Upgrad Lending Club Case Study

- By Abhishek Mukherjee and Abhishek Shukla

September 2024

Contents

- ❑ Project Background
- ❑ Data Collection, Exploration and Cleaning
- ❑ Columns/Variables filtering for EDA
- ❑ Univariate Analysis
- ❑ Bivariate/Multivariate Analysis (including Derived Metrics)
- ❑ Conclusion

Project Background

- This is an Upgrad case study project aimed to do EDA for a given dataset and report analysis results.
- We are given a dataset about a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. 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
- The data provided to us (loan.csv) contains information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate etc.
- In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of "Loan Default".

Data Collection, Exploration and Cleaning .. (1/2)

- The data was read using Pandas function “pd.read_csv()” and has the following inherent characteristics:

Characteristic	Information
Number of observations	39,717
Number of columns/variables	111
Number of duplicate records	0

- We computed Null values and Proportions. There were certain columns with 100% null values. We dropped them and took a decision to drop columns/variables with 30% or more null values as imputing them would introduce significant bias in the dataset. Remaining null percentages are depicted below:

```
Missing data percent greater than 0:

emp_title                    6.19
emp_length                   2.71
title                        0.03
revol_util                   0.13
last_pymnt_d                 0.18
last_credit_pull_d           0.01
collections_12_mths_ex_med   0.14
chargeoff_within_12_mths     0.14
pub_rec_bankruptcies         1.75
tax_liens                    0.10
dtype: float64
```

Data Collection, Exploration and Cleaning .. (2/2)

- Missing Value Treatment
  - 'emp_title' and Loan 'title' had a lot of variation, so we dropped it. There was no way we could impute it with 'mode()'

```
# Lets check the 'emp_title' and 'emp_length' columns
loan_df_clean['emp_title'].value_counts()
```

```
US Army                 134
Bank of America         109
IBM                      66
AT&T                     59
Kaiser Permanente        56
                       ...
Cedar Engineering         1
Tourico Holidays          1
411 Signs & Graphics      1
Cree  Inc                 1
Metro Beverage            1
Name: emp_title, Length: 28820, dtype: int64
```

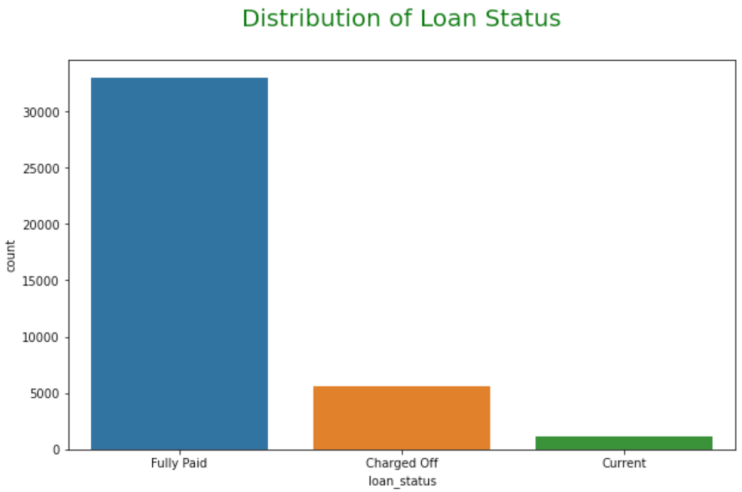
```
# Lets check the Loan Title column to see types of loans
loan_df_clean['title'].value_counts()
```

```
Debt Consolidation                2184
Debt Consolidation Loan           1729
Personal Loan                      659
Consolidation                      517
debt consolidation                 505
                                  ...
Farmer                               1
Simple consolidation of 3 cards      1
cat's debt consolidation             1
$5,000 personal loan                 1
Deb consilidation                    1
Name: title, Length: 19615, dtype: int64
```

  - 'emp_length' had object/string data type and contained more or less equal distribution of data across except '10+ years'. We used 'mode()' function to impute 2.7% of the missing data in this column
  - For numerical columns we imputed them with "median()" function as the missing value percentage was less than 2%
  - For categorical columns we imputed them with "mode()" function

Columns/Variables filtering for EDA

- For performing EDA, we needed to answer the question of identifying patterns for “Loan default”. So, we dropped rows/observations where “loan_status” was “Current” representing ongoing loans.

