Instant Home Loan Approval with Machine Learning.

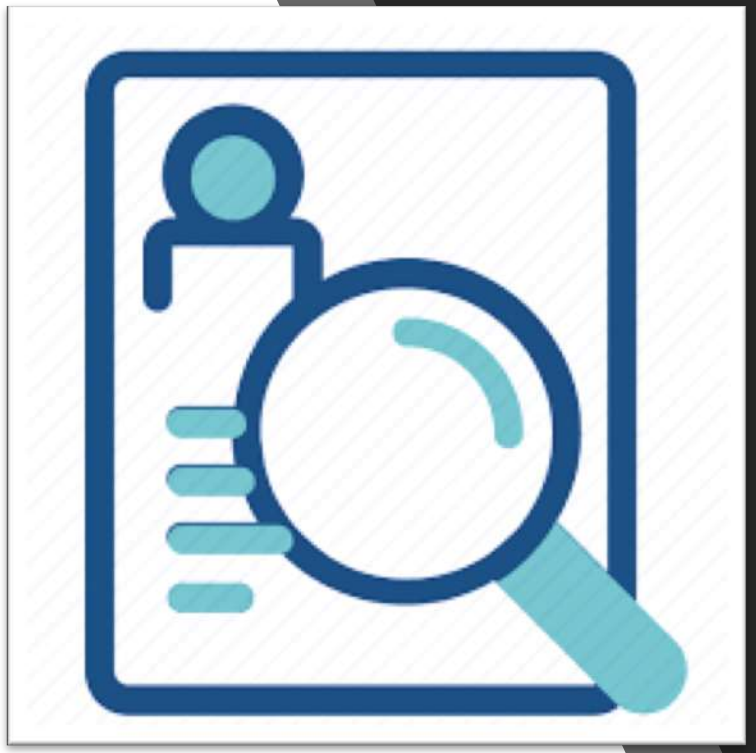
Presented by,

Harsh Raizada

Don't forget to take away !

- 68.72% of the total applicants have loan approved so, Be Happy!
- More down payment and less loan amount will increase chances of getting loan approval.
- Married people has higher chance of getting loan approved.
- Try to buy property in Semi-Urban area.
- Good credit history is definitely be a plus and is highly correlated (+) with loan approval.
- Logistic Regression and SVM model has the best accuracy for automated loan approval.

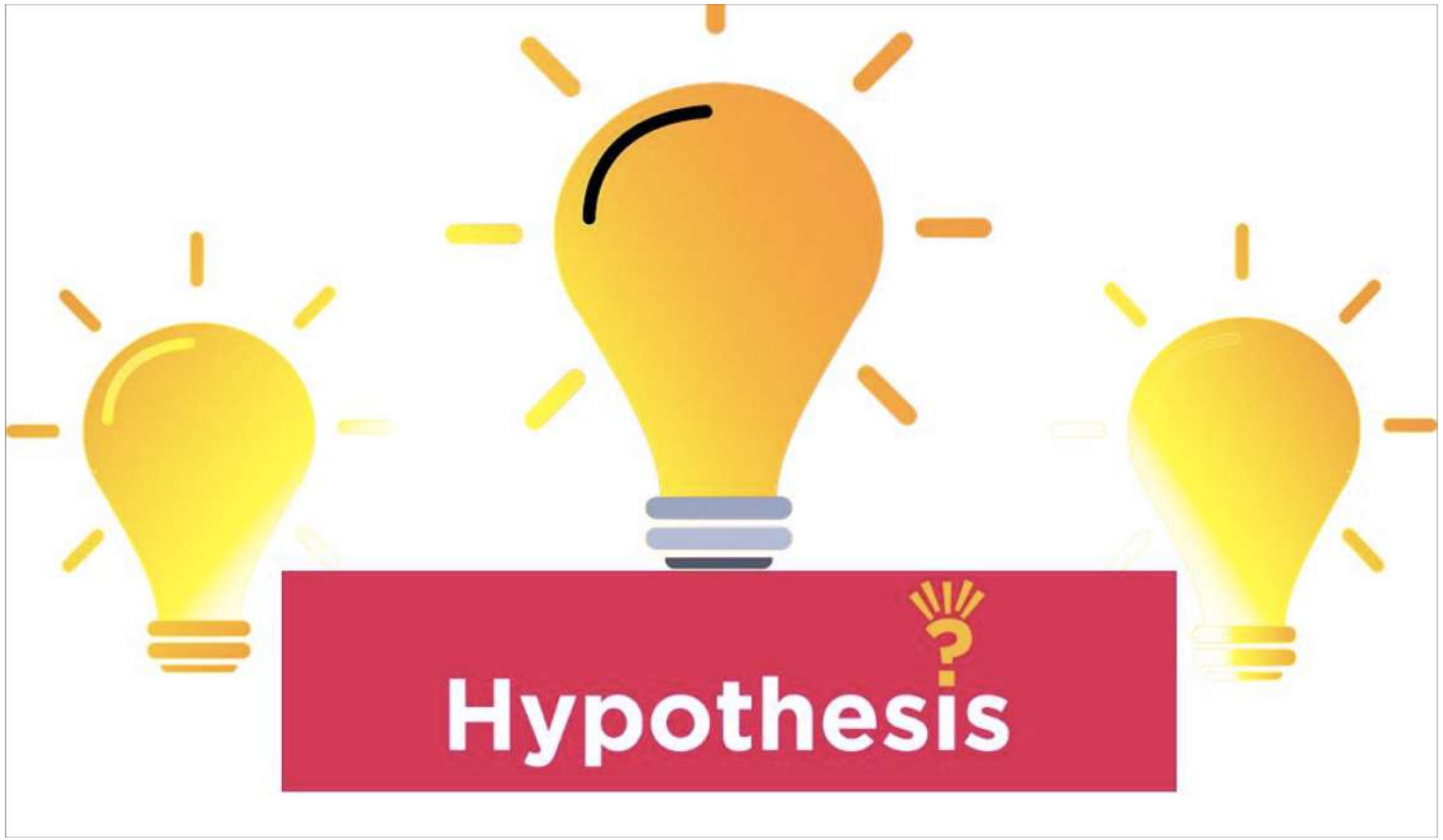
Overview



- ✓ A home loan dataset was given to make a Home Loan Approval prediction.
- ✓ Currently approval process is manual, based on different parameters.
- ✓ These parameters are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others.
- ✓ Now, the company wants to automate the loan eligibility process in real-time based on the customer detail provided while filling online application form.
- ✓ It is a Binary Classification problem where we have to predict whether a loan would be approved or not.
- ✓ In this project we will identify the best algorithm by comparing the results of different classification algorithms for Home Loan approval prediction where based on the parameters applicant will get there loan approved/disapproved and accordingly company will focus only on the approved customers.

The Problem Statement:

A company wants to know and automate the process that which customers or applicant should be given the loans for buying the Home.



The Hypothesis Generation

- ✓ Lower the loan amount will increase the chances of getting loan approved.
- ✓ People who has business has less chances to get their home loan approved.
- ✓ Those applicant who has higher income will have higher chances of getting their loan approved.
- ✓ Females applicants have better chance of getting loan approved.
- ✓ Married applicant will have the lower chances of getting loan approved.
- ✓ Applicants who are buying property in urban areas will easily get loan approved.
- ✓ Co-applicant with higher income will increase the chances of getting loan approved for main applicant.
- ✓ Loan for less time period should have higher chances of approval.
- ✓ Applicants who have repayed their previous debts should have higher chances of loan approval.

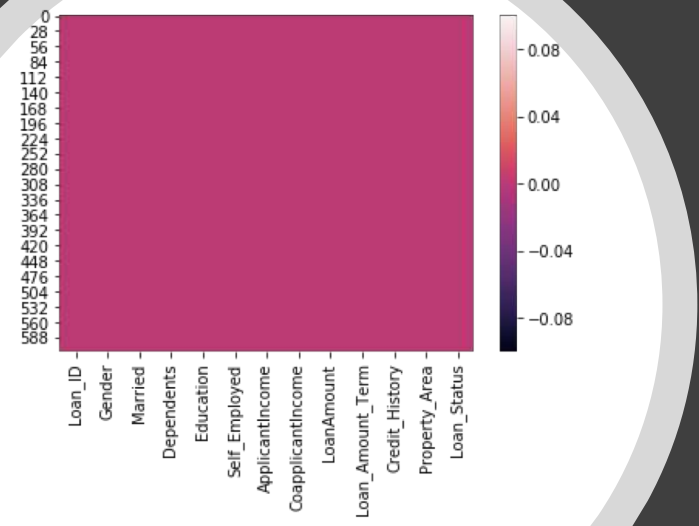
Process & Summary Stats

- All the important packages for analysis, Graphs, Preprocessing and Model building were installed for working on Jupyter notebook.
- We have given two files (Train & Test).
- Training Data set have 614 observation and 13 features (including target variable)
- Test dataset have 367 observation with 12 variables (excluding target variables).

