Table of Contents

INTRODUCTION

The loan application process is a critical aspect of the financial industry, and understanding the factors that influence loan approvals is of utmost importance. In this project, we aim to perform a comprehensive statistical analysis of a loan dataset to gain insights into the characteristics and patterns of loan applicants. By applying various statistical tests and techniques, we will explore the relationships between different variables and identify significant predictors of loan approvals. The primary problem addressed in this project is to understand the factors that contribute to loan approvals. By exploring the relationships between various applicant attributes and loan outcomes, we aim to identify significant predictors and uncover any potential biases or trends in the data.

The dataset used in this project contains information about loan applicants, including their personal and financial attributes. The dataset obtained was from Kaggle titled Analytics Vidhya Loan Prediction (Bora, 2019). It includes variables such as *Loan_ID, Gender, Marital Status, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area,* and *Loan_Status.* Each row represents an individual applicant, providing a rich source of data for analysis.

Our primary objective is to perform an inference statistical analysis on the loan dataset to address several key aspects. Firstly, we will conduct exploratory data analysis to gain a comprehensive understanding of the dataset's structure, summary statistics, and data distributions. This analysis will help us identify any missing values, outliers, or data quality issues that may impact our subsequent analysis. Next, we will delve into hypothesis testing to determine if there are significant differences or relationships between specific variables of interest. We will formulate and test relevant hypotheses using 1-sample tests to evaluate the significance of observed differences. This will enable us to assess the impact of various factors on loan approvals.

Furthermore, we will employ the chi-square goodness of fit test to examine whether the observed frequencies of categorical variables adhere to expected distributions. By identifying significant deviations from expected patterns, we can detect potential biases or underlying trends in the dataset. To explore the association between different categorical variables and loan approvals, we will perform the chi-square test of independence. This test will allow us to

determine if certain personal or financial attributes significantly influence the likelihood of loan approvals and evaluate the strength of these associations.

Additionally, we will conduct correlation analysis to measure the strength and direction of linear relationships between numerical variables. By identifying variables that exhibit significant correlations, we can gain insights into potential predictors for loan approvals. Finally, we will utilise regression models to predict the loan amount based on relevant independent variables. Through this analysis, we will identify significant predictors that impact loan approvals and assess their relative importance.

By conducting this statistical analysis, we aim to uncover valuable insights and patterns within the loan dataset. The significance of these tests lies in their ability to provide evidence-based conclusions regarding the relationships between variables and their impact on loan approvals. The outcomes of this project can assist financial institutions in making informed decisions, improving risk assessment, and enhancing the loan approval process. The statistical analysis of the loan dataset allows us to gain a deeper understanding of the characteristics and patterns of loan applicants. Through the application of exploratory data analysis, hypothesis testing, goodness of fit tests, chi-square tests of independence, correlation analysis, and regression analysis, we aim to uncover significant predictors and patterns that influence loan approvals. The insights obtained from this project can help financial institutions make more informed lending decisions, improve risk assessment, and ultimately enhance the loan application process.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) of loan data involves the examination and exploration of various variables that provide crucial insights into loan applications. The loan dataset consists of essential variables such as Loan_ID, Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area, and Loan_Status. By performing EDA on these variables, we can uncover patterns, trends, relationships, and potential insights that drive loan approval decisions. Through visualizations, summary statistics, and statistical analyses, EDA helps in

gaining deeper insights into the loan data, identifying key factors influencing loan approval, and developing informed strategies for risk assessment and decision-making.

Datatype

```
> head(loandf)
   Loan_ID Gender Married Dependents     Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount
1 LP001002   Male      No          0      Graduate            No            5849                0         NA
2 LP001003   Male     Yes          1      Graduate            No            4583             1508        128
3 LP001005   Male     Yes          0      Graduate           Yes            3000                0         66
4 LP001006   Male     Yes          0 Not Graduate            No            2583             2358        120
5 LP001008   Male      No          0      Graduate            No            6000                0        141
6 LP001011   Male     Yes          2      Graduate           Yes            5417             4196        267
  Loan_Amount_Term Credit_History Property_Area Loan_Status
1              360              1         Urban           Y
2              360              1         Rural           N
3              360              1         Urban           Y
4              360              1         Urban           Y
5              360              1         Urban           Y
6              360              1         Urban           Y
```

Figure 1 Dataframe of The Dataset

The "head()" function displays the first few rows of the dataframe, providing a quick overview of the data and its structure. In this case, it shows the first six rows of the "loandf" dataframe. Each row in the displayed data represents a specific loan application, and the columns provide different details about each application.

```
> str(loandf)
'data.frame':   614 obs. of  13 variables:
 $ Loan_ID          : chr  "LP001002" "LP001003" "LP001005" "LP001006" ...
 $ Gender           : chr  "Male" "Male" "Male" "Male" ...
 $ Married          : chr  "No" "Yes" "Yes" "Yes" ...
 $ Dependents       : chr  "0" "1" "0" "0" ...
 $ Education        : chr  "Graduate" "Graduate" "Graduate" "Not Graduate" ...
 $ Self_Employed    : chr  "No" "No" "Yes" "No" ...
 $ ApplicantIncome  : int  5849 4583 3000 2583 6000 5417 2333 3036 4006 12841 ...
 $ CoapplicantIncome: num  0 1508 0 2358 0 ...
 $ LoanAmount       : int  NA 128 66 120 141 267 95 158 168 349 ...
 $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 360 360 ...
 $ Credit_History   : int  1 1 1 1 1 1 1 0 1 1 ...
 $ Property_Area    : chr  "Urban" "Rural" "Urban" "Urban" ...
 $ Loan_Status      : chr  "Y" "N" "Y" "Y" ...
```

Figure 2 Structure of The Dataframe

The "str(loandf)" command provides an overview of the structure of the dataframe and the data types of its variables. Based on the output, The dataframe has 614 observations (rows) and 13 variables (columns). Each row represents a different loan application, while each column represents a specific attribute or characteristic of the loan application. The variable "Loan_ID" is of character (chr) type, meaning it stores alphanumeric identifiers for each loan application. The variables "Gender", "Married", "Dependents", "Education", "Self_Employed",

"Property_Area", and "Loan_Status" are also of character (chr) type.These variables capture categorical information such as the gender of the applicant, marital status, number of dependents, education level, self-employment status, property area, and loan status (approved or not). The variable "ApplicantIncome" is of integer (int) type, representing the income of the loan applicants. This variable store whole numbers. The variable "CoapplicantIncome" is of numeric (num) type, indicating the income of any co-applicants associated with the loan application. The numeric data type can represent both whole numbers and decimals.

The variables "LoanAmount" and "Loan_Amount_Term" are both of integer (int) type. "LoanAmount" represents the amount of loan requested by the applicants, and "Loan_Amount_Term" indicates the duration or term of the loan. These variables store whole numbers. However, based on observation, "Loan_Amount_Term" should be stored as a character (chr) type instead of an integer (int) type. This is because the values for "Loan_Amount_Term" are all the same and do not represent distinct numeric values. Storing it as a character type will preserve the original format and avoid potential misinterpretation of the data. The variable "Credit_History" is of integer (int) type, storing binary values (0 or 1) to indicate whether the loan applicants have a credit history or not. Since the variable "Credit_History" represents binary data (0 or 1), it would be more appropriate to store it as a binary variable.

```
#changing Y and N in Loan_status to yes and no
loandf$Loan_Status <- ifelse(loandf$Loan_Status == "Y", "Yes", "No")

#changing Loan_Amount_Term from Int to Categorical
loandf$Loan_Amount_Term <- as.character(loandf$Loan_Amount_Term)

#change datatype for credit history from int to categorical
loandf$Credit_History <- as.character(loandf$Credit_History)
```

Figure 3 Data Modifying

Using the above code, we modify the values in the "Loan_Status" column. It checks each value in the column and if the value is "Y" (indicating loan approval), it assigns the corresponding row with "Yes". If the value is "N" (indicating loan rejection), it assigns "No" to that row. This code effectively replaces the original "Y" and "N" values with "Yes" and "No" respectively, making the meaning more explicit and easier to interpret. For "Loan_Amount_Term", by converting the column to character type, the numerical values representing the loan term or duration are now treated as text. This change is useful when the column contains categorical or non-numeric data, allowing for better handling and analysis. The same steps were applied to

column "Credit_History" since the column contains binary data with values of 0 and 1 representing the absence or presence of a credit history, converting it to character type allows for better representation and enables treating the values as text rather than numerical values.

Missing and Duplicating Values

```
> # Replace empty strings with NA in all columns of the dataframe 'loandf'
> loandf[loandf == ""] <- NA
> # Check missing value for all
> sum(is.null(loandf)| is.na(loandf))  # Check missing value in a data frame or column
[1] 149
> # Count the number of missing values in each column
> missing_count <- colSums(is.null(loandf) | is.na(loandf))
> print(missing_count)
         Loan_ID            Gender           Married        Dependents         Education     Self_Employed
               0                13                 3                15                 0                32
  ApplicantIncome CoapplicantIncome        LoanAmount  Loan_Amount_Term    Credit_History     Property_Area
               0                 0                22                14                50                 0
     Loan_Status
               0
```

Figure 4 Missing Values in Dataset

The first step to check for missing value, the empty string "" was replaced with NA to standardise all missing values across the dataframe. Upon checking the missing values across all columns, there are a total of 149 missing values in this data set with the highest number of missing values being 'Credit_History' with 50 occurrences. Other columns with missing values included 'Self_Employed' (32 missing values), 'LoanAmount' (22 missing values), 'Dependents' (15 missing values), 'Loan_Amount_Term' (14 missing values), 'Gender' (13 missing values), and 'Married' (3 missing values). 149 of missing values contribute to almost 25% of the data hence it is not significant to remove all the missing values. Instead, all the missing values from categorical data will be replaced with the mode value, while for numerical data, will be replaced with median value.

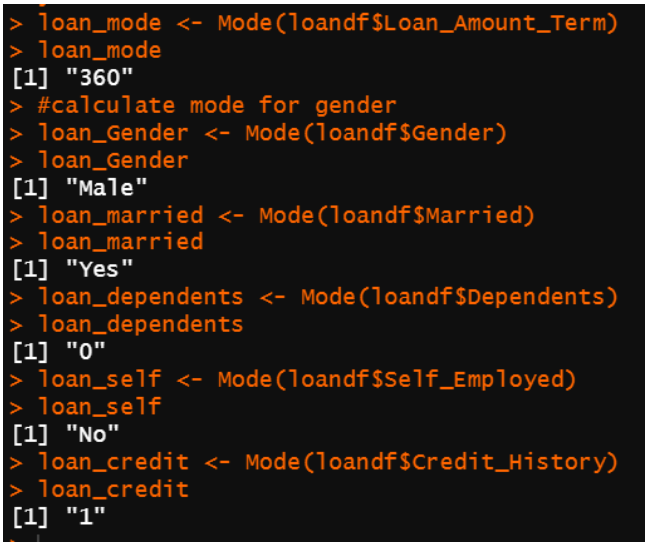


Figure 5 Calculating Mode Value For Each Categorical Variables

The modes calculated for the categorical variables in the 'loandf' dataframe can be used to fill in the missing values for each respective column. For the 'Loan_Amount_Term' column, the missing values can be replaced with the mode value of "360". Similarly, for the remaining categorical columns, the missing values can be filled as follows: 'Gender' with the mode value of "Male", 'Married' with "Yes", 'Dependents' with "0", 'Self_Employed' with "No", and 'Credit_History' with "1". By utilizing the modes, which represent the most common categories, the missing values are replaced with values that align with the dominant patterns observed in the dataset. For numerical data, the only column that has missing value is 'LoanAmount'. The data for 'LoanAmount' is not normally distributed hence the missing value is replaced with the median value.

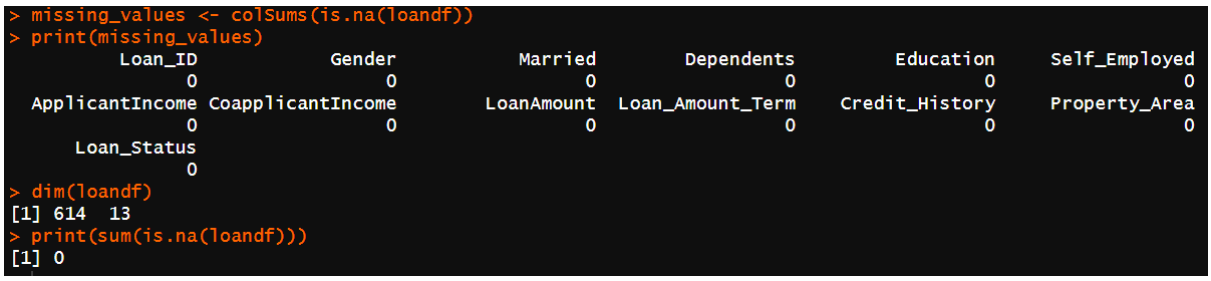


Figure 6 Checking Missing Values After Replacing With Respective Mode And Median Values

To reconfirm that all missing values have been successfully replaced, we can refer to the information provided. The variable "missing_values' shows that for each column in the 'loandf'

dataframe, the count of missing values is zero. This implies that no missing values exist in any of the columns. Additionally, the dimensions of the dataframe, as given by the "dim(loandf)" code, confirm that there are 614 rows and 13 columns of data the same as given dimension in the original dataset.

```
> # Check duplicated value
> duplicates <- duplicated(loandf)
> sum(duplicates) # no duplicated value
[1] 0
>
> #data summary
> summary(loandf)
   Loan_ID            Gender            Married           Dependents         Education         Self_Employed
 Length:614         Length:614         Length:614         Length:614         Length:614         Length:614
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


 ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term   Credit_History     Property_Area
 Min.    :  150   Min.   :    0     Min.   :  9.0   Length:614         Length:614         Length:614
 1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.2   Class :character   Class :character   Class :character
 Median : 3812   Median : 1188     Median :128.0   Mode  :character   Mode  :character   Mode  :character
 Mean   : 5403   Mean   : 1621     Mean   :145.8
 3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:164.8
 Max.   :81000   Max.   :41667     Max.   :700.0
 Loan_Status
 Length:614
 Class :character
 Mode  :character
```

Figure 7 Checking For Duplicating Values

Upon checking for duplicating values, this dataset shows that there are no duplicated rows or entries. Hence, there are no additional steps needed to handle the duplicating values. The summary statistics provide an overview of all the variables in the 'loandf' dataframe. This extra step to verify all data was stored in the desired data type also to check for the numerical statistics for each numeric column.

Outliers and Boxplot

```
num_outliers
ApplicantIncome CoapplicantIncome        LoanAmount
             16                 11                25
```

Figure 8 Number of Outliers in Every Numerical Variables

Upon checking the numbers of outliers, there are 16 outliers in the 'ApplicantIncome' variable, 11 outliers in the 'CoapplicantIncome' variable, and 25 outliers in the 'LoanAmount' variable. Outliers are data points that deviate significantly from the rest of the observations and may have a substantial impact on statistical analysis or modelling results. Therefore, it is important

to carefully evaluate the nature and potential impact of these outliers. To further evaluate the outliers, a boxplot for each numerical variable was plotted.

