**Name-Kushagra Srivastava**

**Intern-ID=EI0050**

**Project-4 – Prediction of Consumer Personal Loan-REPORT**

**DATA TYPES OF EACH FEATURE INVOLVED IN DATASET:**

**Loanapp\_ID object**

**Sex object**

**Marital\_Status object**

**first\_name object**

**last\_name object**

**email object**

**address object**

**Dependents object**

**Qual\_var object**

**SE object**

**App\_Income\_1 float64**

**App\_Income\_2 float64**

**CPL\_Amount float64**

**CPL\_Term float64**

**Credit\_His float64**

**Prop\_Area object**

**INT\_ID int64**

**Prev\_ID object**

**AGT\_ID object**

**CPL\_Status object**

**dtype: object**

**CPL\_Status is our target variable**

**DESCRIPTION ABOUT DATA TYPES INVOLLVED IN DATASET:**

**object: Instance of a class**

**int64: It represents the integer variables**

**float64: It represents the variable which have some decimal values involved**

**Dimensionality of train dataset**

**(614, 20)**

**UNIVARIATE ANALYSIS**

**CPL\_STATUS - FREQUENCY**

**Y 422**

**N 192**

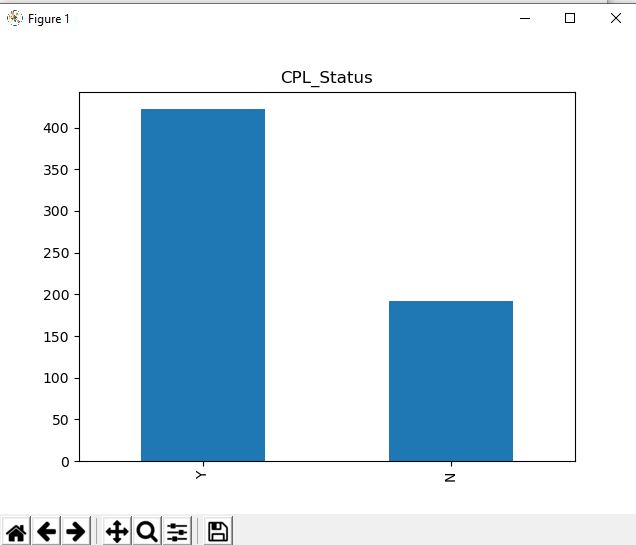
**Name: CPL\_Status, dtype: int64**

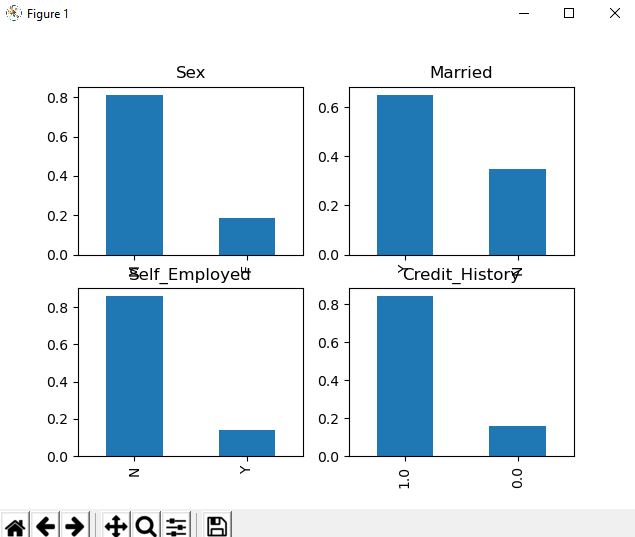
**LOAN\_STATUS - PROPORTION**

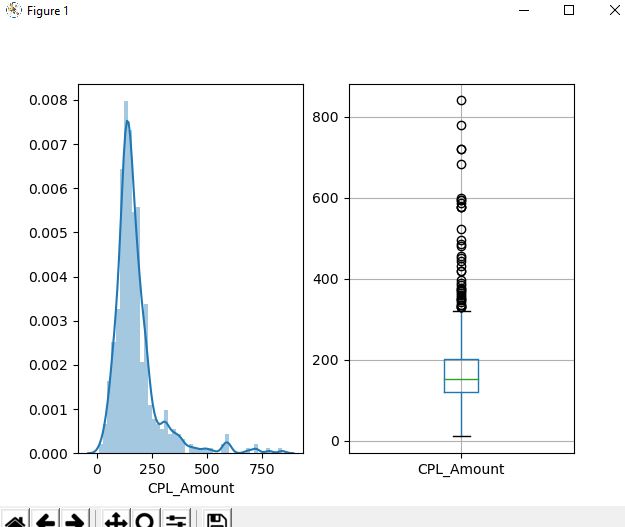
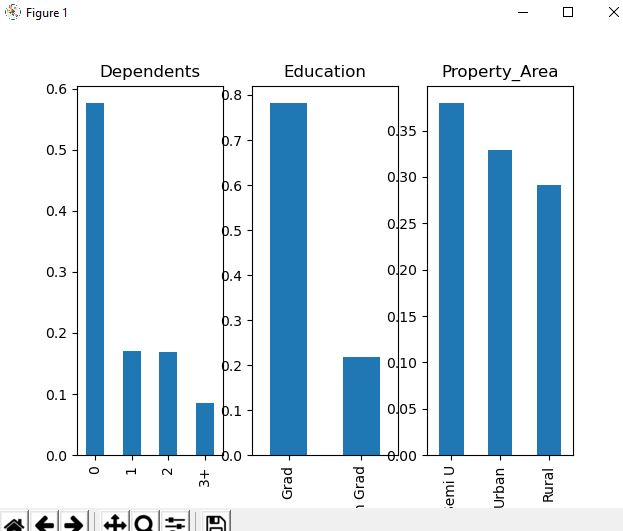
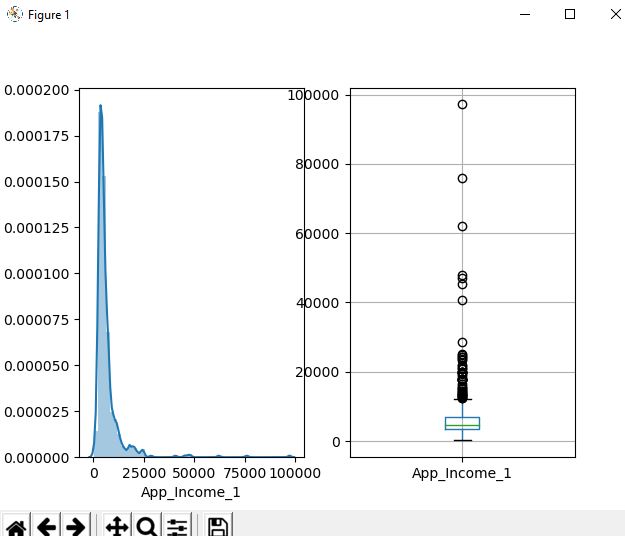
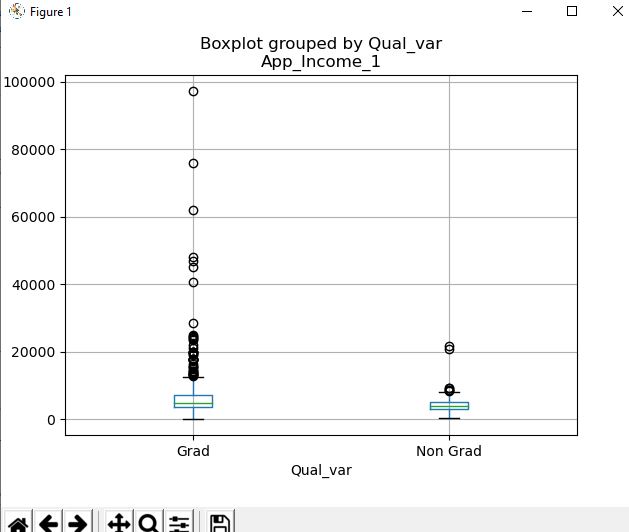
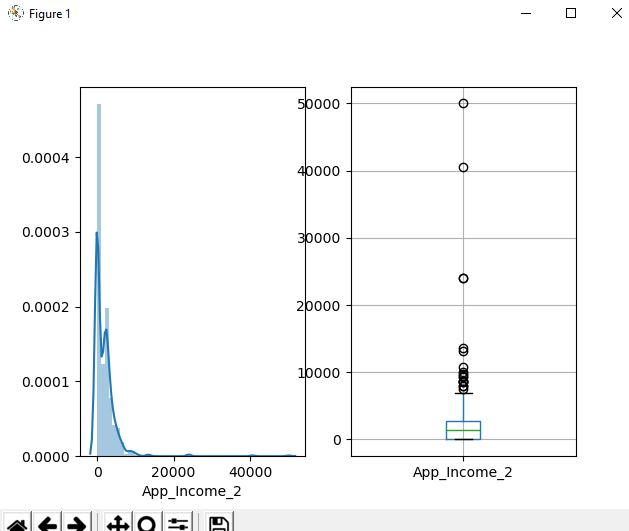
**Y 0.687296**

