LOAN ELIGIBILITY STATUS

SUBMITTED TO THE
DEPARTMENT OF STATISTICS

Loyola College (Autonomous),
Chennai-600034

In
Partial fulfillment for the requirement of Bachelor of Science

In Statistics

By
ANIKET GOSWAMI

(20-UST-064)

Under the Supervision of
DR. M. JEEVITHA

DECLARATION

I ANIKET GOSWAMI (20-UST-064) hereby declare that the work entitled "**LOAN ELIGIBILITY STATUS**" is an authentic record of my own work. It is a major project for the period of last semester, as the requirement for the completion of the course B.sc(statistics) under theguidance of Dr **M. JEEVITHA**, Department of Statistics, Loyola College, Chennai – 34.

Signature
ANIKET GOSWAMI
(20-UST-064)

ACKNOWLEDGEMENT

My gratitude is reserved to Dr. Edwin Prabhakaran Head of the Department for his support throughout this project. I am extremely thankful to my guide Dr. M. Jeevitha, for her sincere guidance and timely encouragement throughout the project. I express my sincere thanksto my lovable parents, staff and friends who helped me in carrying out my project work successfully.

BONAFIDE CERTIFICATE

This is to certify that "**LOAN ELIGIBILITY STATUS** " is a bonafide record done by ANIKET GOSWAMI (20-UST-064) in the partial fulfillment of the requirement for the degree of BSc Statistics , Loyola college(autonomous), Chennai – 34 under the guidance of Dr. M. Jeevitha, during the year 2022- 2023.

Dr. Edwin Prabhakaran

Head of the Department

Department of Statistics

Loyola College, Chennai-34

Dr. M. Jeevitha

Department of Statistics

Loyola College-34

INDEX

ABOUT THE STUDY

Loan eligibility refers to the criteria that borrowers need to meet to be considered for a loan. These criteria are set by lenders and may vary depending on the type of loan, the lender's policies, and the borrower's financial circumstances. The eligibility requirements typically include factors such as credit score, income, employment history, debt-to-income ratio, and collateral. Meeting these requirements does not guarantee loan approval, but it increases the likelihood of being approved.

Currently there are a number of criterion that can be taken in order to decide whether you are eligible for that particular loan. It is done by predicting if the loan can be given to that person on the basis of various parameters like credit score, income, age, marital status, gender, etc. The prediction model not only helps the applicant but also helps the bank by minimizing the risk and reducing the number of defaulters.

The aim of this project is to develop a system that can perform a prediction based on categories like Gender, Dependents, Income etc., by combining the results made by different machine learning models. This project aims to predict the Loan Approval status by two methods namely: Logistic Regression and Decision Tree model.

ABOUT THE DATA

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

The company wants to automate the loan eligibility process (real-time) based on customer detail provided while filling the online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. To automate this process, they have given a problem to identify the customer's segments, those are eligible for loan amount so that they can specifically target these customers. There are a total of 481 observations. Here they have provided a partial data set.

KEY IMFORMATION REGARDING THE DATA:

KEY IMFORMATION	DESCRIPTION
Gender	Male/Female
Married	Y/N
Dependents	Number of dependents
Education	Graduate/Under-graduate
Self-Employed	Y/N
Applicant’s Income	Applicant income
Co-applicant’s Income	Co-applicant’s income
Loan amount	Loan Amount in thousands
Loan amount term	Term of a loan in months
Credit-history	Credit history meets guidelines
Property Area	Urban/Semi-urban/Rural
Loan-Status	Loan Approval(Y/N)

PROPOSED METHODOLOGY

I) DATA CLEANING:

It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluate the results. Dataset carries 481 rows i.e., the total number of data, and 12 columns i.e., Gender, Married, Dependents, Education, Self-Employed, Applicant's Income, Co-applicant's income, Loan Amount, Loan amount Term, Credit-history, Property Area and Loan Status.

II) DATA PRE-PROCESSING:

This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id was inconsistent so we dropped the feature. This data did not contain any missing observation, so no data-cleaning process was needed. Since, there were very less features and all of them were equally important so no feature selection was done.

III) SPLITTING OF THE DATASET:

After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train / split method, we split the dataset randomly into the training and testing set. For Training, we took 80% of the sample and for testing, we took20% of the sample. Attached is the code for splitting and training the dataset:

```
set.seed(123)
train_index <- sample(1:nrow(data), size = 0.7 * nrow(data), replace = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
```

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is an approach to analyzing and understanding datasets to identify patterns, relationships, and insights that may be hidden within the data. It involves using various statistical and visualization techniques to summarize the main characteristics of the data and gain a deeper understanding of the underlying structure of the data.

The main objective of EDA is to discover interesting and unexpected insights in the data that can guide further analysis or provide new perspectives on a problem. It often involves examining the distribution of the data, identifying outliers and missing values, and exploring relationships between variables.

EDA is commonly used in fields such as statistics, data science, and machine learning to help researchers and analysts make better decisions and develop more effective models. By understanding the underlying patterns and relationships within the data, analysts can make better predictions, identify potential problems, and design more effective interventions.

- UNIVARIATE ANALYSIS:
  Univariate analysis is a statistical technique that involves analyzing a single variable in isolation to gain insights into its characteristics and properties. It is often used to describe the distribution of a variable, such as its mean, median, mode, range, standard deviation, and variance. Univariate analysis can also be used to identify outliers and to test hypotheses about the distribution of the variable, such as whether it is normally distributed or not. In summary, univariate analysis involves exploring the characteristics and behavior of a single variable without considering any other variables.

- BIVARIATE ANALYSIS:
  Bivariate analysis is a statistical method used to analyze the relationship between two variables. It is a form of data analysis that involves examining two variables simultaneously to determine if there is a correlation or association between them. Bivariate analysis can be used to explore the relationship between variables such as age and income, or height and weight. It is often used in research studies to identify patterns and relationships between variables and to test hypotheses about the

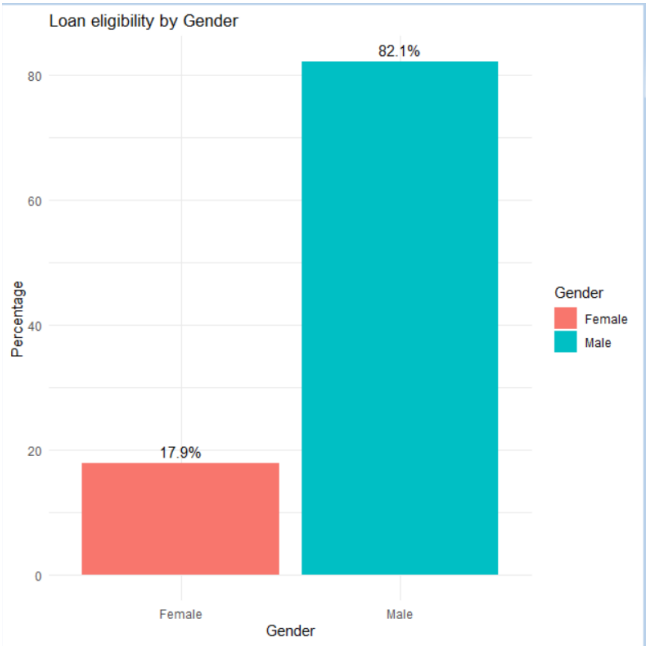
relationship between them. Bivariate analysis can be conducted using a variety of statistical methods, including correlation analysis, regression analysis, and chi-square tests.

In conclusion, exploratory data analysis includes various techniques, such as univariate and bivariate analysis and box plots, which help in understanding the distribution and relationships of variables in the data. Univariate analysis summarizes the central tendency, dispersion, and shapeof the distribution of a single variable, while bivariate analysis examines therelationship between two variables. Box plots provide a graphical representation of the distribution of a variable and help in visualizing the distribution, identifying outliers, and comparing the distributions of different variables. EDA is an essential step in data analysis, as it helps in gaining insights and formulating hypotheses before conducting formal statistical inference or modeling.