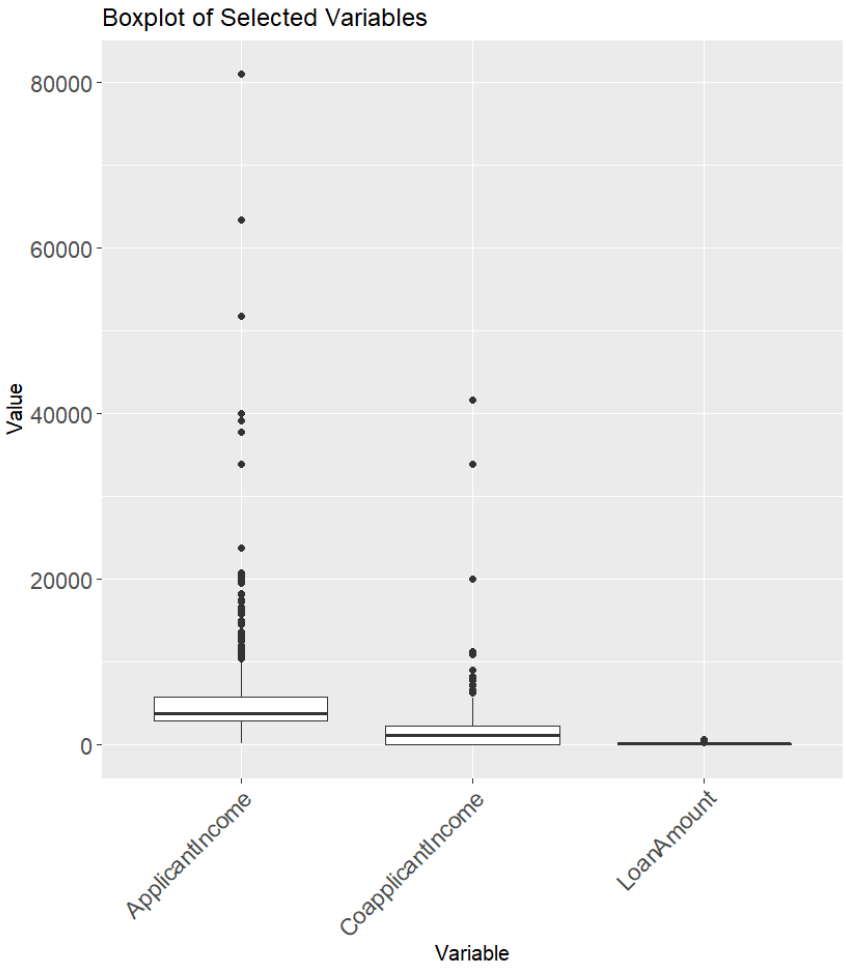
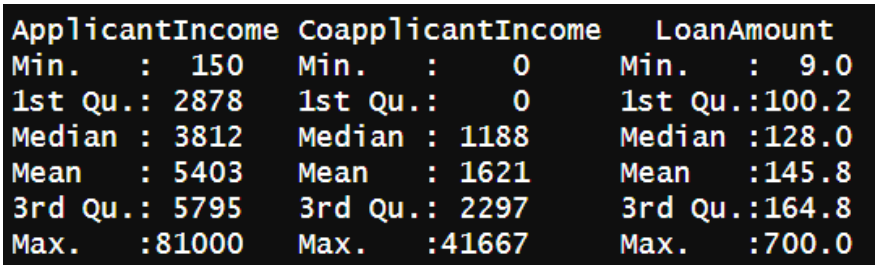


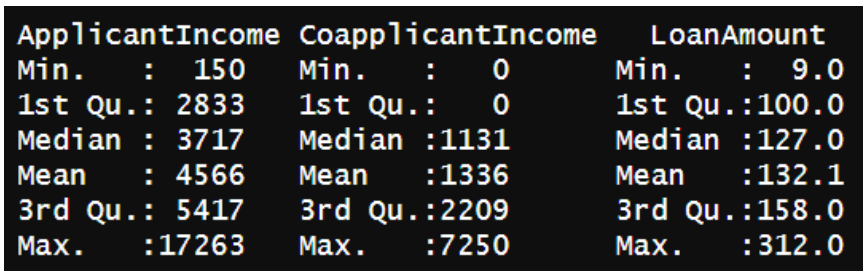
Figure 9 Boxplot Before Removing Outliers

Upon evaluation, the numbers of outliers can skew statistical measures and modelling results, leading to biassed conclusions. Hence it needs to be removed in order to obtain a more representative and robust dataset that better captures the underlying patterns and relationships in the data.



ApplicantIncome	CoapplicantIncome	LoanAmount
Min. : 150	Min. : 0	Min. : 9.0
1st Qu.: 2878	1st Qu.: 0	1st Qu.:100.2
Median : 3812	Median : 1188	Median :128.0
Mean : 5403	Mean : 1621	Mean :145.8
3rd Qu.: 5795	3rd Qu.: 2297	3rd Qu.:164.8
Max. :81000	Max. :41667	Max. :700.0

Figure 10 Summary of Data Statistical Values Before Removing Outliers



ApplicantIncome	CoapplicantIncome	LoanAmount
Min. : 150	Min. : 0	Min. : 9.0
1st Qu.: 2833	1st Qu.: 0	1st Qu.:100.0
Median : 3717	Median :1131	Median :127.0
Mean : 4566	Mean :1336	Mean :132.1
3rd Qu.: 5417	3rd Qu.:2209	3rd Qu.:158.0
Max. :17263	Max. :7250	Max. :312.0

Figure 11 Summary of Data Statistical Values After Removing Outliers

There are changes in the summary statistics that can be seen before and after removing the outliers such as the mean, median, and quartiles for the respective variables. The mean and median values may show variations, reflecting the influence of the outliers on the central tendency measures. Overall, removing outliers has resulted in a narrower range and potentially more representative summary statistics for the 'ApplicantIncome', 'CoapplicantIncome', and 'LoanAmount' variables.

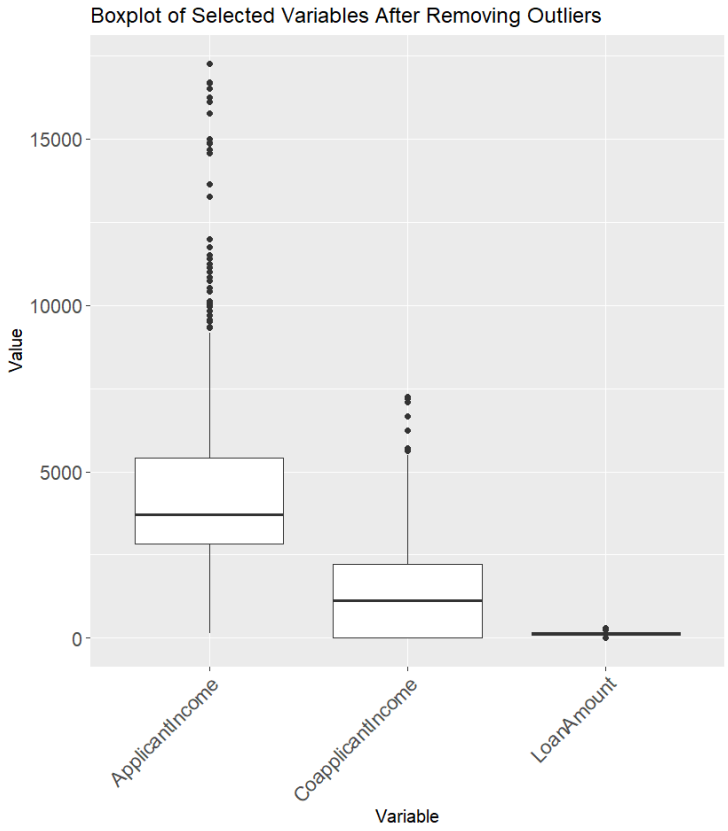


Figure 12 Boxplot After Removing Outliers

In the case of 'ApplicantIncome', 'CoapplicantIncome', and 'LoanAmount', the boxplot after removing outliers shows a more concentrated box with less extreme values, resulting in a more compact representation of the data distribution. This provides a clearer visual understanding of the central tendency and spread of the data, as well as any remaining potential outliers. After removing outliers, the boxplot for each variable shows a clearer representation of the distribution of the data.

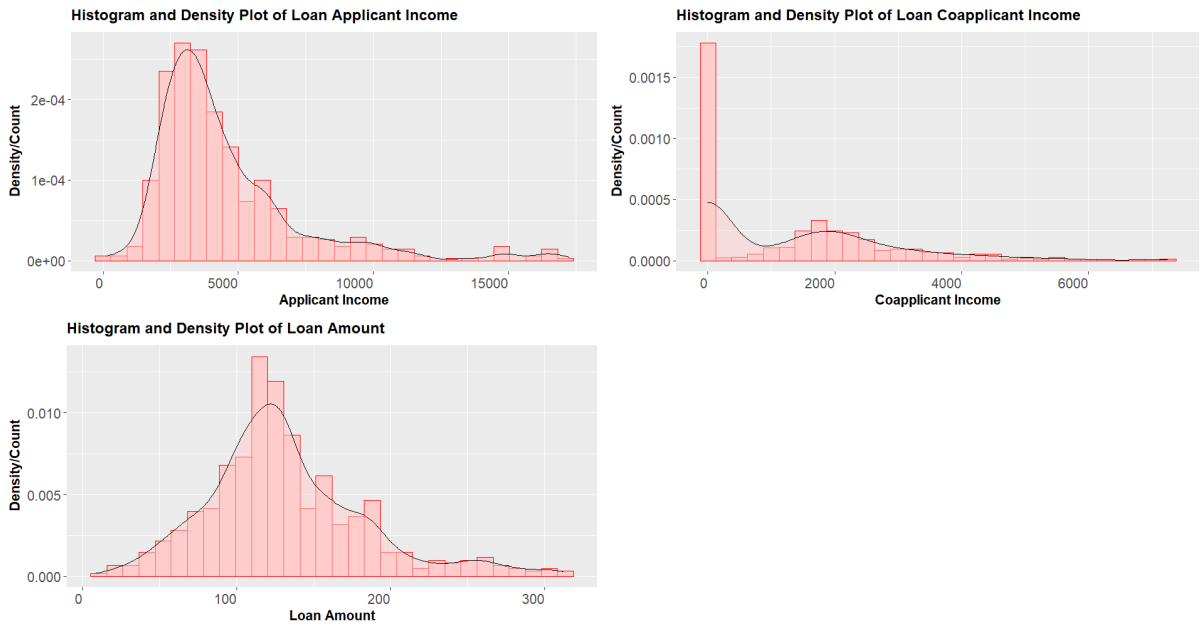


Figure 13 Combine Histogram And Density Plot For All Numerical Variables

The histogram and density plot provide valuable insights into the distributions of the variables ApplicantIncome, CoapplicantIncome, and LoanAmount. For the ApplicantIncome variable, the histogram shows that the majority of applicants fall within the income range of approximately 150 to 5500. The distribution is right-skewed, indicating that a significant portion of applicants have lower incomes, while a smaller proportion have higher incomes. The density plot further confirms this pattern, showing a smooth estimate of the distribution with a peak around the 2000-3000 range. In terms of CoapplicantIncome, the histogram reveals that many applicants have either zero or very low coapplicant income, as seen by the tall bar at the leftmost side. There is also a noticeable peak in the 1000-2000 range, suggesting a significant number of applicants have coapplicants with incomes within that range. The distribution is right-skewed, with a long tail extending towards higher coapplicant income values. The density plot aligns with this observation, presenting a smooth estimate of the distribution that reflects the skewed nature of the data.

Moving on to LoanAmount, the histogram demonstrates a relatively symmetrical distribution, with a peak around the 100-150 range. This indicates that a considerable number of applicants are requesting loan amounts falling within this range. The distribution tapers off towards both lower and higher loan amounts. The density plot supports this finding, presenting a smooth estimate of the distribution that confirms the symmetrical nature of the data. Overall, the histogram and density plot analyses shed light on the distributions of ApplicantIncome,

CoapplicantIncome, and LoanAmount. They reveal important insights about the ranges, concentrations, and shapes of these variables, allowing for a deeper understanding of the data and its characteristics. These visualisations aid in identifying patterns and trends, which can be valuable for making informed decisions and drawing meaningful conclusions from the dataset.

Categorical Variables Analysis

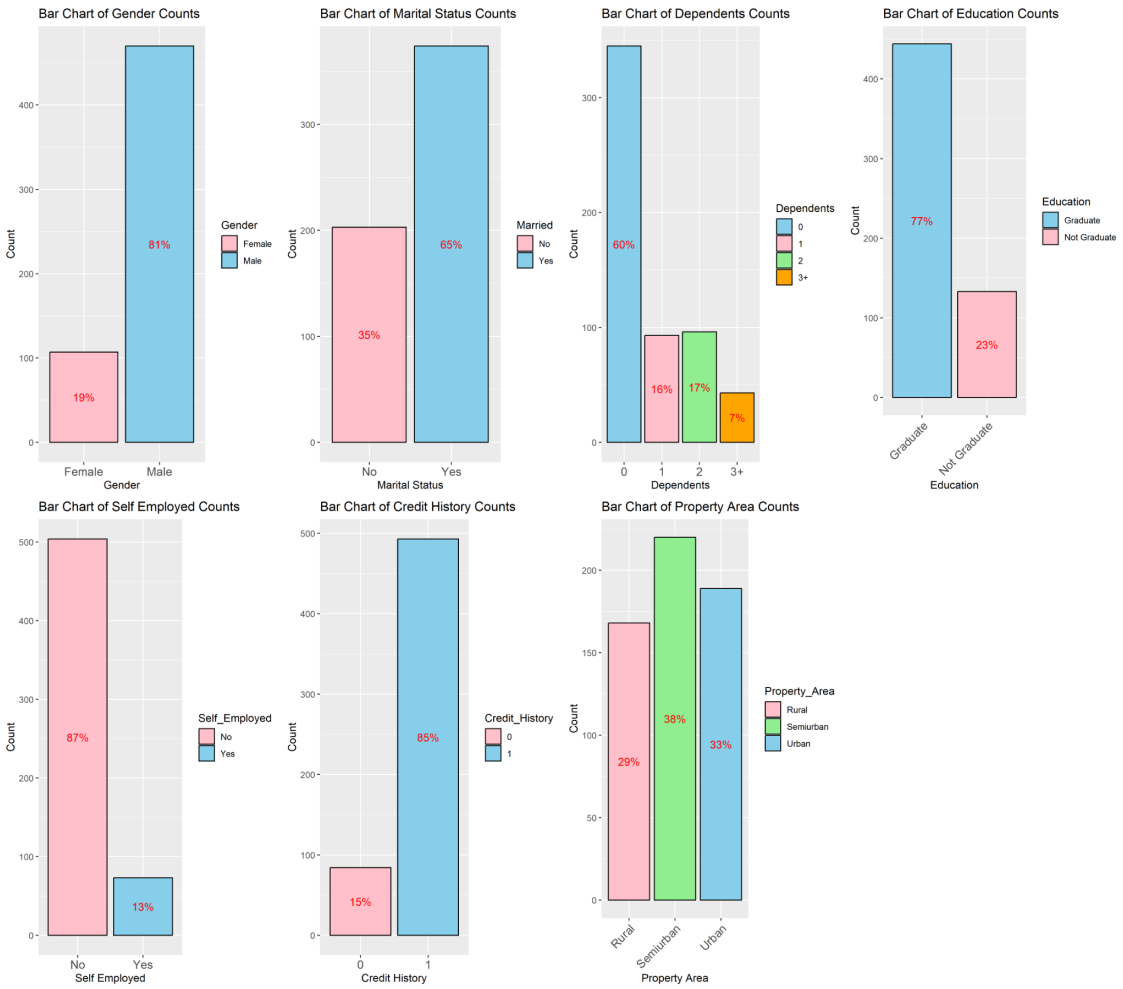


Figure 14 Combine Bar Chart For All Categorical Variables

The figure above shows a bar chart of various categorical variables in the dataset. A comprehensive analysis of the chart reveals significant insights about the loan applicants. From the figure, it is evident that a substantial majority of the applicants, approximately 81%, is male, while the remaining 19% are female. For marital status, around 65% of the applicants are married, indicating a significant proportion of individuals seeking loans have entered into

matrimony, whereas the remaining applicants are yet to tie the knot. When examining the number of dependents, it becomes apparent that the majority, constituting 60% of the applicants, have no dependents. In contrast, a smaller proportion, accounting for only 7%, indicates having three or more dependents, signifying the presence of larger families within the dataset. Analysing the education category reveals that approximately 77% of the applicants have graduated, indicating a higher level of educational attainment among loan seekers. Conversely, around 23% of the applicants are yet to graduate, suggesting a diverse educational background within the dataset.

