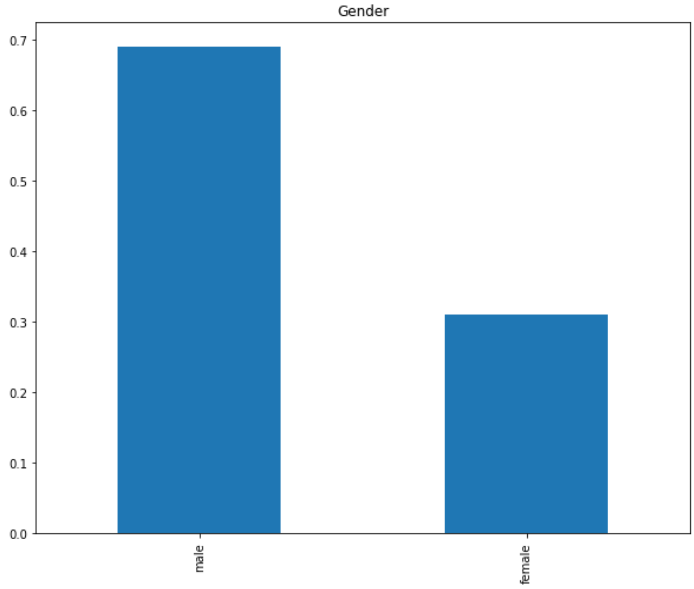
Data Science Assignment

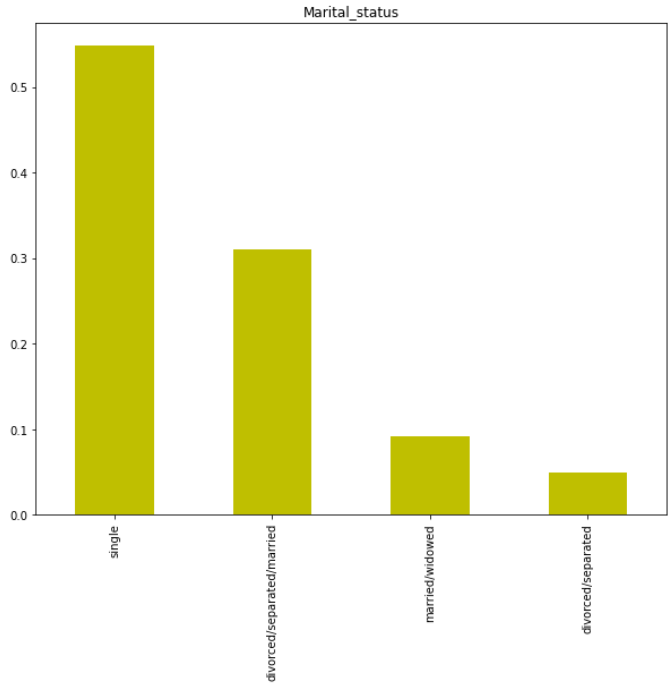
Task-1

1. Do the Exploratory Data Analysis & share the insights.

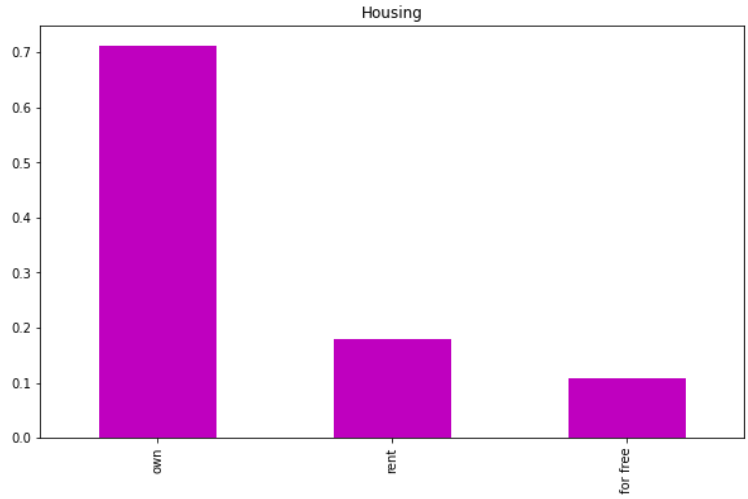
Categorical Features:



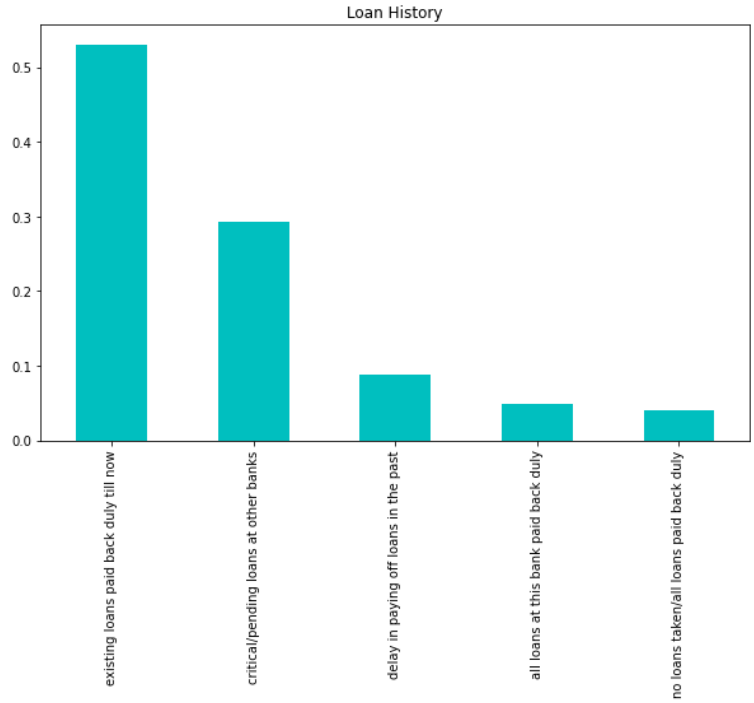
70% of applicants in the dataset are male and 30% of applicants are female.



