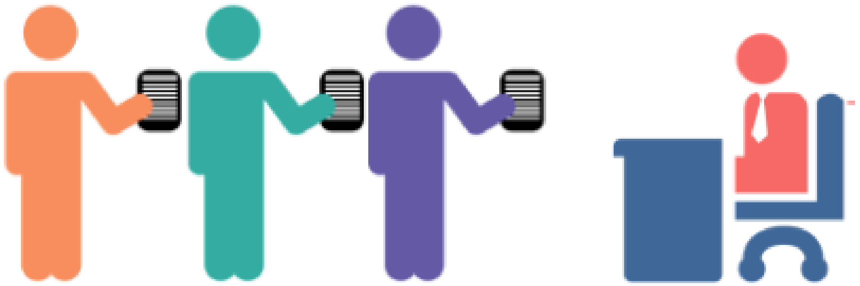
Lending Club Case Study

Exploratory Data Analysis

Submitted by : Vipul Acharya

Problem Statement

- The Company is a Consumer finance company that lends various type of loans to urban customers
- Whenever a new loan application comes to the bank, it is crucial for the bank to decide whether or not to approve the loan
- There are two types of risks are associated with the bank's decision:
    - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
    - If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company
- Hence it is important for the banks to take the right decision and approve the right loans
- In this case study, objective is to understand which **consumer attributes and loan attributes** influence the tendency of defaulting on a loan
- This will help the banks in making quicker, smarter and less risky decisions while approving the loan



Analysis Approach

- ***Data Understanding***
  - Checking the number of records
  - Checking the number of columns
  - Understanding the datatypes of the different columns in dataset
  - Checking if there are any duplicate entries
  - Checking for columns with only one value

- ***Data Cleaning***
  - Checking missing values in the dataset
  - Treating the missing values, if possible, else removing the columns having most of the records as missing values
  - Formatting the data wherever required
  - Deriving new columns for analysis if required

- ***Data Analysis and Visualization***
  - Univariate Analysis
  - Bivariate Analysis

- ***Conclusion***

Data Understanding

- Loan Dataset:

  No of rows : 39717

  No of columns : 111

- Understanding datatype of columns:

  <u>Sample shown as below:</u>

```
 #    Column                          Non-Null Count  Dtype
---   ------                          --------------  -----
 0    id                              39717 non-null  int64
 1    member_id                       39717 non-null  int64
 2    loan_amnt                       39717 non-null  int64
 3    funded_amnt                     39717 non-null  int64
 4    funded_amnt_inv                 39717 non-null  float64
 5    term                            39717 non-null  object
 6    int_rate                        39717 non-null  object
 7    installment                     39717 non-null  float64
 8    grade                           39717 non-null  object
 9    sub_grade                       39717 non-null  object
 10   emp_title                       37258 non-null  object
 11   emp_length                      38642 non-null  object
 12   home_ownership                  39717 non-null  object
```

<u>Identifying Categorical and Numerical data from data type:</u>

```
categorical_cols
```

```
Index(['term', 'int_rate', 'grade', 'sub_grade', 'emp_title', 'emp_length',
       'home_ownership', 'verification_status', 'issue_d', 'loan_status',
       'pymnt_plan', 'url', 'desc', 'purpose', 'title', 'zip_code',
       'addr_state', 'earliest_cr_line', 'revol_util', 'initial_list_status',
       'last_pymnt_d', 'last_credit_pull_d', 'application_type',
       'pub_rec_bankruptcies'],
      dtype='object')
```

```
numerical_cols
```

```
Index(['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv',
       'installment', 'annual_inc', 'dti', 'delinq_2yrs', 'inq_last_6mths',
       'open_acc', 'pub_rec', 'revol_bal', 'total_acc', 'out_prncp',
       'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp',
       'total_rec_int', 'total_rec_late_fee', 'recoveries',
       'collection_recovery_fee', 'last_pymnt_amnt',
       'collections_12_mths_ex_med', 'policy_code', 'acc_now_delinq',
       'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens'],
      dtype='object')
```

- Checked if we have duplicate records that is multiple records with same load id. We do not have such case

Data Understanding (Continued)

- Removing the below columns which have only single value:

```
Index(['pymnt_plan', 'initial_list_status', 'collections_12_mths_ex_med',
       'policy_code', 'application_type', 'acc_now_delinq',
       'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens'],
      dtype='object')
```

- Since the objective is to find the factors which result in default. Therefore, the fields that are created after a loan application is approved might not be very helpful.

```
## Removing columns that are created post loan approval since it wont be useful for our analysis
cols = ['delinq_2yrs','revol_bal', 'out_prncp','total_pymnt','total_rec_prncp','total_rec_int','total_rec_late_fee'
,'collection_recovery_fee', 'last_pymnt_d','last_pymnt_amnt']
loan_df.drop(columns = cols, inplace=True)
```

- Fields like id, member_id & url will not help us determine if the person will default
- Therefore, removing these columns which are not useful for our analysis
- Since analysis is to understand who will default, considering only those records which are either fully paid or charged off. Removing the rest of records

Data Cleaning

- Checking Missing Values:
  - We have zero records which do not have any information in all the columns
  - We have 54 columns which is 100% empty. These columns would not be useful in any way
  - As a standard, we are removing any column which has more than 50% of information missing as it would not help us in any way

  <u>Sample Snap with columns more than 50% values missing:</u>

```
missing_values[missing_values['missing %'] >= 50]['column name']
```

```
28             mths_since_last_delinq
29             mths_since_last_record
47                       next_pymnt_d
50        mths_since_last_major_derog
53                   annual_inc_joint
54                          dti_joint
55          verification_status_joint
57                       tot_coll_amt
58                        tot_cur_bal
59                        open_acc_6m
60                         open_il_6m
61                        open_il_12m
62                        open_il_24m
63                 mths_since_rcnt_il
64                       total_bal_il
65                            il_util
66                        open_rv_12m
```

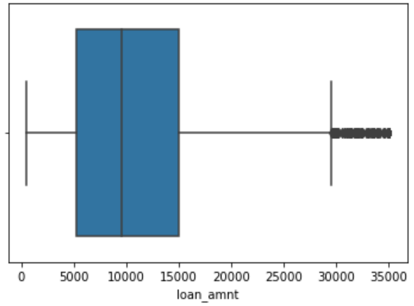
- Standardizing (Formatting) columns/ treating missing values:
  - Some columns like interest rate, employment length etc. can be converted as numerical variables after some data cleaning
  - If the missing value percentage is very less and if it is numerical variable, we can impute the missing value with mode/median/0 whichever looks correct
  - And incase of categorical variable, we can impute it as not available if the % missing values is less
  - If the columns are identified as objects but the values indicate numerical, convert those variables to numerical datatype

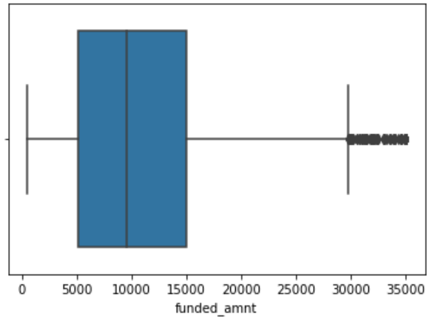
- Deriving year and month from date column for further analysis

