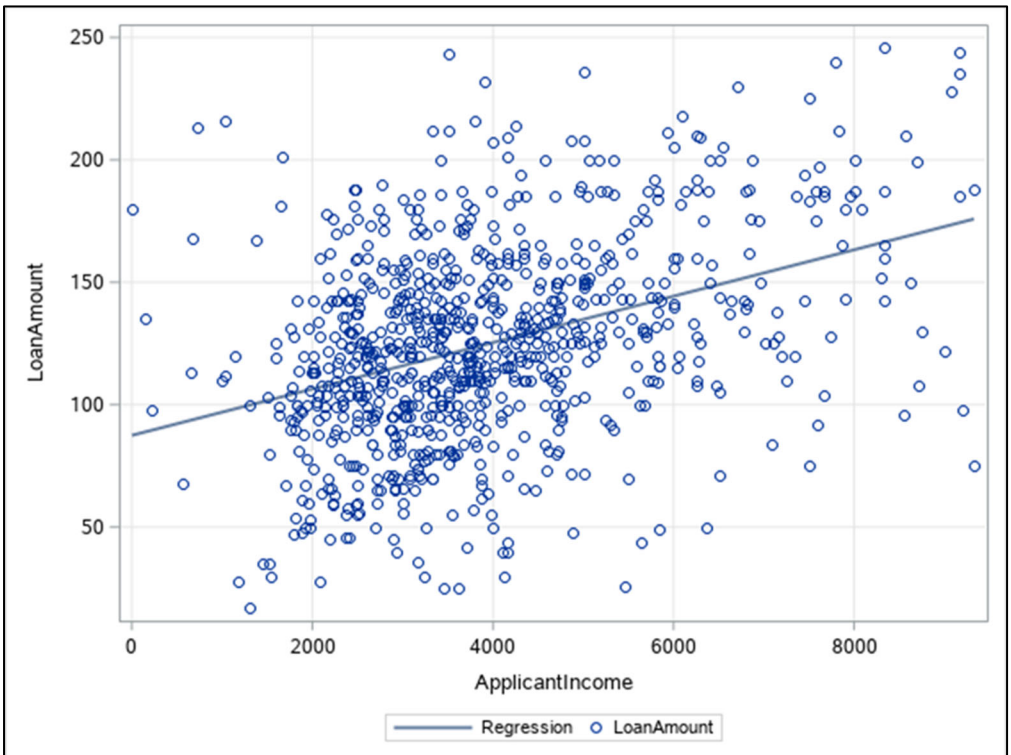
Figure 5.13: Correlation analysis for numerical variables

Figure 5.13 shows the correlation analysis on numerical variables. In addition, code snippet for generating the figure is attached below the histogram. It is observed that applicant income is negatively and weakly correlated with co-applicant income. Where a point increase in applicant income would result in a small decrease in co-applicant income. Second observation is that applicant income is positively and moderately correlated with loan amount. Where a point increase in applicant income would result in a moderate increase in loan amount. Third observation is that co-applicant income is positively and moderately correlated with loan amount. Where a point increase in co-applicant income would result in a moderate increase in loan amount.

5.3 HYPOTHESIS VALIDATION

This section describes the validation of the hypothesis formulated in the earlier section. Graphical and numerical examination of the dataset would be performed to validate the hypotheses.

Hypothesis 1: Applicants with higher income would apply for a higher loan amount.

