7/25/2022

Sumit Kumar Kanaujiya

roll no. 215280025

PROJECT REPORT: LOGISTIC REGRESSION MODEL

**PROJECT GUIDE: PROF. SANJEEV V. SABNIS**

DATA INFORMATION

Data columns (total 13 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Loan\_ID 614 non-null object

1 Gender 601 non-null object

2 Married 611 non-null object

3 Dependents 599 non-null object

4 Education 614 non-null object

5 Self\_Employed 582 non-null object

6 ApplicantIncome 614 non-null int64

7 CoapplicantIncome 614 non-null float64

8 LoanAmount 592 non-null float64

9 Loan\_Amount\_Term 600 non-null float64

10 Credit\_History 564 non-null float64

11 Property\_Area 614 non-null object

12 Loan\_Status 614 non-null object

Conclusion

1. 614 Observations & 13 features
2. Loan\_ID is an unnecessary feature.
3. Challenge1 - Missing values present in the dataset.
4. Challenge2 - values of 'Dependents' are numerical type but its datatype is an object.

Challenge1 -Determine the Missing values present in the dataset.

**Loan\_ID False Loan\_ID 0**

**Gender True Gender 13**

**Married True Married 3**

**Dependents True Dependents 15**

**Education False Education 0**

**Self\_Employed True Self\_Employed 32**

**ApplicantIncome False ApplicantIncome 0**

**CoapplicantIncome False CoapplicantIncome 0**

**LoanAmount True LoanAmount 22**

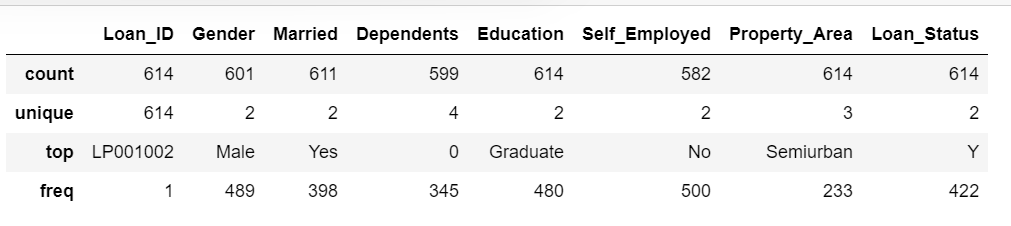
**Loan\_Amount\_Term True Loan\_Amount\_Term 14**

**Credit\_History True Credit\_History 50**

**Property\_Area False Property\_Area 0**

**Loan\_Status False Loan\_Status 0**

**dtype: bool dtype: int64**

 Overall Information about the categorical variables

The Unique values of categorical variables are:

* **Gender :**

1. **Male b) Female 3) None**

* **Married:**

1. **No b) Yes c) None**

* **Dependents:**

1. **0 b) 1 c) 2 d) 3+**

* **Education:[**

1. **Graduate b) Not Graduate**

* **Self\_Employed:**

1. **No b) Yes c) None**

* **Property\_Area:**

1. **Urban b) Rural c) Semi-Urban**

* **Loan\_Status:**

1. **Yes='Y' b) No='N'**

Overall information about the value counts

* **GENDER**

Male 489

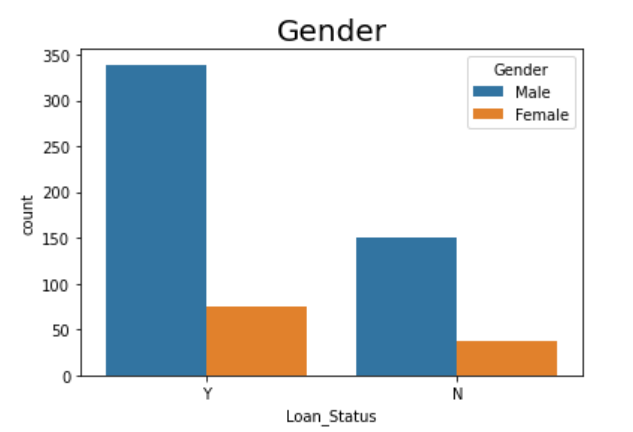
Female 112

Name: Gender, dtype: int64

Male 0.813644

Female 0.186356

Name: Gender, dtype: float64

****

* **MARRIED**

Yes 398

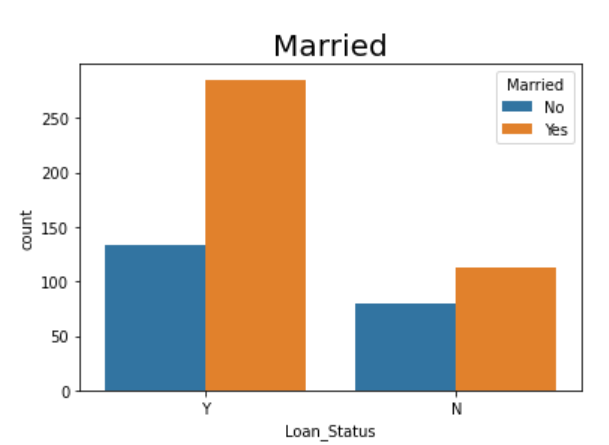
No 213

Name: Married, dtype: int64

Yes 0.651391

No 0.348609

Name: Married, dtype: float64



* **DEPENDENTS**

0 345

1 102

2 101

3+ 51

Name: Dependents, dtype: int64

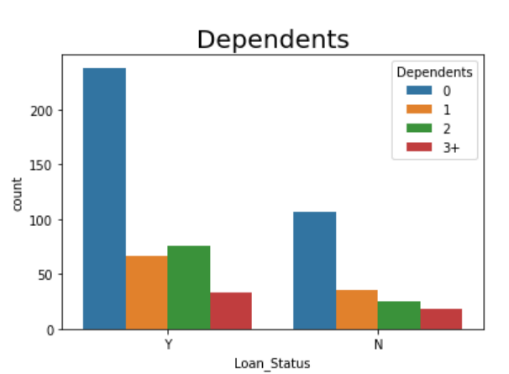
0 0.575960

1 0.170284

2 0.168614

3+ 0.085142

Name: Dependents, dtype: float64



* **EDUCATIONS**

Graduate 480

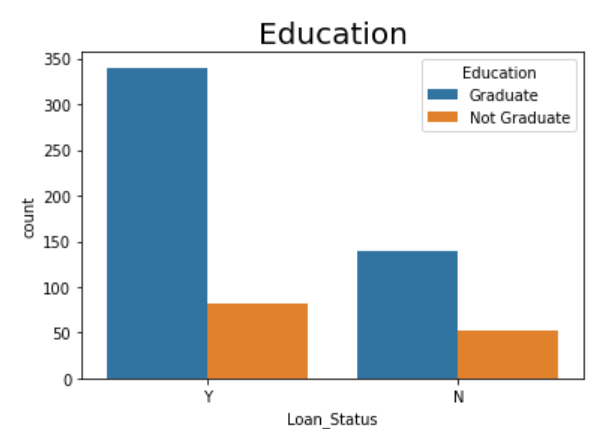
Not Graduate 134

Name: Education, dtype: int64

Graduate 0.781759

Not Graduate 0.218241

Name: Education, dtype: float64



* **SELF-EMPLOYED**

No 500

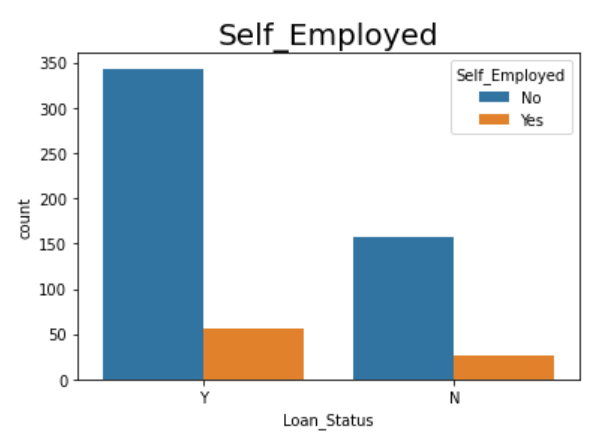
Yes 82

Name: Self\_Employed, dtype: int64

No 0.859107

Yes 0.140893

Name: Self\_Employed, dtype: float64



* **PROPERTY AREA**

Semiurban 233

Urban 202

Rural 179

Name: Property\_Area, dtype: int64

Semiurban 0.379479

Urban 0.328990

Rural 0.291531

Name: Property\_Area, dtype: float64



* **LOAN STATUS**

Y 422

N 192

Name: Loan\_Status, dtype: int64

Y 0.687296

N 0.312704

Name: Loan\_Status, dtype: float64

CONCLUSION

* Male are getting more loan approval.
* Married persons are getting more loan approval.
* Graduates are more prone to loan approval.
* No self employed person are getting more loan approval.
* Semiurban people are getting more loan approval whereas urban & rural people are almost equal.