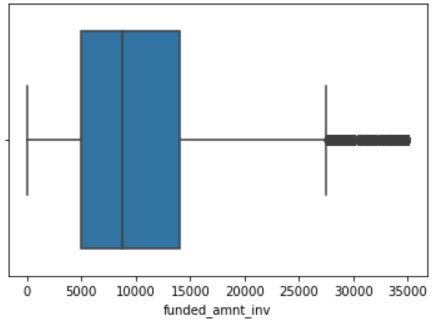
- Dividing the data in columns Loan Amount, Annual Income, Interest rate, dti etc. in to bins to analyze better

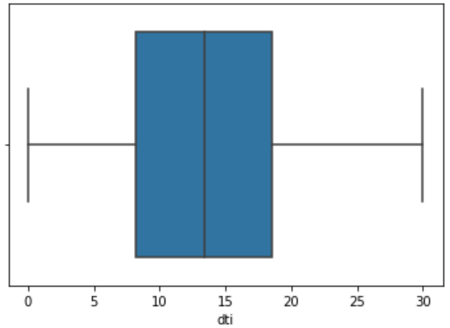
Data Analysis and Visualization

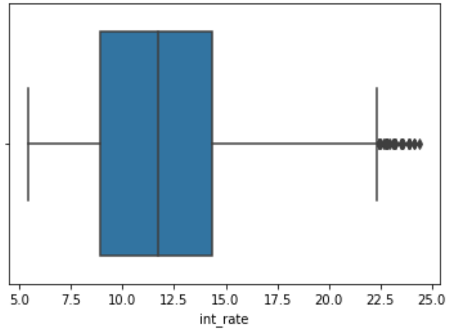
<u>Outlier Treatment</u>



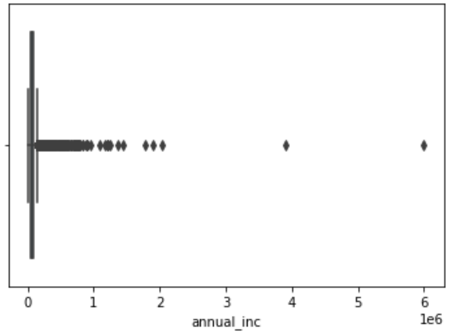




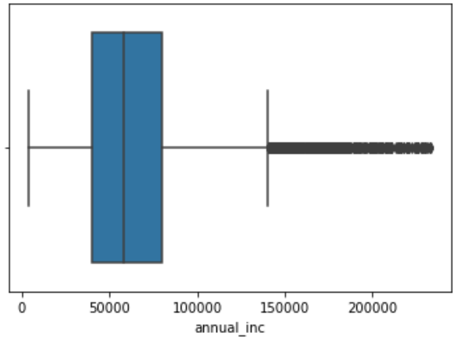




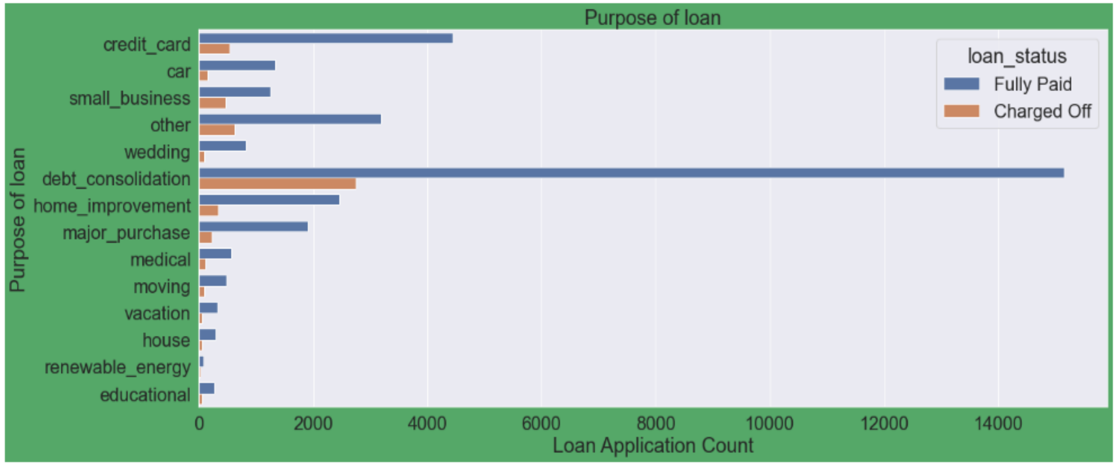
Though there are some outliers in the above columns, distribution is continuous and hence outlier treatment is not reqyured



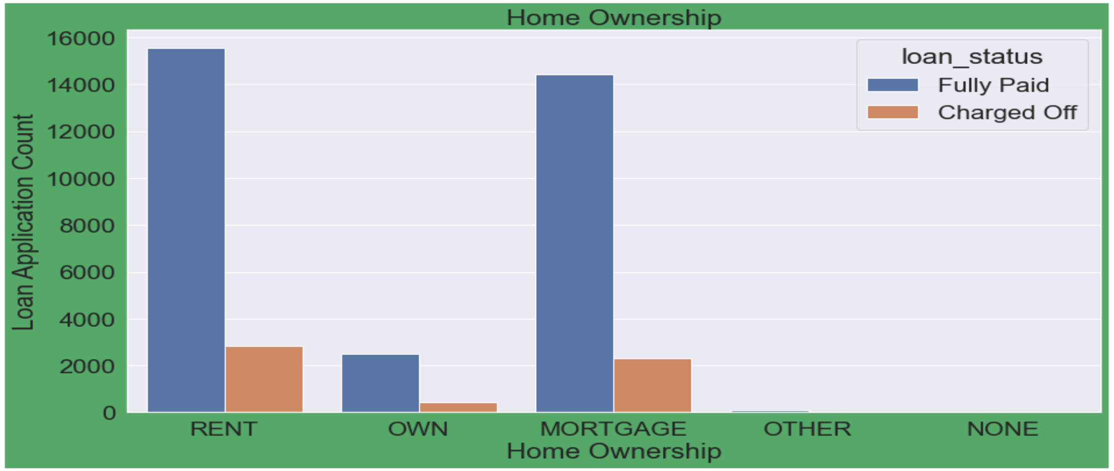
Remove Outliers quantile .99
from Annual Income



Univariate Analysis

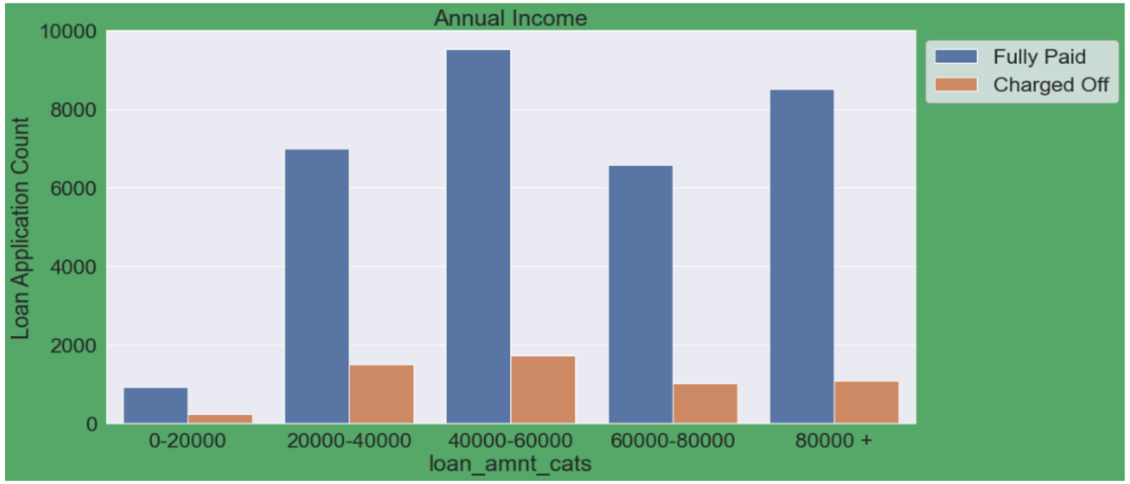


- We can see that most of the purpose is of debt consolidation & paying credit card bill.
- Number of charged off also are too high for these loans