Description of Data

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Loan_ID	614	614	LP001691	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	601	2	Male	489	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Married	611	2	Yes	398	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Dependents	599	4	0	345	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Education	614	2	Graduate	480	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Self_Employed	582	2	No	500	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ApplicantIncome	614	NaN	NaN	NaN	5403.46	6109.04	150	2877.5	3812.5	5795	81000
CoapplicantIncome	614	NaN	NaN	NaN	1621.25	2926.25	0	0	1188.5	2297.25	41667
LoanAmount	592	NaN	NaN	NaN	146.412	85.5873	9	100	128	168	700
Loan_Amount_Term	600	NaN	NaN	NaN	342	65.1204	12	360	360	360	480
Credit_History	564	NaN	NaN	NaN	0.842199	0.364878	0	1	1	1	1
Property_Area	614	3	Semiurban	233	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Loan_Status	614	2	Y	422	NaN	NaN	NaN	NaN	NaN	NaN	NaN

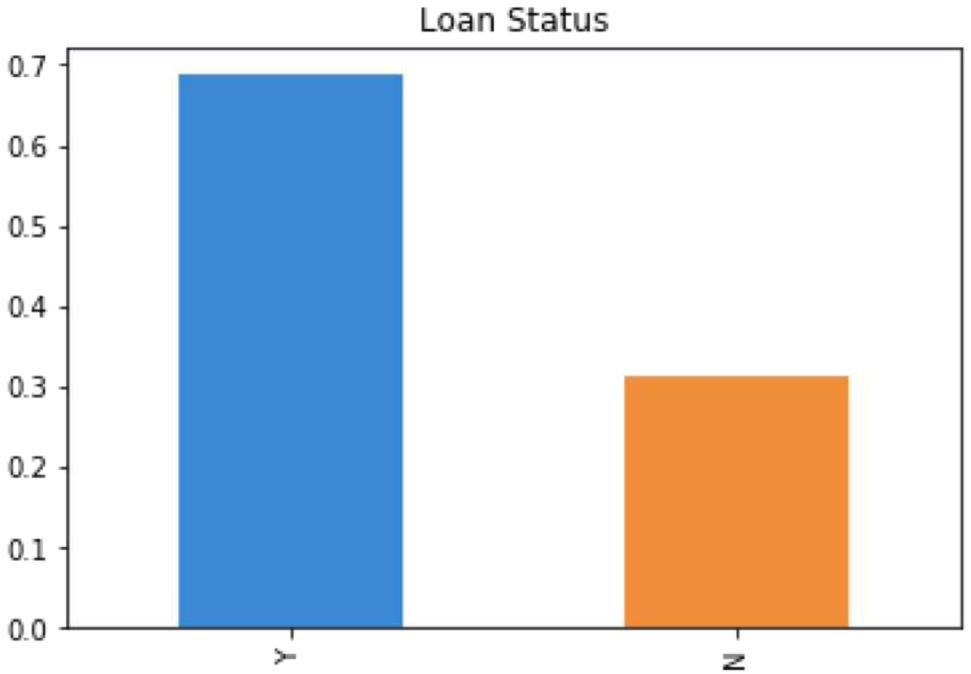
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
Loan_ID              614 non-null object
Gender               614 non-null int64
Married              614 non-null int64
Dependents           614 non-null int64
Education            614 non-null int64
Self_Employed        614 non-null int64
ApplicantIncome      614 non-null int64
CoapplicantIncome    614 non-null float64
LoanAmount           614 non-null float64
Loan_Amount_Term     614 non-null float64
Credit_History       614 non-null float64
Property_Area        614 non-null int64
Loan_Status          614 non-null int64
dtypes: float64(4), int64(8), object(1)
memory usage: 62.4+ KB
```



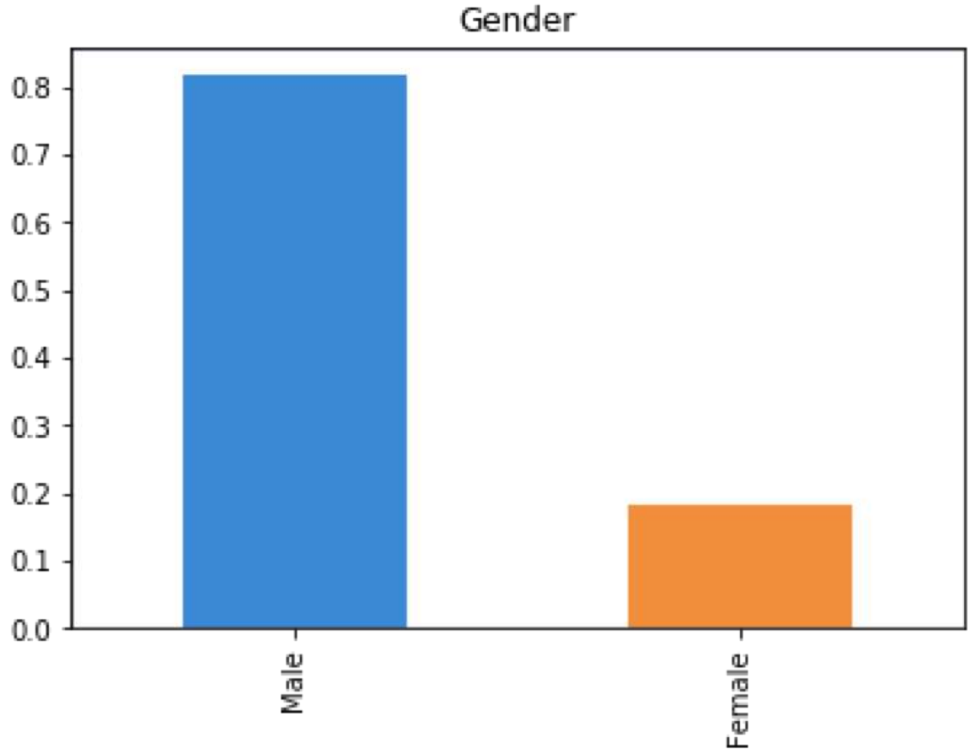
Data Issues & solutions

- Training data were having 149 missing values, while
- Test data have 84 missing values.