In terms of self-employment, a striking 87% of the applicants respond negatively, indicating that they are not self-employed, while the remaining applicants express self-employment as their occupation. Considering credit history, an overwhelming 85% of the loan applicants have a documented credit history, signifying a previous borrowing experience or financial transactions that have been recorded. Conversely, the remaining 15% lack any record of credit history, potentially indicating a lack of prior engagement with any formal financial systems. Regarding the preferred property area, a majority of loan applicants, the highest percentage being in the range of 85%, seek property in semi-urban areas. Comparatively, a lower percentage of applicants express interest in rural areas, representing the least preferred location among the dataset.

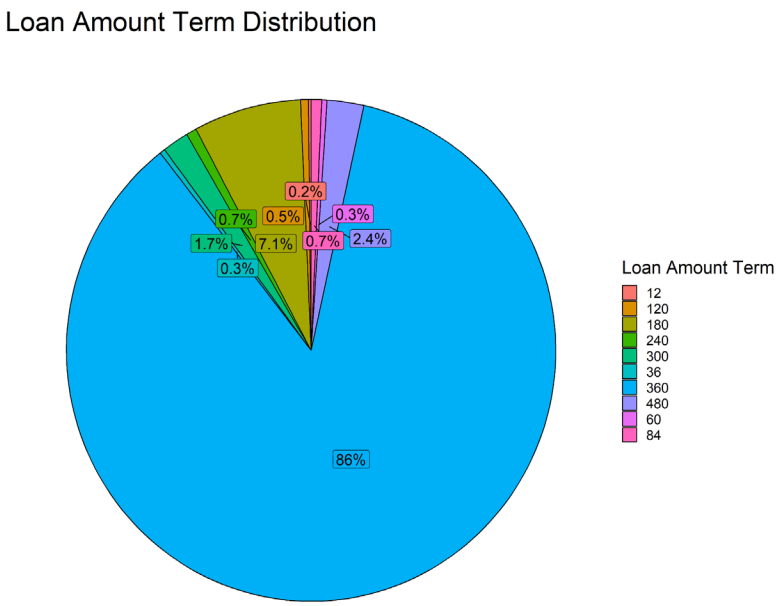


Figure 15 Pie Chart For Loan Amount Term Distribution

The pie chart represents the distribution of loan amount terms for loan applicants in the dataset. The loan amount term refers to the duration, in months, over which the loan is scheduled to be repaid. The chart displays 10 different loan amount terms: 12, 120, 180, 240, 300, 36, 360, 480, 60, and 84 months. The dominant loan amount term is 360 months, accounting for the majority with 86% of the loan applicants. This suggests that a significant proportion of the applicants opt for a longer-term repayment plan. On the other hand, the loan amount term of 12 months represents a minimal percentage, only 0.2%, indicating that very few applicants select this short-term option. The remaining loan amount terms, including 120, 180, 240, 300, 36, 480, 60, and 84 months, make up the remaining portion of the pie chart. Each of these terms accounts for a relatively smaller percentage, ranging from 0.3% to 7.1%. These figures demonstrate that loan applicants have diverse preferences when it comes to choosing the duration of their loan repayment plan.

Descriptive Analysis

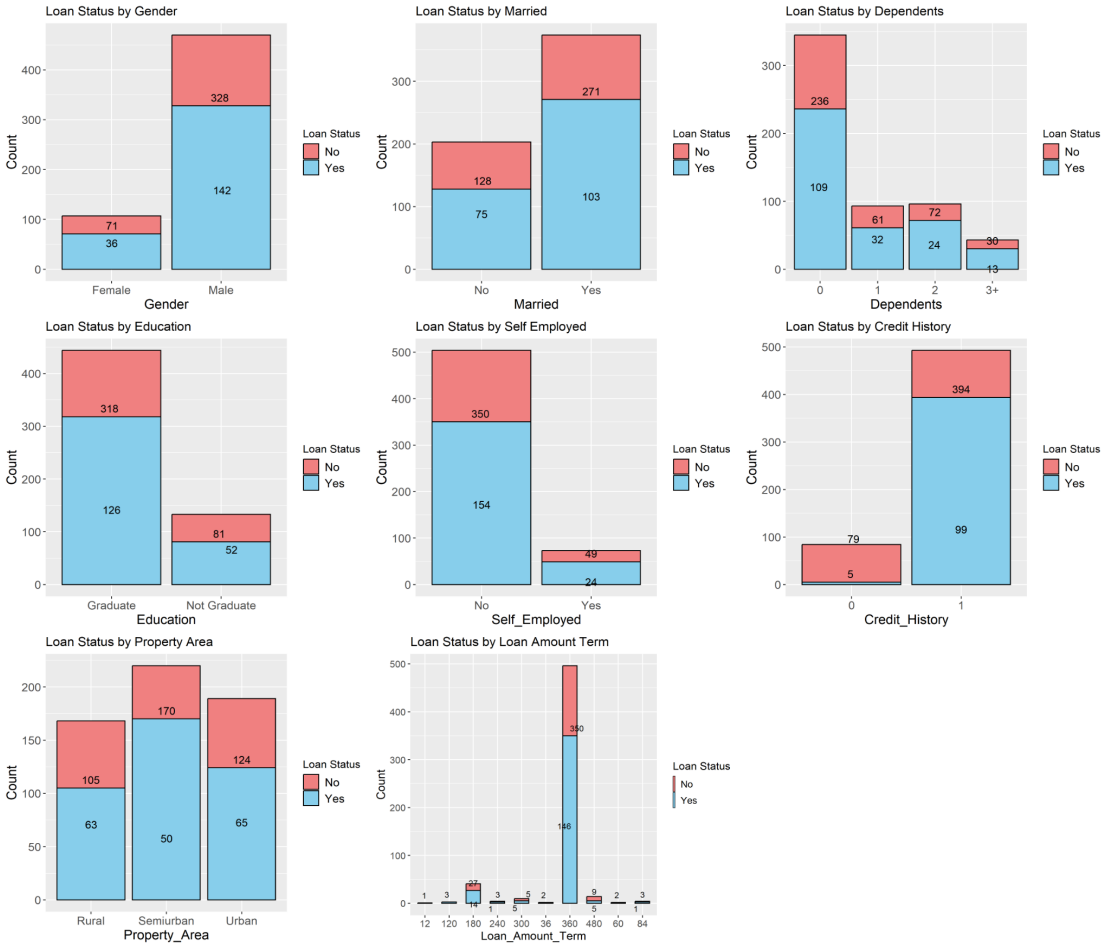


Figure 16 Combine Bar Chart For Categorical Variables Group By Loan Status

The combination bar chart illustrates the relationship between the categorical variables and the loan status of applicants. The first bar chart represents the gender distribution categorised by loan status. Among female applicants, 36 individuals have been approved for loans, while 71 females have had their loan applications denied. On the other hand, among male applicants, 142 individuals have been approved for loans, while 328 males have faced loan rejections. This chart indicates that a higher number of males have applied for loans compared to females, and the approval rates for males are also higher. The second bar chart examines the loan status based on the marital status of the applicants. Among those who are not married, 128 individuals have successfully obtained loans, while 75 individuals have faced loan rejections. In the married category, 103 individuals have been approved for loans, while 271 individuals have not been successful. This chart indicates that individuals who are married have a higher number of loan applications, but their approval rates are comparatively lower than those who are not married. The third bar chart focuses on the number of dependents and their loan status. Among applicants with no dependents, 109 individuals have been approved for loans, while 236 have faced loan rejections. For applicants with one dependent, 32 individuals have received loan approvals, and 61 have faced rejections. For those with two dependents, 24 individuals have been approved, while 72 have faced rejections. Finally, for applicants with three or more dependents, 13 individuals have been approved for loans, while 30 have been denied. This chart demonstrates that the number of dependents can influence loan approval, with higher numbers of dependents correlating with lower approval rates.

The bar chart for education and loan status reveals insightful patterns. Among graduates, 126 individuals have been approved for loans, while 318 graduates have faced loan rejections. Similarly, for non-graduates, 52 individuals have received loan approvals, while 81 individuals have experienced loan denials. This chart indicates that a higher number of loan applicants are graduates, but they also face a significant number of loan rejections compared to non-graduates. For self-employment variables, among individuals who are not self-employed, 154 individuals have been approved for loans, while 350 individuals have faced loan rejections. On the other hand, among self-employed individuals, 24 individuals have received loan approvals, while 49 individuals have faced loan denials. This chart suggests that non-self-employed applicants have higher loan application rates and a relatively higher number of loan rejections compared to self-employed individuals. For applicants with a credit history of 0, only 5 individuals have been approved for loans, while 79 individuals have faced loan rejections. In contrast, among applicants with a credit history of 1, 99 individuals have received loan approvals, and 394

individuals have encountered loan denials. This chart demonstrates the critical impact of credit history on loan approval, with a higher likelihood of approval for individuals with a credit history of 1.

Among applicants residing in rural areas, 63 individuals have been approved for loans, while 105 individuals have faced loan rejections. In the semiurban category, 50 individuals have received loan approvals, while 170 individuals have experienced loan denials. For urban applicants, 65 individuals have been approved for loans, while 124 individuals have encountered loan rejections. This chart highlights that loan approval rates vary across different property area statuses, with urban areas having a higher approval rate compared to semiurban and rural areas. The bar chart for the loan amount term and loan status demonstrates important trends. For a loan amount term of 180 months, 14 individuals have received loan approvals, while 27 individuals have faced loan rejections. Similarly, for a loan amount term of 300 months, 5 individuals have been approved for loans, while 5 individuals have been denied. In the case of a loan amount term of 360 months, 146 individuals have obtained loan approvals, and 350 individuals have been denied. Lastly, for a loan amount term of 480 months, 5 individuals have been approved, while 9 individuals have faced loan rejections. This chart reveals that loan approval rates differ based on the loan amount term, with a higher likelihood of approval for terms of 180 and 360 months.

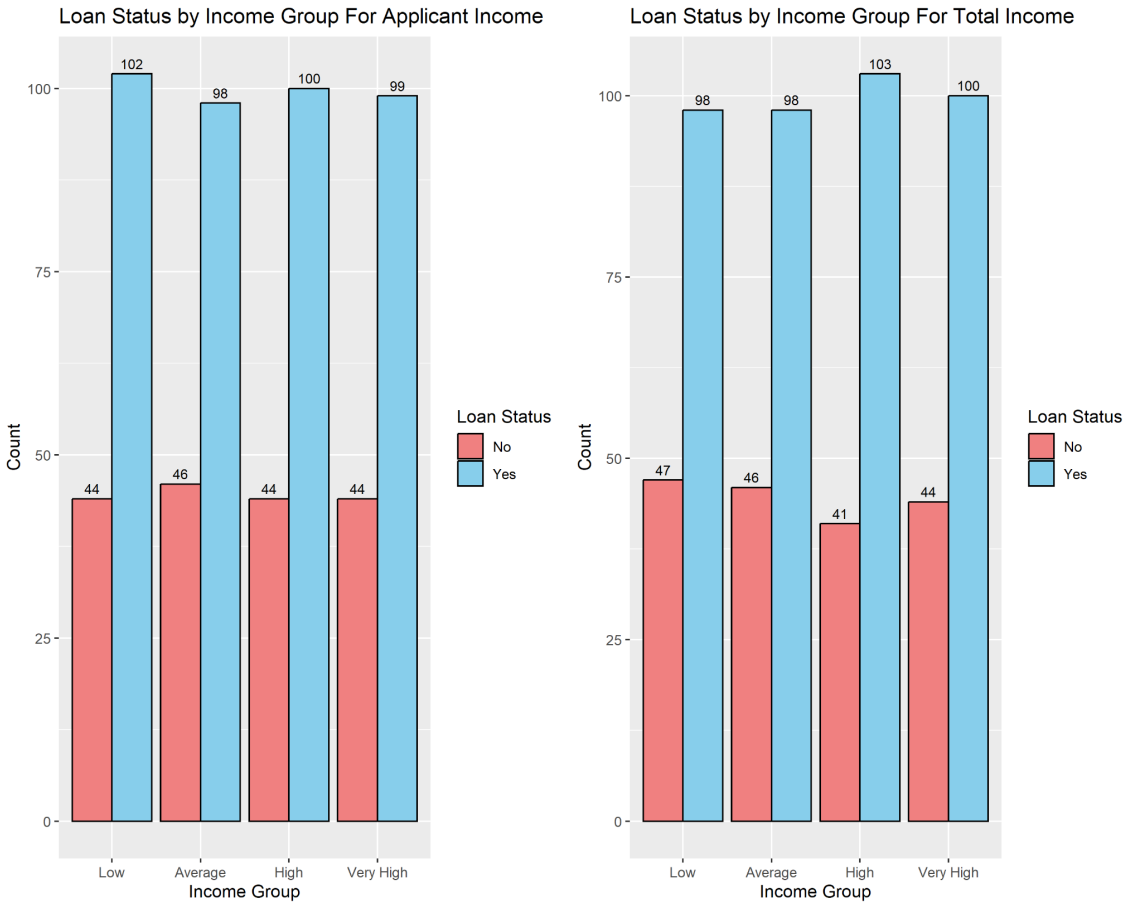


Figure 17 Combine Bar Chart For Numerical Variables Group By Loan Status

The first bar chart examines the relationship between loan status and income group for applicant income. Among individuals classified as low income, 102 have been approved for loans, while 44 have faced loan rejections. In the average income category, 98 individuals have received loan approvals, and 46 have encountered loan denials. For high-income applicants, 100 individuals have been approved for loans, while 44 have been denied. Among those with very high incomes, 99 individuals have received loan approvals, and 44 have faced loan rejections. This chart indicates that loan approval rates vary across different income groups, with higher approval rates for individuals in the average and high-income categories.

The second bar chart analyses loan status by income group for total income, which combines both the applicant's income and co applicant's income. Among individuals categorised as low income, 98 individuals have been approved for loans, while 47 have faced loan rejections. In the average income group, 98 individuals have received loan approvals, and 46 have been denied. For high-income applicants, 103 individuals have been approved for loans, while 41 have faced loan rejections. Among those with very high incomes, 100 individuals have received

loan approvals, and 44 have encountered loan denials. This chart reveals that loan approval rates remain consistent across income groups when considering the total income of the applicants.

Both bar charts demonstrate the impact of income on loan approval rates. They reveal that higher-income individuals generally have a higher likelihood of loan approval, with average and high-income groups consistently experiencing higher approval rates compared to the low-income group. These insights highlight the significance of income as a crucial factor in the loan approval process and its influence on individuals' borrowing capabilities.

