HO CHI MINH CITY

UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

COURSE: PROBABILITY AND STATISTICS

(Faculty of Applied Science)

DREAM HOUSING FINANCE'S LOAN DATASET ANALYSIS WITH LOAN APPROVAL PREDICTION USING MACHINE LEARNING MODEL

Class: CC04 Group: 13

May 2023

Instructor: Dr. Phan Thi Huong

Student	Student ID	Score
Nguyễn Kiều Bảo Khánh	2152654	
Trần Trường Giang	2152534	
Nguyễn Nhất Huy	2053042	
Trần Nhật Tân	2112259	
Phương Xương Thịnh	2012479	

Ho Chi Minh City – 2023

Team Members and Workload

This is our team information and percentage of work. Each member commits the same amount of work for the assignment.

No.	Full name	Student ID	Workload
1	Nguyễn Kiều Bảo Khánh	2152654	20%
2	Trần Trường Giang	2152534	20%
3	Nguyễn Nhất Huy	2053042	20%
4	Trần Nhật Tân	2112259	20%
5	Phương Xương Thịnh	2012479	20%

I. Data introduction

1. Introduction of dataset

Dream Housing Finance company deals in all kinds of home loans. They have a presence across all urban, semi-urban and rural areas. The customer first applies for a home loan and after that, the company validates the customer eligibility for the loan.

The company wants to automate the loan eligibility process (real-time) based on customer detail provided while filling out online application forms. These details are Gender, Marital Status, Education, number of Dependents, Income, Loan Amount, Credit History, and others.

To automate this process, they have provided a dataset to identify the customer segments that are eligible for loan amounts so that they can specifically target these customers.

2. Datasets properties

The Loan Prediction Dataset is a popular dataset (history of the loan applications of Dream Housing Finance) used in machine learning to predict whether a loan application will be approved or not based on various factors. The dataset contains information about loan applicants such as their gender, marital status, education, income, loan amount, loan term, credit history, and more.

The data set we got is the history of loan applicants with all the information of the applicants. The details are:

- **Gender**: Refers to the applicant's gender, which can be male or female. This information can be used to analyze the loan application trends based on gender.
- **Marital Status**: Refers to the applicant's current marital status, which can be *Married or Not Married*. This information can be used to analyze the loan application trends based on marital status.
- **Education**: Refers to the applicant's educational qualification, which can be *Graduate or Not Graduate*. This information can be used to analyze the loan application trends based on education levels.
- **Number of Dependents**: Refers to the number of people *(1, 2, or 3+)* who depend on the applicant for support. This information can be used to analyze the loan application trends based on the number of dependents.
- **Income (Applicant Income and Co-Applicant Income)**: Refers to the amount of money the applicant earns annually or monthly from all sources of income. This information can be used to determine the applicant's ability to repay the loan.

- **Loan Amount**: Refers to the amount of money the applicant is applying for as a loan. This information can be used to analyze the loan application trends based on loan amounts.
- **Loan Amount Term:** Refers to the length of time over which a borrower will repay their loan. It is usually expressed in months and is one of the factors that lenders consider when determining the eligibility of a borrower for a loan.
- **Credit History**: Refers to the applicant's past credit record including their repayment status. This information can be used to determine the applicant's creditworthiness and likelihood of repaying the loan.
- **The Property Area:** Refers to the type of location where the property for which a loan is being requested is situated. There are three categories in this column: *Urban, Semiurban, and Rural.*
- **Loan Status**: Refers to whether the applicant is eligible for the loan or not based on their overall profile and creditworthiness. This information can be used to determine the loan approval rate and identify any patterns in loan rejections.

II. Background theory

1. Logistic Regression

The principal use of the supervised machine learning technique, logistic regression, is classification assignments where the objective is to estimate the likelihood that a given instance belongs to a certain class. Its term is logistic regression, and it is utilized for classification methods. Regression is used because it employs a sigmoid function to estimate the probability for the given class using the result of a linear regression function as input. Logistic regression differs from linear regression in that it predicts the likelihood that an instance will belong to a specific class or not, whereas the output of the former is a continuous number that can be anything.

Terminologies involved in Logistic Regression:

- **Independent variables**: The input characteristics or predictor factors applied to the dependent variable's predictions.
- **Dependent variable**: The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function**: The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.
- **Odds**: It is the ratio of something occurring to something not occurring. It is different from probability as probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds**: The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of

the dependent variable are modeled as a linear combination of the independent variables and the intercept.

- **Coefficient**: The logistic regression model's estimated parameters, show how the independent and dependent variables relate to one another.
- **Intercept**: A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- **Maximum likelihood estimation**: The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

How Logistic Regression works:

The logistic regression model uses a sigmoid function, which converts any real-valued collection of independent variables input into a value between 0 and 1, to convert the continuous value output of the linear regression function into categorical value output. The logistic function is the name of this process.

Let the independent input features be :

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

and the dependent variable is Y having only binary value i.e 0 or 1.

$$Y = \begin{cases} 0 & \text{if } Class\ 1 \\ 1 & \text{if } Class\ 2 \end{cases}$$

then apply the multi-linear function to the input variables X

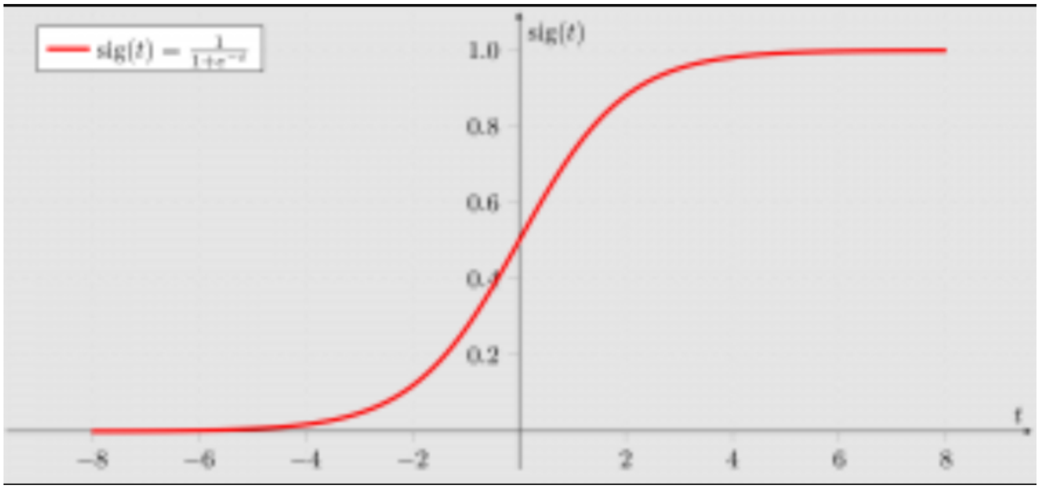
$$z = \left(\sum_{i=1}^{n} w_i x_i\right) + b$$

Here Xi is the ith observation of X, $w_i = [w_1, w_2, w_3, \cdots, w_m]$ is the weights or Coefficient and b is the bias term also known as intercept, simply this can be represented as the dot product of weight and bias.

$$z = w \cdot X + b$$

Whatever we discussed above is linear regression. Now we use the sigmoid function where the input will be z and we find the probability between 0 and 1. i.e predicted y.

$$\sigma(z) = \frac{1}{1-e^{-z}}$$



As shown above the fig sigmoid function converts the continuous variable data into the probability i.e between 0 and 1.

- $\sigma(z)$ tend towards 1 as $|z \rightarrow \infty$
- $\sigma(z)$ tends towards 0 as $|z \rightarrow -\infty$
- $\sigma(z)$ is always bounded between 0 and 1.

where the probability of being a class can be measured as:

$$P(y=1) = \sigma(z)$$

$$P(y=0) = 1 - \sigma(z)$$

Logistic Regression Equation:

The odd is the ratio of something occurring to something not occurring. It is different from probability as probability is the ratio of something occurring to everything that could possibly occur. so odd will be

$$\frac{p(x)}{1-p(x)} = e^{z}$$

Applying natural log on odd. then log odd will be

$$\log\left[\frac{p(x)}{1-p(x)}\right] = z$$

$$\log\left[\frac{p(x)}{1-p(x)}\right] = w \cdot X + b$$

then the final logistic regression equation will be:

$$p(X; b, w) = \frac{e^{w \cdot X + b}}{1+e^{w \cdot X+b}} = \frac{1}{1+e^{-w \cdot X+b}}$$

The likelihood function for Logistic Regression:

The predicted probabilities will p(X;b,w) = p(x) for y=1 and for y = 0 predicted probabilities will 1-p(X;b,w) = 1-p(x)

$$L(b, w) = \prod_{i=1} np(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

Taking natural logs on both sides

$$\begin{aligned}
l(b,w) = \log(L(b,w)) &= \sum_{i=1}^{n} y_i \log p(x_i) \; + \; (1-y_i)\log(1-p(x_i)) \\
&= \sum_{i=1}^{n} y_i \log p(x_i) + \log(1-p(x_i)) - y_i \log(1-p(x_i)) \\
&= \sum_{i=1}^{n} \log(1-p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1-p(x_i} \\
&= \sum_{i=1}^{n} -\log 1 - e^{-(w \cdot x_i + b)} + \sum_{i=1}^{n} y_i(w \cdot x_i + b) \\
&= \sum_{i=1}^{n} -\log 1 + e^{w \cdot x_i + b} + \sum_{i=1}^{n} y_i(w \cdot x_i + b)
\end{aligned}$$

The gradient of the log-likelihood function

To find the maximum likelihood estimates, we differentiate w.r.t w

$$\frac{\partial J(l(b,w)}{\partial w_j} = -\sum_{i=n}^{n} \frac{1}{1+e^{w\cdot x_i+b}} e^{w\cdot x_i+b} x_{ij} + \sum_{i=1}^{n} y_i x_{ij}$$

$$= -\sum_{i=n}^{n} p(x_i; b, w) x_{ij} + \sum_{i=1}^{n} y_i x_{ij}$$

$$= \sum_{i=n}^{n} (y_i - p(x_i; b, w)) x_{ij}$$

Assumptions for Logistic regression:

The assumptions for Logistic regression are as follows:

- **Independent observations**: Each observation is independent of the other. meaning there is no correlation between any input variables.
- **Binary dependent variables**: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories softmax functions are used.
- **Linearity relationship between independent variables and log odds**: The relationship between the independent variables and the log odds of the dependent variable should be linear.
- **No outliers**: There should be no outliers in the dataset.
- **Large sample size**: The sample size is sufficiently large.

2. Hypothesis testing

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

Let's discuss few examples of statistical hypothesis from real-life -

- A teacher assumes that 60% of his college's students come from lower-middle-class families.
- A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

Now that you know about hypothesis testing, look at the two types of hypothesis testing in statistics.

How Hypothesis Testing works?

To show that the null hypothesis is plausible, an analyst tests the null hypothesis using a statistical sample. To test a hypothesis, measurements and analysis are performed on a randomly chosen sample of the population. Researchers test the null and alternative hypotheses using a random population sample.

The null hypothesis is typically an equality hypothesis between population parameters; for example, a null hypothesis may claim that the population means return equals zero. The alternate hypothesis is essentially the inverse of the null hypothesis (e.g., the population means the return is not equal to zero). As a result, they are mutually exclusive, and only one can be correct. One of the two possibilities, however, will always be correct.

Null Hypothesis and Alternative Hypothesis

The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.

H0 is the symbol for it, and it is pronounced H-naught.

The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.

Let's understand this with an example.

A sanitizer manufacturer claims that its product kills 95 percent of germs on average.

To put this company's claim to the test, create a null and alternate hypothesis.

H0 (Null Hypothesis): Average = 95%.

Alternative Hypothesis (H1): The average is less than 95%.

Another straightforward example to understand this concept is determining whether or not a coin is fair and balanced. The null hypothesis states that the probability of a show of heads is equal to the likelihood of a show of tails. In contrast, the alternate theory states that the probability of a show of heads and tails would be very different.

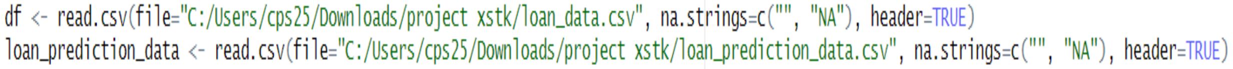
III. Descriptive statistics

1. Import data

The dataset is a history of the loan applications of Dream Housing Finance. The applicant was examined by some specialists and was determined to be eligible for the loan or not. Data was then collected along with the loan status.

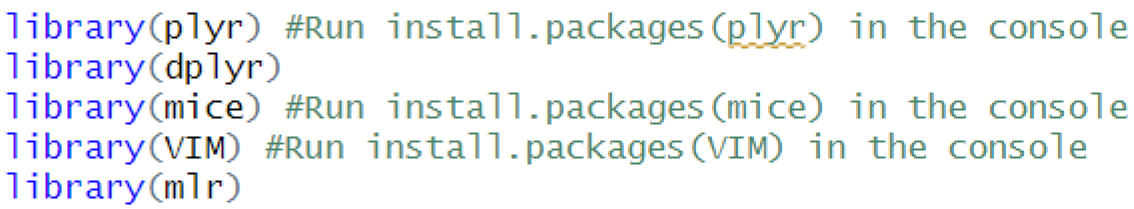
This dataset was collected for us to test our hypothesis that machine learning can understand. We are trying a new approach different from the previous intuitive and human based approach.

