LOAN PREDICTION

PREDICTIVE MODELING

Presented By:

- Khushi Goyal 24034
- Tanisha Sighania 24057
- Sakshi Saraiya 24011
- Krish Garg 24042
- Ankit Mittal 24055

INTRODUCTION

Objective:

- To develop a predictive model to determine loan approval status based on customer attributes.

Problem Statement:

- Predicting loan status (approved or not) based on features such as income, credit history, and property area.

Dataset Source:

- Kaggle's Loan Prediction Dataset.

DATASET DESCRIPTION

Key Highlights:

- Rows: 10,439
- Columns: 12
- Target Variable: Loan_Status (Approved/Not Approved).
- **Variable Types:**
  - Numerical: Applicant_Income, Loan_Amount, etc.
  - Categorical: Gender, Married, Education, etc.

Objective: Predict the likelihood of loan approval.

DISTRIBUTION OF LOAN STATUS

The majority of loan applications were approved (Loan_Status = Y), indicating a higher approval rate overall.

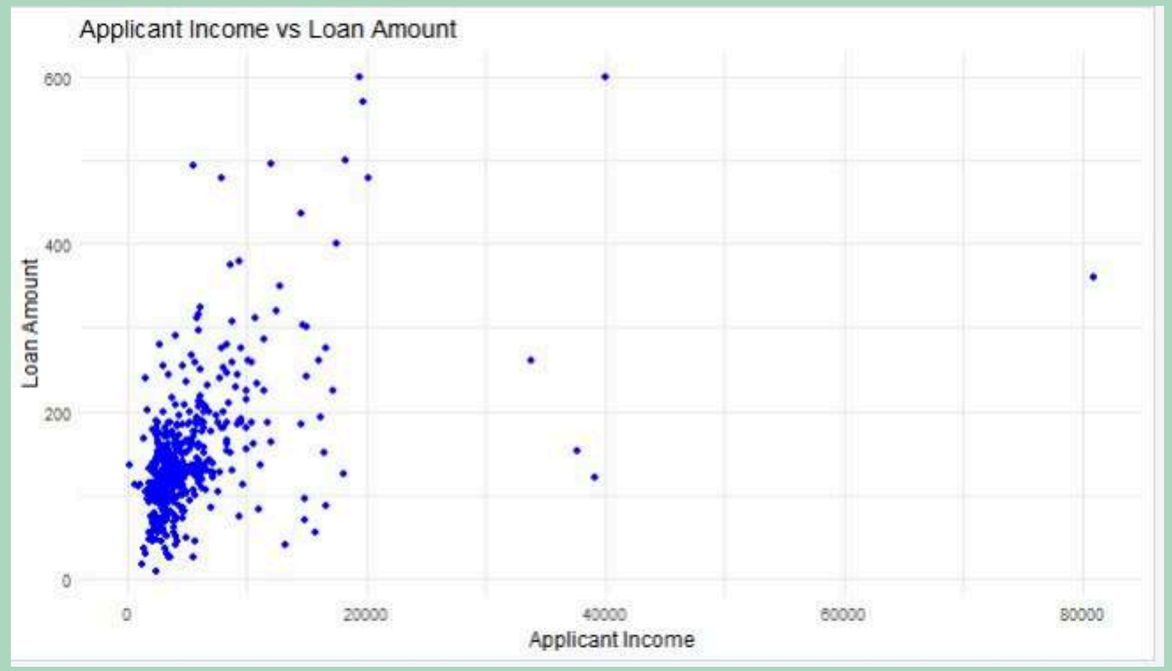
```
#Visualization (Bar Plot)
ggplot(Data,aes(x=Loan_Status)) +
geom_bar(fill="skyblue",color="black")+
  labs(title="Distribution of Loan
Status",x="Loan Status",y="Count") +
  theme_minimal()
```

AVERAGE LOAN AMOUNT BY PROPERTY AREA

- The average loan amount is highest in Rural areas (155.64).
- It is followed by Semiurban areas (148.06).
- The lowest average loan amount is observed in Urban areas (132.95).

```
ggplot(Avg_Loan_Amount,aes(x=Property_Area,
y=Loan_Amount,fill=Property_Area)) +
  geom_bar(stat="identity",color="black") +

geom_text(aes(label=Loan_Amount),vjust=-0.5,
color="black",size=3)+
  labs(title="Average Loan Amount by Property
Area",x="Property Area",y="Average Loan
Amount") +
  theme_minimal()
```

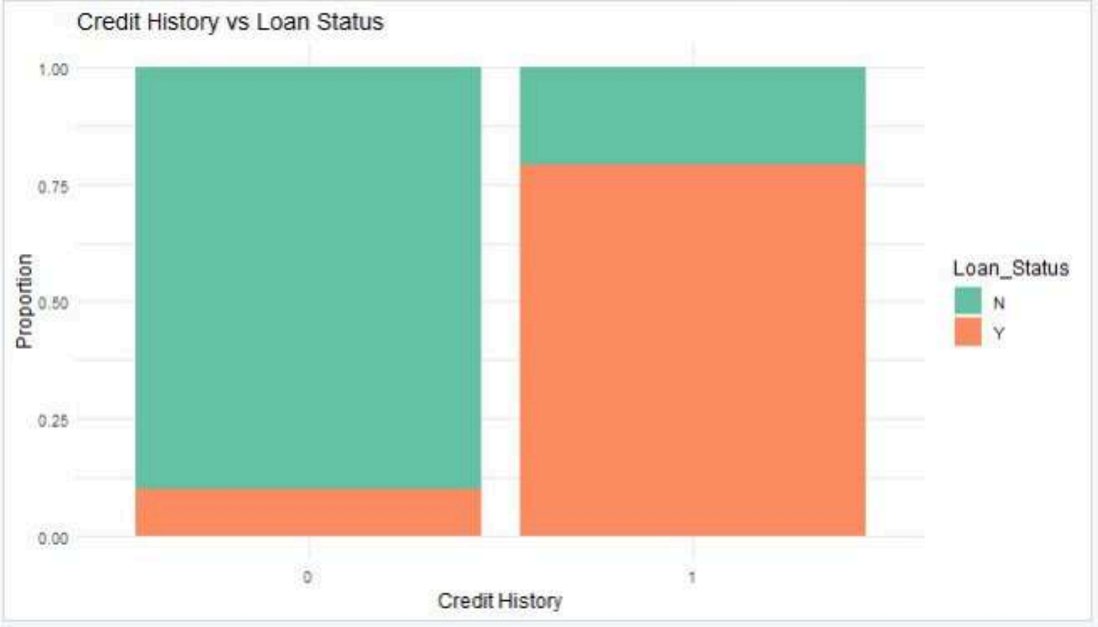
APPLICANT INCOME VS LOAN AMOUNT



- There is no strong linear correlation between Applicant Income and Loan Amount, as evident from the scattered distribution.
- Most applicants have lower incomes (under 20,000), and the majority of loan amounts are clustered below 200 units.
- A few outliers show very high incomes and loan amounts.

```
ggplot(Data,aes(x=Applicant_Income,y=Loan_Amount)) +
  geom_point(color="blue",alpha=0.6) +
  labs(title="Applicant Income vs Loan
Amount",x="Applicant Income",y="Loan Amount")+
  theme_minimal()
```