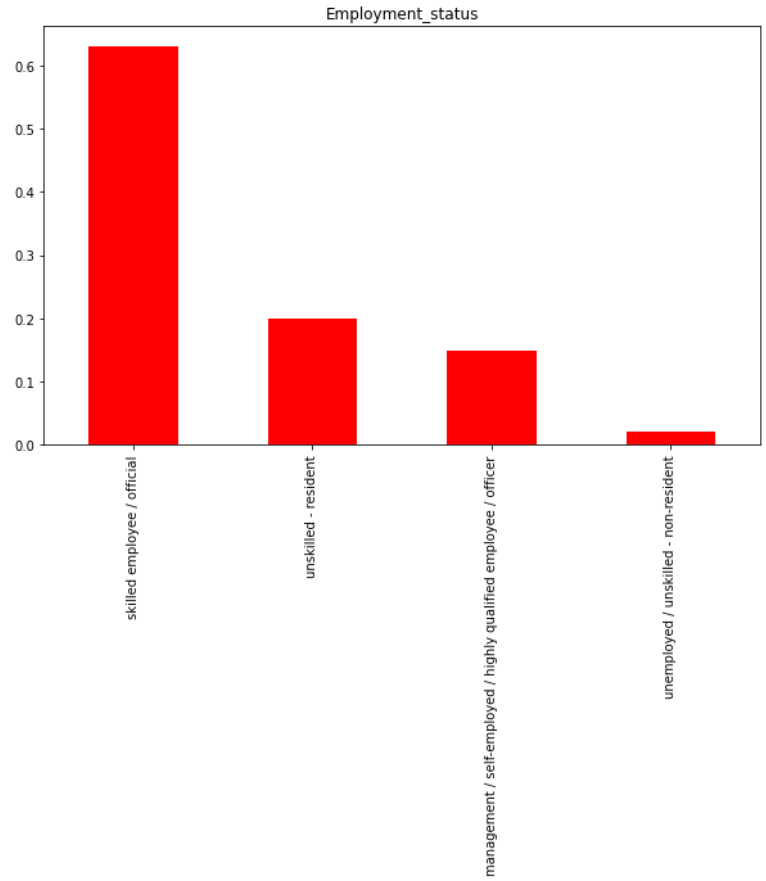
53% of applicants in the dataset are Single and 30% of applicants are married.



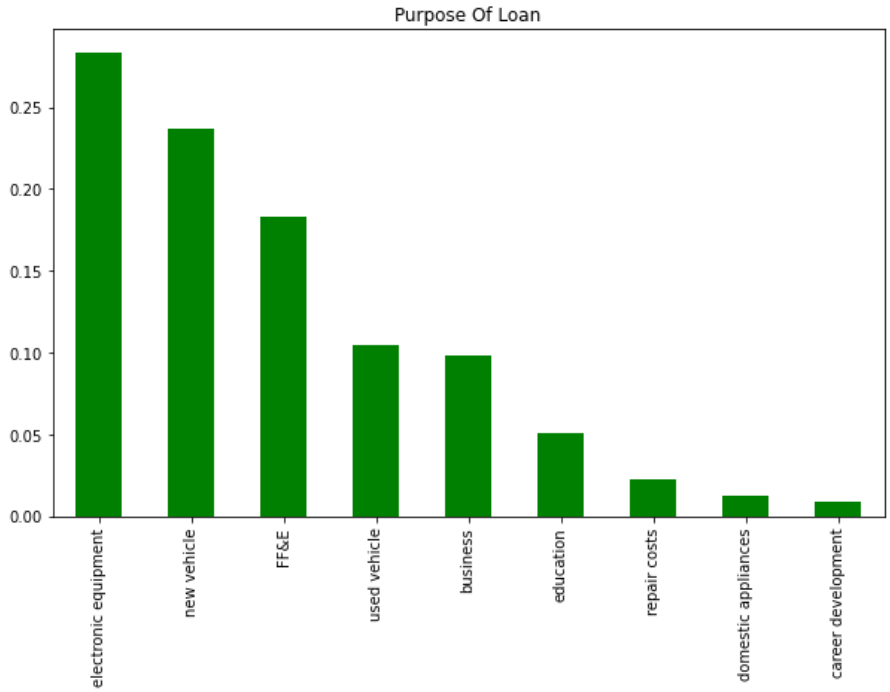
70% of applicants in the dataset are have own house and 20% of applicants are lived on rent.



52% of applicants have paid the existing loans back duly till now and 30% of applicant are critical/pending loans at the other banks.



60% of applicants are skilled employee and 20% of applicants are unskilled.

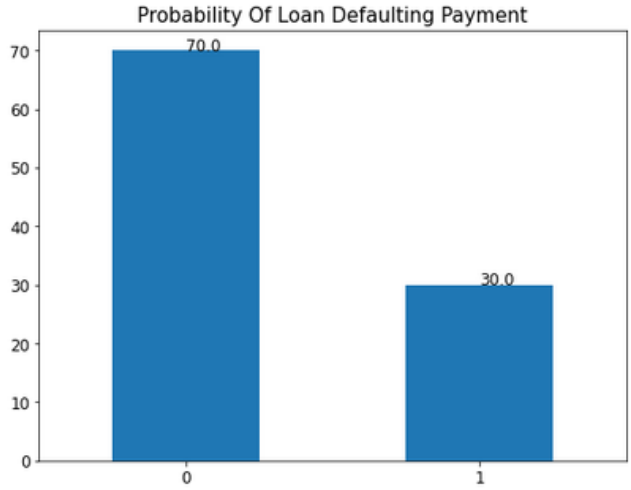


27% of applicant taken loan for electronic equipment and least taken for career development is 2%-3%.