Before analyzing, we must read the file loan_data.csv by df and the file loan_prediction_data.csv by loan_prediction_data by following function:



```
df <- read.csv(file="C:/Users/cps25/Downloads/project xstk/loan_data.csv", na.strings=c("", "NA"), header=TRUE)
loan_prediction_data <- read.csv(file="C:/Users/cps25/Downloads/project xstk/loan_prediction_data.csv", na.strings=c("", "NA"), header=TRUE)
```

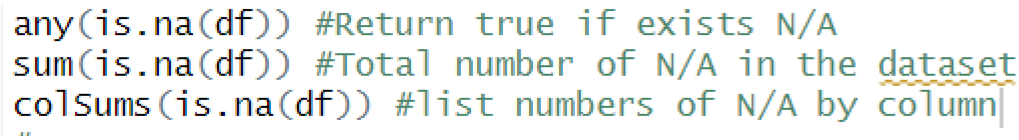
Then we have to declare some library:



```
library(plyr) #Run install.packages(plyr) in the console
library(dplyr)
library(mice) #Run install.packages(mice) in the console
library(VIM) #Run install.packages(VIM) in the console
library(mlr)
```

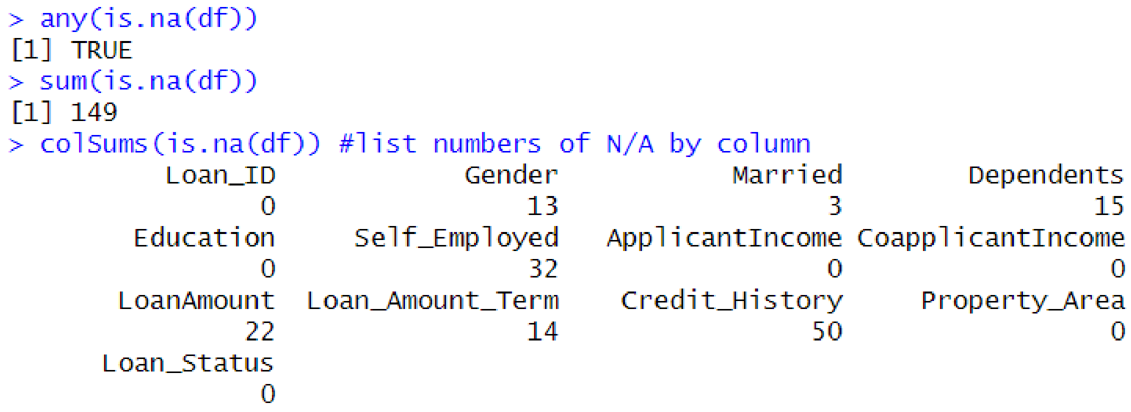
2. Data exploratory

Before analyzing the data, we must check the data for any missing values and outliers. First, we will look over the dataset and see if there is any N/A value. This can be done with the following functions:



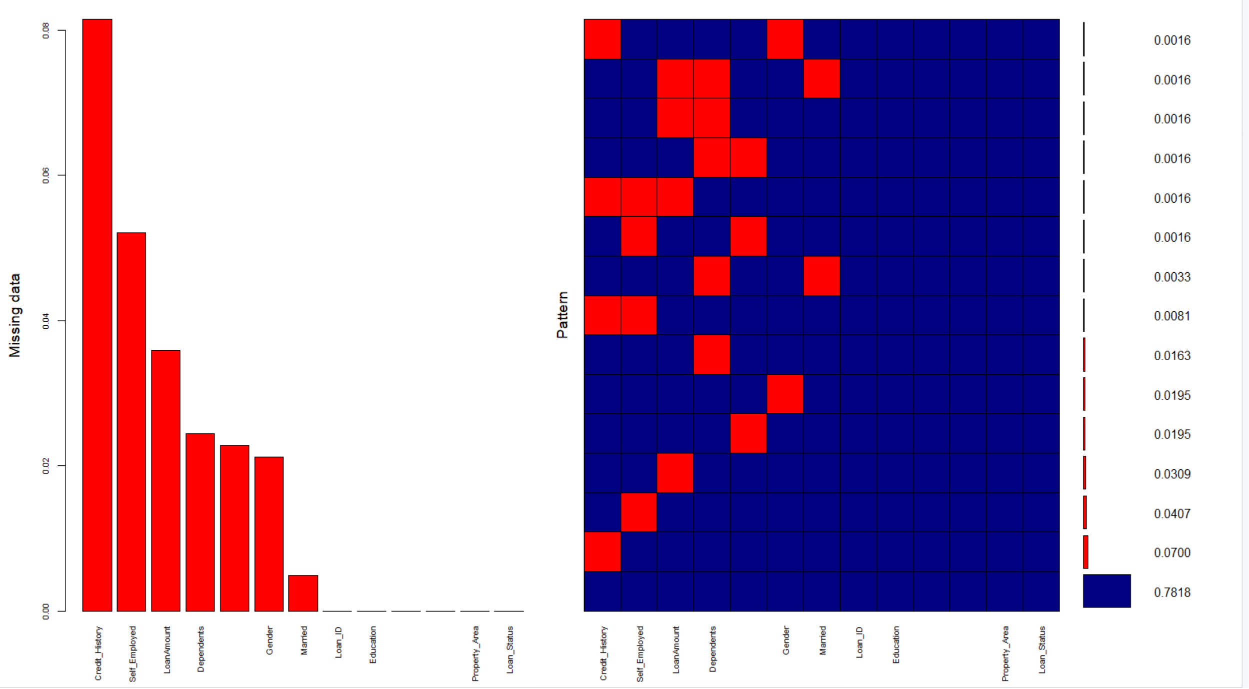
```
any(is.na(df)) #Return true if exists N/A
sum(is.na(df)) #Total number of N/A in the dataset
colSums(is.na(df)) #list numbers of N/A by column
```

This function returns the number of N/A values in each column.

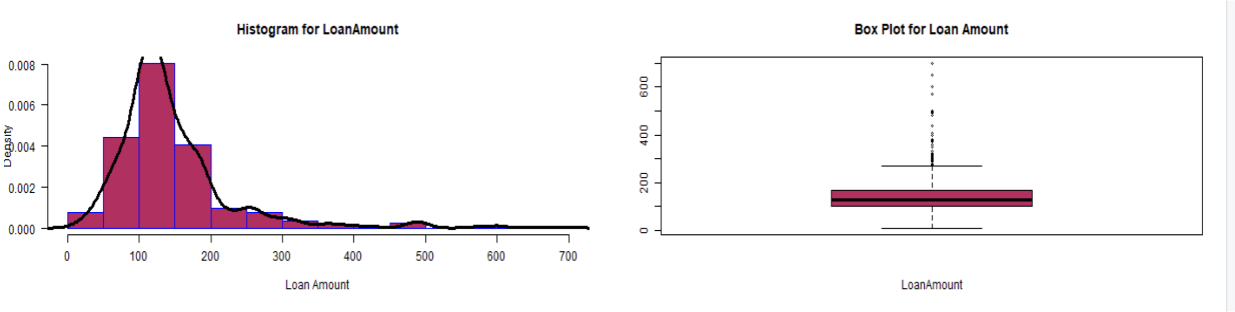


```
> any(is.na(df))
[1] TRUE
> sum(is.na(df))
[1] 149
> colSums(is.na(df)) #list numbers of N/A by column
          Loan_ID            Gender           Married        Dependents 
                0                13                 3                15 
        Education     Self_Employed   ApplicantIncome CoapplicantIncome 
                0                32                 0                 0 
       LoanAmount  Loan_Amount_Term    Credit_History     Property_Area 
               22                14                50                 0 
      Loan_Status 
                0 
```

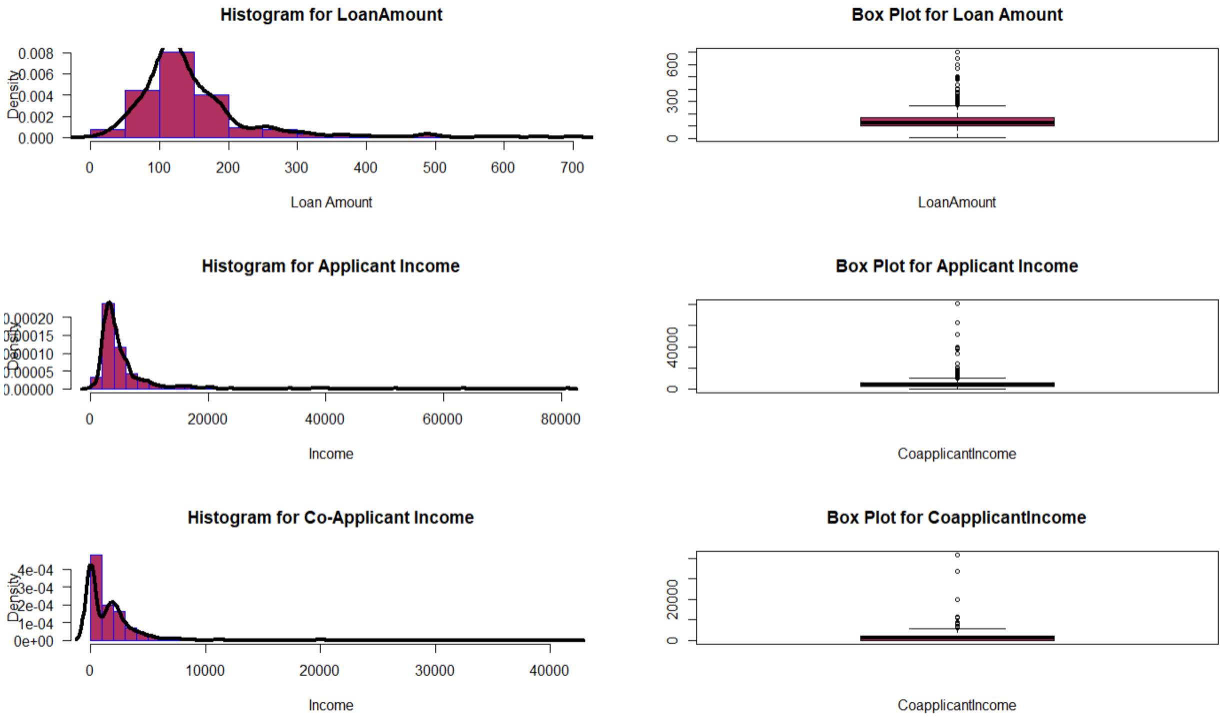
We can see that there are missing values in the Gender, Married, Dependents, Self_Employed, LoanAmount, Loan_Amount_Term, Credit_History category. Below is a distribution of founded missing values from the dataset:



Next, we check for outliers in numerical categories. To do that, we draw the distribution chart of these categories: LoanAmount, ApplicantIncome , CoapplicantIncome:



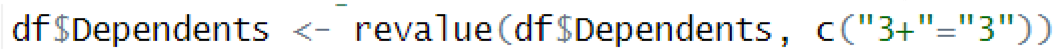
We do the same for ApplicantIncome and CoapplicantIncome:



We can clearly see that there are outliers in these categories, since their distribution charts are right-skewed and the box plot is showing many outliers as well. Therefore, we need to take appropriate steps to fix this data by transforming the data to reduce their impact on the analysis. This will ensure that the results obtained from the analysis are accurate and reliable.

3. **Data cleaning and transformation**
    a. **Handling missing values**
    Categorical Values

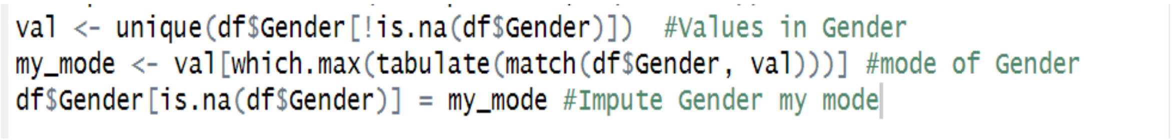
Currently the values in Dependents are categorical; however the number of dependents of a person should be converted to numerical. Therefore we alter the only non-numerical value in that column “3+” to “3” with the revalue() function:



```
df$Dependents <- revalue(df$Dependents, c("3+"="3"))
```

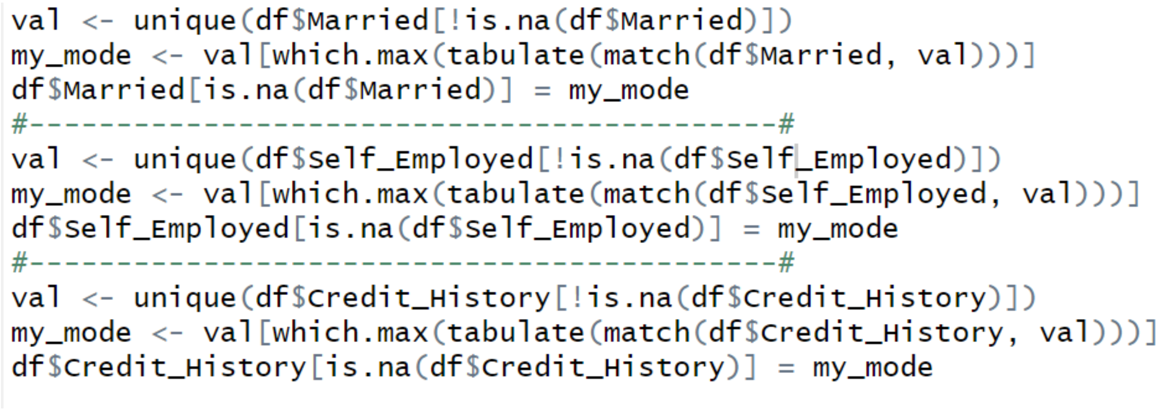
The values in Gender, Married, Self_Employed, Credit_History are categorical, while the values in Dependents LoanAmount and Loan_Amount_Term are numerical.

