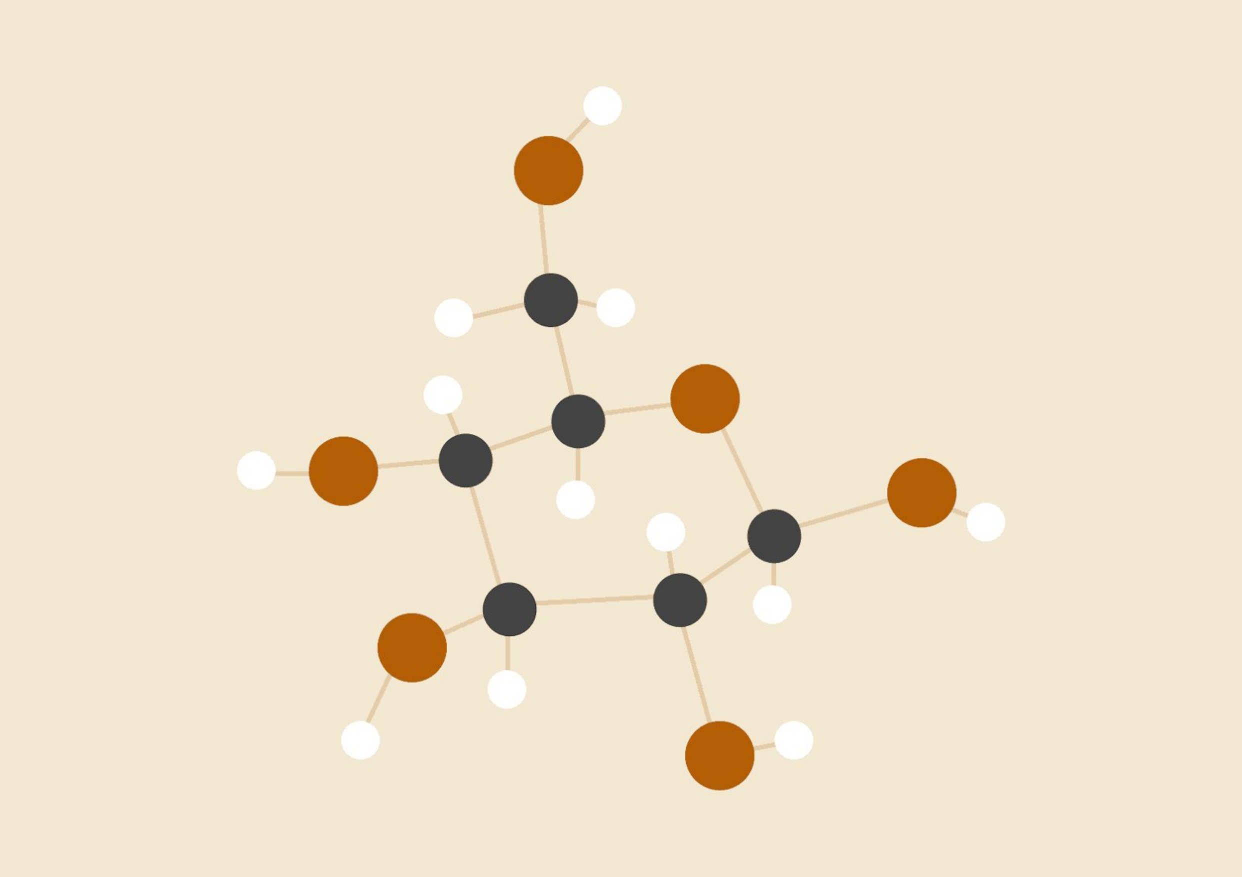
Exploratory Data Analysis REPORT

Bank Loan Prediction Dataset



Azlaan Mustafa Samad

Exploratory Data Analysis
Coursera

INTRODUCTION

This report is about Exploratory Data Analysis and I chose the dataset available for free on kaggle for Loan Prediction.

The name of the dataset is: ***train_ctrUa4K.csv***

EXPLORATORY DATA ANALYSIS

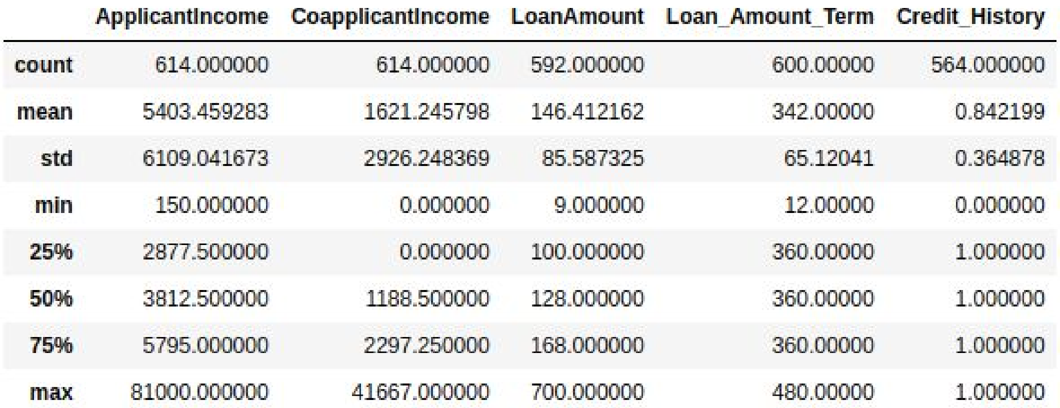
The different columns present in the dataset are:

1. **'Loan_ID'**: Unique ID for every loan request by a customer
2. **'Gender'**: Male, Female
3. **'Married'**: Yes or No
4. **'Dependents'**: Number of dependents: 0, 1, 2, 3+
5. **'Education'**: Graduate, Non-Graduate
6. **'Self_Employed'**: Yes or No
7. **'ApplicantIncome'**: Annual Income of the Applicant.
8. **'CoapplicantIncome'**: Annual Income of the Applicant partner if there is one.
9. **'LoanAmount'**: The amount of loan requested by the applicant.
10. **'Loan_Amount_Term'**: Term of loan in months/
11. **'Credit_History'**: 1 (The previous loan is paid) or 0 (The previous are unpaid).
12. ‘**Property_Area’:** 'Urban', 'Rural', 'Semiurban'
13. **‘Loan_Status**: This is the target variable (dependent variable) denoted with ‘Y’ (loan approved) or ‘N’ (loan rejected)

The following columns with the number of **missing values** are:

1.	Gender	13
2.	Married	3
3.	Dependents	15
4.	Self_Employed	32
5.	LoanAmount	22
6.	Loan_Amount_Term	14
7.	Credit_History	50

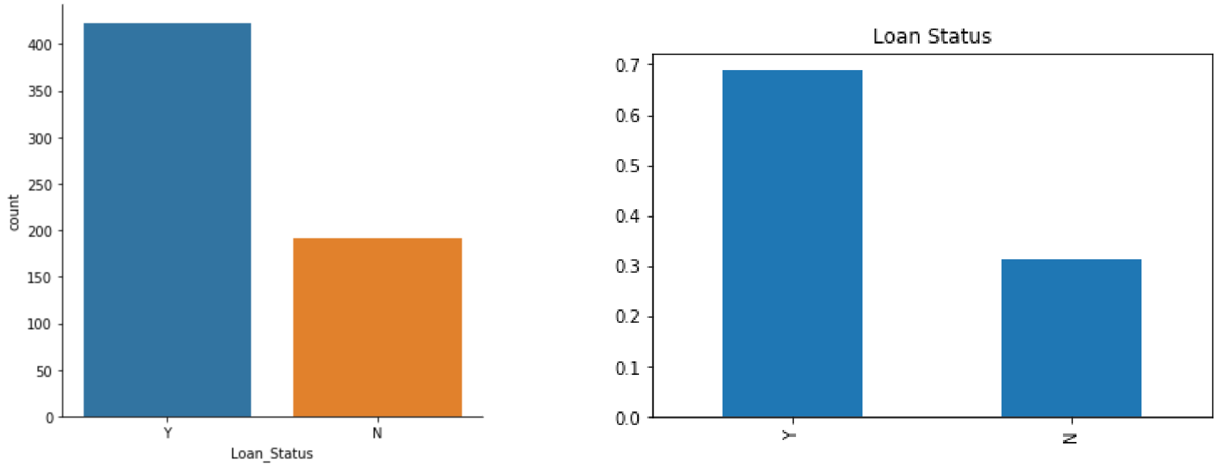
The following table describes the statistics of the numerical variables:



	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

Univariate Analysis:

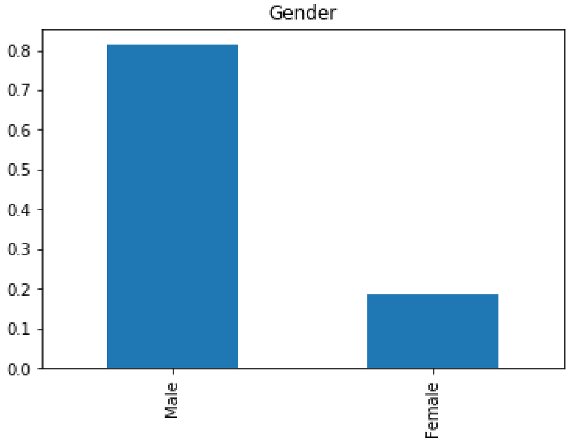
Loan_Status: Dependent/Target Variable

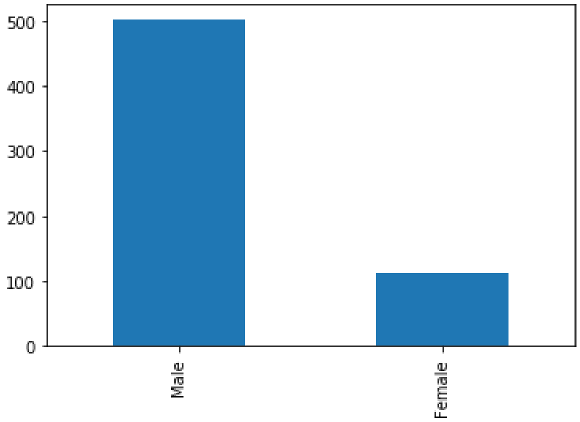


The above plot (L) shows the number of Loan approved and rejected. The probability of loan approval is 0.687296 while rejection is 0.312704 shown in the left plot.

INDEPENDENT VARIABLE (categorical):

Analysis of Gender variable:

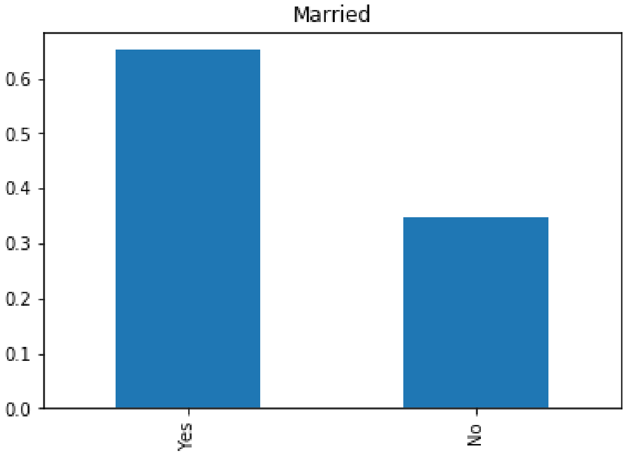




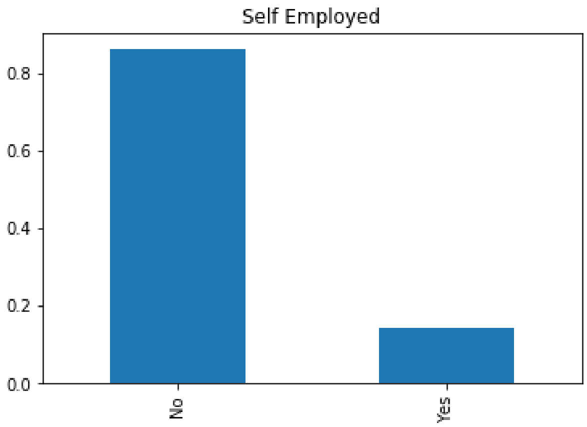
In our train dataset the "Gender" variable contains Male : 81% Female: 19%. The right plot shows the actual count of Male and Females while the left plot shows the normalised count which is for Male 0.81759 and for females 0.18241.

Analysis of Married variable:

The number of married and unmarried applicants are 398 and 213 respectively. The plot below shows the normalised count in the entire dataset which for married and unmarried applicants are 0.653094 and 0.346906 respectively.



Analysis of Self_Employed Variable:

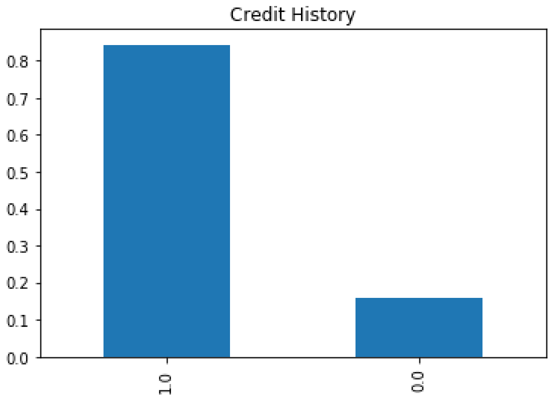


The above plot shows the normalised count

# of Self_Employed : 82
# of Not_Self_Employed : 500

In the dataset only 14% are Self Employed and the rest of the 86% are not Self Employed.

Analysis of Credit_History variable :



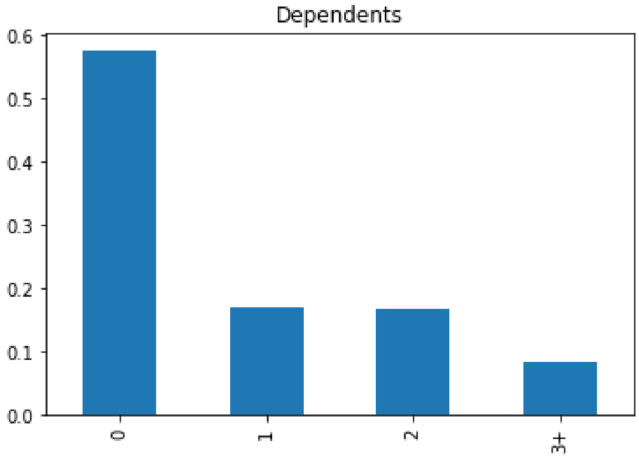
The above plot shows normalised count. Around 84% applicants have repaid their debts.

# of people with credit history equal to 1 (paid debts): 475
# of people without any credit history (unpaid debts): 89

INDEPENDENT VARIABLE (ORDINAL):

Ordinal features: Variables in categorical features having some order involved (Dependents, Education, Property_Area)

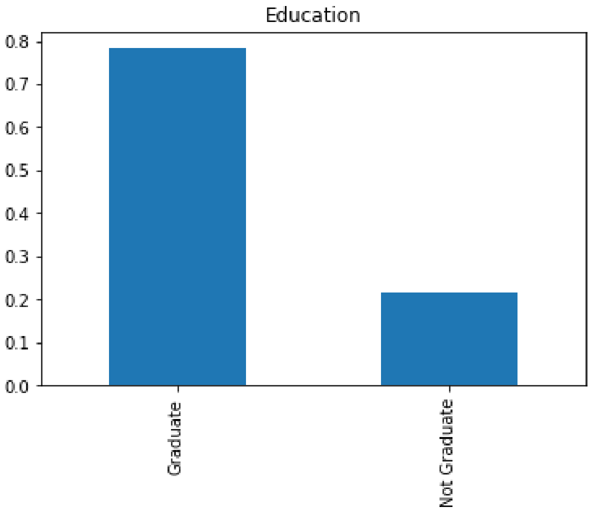
Analysis of Dependents variable:



The above plot shows the normalised count for different number of dependents.

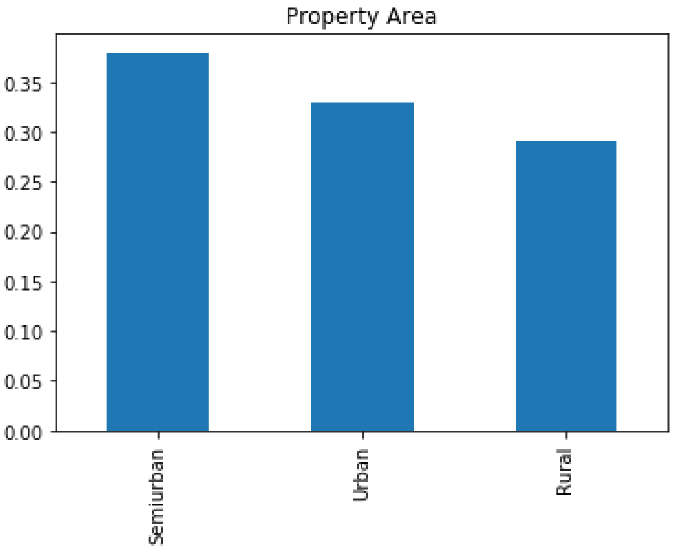
Number of 0 Dependent : 345 or 58%
Number of 1 Dependent : 102 or 17%
Number of 2 Dependent : 101 or 17%
Number of 3+ Dependent : 51 or 8%