Bivariate Analysis of Loan Approval Prediction Model

* **GENDER :**

**Loan\_Status N Y**

**Gender**

Female 37 75

Male 150 339

* **MARRIED :**

**Loan\_Status N Y**

**Married**

No 79 134

Yes 113 285

* **DEPENDENTS :**

**Loan\_Status N Y**

**Dependents**

0 107 238

1 36 66

2 25 76

3+ 18 33

* **EDUCATION :**

**Loan\_Status N Y**

**Education**

Graduate 140 340

Not Graduate 52 82

* **SELF-EMPLOYED :**

**Loan\_Status N Y**

**Self\_Employed**

No 157 343

Yes 26 56

* **PROPERTY AREA :**

**Loan\_Status N Y**

**Property\_Area**

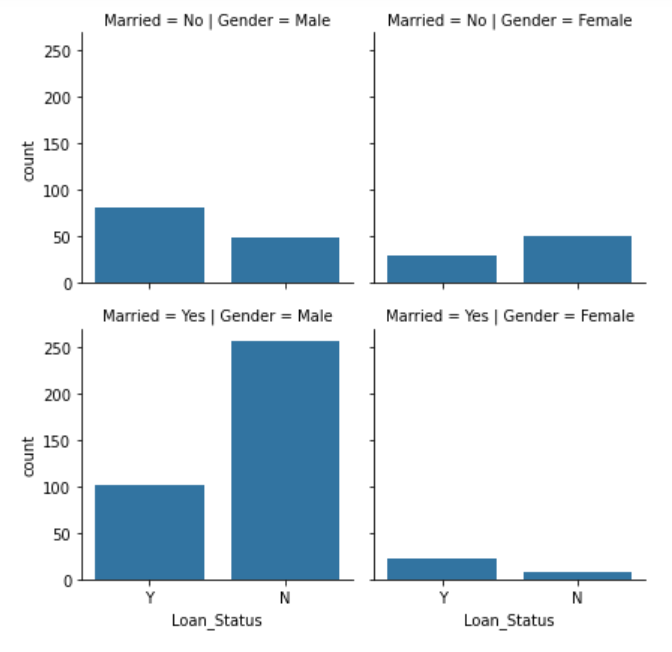
Rural 69 110

Semiurban 54 179

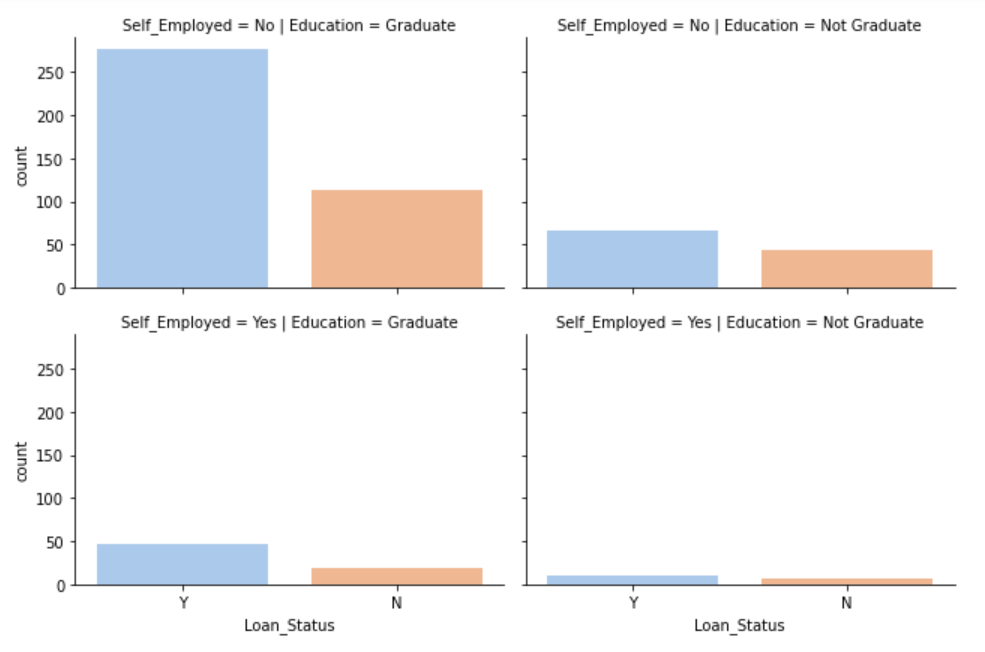
Urban 69 133

Visualization of Categorical Variables by Using Facetgrid

# GENDER,MARRIED Vs LOAN APPROVAL



# EDUCATION, SELF\_EMPLOYED Vs LOAN APPROVAL



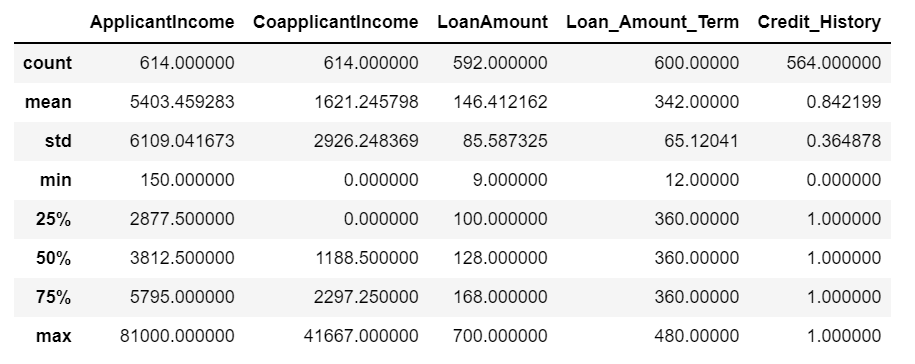
# PROPERTY AREA, GENDER Vs LOAN APPROVAL



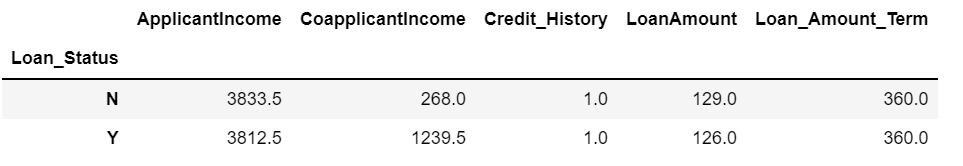