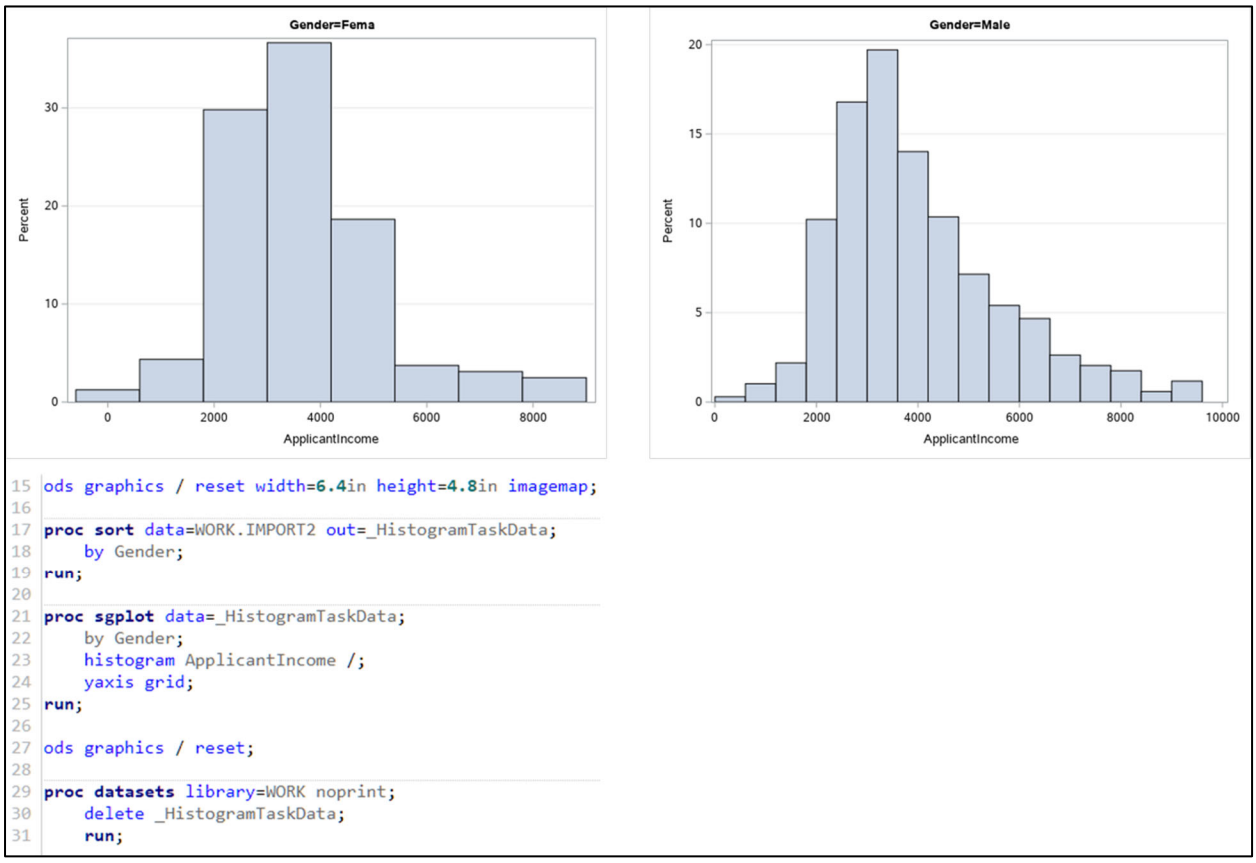



```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sgplot data=WORK.IMPORT2;
18     reg x=ApplicantIncome y=LoanAmount / nomarkers;
19     scatter x=ApplicantIncome y=LoanAmount /;
20     xaxis grid;
21     yaxis grid;
22 run;
23
24 ods graphics / reset;
```

Figure 5.14: Scatter plot between “ApplicantIncome” and “LoanAmount”

Figure 5.14 shows a scatter plot between “ApplicantIncome” and “LoanAmount”. In addition, code snippet for generating the figure is attached below the scatter plot. Based on Figure 5.14, a simple linear regression line is plotted within the scatter plot which indicates a positive gradient for the relationship between “ApplicantIncome” and “LoanAmount”. This signify that an increase in the income of applicants would result in an increase in the loan amount borrowed. Therefore, the hypothesis is validated.

Hypothesis 2: Male applicants have a higher income than the female applicants.