Analysis of Education variable:



# of Graduates: 480 (78%)
# of Non-Graduate: 134 (21%)

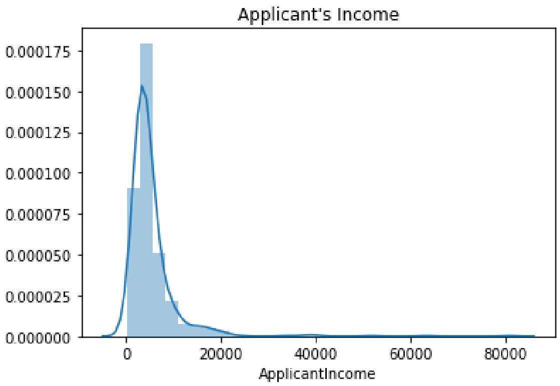
Analysis on Property_Area variable :

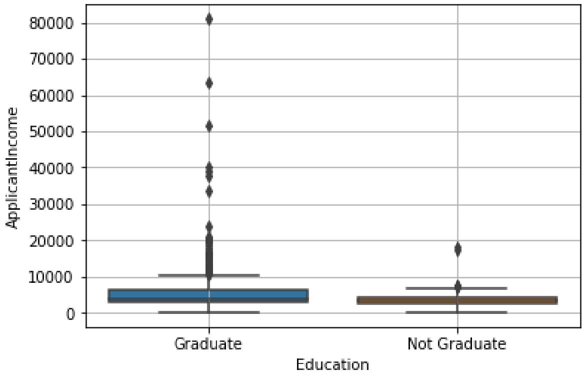


# of Semiurban: 233 (37.94%)
# of Urban: 202 (32.89%)
# of Rural: 179 (29.15%)

INDEPENDENT VARIABLE (NUMERICAL):

Analysis of Applicant's Income:

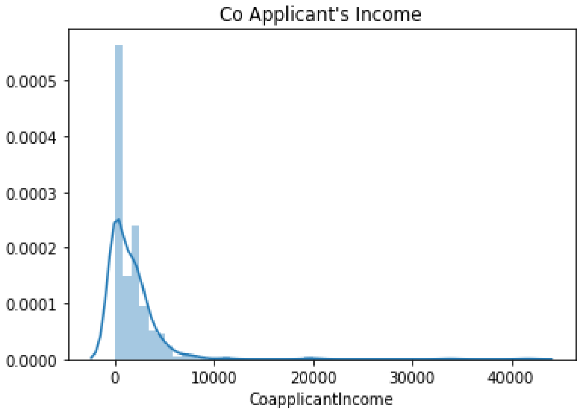


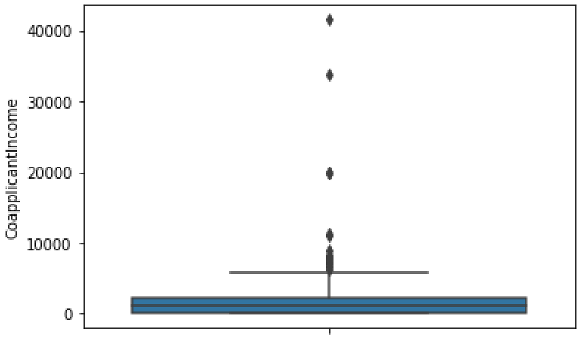


The distribution is not normally distributed. It is skewed to the left. The boxplot confirms the presence of a lot of outliers/extreme values. This can be attributed to the income disparity in the society.

We can see that there are a higher number of graduates with very high incomes, which are appearing to be the outliers.

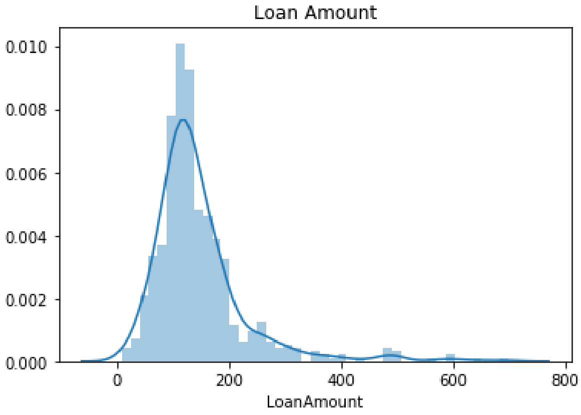
Analysis of Coapplicant's Income:

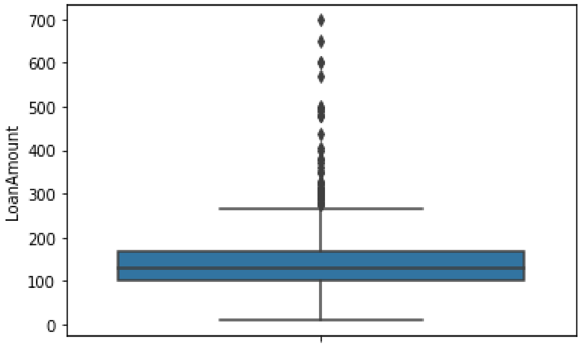




We see a similar distribution as that of the applicant income. Majority of the co-applicant's income ranges from 0 to 500. We also see a lot of outliers in the co-applicant's income and it is not normally distributed.

Analysis of Loan Amount:





We see a lot of outliers in this variable and the distribution is fairly normal.