CONCLUSION: IMPACT OF CATEGORICAL PREDICTOR VARIABLES ON TARGET VARIABLE

* **LOAN APPROVAL STATUS**: About 2/3rd of applicants have been granted loan.
* **GENDER**: There are more Men than Women (approx. 3x)
* **MARTIAL STATUS**: 2/3rd of the population in the dataset is Marred; Married applicants are more likely to be granted loans.
* **DEPENDENTS**: Majority of the population have zero dependents and are also likely to accepted for loan.
* **EDUCATION**: About 5/6th of the population is Graduate and graduates have higher propotion of loan approval
* **SELF-EMPLOYMENT**: 5/6th of population is not self employed.
* **PROPERTY AREA**: More applicants from Semi-urban and also likely to be granted loans.
* **CREDIT HISTORY: Applicant** with credit history are far more likely to be accepted.
* **LOAN AMOUNT TERMS:** Majority of the loans taken are for 360 Months (30 years).

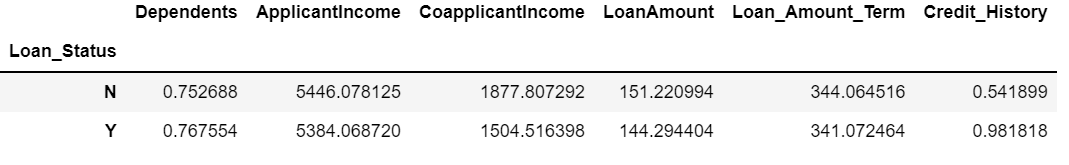
Overall Information about the Numerical variables