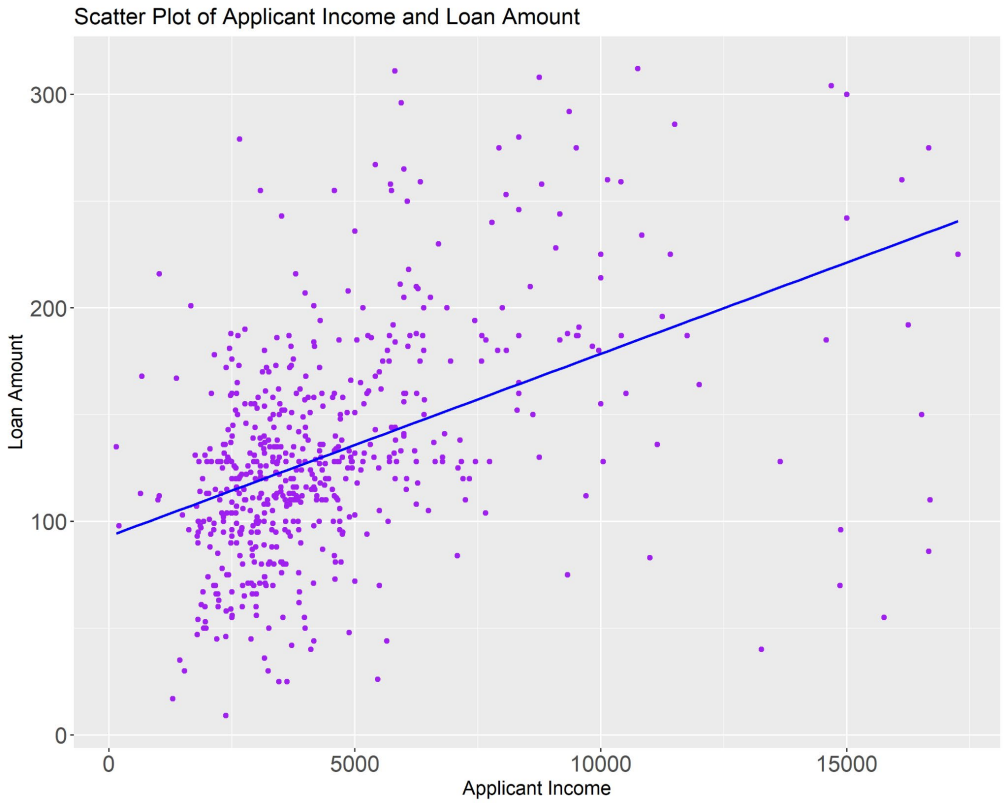


Figure 18 Scatter Plot of Applicant Income and Loan Amount

From the scatter plot of loan amount and applicant income, we can observe a positive relationship between the two variables. As the applicant income increases, the loan amount also tends to increase. This indicates that individuals with higher incomes are more likely to apply for larger loan amounts. Additionally, the scatter plot reveals a concentration of data points in the range of applicant income from 1000 to 5000 and loan amount from 50 to 200. This concentration suggests that a significant number of loan applicants fall within this income

and loan amount range. The positive relationship and concentration of data points in the scatter plot indicate that there is a tendency for individuals with higher incomes to apply for larger loan amounts. However, it is important to note that other factors such as credit history, loan eligibility criteria, and borrower preferences may also influence the loan amount decision. Understanding the relationship between applicant income and loan amount can help financial institutions and lenders assess risk, determine loan eligibility, and make informed decisions regarding loan approvals and terms.

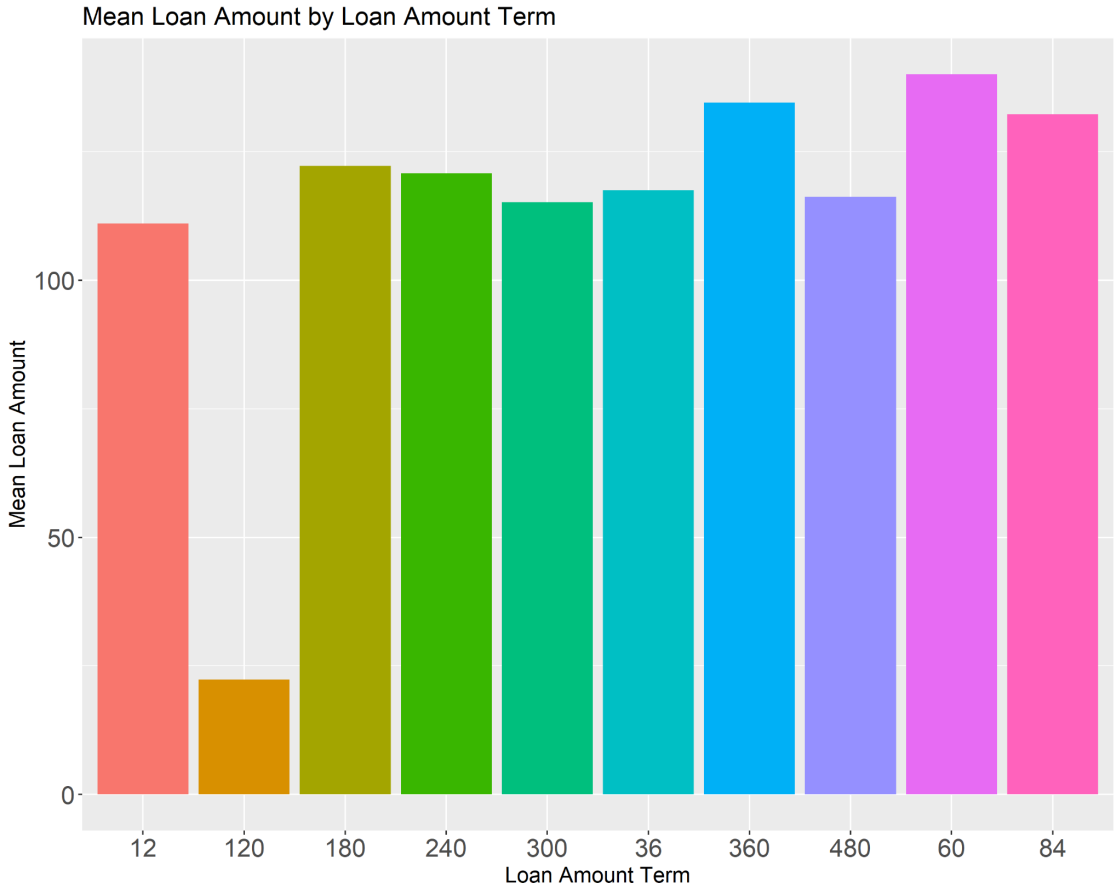


Figure 19 Bar Chart For The Mean Of Loan Amount Group By Loan Amount Term

The bar chart displays the mean loan amount grouped by the loan amount term. It provides insights into the average loan amounts for different loan durations. For a loan amount term of 12 months, the mean loan amount is 115. Similarly, for a term of 120 months, the mean loan amount is 23. The loan amount term of 180 months has a mean loan amount of 123, while the term of 240 months has a mean loan amount of 121. A term of 300 months corresponds to a mean loan amount of 117, and for a term of 36 months, the mean loan amount is 119. The highest loan amount term in the dataset, 360 months, has a relatively higher mean loan amount

of 130. The loan amount term of 480 months has a mean loan amount of 118, while the terms of 60 and 84 months have mean loan amounts of 140 and 127, respectively. This bar chart allows for a visual comparison of the mean loan amounts across different loan amount terms. It demonstrates the variation in average loan amounts based on the duration of the loan.

HYPOTHESIS TESTING (1 SAMPLE)

Hypothesis testing plays a crucial role in statistical analysis by providing a systematic framework to assess the validity of assumptions and draw meaningful conclusions. One key aspect of hypothesis testing is the evaluation of normality, as many statistical tests assume normally distributed data. In this essay, we explore the process of evaluating normality and applying transformations in hypothesis testing using a dataset of loan data.

To assess the normality of the data, the Shapiro-Wilk test was conducted, and the obtained p-values were examined. The Shapiro-Wilk test is a statistical test used to assess whether a given dataset follows a normal distribution. It is based on the null hypothesis that the population from which the dataset is drawn follows a normal distribution. The Shapiro-Wilk test calculates a test statistic that measures the discrepancy between the observed data and the expected values under the assumption of normality. The test produces a p-value, which represents the probability of obtaining the observed data (or more extreme) if the null hypothesis of normality is true. If the p-value is greater than a chosen significance level (commonly 0.05), we fail to reject the null hypothesis and conclude that the data is reasonably consistent with a normal distribution. On the other hand, if the p-value is less than the significance level, we reject the null hypothesis and conclude that there is evidence to suggest that the data does not follow a normal distribution.

```
Variable: ApplicantIncome
Shapiro-Wilk p-value: 3.945229e-26
This variable is not normal with p = 3.945229e-26

Variable: CoapplicantIncome
Shapiro-Wilk p-value: 3.470039e-24
This variable is not normal with p = 3.470039e-24

Variable: LoanAmount
Shapiro-Wilk p-value: 1.790253e-12
This variable is not normal with p = 1.790253e-12
```

Figure 20 Shapiro test result of *ApplicantIncome, CoapplicantIncome* and *LoanAmount.*

In our analysis, we conducted the Shapiro-Wilk test on three numerical variables: *ApplicantIncome, CoapplicantIncome*, and *LoanAmount.* The results revealed extremely low p-values for all variables, indicating that the data is not normally distributed. Specifically, the p-values were 3.945229e-26 for *ApplicantIncome*, 3.470039e-24 for *CoapplicantIncome,* and 1.790253e-12 for *LoanAmount*. These results provide evidence to reject the assumption of normality in the original dataset. The reason for the variables not being normally distributed could be due to several factors. For example, *ApplicantIncome* and *CoapplicantIncome* might be skewed by a few high-income or low-income individuals, leading to a departure from normality. *LoanAmount* could also be influenced by certain factors such as loan policies or borrower characteristics, causing deviations from a normal distribution.

To address the violation of normality assumption, transformation techniques were applied to the numerical variables. Two common transformations used in such scenarios are the square root transformation and the log transformation. These transformations aim to achieve a more symmetric and normally distributed dataset, which can subsequently improve the validity of hypothesis testing. For each numerical variable, we performed both the square root and log transformations. The square root transformation was applied using the `sqrt()` function, while the log transformation was applied using the `log()` function. It is worth noting that a small constant (1) was added to handle zero values in the log transformation.

After applying the transformations, the normality of the transformed data was evaluated using the Shapiro-Wilk test. The Shapiro-Wilk test was conducted on the square root transformed data and the log transformed data for each variable. The resulting p-values were then compared to determine if the transformations successfully improved normality. The difference of Shapiro test for both log and square root test can be seen based on the Figure 2.2 below.

```
Variable: ApplicantIncome
Square Root Transformation - Shapiro-Wilk p-value: 3.299242e-17
Log Transformation - Shapiro-Wilk p-value: 1.573684e-12
Variable: CoapplicantIncome
Square Root Transformation - Shapiro-Wilk p-value: 7.984009e-25
Log Transformation - Shapiro-Wilk p-value: 7.76652e-31
Variable: LoanAmount
Square Root Transformation - Shapiro-Wilk p-value: 1.784973e-05
Log Transformation - Shapiro-Wilk p-value: 2.737515e-13
```

Figure 21 Shapiro-Wilk Test Result After Transformation

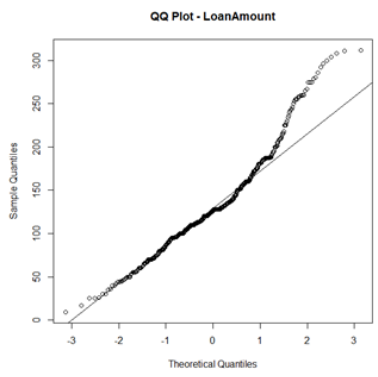
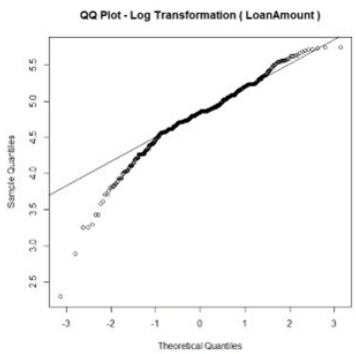
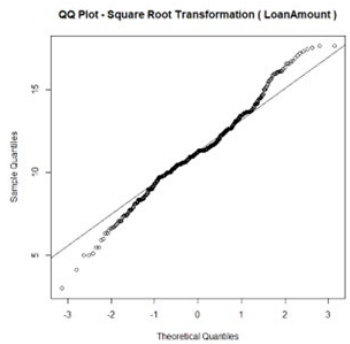
The square root transformation yielded improved normality for all three variables. The Shapiro-Wilk test p-values for the square root transformed data were significantly higher than the p-

values for the original data, indicating a better approximation to normality. Similarly, the log transformation also resulted in improved normality for all variables, with higher Shapiro-Wilk test p-values compared to the original data.

For *ApplicantIncome*, both the square root and log transformations resulted in improved normality compared to the original data. The Shapiro-Wilk p-values for the square root transformed data and log transformed data were 3.299242e-17 and 1.573684e-12, respectively. However, the log transformation yielded the highest p-value, indicating the best improvement in normality for this variable.

Similarly, for *CoapplicantIncome*, both the square root and log transformations improved normality. The Shapiro-Wilk p-values for the square root transformed data and log transformed data were 7.984009e-25 and 7.76652e-31, respectively. Once again, the log transformation exhibited the highest p-value, indicating the most substantial improvement in normality.

Finally, for *LoanAmount*, both the square root and log transformations improved normality compared to the original data. The Shapiro-Wilk p-values for the square root transformed data and log transformed data were 1.784973e-05 and 2.737515e-13, respectively. Here as well, the log transformation yielded the highest p-value, indicating the most significant improvement in normality. We can see for all three variables; the log transformation shows the best improvement in achieving normality based on the Shapiro-Wilk test results.