For the categorical values, we impute the missing data by mode. Since R doesn’t have a built-in function to find and impute using the mode, we have to code it ourselves.



```
val <- unique(df$Gender[!is.na(df$Gender)])  #Values in Gender
my_mode <- val[which.max(tabulate(match(df$Gender, val)))] #mode of Gender
df$Gender[is.na(df$Gender)] = my_mode #Impute Gender my mode
```

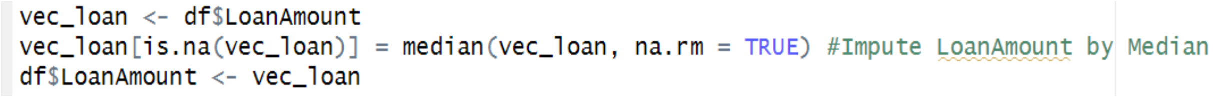
Use the same three functions for the other categorical values in order to achieve the desired result:



```
val <- unique(df$Married[!is.na(df$Married)])
my_mode <- val[which.max(tabulate(match(df$Married, val)))]
df$Married[is.na(df$Married)] = my_mode
#------------------------------------------#
val <- unique(df$Self_Employed[!is.na(df$Self_Employed)])
my_mode <- val[which.max(tabulate(match(df$Self_Employed, val)))]
df$Self_Employed[is.na(df$Self_Employed)] = my_mode
#------------------------------------------#
val <- unique(df$Credit_History[!is.na(df$Credit_History)])
my_mode <- val[which.max(tabulate(match(df$Credit_History, val)))]
df$Credit_History[is.na(df$Credit_History)] = my_mode
```

Numerical Values

For the numerical values, we impute the missing data by median.



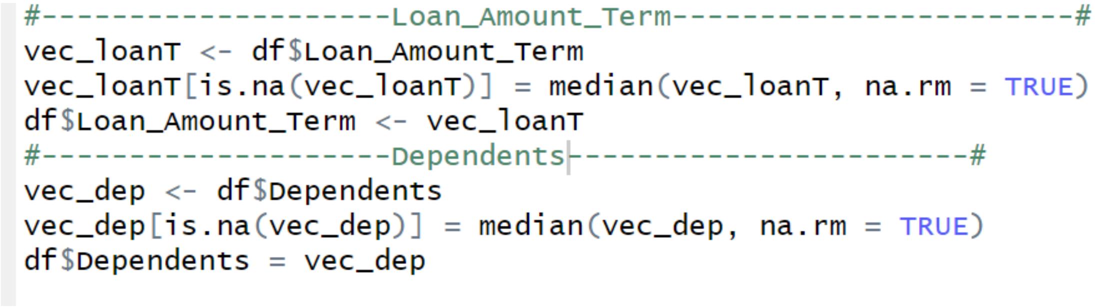
```
vec_loan <- df$LoanAmount
vec_loan[is.na(vec_loan)] = median(vec_loan, na.rm = TRUE) #Impute LoanAmount by Median
df$LoanAmount <- vec_loan
```

Line 1: Transfer the values into a variable “vec_loan”.

Line 2: Replace all N/A values in “vec_loan” with the median in the column which is calculated using the median() function (with parameter na.rm = TRUE to ignore all N/A values while calculating).

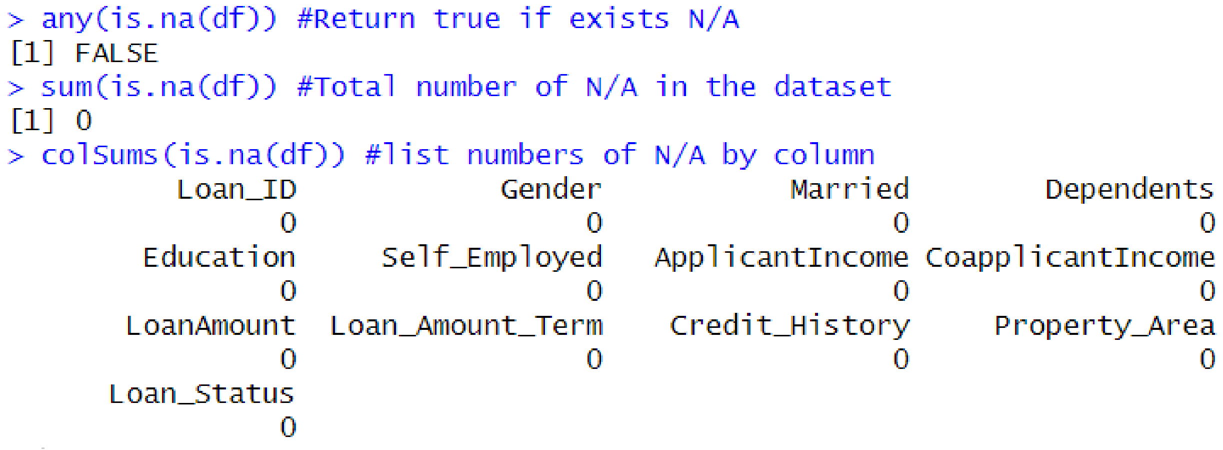
Line 3: Transfer the imputed values into our original table.

The same thing is done on the other two numerical columns



```
#--------------------Loan_Amount_Term-----------------------#
vec_loanT <- df$Loan_Amount_Term
vec_loanT[is.na(vec_loanT)] = median(vec_loanT, na.rm = TRUE)
df$Loan_Amount_Term <- vec_loanT
#--------------------Dependents-----------------------#
vec_dep <- df$Dependents
vec_dep[is.na(vec_dep)] = median(vec_dep, na.rm = TRUE)
df$Dependents = vec_dep
```

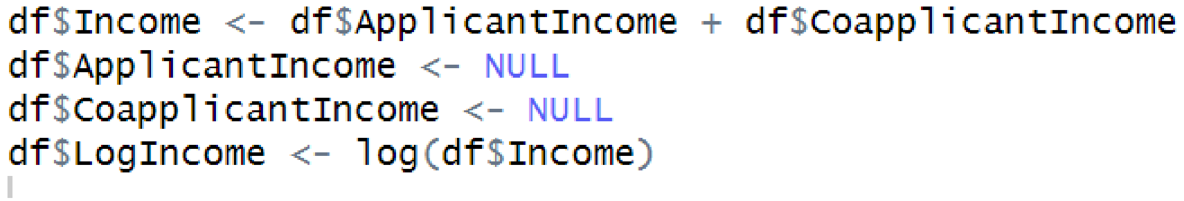
After handling everything, check if we have imputed all N/A values:



```
> any(is.na(df)) #Return true if exists N/A
[1] FALSE
> sum(is.na(df)) #Total number of N/A in the dataset
[1] 0
> colSums(is.na(df)) #list numbers of N/A by column
          Loan_ID            Gender           Married        Dependents 
                0                 0                 0                 0 
        Education     Self_Employed   ApplicantIncome CoapplicantIncome 
                0                 0                 0                 0 
       LoanAmount  Loan_Amount_Term    Credit_History     Property_Area 
                0                 0                 0                 0 
      Loan_Status 
                0 
```

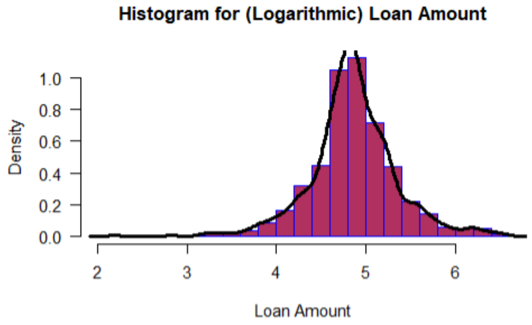
b. Outlier treatment

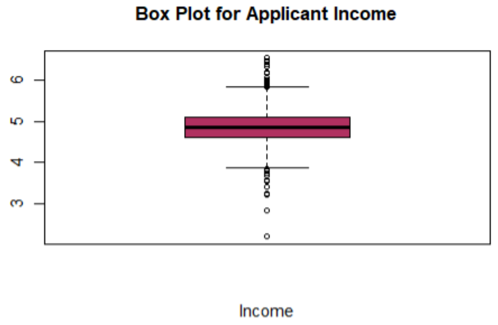
We apply the log transformation for LoanAmount, ApplicantIncome and CoapplicantIncome for outlier treatment. For ApplicantIncome and CoapplicantIncome, we can combine both of them as total income and then perform log transformation of the combined variable:

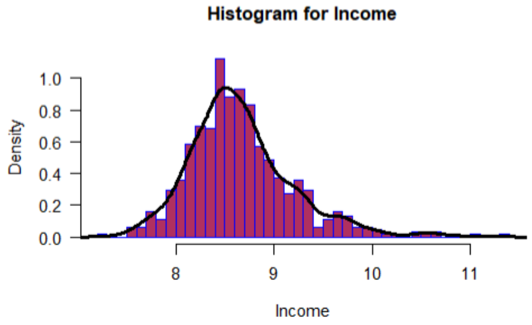


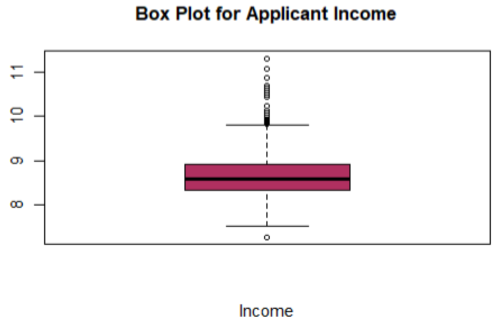
```
df$Income <- df$ApplicantIncome + df$CoapplicantIncome
df$ApplicantIncome <- NULL
df$CoapplicantIncome <- NULL
df$LogIncome <- log(df$Income)
```

We plot the distribution chart of the transformed variables:









After transformation, the distribution of these variables resembles more of a normal distribution.

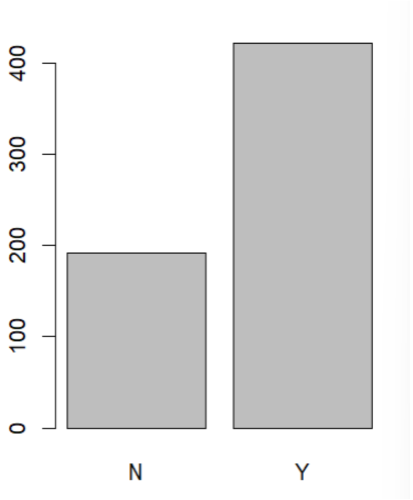
4. Data statistics

a. Univariate analysis

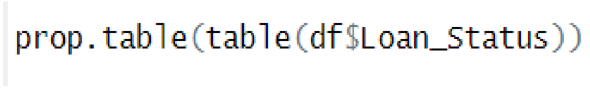
By analyzing each variable in the dataset individually, Univariate Analysis helps in understanding the distribution, central tendency, and spread of each variable. Additionally, Univariate Analysis helps in identifying the key features of the dataset that are most relevant to the prediction of loan eligibility.
We will collect and show the data of Loan_Status by using bar charts by the following function:

```
barplot(table(df$Loan_Status))
```

This function shows the numbers of Loan_Status:

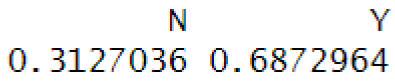


Next, we calculate the percentage of Loan_Status by the following function:



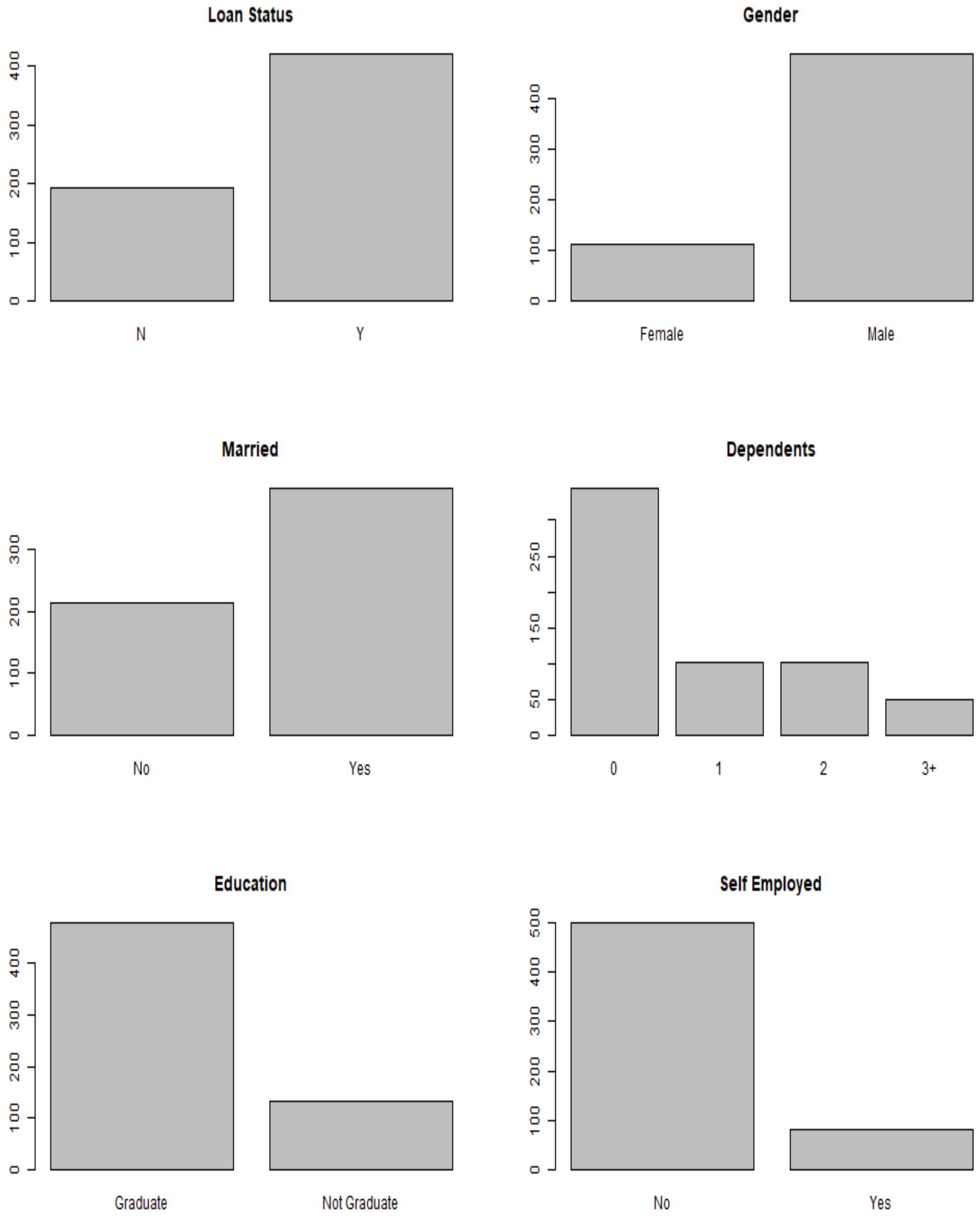
```
prop.table(table(df$Loan_Status))
```

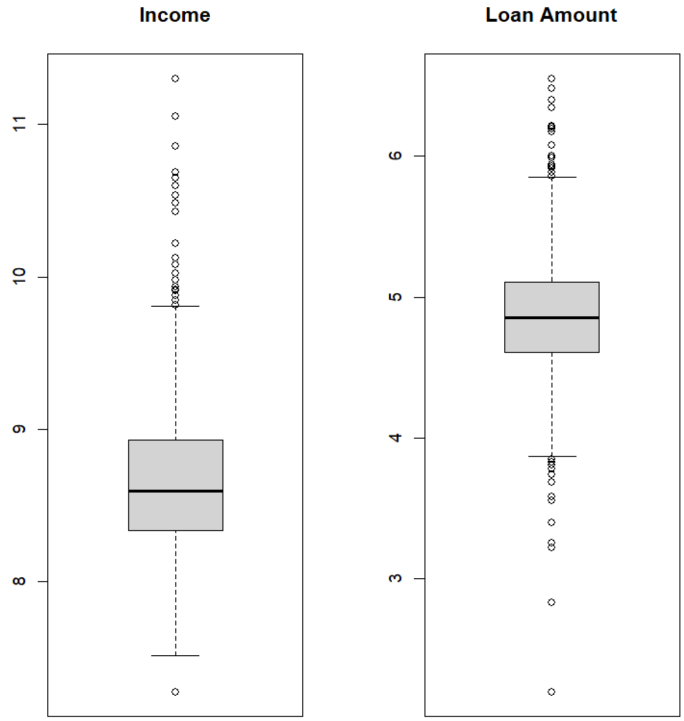
The previous function will show the percentage of Loan_Status in the data set.



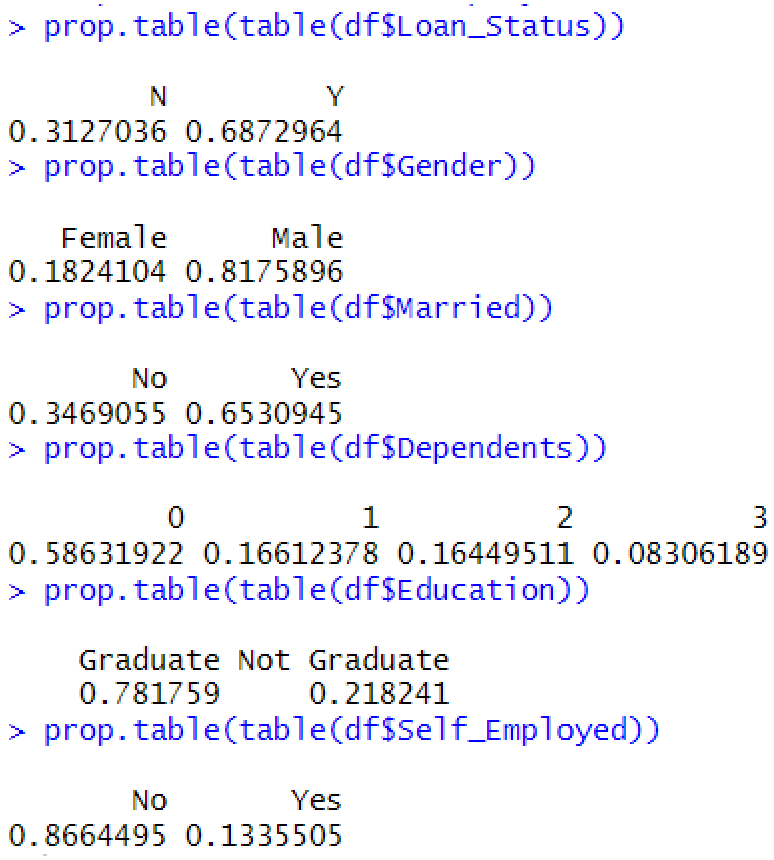
```
        N         Y 
0.3127036 0.6872964 
```

Using those same functions, we have the following charts:


Loan Status
0
100
200
300
400
N
Y
Gender
0
100
200
300
400
Female
Male
Married
0
100
200
300
No
Yes
Dependents
0
50
150
250
0
1
2
3+
Education
0
100
200
300
400
Graduate
Not Graduate
Self Employed
0
100
200
300
400
500
No
Yes



The percentage of the above categories:

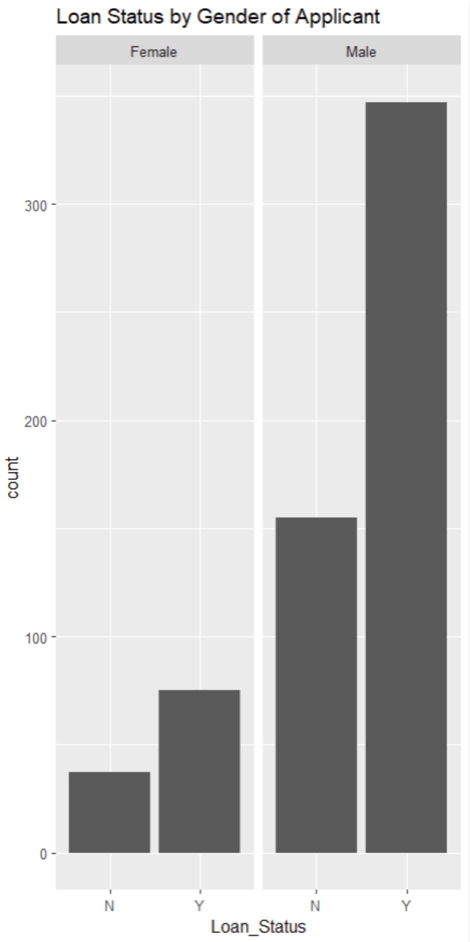


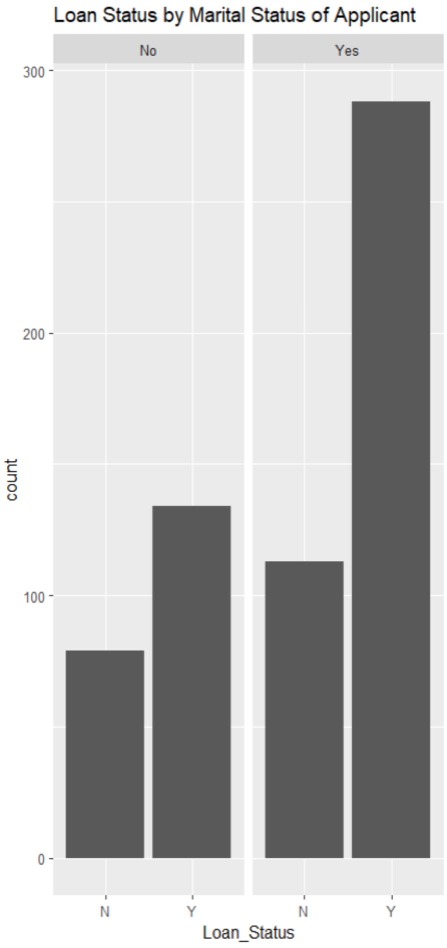
```
> prop.table(table(df$Loan_Status))

        N         Y 
0.3127036 0.6872964 
> prop.table(table(df$Gender))

   Female      Male 
0.1824104 0.8175896 
> prop.table(table(df$Married))

       No       Yes 
0.3469055 0.6530945 
> prop.table(table(df$Dependents))

         0          1          2          3 
0.58631922 0.16612378 0.16449511 0.08306189 
> prop.table(table(df$Education))

    Graduate Not Graduate 
    0.781759     0.218241 
> prop.table(table(df$Self_Employed))

       No       Yes 
0.8664495 0.1335505 
```

b. Bivariate analysis

By analyzing the relationship between two variables or more, this helps in understanding how one variable affects the other and helps in identifying any patterns or trends in the data.

Bivariate Analysis is particularly useful in identifying the most relevant features for predicting loan eligibility, as it helps in identifying which variables are most strongly correlated with the target variable.



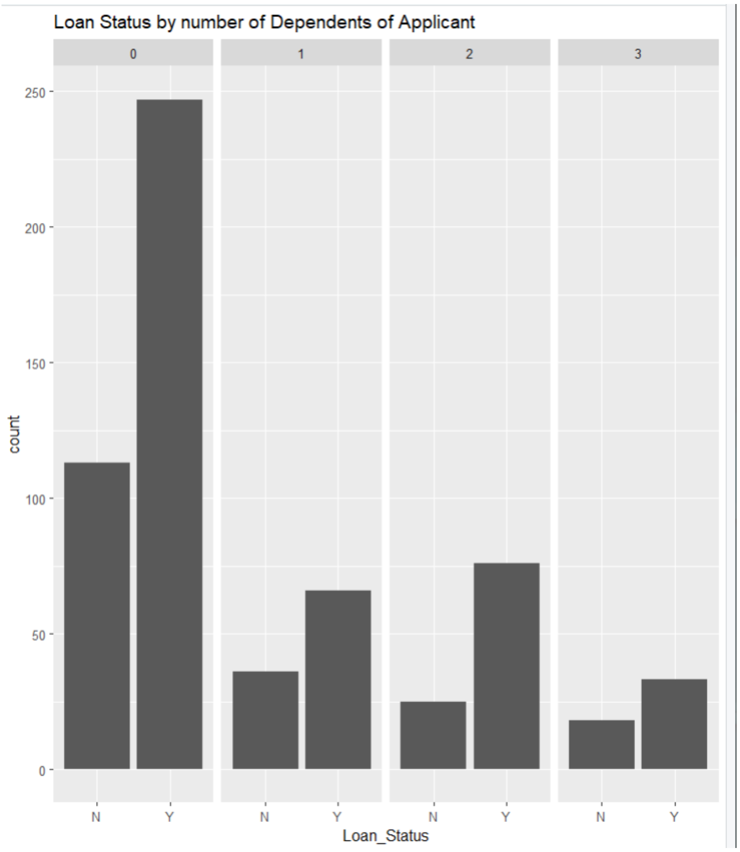


Insight 1:

We found that a larger proportion of unmarried applicants are refused by lenders compared to married applicants.