Median of Loan Approval

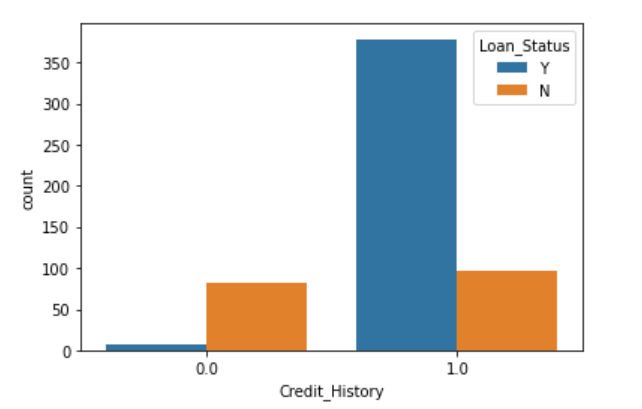


Mean of Loan Approval

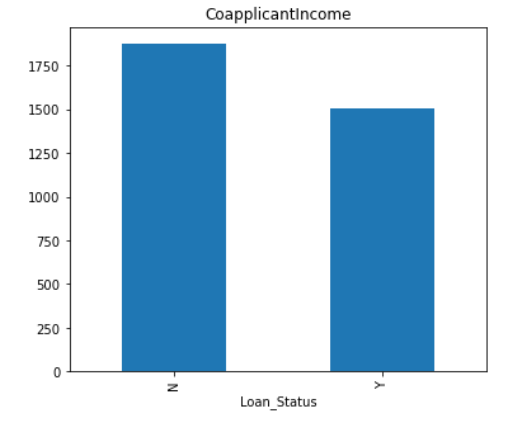


Visualization of Numerical Variables by Using Facetgrid

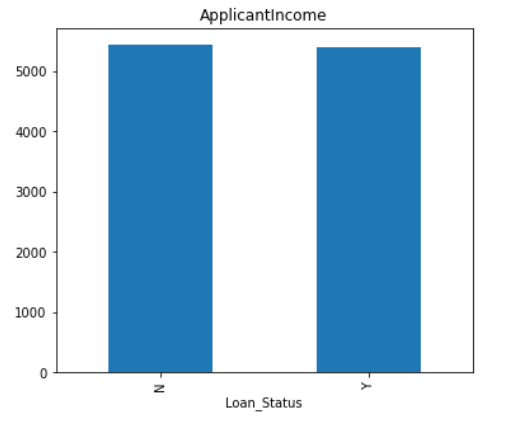
# Credit History Vs Loan Approval



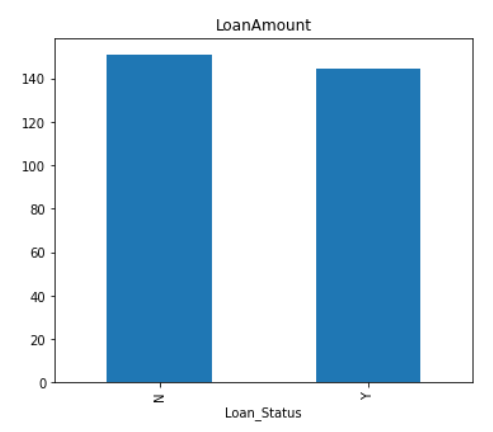
# Coapplicant Income Vs Loan Approval



# Applicant Income Vs Loan Approval

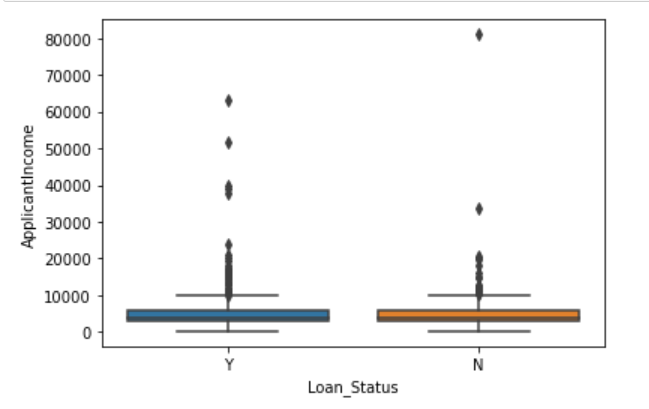