CREDIT HISTORY VS LOAN STATUS



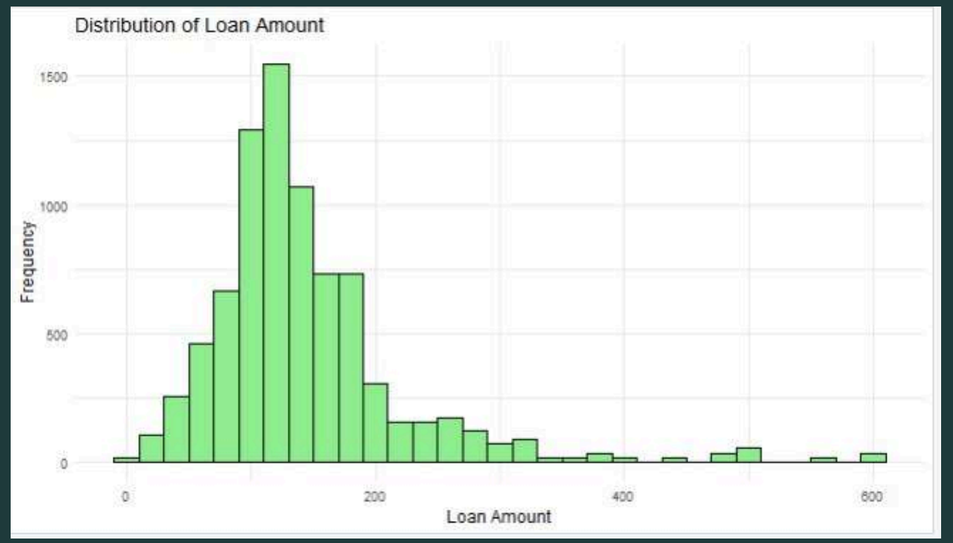
Credit history is a critical determinant in loan approval decisions.
Applicants with a credit history (Credit_History = 1) have a significantly higher probability of loan approval (Loan_Status = Y).
Applicants without a credit history (Credit_History = 0) are more likely to face rejection (Loan_Status = N).

```
ggplot(Data,aes(x=factor(Credit_History),fill=Loan_Status))+
  geom_bar(position="fill")+
        labs(title="Credit History vs Loan Status",x="Credit History",y="Proportion") +
  scale_fill_brewer(palette="Set2")+
  theme_minimal()
```

DISTRIBUTION OF LOAN AMOUNT

- The distribution of loan amounts is right-skewed, indicating that most loan amounts are smaller.
- The majority of the loans are in the range of 100 to 200 units (currency unspecified) with very few loans exceeding 600 units.

```
ggplot(Data,aes(x = Loan_Amount))+
geom_histogram(binwidth=20,fill="lightgreen",color="black") +
  labs(title="Distribution of Loan Amount",x="Loan Amount",y="Frequency")+
  theme_minimal()
```



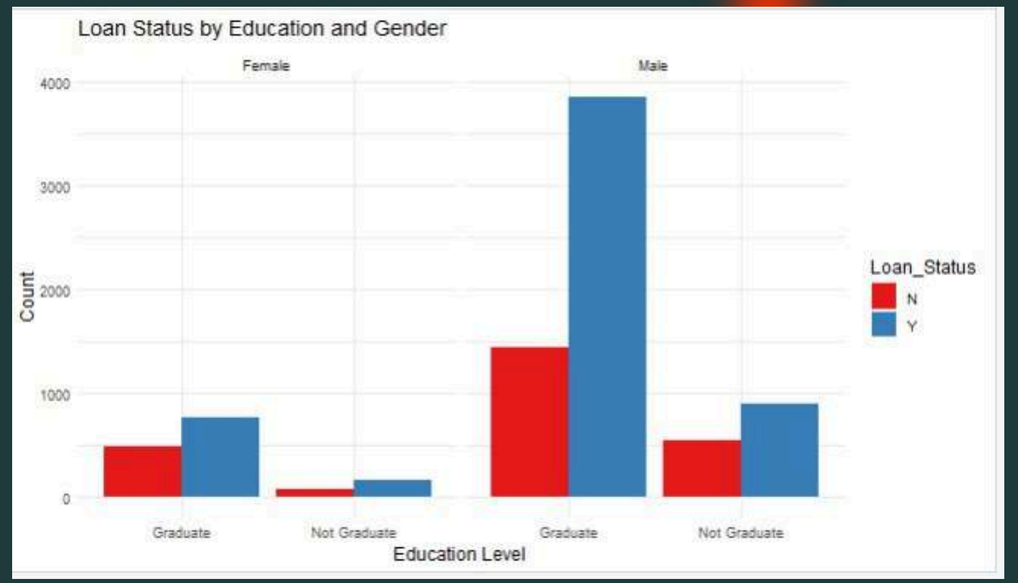
LOAN STATUS BY EDUCATION AND GENDER

• Male graduates have the highest number of loan approvals (marked by the blue bar under “Y”) compared to other groups.

• Female graduates also show a higher loan approval rate than non-graduates, but their overall numbers are lower compared to males.

• Non-graduates, both male and female, show significantly fewer loan approvals. However, the loan rejection rate is notably higher for male non-graduates.

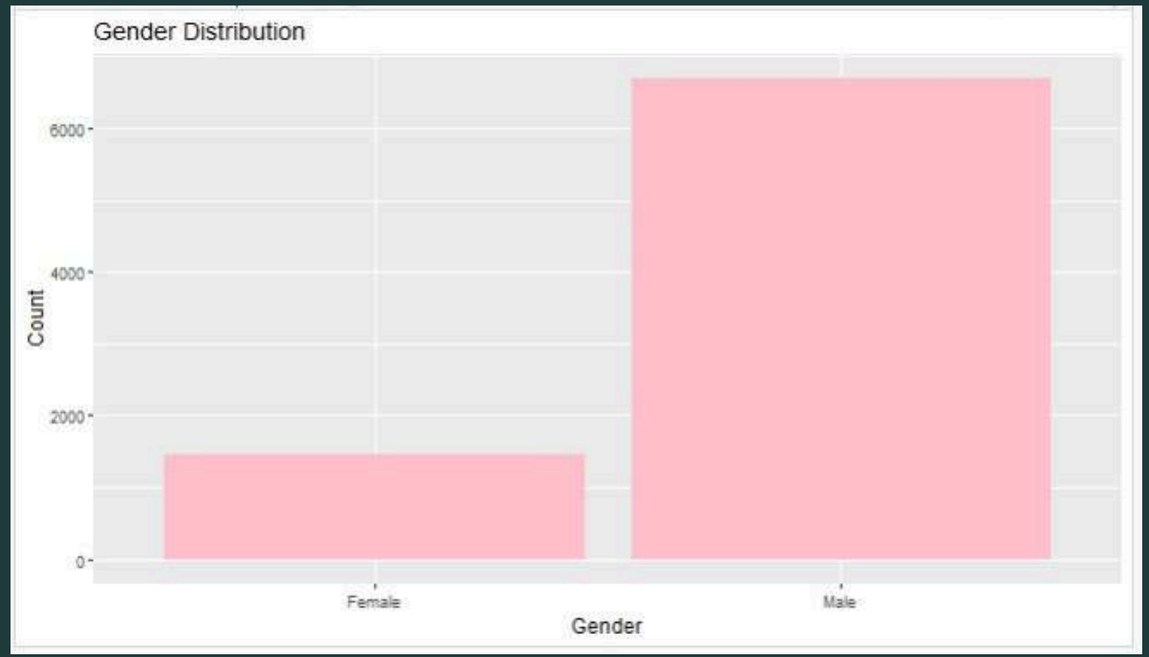
```
ggplot(Data,aes(x=Education,fill=Loan_Status))
+
  geom_bar(position="dodge")+
  facet_wrap(~ Gender)+
  labs(title="Loan Status by Education and
Gender",x="Education Level",y="Count")+
  scale_fill_brewer(palette="Set1")+
  theme_minimal()
```



GENDER DISTRIBUTION

Dominance of Males: The chart highlights a significant gender imbalance, with the male count (over 6,000) being much higher than the female count (around 2,000).