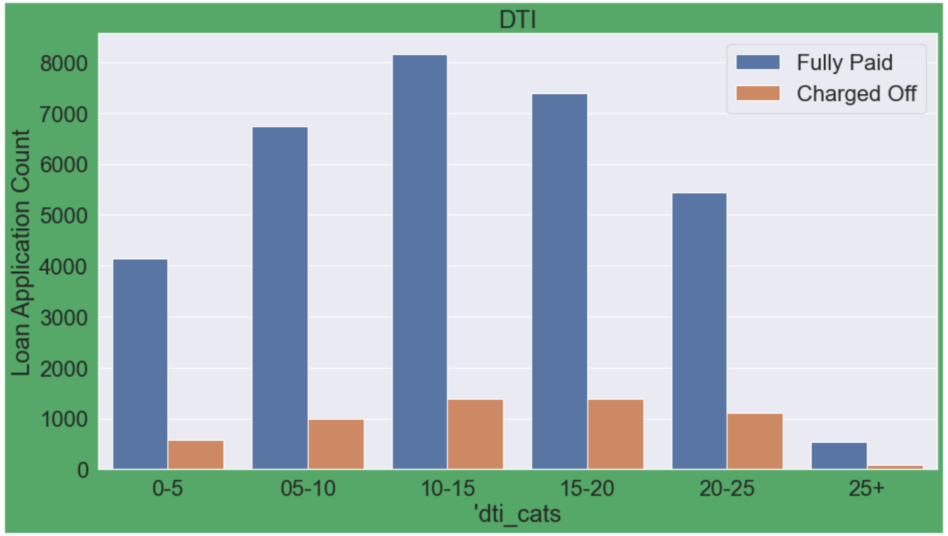


- we can see that most of them are living in rent & mortgage house

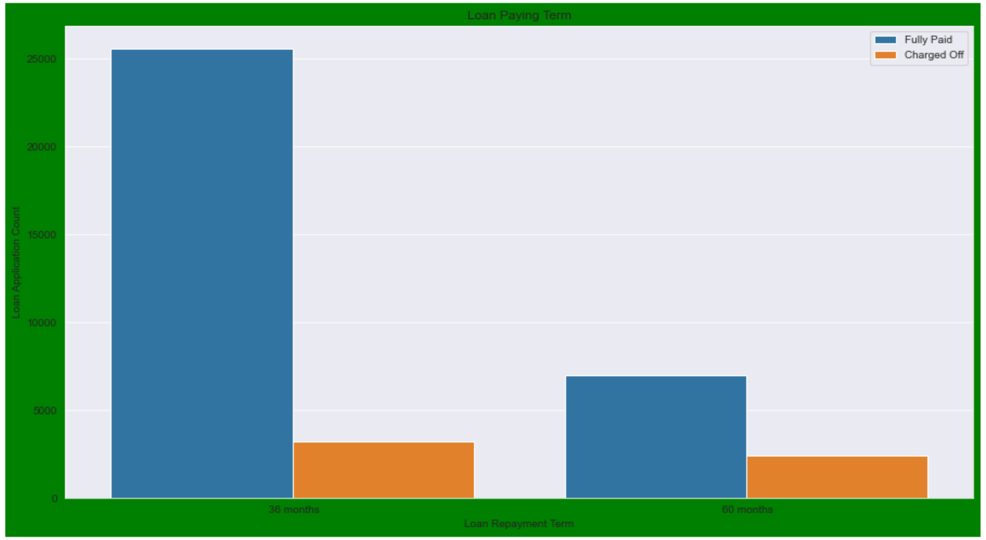
- Higher the rate of interest, higher the chance of default



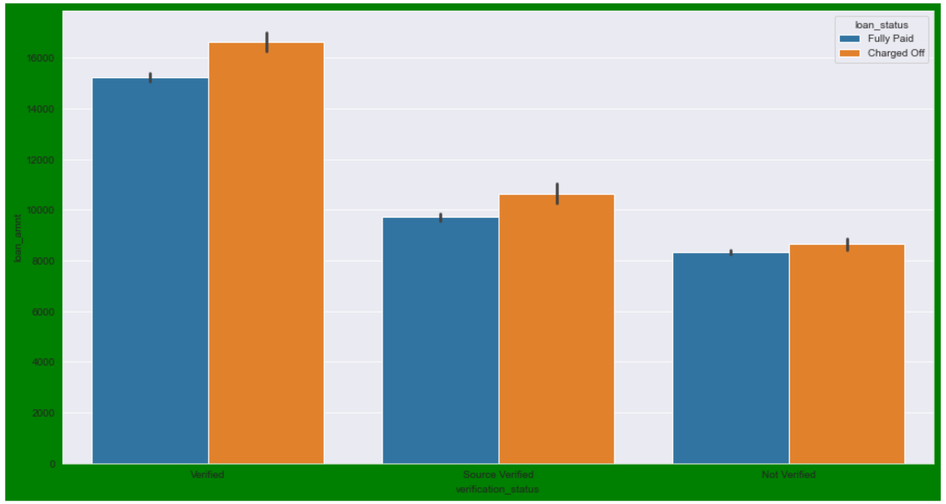
- Higher the income, less chances of default