Before Transformation	Log transformation	Square Root Transformation
		

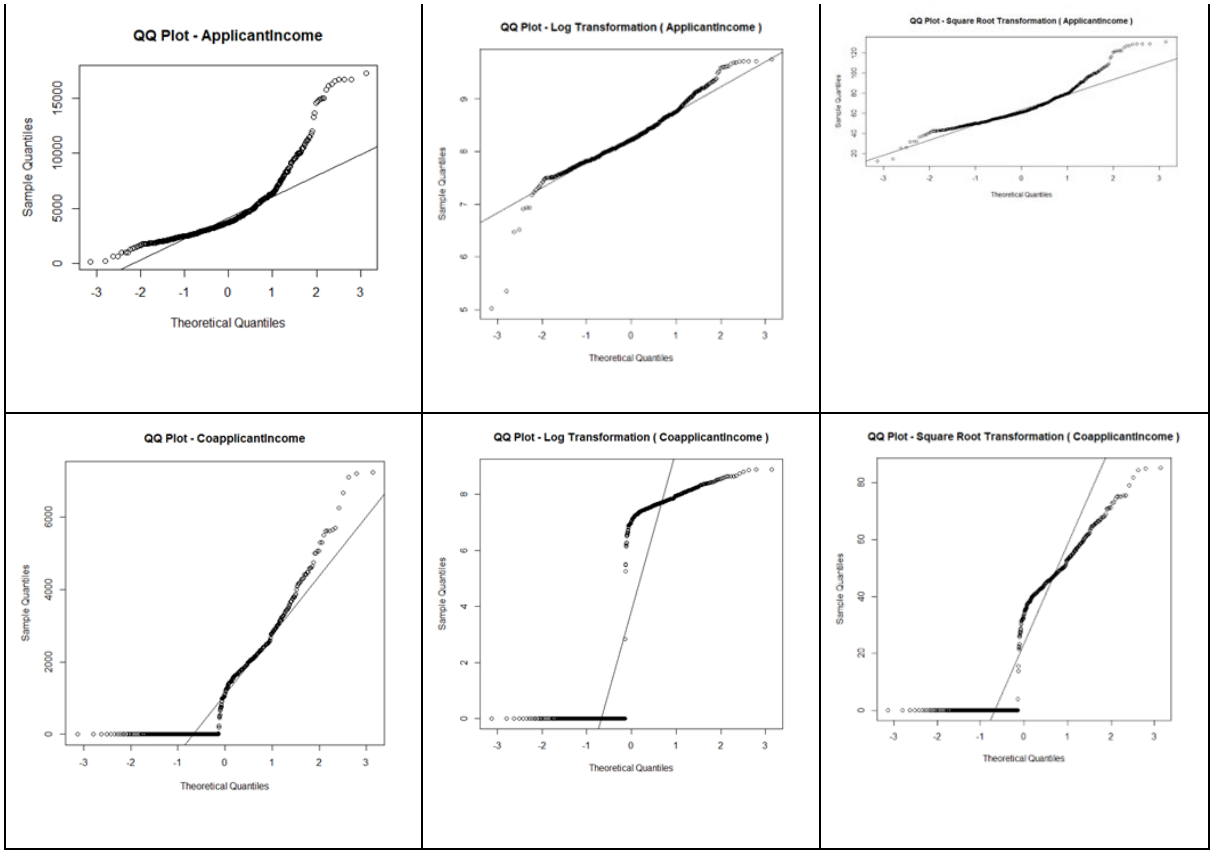


Table 1 QQ plot for all three Shapiro-Wilk test starting from before transformation, Log Transformation and Square Root Transformation.

The QQ plot, short for quantile-quantile plot, is a graphical tool used to assess the similarity between the observed data and a theoretical distribution, typically the normal distribution. The main function of a QQ plot is to visually examine whether the data follows a particular distribution or if it deviates from it. In a QQ plot, the observed data values are plotted against the corresponding quantiles of the theoretical distribution. If the data points fall approximately along a straight line, it suggests that the data is well approximated by the theoretical distribution. On the other hand, if the data points deviate from the straight line, it indicates a departure from the assumed distribution.

QQ plots are particularly useful for evaluating the normality assumption for hypothesis testing. If the QQ plot for a dataset shows a linear pattern, it suggests that the data can be reasonably assumed to follow a normal distribution. Conversely, if the QQ plot shows significant deviations from a straight line, it indicates departures from normality, which may require data transformation or the use of non-parametric tests.

In the case of *CoapplicantIncome*, the original data contains 0 values, it's possible that the Shapiro-Wilk test is detecting this departure from normality. The transformation methods applied (such as log transformation or square root transformation) can help address this issue by shifting the distribution and reducing the impact of the 0 values on normality assumptions. Overall, the presence of 0 values in the *CoapplicantIncome* column can certainly be a factor contributing to the departure from normality.

Therefore, evaluating normality and applying transformations are critical steps in hypothesis testing when dealing with non-normally distributed data. Our analysis of the loan dataset demonstrated the violation of normality assumption in the original variables. However, by applying the square root and log transformations, we were able to achieve improved normality.

Since the transformed data has shown improved normality, t-test is used instead of the z-test. The t-test is more appropriate when the sample size is small or when the population standard deviation is unknown. The dataset is big and standard deviation can be calculated but it not normally distributed and even after transformation, the Shapiro test shown it is improved but the p-value is still far from normal distribution, hence due to the normality assumption is still not fully met even with the transformed data, it is generally safer to rely on the t-test, which is more robust to deviations from normality.

Hypothesis testing was performed using the one-sample t-test for mean to evaluate the mean of the data before and after transforming it to a normal distribution. This type of hypothesis testing is suitable for numerical data, as it allows us to determine if the mean of a single sample significantly differs from a hypothesized population mean. There are 3 numerical data from the dataset which are *ApplicantIncome, CoapplicantIncome* and *LoanAmount.*

The first hypothesis is to find out whether the average normalized *ApplicantIncome* significantly different from unnormalized sample mean? For null hypothesis (H0) is the average Applicant Income is not significantly different from the unnormalized sample mean. Meanwhile, alternative hypothesis (H1) is the average Applicant Income is significantly different from the unnormalized sample mean. Symbolically, it can be represented as:

H0: $\mu = \mu 0$

H1: $\mu \neq \mu 0$

Where:

- H0 represents the null hypothesis.
- H1 represents the alternative hypothesis.
- μ represents the population mean of the transformed Applicant Income.
- $\mu 0$ represents the hypothesized population mean.

The goal of the hypothesis test is to assess whether there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis, indicating a significant difference between the average Applicant Income and the unnormalized sample mean.

Alpha (α) is the significance level, which represents the predetermined threshold for accepting or rejecting the null hypothesis. It indicates the level of risk or probability of making a Type I error, which is the incorrect rejection of the null hypothesis when it is true. The alpha used is 0.05 which corresponds to a 5% significance level (a commonly used in many fields as a standard practice). This means that if the p-value obtained from the hypothesis test is less than 0.05, we would reject the null hypothesis and conclude that there is significant evidence to support the alternative hypothesis. Conversely, if the p-value is greater than or equal to 0.05, we would fail to reject the null hypothesis due to insufficient evidence.

```
result <- t.test(appincome_trans, mu = mu0)
```

Figure 22 R-code used to do t-test for *ApplicationIncome* data

The t.test() function is a one-line code that calculates the test statistic, degrees of freedom, and p-value for conducting a one-sample t-test, allowing us to evaluate whether the mean of a sample significantly differs from a hypothesized population mean. This function will be used for the following 1 sample hypothesis testing. The results of the one-sample t-test fo*r ApplicantIncome* are as follows:

- Sample mean: 8.277816

- Hypothesized population mean: 4566.17

- Test statistic: -201236.7

- Degrees of freedom: 576

- p-value: 0

With a significance level of 0.05, the p-value obtained from the test is less than the significance level. Therefore, we have significant evidence to reject the null hypothesis. This means that the Applicant Income is significantly different from the hypothesized population mean. In other words, based on the provided data, there is strong evidence to suggest that the average Applicant Income differs significantly from the hypothesized value of 4566.17.

Next hypothesis testing is for *LoanAmoun*t. The null and alternative are the same, and we want to see whether there is significant difference between the null and alternative hypothesis. All the alpha used in the hypothesis is 0.05.

H0: $\mu = \mu 0$

H1: $\mu \neq \mu 0$

The results of the one-sample t-test for Loan Amount are as follows:

- Sample mean: 4.799948

- Hypothesized population mean: 132.0919

- Test statistic: -7061.454

- Degrees of freedom: 576

- p-value: 0

In the case of the one-sample t-test for Loan Amount, the obtained p-value from the test is 0, which is less than the significance level of 0.05. This means that the probability of observing the obtained sample mean (4.799948) or a more extreme value, assuming the null hypothesis is true, is very low.

When the p-value is less than the significance level, it provides strong evidence to reject the null hypothesis. Therefore, we can confidently conclude that the Loan Amount is significantly different from the hypothesized population mean of 132.0919. This implies that, based on the given data, the average Loan Amount deviates significantly from the value of 132.0919 that was hypothesized as the population mean. In simpler terms, the results indicate that there is strong statistical evidence to support the claim that the average Loan Amount is significantly different from the hypothesized value. The difference observed in the data is unlikely to have occurred by chance, suggesting a meaningful distinction between the average Loan Amount and the hypothesized population mean.

The one-sample t-test was also conducted to investigate whether the average *TotalIncome* significantly differs from a hypothesized population mean. The significance level was set at 0.05, and the test was performed on the provided dataset. With similar hypothesis, we can write the hypothesis in the formula:

H0: $\mu = \mu 0$

H1: $\mu \neq \mu 0$

The results of the one-sample t-test for Total Income are as follows:

- Sample mean: 8.58697
- Hypothesized population mean: 5902.423
- Test statistic: -328623.1
- Degrees of freedom: 576
- p-value: 0

Using a significance level of 0.05, the obtained p-value from the test is less than the significance level. Therefore, there is significant evidence to reject the null hypothesis. This indicates that the Total Income is significantly different from the hypothesized population mean. In other words, based on the provided data, there is strong evidence to suggest that the average Total Income significantly differs from the hypothesized value of 5902.423.

In the context of the given data, the results indicate that all three variables (Applicant Income, Loan Amount, and Total Income) exhibit significant differences from their respective hypothesized population means. This implies that the average values of these variables, as observed in the sample, deviate significantly from the expected values. These findings have practical implications for various domains. For example, in the case of Applicant Income, it suggests that the average income of loan applicants is significantly different from the hypothesized value. This information can be used by financial institutions to better understand the income distribution of loan applicants and make informed decisions regarding loan approvals, interest rates, or eligibility criteria. Similarly, for Loan Amount and Total Income, the significant differences highlight variations in the loan amounts requested and the total income reported by individuals compared to the expected values. This knowledge can aid in making informed decisions related to loan disbursement, risk assessment, or financial planning.

Overall, the hypothesis results provide statistical evidence that helps in understanding the characteristics of the variables under investigation and informs decision-making processes in various fields. They contribute to a deeper understanding of the population from which the sample is drawn and can guide further analysis or actions based on the observed deviations.

GOODNESS FIT TEST

The chi-square goodness of fit test is a statistical test used to determine if the observed frequencies of categorical variables significantly differ from the expected frequencies. It is commonly employed when we want to examine whether the distribution of categories in a sample follows a particular theoretical distribution or when comparing observed frequencies to specified target proportions.

In this analysis, the chi-square goodness of fit test was applied to several categorical variables in the loan dataset. By conducting this test, we aimed to investigate if the observed frequencies of different categories within each variable deviated significantly from what would be expected under the assumption of no relationship or specific target proportions.

```
# Create a contingency table of observed frequencies
observed <- table(loandf_cleaned[[var]])

# Perform the chi-square goodness of fit test
result <- chisq.test(observed)
```

Figure 23 Code for doing Chi-Square goodness of fit test

From Figure 23, table() is used to create a contingency table. The table() function takes a vector or a combination of vectors as input and constructs a contingency table, which shows the count of each combination of values in the specified variables. The `chisq.test()` function calculates the chi-square test statistic, degrees of freedom, and p-value associated with the test. It compares the observed frequencies in the contingency table to the expected frequencies under the null hypothesis of independence or a specific target distribution.

The result of the chi-square test is stored in the `result` object, which contains information such as the test statistic, degrees of freedom, and p-value. These values are then used for further analysis or reporting. By using these two lines of code, a contingency table can be quickly generated, and p chi-square goodness of fit test be performed to assess the association or departure from the expected distribution for each categorical variable in the dataset.

The significance of the chi-square goodness of fit test lies in its ability to provide evidence of association or departure from the expected distribution. A small p-value indicates that there is strong evidence to reject the null hypothesis, suggesting that the observed frequencies are significantly different from the expected frequencies. On the other hand, a large p-value suggests that there is not enough evidence to reject the null hypothesis, indicating that the observed frequencies are consistent with the expected frequencies.

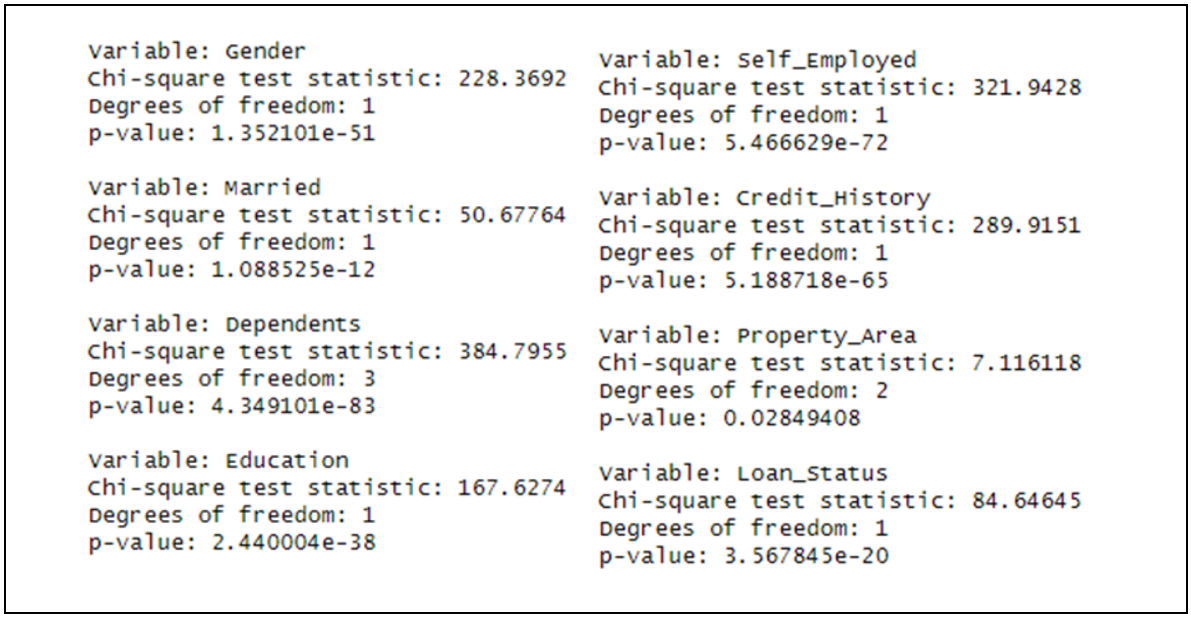


Figure 24 Chi-Square goodness of fit test results

Based on Figure 24, for gender the chi-square test for the Gender variable indicates a highly significant association between gender and the observed frequencies. The chi-square test statistic is a measure of the difference between the observed frequencies and the frequencies that would be expected if there were no association between gender and the categories. In this case, the high chi-square test statistic suggests a substantial difference between the observed frequencies and the expected frequencies under the null hypothesis. The degrees of freedom (df) indicate the number of categories minus 1. Since there are two categories (male and female) for gender, the df is 1. The extremely small p-value (1.352101e-51) suggests strong evidence against the null hypothesis of no association between gender and the observed frequencies. It indicates that the observed frequencies are highly unlikely to occur by chance alone if gender and the categories were not associated.

Figure 24's chi-square test for the Married variable shows a significant relationship between marital status and observed frequencies. The chi-square test statistic (50.67764) is relatively high, indicating that there is a significant difference between the observed and predicted frequencies under the null hypothesis of no connection. There are two marital status categories with one degree of freedom (married and not married). The low p-value (1.088525e-12) indicates that there is substantial evidence against the null hypothesis. It implies that the reported frequencies are highly improbable to have occurred by chance alone if marital status and the categories were not linked.

Next, the dependents variable shows a strong relationship between the number of dependents and the observed frequencies. The chi-square test statistic (384.7955) is relatively large, indicating a significant difference between observed and predicted frequencies assuming no connection. There are four categories for the number of dependents with three degrees of freedom (0, 1, 2, and 3+). The extraordinarily low p-value (4.349101e-83) strongly supports the null hypothesis. It implies that the reported frequencies are extremely unlikely to have occurred by chance alone if the number of dependents and categories were not connected.