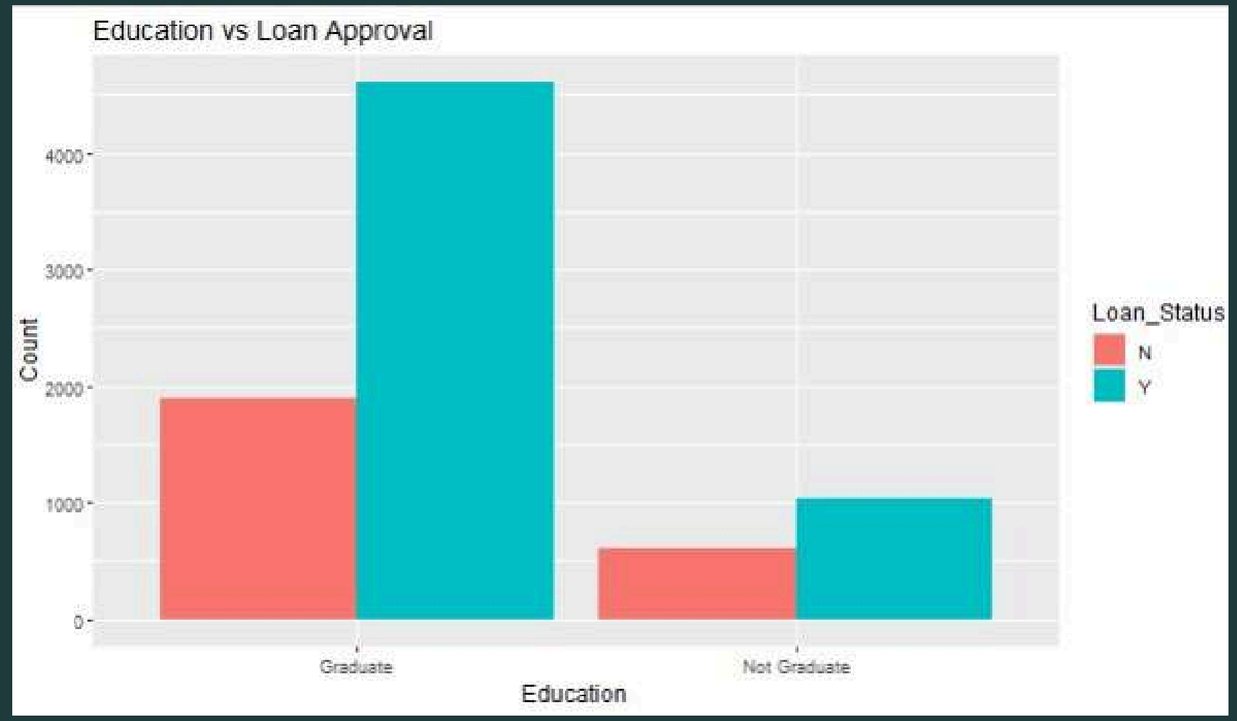
```
ggplot(Data,aes(x=Gender)) +
  geom_bar(fill="pink") +
  labs(title="Gender
Distribution",x="Gender",y="Count")
```



EDUCATION VS LOAN APPROVAL

Dominance of Males: The chart highlights a significant gender imbalance, with the male count (over 6,000) being much higher than the female count (around 2,000).

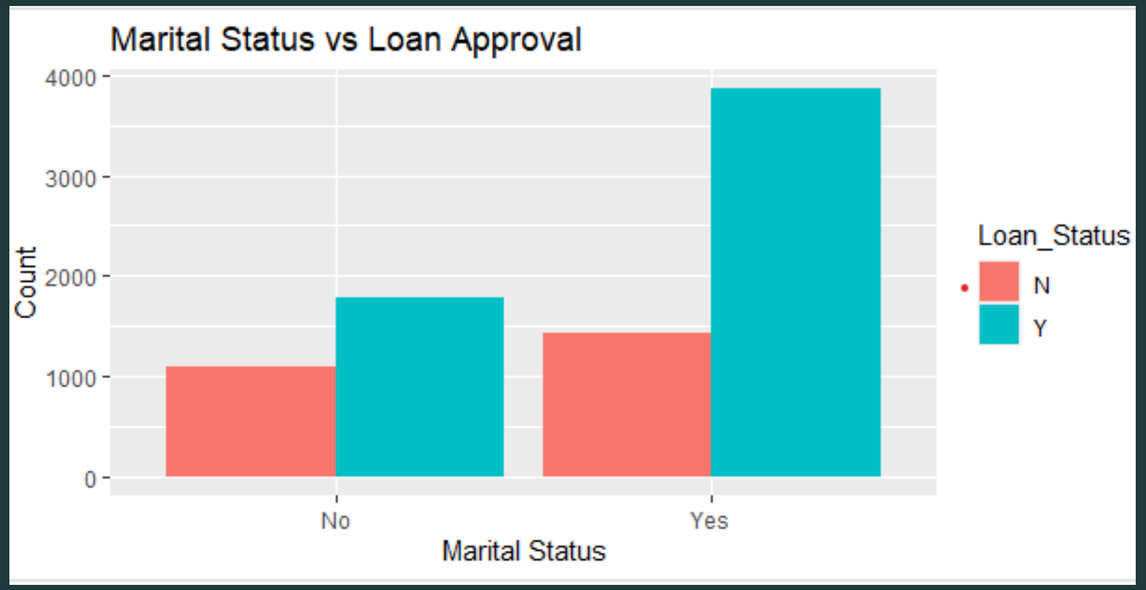
```
ggplot(Data,aes(x=Education,fill=Loan_Status))+
  geom_bar(position="dodge")+
  labs(title="Education vs Loan Approval",x="Education",y="Count")
```



MARITAL STATUS VS LOAN APPROVAL

- Total Applications: The total number of loan applications considered in this dataset is 8160.
- Married Applicants: The majority of loan applications (4000) are from individuals who are married.
- Loan Approvals: A higher proportion of married individuals (5252) received loan approval compared to unmarried individuals (1445).