Hypothesis:

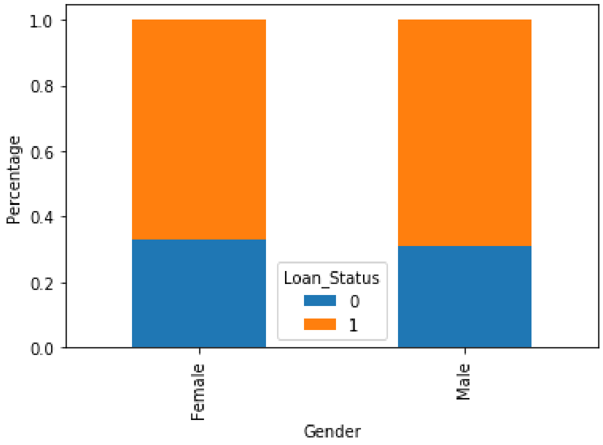
Based on the analysis of variables individually the following hypothesis is generated:

1. Applicants with high income should have more chances of loan approval.

2. Applicants who have repaid their previous debts should have higher chances of loan approval.
3. Loan approval should also depend on the loan amount. If the loan amount is less, chances of loan approval should be high.
4. Lesser the amount to be paid monthly to repay the loan, higher the chances of loan approval.

Bivariate Analysis:

Analysis of Loan Status and Gender:



Number of Female whose Loan was approved : 75

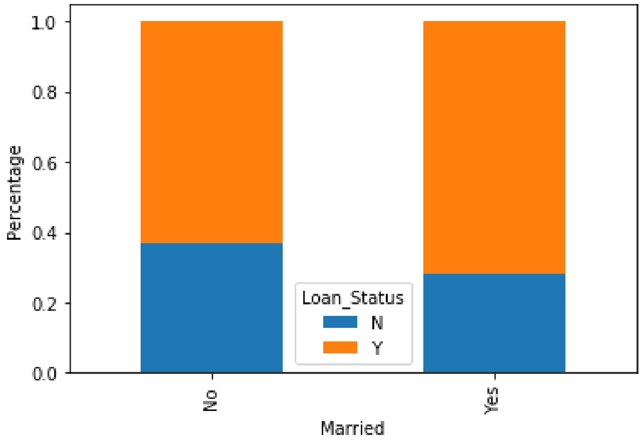
Number of Male whose Loan was approved : 339

Number of Female whose Loan was not approved : 37

Number of Male whose Loan was not approved : 150

Proportion of Male applicants is higher for the approved loans. The probability of loan denial for both Male and Females is the same as the total probability of Loan denial.And it is similar for the approval case. Thus we can for now say that Gender doesn't really matter in case of Loan Approval. We have to conduct statistical tests to confirm this.

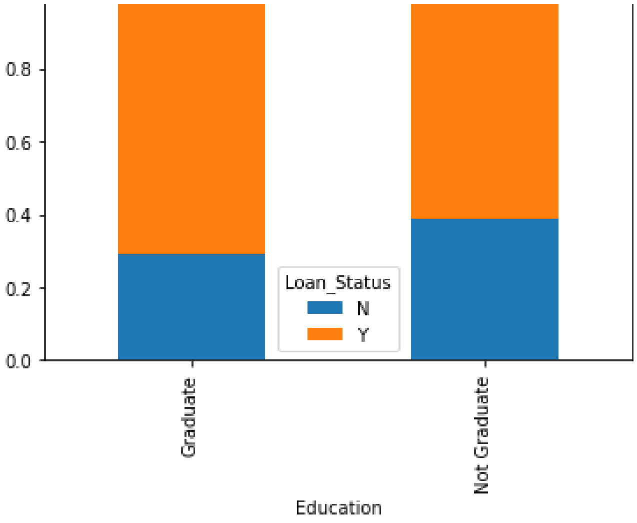
Analysis of Loan Status and Married:



Number of married people whose Loan was approved : 285
Number of married people whose Loan was not approved : 113
Number of unmarried people whose Loan was approved : 134
Number of unmarried people whose Loan was not approved : 790

Proportion of Married applicants is higher for the approved loans.

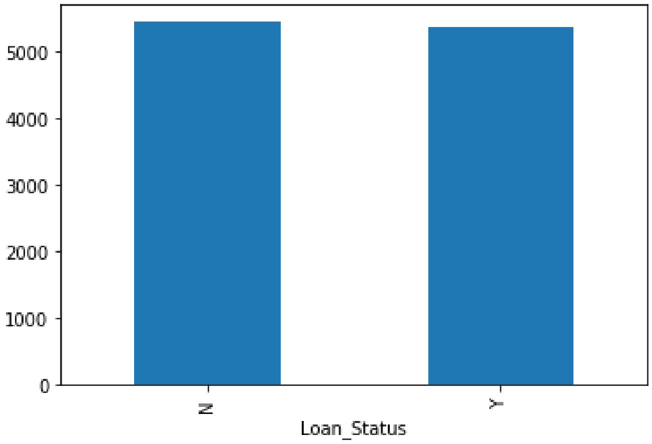
Analysis of Loan Status and Education:



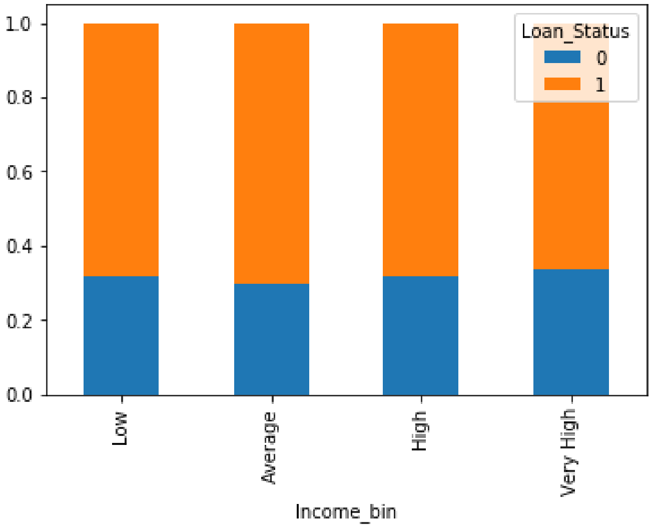
Number of people who are Graduate and Loan was approved : 340
Number of people who are Graduate and Loan was no approved : 140
Number of people who are Not Graduate and Loan was approved : 82

Number of people who are Not Graduate and Loan was not approved : 52
We can clearly see the Loan denial in case of Non-Graduates is higher than Graduates.
We have to conduct statistical tests to confirm this.