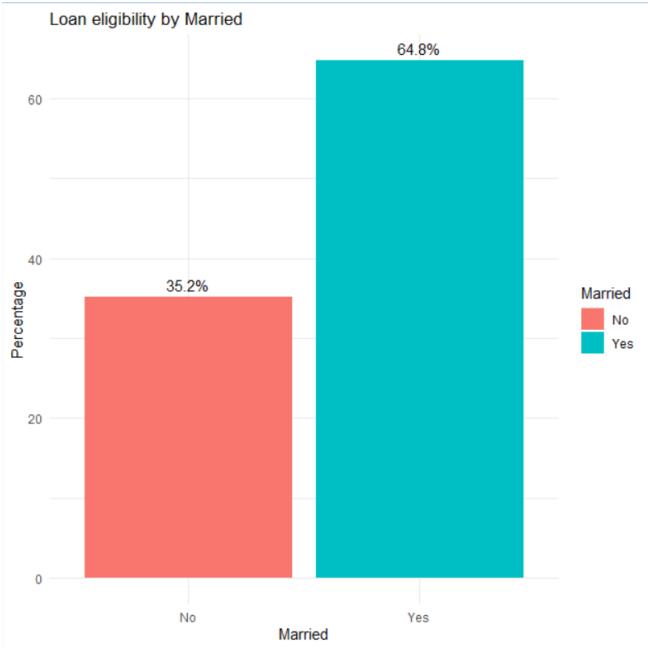
UNIVARIATE ANALYSIS OF THE DATA

Under the Univariate analysis, a simple bar plot was performed for categorical data and for continuous data histogram plot was performed.

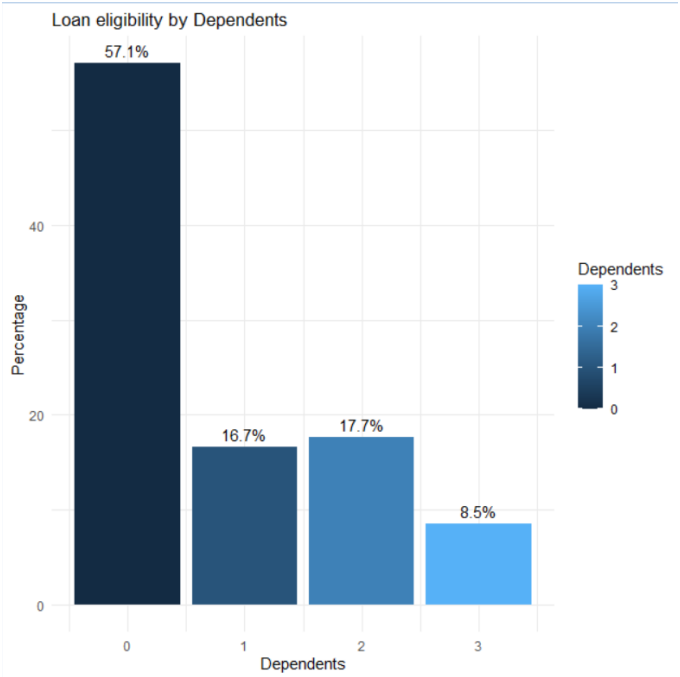
FOR CATEGORICAL DATA:



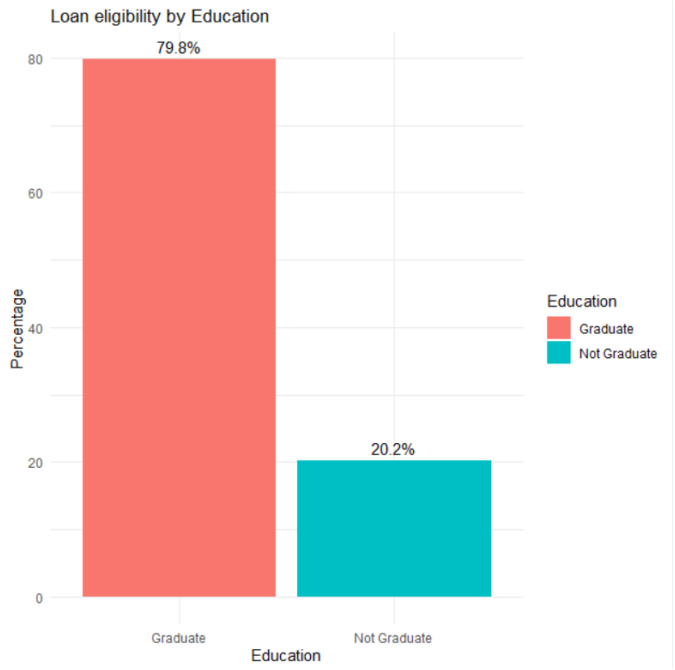
82.1% male are eligible for loan whereas only 17.9% female are eligible.



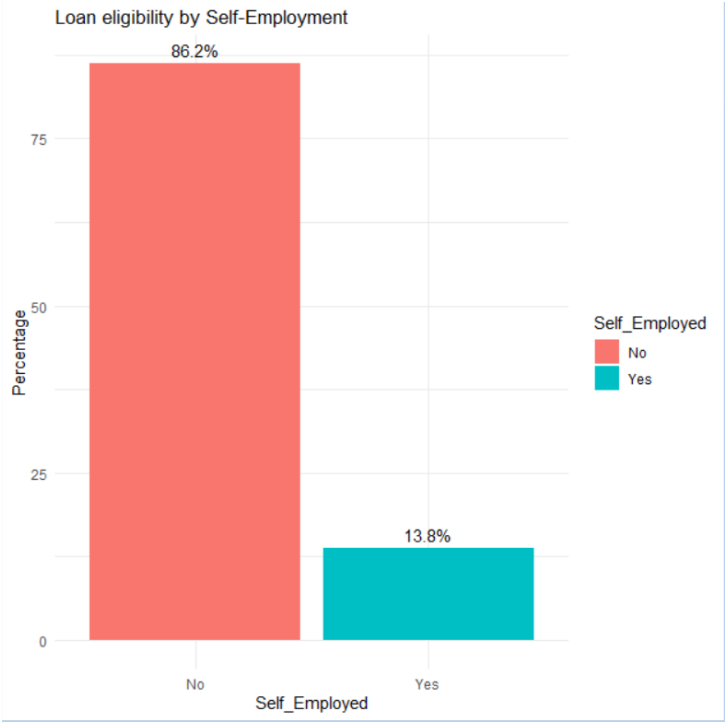
64.8% people who are married were eligible for loan whereas 35.2% who were unmarried were eligible.



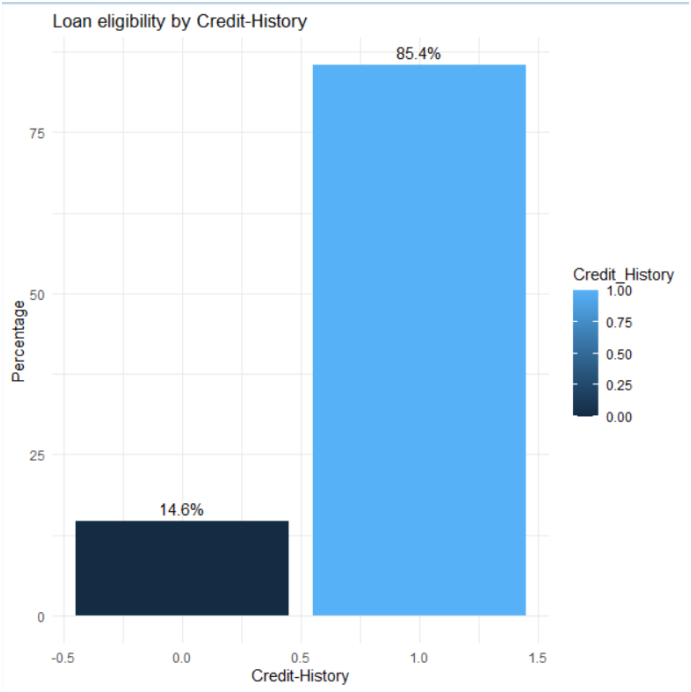
57.1% people whose loan got approved have dependents 0 whereas 16.7% have dependents 1, 17.7% have dependents 2 and 8.5% have dependents 3.