- Variables Gender, Married, Dependents, Self Employed, Credit History, Loan Amount Term missing values were imputed with mode and Loan amount missing values with median.
- Some variables were categorical, those were converted into numerical.

Exploring Data: Univariate analysis

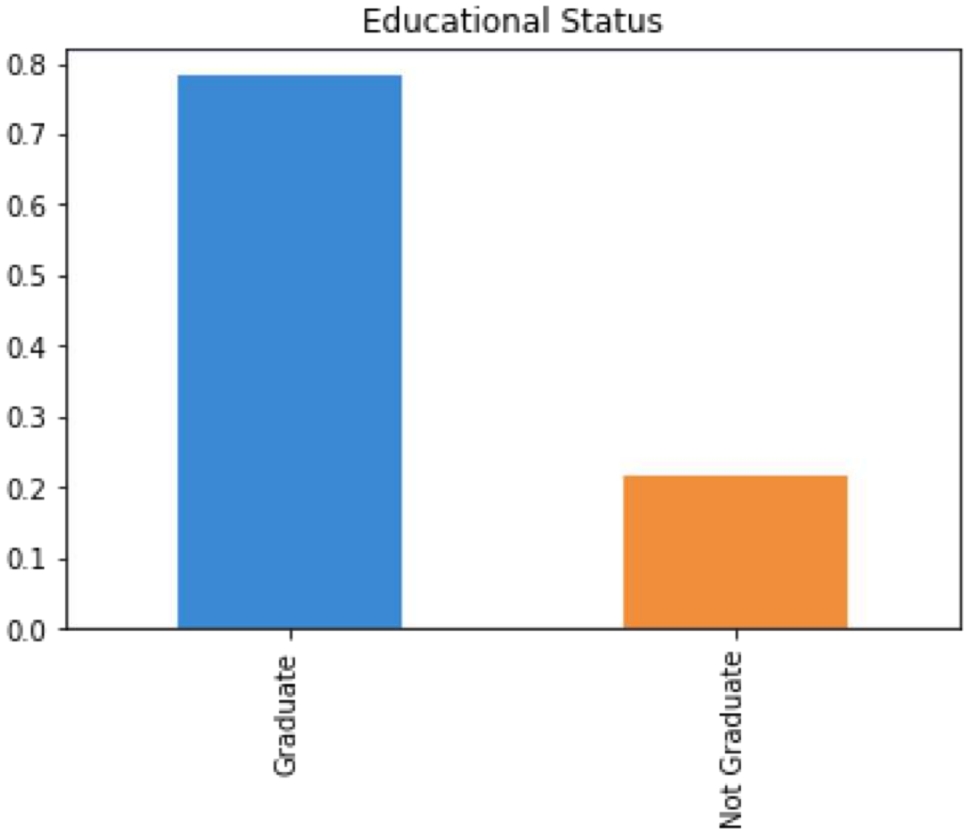


68.72% of the applicants have loan approved i.e most of the applicants are getting there loan approved

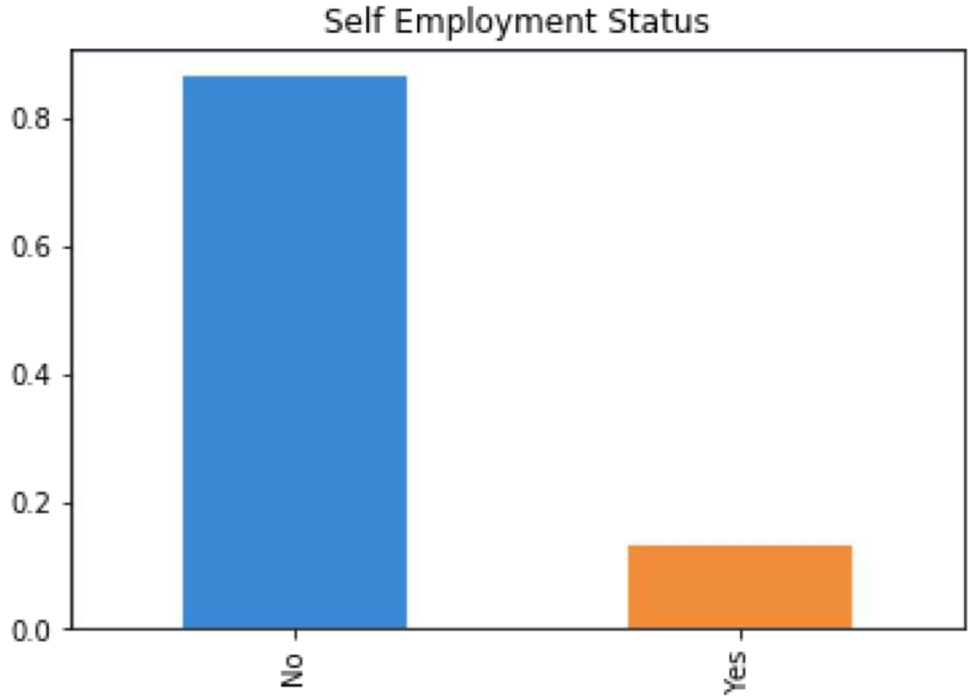


81.75% of applicant are male and 18.24% of applicants are female

Exploring Data: Univariate analysis

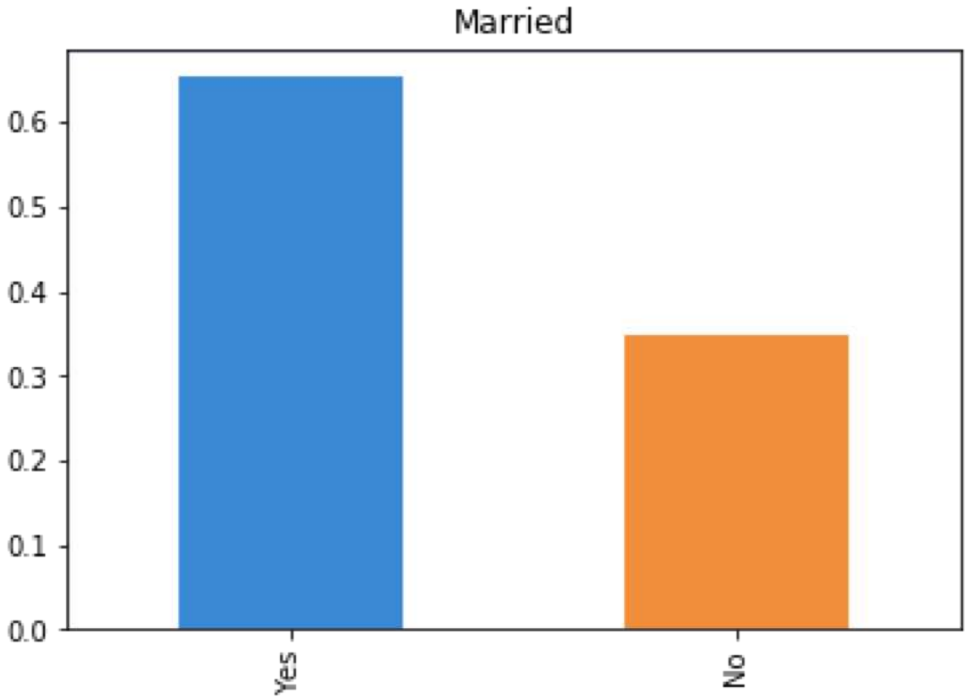


Among all the applicants 78% are graduate and rest are non-graduate

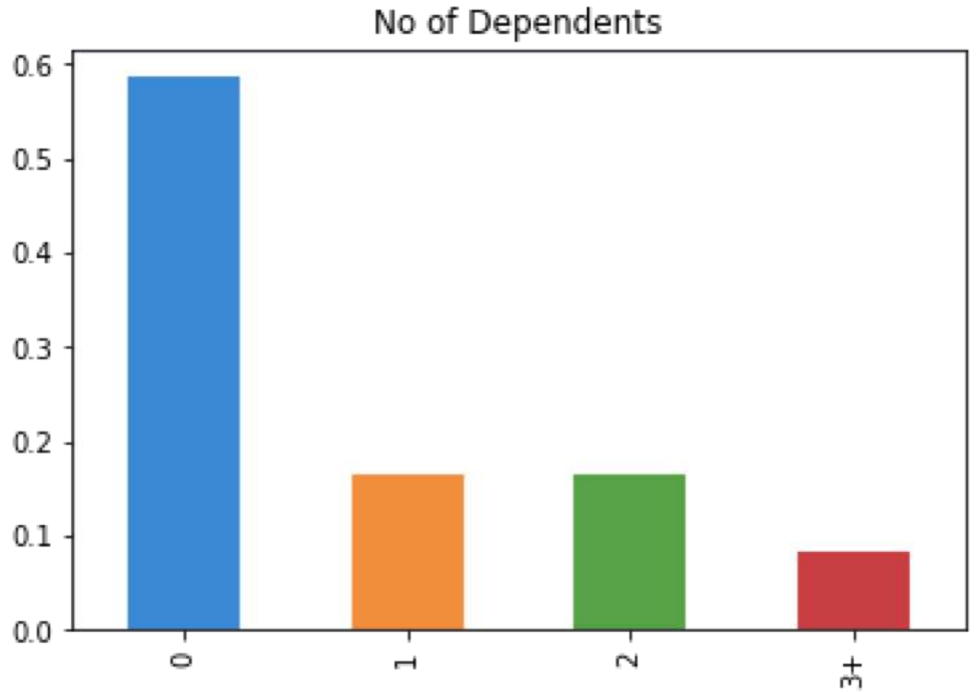


Among all the applicants 86.64% are salaried people or those who are not self employed

Exploring Data: Univariate analysis

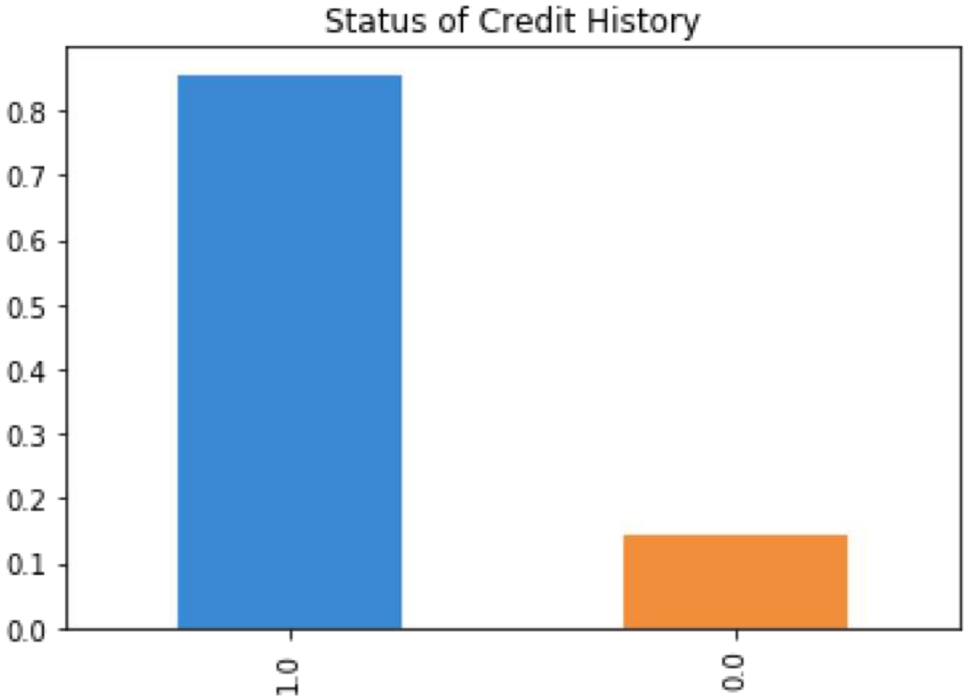


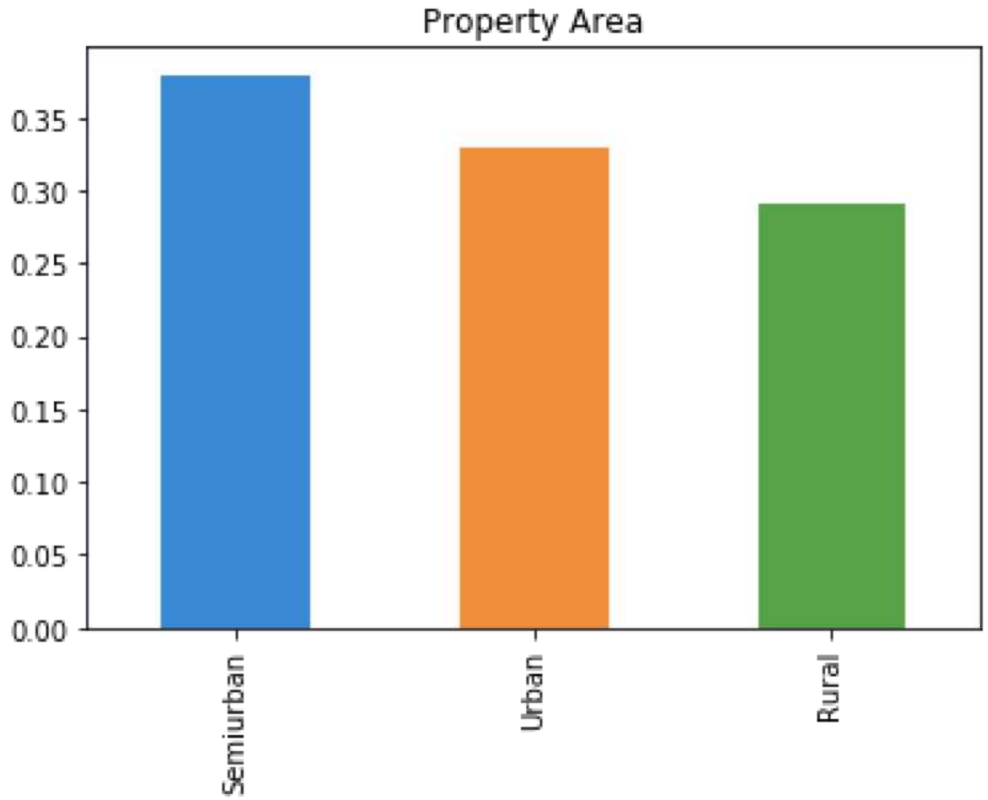