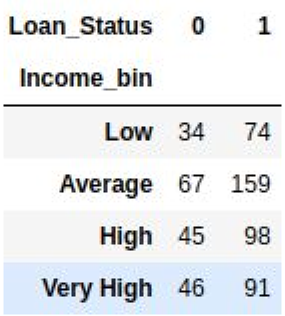
Analysis of Loan Status and Applicant's Income:



Here the y-axis represents the mean of the applicant's income. We don't see any change in the mean income. So, let's make bins for the applicant's income variable based on the values in it and analyze the corresponding loan status for each bin.



The following table displays the Loan status corresponding to the income range:



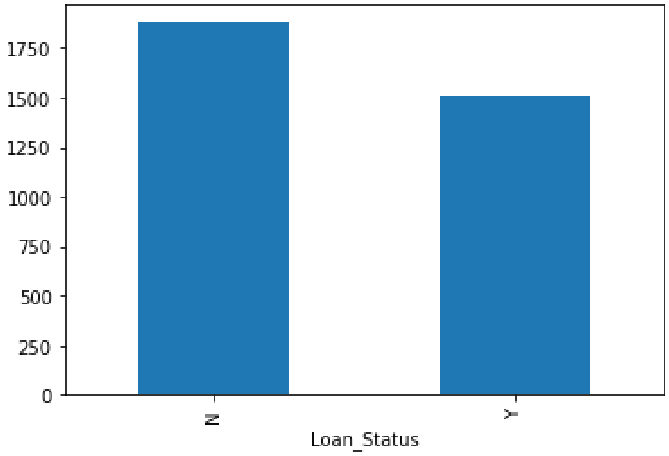
Loan_Status Income_bin	0	1
Low	34	74
Average	67	159
High	45	98
Very High	46	91

The bins created are:

bins = 0-2500 (low),
2500-4000 (Average),
4000-6000 (high),
6000-81000 (very high)

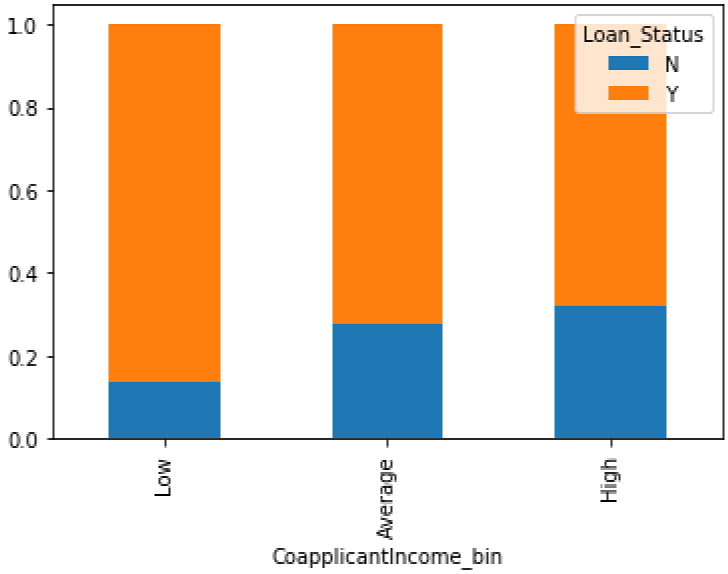
It can be inferred that Applicant's income does not affect the chances of loan approval which contradicts our hypothesis in which we assumed that if the applicant's income is high the chances of loan approval will also be high.

Analysis of Loan Status and Coapplicant's Income:

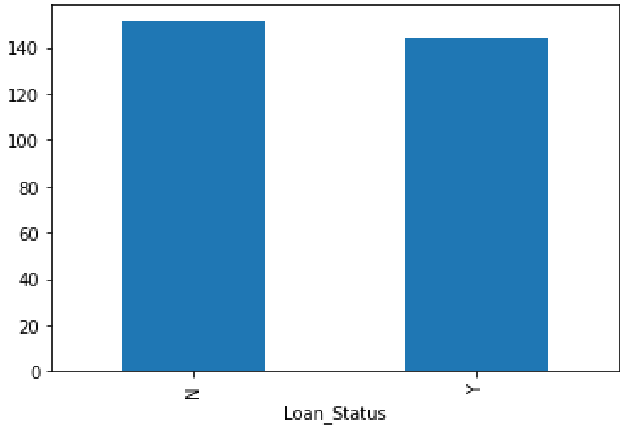


The mean co-applicant's income of Loan not approved is higher than the approved ones. The co-applicant's income is further analysed with respect to different income range and Loan status. The following income range was created and analysed: 0-1000 (low), 1000-3000(average), 3000-4200(high).

It shows that if the co applicant's income is less the chances of loan approval are high. But this does not look right. The possible reason behind this may be that most of the applicants don't have any co applicant so the co applicant income for such applicants is 0 and hence the loan approval is not dependent on it. So we can make a new variable in which we will combine the applicant's and co applicant's income to visualize the combined effect of income on loan approval. We will analyse the total income in the feature engineering section of the report.