**N 0.312704**

**Name: CPL\_Status, dtype: float64**

****

****

****

**BIVARIATE ANALYSIS**

**Categorical Independent Variable vs Target Variable**

**Split up of Loan\_Status based on Gender**

**CPL\_Status N Y**

**Sex**

**F 37 75**

**M 150 339**

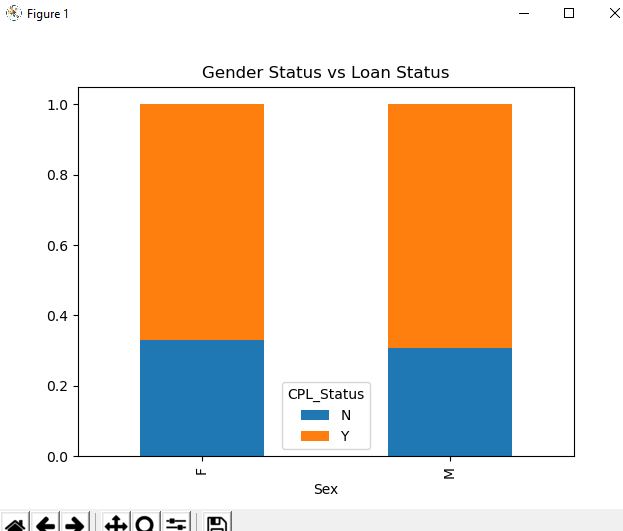
**Split up of Loan\_Status based on Gender**

**CPL\_Status N Y**

**Sex**

**F 0.330357 0.669643**

**M 0.306748 0.693252**

****

**Split up of Loan\_Status based on Married or not**

**CPL\_Status N Y**

**Marital\_Status**

**N 79 134**

**Y 113 285**

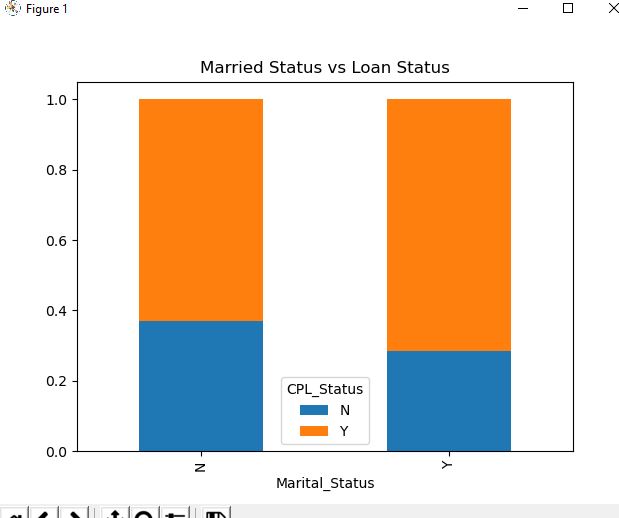
**Split up of CPL\_Status based on Married or not - Probabillity**

**CPL\_Status N Y**

**Marital\_Status**

**N 0.370892 0.629108**

**Y 0.283920 0.716080**

****

**Split up of Loan\_Status based on Education**

**CPL\_Status N Y**

**Qual\_var**

**Grad 140 340**

**Non Grad 52 82**

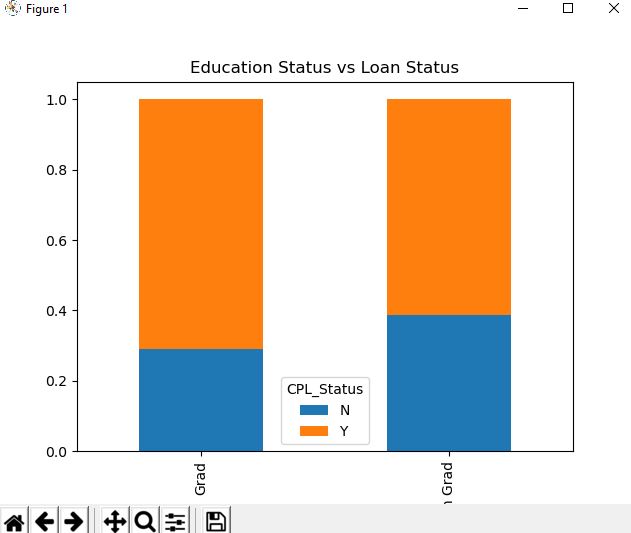
**Split up of CPL\_Status based on Education - Probabillity**