79.8% Graduate people were eligible for loan whereas 20.2% non-graduate were eligible.



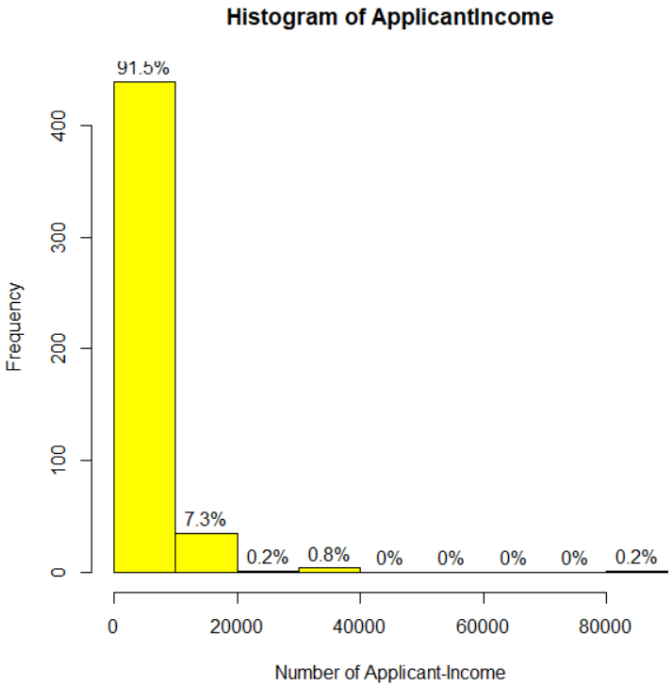
86.2% people who are not self-employed were eligible for loan status whereas only 13.8% people who were self-employed were eligible.



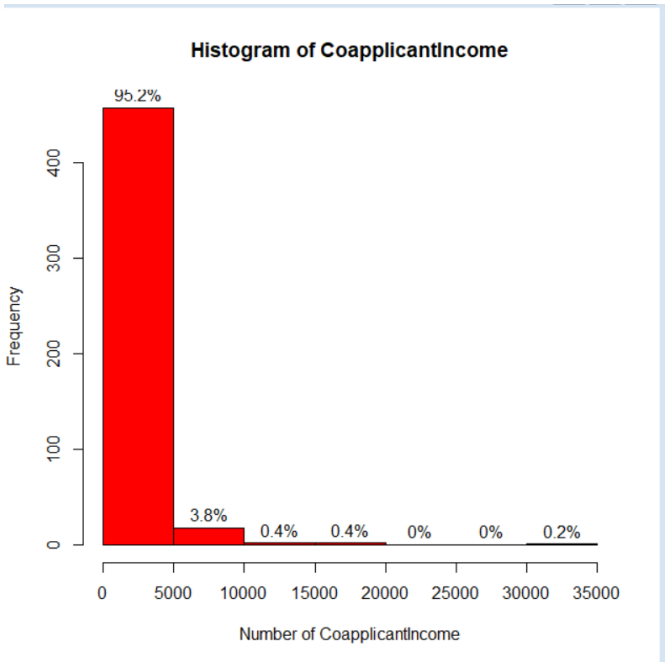
85.4% people who has credit-history 1 were eligible for loan whereas 14.6% people who has 0 credit-history were eligible for loan.

39.8% people who were eligible for loan lives in semi-urban areas, 31.2% people who were eligible for loan lives in urban areas and 29% people who were eligible for loan lives in rural areas.

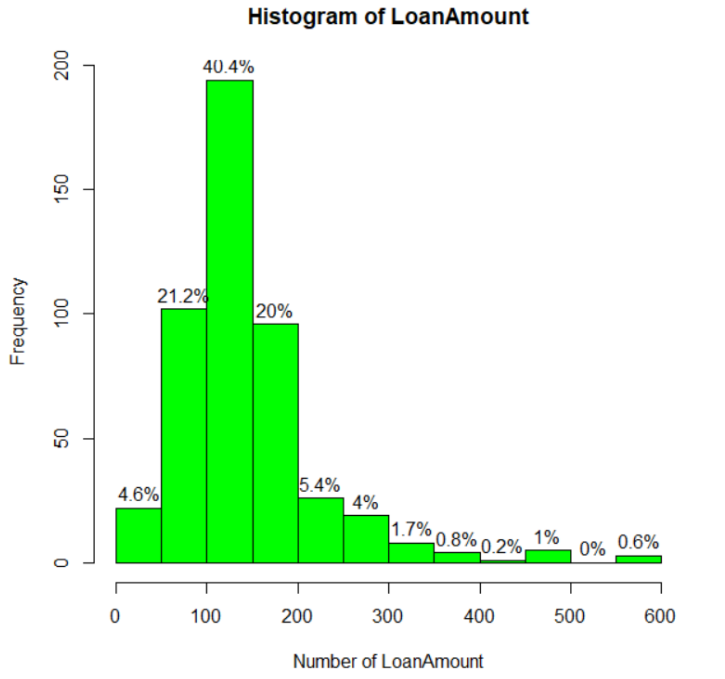
FOR CONTINUOUS DATA:



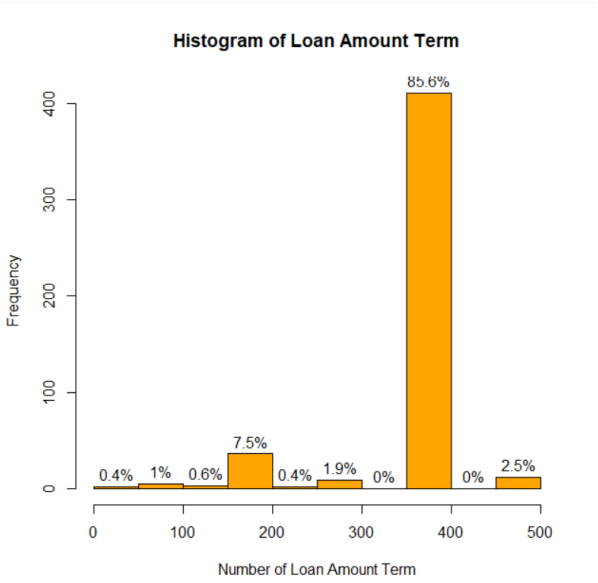
91.2% people have income between 0-20000.



95.2% co-applicant's have income between 0-5000



40.4% people can have a loan amount between 100000-200000.

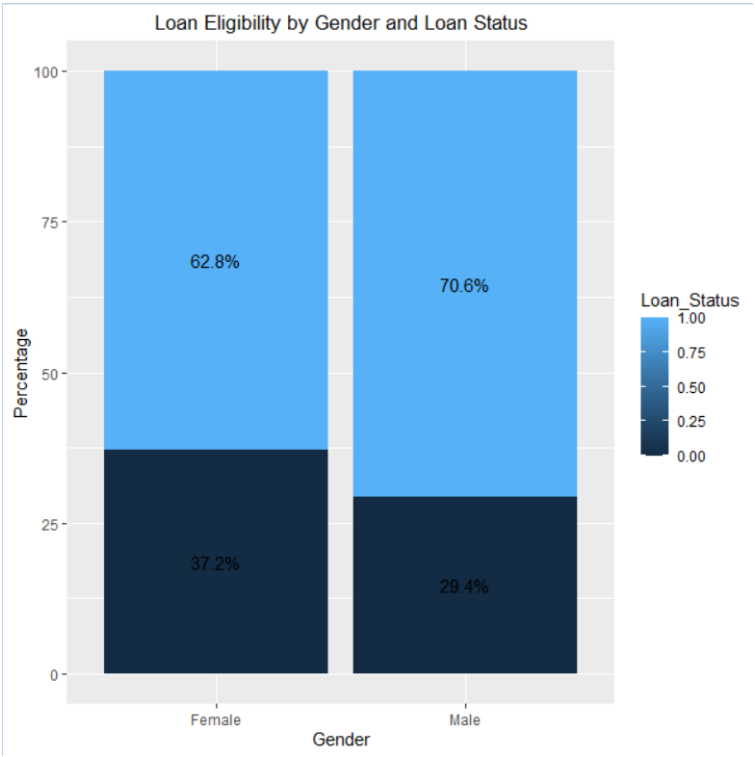


85.6% people can have a loan period between 380-400 months.