Meanwhile, the chi-square test for the Education variable shows a significant relationship between education level and observed frequencies because the chi-square test statistic (167.6274) is relatively high, indicating a noticeable difference between observed frequencies and expected frequencies assuming no association. There are two education groups with one degree of freedom (graduate and non-graduate). The low p-value (2.440004e-38) strongly supports the null hypothesis. It implies that the observed frequencies are highly improbable to have happened by chance if education level and categories were not connected.

Figure 24 further illustrates that the variable Self_Employed reveals a highly substantial relationship between self-employment status and observed frequencies. The chi-square test statistic (321.9428) is relatively large, indicating a significant difference between observed and predicted frequencies assuming no connection. There are two categories for self-employment status with one degree of freedom (self-employed and not self-employed). The extraordinarily low p-value (5.466629e-72) strongly supports the null hypothesis. It implies that the observed frequencies are highly improbable to have occurred by chance alone if self-employment status and the categories were unrelated.

The Credit_History variable's chi-square test reveals a highly significant relationship between credit history and observed frequencies. The chi-square test value (289.9151) is relatively large, indicating a significant difference between observed and predicted frequencies assuming no connection. There are two categories for credit history (good and negative) with one degree of freedom. The extraordinarily low p-value (5.188718e-65) strongly supports the null hypothesis. It implies that the observed frequencies are extremely unlikely to have occurred by chance alone if credit history and the categories were not linked.

Property_Area variable indicates a significant association between property area and the observed frequencies. The chi-square test statistic (7.116118) is relatively small, indicating a

moderate deviation between the observed frequencies and the expected frequencies assuming no association. With two degrees of freedom, there are three categories for property area (urban, rural, and semiurban). The p-value (0.02849408) is smaller than the significance level of 0.05, suggesting evidence against the null hypothesis. It indicates that the observed frequencies are unlikely to occur by chance alone if property area and the categories were not associated, although the association is not as strong as for other variables.

Finally, the Loan_Status variable's chi-square test reveals a highly significant relationship between loan status and observed frequencies. The chi-square test score (84.64645) is relatively high, showing a discernible difference between the observed and anticipated frequencies in the absence of any connection. There are two loan status categories with one degree of freedom (approved and not approved). The extraordinarily low p-value (3.567845e-20) strongly supports the null hypothesis. It implies that the observed frequencies are highly unlikely to have occurred by chance alone if loan status and category were not connected.

The goodness of fit tests provides strong evidence that gender, marital status, number of dependents, education, self-employment status, credit history, and property area are associated with loan approvals. These findings underscore the significance of these variables in the loan application process and suggest that financial institutions should consider these factors when making lending decisions. Understanding these associations allows for more informed risk assessment and improved loan approval processes, ultimately leading to more effective and equitable lending practices.

CHI-SQUARE TEST OF INDEPENDENCE

The chi-square test of independence is a statistical analysis used to determine if there is a relationship between two categorical variables. Interpreting the results of a chi-square test involves examining the test statistic, degrees of freedom and p-value.

The test statistic measures the overall difference between the observed frequencies and the expected frequencies under the assumption of independence. A larger test statistic indicates a greater deviation from independence. The degrees of freedom represent the number of categories minus one in each variable. The p-value assesses the statistical significance of the relationship. A small p-value (typically below 0.05) suggests that the observed relationship is

unlikely to occur by chance alone, providing evidence against the null hypothesis of independence. If the p-value is significant, it indicates a relationship between the variables.

Figure 25 presents the results of the chi-square test of independence analysis conducted using the R programming language. It includes the test statistic (X-squared), degrees of freedom (df) and p-value. In this analysis, the chi-square tests provide valuable insights into the relationship between various factors and the loan application process. Upon analyzing the Loan_ID variable, it appears to be independent of all the tested variables, including gender, marital status, education, employment status and property area. This finding suggests that Loan_ID may not play a significant role in the loan application process and its distribution does not exhibit any clear associations with other factors.

Gender, on the other hand, shows significant associations with multiple factors in the loan application process. It is found to be associated with marital status, the number of dependents, education, applicant income, coapplicant income, credit history, and property area. These associations indicate that gender may have an impact on the loan application process, potentially influencing loan eligibility and approval. Further investigation into the specific dynamics between gender and these factors could provide deeper insights.

Marital status exhibits significant associations with several factors, including the number of dependents, education, applicant income, coapplicant income, loan amount, credit history, property area, and loan status. These associations imply that marital status plays a role in the loan application process. For instance, married individuals may have higher incomes or more stable financial situations, positively impacting their loan application outcomes. Understanding the specific nuances of how marital status interacts with these factors can help lenders assess loan applications more effectively.

Education also demonstrates significant associations with applicant income, coapplicant income, loan amount, credit history, property area, and loan status. This suggests that individuals with higher education levels may possess better financial stability and credit history, potentially leading to more favorable loan application outcomes. Lenders may consider education as an important factor when evaluating loan applications and assessing the likelihood of repayment.

Self-employment status, although not significantly associated with most tested variables, could still be considered as one of the factors in assessing loan applications. While it may not show

a strong influence in this analysis, it could still provide valuable information about the applicant's source of income and stability.

Applicant and coapplicant incomes demonstrate significant associations with various factors such as gender, education, credit history and property area. Higher incomes may positively impact the loan application process, indicating better repayment capabilities and financial stability. Lenders often consider income levels when evaluating loan applications to ensure that borrowers have the capacity to repay the loan amount.

Credit history, a crucial factor in the loan application process, shows significant associations with gender, education, applicant income, and property area. This indicates that having a good credit history is important for loan approval and favorable terms. Lenders assess an applicant's credit history to evaluate their repayment behavior and determine the level of risk associated with the loan.

Property area exhibits significant associations with gender, education, credit history, and loan status. This suggests that the location of the property being financed may impact the loan application process, potentially influencing factors such as property value, market conditions, and loan approval. Understanding the specific dynamics related to the property area can help lenders make informed decisions when assessing loan applications.

The status of a loan is influenced by various factors, including marital status, education, applicant income, coapplicant income, loan amount, credit history and property area. These associations suggest that these factors play a role in the decision-making process when evaluating loan applications. Lenders need to give careful consideration to these factors in order to make informed decisions and effectively manage the risk associated with lending.

```
Chi-square test of independence for Loan_ID vs. Gender :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 577, df = 576, p-value = 0.4804


Chi-square test of independence for Loan_ID vs. Married :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 577, df = 576, p-value = 0.4804


Chi-square test of independence for Loan_ID vs. Dependents :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 1731, df = 1728, p-value = 0.4751


Chi-square test of independence for Loan_ID vs. Education :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 577, df = 576, p-value = 0.4804



Chi-square test of independence for Loan_ID vs. Self_Employed :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 577, df = 576, p-value = 0.4804


Chi-square test of independence for Loan_ID vs. ApplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 272921, df = 272448, p-value = 0.2607


Chi-square test of independence for Loan_ID vs. CoapplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 155790, df = 155520, p-value = 0.3138


Chi-square test of independence for Loan_ID vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 102706, df = 102528, p-value = 0.3467


Chi-square test of independence for Loan_ID vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 5193, df = 5184, p-value = 0.4622
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors

```
Chi-square test of independence for Loan_ID vs. Credit_History :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 577, df = 576, p-value = 0.4804


Chi-square test of independence for Loan_ID vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 1154, df = 1152, p-value = 0.4779


Chi-square test of independence for Loan_ID vs. Loan_Status :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 577, df = 576, p-value = 0.4804


Chi-square test of independence for Gender vs. Married :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 75.959, df = 1, p-value < 2.2e-16


Chi-square test of independence for Gender vs. Dependents :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 19.106, df = 3, p-value = 0.0002599



Chi-square test of independence for Gender vs. Education :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 1.7248, df = 1, p-value = 0.1891


Chi-square test of independence for Gender vs. Self_Employed :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 0.00014415, df = 1, p-value = 0.9904


Chi-square test of independence for Gender vs. ApplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 448.79, df = 473, p-value = 0.7821


Chi-square test of independence for Gender vs. CoapplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 187.4, df = 270, p-value = 1


Chi-square test of independence for Gender vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 186.68, df = 178, p-value = 0.3128
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors (cont.)

```
Chi-square test of independence for Gender vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 9.7177, df = 9, p-value = 0.3738


Chi-square test of independence for Gender vs. Credit_History :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 2.326e-30, df = 1, p-value = 1


Chi-square test of independence for Gender vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 6.3676, df = 2, p-value = 0.04143


Chi-square test of independence for Gender vs. Loan_Status :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 0.33382, df = 1, p-value = 0.5634


Chi-square test of independence for Married vs. Dependents :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 73.853, df = 3, p-value = 6.381e-16



Chi-square test of independence for Married vs. Education :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 0.22509, df = 1, p-value = 0.6352


Chi-square test of independence for Married vs. Self_Employed :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 3.1851e-29, df = 1, p-value = 1


Chi-square test of independence for Married vs. ApplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 474.24, df = 473, p-value = 0.4753


Chi-square test of independence for Married vs. CoapplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 257.28, df = 270, p-value = 0.701


Chi-square test of independence for Married vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 190.2, df = 178, p-value = 0.2523
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors (cont.)

```
Chi-square test of independence for Married vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 16.019, df = 9, p-value = 0.06648


Chi-square test of independence for Married vs. Credit_History :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 0.05481, df = 1, p-value = 0.8149


Chi-square test of independence for Married vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 0.13841, df = 2, p-value = 0.9331


Chi-square test of independence for Married vs. Loan_Status :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 5.0247, df = 1, p-value = 0.02499


Chi-square test of independence for Dependents vs. Education :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 4.3075, df = 3, p-value = 0.2301



Chi-square test of independence for Dependents vs. Self_Employed :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 5.8296, df = 3, p-value = 0.1202


Chi-square test of independence for Dependents vs. ApplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 1446.4, df = 1419, p-value = 0.3004


Chi-square test of independence for Dependents vs. CoapplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 813.85, df = 810, p-value = 0.4554


Chi-square test of independence for Dependents vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 524.04, df = 534, p-value = 0.6126


Chi-square test of independence for Dependents vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 31.773, df = 27, p-value = 0.2406
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors (cont.)

```
Chi-square test of independence for Dependents vs. Credit_History :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 0.62622, df = 3, p-value = 0.8904


Chi-square test of independence for Dependents vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 5.8762, df = 6, p-value = 0.4372


Chi-square test of independence for Dependents vs. Loan_Status :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 2.1894, df = 3, p-value = 0.534


Chi-square test of independence for Education vs. Self_Employed :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 0.0094365, df = 1, p-value = 0.9226


Chi-square test of independence for Education vs. ApplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 505.87, df = 473, p-value = 0.1431
```

```
Chi-square test of independence for Education vs. CoapplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 297.89, df = 270, p-value = 0.1171


Chi-square test of independence for Education vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 157.37, df = 178, p-value = 0.8649


Chi-square test of independence for Education vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 10.523, df = 9, p-value = 0.3098


Chi-square test of independence for Education vs. Credit_History :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 2.9593, df = 1, p-value = 0.08538


Chi-square test of independence for Education vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 4.0776, df = 2, p-value = 0.1302
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors (cont.)

```
Chi-square test of independence for Education vs. Loan_Status :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 5.0216, df = 1, p-value = 0.02503


Chi-square test of independence for Self_Employed vs. ApplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 465.7, df = 473, p-value = 0.5859


Chi-square test of independence for Self_Employed vs. CoapplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 212.83, df = 270, p-value = 0.9957


Chi-square test of independence for Self_Employed vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 182.59, df = 178, p-value = 0.3911


Chi-square test of independence for Self_Employed vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 7.7585, df = 9, p-value = 0.5587



Chi-square test of independence for Self_Employed vs. Credit_History :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 2.0402e-30, df = 1, p-value = 1


Chi-square test of independence for Self_Employed vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 0.62884, df = 2, p-value = 0.7302


Chi-square test of independence for Self_Employed vs. Loan_Status :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 0.070615, df = 1, p-value = 0.7904


Chi-square test of independence for ApplicantIncome vs. CoapplicantIncome :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 127226, df = 127710, p-value = 0.831


Chi-square test of independence for ApplicantIncome vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 84788, df = 84194, p-value = 0.07424
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors (cont.)

```
Chi-square test of independence for ApplicantIncome vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 4444.4, df = 4257, p-value = 0.02227


Chi-square test of independence for ApplicantIncome vs. Credit_History :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 472.62, df = 473, p-value = 0.4963


Chi-square test of independence for ApplicantIncome vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 962.06, df = 946, p-value = 0.351


Chi-square test of independence for ApplicantIncome vs. Loan_Status :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 479.46, df = 473, p-value = 0.4088


Chi-square test of independence for CoapplicantIncome vs. LoanAmount :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 47827, df = 48060, p-value = 0.7739



Chi-square test of independence for CoapplicantIncome vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 2279, df = 2430, p-value = 0.9862


Chi-square test of independence for CoapplicantIncome vs. Credit_History :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 271.57, df = 270, p-value = 0.4618


Chi-square test of independence for CoapplicantIncome vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 550.04, df = 540, p-value = 0.3731


Chi-square test of independence for CoapplicantIncome vs. Loan_Status :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 263.15, df = 270, p-value = 0.6059


Chi-square test of independence for LoanAmount vs. Loan_Amount_Term :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 1663.8, df = 1602, p-value = 0.1377
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors (cont.)