Among all the applicants 65.30% of the applicants are married and 34% are unmarried



among all the applicants around 58.63% of the applicants have no dependents and the applicants with highest 3+ dependents is very less i.e. 8%

Exploring Data: Univariate analysis

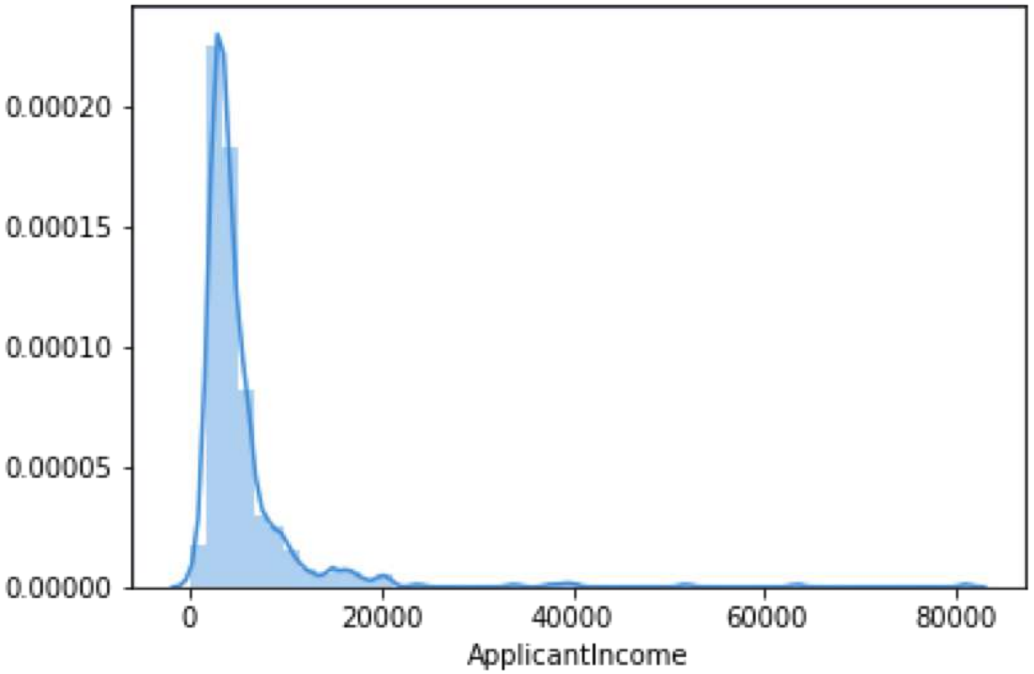


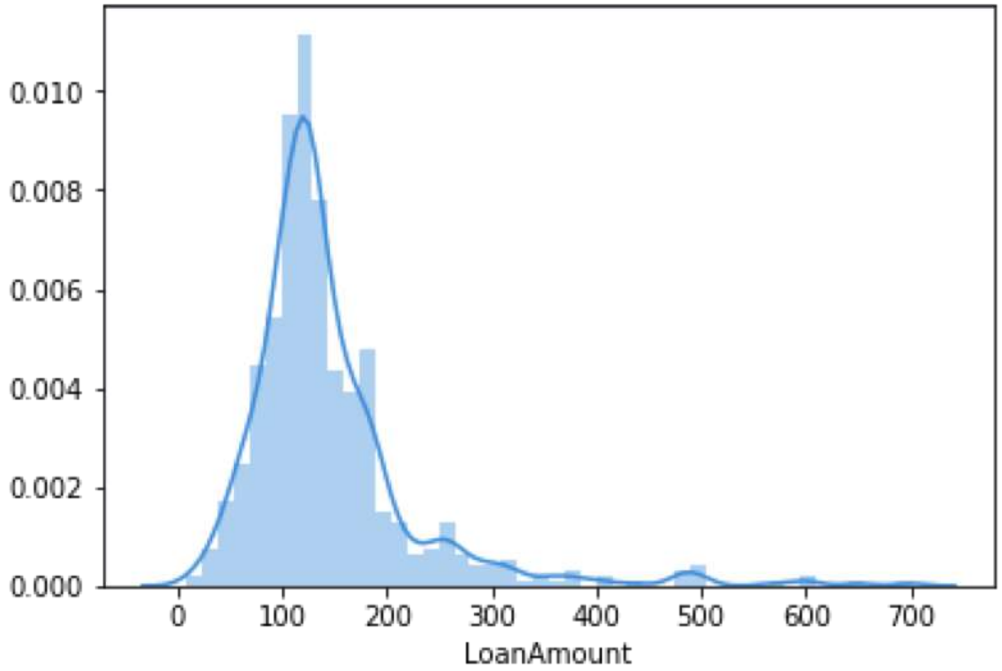


Among all the applicants 85% have good credit history

37.94% of the applicants are taking Home Loan for the properties located in SemiUrban are followed by Urban and Rural area

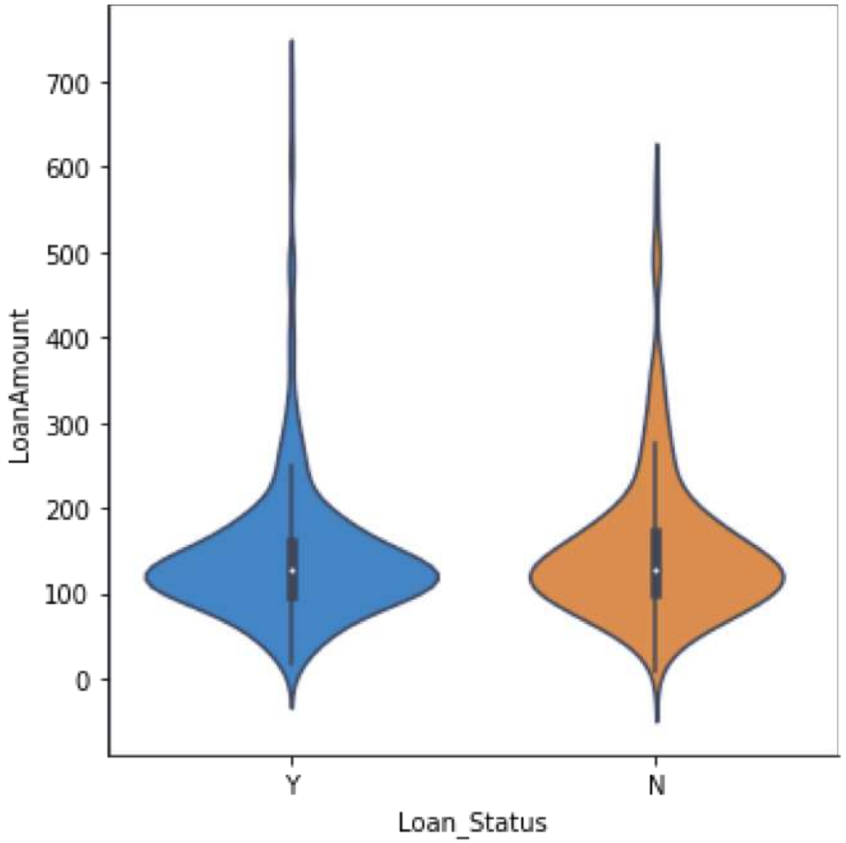
Exploring Data: Univariate analysis



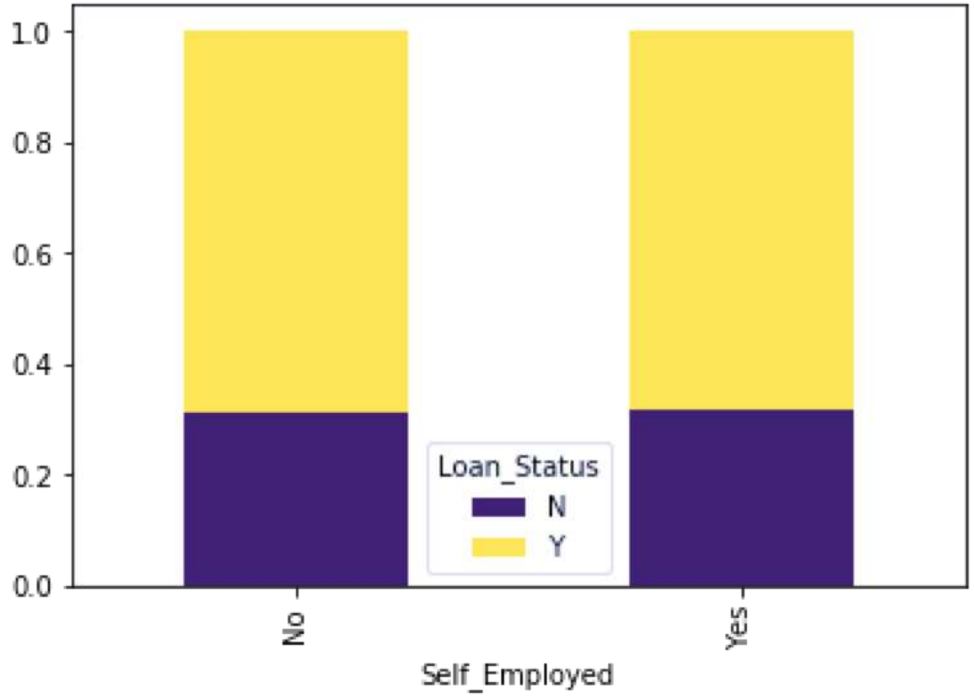


Its quite visible from above applicant income graph that it is positively skewed and not normally distributed while Loan amount graph is slightly right skewed but in general it is normally distributed. Mean Applicant income is 5403.

Exploring Data: Bivariate analysis

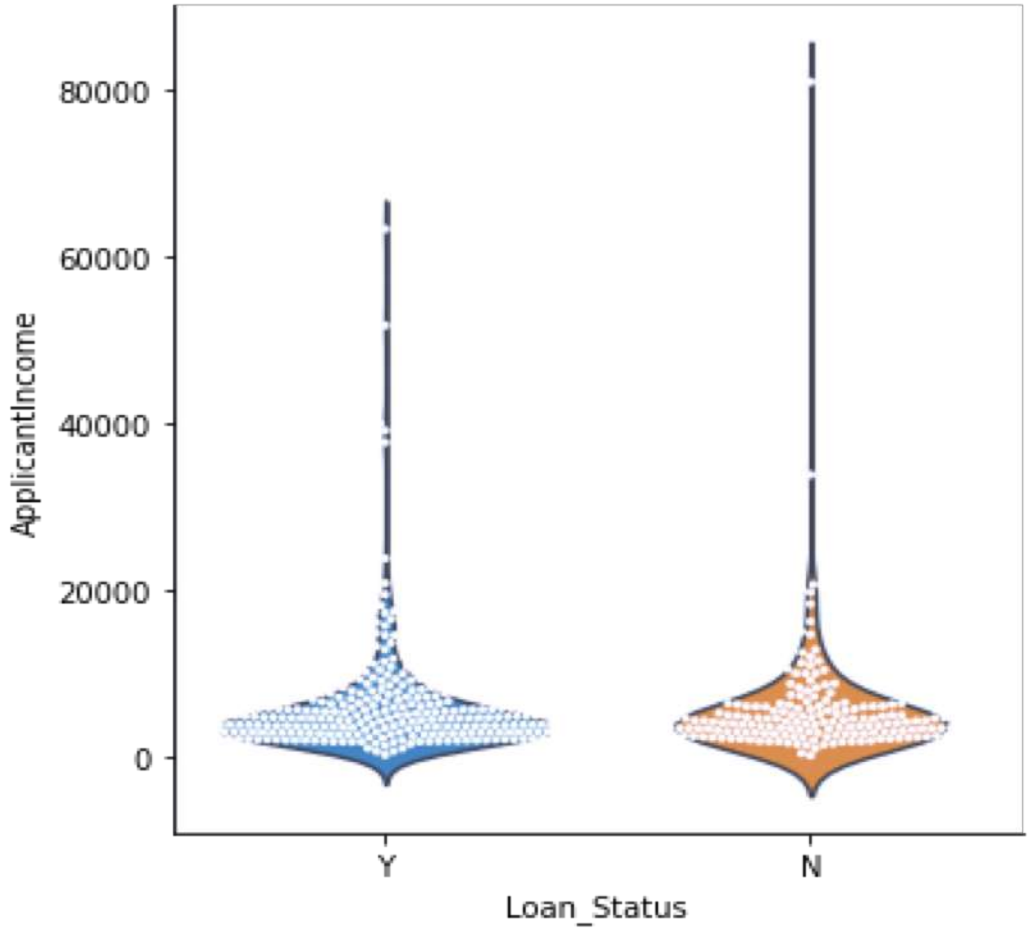