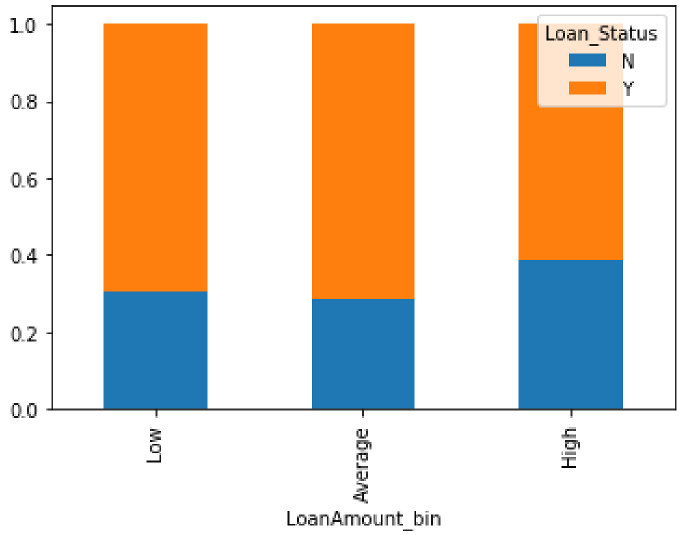


Analysis of the Loan Status and Loan Amount:



It seems the average Loan Amount is the same for approved and non approved Loans. Let's do further analysis by dividing into bins and seeing how the Loan approved changes over different groups.

The following loan amount range was chosen and analysed:
Low: 0-100
Average: 100-200
High: 200-700



It can be seen that the proportion of approved loans is higher for Low and Average Loan Amount as compared to that of High Loan Amount which supports our hypothesis in which we considered that the chances of loan approval will be high when the loan amount is less.

Feature Engineering and Variable Transformation

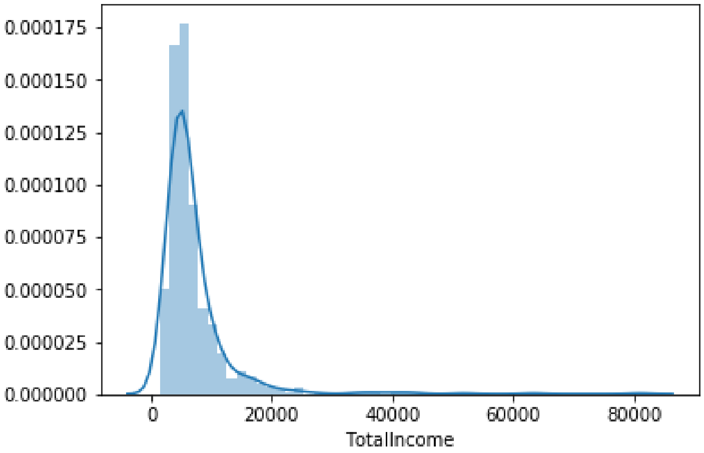
Based on the domain knowledge, we can come up with new features that might affect the target variable. We will create the following three new features:

Total Income - As discussed during bivariate analysis we will combine the Applicant Income and Co-applicant Income. If the total income is high, chances of loan approval might also be high.

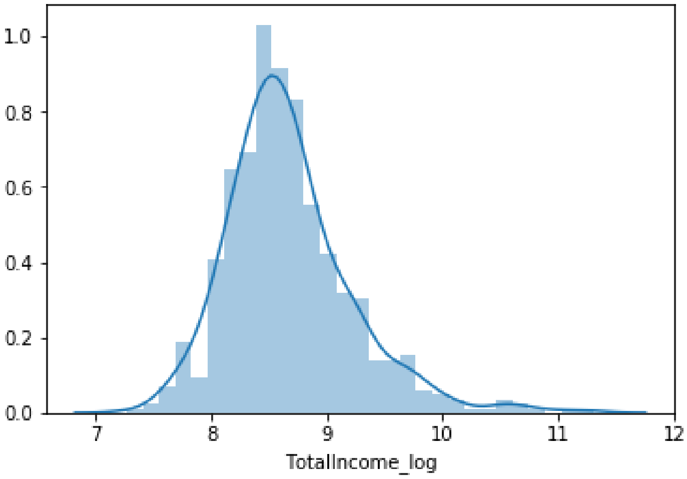
EMI - EMI is the monthly amount to be paid by the applicant to repay the loan. Idea behind making this variable is that people who have high EMI's might find it difficult to pay back the loan. We can calculate the EMI by taking the ratio of loan amount with respect to loan amount term.

Balance Income - This is the income left after the EMI has been paid. Idea behind creating this variable is that if this value is high, the chances are high that a person will repay the loan and hence increasing the chances of loan approval.

Analysis of Total Income:

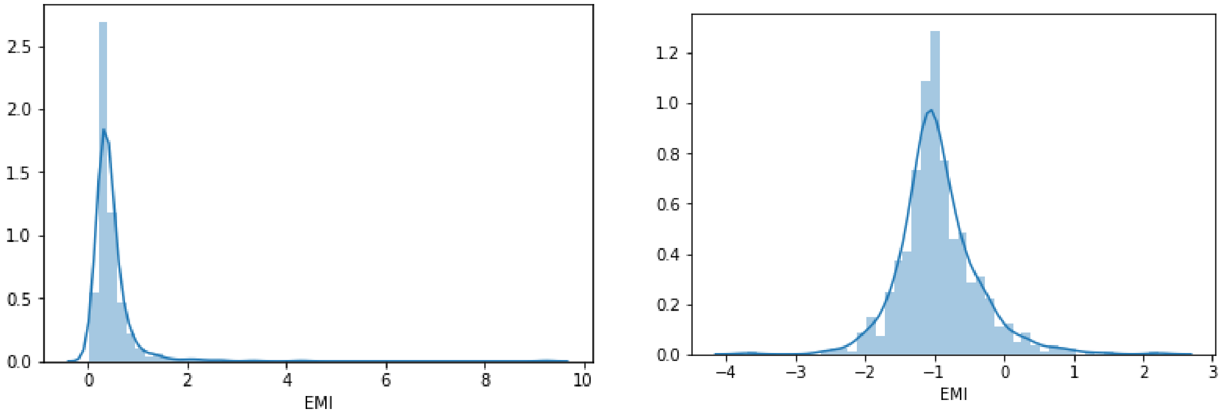


We can see it is shifted towards left, i.e., the distribution is right skewed. So, let’s take the log transformation to make the distribution normal.