**CPL\_Status N Y**

**Qual\_var**

**Grad 0.291667 0.708333**

**Non Grad 0.388060 0.611940**

****

**Split up of Loan\_Status based on Selfemployed or not**

**CPL\_Status N Y**

**SE**

**N 157 343**

**Y 26 56**

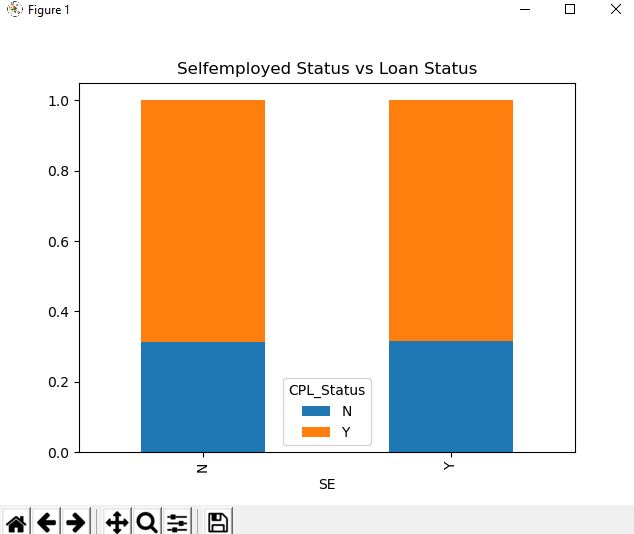
**Split up of Loan\_Status based on Self employed or Not - Probabillity**