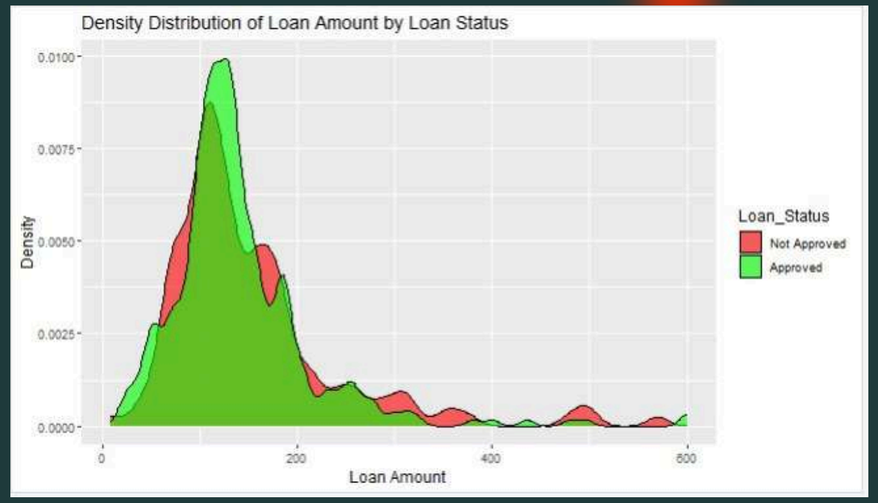
```
ggplot(Data,aes(x=Married,fill=Loan_Status))
+
  geom_bar(position="dodge")+
  labs(title="Marital Status vs Loan
Approval",x="Marital Status",y= "Count")
```



DENSITY DISTRIBUTION OF LOAN AMOUNT BY LOAN STATUS

- Bimodal Distribution: The graph displays two peaks, suggesting a bimodal distribution for both approved and not approved loans. This indicates that there might be two distinct groups of loan amounts within each category.
- Overlap: The distributions of approved and not approved loans overlap significantly. This suggests that loan amount alone might not be a strong predictor of loan approval.
- Loan Amount Range: The majority of loan amounts seem to fall within the range of 0 to 200, with some loans extending up to 600.

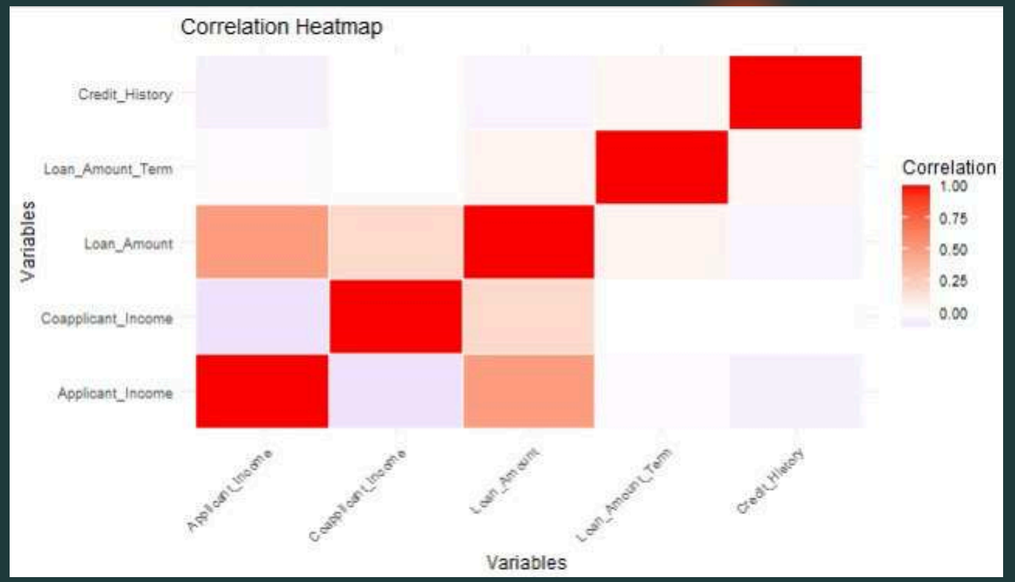
```
ggplot(Data,aes(x=Loan_Amount,fill=Loan_Status)) +
  geom_density(alpha=0.6)+
  labs(title="Density Distribution of Loan Amount by Loan Status",
       x="Loan Amount",
       y="Density")+
scale_fill_manual(values=c("N"="red","Y"="green"),labels=c("Not Approved","Approved"))
```



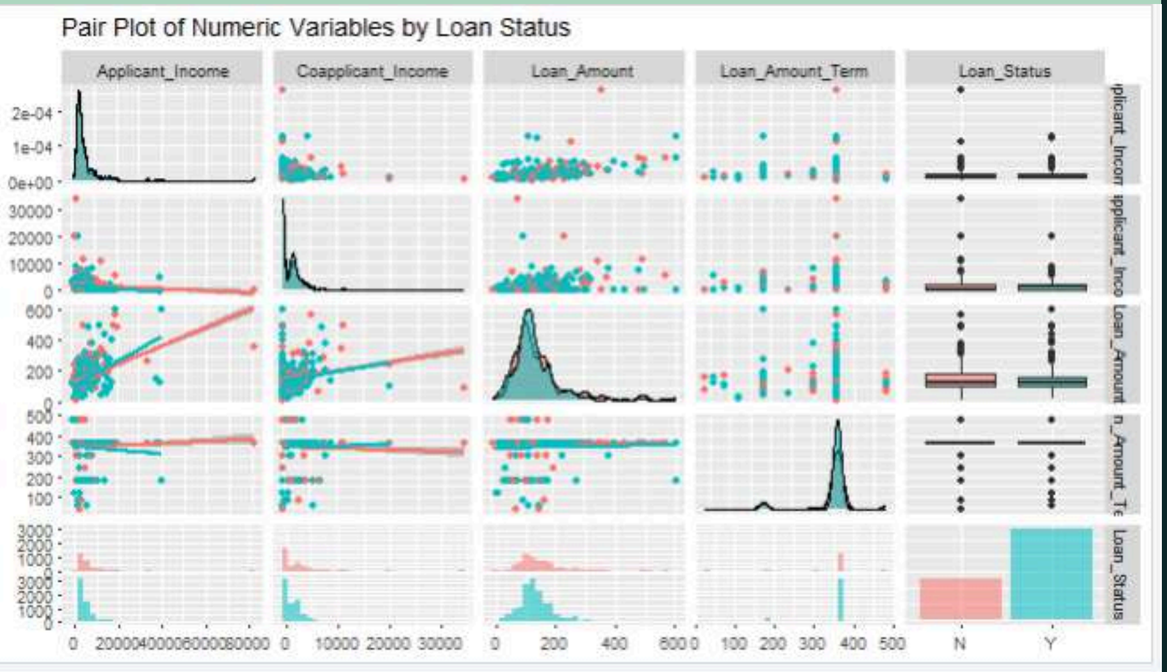
CORRELATION HEATMAP

- Loan Amount Term and Credit History: The most prominent feature is the strong positive correlation between "Loan_Amount_Term" and "Credit_History." This suggests that individuals with longer loan terms tend to have better credit history.
- Loan Amount and Applicant Income: A moderate positive correlation exists between "Loan_Amount" and "Applicant_Income." This indicates that individuals with higher incomes tend to apply for larger loans.
- Weaker Correlations: There appear to be weaker correlations between "Loan_Amount" and "Loan_Amount_Term," and between "Coapplicant_Income" and other variables.

```
ggplot(Cor_Melt,aes(x=Var1,y=Var2,fill=value))+
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red",
mid = "white", midpoint = 0) +
  labs(title = "Correlation Heatmap", x =
"Variables", y = "Variables", fill = "Correlation")
+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45,
hjust = 1))
```