- high DTI leads to high % of charge off

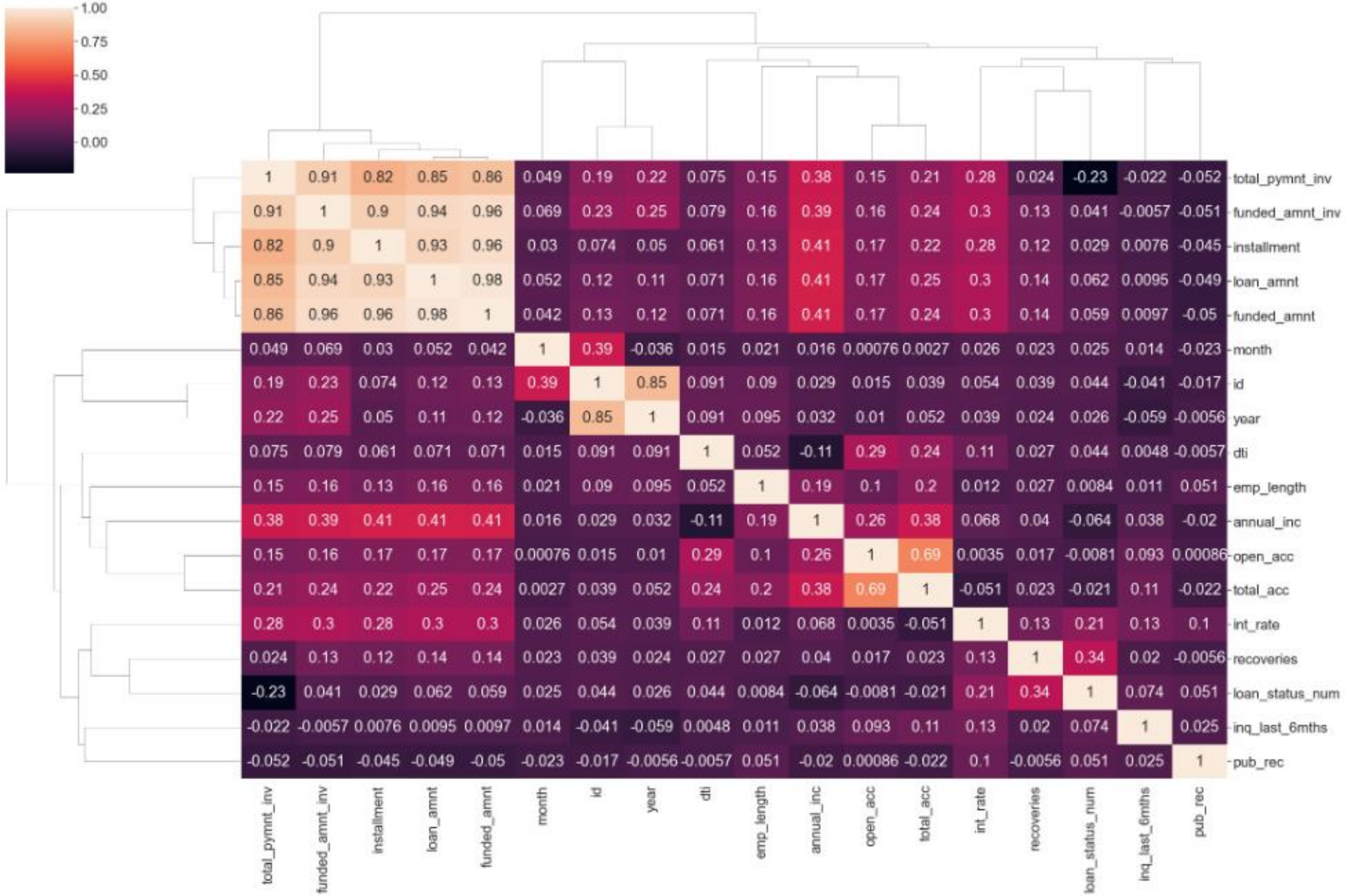


- - For loans with 5 year repayment term, the default percent is 25%.
- - for 3 year loan repayment term, the default is only for 11% of the cases.
- - Therefore, loan repayment is an important factor



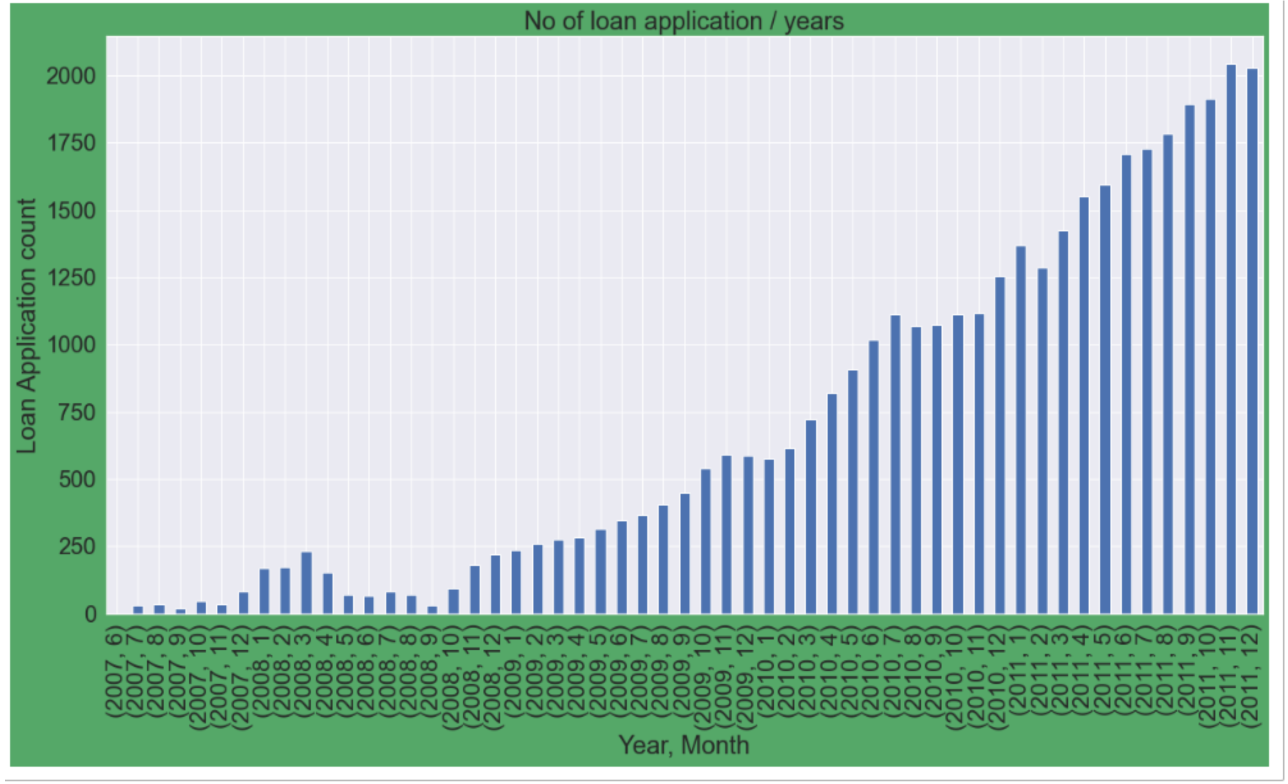
- Verified loans have higher loan amounts.