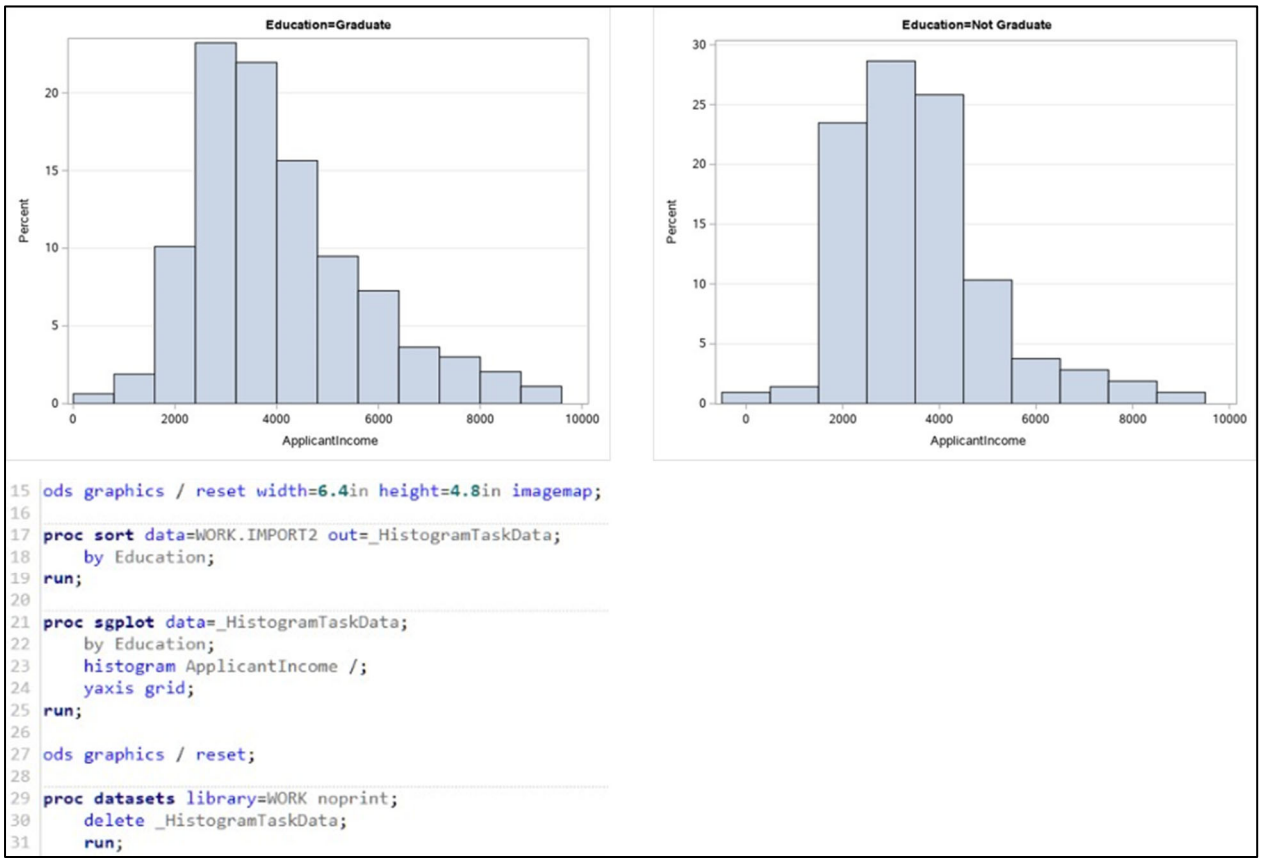



```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sort data=WORK.IMPORT2 out=_HistogramTaskData;
    by Gender;
run;

proc sgplot data=_HistogramTaskData;
    by Gender;
    histogram ApplicantIncome /;
    yaxis grid;
run;

ods graphics / reset;

proc datasets library=WORK noprint;
    delete _HistogramTaskData;
    run;
```

Figure 5.15: Histogram for “ApplicantIncome” based on “Gender”

Figure 5.15 shows the histogram for “ApplicantIncome” segmented by “Gender” where the left histogram shows the income for the female applicants and the right histogram shows the income for the male applicants. In addition, code snippet for generating the figure is attached below the histograms. It is observed that the female applicants income has a normal distribution while the male applicants income is having a positive skewed distribution. Based on the histogram, majority of the income of female applicants are between 2400 to 3600 while majority of income for male applicants are between 2700 to 3300. This shows on average that the female applicants have a higher income than the male applicants. Therefore, the hypothesis is invalidated.