As shown in above graph that at less loan amount the chances of getting loan approved is higher than at high loan amount

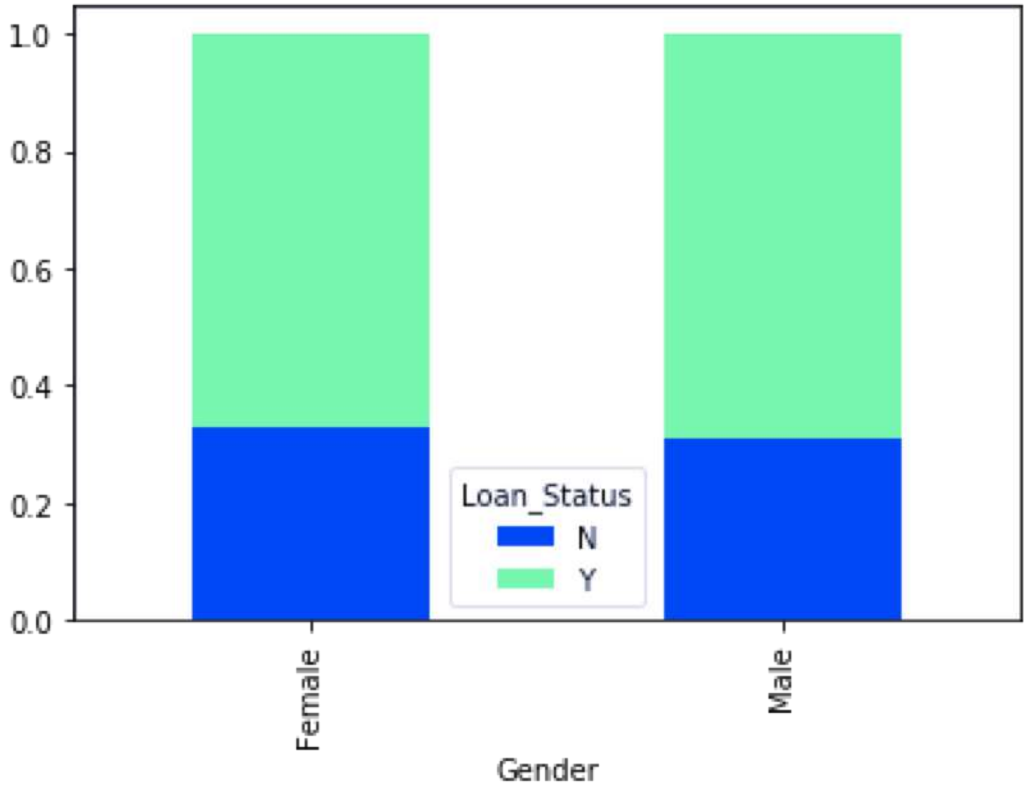


As per above graph it is clearly visible that loan approval status is same for the self employed and not self employed

Exploring Data: Bivariate analysis

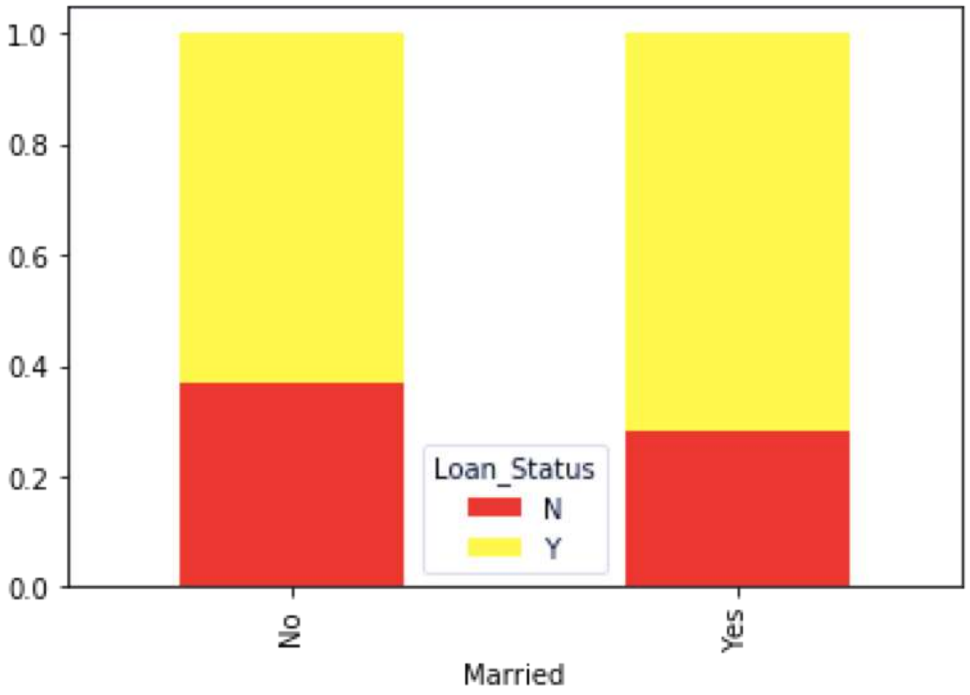


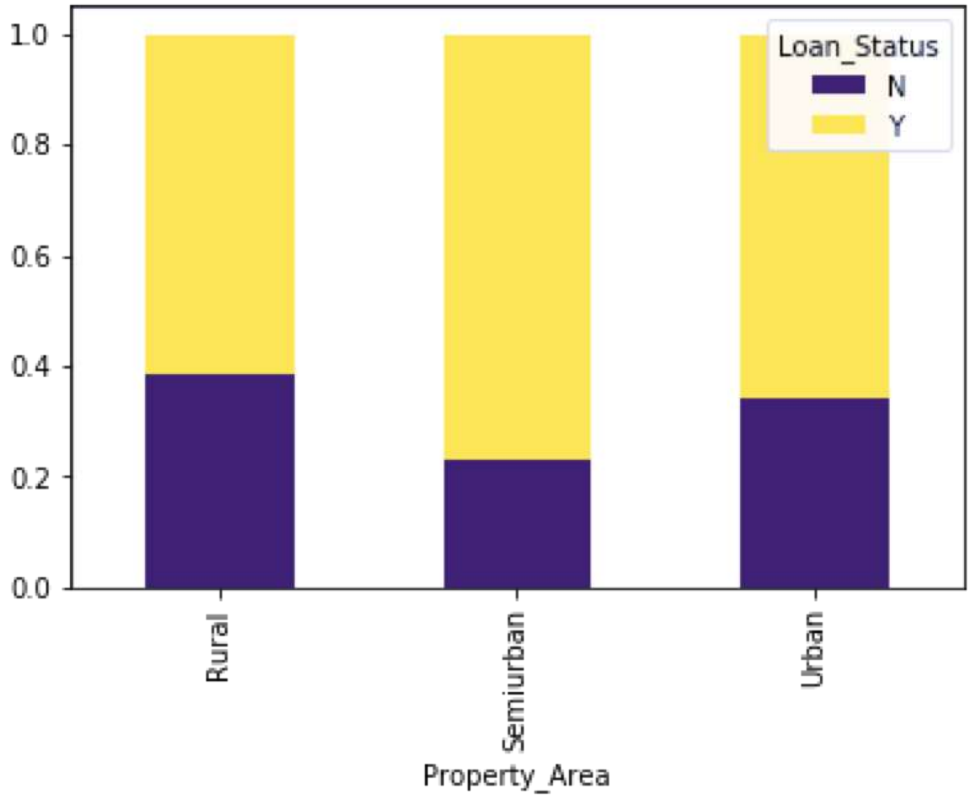
As per above graph we can say that at higher income level number of loan approvals are high



As per above graph it is clearly visible that loan approval ratio is mor or less same for both the gender

Exploring Data: Bivariate analysis

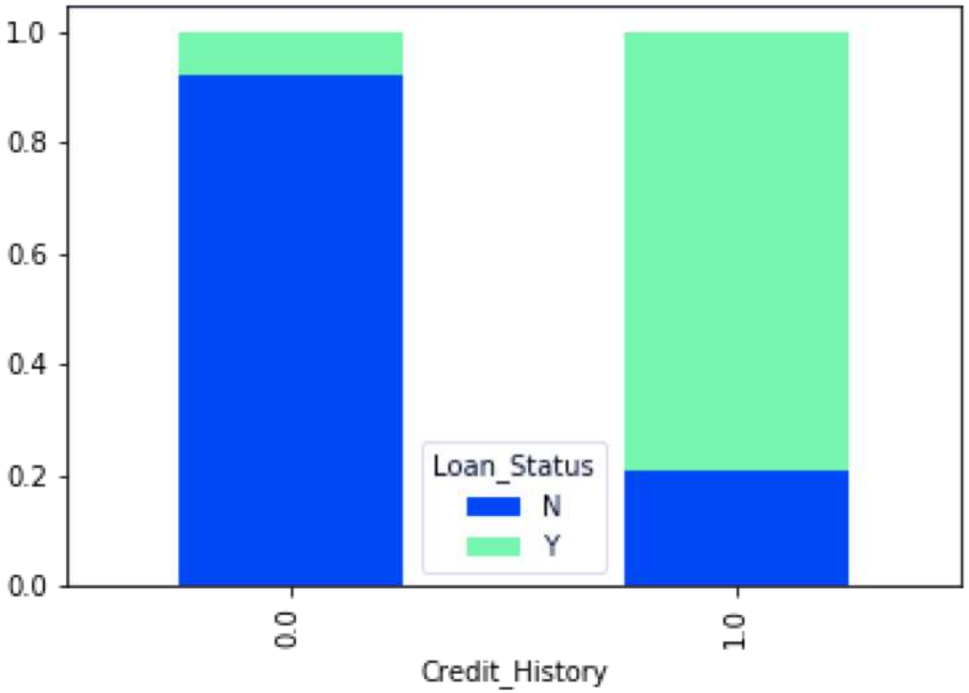




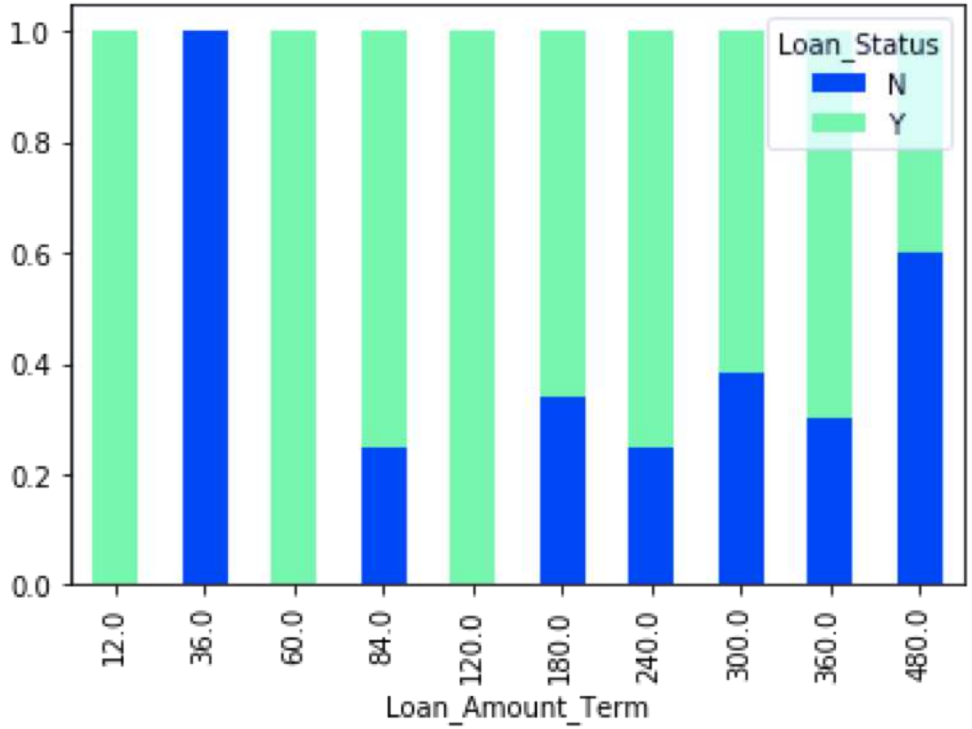
As per above graph our hypothesis regarding marital status was wrong in fact married people has higher chance of getting loan approved

Applicants who are buying property in Semi-Urban area having more loan approved