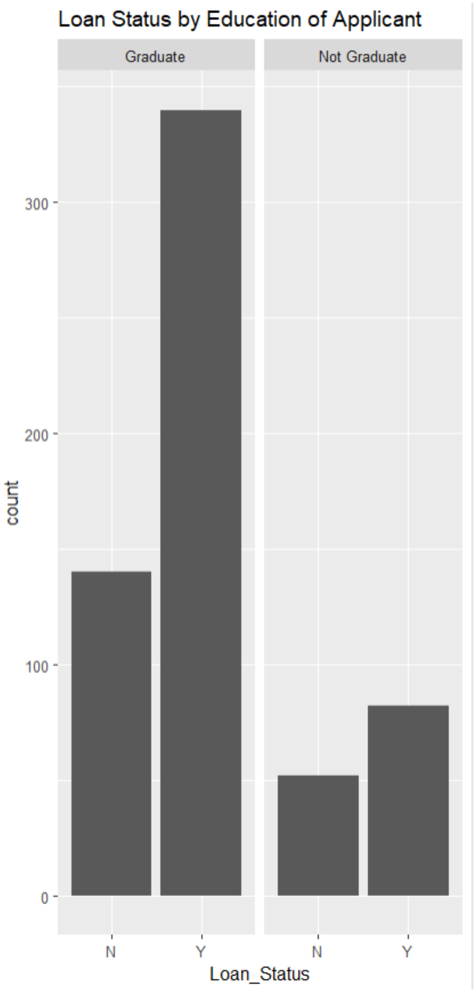
We also found that there is a weak correlation between Gender and Loan_Status. Specifically, we found that a slightly higher proportion of male applicants are approved for loans compared to female applicants. However, this correlation is weak and may not be statistically significant.

However, it is important to note that Gender should not be used as a sole factor in determining loan eligibility, as this could lead to biased and unfair output.

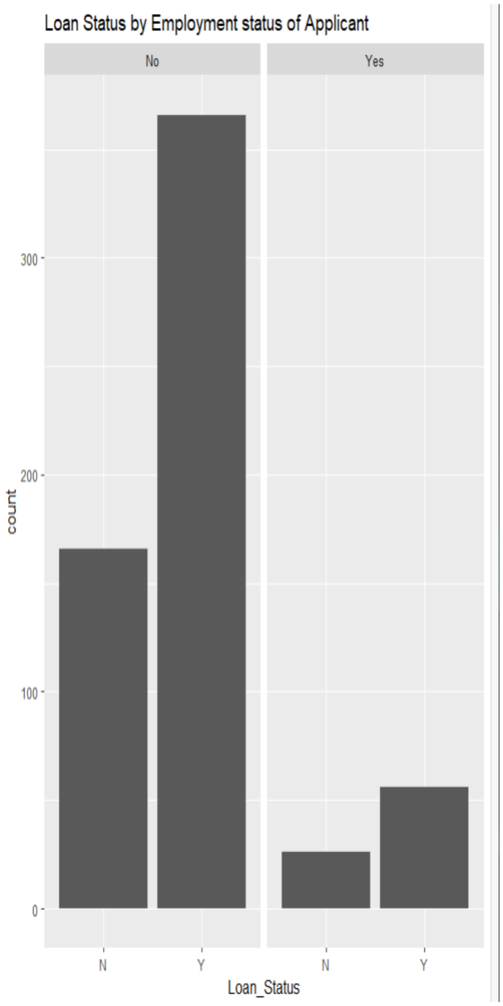


Insight 2:

We found that applicants with 2 dependents have a higher likelihood of loan approval compared to those with other numbers of dependents.



Loan Status by Education of Applicant



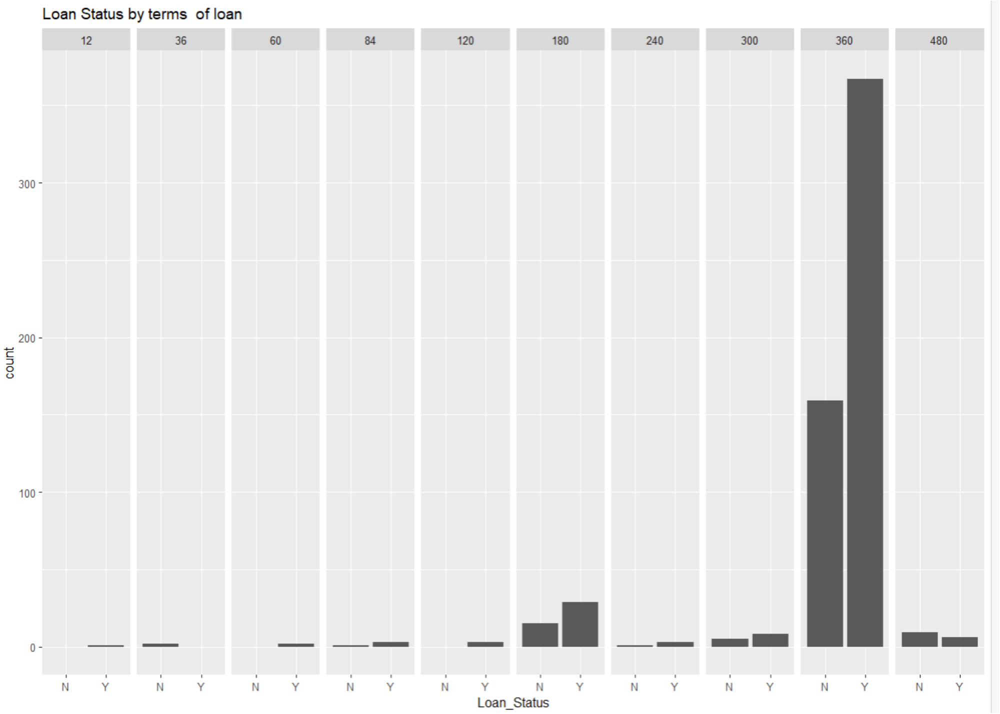
Loan Status by Employment status of Applicant

Insight 3:

We found that a larger proportion of non-graduate applicants are refused loans compared to graduate applicants.

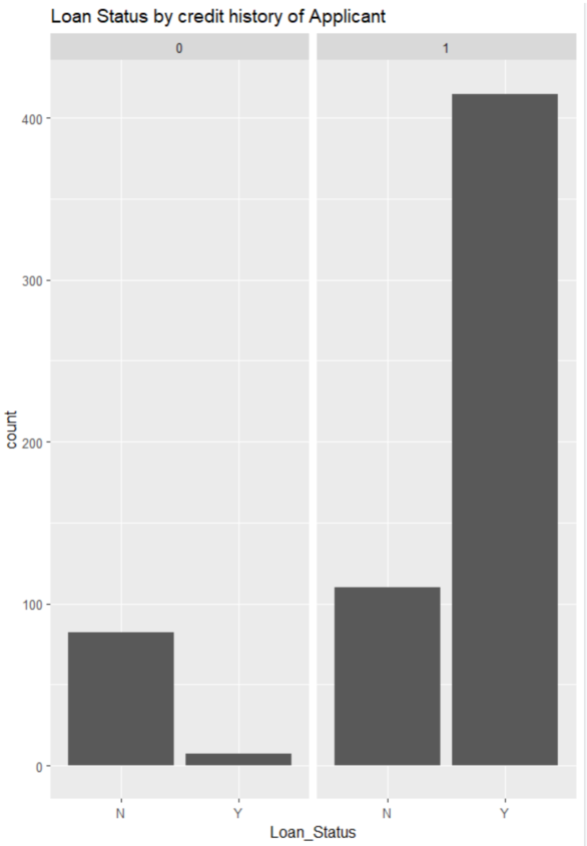
We also found that there is a slightly higher likelihood of loan approval for applicants who are not self-employed compared to those who are self-employed.

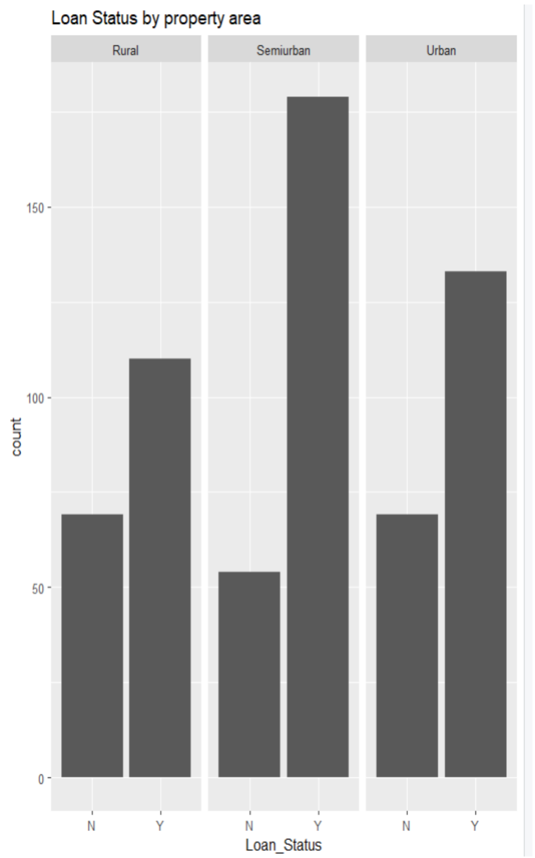
Additionally, these findings highlight the need for further research into the factors that influence Loan_Status and the potential impact of education level and Employment status on Loan_Status.



Insight 4:

We found that most loans have a term of 360 months, making it difficult to identify any clear patterns in loan duration.



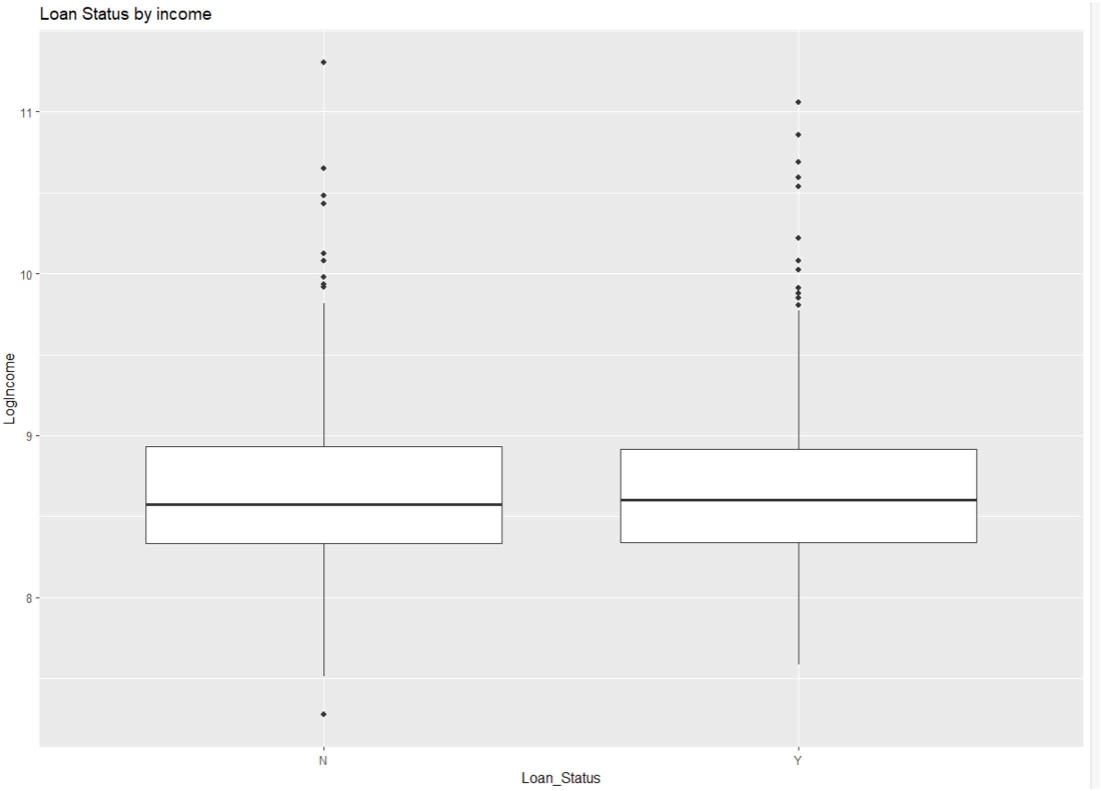


Insight 5:

Another important finding from our analysis is that almost all applicants with a credit history of 0 were refused loans. This suggests that credit history is a critical factor in determining loan eligibility, and applicants with no credit history may face additional challenges in obtaining loans.

Another finding is that loan approval rates vary depending on the type of property. Specifically, it is easiest to obtain a loan for a property in a semi-urban area, while it is hardest to obtain a loan for a property in a rural area.