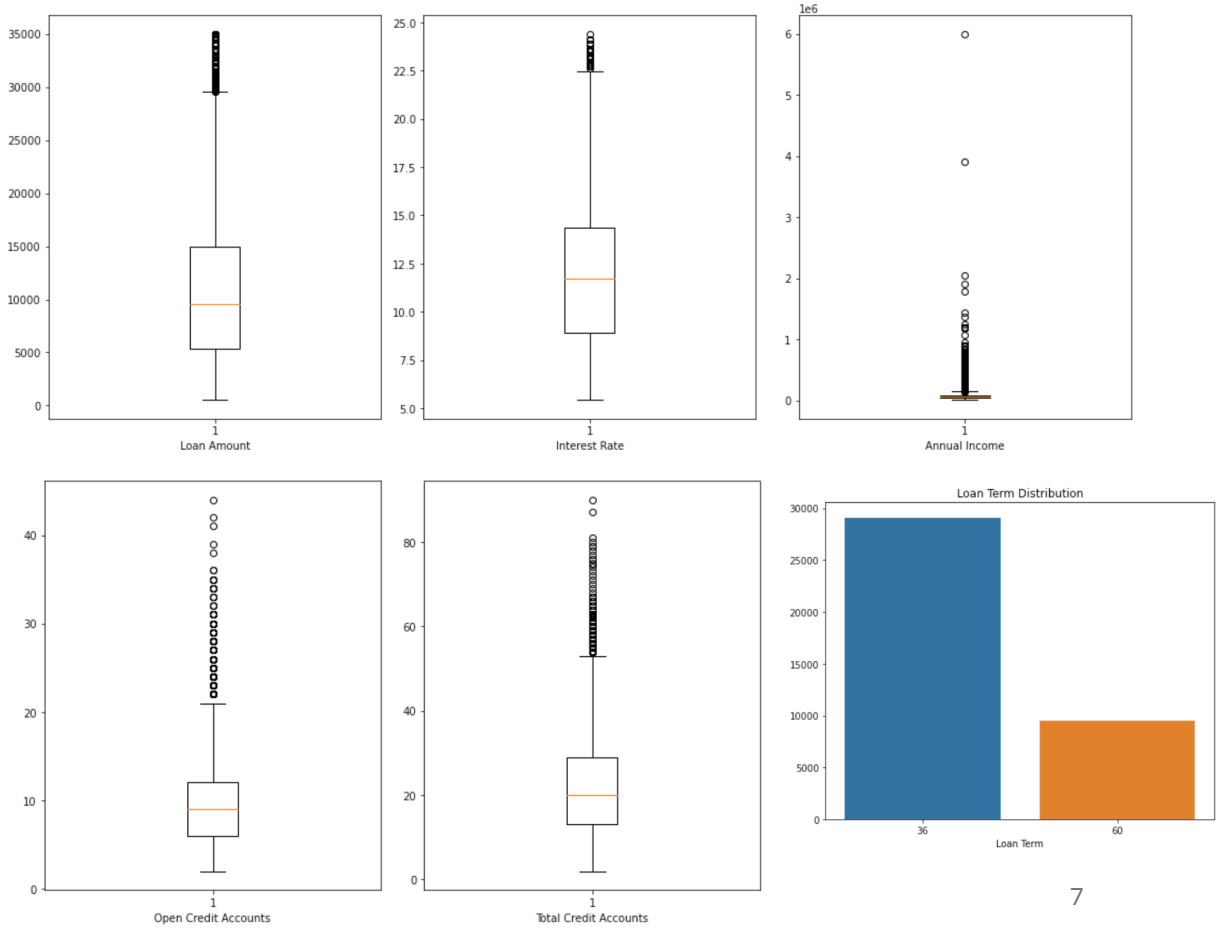


- Other columns were dropped that contained IDs, columns related to post charge-off details, columns related to current loan payments or interests, few date columns that only had month and year, columns that had only 1 unique value across the dataset etc.