**CPL\_Status N Y**

**SE**

**N 0.314000 0.686000**

**Y 0.317073 0.682927**

****

**Split up of Loan\_Status based on number of Dependents**

**CPL\_Status N Y**

**Dependents**

**0 107 238**

**1 36 66**

**2 25 76**

**3+ 18 33**

**Split up of Loan\_Status based on number of Dependents - Probabillity**

**CPL\_Status N Y**

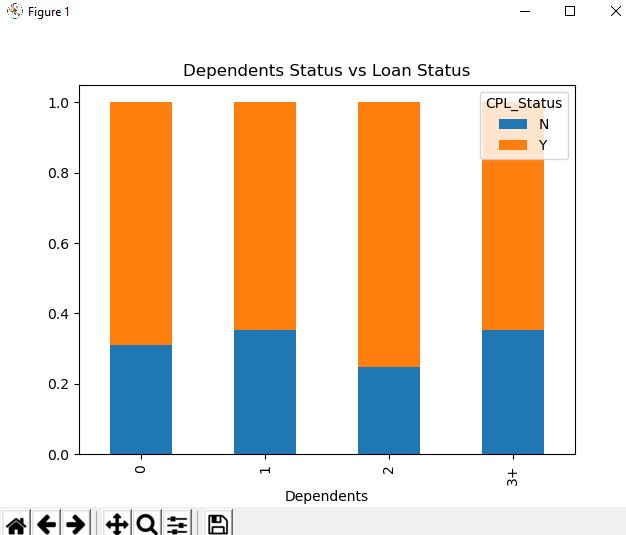
**Dependents**

**0 0.310145 0.689855**

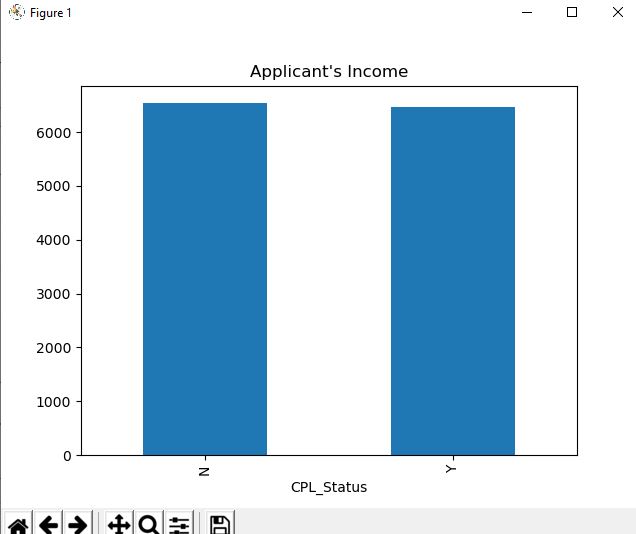
**1 0.352941 0.647059**

**2 0.247525 0.752475**

**3+ 0.352941 0.647059**

****

**Numerical Independent Variable vs Target Variable**

****

**From Appliant Income vs Loan Status, we cannot come to a conclusion**

**So, we create several bins and try comparing the loan status**

**Samples of Income after segregating into different bins**

**0 Very high**

**1 High**

**2 Average**

**3 Average**

**4 Very high**

**5 Very high**

**6 Average**

**7 Average**

**8 High**

**9 Very high**

**Name: Income\_bin, dtype: category**

**Categories (4, object): [Low < Average < High < Very high]**

**Split up of Loan\_Status based on Income**

**CPL\_Status N Y**

**Income\_bin**

**Low 17 33**

**Average 51 135**

**High 65 122**

**Very high 58 132**

**Split up of Loan\_Status based on Income - Probability**

**CPL\_Status N Y**

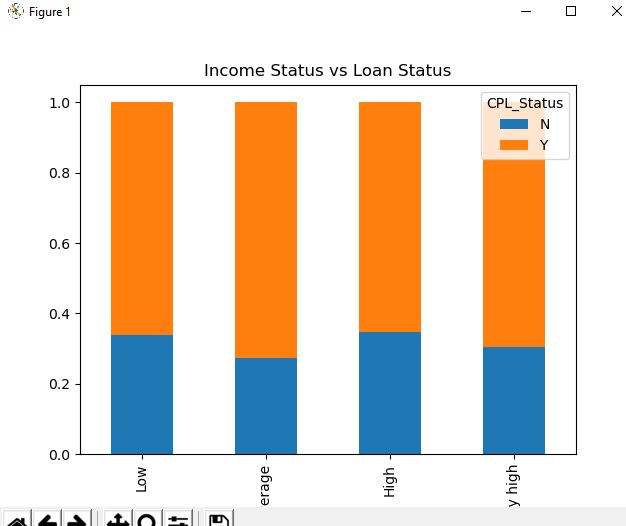
**Income\_bin**

**Low 0.340000 0.660000**

**Average 0.274194 0.725806**

**High 0.347594 0.652406**

**Very high 0.305263 0.694737**

****

**It can be inferred that Applicant income does not affect the chances of loan approval**

**Samples of Co-Applicant Income after segregating into different bins**

**0 NaN**

**1 Average**

**2 NaN**

**3 Average**

**4 NaN**

**5 High**

**6 Average**

**7 High**

**8 Average**

**9 High**

**Name: Coapplicant\_Income\_bin, dtype: category**

**Categories (3, object): [Low < Average < High]**

**Split up of Loan\_Status based on Coapplicant Income**

**CPL\_Status N Y**

**Coapplicant\_Income\_bin**

**Low 2 14**

**Average 51 147**

**High 42 84**

**Split up of Loan\_Status based on Coapplicant Income - Probability**

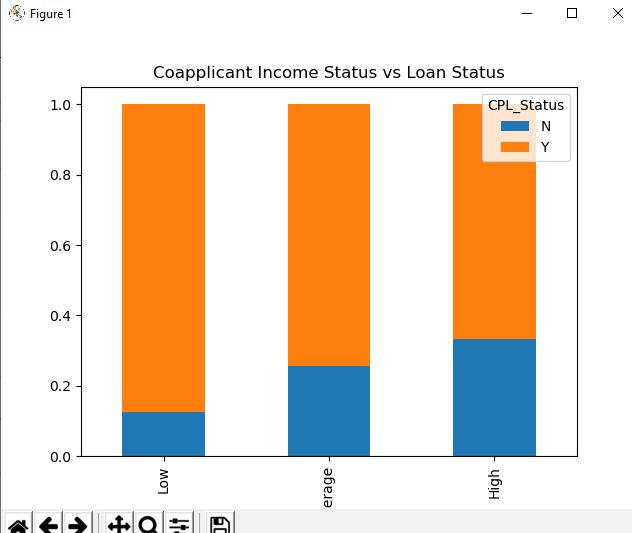
**CPL\_Status N Y**

**Coapplicant\_Income\_bin**

**Low 0.125000 0.875000**

**Average 0.257576 0.742424**

**High 0.333333 0.666667**

****

**It can be inferred that if coapplicant’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 coapplicant.**

**So, the coapplicant income for such applicants is 0 and hence the loan approval is not dependent on it.**

**So, we will combine the applicant’s and coapplicant’s income to visualize the combined effect of income on loan approval**