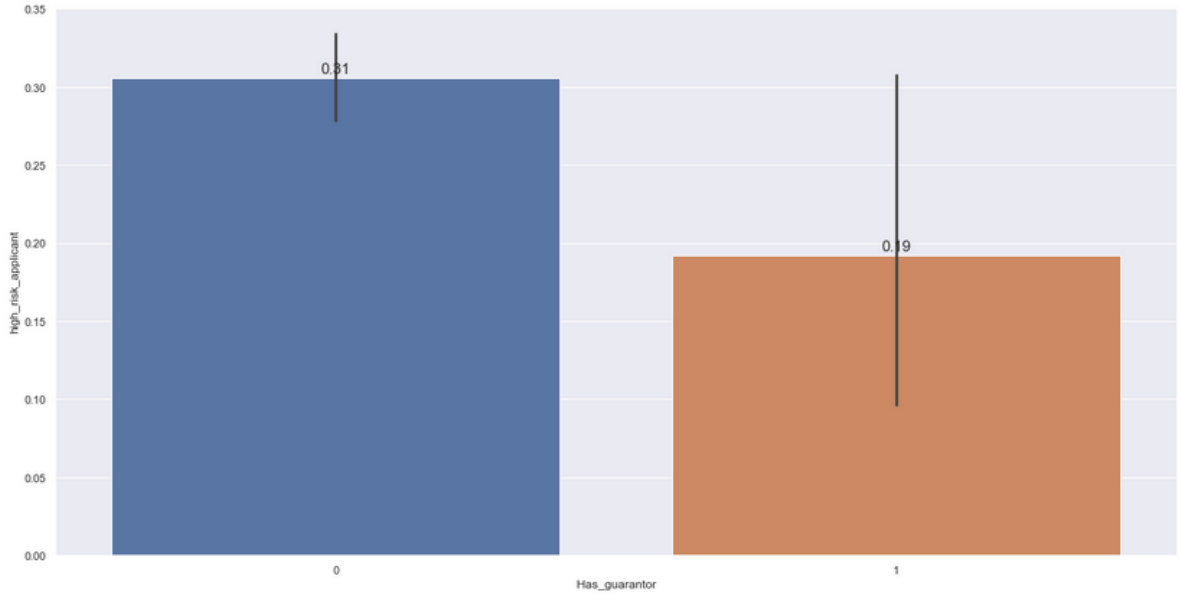
Numerical Features:

- There is linear relationship between Principal_loan_amount and months_loan_taken_for.
- There are some outliers in Principal_loan_amount.
- There is slight inverse relationship between Principal_loan_amount and Primary_applicant_age_in_years which makes sense as more elderly people tend to avoid bigger loans and young people take bigger loans for house and vehicles.
- There is inverse relationship between Principal_loan_amount and Number_of_existing_loans_at_this_bank. The more loans are open, the smaller is the ability of people to take big loan amounts.
- Linear relationship between Primary_applicant_age_in_years and Years_at_current_residence.

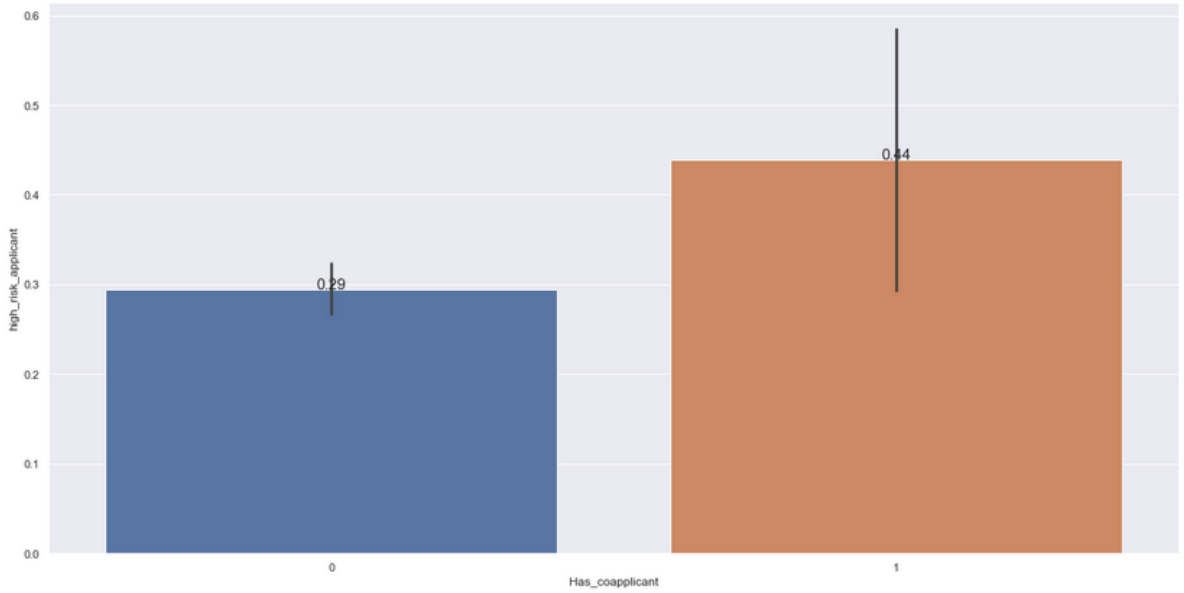
Independent Variables:



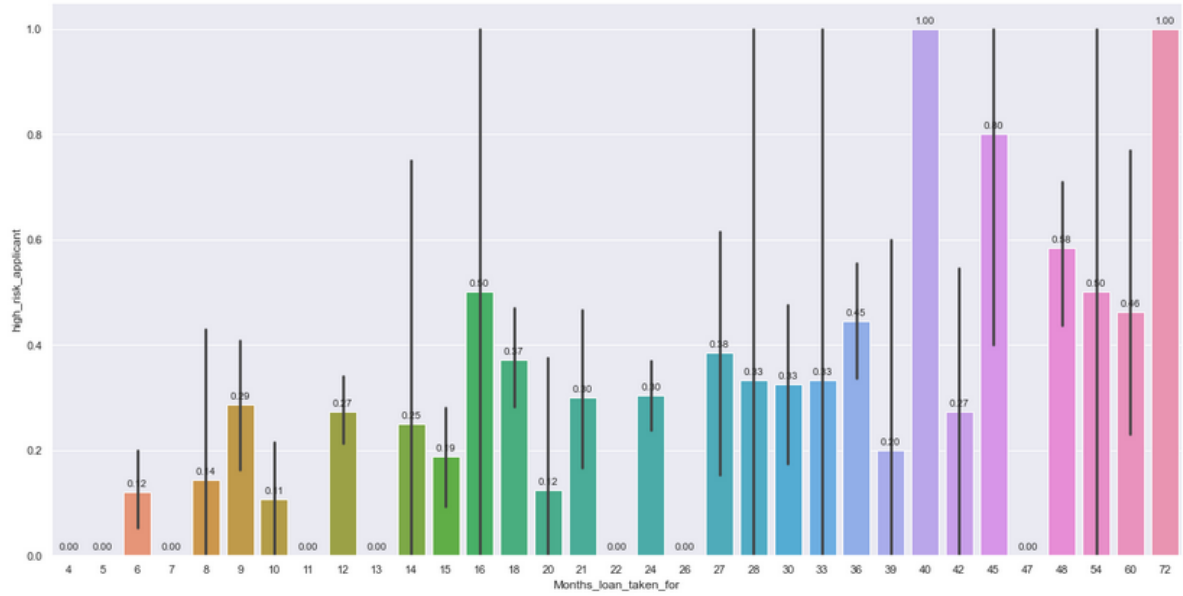
We can see that the datasets consist of 70% applicants are not expected to default payment whereas 30% applicants are expected to default the payment.



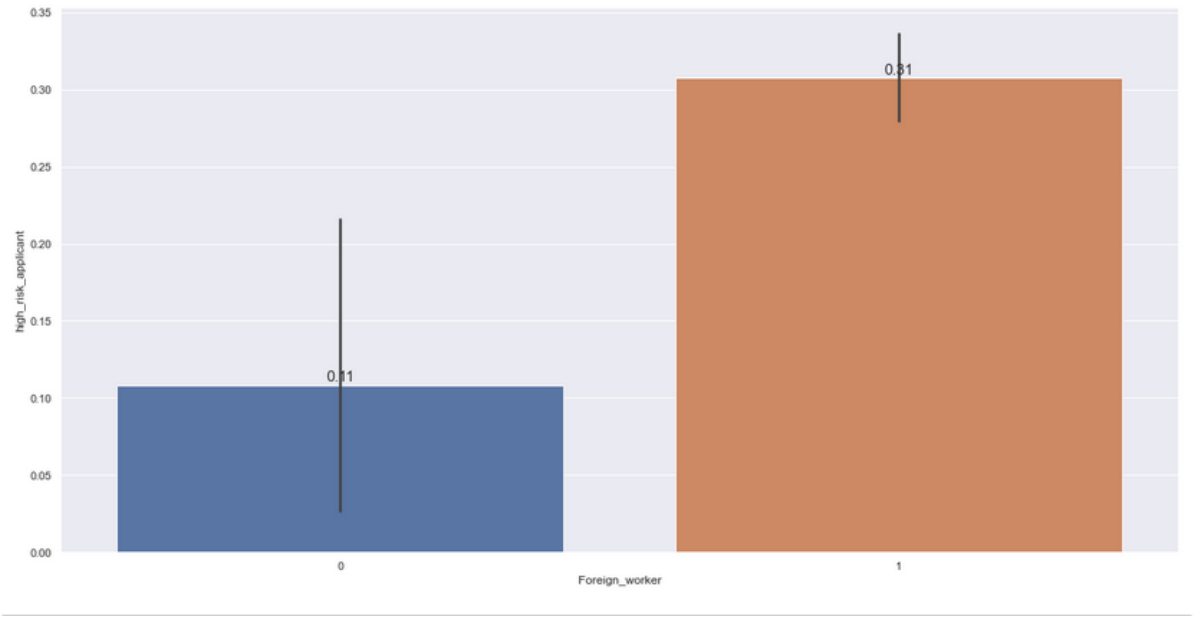
Applicants with the guarantor has less chance of defaulting compare to not having the guarantors.



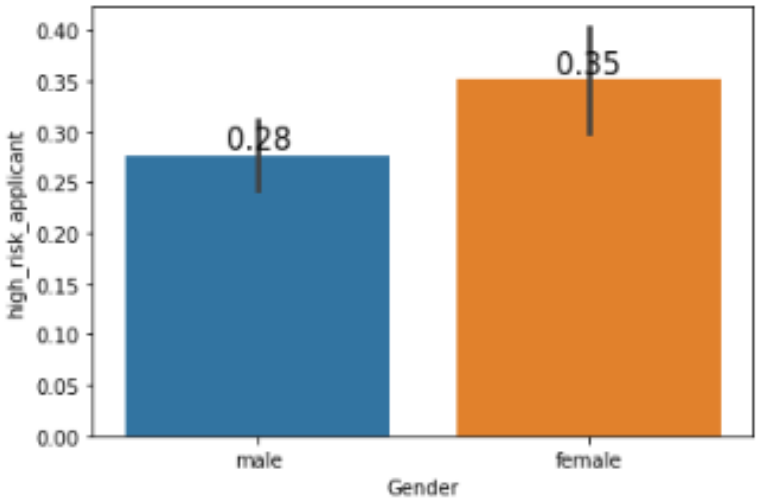
Applicants who don't have co-applicants have high chance of defaults is 44%.



