DANA 4820 Group 5 Project Report

Full Name	Student ID
Chin Wei Shih (David)	#100334127
Zhixuan Zhang (Eric)	#100338057
Chun Ching Look (Cyrus)	#100347726
Simranjit Singh (Simran)	#100348495
Kailash Sukumaran (Kailash)	#100350193

Goal: To Examine the Loan Approval Possibility for each Observation

Dataset Description

Our dataset contains 1 response variable along with 11 explanatory variables. Loan_ID is the identifier and will be excluded in our analysis.

Columns	**Description**
Loan_ID	Unique # assigned to each loan
Gender	Male/Female
Married	Yes/No
Dependents	# of persons depending on client
Education	Education level (graduate/undergrad)
Self_Employed	Yes/No
ApplicantIncome	Applicant income
Coapplicant income	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	Yes/No
Property_Area	Urban, Semi or Rural
Loan_Status	Approved (yes) or Not Approved (No)

The breakdown of the variables are as follows:

1) Which Variable is the dichotomous variable?

Name of Variable	**What are the 2 Categories (levels)?**
Loan_Status	Y, N

2) Our continuous/discrete variables:

Name of Variable	**Continuous or Discrete?**
Applicant Income	Continuous
Co-Applicant Income	Continuous
LoanAmount	Continuous
Loan_Amount_Term	Discrete

3) Our categorical variables

Name of Categorical Variable	**Range**
Education	Graduate, Not Graduate
Gender	Male, Female
Married	Yes, No
Dependents	0, 1, 2, 3+
Self_Employed	No, Yes
Property_Area	Rural, SemiUrban, Urban
Credit_History	0, 1

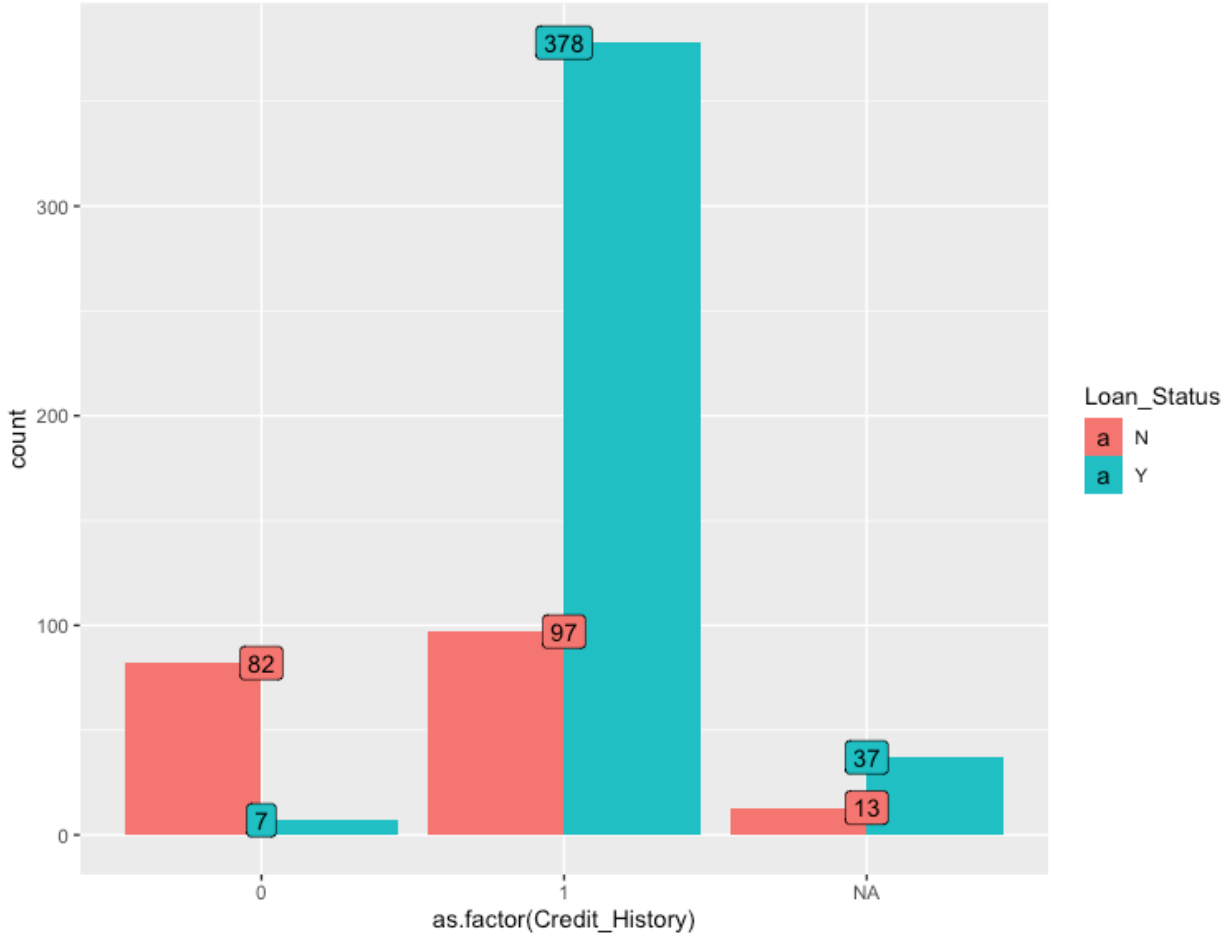
```
    Loan_ID       Gender    Married   Dependents        Education   Self_Employed
LP001002:  1         : 13      :  3     : 15     Graduate    :480      : 32
LP001003:  1   Female:112   No :213   0 :345     Not Graduate:134   No :500
LP001005:  1   Male  :489   Yes:398   1 :102                        Yes: 82
LP001006:  1                           2 :101
LP001008:  1                           3+: 51
LP001011:  1
(Other) :608
ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term Credit_History
Min.    :  150   Min.    :    0     Min.    :  9.0   Min.    : 12      Min.   :0.0000
1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.0   1st Qu.:360      1st Qu.:1.0000
Median : 3812   Median : 1188     Median :128.0   Median :360      Median :1.0000
Mean   : 5403   Mean   : 1621     Mean   :146.4   Mean   :342      Mean   :0.8422
3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:168.0   3rd Qu.:360      3rd Qu.:1.0000
Max.   :81000   Max.   :41667     Max.   :700.0   Max.   :480      Max.   :1.0000
                                  NA's   :22      NA's   :14       NA's   :50
  Property_Area Loan_Status
Rural    :179   N:192
Semiurban:233   Y:422
Urban    :202
```