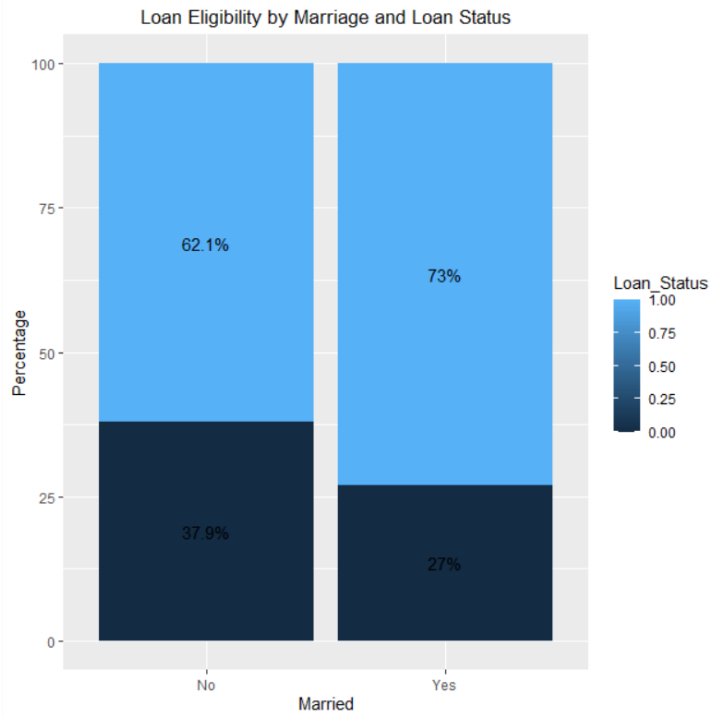
BIVARIATE ANALYSIS

In order to perform bivariate analysis, we used stacked- bar charts for categorical variable and box-plot for continuous variable.

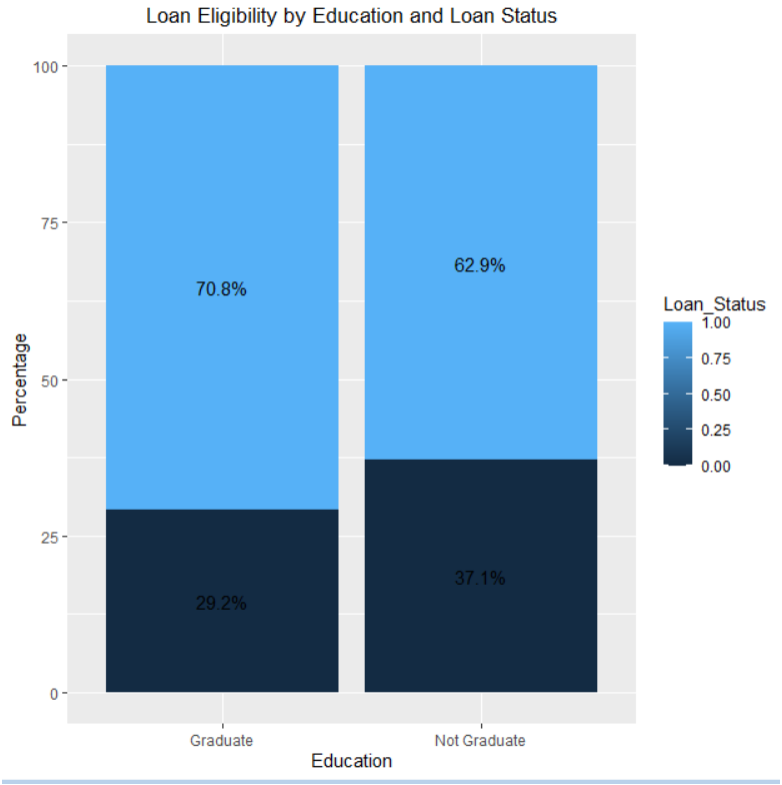
FOR CATEGORICAL DATA:



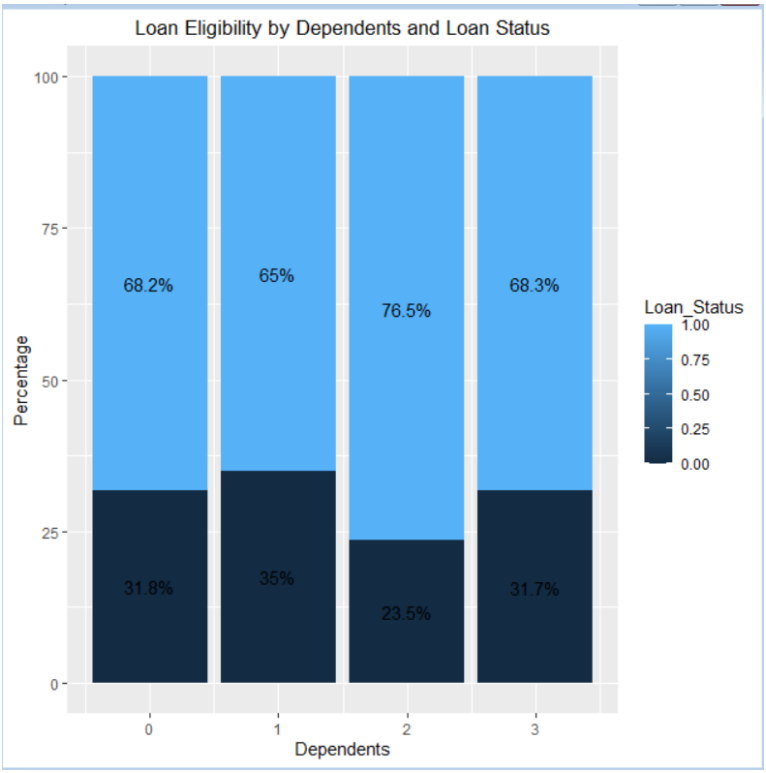
Of the male population,70.6%of them got their loan status approved while 62.8% of the female population got their status approved.



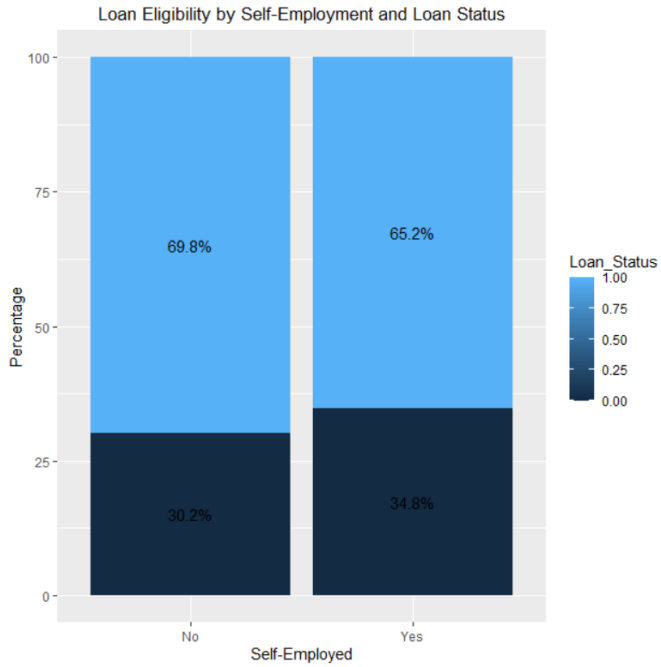
73% of the people who are married got their loan status approved while those who are not married 62.1% of them got their status approved.