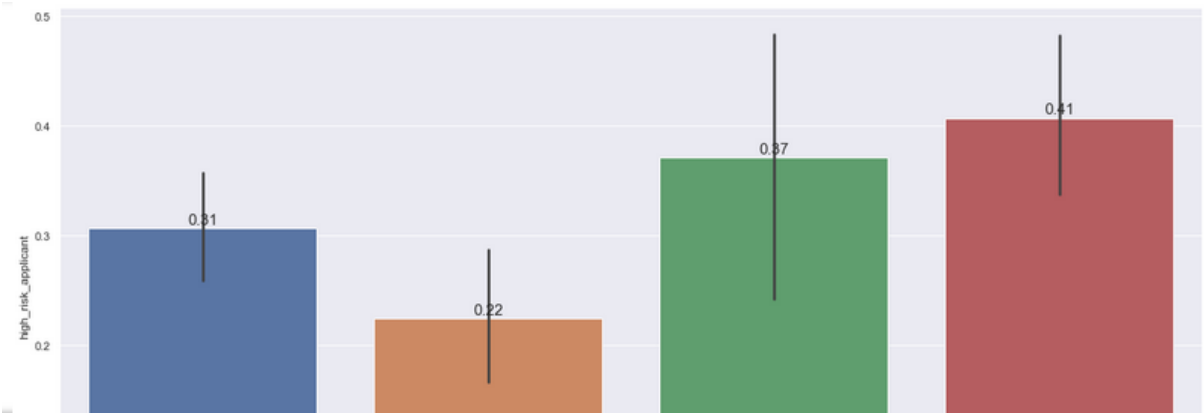
We can clearly see if loan period is more than 40 months than the chance of defaulting is increased.



We can see applicant working in foreign has 31% chance of defaulting.

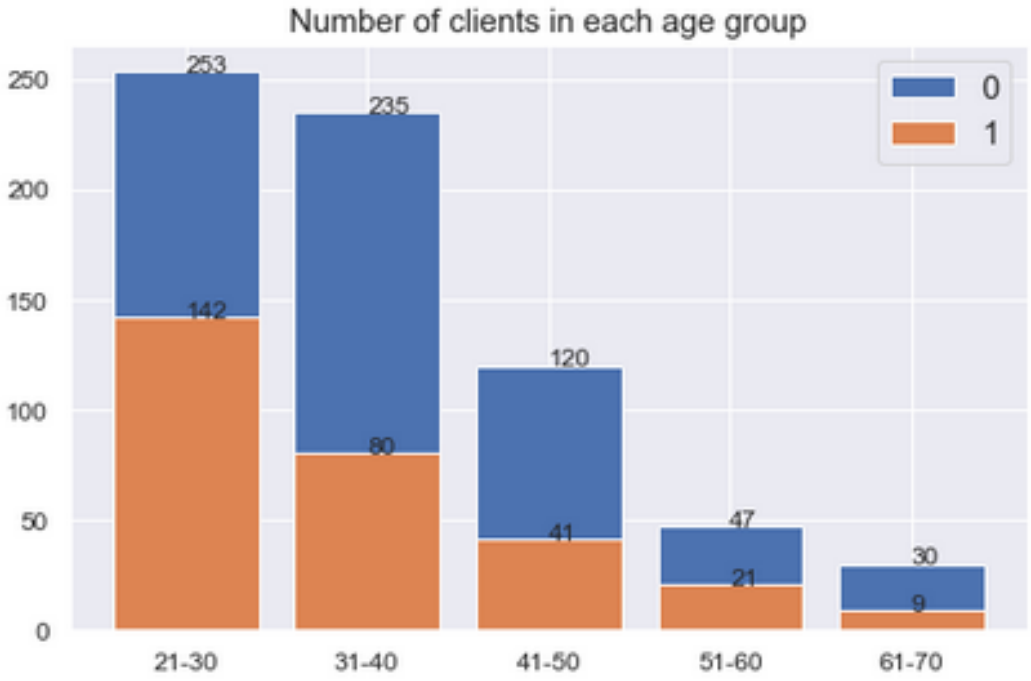


In Male 28% applicants are chance for default the payment and 35% for female.



If person works for more than 4 years have less chance to defaults.