Data Cleaning

Handling Missing Values

Credit History (0 or 1)

There are total 50 Null values in the credit history column

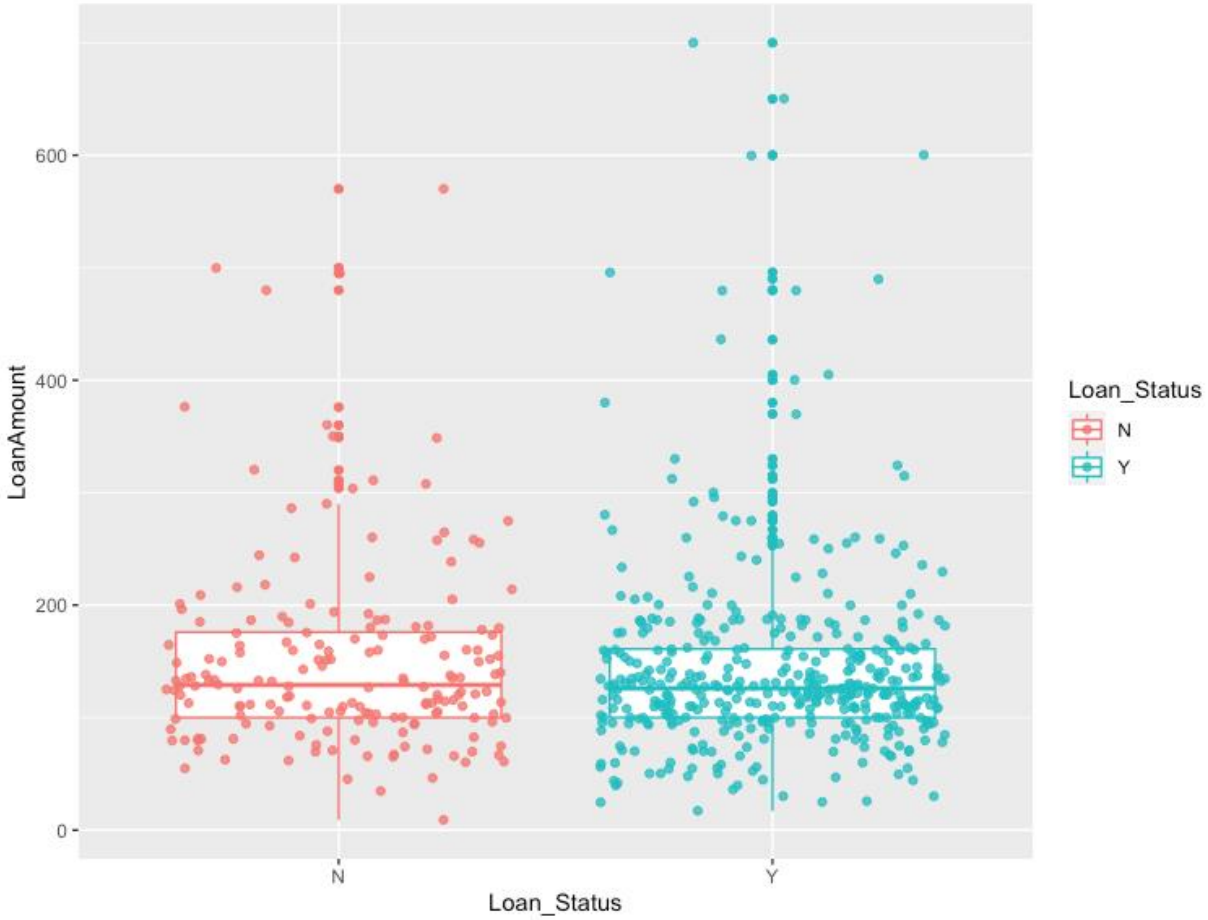


For Credit_History, we took a look at the overall pattern in the data set. Out of the 89 variables with no (0) credit history, only 7 had their loan approved. We decided to take a look at the NA's, if their loan was approved, we assigned them as yes (1) for credit history. If their loan was not approved, we assign the NA as no (0).

```
loan_data$Credit_History <- ifelse(loan_data$Credit_History %in% c(0, 1),
                                   loan_data$Credit_History,
                                   ifelse(loan_data$Loan_Status == 'Y', 1, 0))
```