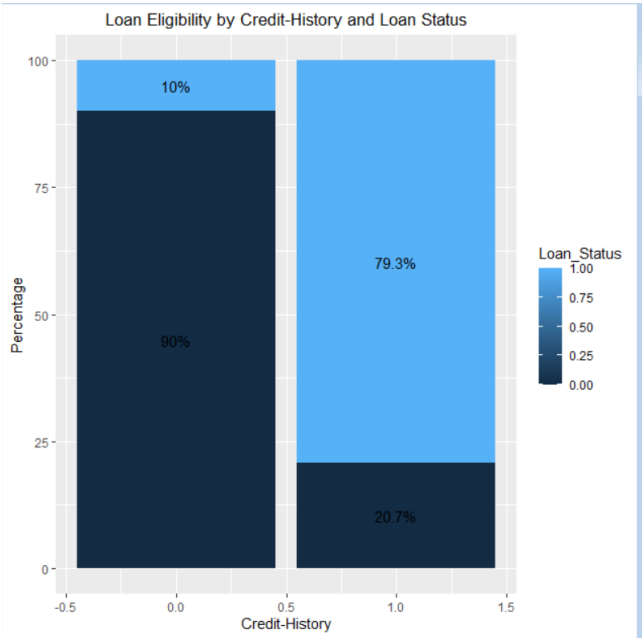
70.8% people who are graduate got their loan approved while 62.9% people who are not graduate got their loan approved.



75.6% of people who have dependents 2 got their loan approved following which 68.3% with dependents 4, 68.2% with dependents 2 and 65% percent with dependents 1.

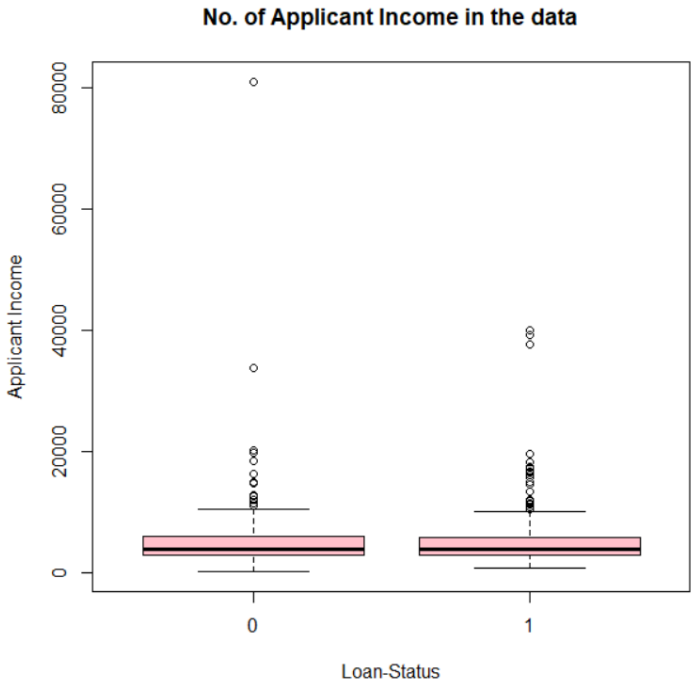


69.8% people who are not self-employed got their loan approved while 65.2% people who are self-employed got their status approved.

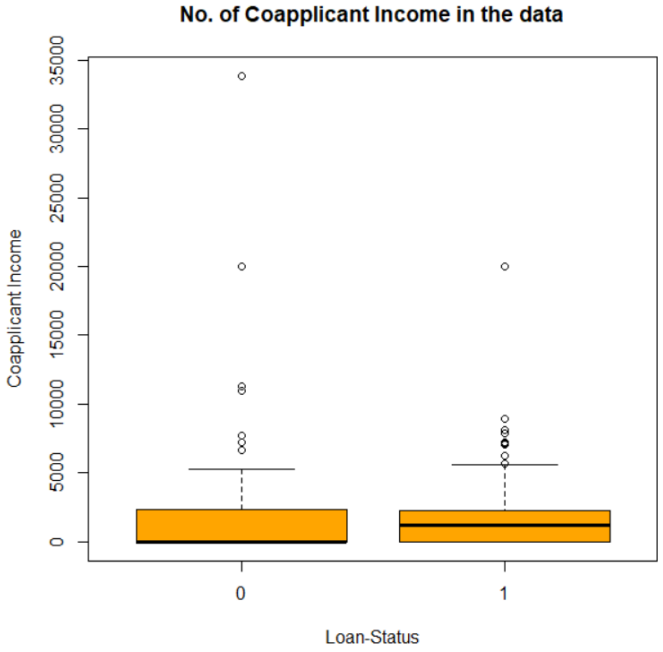


79.3% of people with credit-history 1 got their loan approved while only 10% of people with 0 credit-history got their loan approved.

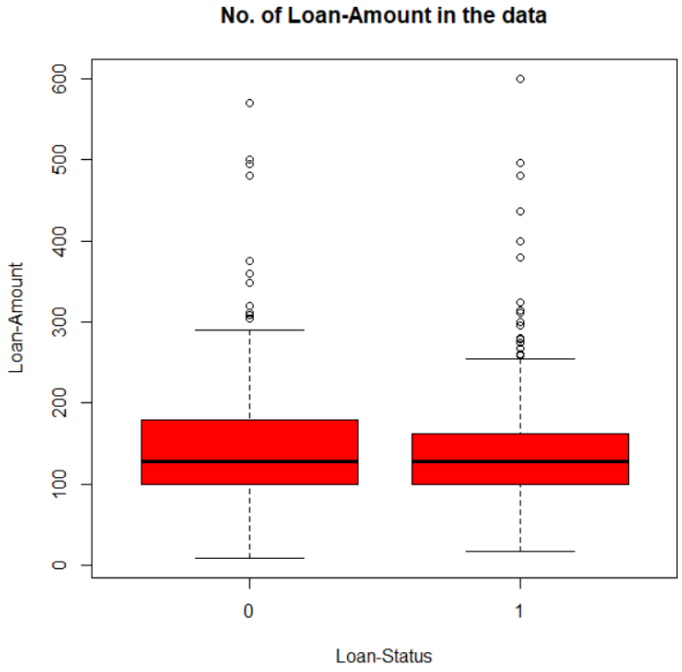
FOR CONTINUOUS DATA:



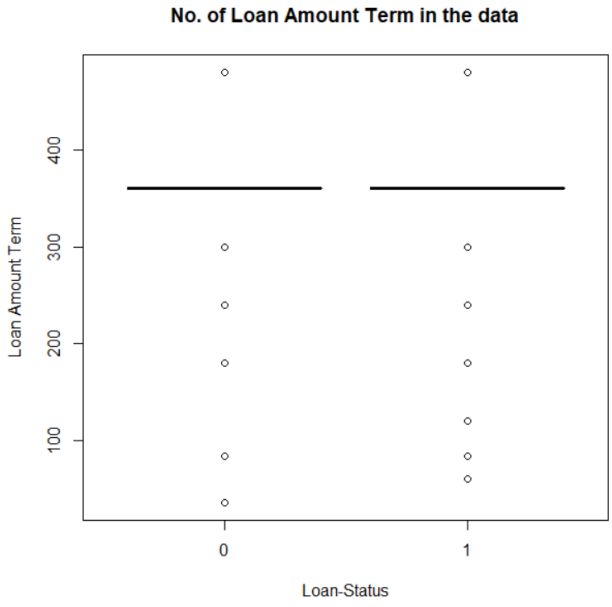
The p-value is too small and since it is less than 0.05 we can conclude that number of applicant's income is significant.
There are outliers in the data.