**(Nan for some records represents that those applicants don’t have any coapplicant).**

**Samples of Total Income (Applicant and Co-Applicant) after segregating into different bins**

**0 Very high**

**1 Very high**

**2 Average**

**3 High**

**4 Very high**

**5 Very high**

**6 High**

**7 Very high**

**8 Very high**

**9 Very high**

**Name: Total\_Income\_bin, dtype: category**

**Categories (4, object): [Low < Average < High < Very high]**

**Split up of Loan\_Status based on Total Income (Applicant and Co-Applicant)**

**CPL\_Status N Y**

**Total\_Income\_bin**

**Low 5 1**

**Average 22 41**

**High 62 139**

**Very high 102 241**

**Split up of Loan\_Status based on Coapplicant Income (Applicant and Co-Applicant) - Probability**

**CPL\_Status N Y**

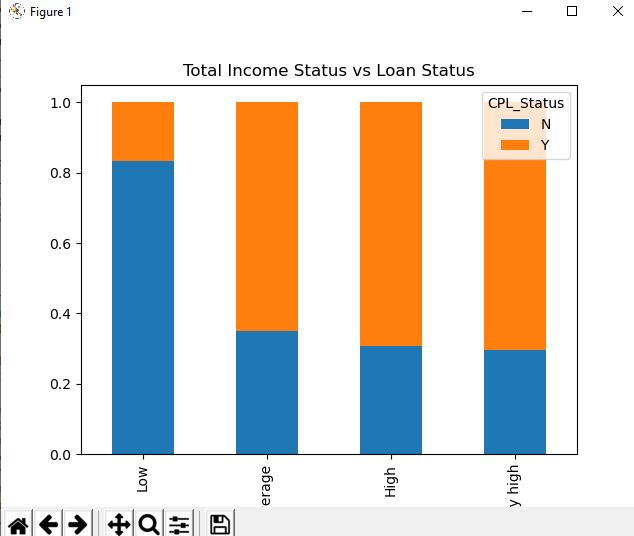
**Total\_Income\_bin**

**Low 0.833333 0.166667**

**Average 0.349206 0.650794**

**High 0.308458 0.691542**

**Very high 0.297376 0.702624**

****

**It can be inferred that proportion of loans getting approved for applicants having low Total Income is the least.**

**Samples of Loan amount after segregating into different bins**

**0 NaN**

**1 Average**

**2 Low**

**3 Average**

**4 Average**

**5 High**

**6 Average**

**7 Average**

**8 High**

**9 High**

**Name: LoanAmount\_bin, dtype: category**

**Categories (3, object): [Low < Average < High]**

**Split up of Loan\_Status based on LoanAmount**

**CPL\_Status N Y**

**LoanAmount\_bin**

**Low 30 58**

**Average 113 257**

**High 47 103**

**Split up of Loan\_Status based on LoanAmount - Probability**

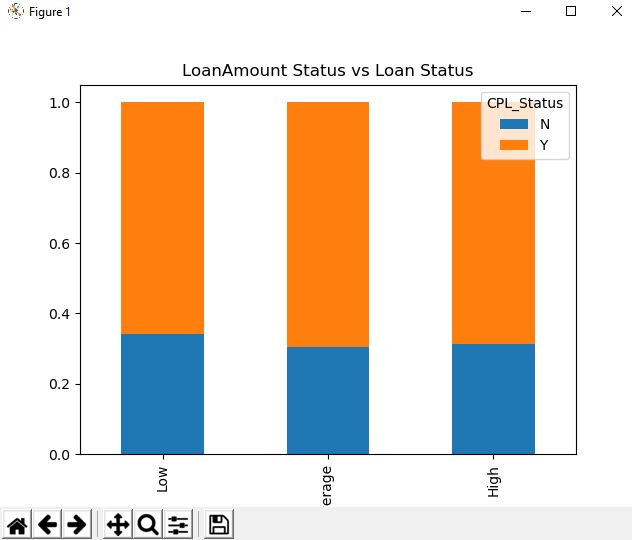
**CPL\_Status N Y**

**LoanAmount\_bin**

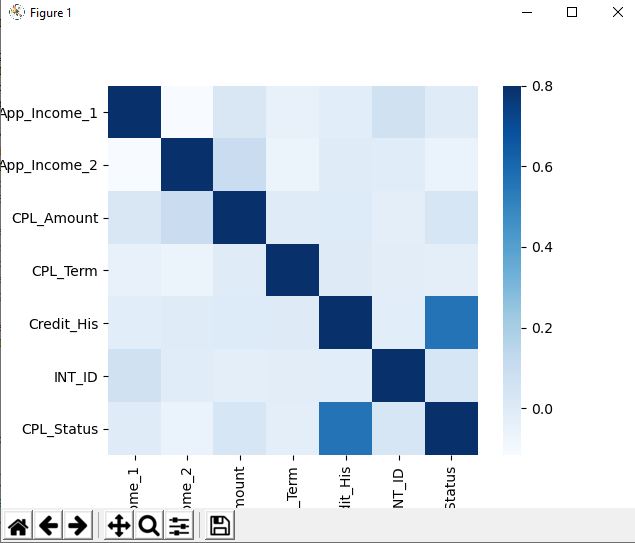
**Low 0.340909 0.659091**

**Average 0.305405 0.694595**

**High 0.313333 0.686667**

****

**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**

****

**We see that the most correlated variables are**

**1. ApplicantIncome - LoanAmount**

**2. Credit\_History - Loan\_Status**

**Number of null values in each field**

**Loanapp\_ID 0**

**Sex 13**

**Marital\_Status 