Loan Amount

There are a total 22 missing values in the Loan Amount column. In the real life approval of the loan is dependent on the loan amount. As per this dataset, it is clear from the below plot that approval of a loan is not dependent on the Loan Amount:



We can even verify it using one-way anova:

```
> res.aov <- aov(LoanAmount ~ Loan_Status,loan_data[!is.na(loan_data$LoanAmount),])
> summary(res.aov)
             Df  Sum Sq Mean Sq F value Pr(>F)
Loan_Status   1    6029    6029   0.823  0.365
Residuals   590 4323159    7327
```

Null hypothesis - Mean value for loan status Yes and No is equal
Alternate hypothesis - Mean is not equal
As p-value is not significant, we fail to reject the null hypothesis, and cannot conclude that the mean is unequal.

```
> res.aov <- aov(LoanAmount ~ Property_Area,loan_data[!is.na(loan_data$LoanAmount),])
> summary(res.aov)
              Df  Sum Sq Mean Sq F value Pr(>F)
Property_Area   2    9495    4747   0.647  0.524
Residuals     589 4319693    7334
```

Above is another anova test for Loan Amount with Property_Area.

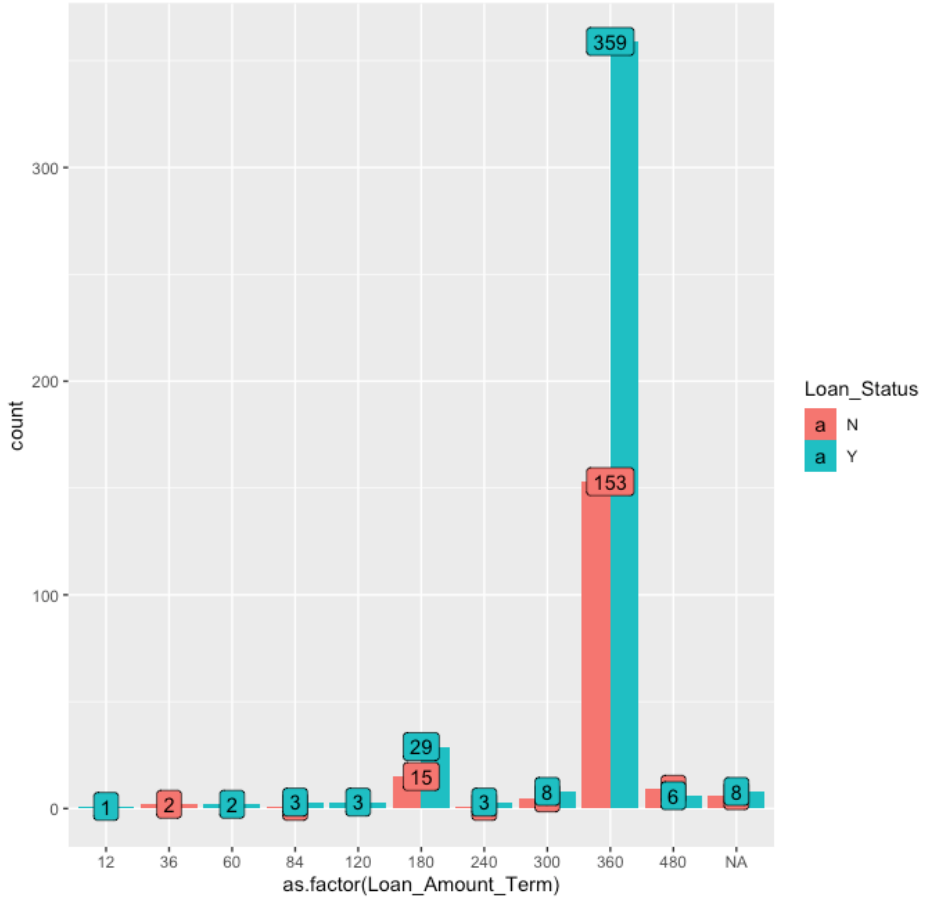
It is clear from the test that p-value is not significant. So we cannot conclude that loan amount and Property_Area are dependent.