The p-value is too small and since it is less than 0.05 we can conclude that number of Co-applicant's income is significant.
There are significant outliers in the data.



The p-value is too small and since it is less than 0.05 we can conclude that number of Loan Amount is significant. There are quite a few outliers in the data.



The p-value is too small and since it is less than 0.05 we can conclude that number of Loan Amount Term is significant. There are a few outliers in the data.

TESTS FOR VARIOUS DATA

We have applied various statistical tests for the data. Listed below are the tests along with their results.

- FOR CATEGORICAL VS CATEGORICAL DATA CHI-SQUARE TEST IS USED:

```
        Pearson's Chi-squared test

data:  data$Gender and data$Loan_Status
X-squared = 1.9972, df = 1, p-value = 0.1576
```

Since the p-value is greater than 0.05, it is not significant.

```
        Pearson's Chi-squared test

data:  data$Married and data$Loan_Status
X-squared = 6.0557, df = 1, p-value = 0.01386
```

Since the p-value is less than 0.05, it is significant.

```
        Pearson's Chi-squared test

data:  data$Dependents and data$Loan_Status
X-squared = 2.9006, df = 3, p-value = 0.4072
```

Since the p-value is greater than 0.05, it is not significant.

```
        Pearson's Chi-squared test

data:  data$Education and data$Loan_Status
X-squared = 2.2482, df = 1, p-value = 0.1338
```

Since the p-value is greater than 0.05, it is not significant.

```
        Pearson's Chi-squared test

data:  data$Self_Employed and data$Loan_Status
X-squared = 0.57846, df = 1, p-value = 0.4469
```

Since the p-value is greater than 0.05, it is not significant.

```
        Pearson's Chi-squared test

data:  data$Credit_History and data$Loan_Status
X-squared = 134.52, df = 1, p-value < 2.2e-16
```

Since the p-value is less than 0.05, it is significant.

```
        Pearson's Chi-squared test

data:  data$Property_Area and data$Loan_Status
X-squared = 12.226, df = 2, p-value = 0.002214
```

Since the p-value is less than 0.05, it is significant.

- FOR CONTINUOUS VS CATEGORICAL WE USE ANOVA:

```
> result<- aov(ApplicantIncome~Loan_Status,data=data)
> summary(result)
             Df    Sum Sq  Mean Sq F value Pr(>F)
Loan_Status   1 2.866e+07 28656769   0.892  0.345
Residuals   478 1.536e+10 32136336
```

than 0.05, it is not significant.

```
> result<- aov(CoapplicantIncome~Loan_Status,data=data)
> summary(result)
             Df    Sum Sq Mean Sq F value Pr(>F)
Loan_Status   1 7.887e+06 7886980   1.151  0.284
Residuals   478 3.274e+09 6850148
```

Since the p-value is greater than 0.05, it is not significant.

```
> result<- aov(LoanAmount~Loan_Status,data=data)
> summary(result)
             Df  Sum Sq Mean Sq F value Pr(>F)
Loan_Status   1   15984   15984   2.474  0.116
Residuals   478 3088685    6462
```

Since the p-value is greater than 0.05, it is not significant.

```
> result<- aov(Loan_Amount_Term~Loan_Status,data=data)
> summary(result)
             Df  Sum Sq Mean Sq F value Pr(>F)
Loan_Status   1     124     124   0.029  0.865
Residuals   478 2036899    4261
```

Since the p-value is greater than 0.05, it is not significant.

MODEL BUILDING

Model building in statistics refers to the process of developing a mathematical representation of a system or phenomenon based on available data. The goal is to create a model that accurately captures the behavior of the system or phenomenon and can be used to make predictions or gain insights.

There are several steps involved in building a statistical model:

- Define the problem: Clearly define the research question or problem you are trying to solve.
- Collect data: Gather relevant data that can be used to build the model.
- Choose a modeling technique: Select a modeling technique that is appropriate for the problem at hand. Common techniques include linear regression, logistic regression, decision trees, and neural networks.
- Prepare the data: Preprocess and clean the data, including handling missing values, outliers, and categorical variables.
- Build the model: Use the selected modeling technique to build the model and estimate the model parameters.
- Evaluate the model: Assess the model's performance using various metrics such as mean squared error, accuracy, precision, recall, and F1 score.
- Validate the model: Test the model's performance on new, unseen data to ensure that it is generalizable and not overfitting to the training data.
- Refine the model: Iterate and refine the model as needed, making changes to the modeling technique or data preparation steps to improve performance.
- Deploy the model: Use the model to make predictions or gain insights about the system or phenomenon.

Overall, model building in statistics is an iterative process that involves careful attention to the data and the problem at hand, as well as a deep understanding of the modeling techniques and their assumptions.

1. LOGISTIC REGRESSION:

Logistic regression is a statistical method used for analyzing data in which there are one or more independent variables that determine an outcome or dependent variable. It is a type of regression analysis that is used to predict a binary outcome (i.e., yes or no, success or failure, etc.).

The goal of logistic regression is to find the relationship between the independent variables and the dependent variable by estimating the probabilities of the outcomes. It is called "logistic" because it uses a logistic function to model the probability of the outcome variable.

Logistic regression is commonly used in various fields such as finance, healthcare, and marketing. In finance, it can be used to predict whether a customer will default on a loan or not. In healthcare, it can be used to predict the likelihood of a patient developing a particular disease. In marketing, it can be used to predict whether a customer will purchase a product or not.

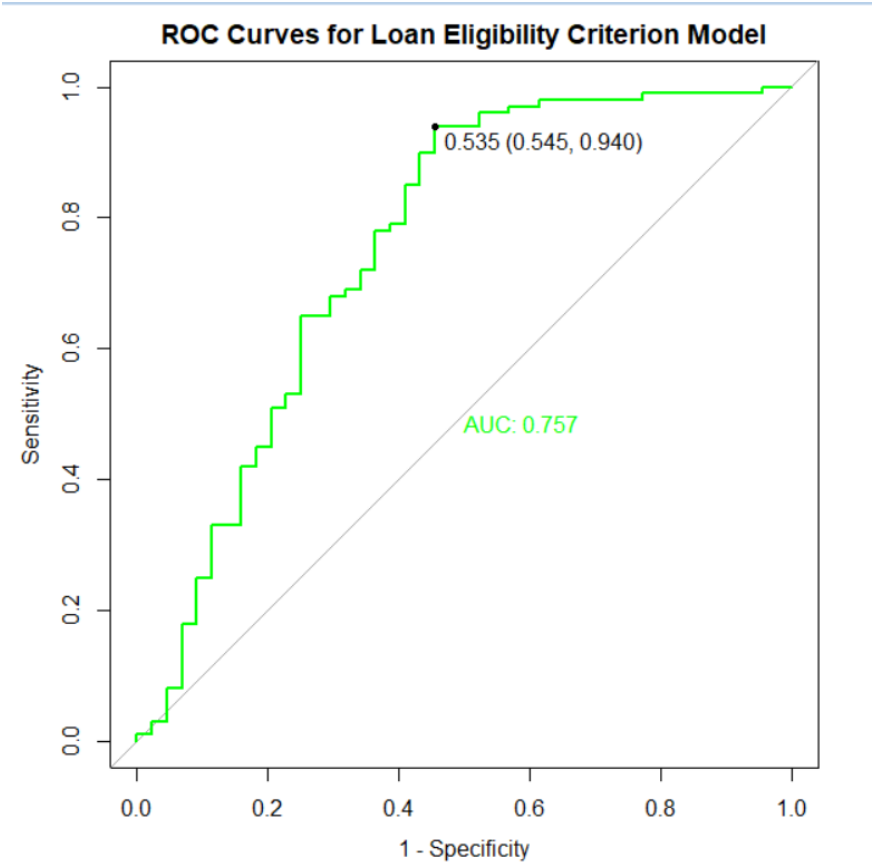
One of the advantages of logistic regression is that it provides interpretable coefficients, which can be used to determine the strength and direction of the relationship between the independent variables and the dependent variable. Additionally, logistic regression can be easily implemented in most statistical software packages.