3**

**first\_name 0**

**last\_name 0**

**email 0**

**address 0**

**Dependents 15**

**Qual\_var 0**

**SE 32**

**App\_Income\_1 0**

**App\_Income\_2 0**

**CPL\_Amount 2**

**CPL\_Term 14**

**Credit\_His 50**

**Prop\_Area 0**

**INT\_ID 0**

**Prev\_ID 0**

**AGT\_ID 0**

**CPL\_Status 0**

**dtype: int64**

**Training data after performing One-hot encoding**

**App\_Income\_1 ... AGT\_ID\_Mozilla/6.0 (Windows NT 6.2; WOW64; rv:16.0.1) Gecko/20121011 Firefox/16.0.1**

**0 7018.8 ... 0**

**1 5499.6 ... 0**

**2 3600.0 ... 0**

**3 3099.6 ... 0**

**4 7200.0 ... 0**

**[5 rows x 3421 columns]**

**Dummy columns that are automatically generated for the purpose of One hot encoding**

**Index(['App\_Income\_1', 'App\_Income\_2', 'CPL\_Amount', 'CPL\_Term', 'Credit\_His',**

**'INT\_ID', 'CPL\_Status', 'CPL\_Amount\_log', 'Sex\_F', 'Sex\_M',**

**...**

**'AGT\_ID\_Mozilla/5.0 (iPhone; U; ru; CPU iPhone OS 4\_2\_1 like Mac OS X; ru) AppleWebKit/533.17.9 (KHTML, like Gecko) Version/5.0.2 Mobile/8C148a Safari/6533.18.5',**

**'AGT\_ID\_Mozilla/5.0 (iPod; U; CPU iPhone OS 4\_2\_1 like Mac OS X; he-il) AppleWebKit/533.17.9 (KHTML, like Gecko) Version/5.0.2 Mobile/8C148 Safari/6533.18.5',**

**'AGT\_ID\_Mozilla/5.0 (iPod; U; CPU iPhone OS 4\_3\_1 like Mac OS X; zh-cn) AppleWebKit/533.17.9 (KHTML, like Gecko) Version/5.0.2 Mobile/8G4 Safari/6533.18.5',**

**'AGT\_ID\_Mozilla/5.0 (iPod; U; CPU iPhone OS 4\_3\_3 like Mac OS X; ja-jp) AppleWebKit/533.17.9 (KHTML, like Gecko) Version/5.0.2 Mobile/8J2 Safari/6533.18.5',**

**'AGT\_ID\_Mozilla/5.0 ArchLinux (X11; Linux x86\_64) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/13.0.782.41 Safari/535.1',**

**'AGT\_ID\_Mozilla/5.0 ArchLinux (X11; U; Linux x86\_64; en-US) AppleWebKit/534.30 (KHTML, like Gecko) Chrome/12.0.742.100',**

**'AGT\_ID\_Mozilla/5.0 ArchLinux (X11; U; Linux x86\_64; en-US) AppleWebKit/534.30 (KHTML, like Gecko) Chrome/12.0.742.60 Safari/534.30',**

**'AGT\_ID\_Mozilla/5.0 Slackware/13.37 (X11; U; Linux x86\_64; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/12.0.742.91',**