As we can see Loan Amount has no relationship with any of the variables. We can simply use group_by Loan status to replace null values of loan amount with the median value. So that there wouldn't be any outlier problem.

```
> loan_data <- loan_data %>%
+   group_by(Loan_Status) %>%
+   mutate(LoanAmount = ifelse(
+     is.na(LoanAmount) ,
+     median(LoanAmount, na.rm = TRUE),
+     LoanAmount
+   ))
```

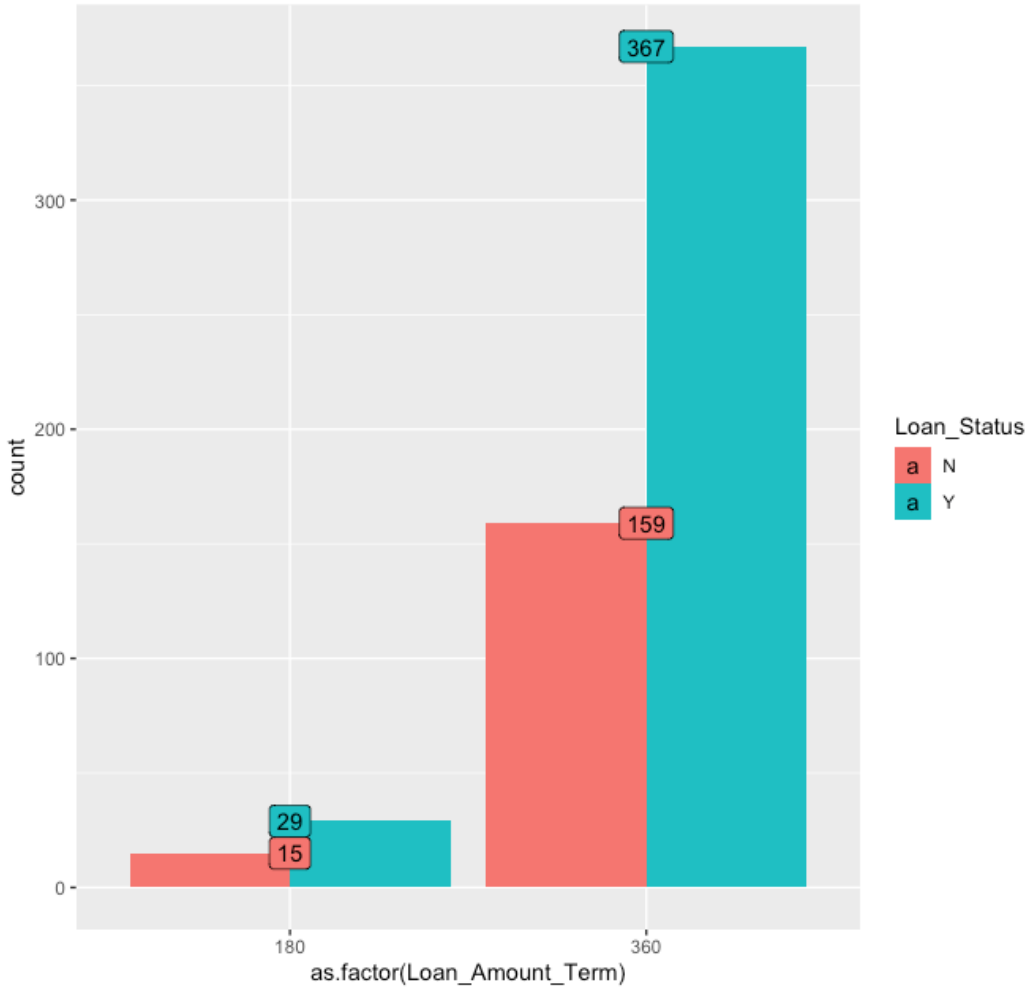
Loan_amount_term

There are a total 14 null values in the loan amount term.



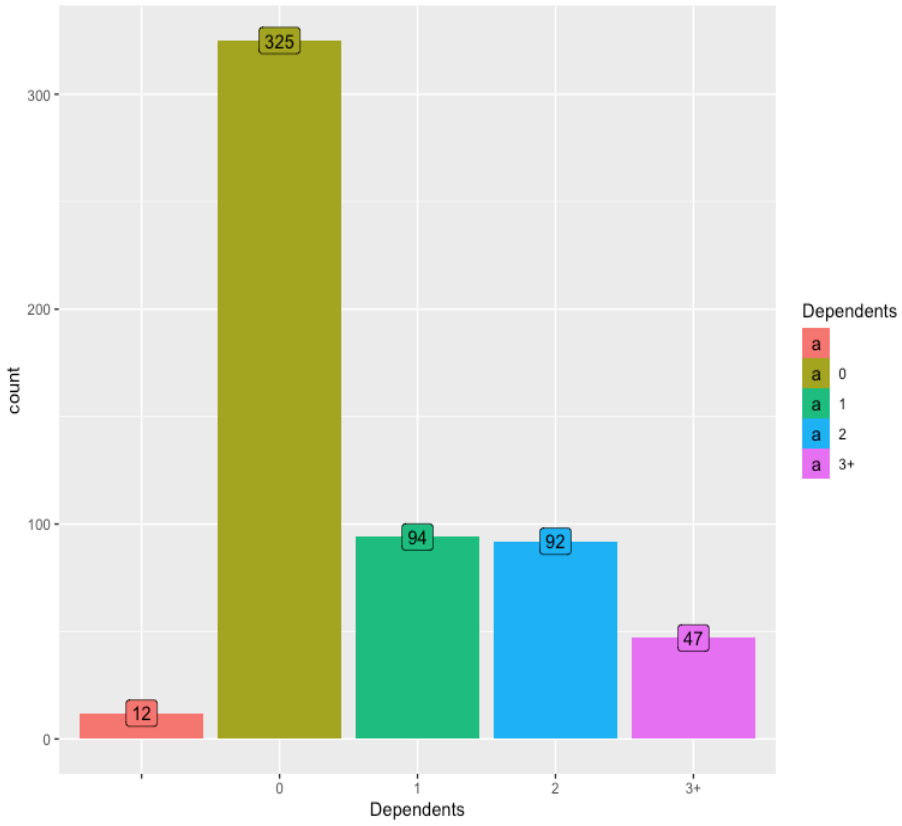
We have considered this variable as a discrete. It is clear from the above plot that frequency of values except 180 & 360 (days) are significantly low. To clarify, the data points are enough to train our model. So we can simply remove the rows which don't have values 180 or 360.