# Loan Amount Vs Loan Approval

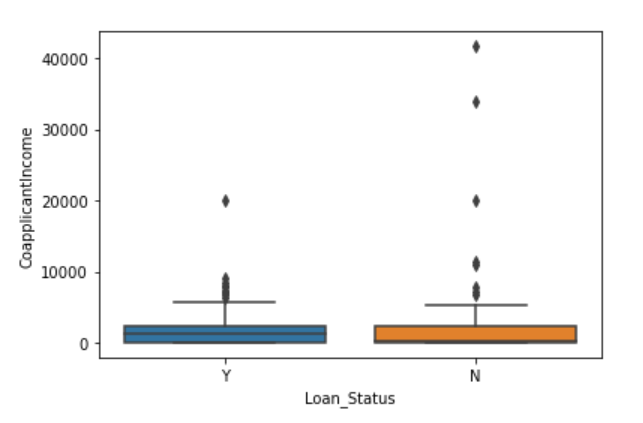


Outliers using Box Plot

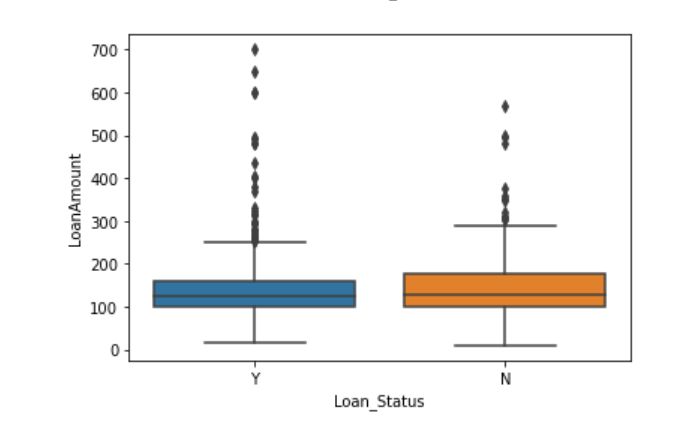
# Applicant Income:



# Co-Applicant Income:

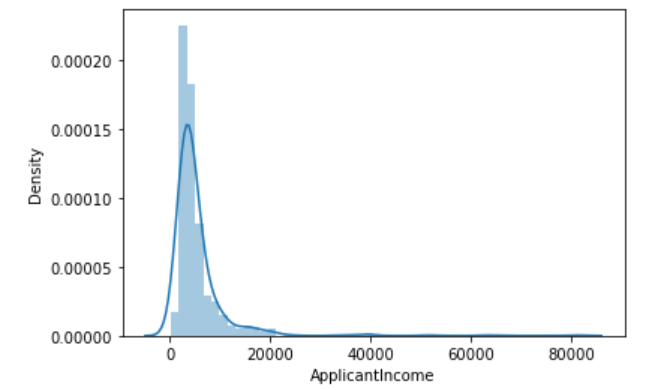


# Loan Amount Income:

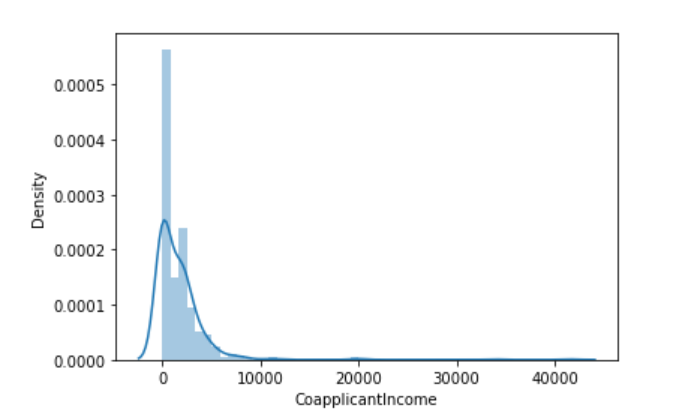


Normalise Behaviour of Continous Variables

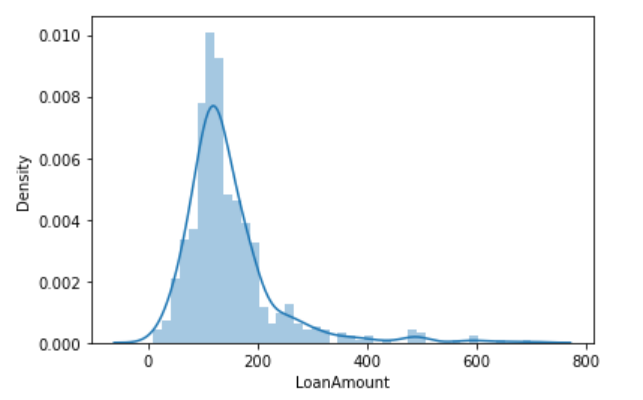
# Applicant Income:



# Co-Applicant Income:

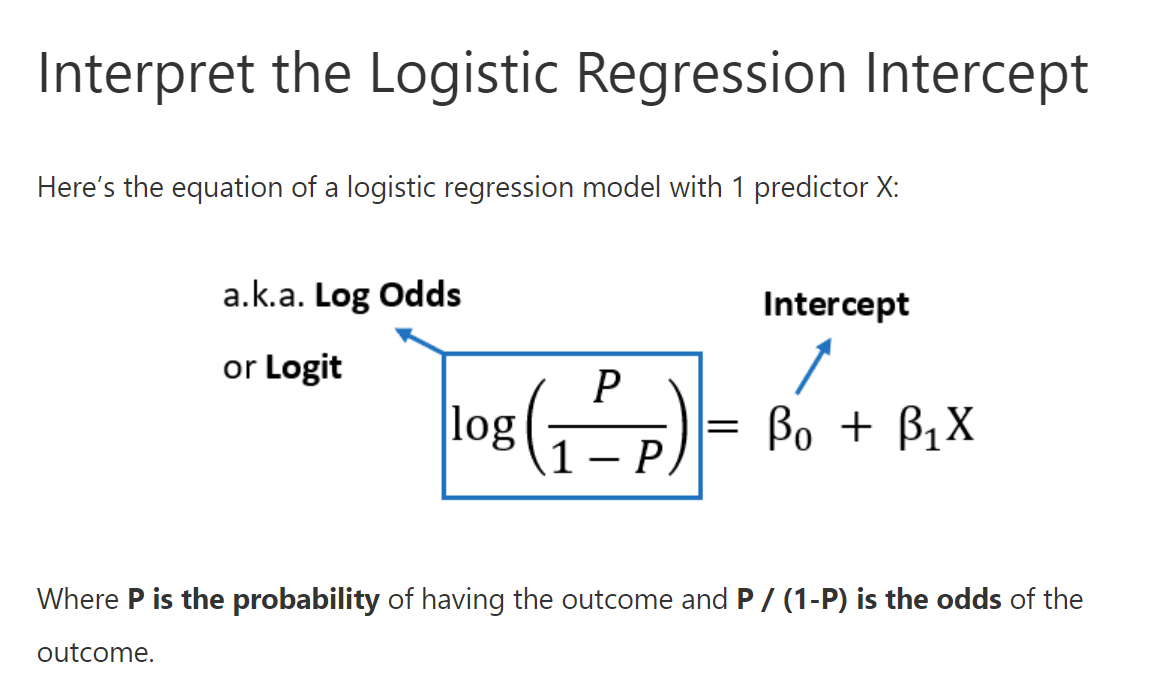


# Loan Amount Income:



STATISTICAL MODEL: LOGISTIC REGRESSION

Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function



**ABSTRACT:**

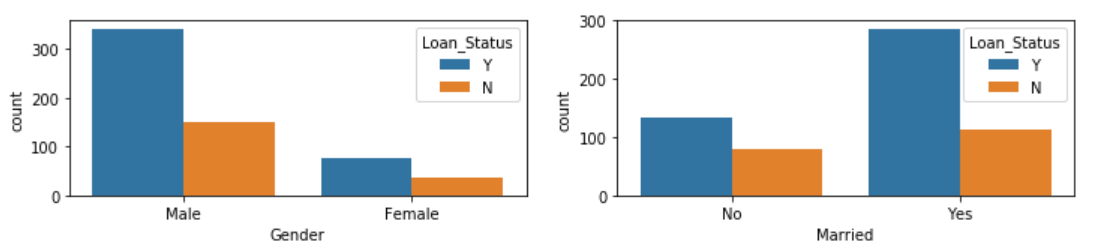
We wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. work on dealing in all home loan data, which contain various detail regarding the home loans. Also, with the logistic regression model, we will try to analyze which are the most important predictors for Loan Approval.