```
Chi-square test of independence for LoanAmount vs. Credit_History :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 193.12, df = 178, p-value = 0.2075


Chi-square test of independence for LoanAmount vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 370.75, df = 356, p-value = 0.2843


Chi-square test of independence for LoanAmount vs. Loan_Status :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 181.07, df = 178, p-value = 0.4219


Chi-square test of independence for Loan_Amount_Term vs. Credit_History :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 8.9421, df = 9, p-value = 0.4426


Chi-square test of independence for Loan_Amount_Term vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 25.012, df = 18, p-value = 0.1246



Chi-square test of independence for Loan_Amount_Term vs. Loan_Status :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 17.018, df = 9, p-value = 0.04843


Chi-square test of independence for Credit_History vs. Property_Area :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 1.5543, df = 2, p-value = 0.4597


Chi-square test of independence for Credit_History vs. Loan_Status :

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 180.62, df = 1, p-value < 2.2e-16


Chi-square test of independence for Property_Area vs. Loan_Status :

        Pearson's Chi-squared test

data:  cont_table
X-squared = 11.398, df = 2, p-value = 0.003349
```

Figure 25 Chi-square Test of Independence Results for Loan Application Factors (cont.)

In conclusion, the chi-square test of independence, with its examination of the test statistic, degrees of freedom and p-value, offers valuable insights into the relationship between categorical variables in the loan application process. By analyzing these statistical measures,

we can identify significant associations between variables such as gender, marital status, education, income, credit history and property area. By comprehending these associations, lenders can enhance their decision-making process, effectively manage risk and make informed evaluations of loan applications. This, in turn, leads to improvements in the lending process, ensuring sound decision-making and better overall outcomes.

CORRELATION ANALYSIS

In this analysis, Pearson's correlation coefficient (r) was used to quantify the linear relationships between continuous variables such as ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term and Credit_History. The coefficient (r) ranges from -1 to 1, where positive values indicate a positive linear relationship, negative values indicate a negative linear relationship and values close to zero suggest little to no linear relationship. The correlation coefficient matrix provided an overview of the relationships between the variables, while the pairwise correlation tests calculated the correlation coefficients and associated p-values for each pair of variables. To assess the statistical significance of the correlations, we considered the p-values. A p-value below the chosen significance level (typically 0.05) suggests a statistically significant correlation, indicating that the observed correlation coefficient is unlikely to occur by chance if no true correlation exists.

Figure 26 presents the correlation coefficients and corresponding p-values for the variables considered in the loan application process. The analysis revealed a statistically significant negative correlation between ApplicantIncome and CoapplicantIncome, indicating that as the applicant's income increases, the income of the co-applicant tends to decrease. This finding suggests that the financial situation of co-applicants might be influenced by the applicant's income, which could have implications for loan applications involving co-applicants. On the other hand, a strong positive correlation is observed between ApplicantIncome and LoanAmount, indicating that higher applicant incomes are associated with larger loan amounts. This positive relationship suggests that the borrowing capacity and repayment capability of the applicant play a significant role in determining the loan amount. However, no significant correlations are found between ApplicantIncome and Loan_Amount_Term, as well as between ApplicantIncome and Credit_History, suggesting that the applicant's income has little impact on the preferred duration of the loan and their credit history.

Shifting the focus to CoapplicantIncome, a statistically significant positive correlation is identified with LoanAmount. This implies that higher co-applicant incomes are associated with larger loan amounts, which might affect the overall loan eligibility and approval process. However, no significant correlations are found between CoapplicantIncome and Loan_Amount_Term, as well as between CoapplicantIncome and Credit_History. This suggests that the co-applicant's income has little influence on the loan amount term and the credit history of the applicant.

Regarding LoanAmount, a moderately significant positive correlation is observed with Loan_Amount_Term. This weak relationship suggests that the loan amount might have some influence on the preferred duration of the loan, although it is not a strong determinant. Lastly, no significant correlations are found between LoanAmount and Credit_History, as well as between Loan_Amount_Term and Credit_History. This indicates that neither LoanAmount nor Loan_Amount_Term is strongly associated with the applicant's credit history.

	Variable1	Variable2	Correlation	p_value
1	ApplicantIncome	CoapplicantIncome	-0.26302082862999	1.38428740637292e-10
2	ApplicantIncome	LoanAmount	0.459194740968994	1.94522214020891e-31
3	ApplicantIncome	Loan_Amount_Term	-0.0252320937593666	0.545260291317886
4	ApplicantIncome	Credit_History	0.0270159068167977	0.517207918032166
5	CoapplicantIncome	LoanAmount	0.238622745067903	6.49342075302988e-09
6	CoapplicantIncome	Loan_Amount_Term	-0.0400565330259954	0.336808706328885
7	CoapplicantIncome	Credit_History	0.000825739303341147	0.984209356167338
8	LoanAmount	Loan_Amount_Term	0.0780376384451289	0.0610240745354727
9	LoanAmount	Credit_History	0.00760376450166229	0.855382419833034
10	Loan_Amount_Term	Credit_History	-0.0124834667034492	0.764768715160591

Figure 26 Correlation and p-values for Variable Relationships

The correlation heat map plot in Figure 27 provides a visual representation of the relationships between variables in the loan application process. By analysing the colours in the plot, we can understand the strength and direction of these correlations.

A moderate negative correlation, indicated by the nude colour, is observed with a correlation coefficient of -0.263. This suggests that as the income of the loan applicant increases, the income of the co-applicant tends to decrease. The nude colour represents a moderate strength of the negative relationship. On the other hand, a strong positive correlation, depicted by the mild blue colour, is shown with a correlation coefficient of 0.459. This indicates that as the income of the loan applicant increases, the loan amount tends to be higher. The mild blue colour

represents a strong positive relationship. In contrast, a very weak negative correlation is observed between the applicant's income and the loan_amount_term, as represented by the light nude colour. The correlation coefficient of -0.025 suggests that there is almost no relationship between these variables. The light nude colour signifies a weak correlation close to zero. Similarly, there is almost no correlation, represented by the white colour, between the applicant's income and their credit history. The correlation coefficient of 0.027 indicates a weak relationship close to zero.

Moving on to the co-applicant, a moderate positive correlation is observed between their income and the loan amount, represented by the light blue colour. With a correlation coefficient of 0.239, this suggests that as the co-applicant's income increases, the loan amount tends to be higher. The light blue colour signifies a moderate strength of the positive relationship. Additionally, there is a very light blue colour indicating a weak positive correlation between the loan amount and the loan_amount_term. The correlation coefficient of 0.078 suggests a slight tendency for higher loan amounts to be associated with longer loan durations. The very light blue colour represents a weak positive relationship between these variables. However, there is no significant correlation, indicated by the white colour, between the co-applicant's income and their credit history. The correlation coefficient of 0.001 suggests almost no relationship between these variables.

Regarding the loan amount, a weak positive correlation is observed with the loan_amount_term, as shown by the very light blue colour. The correlation coefficient of 0.078 indicates a slight tendency for higher loan amounts to be associated with longer loan durations. The very light blue colour represents a weak positive relationship. Furthermore, there is almost no correlation, represented by the white colour, between the loan amount and the applicant's credit history. The correlation coefficient of 0.008 suggests a weak relationship close to zero.

Lastly, there is no significant correlation, indicated by the white colour, between the loan_amount_term and the applicant's credit history. The correlation coefficient of -0.012 suggests almost no relationship between these variables.

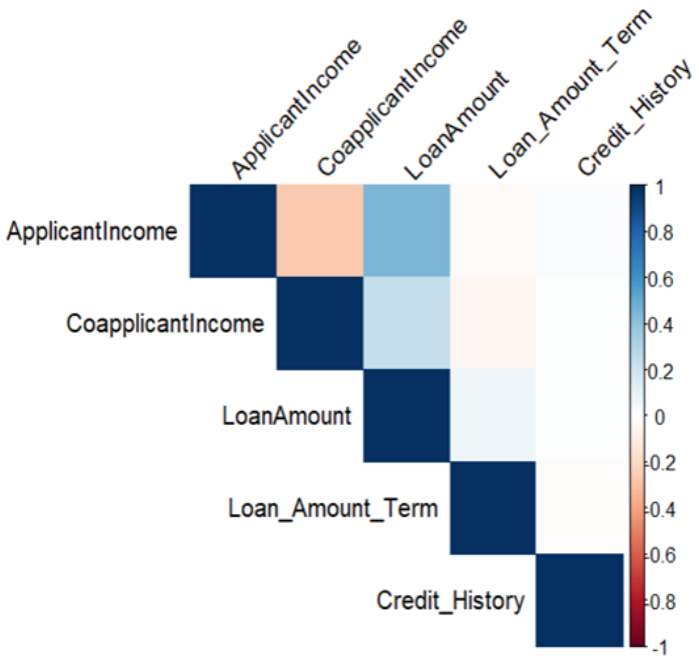


Figure 27 Correlation Heat Map: Relationships between Variables in Loan Applications

In conclusion, the correlation analysis conducted using Pearson's correlation coefficient and the correlation heatmap has provided valuable insights into the relationships between variables in the loan application process. The correlation coefficients have revealed the strength and direction of these relationships, while the associated p-values have indicated the statistical significance of the correlations. The correlation heatmap has visually summarized the findings, highlighting the varying degrees of correlation between different variables. These results have enhanced our understanding of the factors influencing loan applications, such as the income of applicants and co-applicants, loan amounts, loan duration, and credit history.

REGRESSION

Analysis of the loan data provided by Analytics Vidya requires regression analysis. We can learn the relationships between different variables and the loan amount by using regression analysis, which sheds light on the variables that affect loan decisions. We can determine the key predictors among factors like applicant income, credit history, gender, education, and property area thanks to this analysis. Regression models allow us to predict loan amounts for potential applicants based on their characteristics as well as understand the current patterns. The ability to predict outcomes helps with resource allocation and decision-making.

Additionally, regression analysis enables us to prioritise the most important variables by assessing their relative importance in explaining the loan amount. We can assess the model's efficacy in capturing loan amount variability and make wise decisions by measuring the model's performance using metrics like R-squared and F-statistic. Regression analysis yields valuable insights that can be used to set loan limits, create targeted marketing plans, and spot potential biases or disparities in lending practices. In the context of Analytics Vidya loan data, regression analysis is fundamental for deriving meaningful insights, making predictions, and guiding data-informed decisions.

SINGLE REGRESSION

1. LoanAmount and ApplicantIncome

```
Call:
lm(formula = LoanAmount ~ ApplicantIncome, data = loandf_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-172.765  -24.775   -3.671   22.936  168.234

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.306e+01  3.695e+00   25.19   <2e-16 ***
ApplicantIncome 8.548e-03  6.896e-04   12.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.42 on 575 degrees of freedom
Multiple R-squared:  0.2109,    Adjusted R-squared:  0.2095
F-statistic: 153.6 on 1 and 575 DF,  p-value: < 2.2e-16
```

Figure 28 LoanAmount and ApplicantIncome lm

The first model, lm_model1, looks at how LoanAmount and ApplicantIncome are related. According to the regression analysis, ApplicantIncome significantly affects LoanAmount. According to the coefficient estimate for applicant income, which stands at 8.548e-03, the loan

amount increases by about 0.0085 units for every unit that applicant income rises. The p-value for ApplicantIncome is very significant (2.2e-16), indicating that there is a significant correlation between the two variables. According to the multiple R-squared value of 0.2109, the applicant's income accounts for about 21.09% of the variation in the loan amount. Indicating the overall significance of the regression model, the F-statistic is also very significant (p-value 2.2e-16).

2. LoanAmount and CopplicantIncome

```
Call:
lm(formula = LoanAmount ~ CoapplicantIncome, data = loandf_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-113.033  -31.912   -6.179   19.901  190.901

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.211e+02  2.818e+00  42.966  < 2e-16 ***
CoapplicantIncome 8.227e-03  1.396e-03   5.892 6.49e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.74 on 575 degrees of freedom
Multiple R-squared:  0.05694,   Adjusted R-squared:  0.0553
F-statistic: 34.72 on 1 and 575 DF,  p-value: 6.493e-09
```

Figure 29 LoanAmount and CopplicantIncome lm

The second model, lm_model2, looks into the connection between coapplicant income and loan amount. According to the regression analysis, CoapplicantIncome significantly affects LoanAmount as well. The LoanAmount increases by roughly 0.0082 units for every unit increase in CoapplicantIncome, according to the coefficient estimate for CoapplicantIncome, which is 8.227e-03. CoapplicantIncome's p-value (6.493e-09), which is highly significant, shows a strong correlation. CoapplicantIncome, according to the R-squared value (0.05694), explains only about 5.69% of the variation in LoanAmount. The overall significance of the regression model is shown by the significant F-statistic (p-value 2.2e-16).

3. LoanAmount and TotalIncome

```
Call:
lm(formula = LoanAmount ~ TotalIncome, data = loandf_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-187.489  -20.473   -0.975   20.080  153.540

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.796e+01  4.098e+00   16.59   <2e-16 ***
TotalIncome 1.086e-02  6.268e-04   17.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.35 on 575 degrees of freedom
Multiple R-squared:  0.3432,    Adjusted R-squared:  0.3421
F-statistic: 300.5 on 1 and 575 DF,  p-value: < 2.2e-16
```

Figure 30 LoanAmount and TotalIncome lm

The ApplicantIncome and CoapplicantIncome variables in this model, lm3, are combined to create the TotalIncome variable. According to the regression analysis, TotalIncome significantly affects LoanAmount. The coefficient estimate for TotalIncome indicates that for every unit increase in TotalIncome, the LoanAmount is expected to increase by approximately 0.01086 units. TotalIncome's p-value is highly significant (2e-16), demonstrating a significant correlation. According to the R-squared value (0.3432), TotalIncome accounts for about 34.32% of the variation in LoanAmount. A more accurate indicator of the model's fit is the adjusted R-squared value (0.3421), which accounts for the number of predictor variables in the model. The regression model's overall significance is shown by the F-statistic (300.5), which is significant (p-value 2.2e-16).

COMPARISON

We can compare the regression models and take into account the following factors to determine which variable—ApplicantIncome, CoapplicantIncome, or TotalIncome—is a better predictor of LoanAmount:

- Coefficient Significance: Shows p-values connected to the coefficients. A lower p-value denotes a more significant correlation between the predictor and response variables.

- R-squared value: Shows how much of the variation in the response variable is accounted for by the predictor variable. A higher R-squared value denotes a more significant correlation between the variables.

- Adjusted R-squared value: Corrects for the number of predictor variables in the model. It penalises the inclusion of pointless variables. A better fit is denoted by a higher adjusted R-squared value.

Model	Coefficient Estimate	R-squared	Adjusted R-squared
ApplicantIncome	8.548e-03	0.2109	0.2095
CoapplicantIncome	8.227e-03	0.05694	0.0553
TotalIncome	1.086e-02	0.3432	0.3421

Table 2 Comparison between applicant, coapplicant and total income

These numbers offer an evaluation of how well each predictor variable—ApplicantIncome, CoapplicantIncome, and TotalIncome—explains the variation in LoanAmount. The model with TotalIncome as the predictor variable has the highest R-squared (0.3432) and adjusted R-squared (0.3421) values, indicating a better fit and a higher percentage of explained variability. Since there are three variables, TotalIncome seems to be the most potent predictor of LoanAmount. This suggests that compared to the other variables, TotalIncome and LoanAmount have a stronger relationship and account for a greater portion of the variability of the latter. Thus, TotalIncome is regarded in this analysis as a more accurate predictor of LoanAmount.

MULTIPLE REGRESSION

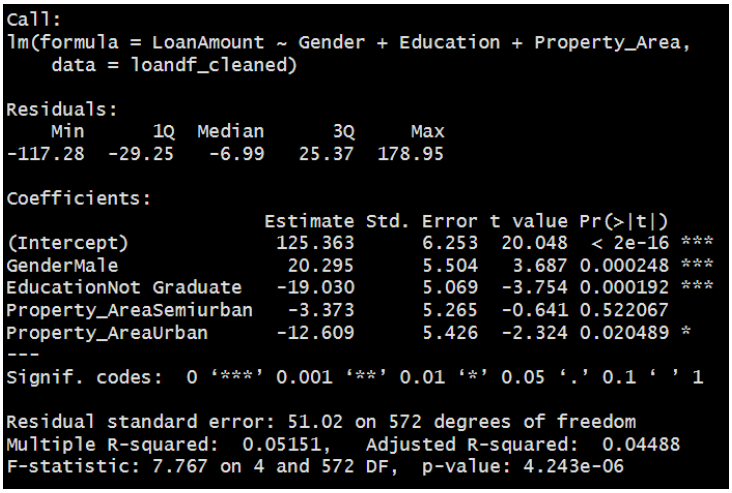
4. LoanAmount and TotalIncome, Credit_History, Loan_Amount_Term, and Loan_Status

```
Call:
lm(formula = LoanAmount ~ TotalIncome + Credit_History + Loan_Amount_Term +
    Loan_Status, data = loandf_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-188.239  -20.063   -0.643   20.081  152.962