2. How would you segment customers based on their risk (of default).



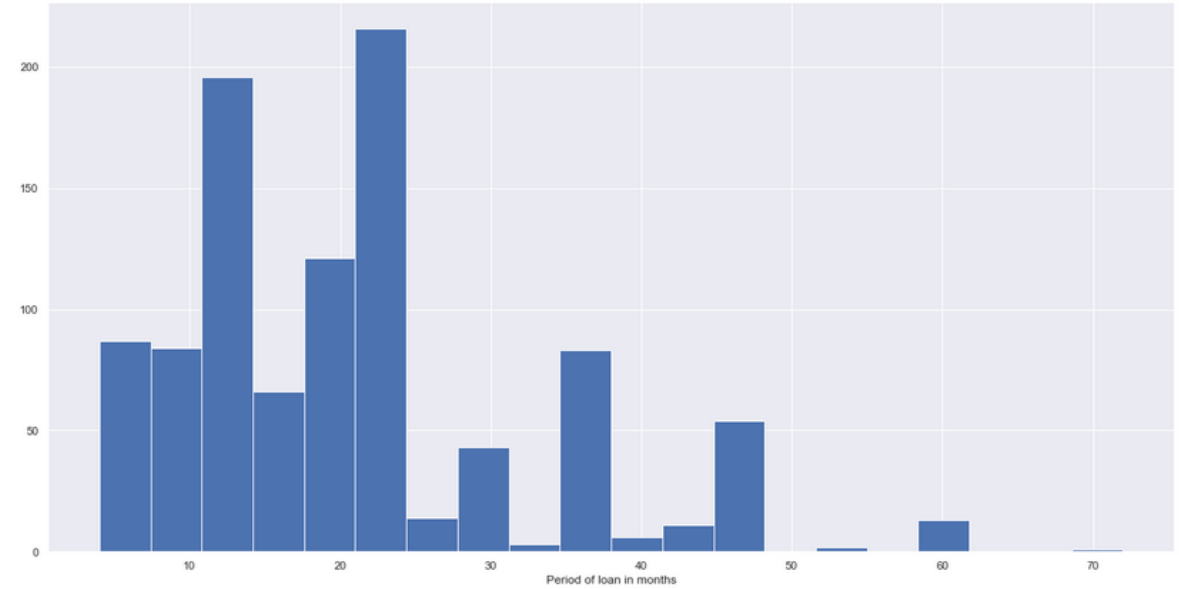
3. Which of these segments / sub-segments would you propose be approved?

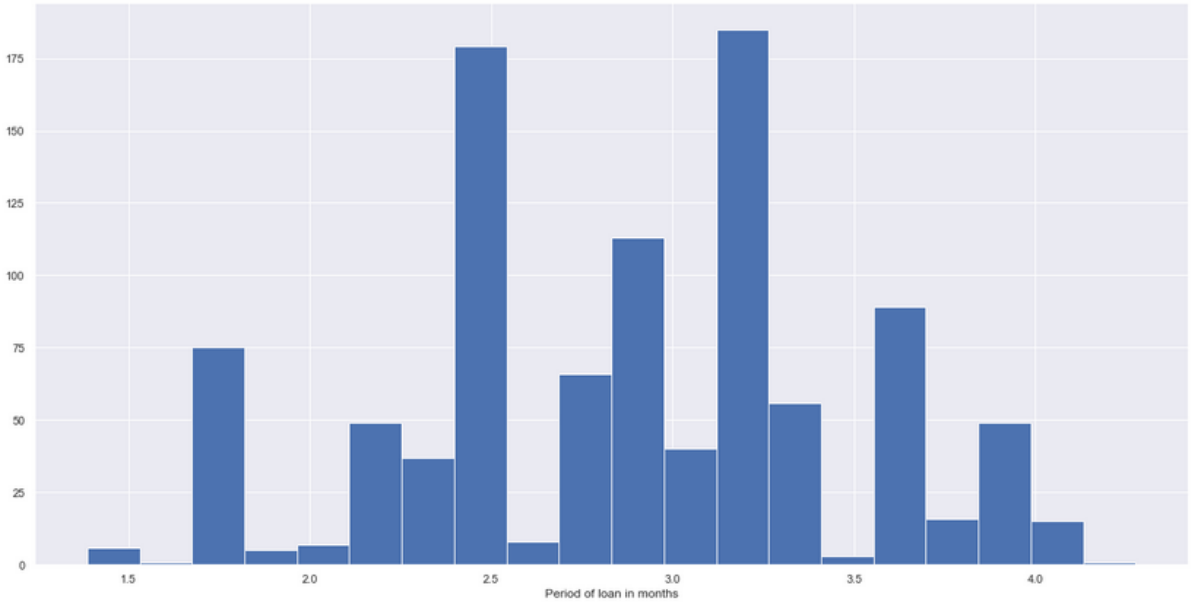
For e.g., Would a person with critical credit history be more creditworthy? Are young people more creditworthy? Would a person with more credit accounts be more creditworthy?

Looking to above segment chart we clearly see Age between 21-30 have high chance of default of Payments 142 applicants default out of 253. Since Person with the more credit accounts are creditworthy because of history payments of loan give confidence to banker to lend moneys. So, if persons is older and have at most 4 years of employees experience and has good creditworthy pay all his loan on duly has higher chance of loan getting.

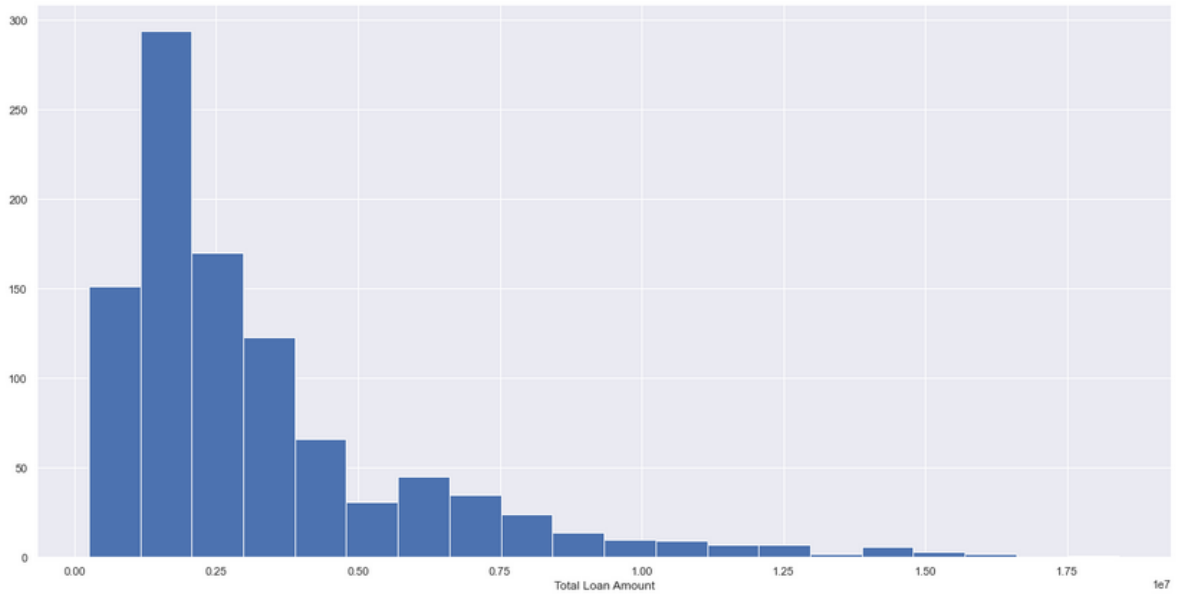
4. Tell us what your observations were on the data itself (completeness, skews).