PAIR PLOT:- VARIABLES BY LOAN STATUS



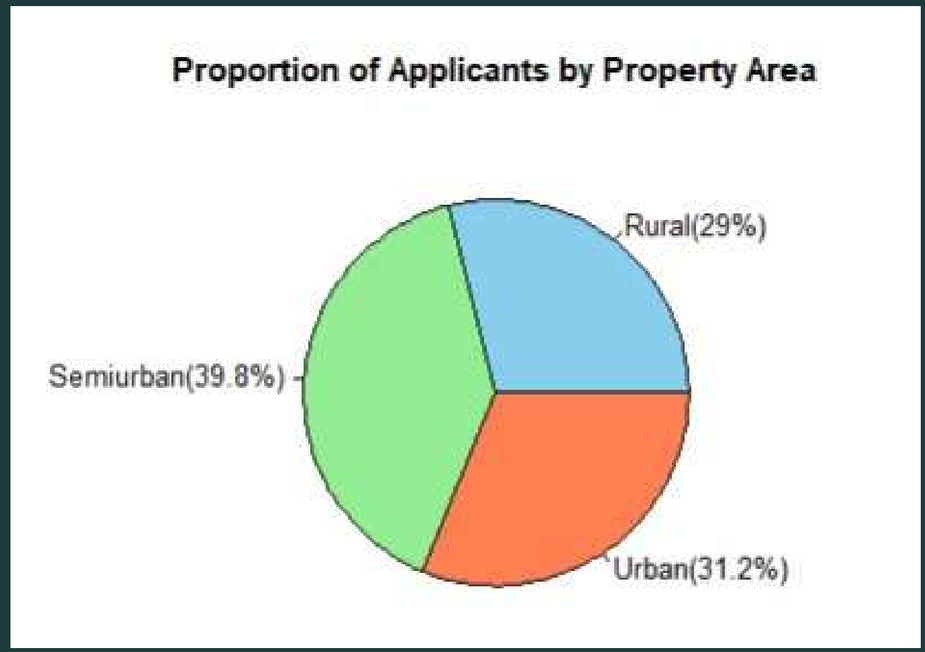
Credit history is a critical determinant in loan approval decisions.
Applicants with a credit history (Credit_History = 1) have a significantly higher probability of loan approval (Loan_Status = Y).
Applicants without a credit history (Credit_History = 0) are more likely to face rejection (Loan_Status = N).

```
ggpairs(
  Selected_Columns,
  aes(color=Loan_Status,alpha=0.6),
  upper=list(continuous="points"),
  lower=list(continuous="smooth")
)+ labs(title="Pair Plot of Numeric Variables by Loan Status")
```

PROPORTION OF APPLICANTS BY PROPERTY AREA

- Semiurban Dominance: The largest proportion of applicants (39.8%) reside in Semiurban areas.
- Rural and Urban: Rural areas account for 29% of applicants, while Urban areas represent 31.2%.

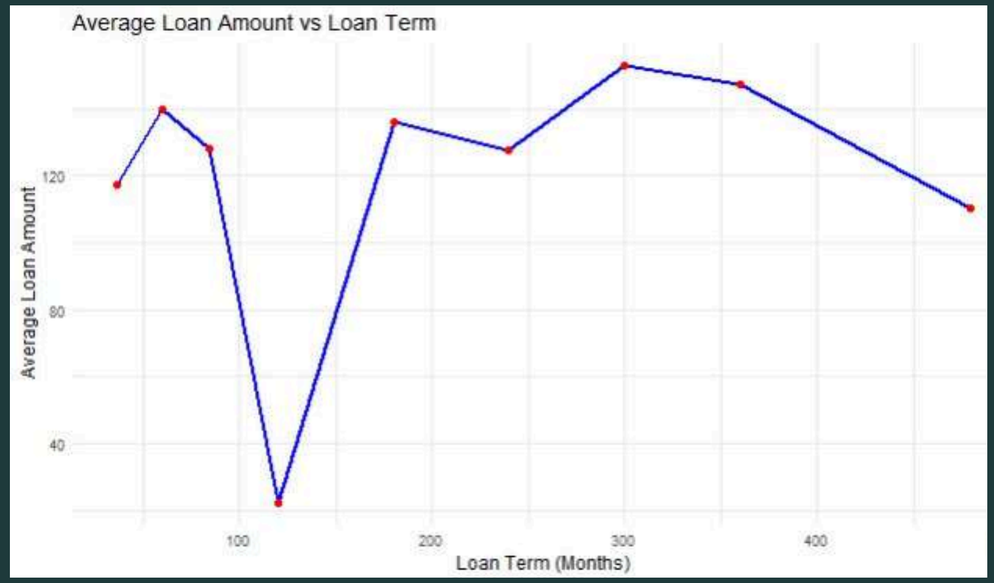
```
pie(Property_Area_Counts,

labels=paste0(names(Property_Area_Counts),"
(",Property_Area_Proportions,"%)"),
    main="Proportion of Applicants by Property
Area",
    col=c("skyblue", "lightgreen", "coral"))
```



CREDIT HISTORY VS LOAN STATUS

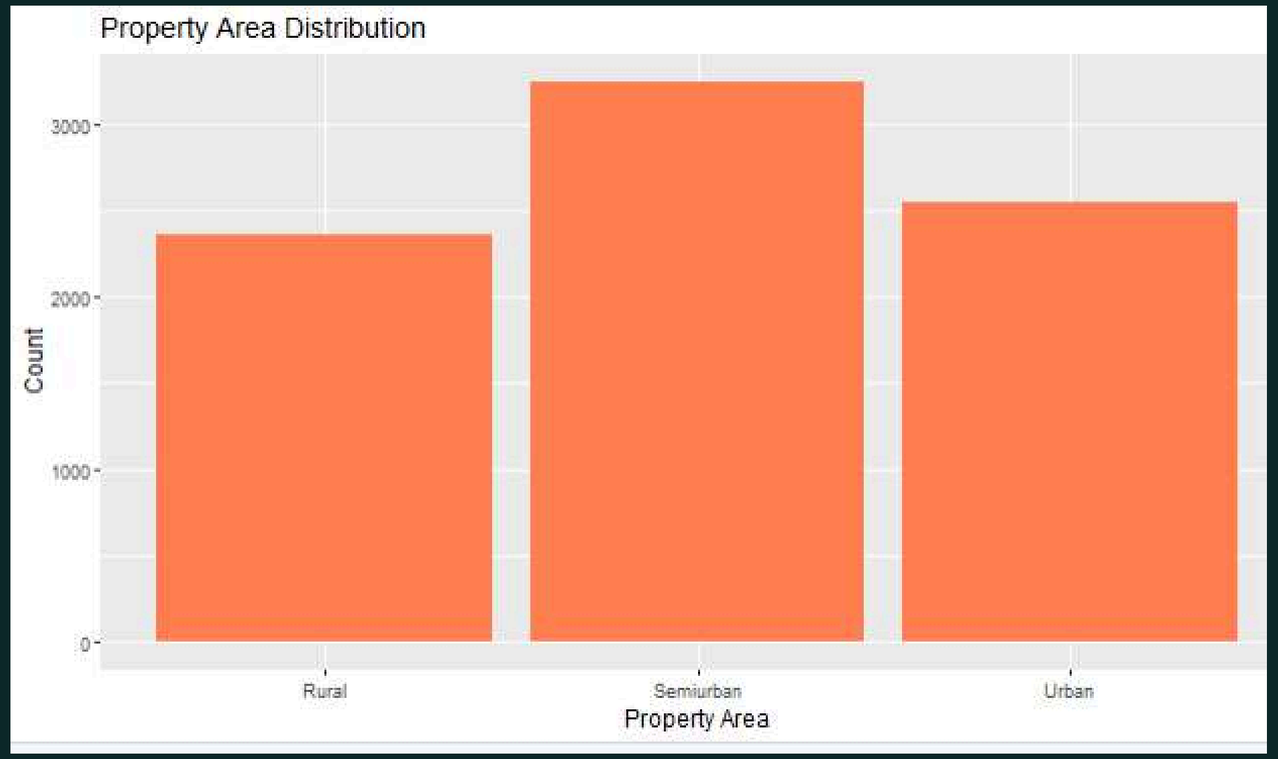
The chart suggests that the relationship between loan term and average loan amount is not straightforward. It's likely influenced by various factors, such as interest rates, borrower risk profiles, and the purpose of the loan.

```
ggplot(Avg_Loan_By_Term,aes(x =
Loan_Amount_Term,y=Average_Loan_Amount))+
  geom_line(color="blue",size=1)+
  geom_point(color="red",size=2)+
  labs(title="Average Loan Amount vs Loan Term",
       x="Loan Term (Months)",
       y="Average Loan Amount") +
  theme_minimal()
```