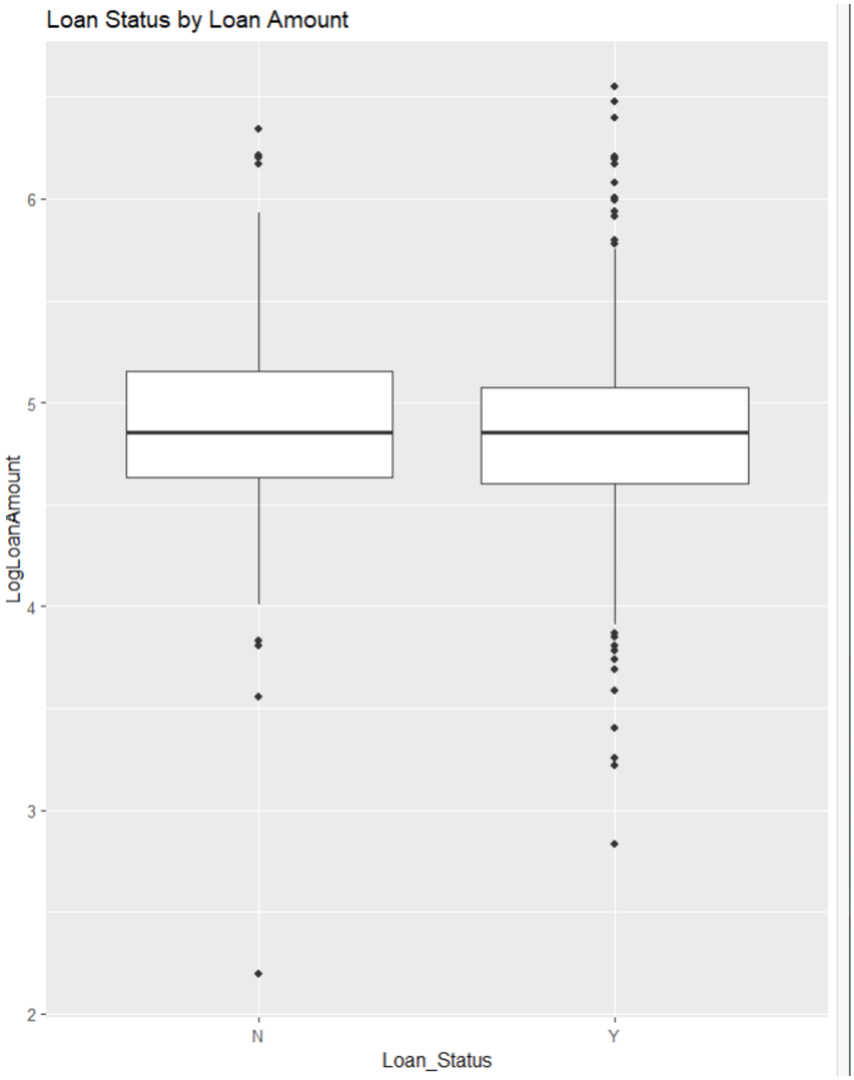
These findings highlight the need for further research into the factors that influence loan eligibility and the potential impact of loan terms, credit history, and geographic location on our problem.



Insight 6:

We found that there is little difference in loan approval rates between different income groups, as suggested by the Loan Status by Applicant Income boxplot.

This finding suggests that income level may not be the sole factor in determining loan eligibility, and that other factors such as credit history or employment status may also be important considerations.



Insight 7:

We found that there is a difference in loan amounts between approved and refused loans, as suggested by the Loan Status by Loan Amount boxplot.

Specifically, we found that the third quartile of refused loans is higher than that of approved loans. This suggests that loan amount may be an important factor in determining loan eligibility, and that applicants who request larger loan amounts may face additional challenges in obtaining loan approval.

IV. Inferential statistics

1. Hypotheses and insights

Dream Housing Finance consider several factors when evaluating a loan application, including the applicant's income, credit history, loan amount, loan term, and Equated Monthly Installment (EMI). The following hypotheses have been proposed to help applicants increase their chances of loan approval:

- **[1] Higher the income, higher the chances of approval**: The higher the applicant's income, the greater the chances of loan approval. Lenders prefer applicants with high incomes because they are more likely to be able to repay the loan. However, income alone does not guarantee loan approval.
- **[2] Good credit history indicates higher chances of approval:** A good credit history with timely payments of previous loans or credit card bills indicates that the applicant is responsible and reliable, making them more likely to be approved for a loan. Applicants with poor credit history or a history of defaulting on loans may find it difficult to get approved for a loan.
- **[3] Lesser the Loan amount, higher the chances of approval:** The loan amount requested by the applicant is an important factor in determining loan approval. In general, lenders prefer to lend smaller amounts as they entail lower risk. Therefore, if the loan amount is less, the chances of loan approval should be higher.
- **[4] Shorter the Loan term, higher the chances of approval:** The loan term is the length of time over which the loan is to be repaid. A longer loan term may result in lower monthly payments, but it also means the loan will accrue more interest over time. Therefore, lenders may be hesitant to approve loans with longer terms as they present a greater risk.
- **[5] EMI:** The EMI is the amount of money the borrower must pay each month to repay the loan. The lower the EMI, the higher the chances of loan approval. This is because a lower EMI indicates that the borrower will have less difficulty repaying the loan.

2. Logistic regression

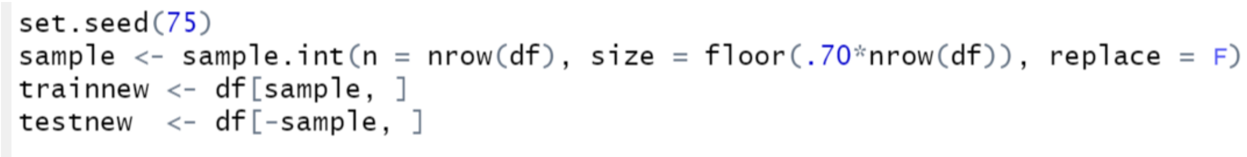
a. Train and Test Dataset Preparation

In order for Modeling, splitting the data into training and test sets is a crucial step in the analysis process. The training set is used to train our models, while the test set is used to validate the models by providing a measure of how well they generalize to new, unseen cases.

Overfitting is a major concern in sparse datasets like ours. Overfitting occurs when the model learns the training set so well that it struggles to handle cases it has never seen before.

To avoid overfitting, we use the test set to score the models and estimate their performance in practice. We treat the test set as if it no longer exists once we split the data, ensuring that we don't unintentionally use it during model training. This helps us to develop more accurate prediction models that can be applied to new, unseen data.

Let us split the data into training and test sets:



```
set.seed(75)
sample <- sample.int(n = nrow(df), size = floor(.70*nrow(df)), replace = F)
trainnew <- df[sample, ]
testnew  <- df[-sample, ]
```

Line 1: This sets the seed for the random number generator, which ensures that the same random numbers are generated each time the code is run.

Line 2: This line of code randomly samples 70% of the rows from the df without replacement and assigns them to the variable sample. The nrow(df) function returns the number of rows in the dataset, and floor(.70*nrow(df)) calculates 70% of the total number of rows and rounds it down to the nearest integer.

Line 3: This line of code creates a new dataset called trainnew that contains only the rows in df that were selected by the sample variable. These rows are used for training the machine learning model.

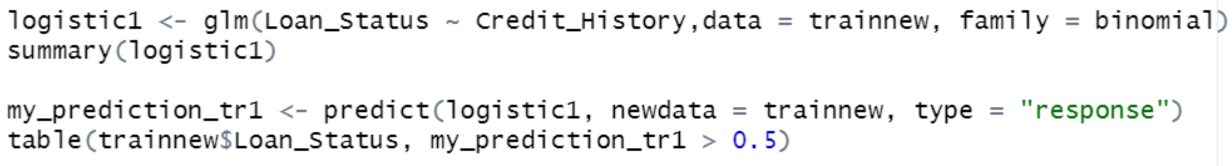
b. The impact of Credit History to Loan Status

In constructing our first logistic regression model, it is important to avoid overfitting the data by carefully selecting the variables to include. Based on our hypotheses, the following variables appear to be influential in predicting loan approval:

- Credit_History: Applicants who have a history of taking loans are more likely to have their loan approved.
- LogIncome: Applicants with higher incomes have a greater likelihood of being approved for a loan.
- Education: Applicants with higher levels of education are more likely to be approved for a loan.
- Self_Employed: Applicants who have a stable job are more likely to have their loan approved.

For our initial logistic regression model, we will focus on the Credit_History variable as it appears to be a critical factor in determining loan approval.

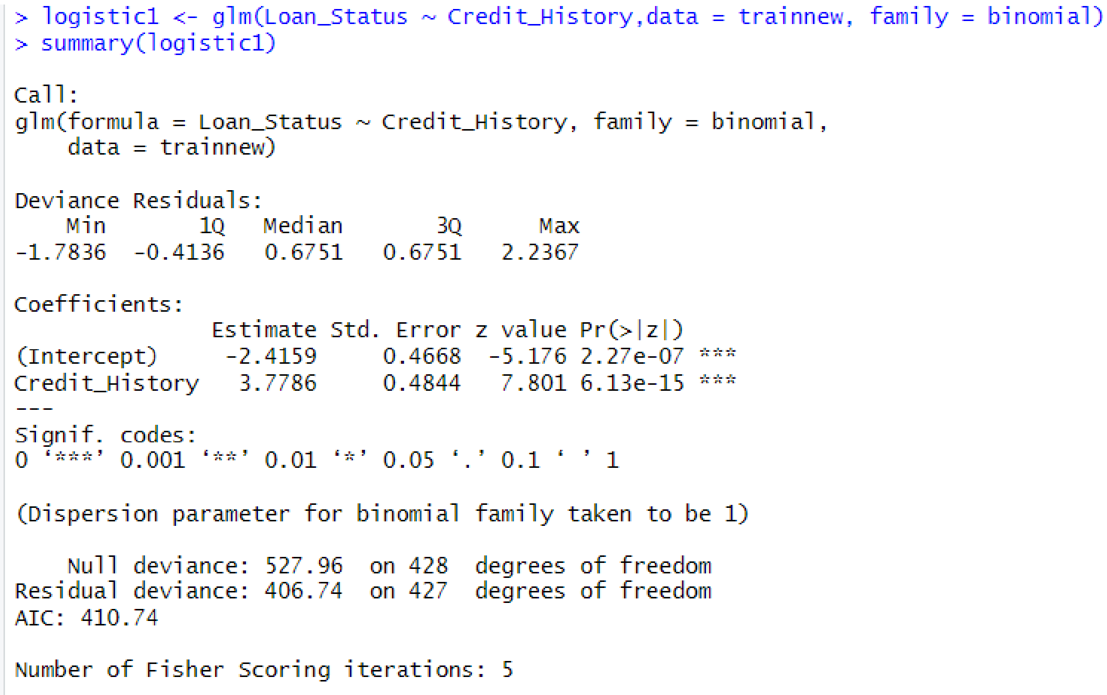
Training Set:



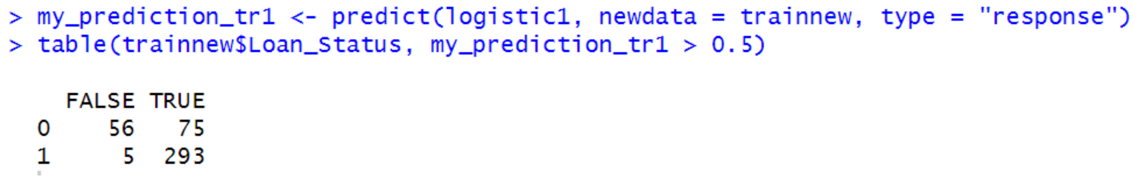
```
logistic1 <- glm(Loan_Status ~ Credit_History,data = trainnew, family = binomial)
summary(logistic1)

my_prediction_tr1 <- predict(logistic1, newdata = trainnew, type = "response")
table(trainnew$Loan_Status, my_prediction_tr1 > 0.5)
```

The above code is used to run a logistic regression model on the trainnew dataset, where we are trying to predict the Loan_Status based on the Credit_History variable.

Output:

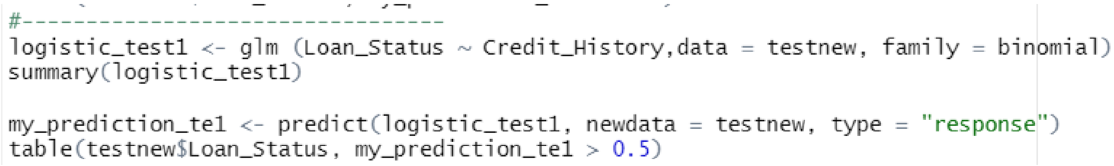


```
> logistic1 <- glm(Loan_Status ~ Credit_History,data = trainnew, family = binomial)
> summary(logistic1)

Call:
glm(formula = Loan_Status ~ Credit_History, family = binomial,
    data = trainnew)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7836  -0.4136   0.6751   0.6751   2.2367

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.4159     0.4668  -5.176 2.27e-07 ***
Credit_History   3.7786     0.4844   7.801 6.13e-15 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 527.96  on 428  degrees of freedom
Residual deviance: 406.74  on 427  degrees of freedom
AIC: 410.74

Number of Fisher Scoring iterations: 5
```