Hypothesis 3: Applicants with higher education level would have a higher income.

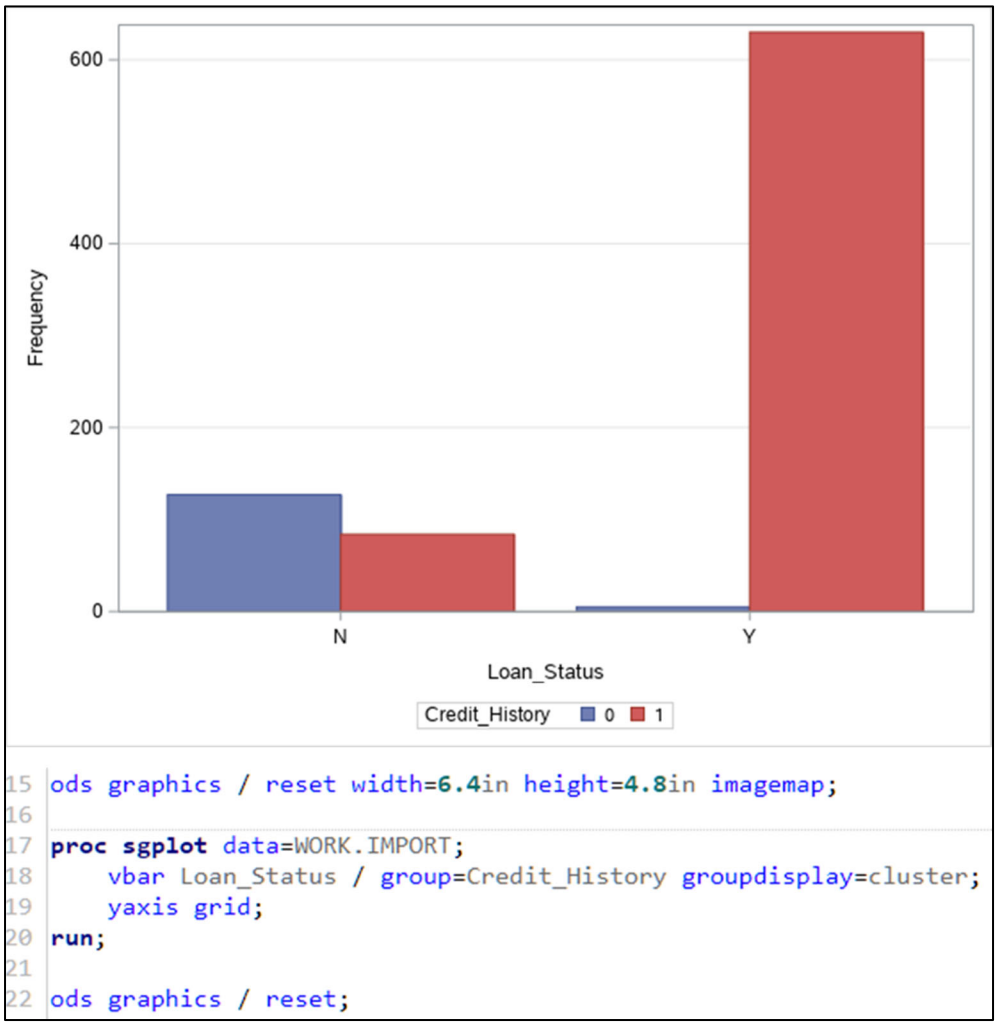



```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sort data=WORK.IMPORT2 out=_HistogramTaskData;
18     by Education;
19 run;
20
21 proc sgplot data=_HistogramTaskData;
22     by Education;
23     histogram ApplicantIncome /;
24     yaxis grid;
25 run;
26
27 ods graphics / reset;
28
29 proc datasets library=WORK noprint;
30     delete _HistogramTaskData;
31     run;
```

Figure 5.16: Histogram for “ApplicantIncome” based on “Education”

Figure 5.16 shows the histogram for “ApplicantIncome” segmented by “Education” where the left histogram shows the income for applicants with high education level and the right histogram shows the income for applicants without high education level. In addition, code snippet for generating the figure is attached below the histograms. Is it observed that both the histograms have a positively skewed distribution. Based on the histograms, majority applicants with high education level have an income range between 2800 to 3600 while majority applicants without high education level have an income range between 2000 to 4000. This shows on average that applicants with higher education level have a higher income as compared to applicants without a high education level. Therefore, the hypothesis is validated.