Data for loan_term after removing insignificant rows:



Dependents (0, 1, 2, 3+)

There are a total 12 missing values in the dependents. In this case, we believe missing values means no dependents. Moreover most of the applicants have 0 dependents (refer to the below plot). So we can simply fill the missing values with 0.



Married (Yes or No)

There is one missing value in the married column. As value of dependent is 0 for that person, we can conclude that the person is not married

Gender (Male or Female) - 12 missing values

Most of the applicants are male. Moreover, Gender is not dependent on our response variable(chi-square is proof). We can simply replace the missing values with male.

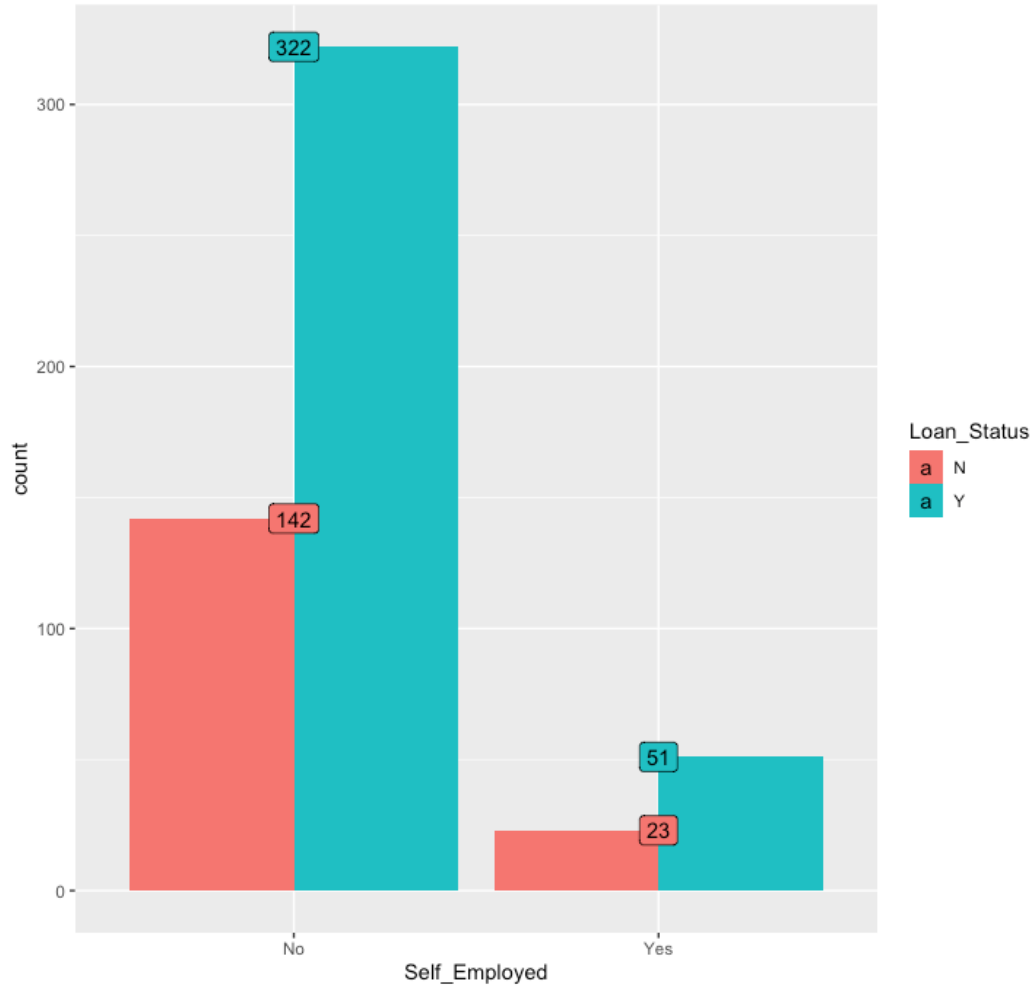
```
> chisq.test(loan_data$Gender, loan_data$Loan_Status)

        Pearson's Chi-squared test

data:  loan_data$Gender and loan_data$Loan_Status
X-squared = 0.84302, df = 2, p-value = 0.6561
```

Self Employed (Yes or NO)

There are a total 32 missing values in the Self Employed. Moreover, it is clear from the below plot (and chi-square test) that loan status is not dependent on the criteria whether a person is self-employed or not.



```
> chisq.test(loan_data$Self_Employed, loan_data$Loan_Status)

        Pearson's Chi-squared test

data:  loan_data$Self_Employed and loan_data$Loan_Status
X-squared = 0.099047, df = 2, p-value = 0.9517
```

Significance Test (Variance test, 2 sample t-test, Chi-Square test)

Chi-Square Test

The Chi-Square Test is used in our dataset to determine the association between the response variable and categorical variables.