Exploring Data: Bivariate analysis



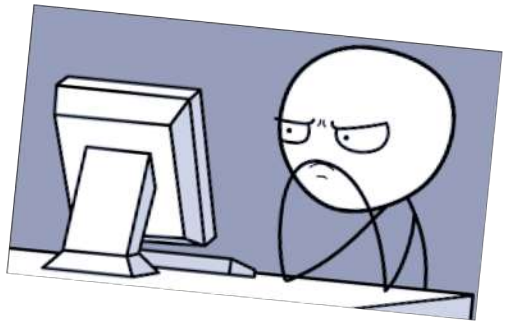
Applicants who have good credit history they have higher proportion of loan approval as compare to the bad credit history applicant.

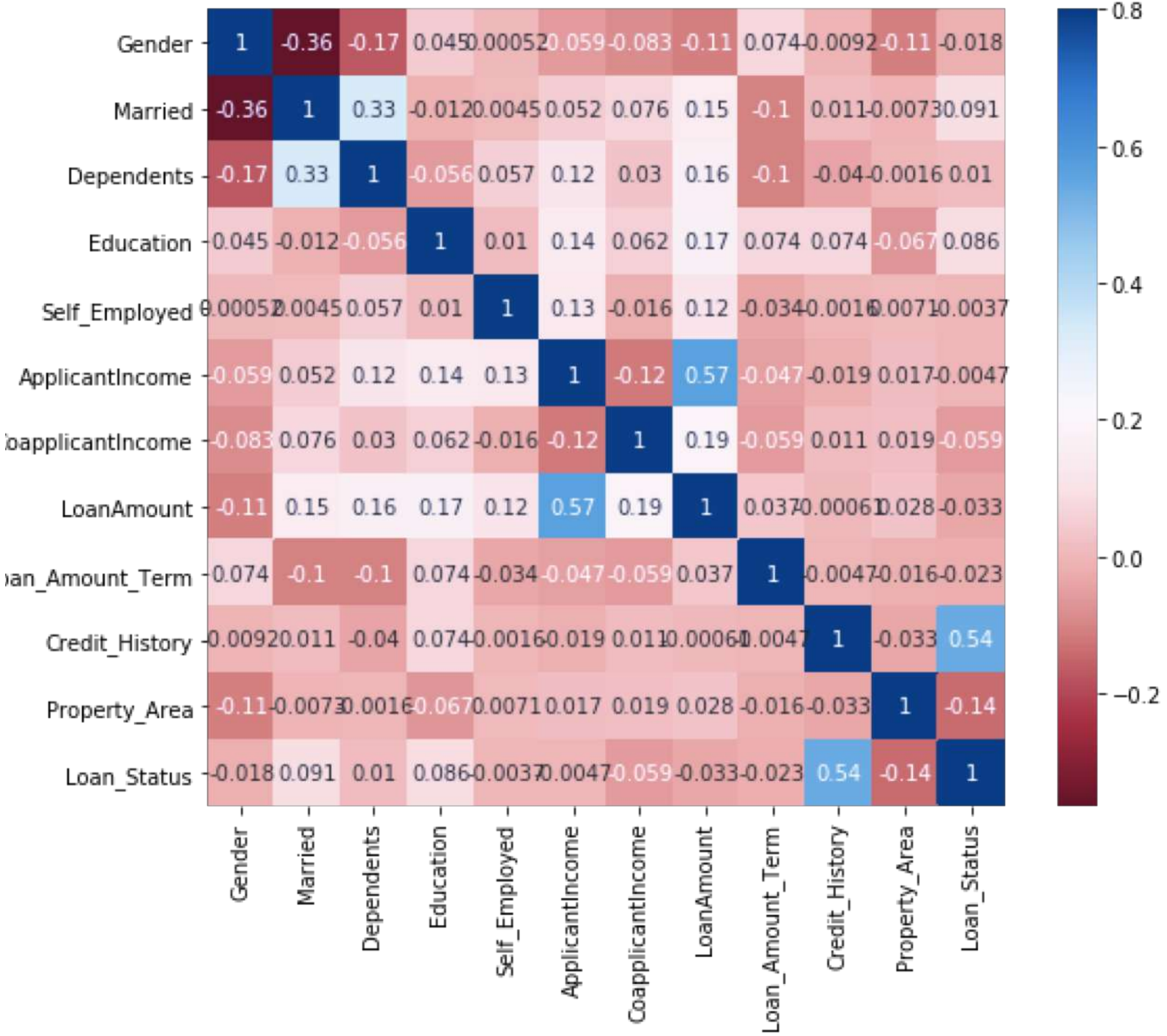


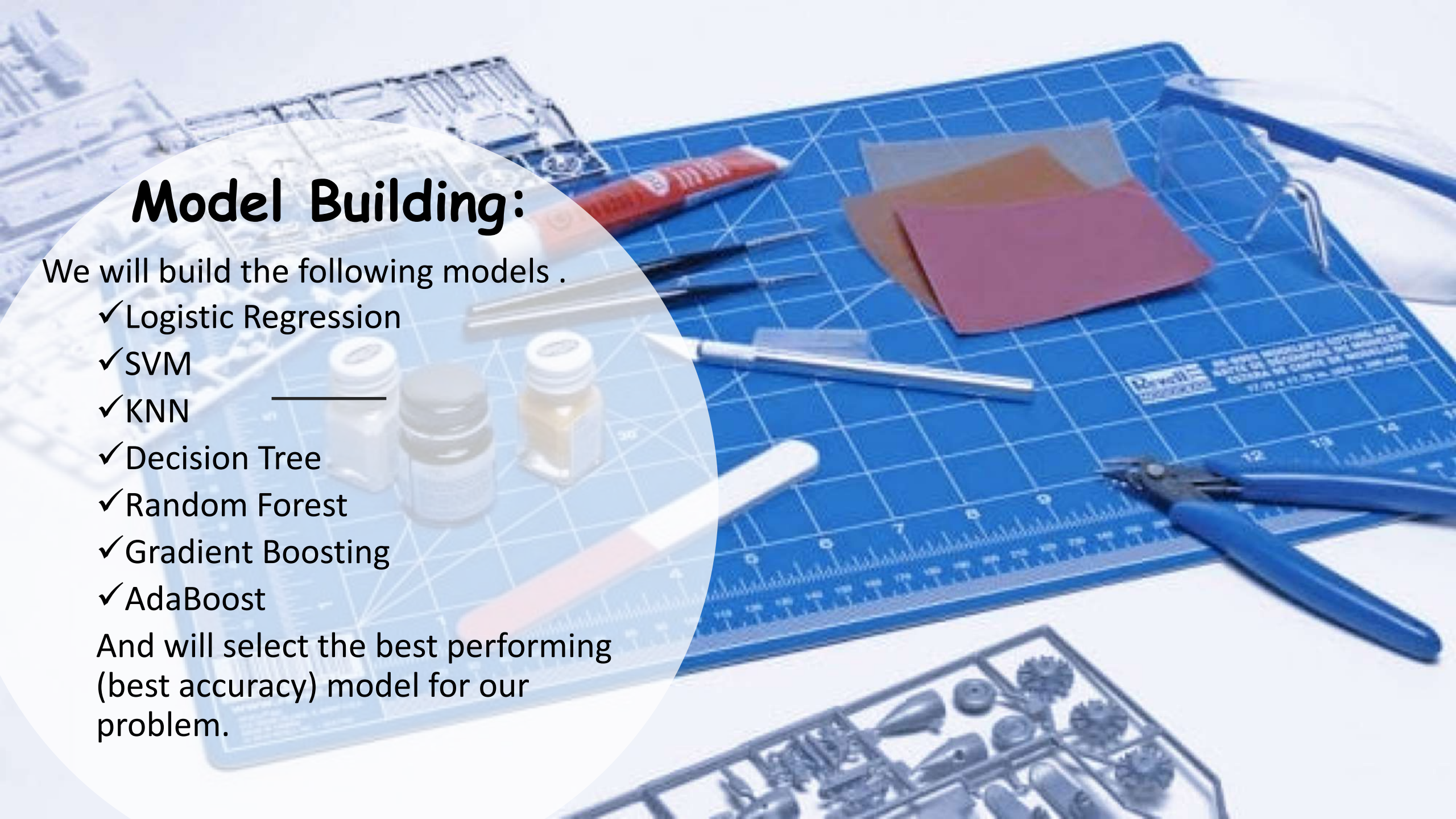
Though its not clear from above graph but holistically we can say that Lower Loan Term period have higher chances of getting loan approved

Correlation Matrix

In above graph The variables with blue color means their correlation is more. Graph shows that Loan status & credit history is correlated and Loan amount & Applicant income is correlated. Marital status and number of Dependents are also related which is quite obvious







Model Building:

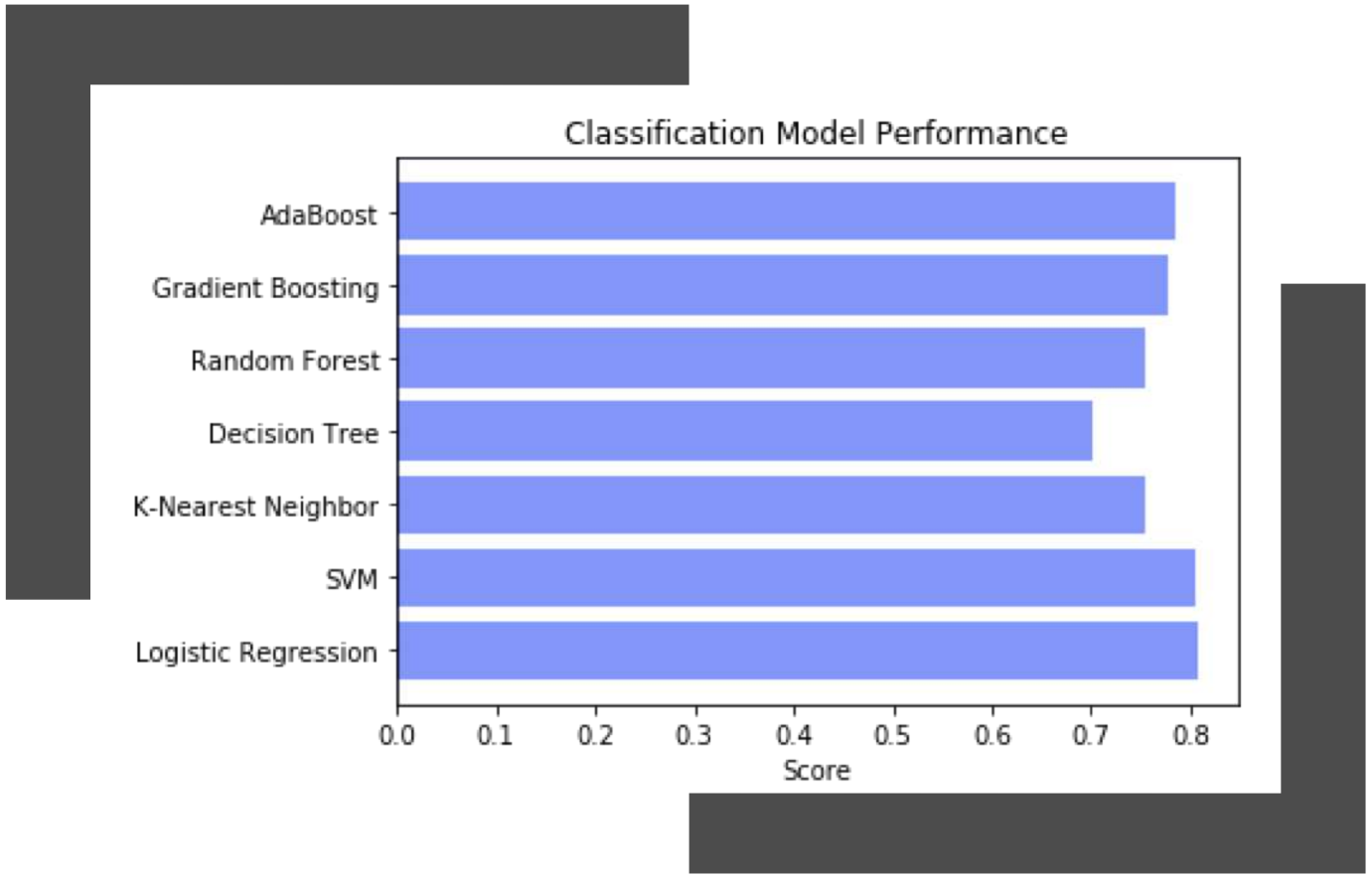
We will build the following models .

- ✓Logistic Regression
- ✓SVM
- ✓KNN
- ✓Decision Tree
- ✓Random Forest
- ✓Gradient Boosting
- ✓AdaBoost

And will select the best performing (best accuracy) model for our problem.

Conclusion:

Accuracy of different models in prediction



After trying and testing 7 different algorithms, the best accuracy is achieved by Logistic Regression (80.80), followed by SVM (80.47) and ADA Boost (78.51).

ML Logistic regression algorithm can be used to automate the loan approval process.

What more can be tried?

- There are still quite a many things that can be tried to improve our models' predictions. We create and add more features, try different models with different subset of features and/or rows, etc. Some of the points are listed below:
- We can train the Adaboost and Random Forest model using grid search to optimize its hyperparameters and improve the accuracy.
- We can also make independent vs independent variable visualizations to discover some more patterns.

THANK YOU.....

DO YOU HAVE ANY QUESTIONS ?

Harsh- harsh.raizada@hotmail.com