Hypothesis 4: Applicants with good repayment history would have a higher loan approval rate.

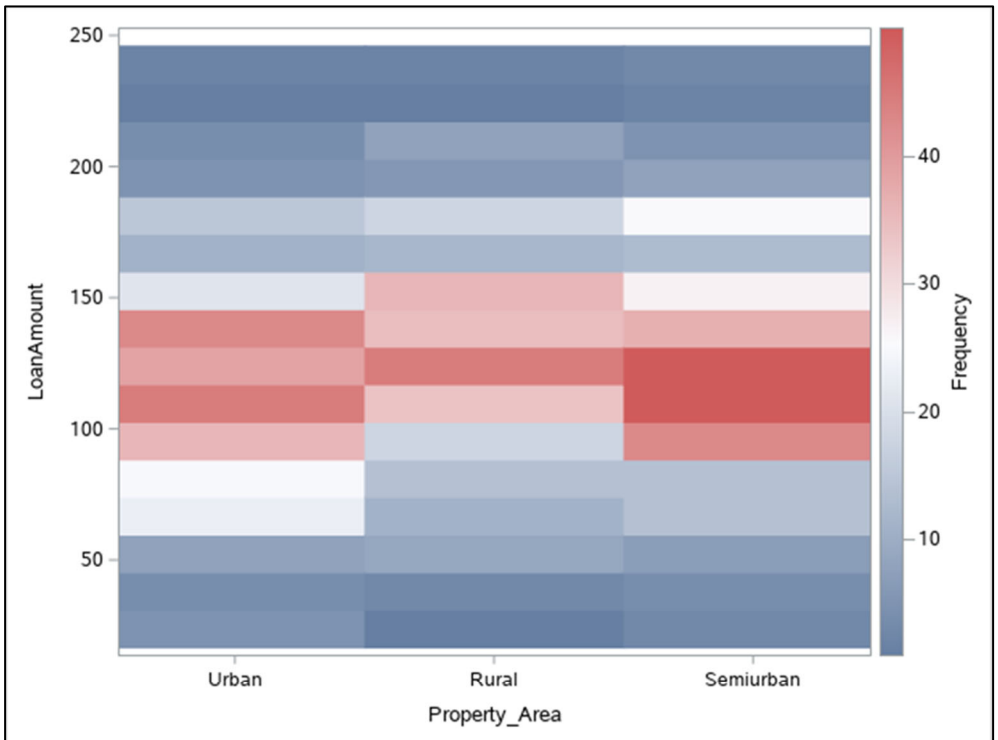



```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sgplot data=WORK.IMPORT;
18     vbar Loan_Status / group=Credit_History groupdisplay=cluster;
19     yaxis grid;
20 run;
21
22 ods graphics / reset;
```

Figure 5.17: Bar chart for “Loan_Status” based on “Credit_History”

Figure 5.17 shows the bar chart for “Loan_Status” clustered by “Credit_History” where the blue bar indicates bad historical repayment performance while the red bar indicates good historical repayment performance. In addition, code snippet for generating the figure is attached below the bar chart. It is observed that applicants with good historical repayment performance generally have approved loan application while applicants with bad historical repayment performance generally have rejected loan application. Although some applicants with good historical repayment performance is observed to have their loan rejected may be due to other reasons. This shows that applicants with good historical repayment performance have a higher loan approval rate. Therefore, the hypothesis is validated.

Hypothesis 5: Housing price in the urban area is higher than the semiurban and rural area.