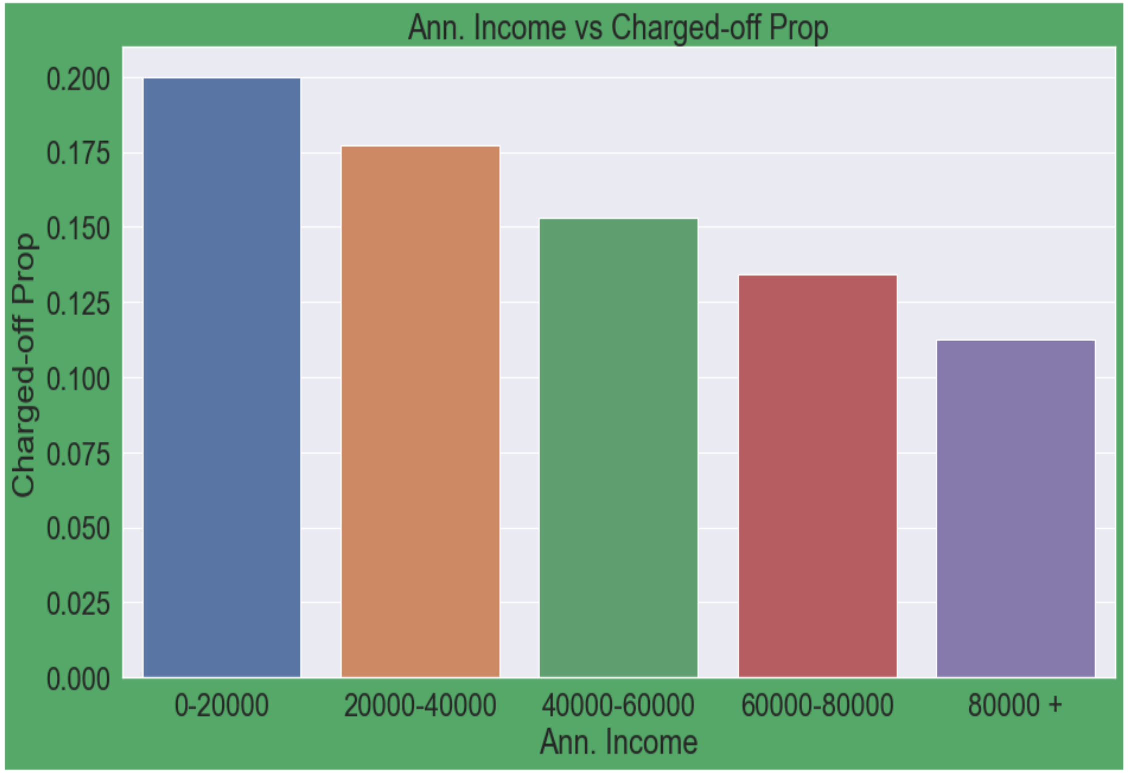
Bivariate Analysis



<u>Understanding strength relationships among different variables:</u>

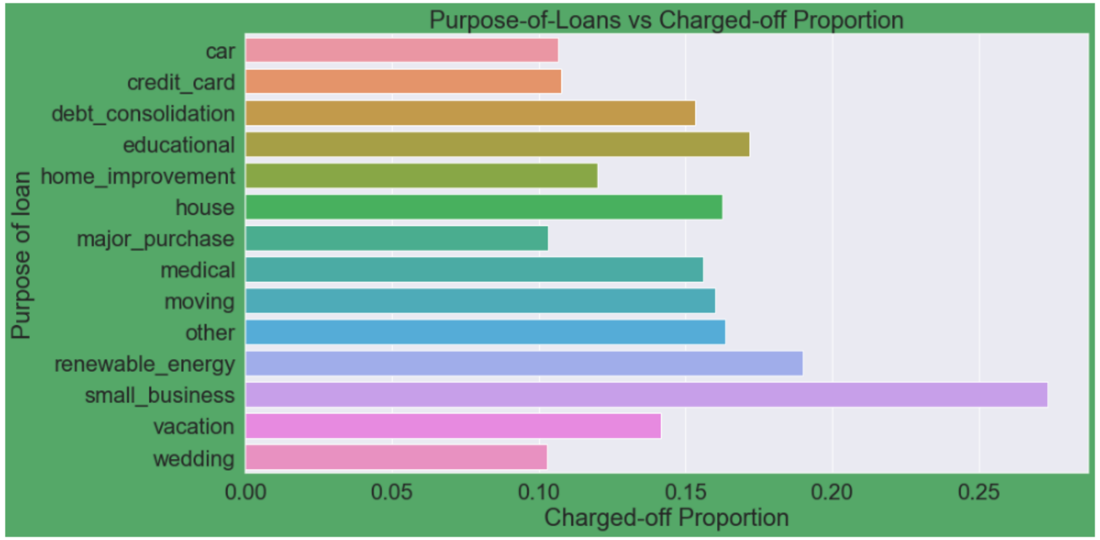
- Total_payment_inv, funded_amnt_inv, installment, loan_amnt, funded amnt show high level of correlation with each other



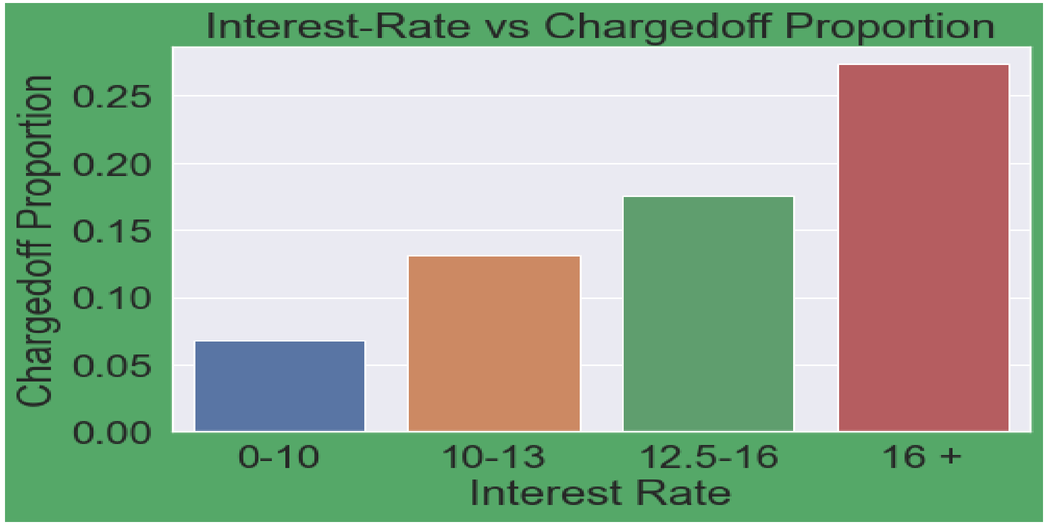


- Looking at the above graph, we can see that loan issued are increasing every passing years
- There is a dip in loan application rate in year 2008, may be due to recession

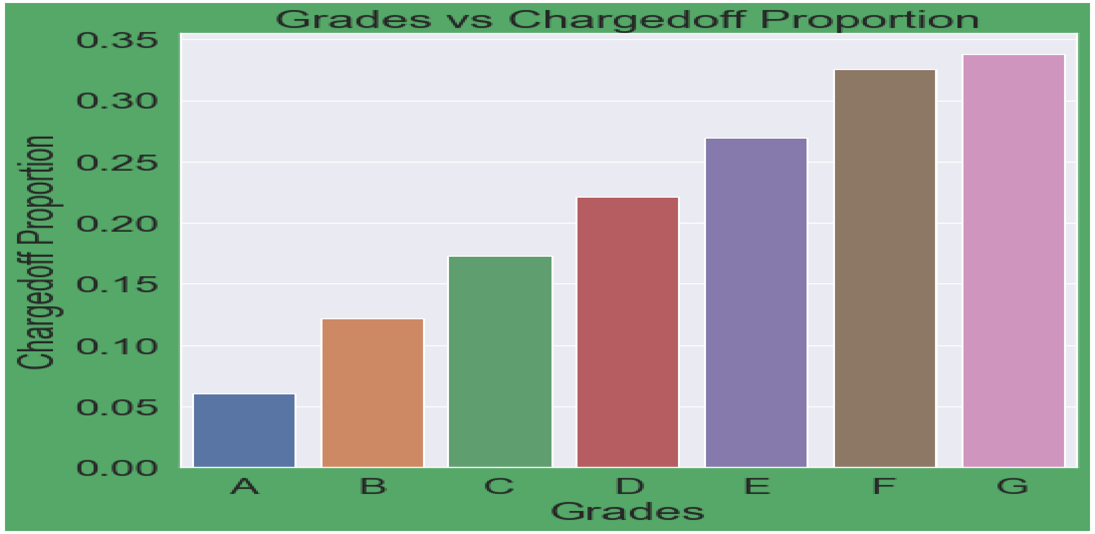
- Annual Income range of 0-20000 has a high chance of charged off.
- one more observation as annual income increases charged off proportion decreases