CREDIT HISTORY VS LOAN STATUS

```
ggplot(Data, aes(x = Property_Area)) +
  geom_bar(fill = "coral") +
  labs(title = "Property Area Distribution", x = "Property Area", y = "Count")
```



01

Highest Concentration: The majority of observations (individuals or loan applications) are concentrated in Semiurban areas.

02

Distribution Pattern: The distribution pattern suggests that there is a higher concentration of individuals or loan applications in Semiurban areas compared to Rural and Urban areas.

OUTLIERS IN APPLICANT INCOME

01

Income Distribution: The box plot shows that the majority of applicant incomes fall within a certain range, with the median income being around 10,000

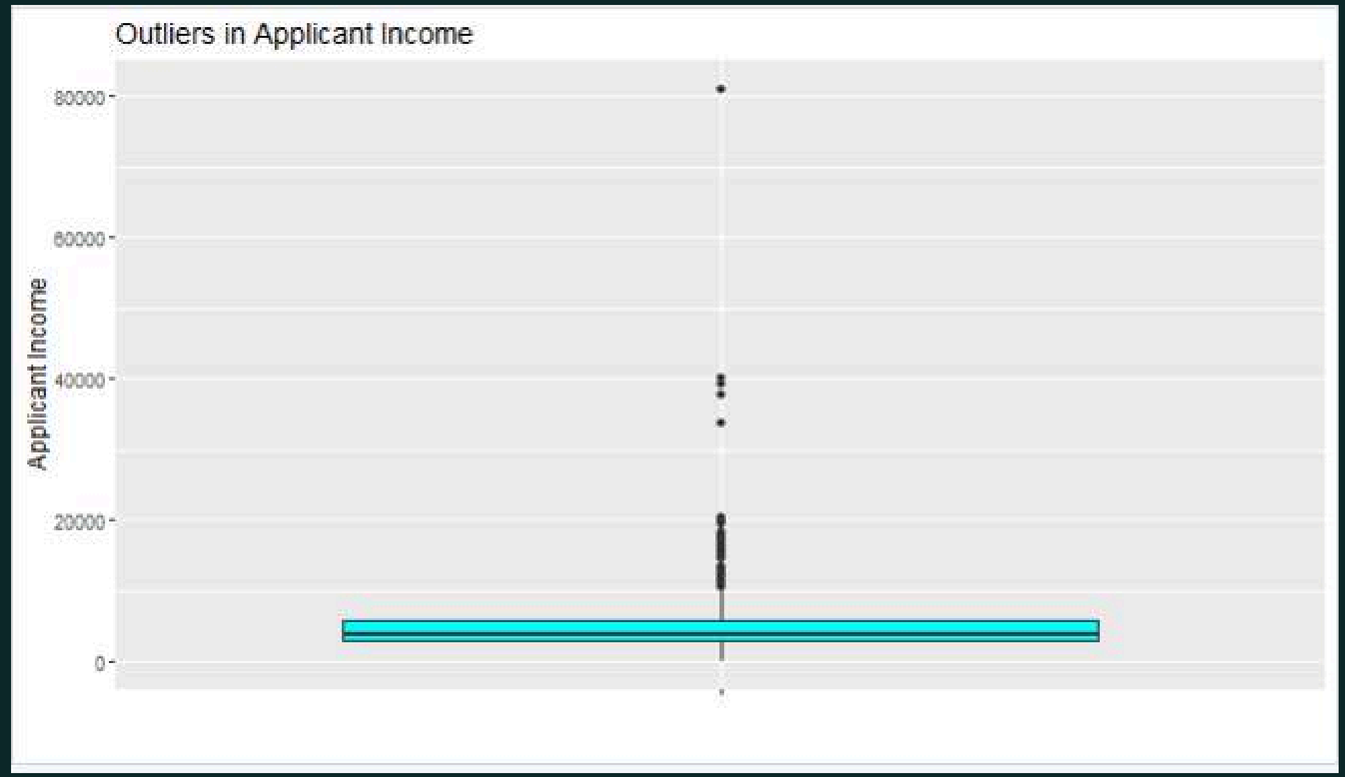
02

Outliers: There are several outliers in the data, indicating that some applicants have significantly higher incomes compared to the majority.

03

Skewness: The longer whisker on the right side suggests that the income distribution might be skewed to the right (positively skewed), meaning there are more applicants with higher incomes than lower incomes.

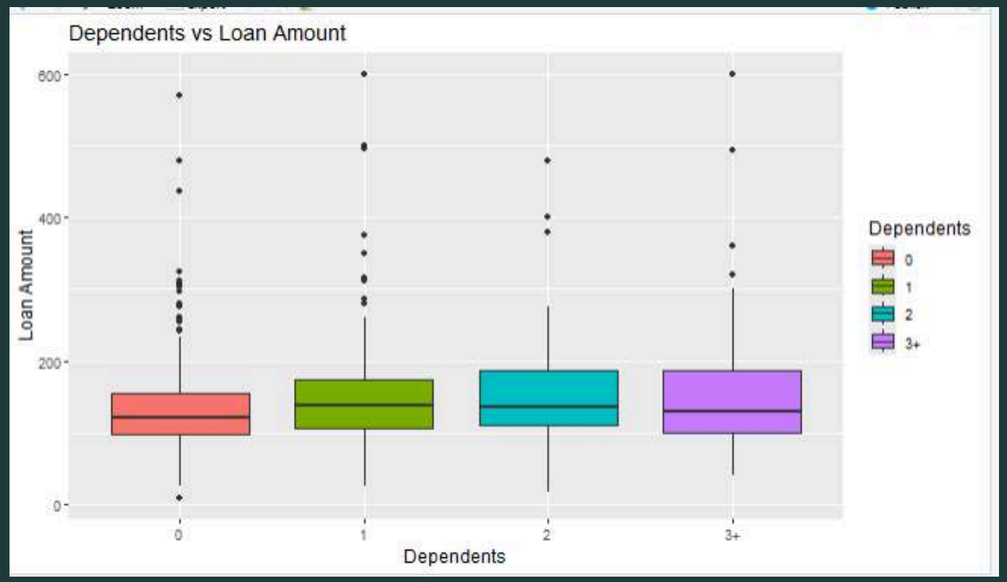
```
ggplot(Data, aes(x = "", y = Applicant_Income)) +
  geom_boxplot(fill = "cyan") +
  labs(title = "Outliers in Applicant Income", x = "", y =
"Applicant Income")
```