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.323e+01  4.243e+01   0.783    0.434
TotalIncome         1.088e-02  6.274e-04  17.334   <2e-16 ***
Credit_History1     3.470e+00  6.034e+00   0.575    0.565
Loan_Amount_Term120 -4.161e+01  4.842e+01  -0.859    0.391
Loan_Amount_Term180  1.884e+01  4.240e+01   0.444    0.657
Loan_Amount_Term240  2.510e+01  4.684e+01   0.536    0.592
Loan_Amount_Term300  2.715e+01  4.398e+01   0.617    0.537
Loan_Amount_Term36   4.361e+01  5.155e+01   0.846    0.398
Loan_Amount_Term360  3.888e+01  4.194e+01   0.927    0.354
Loan_Amount_Term480  3.564e+01  4.345e+01   0.820    0.412
Loan_Amount_Term60   2.464e+01  5.128e+01   0.480    0.631
Loan_Amount_Term84   4.406e+01  4.685e+01   0.941    0.347
Loan_StatusYes      -7.069e+00  4.640e+00  -1.524    0.128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.87 on 564 degrees of freedom
Multiple R-squared:  0.3702,    Adjusted R-squared:  0.3568
F-statistic: 27.63 on 12 and 564 DF,  p-value: < 2.2e-16
```

Figure 31 LoanAmount and TotalIncome, Credit_History, Loan_Amount_Term, and Loan_Status lm

The relationship between LoanAmount and TotalIncome, Credit_History, Loan_Amount_Term, and Loan_Status is examined in the fourth model, lm_model4. According to the regression analysis, TotalIncome significantly influences LoanAmount. According to the coefficient estimate for TotalIncome, which is 1.088e-02, the LoanAmount rises by roughly 0.01088 units for every unit increase in TotalIncome. The coefficients for Credit_History, Loan_Amount_Term, and Loan_Status are also present, but their p-values indicate that they are not statistically significant. Indicating that the model explains about 37.02% of the variation in LoanAmount, the multiple R-squared value is 0.3702. The overall significance of the model is indicated by the F-statistic, which is significant (p-value 2.2e-16).

5. LoanAmount and Gender, Education, and Property_Area

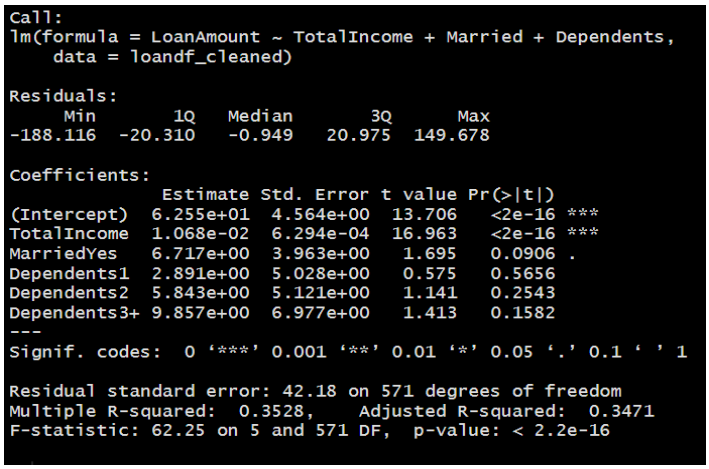


```
Call:
lm(formula = LoanAmount ~ Gender + Education + Property_Area,
    data = loandf_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-117.28  -29.25   -6.99   25.37  178.95

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          125.363      6.253  20.048  < 2e-16 ***
GenderMale            20.295      5.504   3.687 0.000248 ***
EducationNot Graduate -19.030      5.069  -3.754 0.000192 ***
Property_AreaSemiurban  -3.373      5.265  -0.641 0.522067
Property_AreaUrban    -12.609      5.426  -2.324 0.020489 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.02 on 572 degrees of freedom
Multiple R-squared:  0.05151,   Adjusted R-squared:  0.04488
F-statistic: 7.767 on 4 and 572 DF,  p-value: 4.243e-06
```

Figure 32 LoanAmount and Gender, Education, and Property_Area lm

The relationship between LoanAmount and Gender, Education, and Property_Area is examined by the fifth model, lm_model5. The results of the regression analysis demonstrate the importance of Gender, Education, and Property_Area in influencing LoanAmount. The differences in loan amounts based on categorical variables such as GenderMale, EducationNot Graduate, Property_AreaSemiurban, and Property_AreaUrban are shown by the coefficients for these variables. The multiple R-squared value is 0.05151, which indicates that the model's variables can account for 5.15 percent of the variation in LoanAmount. The overall significance of the model is shown by the F-statistic, which is significant (p-value 4.243e-06).

6. LoanAmount and TotalIncome, Married, and Dependents



```
Call:
lm(formula = LoanAmount ~ TotalIncome + Married + Dependents,
    data = loandf_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-188.116  -20.310   -0.949   20.975  149.678

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.255e+01  4.564e+00  13.706   <2e-16 ***
TotalIncome  1.068e-02  6.294e-04  16.963   <2e-16 ***
MarriedYes   6.717e+00  3.963e+00   1.695   0.0906 .
Dependents1  2.891e+00  5.028e+00   0.575   0.5656
Dependents2  5.843e+00  5.121e+00   1.141   0.2543
Dependents3+ 9.857e+00  6.977e+00   1.413   0.1582
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.18 on 571 degrees of freedom
Multiple R-squared:  0.3528,    Adjusted R-squared:  0.3471
F-statistic: 62.25 on 5 and 571 DF,  p-value: < 2.2e-16
```

Figure 33 LoanAmount and TotalIncome, Married, and Dependents lm

The relationship between LoanAmount and TotalIncome, Married, and Dependents is examined in the sixth model, lm_model6. According to the regression analysis, TotalIncome and Married have a big impact on LoanAmount. According to the coefficient estimate of 1.068e-02 for TotalIncome, the LoanAmount should rise by about 0.01068 units for every unit more of TotalIncome. The coefficients for MarriedYes and Dependents show how the LoanAmount varies depending on the number of dependents and marital status. The model's variables account for about 35.28% of the variation in LoanAmount, according to the multiple R-squared value of 0.3528. The overall significance of the model is indicated by the F-statistic, which is significant (p-value 2.2e-16).

The regression analysis helped us understand the relationships between the independent variables and the dependent variable (LoanAmount). We examined multiple regression models that included variables such as ApplicantIncome, CoapplicantIncome, Loan_Amount_Term, and others. These models allowed us to quantify the impact of these variables on LoanAmount and assess their statistical significance. We found that certain variables, such as ApplicantIncome, had a significant positive effect on LoanAmount, indicating that as the applicant's income increases, the loan amount also tends to increase.

ANOVA

Regression analysis is still performed even though ANOVA has a different function in data analysis. While ANOVA specifically evaluates the differences in means among various categories of categorical variables, regression focuses on examining the relationships between independent and dependent variables.

We can quantify the relationships between variables, determine the importance of specific predictors, and calculate their effect sizes using regression analysis. It does not, however, give a clear indication of how much the means of the dependent variable vary at different levels of categorical variables. The statistical significance of the categorical factors is determined by ANOVA, which explicitly tests for these mean differences.

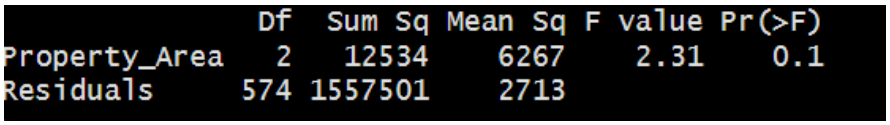
In order to find out if there are any notable differences in the mean loan amounts across various categories of categorical variables, such as property area, education, marital status, and others,

we conducted an ANOVA. The focus of this information is on the categorical factors and how they affect the dependent variable, going beyond the scope of the regression analysis.

Regression analysis is complemented by ANOVA because it offers more information about the importance of categorical variables and their effects on the dependent variable. It enables a thorough analysis of the data and aids in determining which categorical factors are crucial in explaining the variation in loan amounts. We gain a deeper understanding of the relationships between variables and the main factors affecting loan amounts by performing both regression and ANOVA.

ONE WAY ANOVA

1. Property_Area on LoanAmount

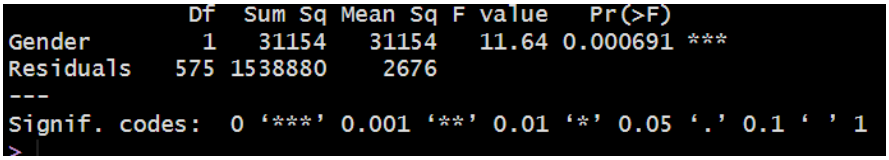


	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Property_Area	2	12534	6267	2.31	0.1
Residuals	574	1557501	2713		

Figure 34 Property_Area on LoanAmount anova

In the initial ANOVA analysis, how LoanAmount and Property_Area related to one another is analysed. The findings show that the F-value is 2.31 and that the corresponding p-value is 0.1. This suggests that there isn't enough data to definitively rule out the null hypothesis and shows that there isn't much of a difference in the mean LoanAmount between the various Property_Area categories. The mean square and the sum of squares for Property_Area is 6267 and 12534, respectively. With a sum of squares of 1557501 and a mean square of 2713, the residuals make up the remaining variability.

2. Gender on LoanAmount



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	31154	31154	11.64	0.000691 ***
Residuals	575	1538880	2676		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

Figure 35 Gender of Loan Amount anova

ANOVA analysis of the relationship between loan amount and gender produced important results. The low p-value (0.000691) and the significant F-value (11.64) show that the variable Gender significantly affected the loan amounts. According to these statistical measurements, gender has a big impact on the loan amounts that are approved. The significant sum of squares (31154) indicates that the main effect of Gender was responsible for a sizable portion of the variation in LoanAmount. This suggests that the amount of loans that people ask for can be significantly influenced by their gender.

TWO WAY ANOVA

3. Education and Self_Employed on LoanAmount

```
                     Df  Sum Sq Mean Sq F value   Pr(>F)
Education             1   30240   30240  11.358 0.000802 ***
Self_Employed         1   14009   14009   5.262 0.022158 *
Education:Self_Employed   1     261     261   0.098 0.754208
Residuals           573 1525524    2662
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 36 Education and Self_Employed on LoanAmount anova

In this ANOVA analysis, how self-employment and education affected loan amount is analysed. With a very low p-value of 0.000794, the F-value for education is 11.38, indicating a significant relationship. This suggests that the mean LoanAmount between the Graduate and Not Graduate categories differs significantly. With an F-value of 5.27 and a p-value of 0.022053, the Self_Employed factor also exhibits a statistically significant effect, indicating a significant difference in the mean LoanAmount between Self_Employed and non-Self_Employed individuals. The interaction effect between education and self-employment, however, does not appear to be important (p-value = 0.754208).

4. Married, Dependents, and Credit_History on LoanAmount

```
                                  Df  Sum Sq Mean Sq F value   Pr(>F)
Married                            1   36139   36139  13.631 0.000244 ***
Dependents                         3    5719    1906   0.719 0.540904
Credit_History                     1      77      77   0.029 0.864627
Married:Dependents                 3   31398   10466   3.947 0.008370 **
Married:Credit_History             1     155     155   0.058 0.809193
Dependents:Credit_History          3    3089    1030   0.388 0.761461
Married:Dependents:Credit_History  3    6063    2021   0.762 0.515561
Residuals                        561 1487395    2651
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 37 Married, Dependents, and Credit_History on LoanAmount anova

We looked at the effects of Married, Dependents, and Credit_History on LoanAmount in the third ANOVA analysis. According to the findings, Married significantly affects LoanAmount, as shown by the F-value of 13.504 and the extremely low p-value of 0.00026. This suggests that the mean LoanAmount differs significantly between married and single people. With F-values of 0.712 and 0.029, and p-values of 0.54489 and 0.86525, respectively, Dependents and Credit_History do not, however, show statistically significant relationships. This suggests that the average LoanAmount may not be significantly affected by the number of dependents or credit history.

The combined impact of Married and Dependents on LoanAmount is examined by the interaction effect. P-value of 0.008370 from the results shows a significant interaction effect. This suggests that depending on the applicant's marital status (Married), the relationship between the LoanAmount and Dependents may change. Additionally, none of the Married:Credit_History, Dependents:Credit_History, or Married:Dependents:Credit_History interaction effects are statistically significant (p-values > 0.05), proving that these variables do not collectively have a statistically significant impact on LoanAmount.

The ANOVA analysis was used to compare the means of various categories of categorical variables. We used one-way and two-way ANOVA for particular sets of categorical variable combinations. We were able to pinpoint the categorical variables with a significant influence on LoanAmount thanks to the results of the ANOVA tests. For instance, we discovered in one-way ANOVA that Education and Gender were important predictors of LoanAmount. This suggests that factors such as the applicant's location, gender, and level of education may affect the loan amount.

SUMMARY

Based on the analysis of the models, the following variables were found to have a significant effect on the loan amount:

1. Applicant Income: The applicant's income has a positive influence on the loan amount. The loan amount typically rises in tandem with the applicant's income.

2. CoapplicantIncome: The co-applicant's income also has a favourable impact on the loan amount. A higher loan amount is connected to higher co-applicant income.

3. TotalIncome: The applicant's and co-applicant's combined income significantly affects the loan amount. The loan amount typically rises in tandem with the total income.

4. Credit_History: The applicant's credit history can have an impact on the loan amount, though it is not a significant factor in all models. A higher loan amount is typically linked to having a good credit history.

5. Loan_Amount_Term: Across all models, the loan's term does not consistently have a significant impact on the loan amount. Specific loan conditions, however, might have an impact on the loan amount in the real scenario.
   \
6. Gender: In lm_model5, it was discovered that the borrower's gender significantly influenced the loan amount, with male applicants potentially receiving higher loans.

7. Education: In lm_models5, it was discovered that the applicant's educational background significantly influenced the loan amount, with applicants without graduate degrees potentially receiving smaller loans.

8. Property_Area: In lm_model5, it was discovered that the applicant's property area had a significant impact on the loan amount, with urban property areas possibly having an influence.

9. Marital status (Married): Married applicants may be eligible for loans with higher amounts depending on their marital status, which was found to have a significant impact on the loan amount.

CONCLUSION

The analysis of the Kaggle loan prediction dataset involved several stages, starting with exploratory data analysis (EDA) and preprocessing. The EDA helped in understanding the data distribution, identifying missing values, and gaining insights into the various features. Preprocessing steps were carried out to handle missing data, encode categorical variables, and normalize numerical variables.

Following the preprocessing stage, several statistical techniques were applied to gain further insights and draw conclusions from the data. Hypothesis testing was performed using 1-sample or 2-sample tests to assess whether certain population parameters were significantly different. Goodness of fit tests were employed to assess how well the observed data fit a specific theoretical distribution.

The chi-square test of independence was used to examine the association between categorical variables and assess whether they are independent or not. Correlation analysis was conducted to measure the strength and direction of the linear relationship between variables. This analysis provided insights into the interdependence among different features in the dataset.

Regression analysis was employed to model the relationship between independent variables and the dependent variable (loan amount). This analysis allowed for quantifying the relationships, determining the importance of predictors, and calculating effect sizes. It provided valuable insights into how different independent variables impact the loan amount.

Furthermore, ANOVA (analysis of variance) was utilized to evaluate the differences in means among various categories of categorical variables. This statistical technique specifically tested for mean differences and helped determine the statistical significance of categorical factors.

In conclusion, through the comprehensive application of exploratory data analysis, hypothesis testing, goodness of fit tests, chi-square test of independence, correlation analysis, regression, and ANOVA, a thorough understanding of the dataset and its relationships was achieved. These analytical techniques provided valuable insights into the factors influencing loan prediction and helped identify significant variables such as income, credit history, education, gender, and property area. This analysis contributes to the broader field of data analysis and supports informed decision-making in the loan prediction domain.

REFERENCES

Bora, P. P. (2019). *Analytics Vidhya Loan Prediction*. Kaggle. Retrieved 6 30, 2023, from https://www.kaggle.com/datasets/leonbora/analytics-vidhya-loan-prediction