- In the end we were left we 21 columns/variables and 38577 rows/observations

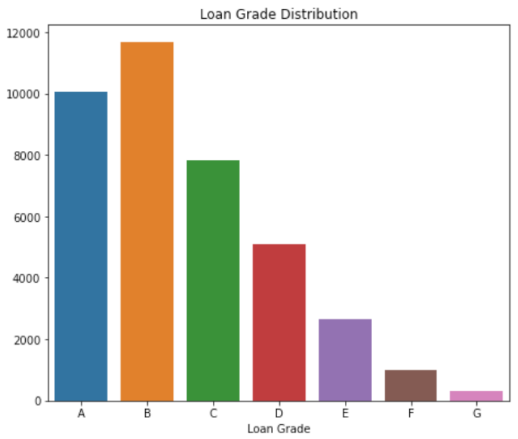
Univariate Analysis..(1/3)

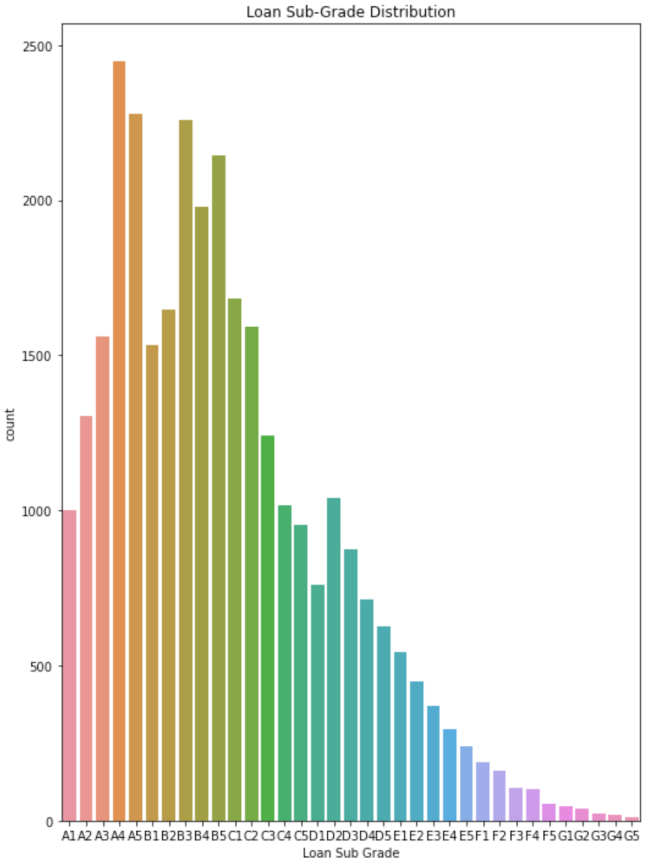
- We see a central tendency of Loan Amount around 10,000 with most loans between 5,300-15,000
- Interest rate central tendency is around 12% with 50% distribution between 9-15%
- Annual Income has a central tendency of around 60k with most incomes between 40k to 80k.
- We also observe that Open Credit Accounts have large number of outliers for higher values and have a central tendency of 9.
- Similar behaviour as Open Credit Account is observed with Total Credit Accounts with a central tendency of around 20. Open credit accounts would be a subset of Total credit accounts. The differential indicates closed/repaid or some other factors etc.
- More people opted for a 36 months term rather than 60 months term.

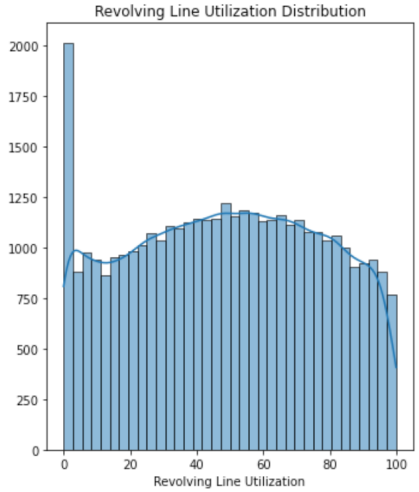


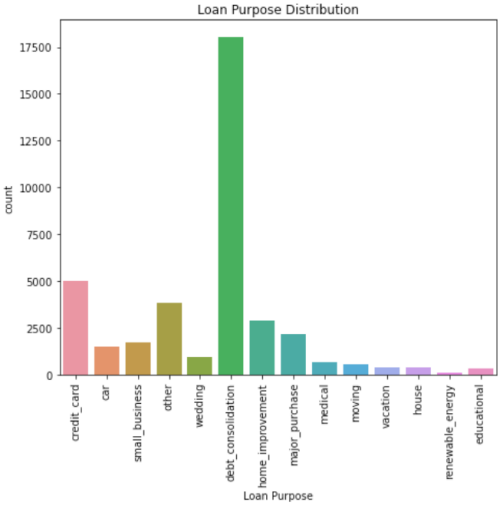
Univariate Analysis..(2/3)