CREDIT HISTORY VS LOAN STATUS

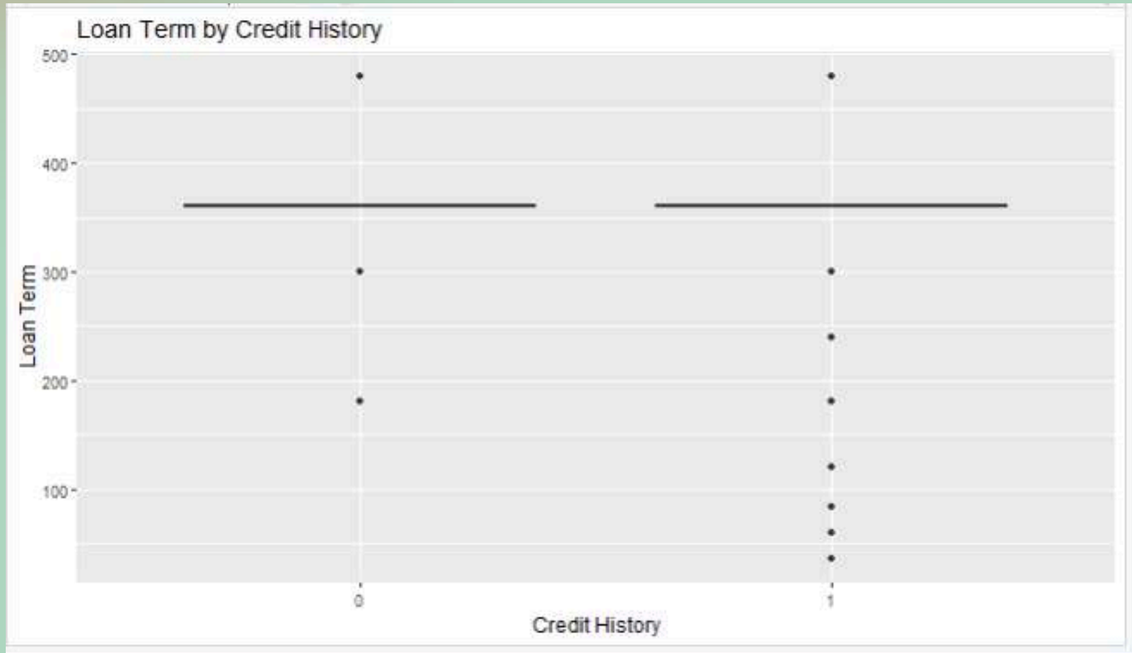
```
ggplot(Data, aes(x = Dependents, y = Loan_Amount, fill = Dependents)) +
  geom_boxplot() +
  labs(title = "Dependents vs Loan Amount", x = "Dependents", y = "Loan Amount")
```

- Distribution: The distribution of loan amounts appears to vary across different numbers of dependents.
- Median Trend: The median loan amount seems to increase with the number of dependents. For instance, the median loan amount for individuals with 3+ dependents is higher than those with 0 dependents.
- Outliers: There are some outliers present in each category, suggesting that there might be some individuals with significantly higher or lower loan amounts compared to the majority in their respective dependent groups.



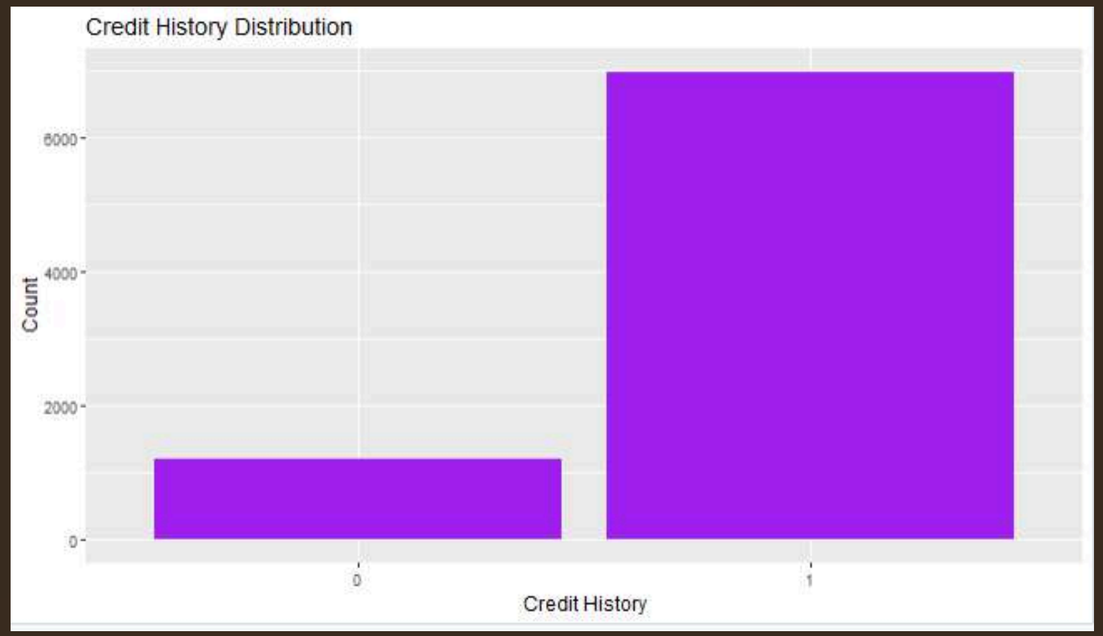
LOAN TERM BY CREDIT HISTORY

```
ggplot(Data, aes(x = as.factor(Credit_History), y = Loan_Amount_Term)) +
  geom_boxplot(fill = "orange") +
  labs(title = "Loan Term by Credit History", x = "Credit History", y = "Loan Term")
```



- Credit History Impact: The graph suggests that credit history plays a role in determining the loan term. Individuals in credit history category "1" appear to have access to longer loan terms compared to those in category "0."
- Loan Term Variation: Within each credit history category, there is a range of loan terms, indicating that other factors besides credit history likely influence the loan term.
- Data Distribution: The data points are clustered around certain loan term values, suggesting that certain loan terms are more common than others.