Categorical Variables	Chi-Square Value	Degree of Freedom	P-Value	Dependency
Gender	0.03	1	0.9	No
Married	2	1	0.2	No
Dependents	3	3	0.5	No
Self_Employed	Close to 0	1	1	No
Education	4	1	0.05	Yes
Loan_Amount_Term	0.1	1	0.7	No
Credit_History	204	1	Close to 0	Yes
Property_Area	15	2	0.0006	Yes

As we can see from the table, in total eight categorical variables, there are only three of them that have dependency with our response variable - loan_status.

Variance Test & Two Sample T-Test

We used the variance test to compare the variances for numeric variables in two groups of loan_status. ApplicantIncome column and CoapplicatIncome column have unequal variances, LoanAmount column has equal variance.
Then we check using Two Sample T-Test to check whether the means in two groups are equal for the numeric variables. According to the P-value, all of them have equal means.

Numeric Variables	P-Value	Equal Variances	P-Value	Equal Means
LoanAmount	0.05322	Yes	0.1607	Yes
ApplicantIncome	Close to 0	No	0.6098	Yes
CoapplicantIncome	Close to 0	No	0.1792	Yes

Logistic Regression

No multicollinearity found in our continuous/discrete variables:

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term
ApplicantIncome	1.00000000	-0.11660458	0.56593666	-0.04532916
CoapplicantIncome	-0.11660458	1.00000000	0.18784494	-0.05974385
LoanAmount	0.56593666	0.18784494	1.00000000	0.03900914
Loan_Amount_Term	-0.04532916	-0.05974385	0.03900914	1.00000000

Split and Partition Data to 2 Parts (Train & Test)

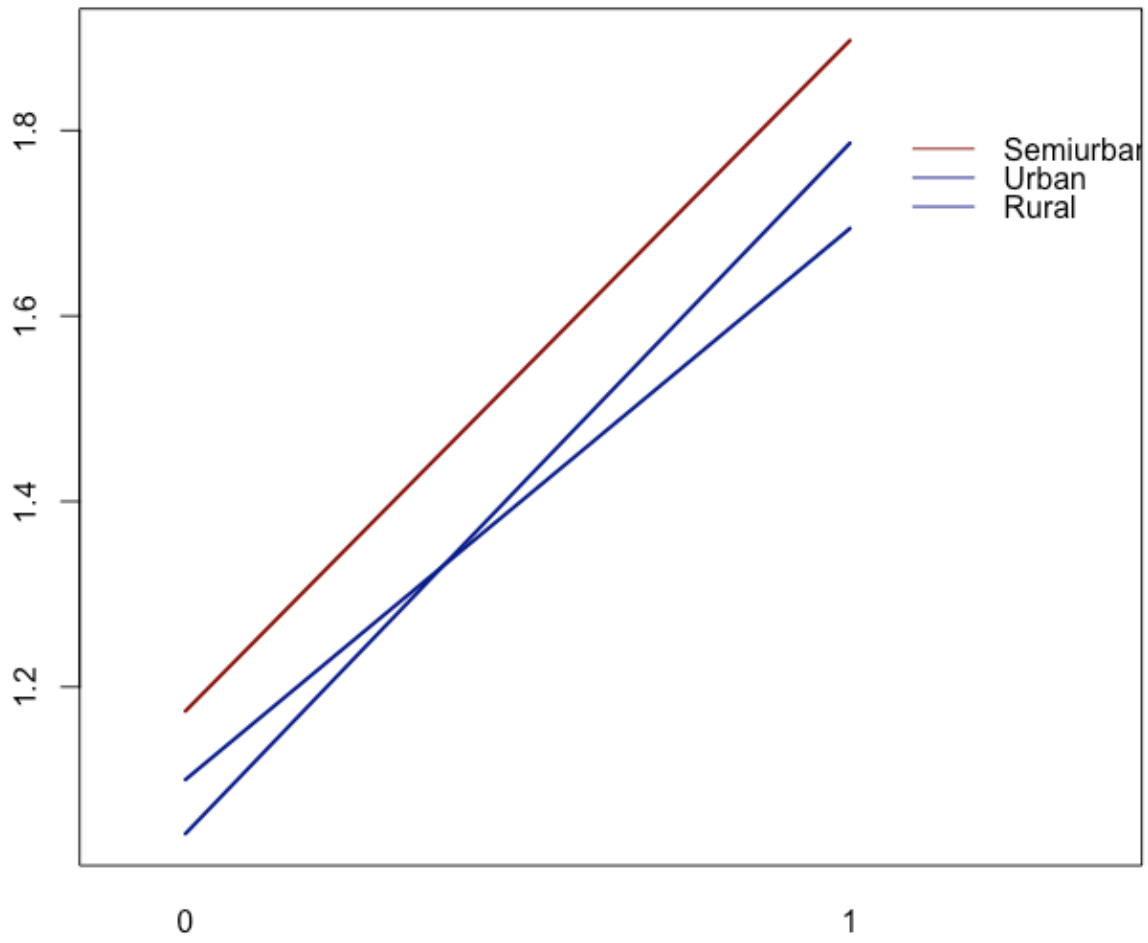
Variable Selection using train data:

After performing stepwise regression, only two predictors are selected (Main effects):

1. Credit_History (Categorical)
2. Property_AreaSemiurban (Categorical)
3. Property_AreaUrban (Categorical)

Interaction Terms:

It seems like the Credit History & Property Area has an interaction effect with loan status:



Model with interaction effects

1. Credit_History (Categorical)
2. Property_AreaSemiurban (Categorical)
3. Property_AreaUrban (Categorical)
4. Credit_History:Property_AreaSemiurban (Categorical)
5. Credit_History:Property_AreaUrban (Categorical)

Likelihood Ratio Test

Apply Likelihood ratio test to check goodness of fit and select best between 2 models:
Residual deviance of the model without interaction = 338.36 & df = 424
Residual deviance of the model with interaction term = 387.08 & df = 422

Wald Test

We used the Wald test to find the significance of each parameter in our final model. In total, there are four parameters in our final model including the intercept term. We use wald.test() function from R 'aod' library. According to the test result, only two out of three parameters are found significant in the final model which are: Credit_History and Property_AreaSemiurban. Property_AreaUrban is proved to be not significant by the Wald test. The result is shows below:

Parameter	X2	P-value	Significant
Credit_History	72	0	Yes
Property_AreaSemiurban	15.3	0.00009	Yes
Property_AreaUrban	1.8	0.18	No

Comparing two models

```
> pchisq(388.36 - 387.08, 424 - 422, lower.tail = F)
[1] 0.5272924
```

Null Hypothesis = Interaction terms have no effect on the model
Alternate Hypothesis = Interaction terms have significant effect on the model

It is clear from the chi-square test that p-value > 0.05, we cannot reject Null Hypothesis and conclude that Interaction terms have not effect on the model

Classification

For training dataset:

```
> confusion_matrix_train
         Actual
Predicted   N   Y
        0  61   7
        1  71 289
> sensitivity <- confusion_matrix_train[2,2]/(confusion_matrix_train[2,2] + confusion_matrix_train[1,2]);sensitivity
[1] 0.9763514
> specificity <- confusion_matrix_train[1,1]/(confusion_matrix_train[1,1] + confusion_matrix_train[2,1]);specificity
[1] 0.4621212
> sum(diag(confusion_matrix_train))/sum(confusion_matrix_train)
[1] 0.817757
```

Sensitivity = 97%;
Specificity = 47%;
Accuracy = 80%

For test dataset:

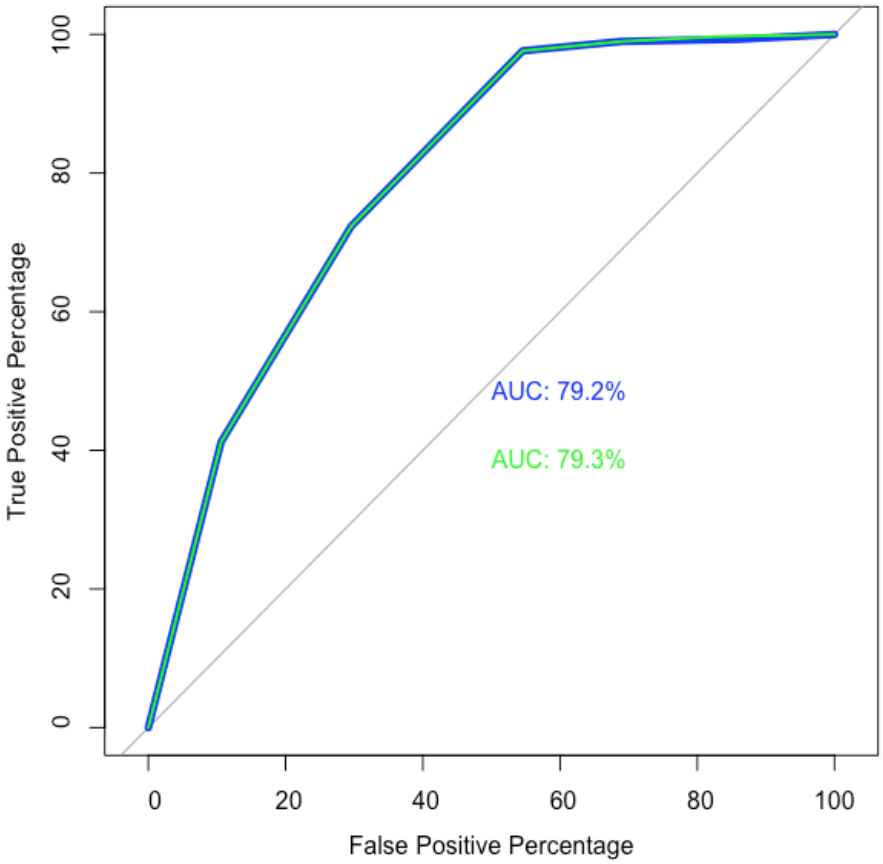
```
> confusion_matrix_test
         Actual
Predicted  N  Y
        0 28  1
        1 14 99
> sensitivity <- confusion_matrix_test[2,2]/(confusion_matrix_test[2,2] + confusion_matrix_test[1,2]);sensitivity
[1] 0.99
> specificity <- confusion_matrix_test[1,1]/(confusion_matrix_test[1,1] + confusion_matrix_test[2,1]);specificity
[1] 0.6666667
> Accuracy <- sum(diag(confusion_matrix_test))/sum(confusion_matrix_test);Accuracy
[1] 0.8943662
```