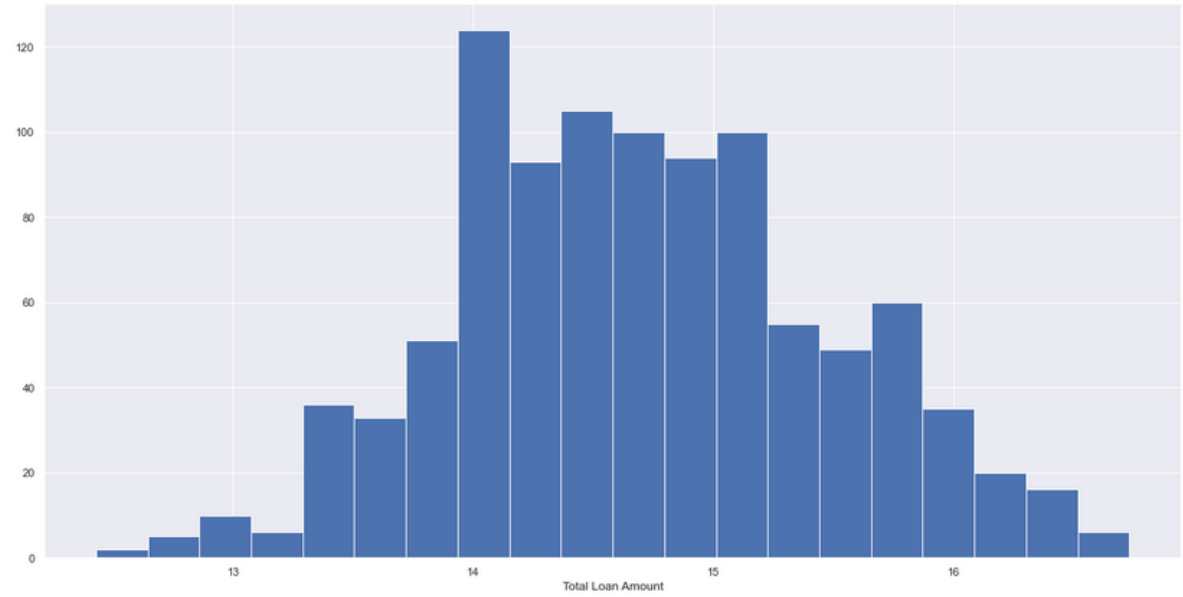
From the correlation value, we can see that Months_loan_taken_for and Principal_loan_amount has a very good correlation with target variable.





The above graph of 'Months_loan_taken_for' having a numerical data and it is Right Skewed so to overcome the skewness, Log function is performed to covert right skewed distribution to normal distribution.





For 'Principal_loan_amount' also, the graph is right skewed so to overcome the skewness, Log function is performed to covert right skewed distribution to normal distribution.

Task-2

1. Explain your intuition behind the features used for modeling.

After doing data pre-processing, we can get the idea about the features that affects the target variable which involves handling categorical variables i.e., converting into dummy variable using One Hot Encoding, dropping some columns, and dealing with null (nan) and different values.

2. Are you creating new derived features? If yes explain the intuition behind them.

There were no derived features but the existing features are added twice as dummy variable are created using One Hot Encoding because target variable has a value in the form of binary classification.

3. Are there missing values? If yes how you plan to handle it.

There are missing values in the dataset which shows error while building model so 0 values are assigned in place of that and some different types of values are also handled like string values converted into integer values.