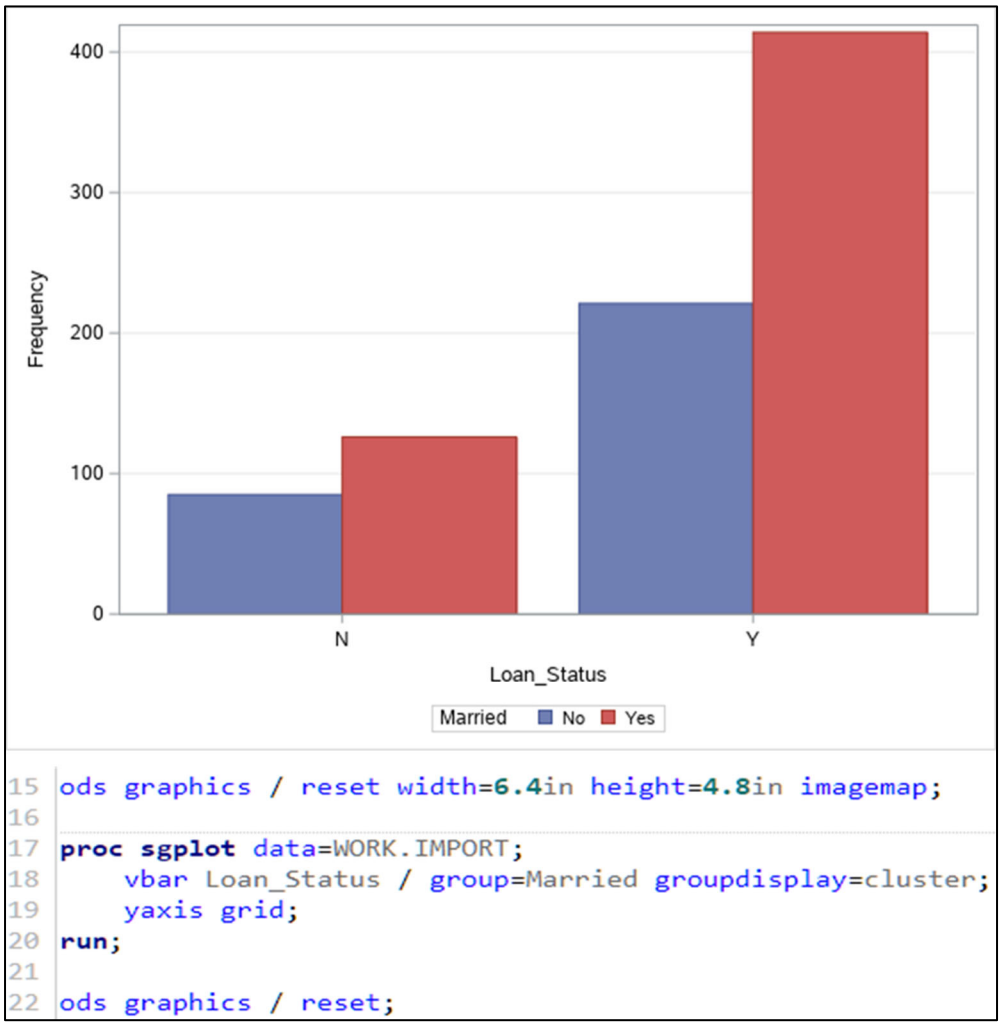



```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sgplot data=WORK.IMPORT;
18     heatmap x=Property_Area y=LoanAmount / name='HeatMap';
19     gradlegend 'HeatMap';
20 run;
21
22 ods graphics / reset;
```

Figure 5.18: Heat map for “LoanAmount” based on “Property_Area”

Figure 5.18 shows the heat map for “LoanAmount” categorized by “Property_Area”. In addition, the code snippet for generating the figure is attached below the heat map. It is observed that applicants are applying higher housing loan for rural area as compared to urban and semiurban area. Typically, housing prices in urban area is higher. However, in this scenario the rural area property might be having a higher housing price as compared to the urban and semiurban area. This may be due to the reason of buying land mass in rural area for other purposes thus a higher amount of loan borrowing is observed in rural area. Therefore, the hypothesis is invalidated.