However, logistic regression also has its limitations. It assumes that the relationship between the independent variables and the dependent variable is linear, and it may not perform well if there are non-linear relationships between the variables. It also assumes that there is no multicollinearity between the independent variables, and it may not be effective if the assumptions are violated.

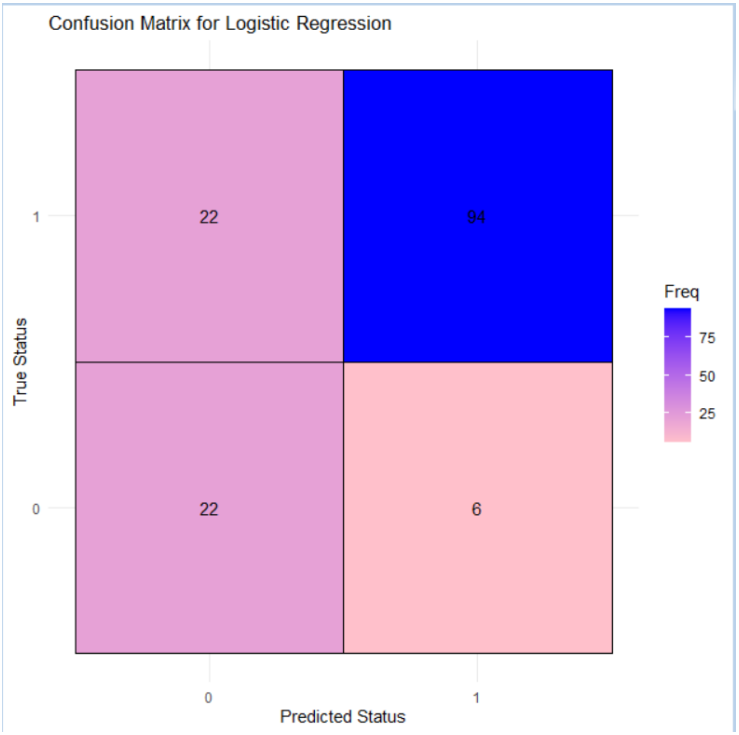
```
Call:
glm(formula = Loan_Status ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2635  -0.3403   0.4951   0.6956   2.3961

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -4.035e+00  1.169e+00  -3.450  0.00056 ***
GenderMale               3.176e-01  3.927e-01   0.809  0.41853
MarriedYes               8.325e-01  3.525e-01   2.362  0.01820 *
Dependents               8.573e-02  1.610e-01   0.532  0.59440
EducationNot Graduate   -4.809e-01  3.581e-01  -1.343  0.17931
Self_EmployedYes        -5.054e-01  4.485e-01  -1.127  0.25980
ApplicantIncome          4.542e-05  6.357e-05   0.714  0.47496
CoapplicantIncome       -1.462e-06  5.373e-05  -0.027  0.97829
LoanAmount              -3.811e-03  2.583e-03  -1.476  0.14000
Loan_Amount_Term         1.188e-03  2.195e-03   0.541  0.58832
Credit_History           4.130e+00  6.434e-01   6.418 1.38e-10 ***
Property_AreaSemiurban   1.233e+00  3.690e-01   3.342  0.00083 ***
Property_AreaUrban       2.917e-01  3.554e-01   0.821  0.41172
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 415.78  on 335  degrees of freedom
Residual deviance: 299.26  on 323  degrees of freedom
AIC: 325.26

Number of Fisher Scoring iterations: 5
```

```
> fitted.results <- predict(model,newdata=test,type='response')
> fitted.results <- ifelse(fitted.results > 0.5,1,0)
> (conf_matrix_logi<-table(fitted.results, test$Loan_Status))

fitted.results  0  1
             0 22  6
             1 22 94
```

```
> misClasificError <- mean(fitted.results != test$Loan_Status)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.805555555555556"
```



CONFUSION MATRIX FOR LOGISTIC REGRESSION:



2. DECISION TREE:

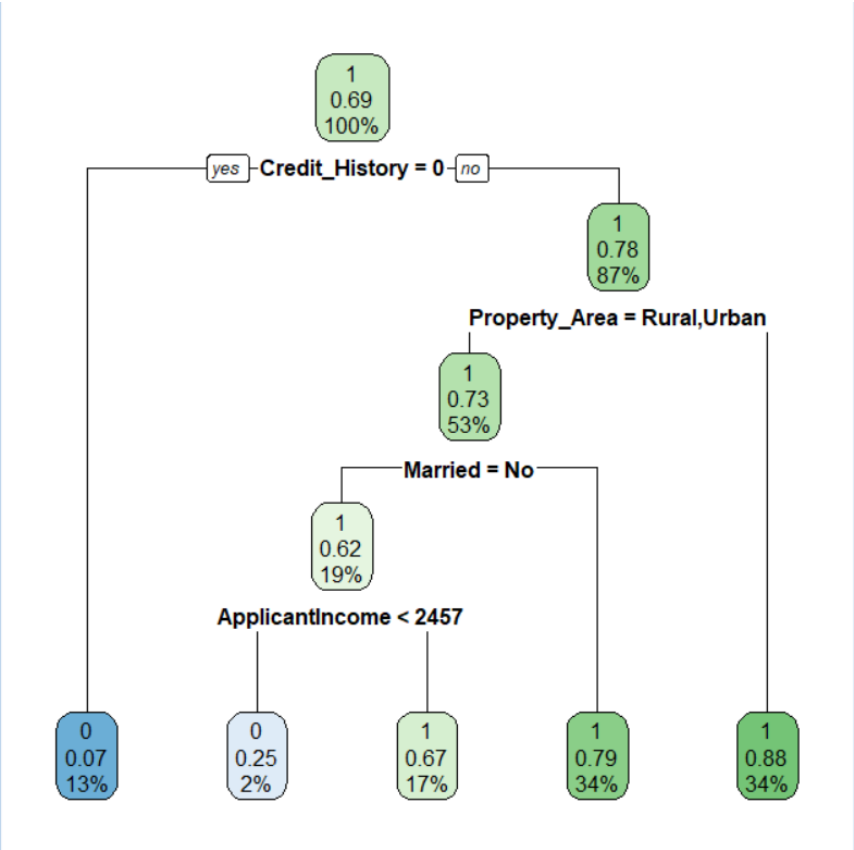
A decision tree is a powerful tool in statistics and data analysis, used for both classification and regression tasks. It is a hierarchical model that represents decisions and their possible consequences in a tree-like structure.

In a decision tree, each internal node represents a test on an attribute, and each branch represents the outcome of the test. The leaves of the tree represent the decision or prediction, based on the values of the attributes in the corresponding path from the root to the leaf.