Sensitivity = 99%;
Specificity = 67%;
Accuracy = 89%

It is clear from the result that our model's accuracy, sensitivity & specificity is more for test dataset. It means our model is well generalized for unseen data.

ROC curve for evaluation

Comparison of ROC (model with and without interaction):



For this dataset, we want the specificity rate to be higher. In other words, a person who is not eligible for a loan, shouldn't have his loan approved. According to the roc curve, 0.77 would be the best threshold.

Train dataset

```
> sum(diag(confusion_matrix_train))/sum(confusion_matrix_train)
[1] 0.7172897
> sensitivity <- confusion_matrix_train[2,2]/(confusion_matrix_train[2,2] + confusion_matrix_train[1,2]);sensitivity
[1] 0.722973
> specificity <- confusion_matrix_train[1,1]/(confusion_matrix_train[1,1] + confusion_matrix_train[2,1]);specificity
[1] 0.7045455
> confusion_matrix_train
         Actual
Predicted   N   Y
        0  93  82
        1  39 214
```

Sensitivity = 72%; Specificity = 70%; Accuracy = 71%

Test dataset

```
> sensitivity <- confusion_matrix_test[2,2]/(confusion_matrix_test[2,2] + confusion_matrix_test[1,2]);sensitivity
[1] 0.71
> specificity <- confusion_matrix_test[1,1]/(confusion_matrix_test[1,1] + confusion_matrix_test[2,1]);specificity
[1] 0.7380952
> Accuracy <- sum(diag(confusion_matrix_test))/sum(confusion_matrix_test);Accuracy
[1] 0.7183099
> confusion_matrix_test
         Actual
Predicted  N  Y
        0 31 29
        1 11 71
```

Sensitivity = 71%; Specificity = 73%; Accuracy = 71%

Hosmer-Lemshow Test

Hosmer-Lemshow test inspects if the observed event rates match expected event rates in subgroups of the model population. We want to cross check whether our training and test models are distributed across 10 deciles similar to actual event rates.

H_0: Actual and predicted event rates are similar across 10 deciles
H_a: Actual and predicted event rates are different across 10 deciles

For the final model using train dataset:

```
> library(generalhoslem)
> logitgof(train$Loan_Status, fitted(final_model_train), g=10, ord=FALSE)

        Hosmer and Lemeshow test (binary model)

data:  train$Loan_Status, fitted(final_model_train)
X-squared = 0.067445, df = 2, p-value = 0.9668
```

The p-value is 0.9668 which is greater than 0.05. Hence, we fail to reject H_0. We can conclude that our training model fits the data and the goodness of fit is high.

For the final model using test dataset:

```
> logitgof(test$Loan_Status, fitted(final_model_test), g=10, ord=FALSE)

        Hosmer and Lemeshow test (binary model)

data:  test$Loan_Status, fitted(final_model_test)
X-squared = 4.807e-08, df = 2, p-value = 1
```

The p-value is 1 which is greater than 0.05. Hence, we do fail to reject the H_0. We can conclude that our test model fits the data and the goodness of fit is high.

Final Model

To determine loan status, we only need 2 explanatory variables which are credit history and property area.

Complete Equation

logit(π(x)) = -2.8549 + 3.7160*Credit_history + 1.3024 * Semiurban + 0.3970* Urban

Coefficients Interpretation

The impact of having credit history on logit(π(x)) is 3.716, the multiplicative effect on the odds is e^3.716=41.1.
The impact of living in a semi urban area on logit(π(x)) is 1.3024, the multiplicative effect on the odds is e^1.3024=3.68.
The impact of living in the urban area on logit(π(x)) is 0.397, the multiplicative effect on the odds is e^0.397=1.49.

```
> confint(final_model)
Waiting for profiling to be done...
                           2.5 %     97.5 %
(Intercept)            -3.8485351 -2.0010822
Credit_History          2.9216930  4.6593724
Property_AreaSemiurban  0.6651354  1.9747670
Property_AreaUrban     -0.1837022  0.9842046
```

Confidence Interval

The 95% Wald confidence interval for the odds ratio corresponding to having credit history is (2.92169, 4.65937), the impact on the odds is e^(2.92169,4.65937).
The 95% Wald confidence interval for the odds ratio corresponding to living in the semi-urban area is (0.66514, 1.97477), the impact on the odds is e^(0.66514,1.97477).
The 95% Wald confidence interval for the odds ratio corresponding to living in the urban area is (-0.18370, 0.98420), the impact on the odds is e^(-0.18370,0.98420).