Hypothesis 6: Applicants who are married, have a higher loan approval rate.




```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sgplot data=WORK.IMPORT;
18     vbar Loan_Status / group=Married groupdisplay=cluster;
19     yaxis grid;
20 run;
21
22 ods graphics / reset;
```

Figure 5.19: Bar chart for “Loan_Status” based on “Married”

Figure 5.19 shows the bar chart for “Loan_Status” grouped by “Married” where the blue bar indicates applicants are not married while the red bar indicates applicants are married. In addition, the code snippet for generating the figure is attached below the bar chart. It is observed that applicants who are married has a higher loan approval rate compared to applicants who are not married. However, quite a number of applicants who are yet to marry also got their loan approved. This can indicate there are other factors affecting the loan approval process. While in average, loan approval for married applicants are two times more than applicants who are not married. Therefore, the hypothesis is validated.

SECTION 6

CONCLUSION

Loan approval process is tedious and time consuming which require the processing of voluminous data supplied by the applicants. The banks have to filter out unfavorable applicants with the potential to default on future loans. However, loan dataset typically contains missing values and outliers which requires processing prior to be used in any predictive model. Therefore, this study performed the appropriate data pre-processing methods on a loan dataset to achieve a more usable dataset for the predictive model. In addition, EDA is performed to identify relationships and trends among the variables.

Missing value imputation and removal of outliers were performed as part of the data pre-processing step. Missing values were imputed using the mode and median of the variables. In addition, utilizing a logistic regression model to predict the missing value of the loan status variable as it contains high number of missing values. The imputation of missing values would help in retaining information without the need to drop any data which is valuable for smaller dataset. Furthermore, outliers are identified using box plots and removed from the dataset. This would minimize the data skew and lead to a better prediction model developed.

Relationships between variables are identified using EDA to validate the formulated hypotheses. Of the six hypotheses formulated, two is invalidated by the results from EDA. Therefore, the EDA provides a valuable tool to identify insights from the dataset. In addition, graphical representations from EDA provide an easy and quick understanding of the relationship from the data.

Further research can explore other data pre-processing methods to further enhance the usability of the dataset in developing predictive models. In addition, exploring other methods in EDA may lead to the discovery of better insights which can be used to better gauge the loan approval rate.

REFERENCES

Ambika, & Biradar, S. (2021). Survey on Prediction of Loan Approval Using Machine Learning Techniques. *International Journal of Advanced Research in Science, Communication and Technology, 5*(1), 449-454. doi:10.48175/IJARSCT-1165

Ashwini S. Kadam, Shraddha R. Nikam, Ankita A. Aher, Gayatri V. Shelke, & Chandgude, A. S. (2021). Prediction for Loan Approval using Machine Learning Algorithm. *International Research Journal of Engineering and Technology (IRJET), 8*(4), 4089-4092.

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications, 128*, 301-315. doi:https://doi.org/10.1016/j.eswa.2019.02.033

Blessie, E. C., & Rekha, R. (2019). Exploring the machine learning algorithm for prediction the loan sanctioning process. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9*(1).

Chen, X., Liu, Z., Zhong, M., Liu, X., & Song, P. (2019). *A Deep Learning Approach Using DeepGBM for Credit Assessment*. Paper presented at the Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence, Shanghai, China. https://doi-org.ezproxy.apu.edu.my/10.1145/3366194.3366333

Fati, S. M. (2021). Machine Learning-Based Prediction Model for Loan Status Approval. *Jorunal of Hunan University (Natural Sciences), 48*(10).

Gupta, K., Chakrabarti, B., Ansari, A. A., Rautaray, S. S., & Pandey, M. (2021). *Loanification-Loan Approval Classification using Machine Learning Algorithms.* Paper presented at the Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021.