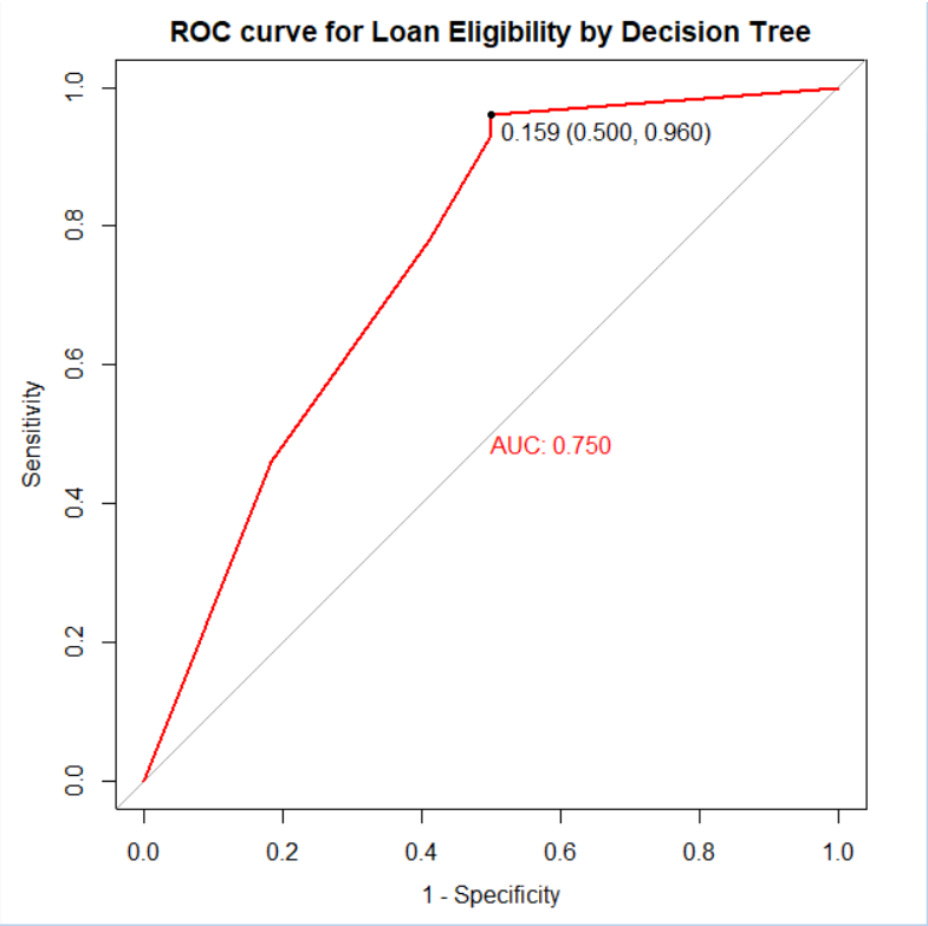
Decision trees are widely used in data mining, machine learning, and predictive analytics, and have several advantages over other classification and regression techniques, such as:

- Easy to interpret: Decision trees provide a clear and intuitive representation of the decision-making process, making it easy to explain and understand.
- Handling both numerical and categorical data: Decision trees can handle both numerical and categorical data, making them useful for a wide range of applications.
- Non-parametric: Decision trees are non-parametric, meaning they do not make any assumptions about the underlying distribution of the data.
- Robust to outliers: Decision trees are robust to outliers and can handle noisy data, making them useful in real-world scenarios where data may not be perfectly clean.

Some of the popular algorithms for decision trees include ID3, C4.5, CART, and Random Forest.




```
> confusion_matrix<- table(predictions,test$Loan_Status)
> accuracy<- sum(diag(confusion_matrix))/sum(confusion_matrix)
> confusion_matrix

predictions  0  1
          0 22  7
          1 22 93
> accuracy
[1] 0.7986111
```

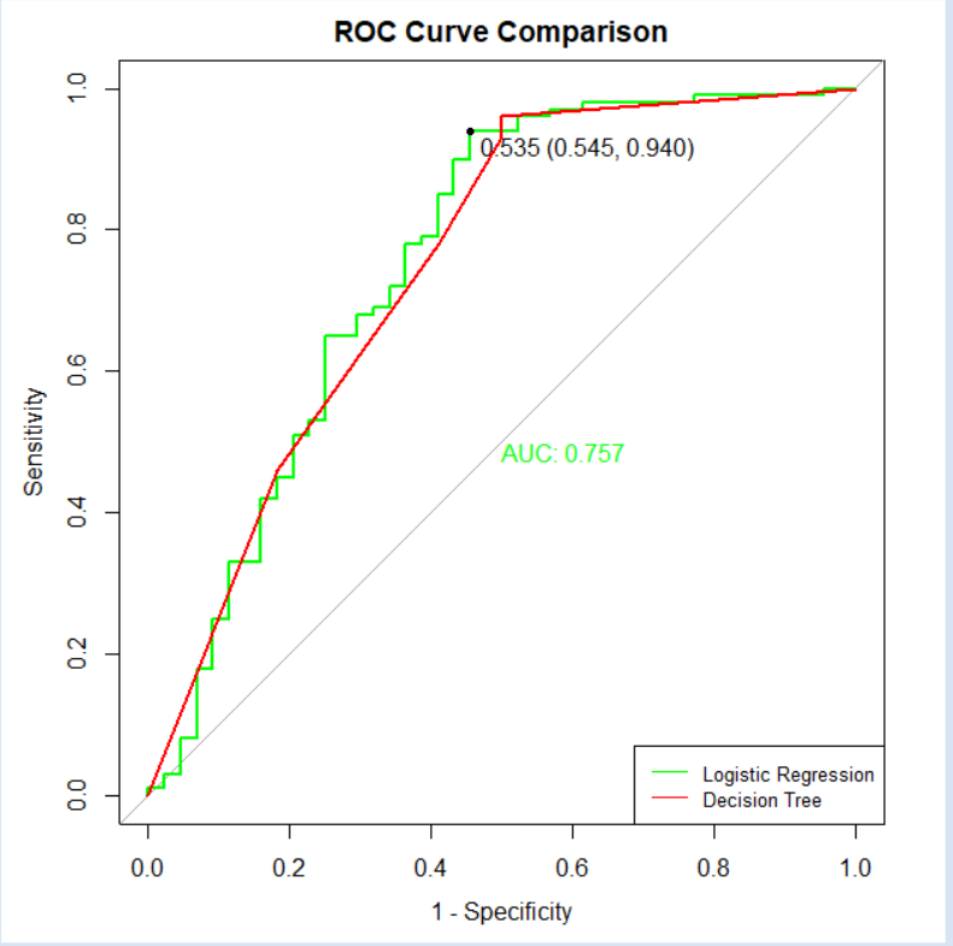

ROC curve for Loan Eligibility by Decision Tree
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.159 (0.500, 0.960)
AUC: 0.750

FINAL RESULT

In this analysis, three different models were used to predict the loan eligibility in a dataset taken from the Kaggle. The models were logistic regression and decision tree. The accuracy and AUC in the ROC curve were used as performance metrics to evaluate and compare the three models.

The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classifier system as the discrimination threshold is varied. It is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
The area under the ROC curve (AUC) is a commonly used metric to evaluate the performance of binary classifiers. A perfect classifier has an AUC of 1.0, while a random guess classifier has an AUC of 0.5. An AUC value of 0.7 or higher is generally considered to be a good classifier. The ROC curve is useful because it allows you to visualize and compare the performance of different classifiers. For example, if you have two classifiers that you want to compare, you can plot their ROC curves on the same graph and compare the AUC values. The classifier with the higher AUC value is generally considered to be the better classifier. In summary, ROC curves and AUC are important tools for evaluating the performance of binary classifiers, allowing you to visualize and compare the trade-off between the true positive rate and the false positive rate at different discrimination thresholds.

The logistic regression model had an accuracy of 80.56% and an AUC in theROC curve of 0.757. The decision tree model had an accuracy of 79.86% and an AUC in the ROC curve of 0.750.



These results suggest that the logistic regression model may be the most effective model in predicting the loan eligibility status of a population.

CONCLUSION

In this data taken from Kaggle on the topic Loan Eligibility, we have compared the performance of two models, that is, Logistic regression and Decision Tree. After comparing, we came to a conclusion that the Logistic Regression has better performance on this data with an accuracy of 80.5% compared to Decision Tree with accuracy of around 79%.

I used R-programming language to complete this project on the data taken from Kaggle. Firstly, I did Exploratory Data Analysis on the data that includes Univariate and Bivariate Analysis along with Cross-Tabulation. After this, I build my model based on the significant data which was tested by Chi-Square and ANOVA test and the insignificant data was removed.

After comparing the results of both the methods we found that the Logistic Regression model gives better performance than the Decision Tree model with an accuracy of 80.5%.