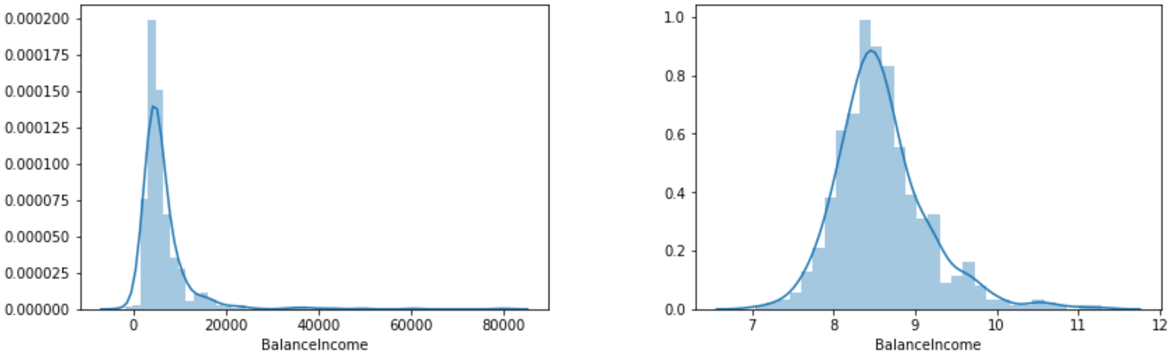
Now the distribution looks much closer to normal and the effect of extreme values has been significantly subsided.

Analysis of EMI:



The left and right are the EMI distribution without and with log transformation.

Analysis of Balanced Income:



The left and right are the Balanced Income distribution without and with log transformation.

Hypothesis Testing

The following are the Null and Alternative hypothesis assumed for the variables being analysed:

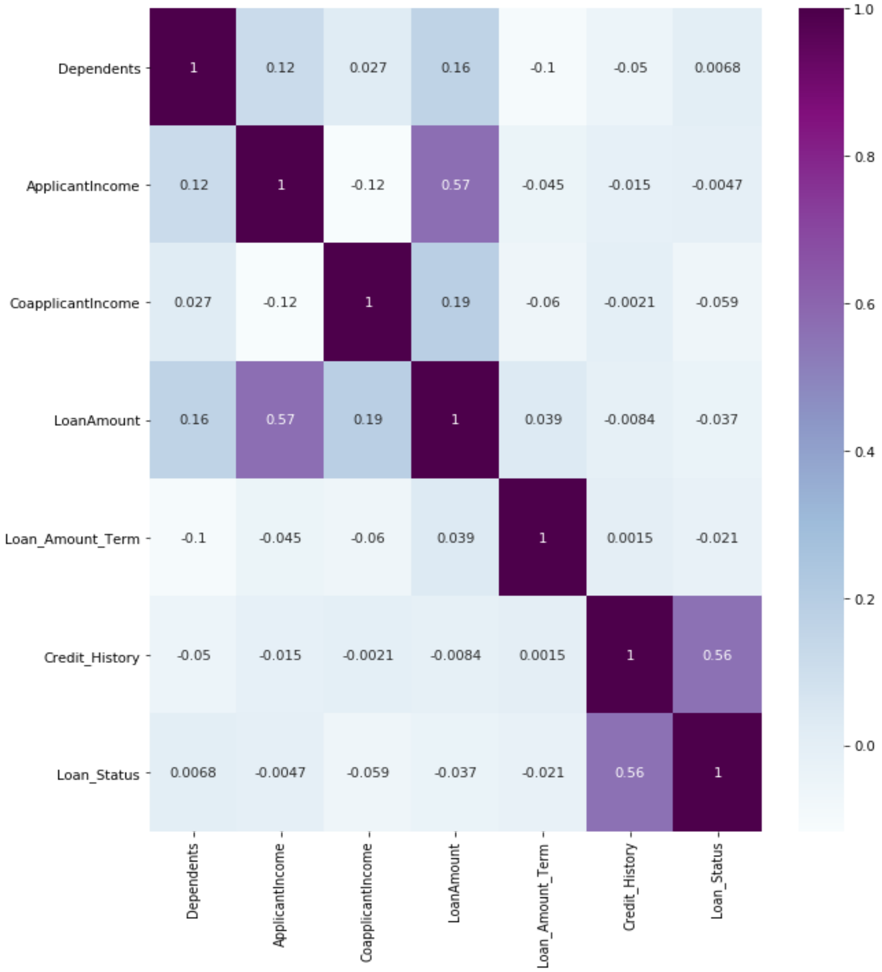
Null hypothesis: every independent variable category has equal chances of getting a loan [m1 == m2]

Alternate hypothesis : not equal chances [m1 != m2]

For all the cases of hypothesis testing the significance level or alpha is chosen as 0.05.

Variable	Chi-squared test statistic value	p-value	Outcome
Property_Area	12.29	0.00213	Reject Null
Gender	0.1108	0.739	Not Reject
Education	4.091	0.043	Reject Null
Married	4.731	0.029	Reject Null
Credit_History	176.114	3.4	Reject Null

Correlation between variables



We see that the most correlated variables are (ApplicantIncome - LoanAmount) and (Credit_History - Loan_Status).

CONCLUSION

The different variables were analysed with respect to the Loan Status. New features were extracted from the already existing ones. Existing features were scaled and transformed.