4. How categorical features are handled for modeling.

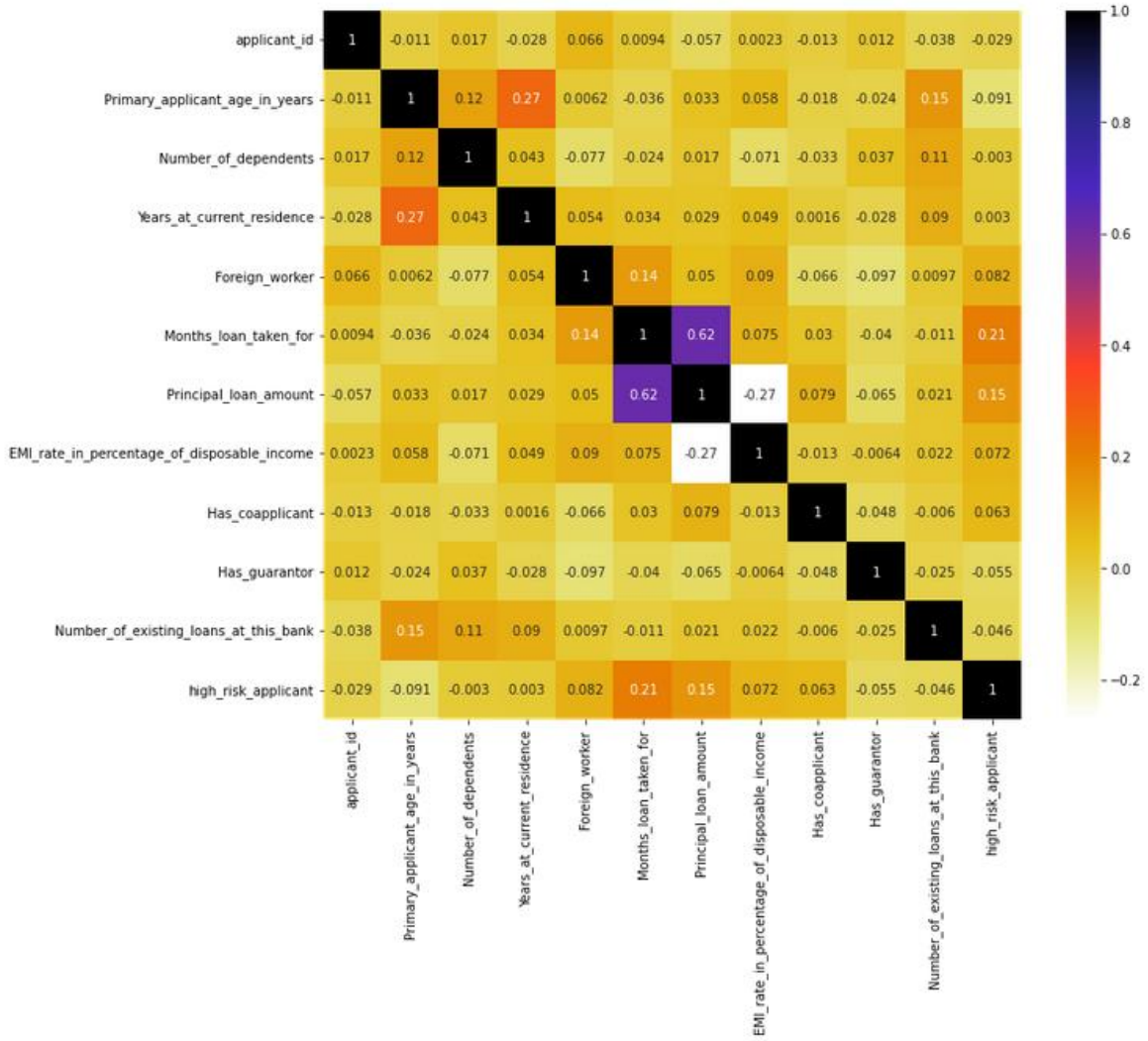
Some of the categorical features are not adding value to target variable so they are dropped and rest features are converted to dummy variable using One Hot Encoding.

5. Describe the features correlation using correlation matrix. Tell us about few correlated feature & share your understanding on why they are correlated.

From the correlated matrix, we can see that Months_loan_taken_for and Principal_loan_amount has the strong correlation with target variable means if a person has taken the loan from very long time and not repaying any amount to the bank so the applicant is not creditworthy in that case.

applicant_id, Primary_applicant_age_in_years, Number_of_dependents, Has_guarantor, Number_of_existing_loans_at_this_bank has a negative correlation with target variable means if we drop this columns than it will not bring any effect on our trained model.

Foreign_worker, Has_coapplicant, EMI_rate_in_percentage_of_disposal_income has not so strong correlation but it will bring about 30% effect on target variable means if a applicant is foreigner and he/she has coapplicant then there are 30% chances of default.



6. Do you plan to drop the correlated feature? If yes then how.

Strongly and moderately correlated feature are not dropped and rest other are dropped as those features will not bring effect in predicting target value. It can be dropped using function drop() in python.

7. Which ML algorithm you plan to use for modeling.

Regression (Logistic) and Classification(Decision Tree, Random Forest, Support Vector Machine).

8. Train two (at least) ML models to predict the credit risk & provide the confusion matrix for each model.

Logistic Regression

```
Confusion Matrix:
 [[560   0]
 [240   0]]
```

Decision Tree Classifier

```
Confusion Matrix:
 [[516  44]
 [ 45 195]]
```

Random Forest Classifier

```
Confusion Matrix:
 [[560   0]
 [  0 240]]
```

Support Vector Machine Classifier

```
Confusion Matrix:
 [[552   8]
 [223  17]]
```

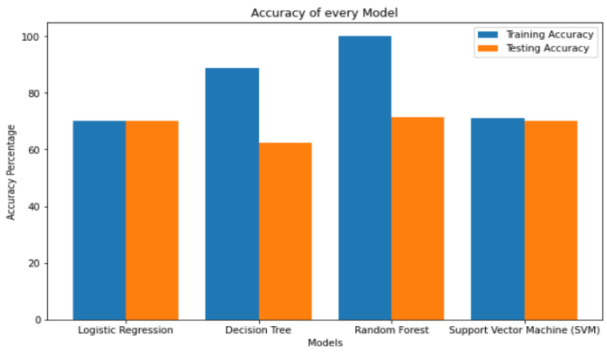
9. How you will select the hyperparameters for models trained in above step.

By training model with the help of certain number of epochs we can look onto the accuracy score and model score value to again train the model as per the need to optimize the existing model.

10. Which metric(s) you will choose to select between the set of models.

Confusion matrix, Training and Testing accuracy score can be used to choose between the model and we can see that desired output is generated close in Logistic Regression and SVM algorithm.

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	70.000	70.0
1	Decision Tree	88.875	62.5
2	Random Forest	100.000	71.5
3	Support Vector Machine (SVM)	71.125	70.0



11. Explain how you will export the trained models & deploy it for prediction in production.

We can save i.e., deploy the trained model in HDF5 form i.e., h5 format file and simply load the model which can directly use for the prediction of credit (low/high risk) of any new or existing data.