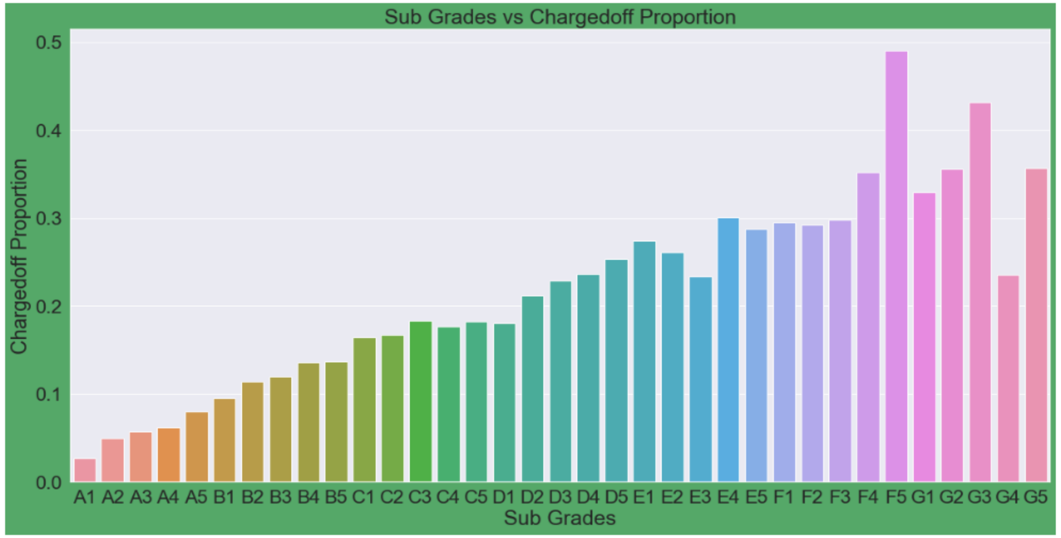
small business has highest charged off proportion



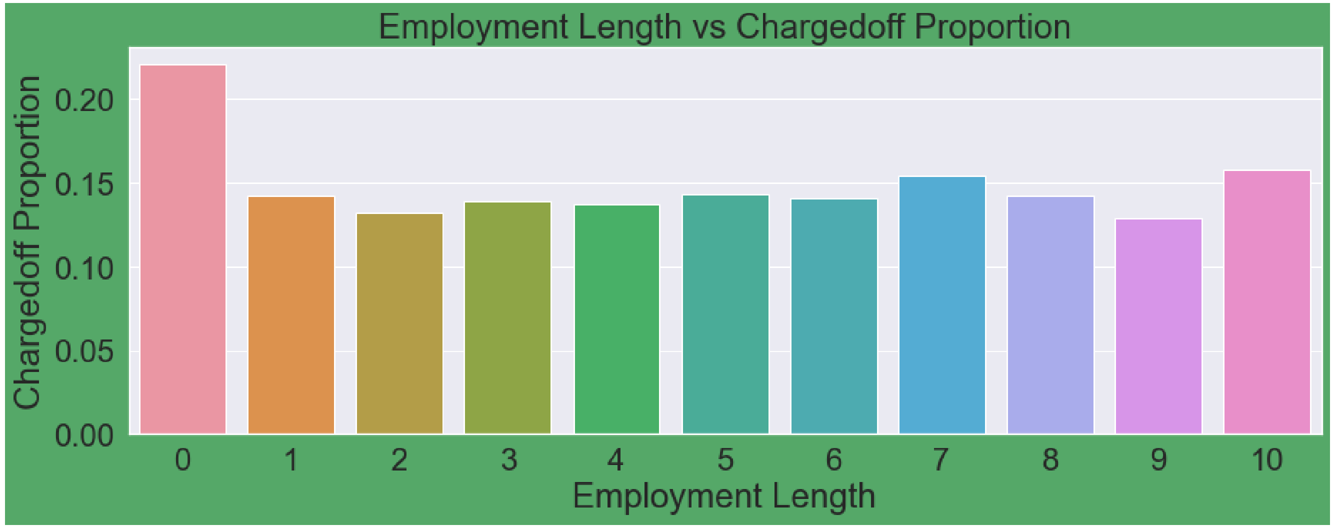
- The rate of interest less than 10% has very less chance of charged off



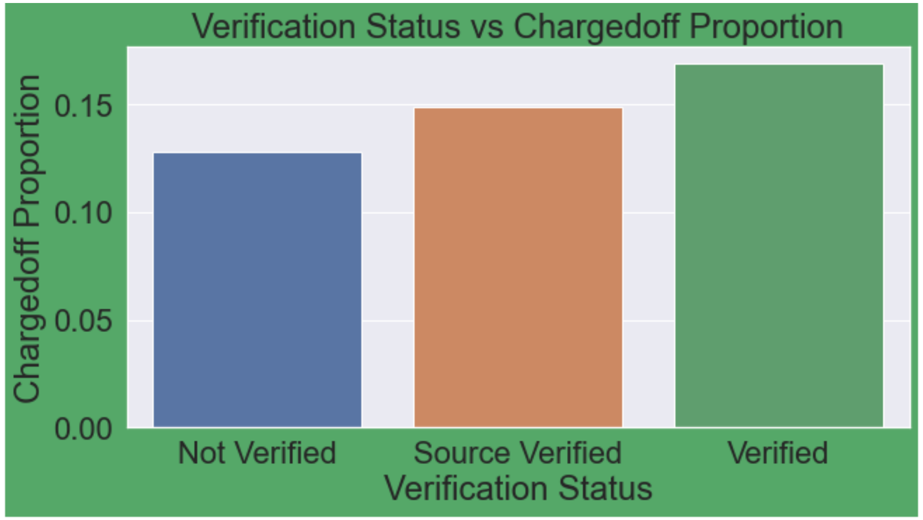
The chance of increase of charge off is toward A-->G



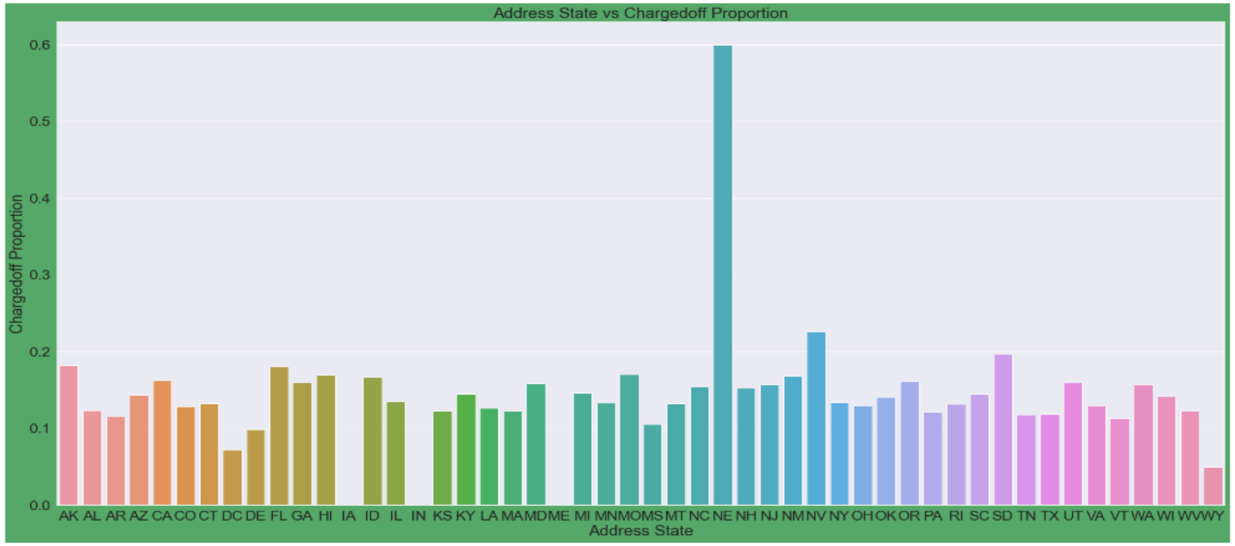
- sub grades of A has very less chances of charged off.
- sub grades of F & G are having very high chances of charged off