**'AGT\_ID\_Mozilla/6.0 (Macintosh; I; Intel Mac OS X 11\_7\_9; de-LI; rv:1.9b4) Gecko/2012010317 Firefox/10.0a4',**

**'AGT\_ID\_Mozilla/6.0 (Windows NT 6.2; WOW64; rv:16.0.1) Gecko/20121011 Firefox/16.0.1'],**

**dtype='object', length=3421)**

**Accuracy of the Logistic Regression model built is**

**0.6702702702702703**

**Logistic Regression - Stratified k-folds cross Validation**

**Iteration 1 of kfold 5**

**accuracy\_score 0.6910569105691057**

**Iteration 2 of kfold 5**

**accuracy\_score 0.6910569105691057**

**Iteration 3 of kfold 5**

**accuracy\_score 0.6829268292682927**

**Iteration 4 of kfold 5**

**accuracy\_score 0.6829268292682927**

**Iteration 5 of kfold 5**

**accuracy\_score 0.6885245901639344**

**Mean validation accuracy for Logistic Regression - Stratified k-folds cross Validation model is 0.6872984139677463**

**Decision Tree - Stratified k-folds cross Validation**

**Iteration 1 of kfold 5**

**accuracy\_score 0.7479674796747967**

**Iteration 2 of kfold 5**

**accuracy\_score 0.7317073170731707**

**Iteration 3 of kfold 5**

**accuracy\_score 0.7398373983739838**

**Iteration 4 of kfold 5**

**accuracy\_score 0.7154471544715447**

**Iteration 5 of kfold 5**

**accuracy\_score 0.7377049180327869**

**Mean validation accuracy for Decision Tree - Stratified k-folds cross Validation model is 0.7345328535252565**

**Random Forest - Stratified k-folds cross Validation**

**Iteration 1 of kfold 5**

**accuracy\_score 0.6910569105691057**

**Iteration 2 of kfold 5**

**accuracy\_score 0.6910569105691057**

**Iteration 3 of kfold 5**

**accuracy\_score 0.6829268292682927**

**Iteration 4 of kfold 5**

**accuracy\_score 0.6829268292682927**

**Iteration 5 of kfold 5**

**accuracy\_score 0.6885245901639344**

**Mean validation accuracy for Random Forest - Stratified k-folds cross Validation model is 0.6872984139677463**

**SVM - Stratified k-folds cross Validation**

**Iteration 1 of kfold 5**

**accuracy\_score 0.6910569105691057**

**Iteration 2 of kfold 5**

**accuracy\_score 0.6910569105691057**

**Iteration 3 of kfold 5**

**accuracy\_score 0.6829268292682927**

**Iteration 4 of kfold 5**

**accuracy\_score 0.6829268292682927**

**Iteration 5 of kfold 5**

**accuracy\_score 0.6885245901639344**

**Mean validation accuracy for SVM - Stratified k-folds cross Validation model is 0.6872984139677463**

**KNN - Stratified k-folds cross Validation**

**Iteration 1 of kfold 5**

**accuracy\_score 0.5691056910569106**

**Iteration 2 of kfold 5**

**accuracy\_score 0.6097560975609756**

**Iteration 3 of kfold 5**

**accuracy\_score 0.5934959349593496**

**Iteration 4 of kfold 5**

**accuracy\_score 0.5772357723577236**

**Iteration 5 of kfold 5**

**accuracy\_score 0.6147540983606558**

**Mean validation accuracy for KNN - Stratified k-folds cross Validation model is 0.592869518859123**

**Mean validation accuracies of algorithms are listed below**

**Mean validation accuracies of four algorithms are listed below**

**Logistic Regression : 0.6872984139677463**

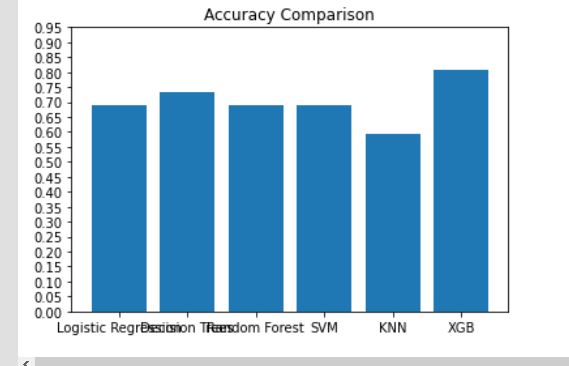
**Decision Trees : 0.7345328535252565**

**Random Forest : 0.6872984139677463**

**SVM0.6872984139677463**

**KNN0.592869518859123**

**XGB0.8094228975076636**

****