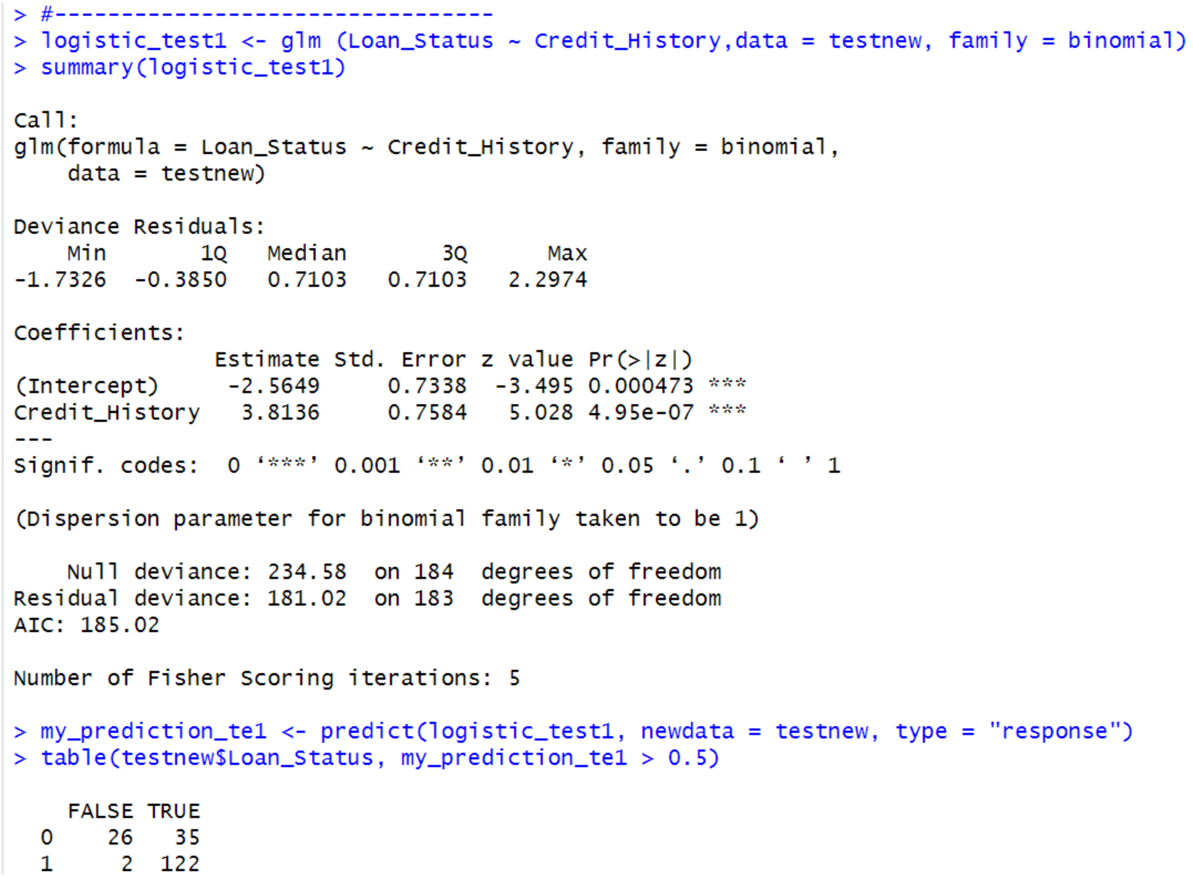
```
> my_prediction_tr1 <- predict(logistic1, newdata = trainnew, type = "response")
> table(trainnew$Loan_Status, my_prediction_tr1 > 0.5)

    FALSE TRUE
  0    56   75
  1     5  293
```

Test dataset:



```
#---------------------------------
logistic_test1 <- glm (Loan_Status ~ Credit_History,data = testnew, family = binomial)
summary(logistic_test1)

my_prediction_te1 <- predict(logistic_test1, newdata = testnew, type = "response")
table(testnew$Loan_Status, my_prediction_te1 > 0.5)
```

Output:



```
> #---------------------------------
> logistic_test1 <- glm (Loan_Status ~ Credit_History,data = testnew, family = binomial)
> summary(logistic_test1)

Call:
glm(formula = Loan_Status ~ Credit_History, family = binomial, 
    data = testnew)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7326  -0.3850   0.7103   0.7103   2.2974  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -2.5649     0.7338  -3.495 0.000473 ***
Credit_History   3.8136     0.7584   5.028 4.95e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.58  on 184  degrees of freedom
Residual deviance: 181.02  on 183  degrees of freedom
AIC: 185.02

Number of Fisher Scoring iterations: 5

> my_prediction_te1 <- predict(logistic_test1, newdata = testnew, type = "response")
> table(testnew$Loan_Status, my_prediction_te1 > 0.5)
   
    FALSE TRUE
  0    26   35
  1     2  122
```

The logistic regression output shows the model's call, which reminds us of the model we've run. Then, the deviance residuals are displayed to measure the model's fit. The coefficients, standard errors, z-statistic, and associated p-values are presented.

The p-value tests the null hypothesis that the coefficient is zero, indicating no effect. A low p-value (< 0.05) suggests the predictor is significant, while a higher p-value indicates it's insignificant. In our case, the p-value for Credit_History is notably small, signifying its significance in the model.

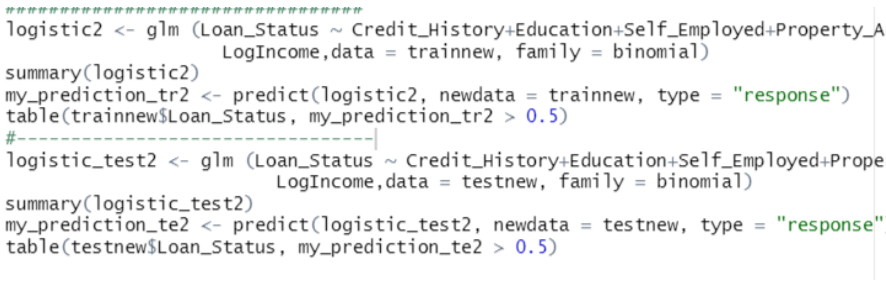
To evaluate the model’s accuracy, we have generated a confusion table for both the training and test data:

- Train data: 79.62% accuracy
- Test data: 77.77% accuracy

c. Incorporating additional variables and investigating the correlation to the Loan Status

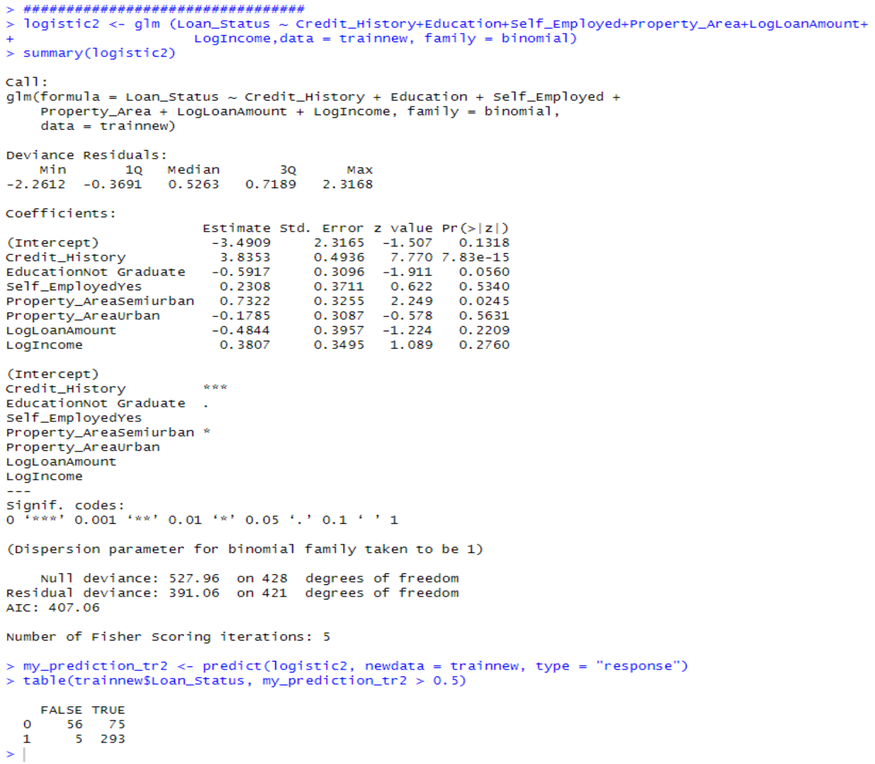
To further improve the model, we can incorporate additional variables and assess their impact on accuracy.

Training Set:



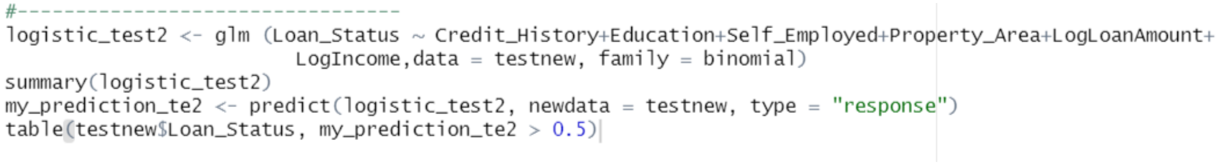
```
#################################
logistic2 <- glm (Loan_Status ~ Credit_History+Education+Self_Employed+Property_Area+LogLoanAmount+
                    LogIncome,data = trainnew, family = binomial)
summary(logistic2)
my_prediction_tr2 <- predict(logistic2, newdata = trainnew, type = "response")
table(trainnew$Loan_Status, my_prediction_tr2 > 0.5)
#-------------------------------|
logistic_test2 <- glm (Loan_Status ~ Credit_History+Education+Self_Employed+Property_Area+LogLoanAmount+
                         LogIncome,data = testnew, family = binomial)
summary(logistic_test2)
my_prediction_te2 <- predict(logistic_test2, newdata = testnew, type = "response")
table(testnew$Loan_Status, my_prediction_te2 > 0.5)
```

Output:



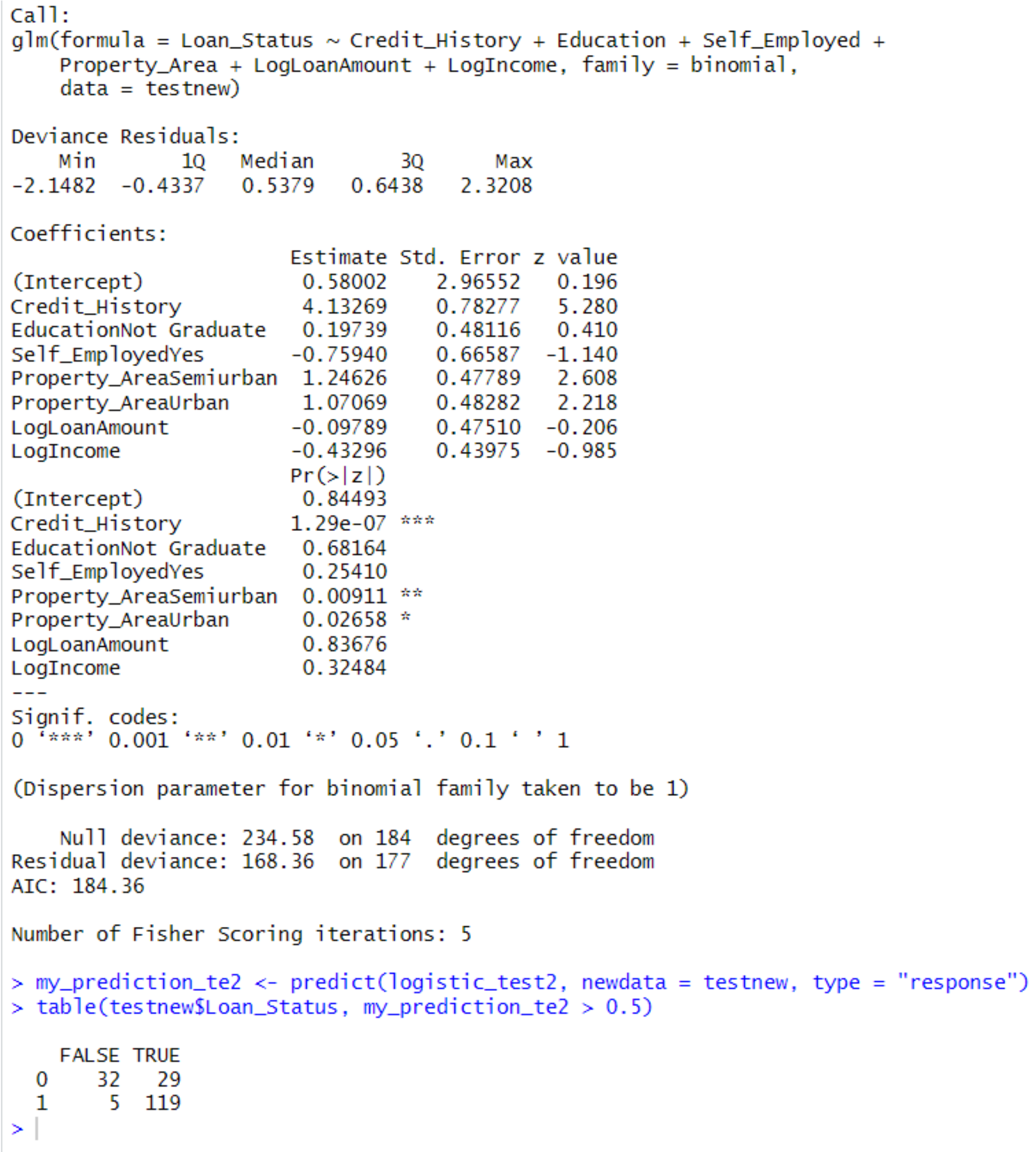
```
> #################################
> logistic2 <- glm (Loan_Status ~ Credit_History+Education+Self_Employed+Property_Area+LogLoanAmount+
+                     LogIncome,data = trainnew, family = binomial)
> summary(logistic2)

Call:
glm(formula = Loan_Status ~ Credit_History + Education + Self_Employed + 
    Property_Area + LogLoanAmount + LogIncome, family = binomial, 
    data = trainnew)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2612  -0.3691   0.5263   0.7189   2.3168  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -3.4909     2.3165  -1.507   0.1318
Credit_History          3.8353     0.4936   7.770 7.83e-15
EducationNot Graduate  -0.5917     0.3096  -1.911   0.0560
Self_EmployedYes        0.2308     0.3711   0.622   0.5340
Property_AreaSemiurban  0.7322     0.3255   2.249   0.0245
Property_AreaUrban     -0.1785     0.3087  -0.578   0.5631
LogLoanAmount          -0.4844     0.3957  -1.224   0.2209
LogIncome               0.3807     0.3495   1.089   0.2760
                         
(Intercept)              
Credit_History        ***
EducationNot Graduate .  
Self_EmployedYes         
Property_AreaSemiurban * 
Property_AreaUrban       
LogLoanAmount            
LogIncome                
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 527.96  on 428  degrees of freedom
Residual deviance: 391.06  on 421  degrees of freedom
AIC: 407.06

Number of Fisher Scoring iterations: 5

> my_prediction_tr2 <- predict(logistic2, newdata = trainnew, type = "response")
> table(trainnew$Loan_Status, my_prediction_tr2 > 0.5)
   
    FALSE TRUE
  0    56   75
  1     5  293
> |
```