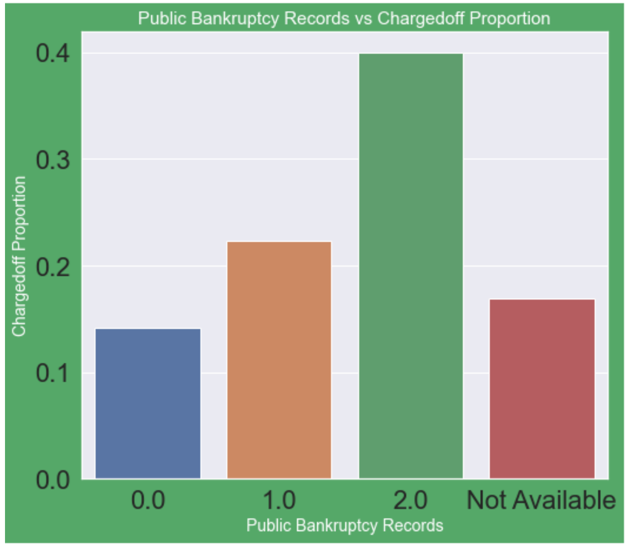
High range of chargedoff in the employees who has 0 level experience to less than 1 year experience



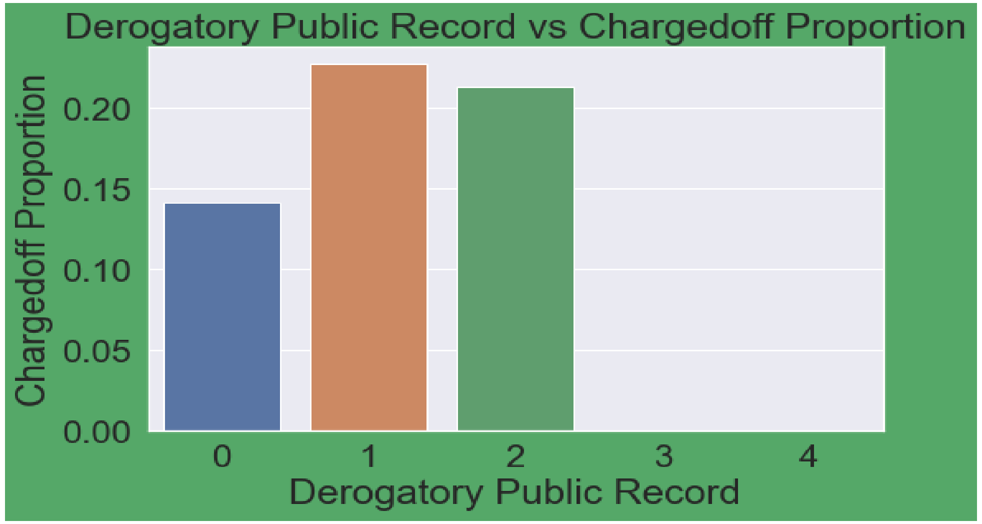
The charged off proportion starts from 0.12. There is not much difference in it but the highest charged off is in verified

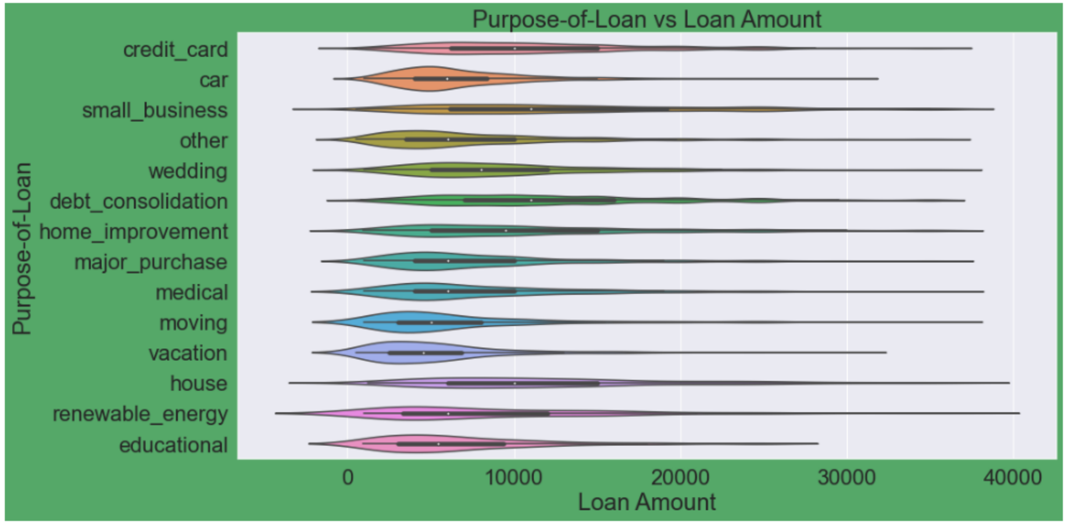


NE has a high range of charged off proportion



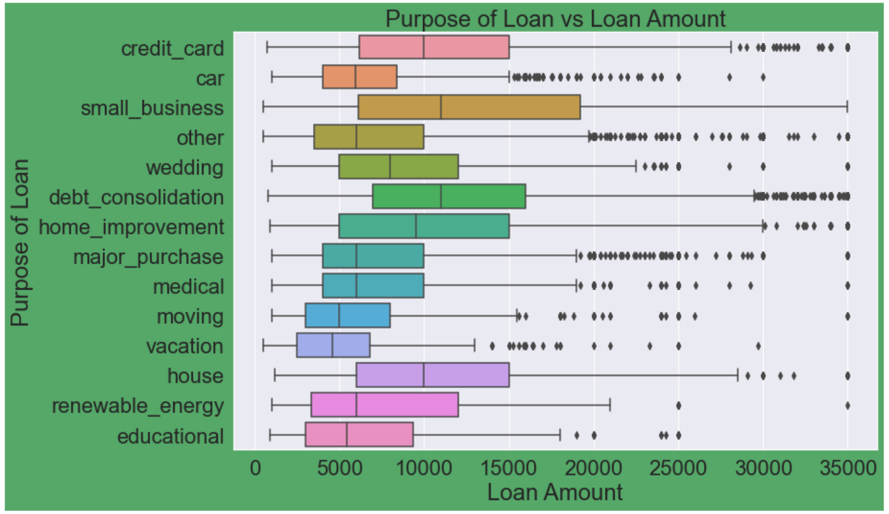
- Higher bankruptcy records, higher the chance of default
- In the NA section, we don't have much information about the users