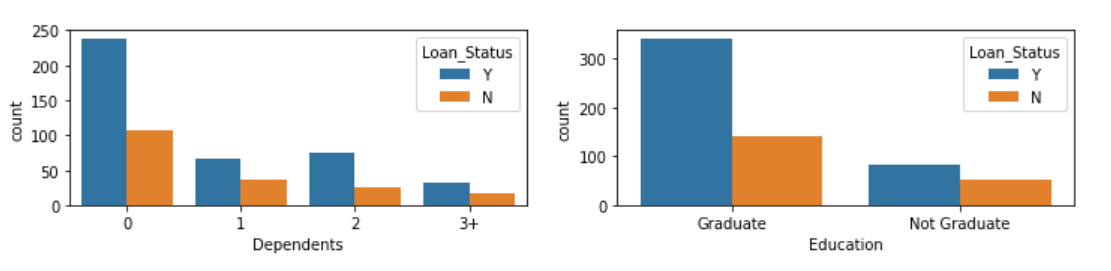
**Methodology:**

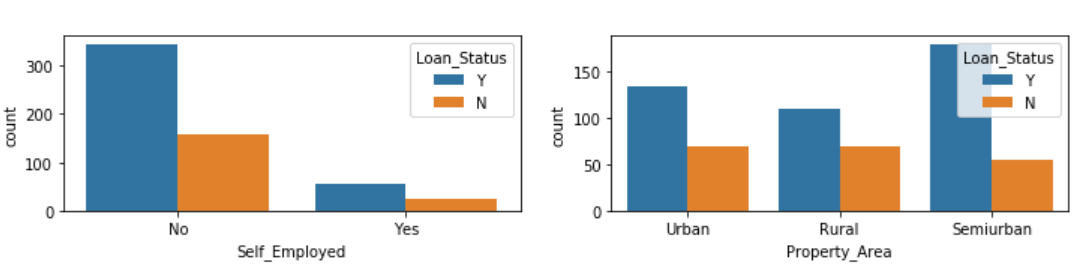
**Cleaning Data:**

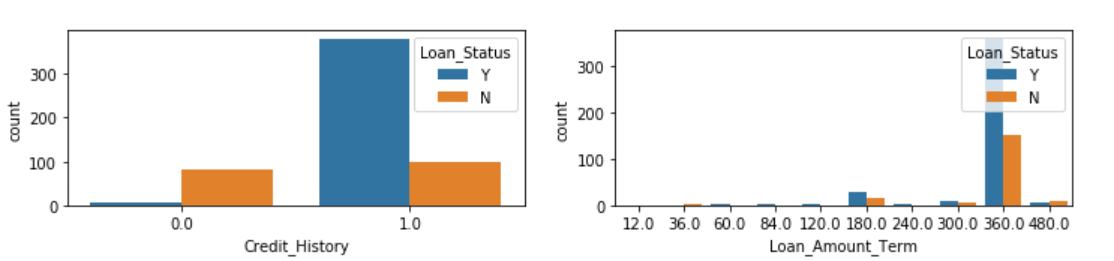
The data consist of numeric values. All categorical variables are encoded aptly. The data contain no duplicate values. There are also no None values found in the data. The target value also consists of 0 and 1 where 0 means not Approved for all home loan and 1 means Approved for all home loan.

**Exploratory data analysis:**

****

****

****

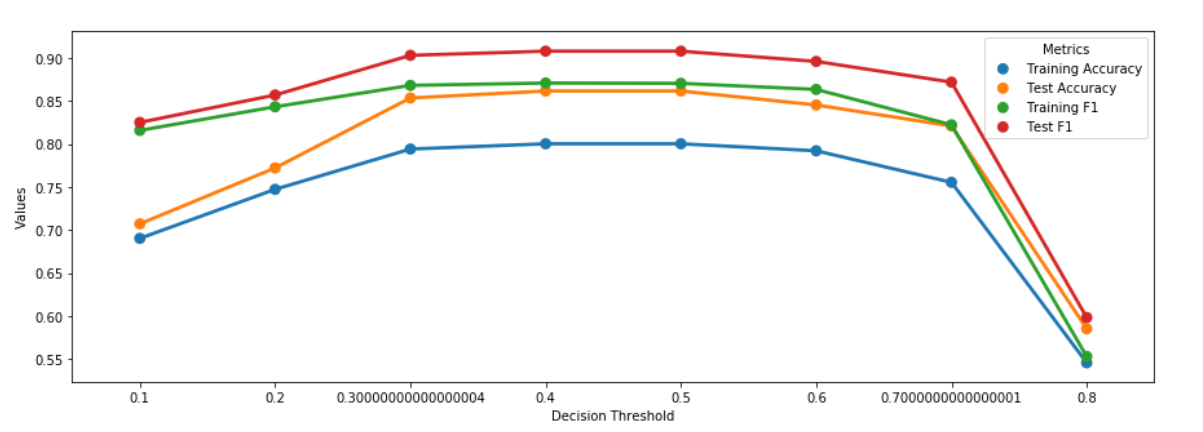
****

In the plot above we can see that there are more people in the dataset who are eligible for approving home loans.

**Model Building**

Since our target value is binary we are using Logistic regression which uses the Logit link function. Also, we don’t need much data preprocessing as all predictors are encoded. Here, we are using multiple logistic regression consisting of some categorical and numerical predictors. There are some assumptions about the model:

* The continuous predictors are linearly related to the log odds of the target variable.
* The observations should be independent.
* There should be no multicollinearity between the independent predictors.



Based on the above Test/Train curves, we can keep threshold to 0.4.  
Now Finally let's look at Logistic Regression Confusion Matrix

* Test Accuracy: 0.8617886178861789
* Test F1 Score: 0.9081081081081082
* Confusion Matrix on Test Data

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted | NO | YES | ALL |
| True |  |  |  |
| NO | 22 | 16 | 38 |
| YES | 1 | 84 | 85 |
| ALL | 23 | 100 | 123 |

CONCLUSION

In the above Logistic Regression Confusion matrix and above analysis, we did extensive analysis of input data and were able to achieve Test Accuracy of **86 %**

**THANKS YOU**