CREDIT HISTORY VS LOAN STATUS



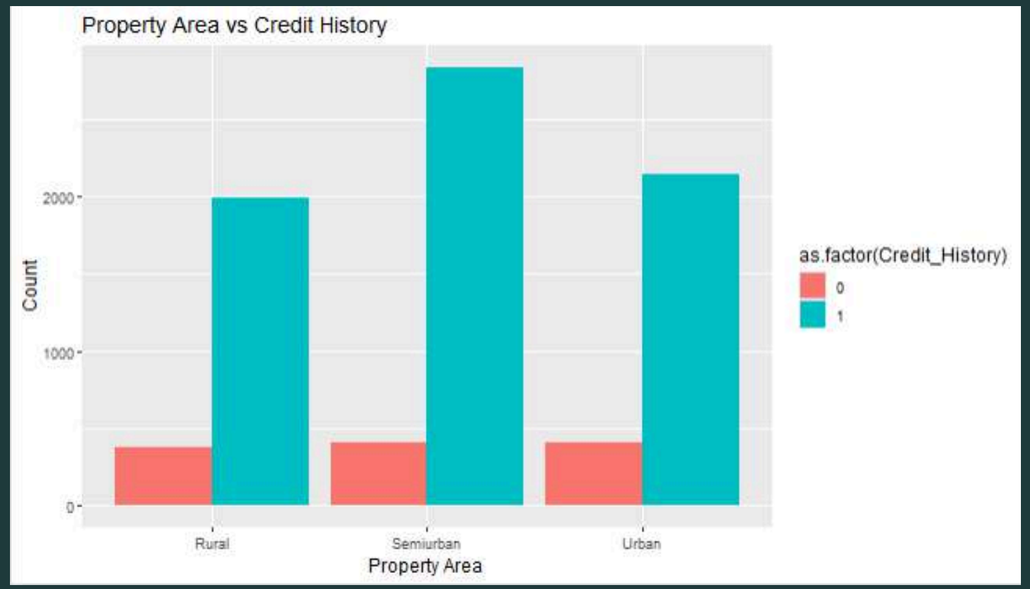
Unequal Distribution: The distribution of observations across credit history categories is highly skewed. A large majority of observations fall into one category (represented by the right bar), while the other category has a much smaller count.

```
ggplot(Data, aes(x = as.factor(Credit_History))) +
geom_bar(fill = "purple") +
labs(title = "Credit History Distribution", x = "Credit History",
y = "Count")
```

CREDIT HISTORY VS LOAN STATUS

- Semiurban Dominance: In Semiurban areas, there is a clear dominance of the blue credit history category. This suggests that a higher proportion of individuals in Semiurban areas have a better credit history compared to Rural and Urban areas.
- Rural and Urban: In Rural and Urban areas, the distribution is less skewed, with a higher proportion of individuals falling into the red credit history category.

```
ggplot(Data, aes(x = Property_Area, fill =
as.factor(Credit_History))) +
  geom_bar(position = "dodge") +
  labs(title = "Property Area vs Credit History", x
= "Property Area", y = "Count")
```



PERFORMING BOSTON ANALYSIS

Encode Categorical Variables as Factors-

- Data$Gender=as.factor(Data$Gender)
- Data$Married=as.factor(Data$Married)
- Data$Dependents=as.factor(Data$Dependents)
- Data$Education=as.factor(Data$Education)
- Data$Self_Employed=as.factor(Data$Self_Employed)
- Data$Property_Area=as.factor(Data$Property_Area)
- Data$Loan_Status=as.factor(Data$Loan_Status)

Analysis-

- Encoded as a factor, even though it contains numbers (e.g. 0, 1, 2, 3+), as the values represent distinct categories rather than continuous data

SPLITING THE DATA INTO 80-20 SETS

```
set.seed(42)
trainIndex=createDataPartition(Data$Loan_Status,p=0.8,list=FALSE)
trainData=Data[trainIndex,]
testData=Data[-trainIndex,]
```

Analysis-

- **The provided code is a data splitting routine, a critical step in preparing a dataset for machine learning or statistical modeling. Below is a detailed analysis of its components, purpose, and implications**

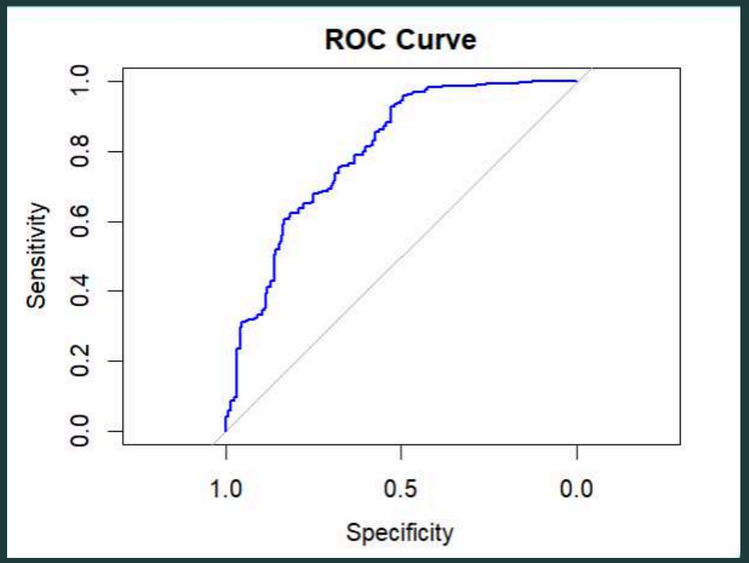
CREATING A MODEL AND EVALUATE IT

```
#Train a logistic regression model
model=glm(Loan_Status ~ ., data = trainData, family = binomial)

#Make predictions on the test set
testData$predicted=predict(model,newdata=testData,type="response")
testData$predicted_class=ifelse(testData$predicted > 0.5, "1", "0")

#Evaluate the model
accuracy=mean(testData$predicted_class==testData$Loan_Status)
roc_curve=roc(as.numeric(testData$Loan_Status),testData$predicted)
auc=auc(roc_curve)

# Print results
print(paste("Accuracy:",round(accuracy, 4)))
print(paste("AUC-ROC:",round(auc, 4)))
```

```
> # Print results
> print(paste("Accuracy:",round(accuracy, 4)))
[1] "Accuracy: NA"
> print(paste("AUC-ROC:",round(auc, 4)))
[1] "AUC-ROC: 0.7992"
>
```

PLOT THE ROC-CURVE

```
# Calculate Accuracy
accuracy=mean(testData$predicted_class==testData$Loan_Status)
print(paste("Accuracy:",round(accuracy,4)))

# Generate a Confusion Matrix
Confusion_Matrix=ConfusionMatrix(as.factor(testData$predicted_class),
testData$Loan_Status)
print(Confusion_Matrix)

# Optional: Plot the ROC Curve
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)
```



RECOMMENDATIONS....

01 **Enhance Data Quality:** Address outliers and imbalances in Loan_Status to improve model robustness.

02 **Model Improvement:** Experiment with advanced algorithms like Random Forest and XGBoost for better predictive performance.

03 **Automation:** Develop an automated pipeline for data preprocessing, analysis, and visualization.

04 **Business Strategy:** Leverage insights to refine loan policies and target high-potential applicants.

CONCLUSION

Summary:-

- Logistic regression effectively predicts loan approval.
- Accuracy and AUC scores suggest the model is reliable for this dataset.

Future Work:-

- Explore other models like Random Forest or Gradient Boosting for improved accuracy.
- Incorporate more features for better prediction.

THANK YOU !!