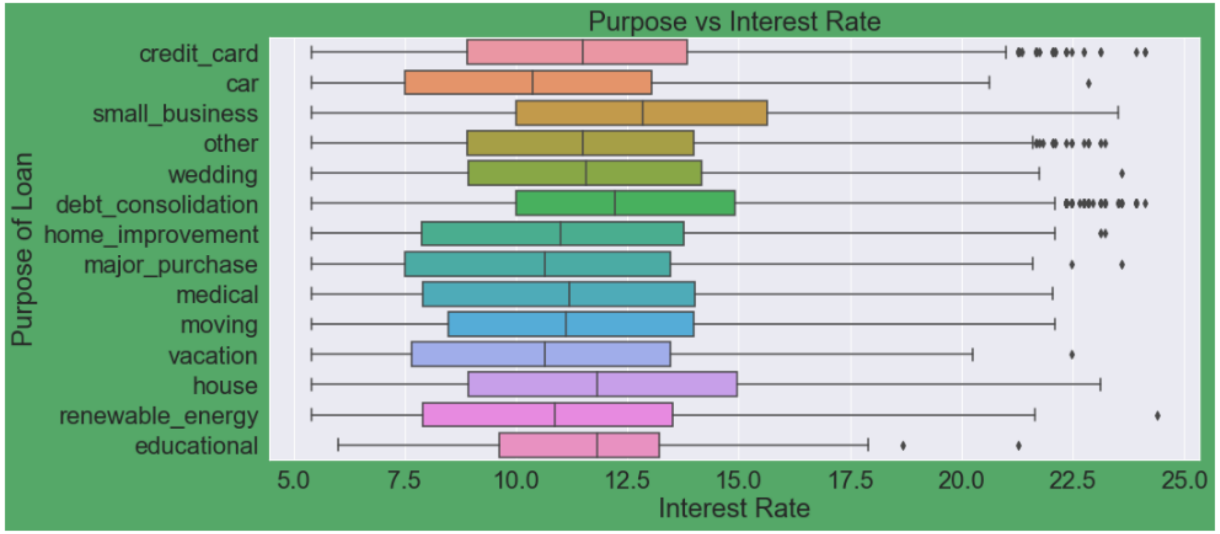




- Derogatory item is an entry that may be considered negative by lenders because it indicates risk and hurts
- Pub_rec count for 3 & 4 has less number
- Pub_rec count for 1 & 2 has higher number of charged off

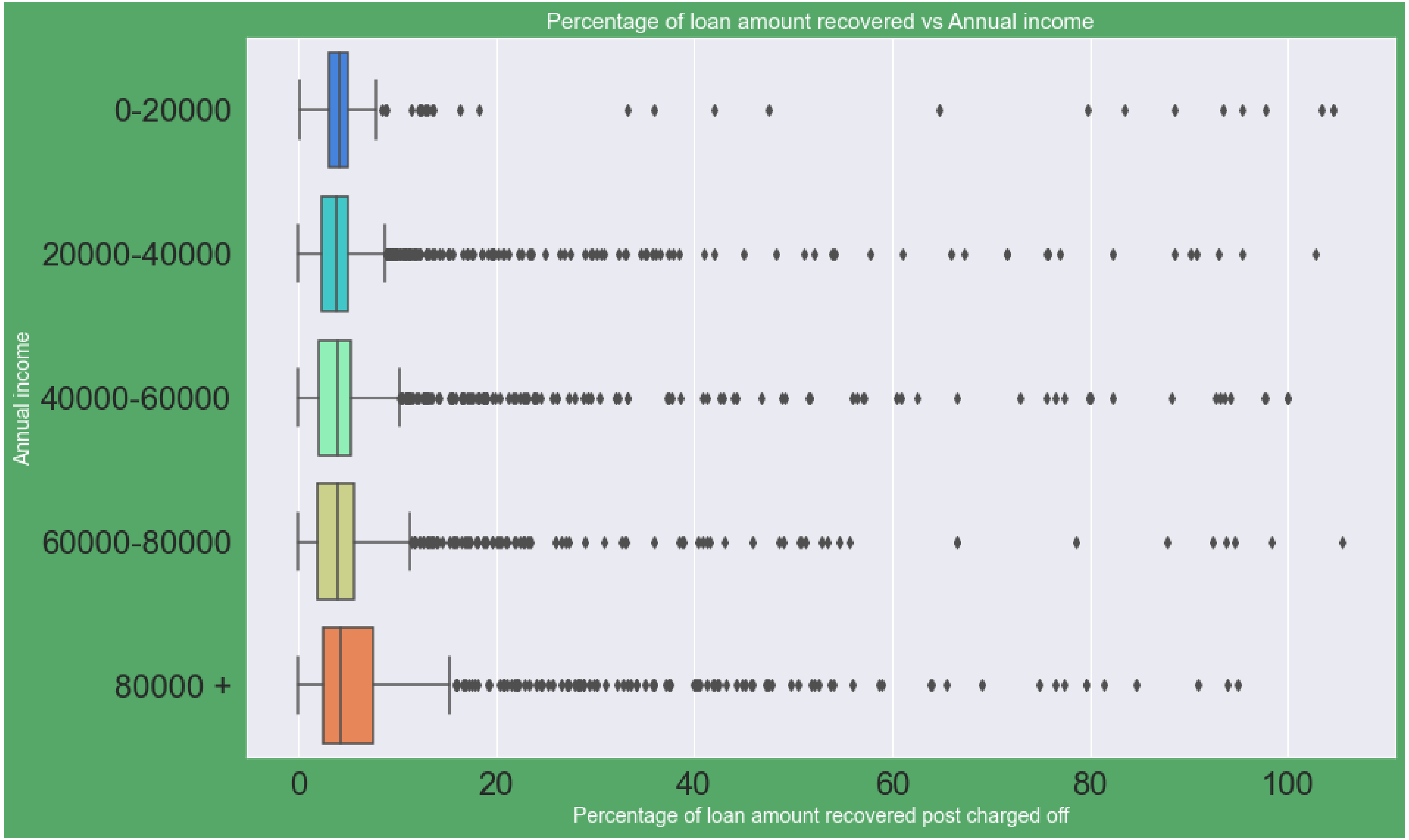
For small business, the loan, debt consolidations and credit card are somewhat evenly distributed





small_business have a higher median , 75th percentile of loan amount compared to others

The average interest rate for small business are high compared to other



The percentage of recovery totally dependent of annual income. The annual income of 80000+ has higher percentage of recovery compared to the others

Conclusion

The detailed data exploration has helped in understanding that there are a few factors which can help us in knowing whether the person will default or not. This can help the bank in knowing the right opportunities and drop the ones that are risky.

Following are few of the factors identified from the analysis that can help identify the people who have a higher chance of default:

Consumer Attributes

- Person living in Rented place or place that has been mortgaged have a higher chance of default as compared to someone who is staying in their own place
- Higher Income, less chances of defaulting. One additional point is that income range less than 20000 has the highest defaults as compared to other income groups
- Annual income greater than 80000 have a higher chance of loan recovery
- High Debt to Income Ratio leads to higher defaults
- If the employee experience is between 0 and 1, high chance of defaults
- State NE has a higher number of defaults as compared to other states
- Higher bankruptcy records, greater the chances of default
- Derogatory public records in the range from 0 to 3 have a high chance of default

Loan Attributes

- Purpose stated for the loan is ‘Debt Consolidation’, ‘Credit card’ or ‘Others’. Small business has the highest charged off percent
- Rate of interest in the range of 12.5 to 16% has higher defaults. ROI less than 10% has the lowest number of defaults
- Loans with higher terms have a higher default percent
- Defaults increases as the grade goes from A to G. Subgrade wise too this holds true

Whenever a new loan application comes in, the bank can check if there are any of the red flags from the above points and decide on loan approval.