Test dataset:



```
#-------------------------------
logistic_test2 <- glm (Loan_Status ~ Credit_History+Education+Self_Employed+Property_Area+LogLoanAmount+
                         LogIncome,data = testnew, family = binomial)
summary(logistic_test2)
my_prediction_te2 <- predict(logistic_test2, newdata = testnew, type = "response")
table(testnew$Loan_Status, my_prediction_te2 > 0.5)|
```

Output:



```
Call:
glm(formula = Loan_Status ~ Credit_History + Education + Self_Employed +
    Property_Area + LogLoanAmount + LogIncome, family = binomial,
    data = testnew)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1482  -0.4337   0.5379   0.6438   2.3208

Coefficients:
                       Estimate Std. Error z value
(Intercept)             0.58002    2.96552   0.196
Credit_History          4.13269    0.78277   5.280
EducationNot Graduate   0.19739    0.48116   0.410
Self_EmployedYes       -0.75940    0.66587  -1.140
Property_AreaSemiurban  1.24626    0.47789   2.608
Property_AreaUrban      1.07069    0.48282   2.218
LogLoanAmount          -0.09789    0.47510  -0.206
LogIncome              -0.43296    0.43975  -0.985
                       Pr(>|z|)
(Intercept)             0.84493
Credit_History         1.29e-07 ***
EducationNot Graduate   0.68164
Self_EmployedYes        0.25410
Property_AreaSemiurban  0.00911 **
Property_AreaUrban      0.02658 *
LogLoanAmount           0.83676
LogIncome               0.32484
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.58  on 184  degrees of freedom
Residual deviance: 168.36  on 177  degrees of freedom
AIC: 184.36

Number of Fisher Scoring iterations: 5

> my_prediction_te2 <- predict(logistic_test2, newdata = testnew, type = "response")
> table(testnew$Loan_Status, my_prediction_te2 > 0.5)

    FALSE TRUE
  0    32   29
  1     5  119
> 
```

- Train data: 88.73%
- Test data: 89.53%

The accuracy of the test set has increased by including more relevant predictors in the model, indicating the importance of incorporating such variables to enhance the model's predictive ability on new data.

3. Summary

Throughout the model suggestion, Credit_History appears to be the most important factor in determining loan approval. It also strengthens the hypothesis [2] (that **Good credit history indicates higher chances of approval).**

By comparing the Credit_History only and the Credit_History and Other factors output, we know that the other factors also put an emphasis on the prediction of the loan eligibility for clients.

In addition, from insights we gained from Descriptive Statistics, we also have:

- **Hypothesis 1 ("Higher the income, higher the chances of approval")** is not supported by Insight 6, which found that there is little difference in loan approval rates between different income groups. However, it is important to note that income alone does not guarantee loan approval, as stated in the hypothesis.

- **Hypothesis 2 ("Good credit history indicates higher chances of approval")** is supported by Insight 5, which found that almost all applicants with a credit history of 0 were refused loans.

- **Hypothesis 3 ("Lesser the Loan amount, higher the chances of approval")** is partially supported by Insight 7, which found that applicants who request larger loan amounts may face additional challenges in obtaining loan approval.

- **Hypothesis 4 ("Shorter the Loan term, higher the chances of approval")** is not supported by any of the insights provided, and the chart indicates that there are but little correlation between the Loan term and chances of approval.

- **Hypothesis 5 ("EMI: The lower the EMI, the higher the chances of loan approval")** is not directly supported by any of the insights provided, but it is reasonable to assume that a lower EMI could be an indication of a smaller loan amount, which could increase the chances of loan approval, as suggested by Insight 7.

V. Discussion and Extension

In this part, we will discuss about pros and cons of the models and test úed in the report:

- About Hypothesis testing:

Pros	Cons
● They provide an accepted convention for statistical analysis ● The techniques are tried and tested ● The alternative hypothesis can be rather vague ● They reflect the same underlying statistical reasoning as confidence intervals	● They are commonly misunderstood and misinterpreted ● Use of a rigid 0.05 level forces a false dichotomy into significant or not significant. ● The *P*-value is uninformative compared to the confidence interval ● The null hypothesis is nearly always false ● The one-tailed or two-tailed decision is profoundly subjective

	● *P*-values take no account of any hypothesis other than the null. ● *P*-values include all values more extreme than the observed result ● The null distribution of the test statistic may not match the actual sampling distribution of the test statistic

Advantages of Hypothesis testing:

- **They provide an accepted convention for statistical analysis.**

It is valuable to have a common approach across different disciplines for analysing data and testing hypotheses. For example, in the field of epidemiology it has been argued that epidemiologists need to agree by consensus on prespecified criteria so that the basis for decisions is explicit. The conventions of significance testing (such as the 0.05 level for significance) then provide a reasonable basis for facilitating scientific decision making. Null hypothesis significance tests are still widely used, and are often insisted upon by referees and journal editors.

- **The techniques are tried and tested.**

Appropriate tests have been devised for a variety of statistics, statistical techniques and statistical models - including many 'pre-cooked' experimental and sampling designs. Formulae and software packages are readily available, as is copious documentation.

- **The alternative hypothesis can be rather vague.**

Although the null model has to be specified with some care, the alternate model can be relatively hazy. This has its down side as well, but some may see it as a plus.

- **They reflect the same underlying statistical reasoning as confidence intervals**

Significance tests and confidence intervals are in fact based on exactly the same underlying theory. Tests not only shed light upon confidence intervals, but also enable some of the more awkward ones to be estimated.

Disadvantages of hypothesis testing:

- **They are commonly misunderstood and misinterpreted**

The main misinterpretations are:

a. A high value of *P* is taken as evidence in favor of the null hypothesis, or worse as proof of the null hypothesis. This is wrong because the *P*-value is not equal to the probability that the null hypothesis is true. It is only a measure of the degree of consistency of the data with the null hypothesis - and a very poor measure at that if the sample size is small!
b. A low value of *P* is taken as evidence in favor of the alternative hypothesis, or worse as proof of the alternative hypothesis. This is also wrong because the *P*-value does not tell you anything directly about your chosen alternative hypothesis. It only tells us about the degree of consistency of

the data with the null hypothesis. In many situations there may be other alternative hypotheses that you have not considered. As above there is also the problem of reliability if the sample size is small.

c. If in one trial the null hypothesis is rejected at *P* = 0.05, it is thought that repeating the experiment many times will produce a significant result on 95% of occasions. Again this is wrong and is known as the 'replication fallacy'. In fact for the usual levels of power in ecological and veterinary research (< 0.5) , repetition is unlikely to produce a significant result on even 50% of occasions.

- **Use of a rigid 0.05 level forces a false dichotomy into significant or not significant.**

The P = 0.05 syndrome is characterized by a slavish adherence to comparing a P-value - that is subject to sampling error like any other statistic - to a fixed significance level - that is entirely arbitrary. If the sample size is small, the null hypothesis is accepted too readily. If the sample size is large, then unimportant differences are accepted. For example, Nester (1996) commented that because (most) biologists always want important differences to be significant and unimportant differences to be non-significant, the biologist is therefore reduced to one of following states of mind:

How biologist view significant tests		
Importance of observed difference	Statistical significance of difference	
	Not significant	Significant
Not important	Happy	Annoyed
Important	Frustrated	Elated

What the biologist should be doing instead is interpreting the result in the light of the experimental design (designs differ in the strength of inference possible from the results) and other research results. In other words they should be thinking about their results!

- **The *P*-value is uninformative compared to the confidence interval**

Most journals now balk at accepting 'naked' *P*-values - in other words where neither the size of the effect, nor its precision are specified. There is a strong case to be made for always estimating the magnitude and the precision of the effect (using the confidence interval or better still the *P*-value function) along with the precise *P*-value. Confidence intervals are not fundamentally different from *P*-values - but they do provide useful additional information.

Unfortunately many (if not most) researchers who use confidence intervals only see them as surrogate null hypothesis significance tests. In other words, if the interval overlaps zero (for a difference) or one (for a ratio), then the effect is dismissed as non-significant - and one is no further forward in a rational approach to evidence. Confidence intervals should be seen as providing additional information on which to base your inferences and conclusions.

- **The null hypothesis is nearly always false.**

It is true to say that nearly all null hypotheses are false on a priori grounds, at least for measurement variables. For example, if we are putting fertilizer on a crop, it would be very surprising if that fertilizer had no effect at all. In this situation we have no interest in disproving the null hypothesis that fertilizer has no effect - what we are interested in is the size of the effect.

Unfortunately the journals are full of tests based on obviously false null hypotheses, such as one spotted by Johnson (1999): "the density of large trees was greater in unlogged forest stands than in logged stands (P = 0.02).". It would indeed be truly amazing if there were no difference! If one looks at it and says - this cannot possibly be true - then it is probably a waste of time trying to disprove it

- **The one-tailed or two-tailed decision is profoundly subjective**

Nearly all statistics textbooks give the usual bland explanation that whether one uses a one-tailed or two-tailed test depends on your initial hypotheses. If a difference is only considered possible in one direction, then it is considered legitimate to use a one-tailed test which halves the *P*-value. But in practice it seems that one-tailed tests are mainly used to push the *P*-value below the magic 0.05 level. These issues are generally avoided in medical research by a (largely unspoken) convention to always use two tailed tests. But occasionally, for good reason, that convention is broken - and invariably leads to disputes in journals.

- ***P*-values take no account of any hypothesis other than the null.**

This is the first of the more fundamental objections to significance tests. If an observation is rare under the null hypothesis, does it necessarily mean we should accept the alternative. Improbable events do happen - people do actually win the lottery on occasions. Do we therefore assume that the lottery has been 'fixed' because an improbable event has happened? Well, no - but if Tony wins the lottery, and we know that Tony's brother runs the lottery, we might feel differently. Now we have a viable alternative hypothesis.

The problem with P-values is that they take no account of any hypothesis other than the null. In other words, only negative non-relative evidence is being used to evaluate evidence. The philosopher Karl Popper might have supported this approach - but many others disagree!

- ***P*-values include all values more extreme than the observed result**

When we work out a *P*-value we are not just asking how unlikely is this result - we are asking how unlikely is a result as extreme or more extreme than this result. But as we have never seen such result, we just have to imagine they exist. Some statisticians argue that *P*-values therefore overstate the degree of conflict with the null hypothesis. Others disagree, but it remains a controversial aspect of *P*-values.

- **The null distribution of the test statistic may not match the actual sampling distribution of the test statistic.**

Calculation of the *P*-value assumes that the null distribution of the test statistic closely matches the actual sampling distribution of the test statistic. Whilst this may be true of some randomized experiments, it is likely to be much less true in observational studies - where all sorts of confounding factors are liable to be operating. Indeed some statisticians argue that significance tests should not be applied in observational studies at all!

- **About Logistic regression:**

Advantages	Disadvantages
Logistic regression is easier to implement, interpret, and very efficient to train.	If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
It makes no assumptions about distributions of classes in feature space.	It constructs linear boundaries.
It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.	The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).	It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
It is very fast at classifying unknown records.	Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.
Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.	Logistic Regression requires average or no multicollinearity between independent variables.
It can interpret model coefficients as indicators of feature importance.	It is tough to obtain complex relationships using logistic regression. More powerful

	and compact algorithms such as Neural Networks can easily outperform this algorithm.
Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets.One may consider Regularization (L1 and L2) techniques to avoid overfitting in these scenarios.	In Linear Regression independent and dependent variables are related linearly. But Logistic Regression needs that independent variables are linearly related to the log odds (log(p/(1-p)).

VI. Code and Data Availability

1. Source code

2. Data

VII. References

[1] Bevans, R. (2022, November 15). Linear Regression in R | A Step-by-Step Guide & Examples. Scribbr. https://www.scribbr.com/statistics/linear-regression-in-r/

[2] Sun, Y. (2019, November 26). Summarizing CPU and GPU Design Trends with Product Data. arXiv.org. https://arxiv.org/abs/1911.11313

[3] Kelly, L., PhD. (2020, November 20). Practice 9 Calculating Confidence Intervals in R | R Practices for Learning Statistics. https://bookdown.org/logan_kelly/r_practice/p09.html