L. Udaya Bhanu, & Narayana, D. S. (2021). Customer Loan Prediction Using Supervised Learning Technique. *International Journal of Scientific and Research Publications, 11*(6). doi: 10.29322/IJSRP.11.06.2021.p11453

Munoz, J., Rezaei, A. A., Jalili, M., & Tafakori, L. (2021). Deep learning based bi-level approach for proactive loan prospecting. *Expert Systems with Applications, 185*, 115607. doi:https://doi.org/10.1016/j.eswa.2021.115607

Tripathi, D., Edla, D. R., Kuppili, V., & Bablani, A. (2020). Evolutionary Extreme Learning Machine with novel activation function for credit scoring. *Engineering Applications of Artificial Intelligence, 96*, 103980. doi:https://doi.org/10.1016/j.engappai.2020.103980

Wang, T., Liu, R., & Qi, G. (2022). Multi-classification assessment of bank personal credit risk based on multi-source information fusion. *Expert Systems with Applications, 191*, 116236. doi:https://doi.org/10.1016/j.eswa.2021.116236

Xia, Y., Zhao, J., He, L., Li, Y., & Niu, M. (2020). A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications, 159*, 113615. doi:https://doi.org/10.1016/j.eswa.2020.113615

Zhang, H., Shi, Y., Yang, X., & Zhou, R. (2021). A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance. *Research in International Business and Finance, 58*, 101482. doi:https://doi.org/10.1016/j.ribaf.2021.101482

Zhang, W., Xu, W., Hao, H., & Zhu, D. (2020). Cost-sensitive multiple-instance learning method with dynamic transactional data for personal credit scoring. *Expert Systems with Applications, 157*, 113489. doi:https://doi.org/10.1016/j.eswa.2020.113489

Zhang, W., Yan, S., Li, J., Tian, X., & Yoshida, T. (2022). Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E: Logistics and Transportation Review, 158*, 102611. doi:https://doi.org/10.1016/j.tre.2022.102611

APPENDIX A

CODE SNIPPETS

1.1 Missing value identification after completion of missing value imputation for categorical variables.

```
ods noproctitle;

proc format;
    value _nmissprint low-high="Non-missing";
    value $_cmissprint " "="" " other="Non-missing";
run;

proc freq data=WORK.COMPLETE_IMPUTATION;
    title3 "Missing Data Frequencies";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format Loan_Amount_Term Credit_History _nmissprint.;
    format Gender Married Dependents Education Self_Employed Property_Area
        Loan_Status $_cmissprint.;
    tables Gender Married Dependents Education Self_Employed Loan_Amount_Term
        Credit_History Property_Area Loan_Status / missing nocum;
run;

proc freq data=WORK.COMPLETE_IMPUTATION noprint;
    table Gender * Married * Dependents * Education * Self_Employed *
        Loan_Amount_Term * Credit_History * Property_Area * Loan_Status / missing
        out=Work._MissingData_;
    format Loan_Amount_Term Credit_History _nmissprint.;
    format Gender Married Dependents Education Self_Employed Property_Area
        Loan_Status $_cmissprint.;
run;

proc print data=Work._MissingData_ noobs label;
    title3 "Missing Data Patterns across Variables";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format Loan_Amount_Term Credit_History _nmissprint.;
    format Gender Married Dependents Education Self_Employed Property_Area
        Loan_Status $_cmissprint.;
    label count="Frequency" percent="Percent";
run;

title3;

/* Clean up */
proc delete data=Work._MissingData_;
run;
```

1.2 Missing value identification after completion of missing value imputation for numerical variables.

```
ods noproctitle;

proc format;
    value _nmissprint low-high="Non-missing";
run;

proc freq data=WORK.COMPLETE_IMPUTATION;
    title3 "Missing Data Frequencies";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format ApplicantIncome CoapplicantIncome LoanAmount _nmissprint.;
    tables ApplicantIncome CoapplicantIncome LoanAmount / missing nocum;
run;

proc freq data=WORK.COMPLETE_IMPUTATION noprint;
    table ApplicantIncome * CoapplicantIncome * LoanAmount / missing
        out=Work._MissingData_;
    format ApplicantIncome CoapplicantIncome LoanAmount _nmissprint.;
run;

proc print data=Work._MissingData_ noobs label;
    title3 "Missing Data Patterns across Variables";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format ApplicantIncome CoapplicantIncome LoanAmount _nmissprint.;
    label count="Frequency" percent="Percent";
run;

title3;

/* Clean up */
proc delete data=Work._MissingData_;
run;
```