- Most borrowers had Loan Grade B, A or C
- ‘Debt Consolidation’ was the most common purpose
- Most common Loan Sub-Grades were A4, A5, B3, B4, B5, C1 and C2
- Revolving utilization (the percentage of available credit that is used) shows a distribution where many borrowers are using a moderate to high percentage of their available credit. It appears that some borrowers may even be using nearly all of their available credit (100% utilization)





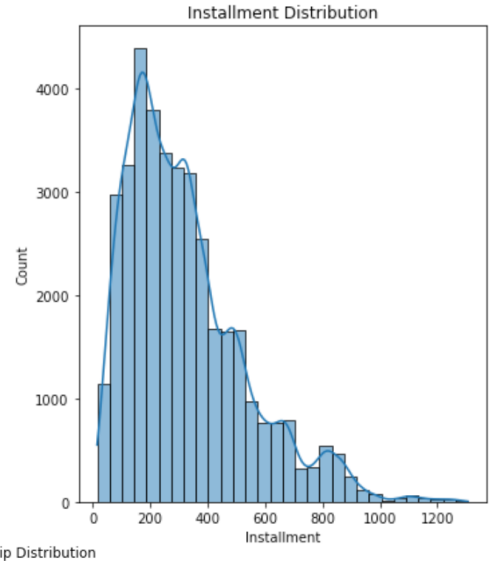


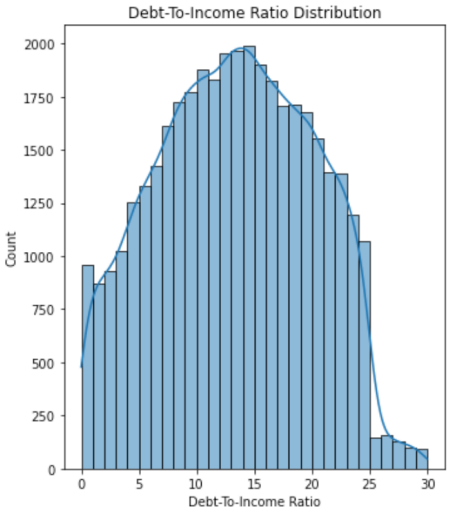


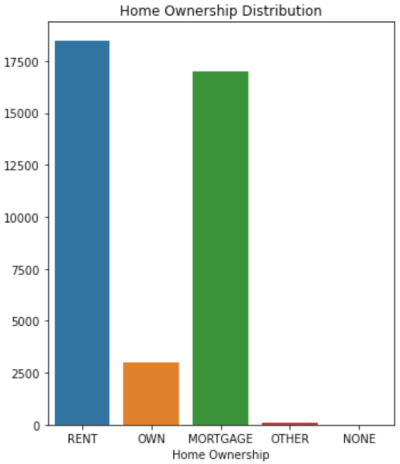
Univariate Analysis..(3/3)

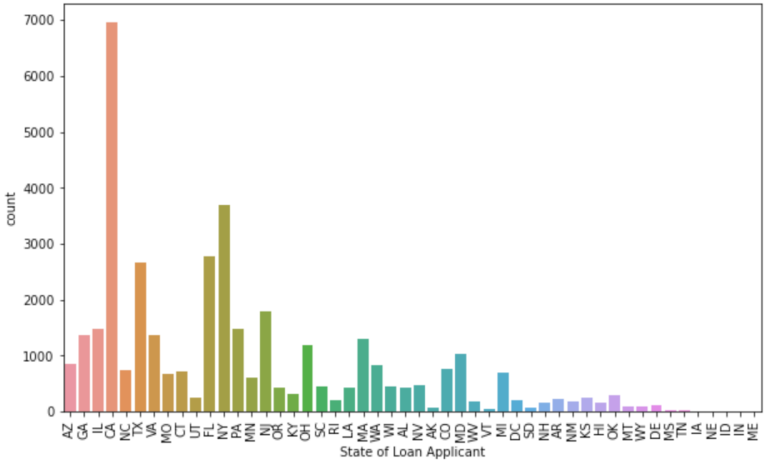
- The graph is right skewed, and the maximum number of installments seem to be around 200-300.
- A higher DTI may indicate an already burdened borrower before current loan and from the graph it appears there are large number of borrowers peaking between 10-16. Very high DTI 25-30 have lower number of borrowers.
- Most people borrowing are on rent. Now this may indicate they may be interested in applying for home mortgage/loan and that would need deeper analysis. The number of homeowners with mortgage is also high.
- Most Loan applicants seem to be from the state of California followed by New York and Florida

We proceeded with removing outliers from the numerical data and moved to Bivariate/multivariate analysis





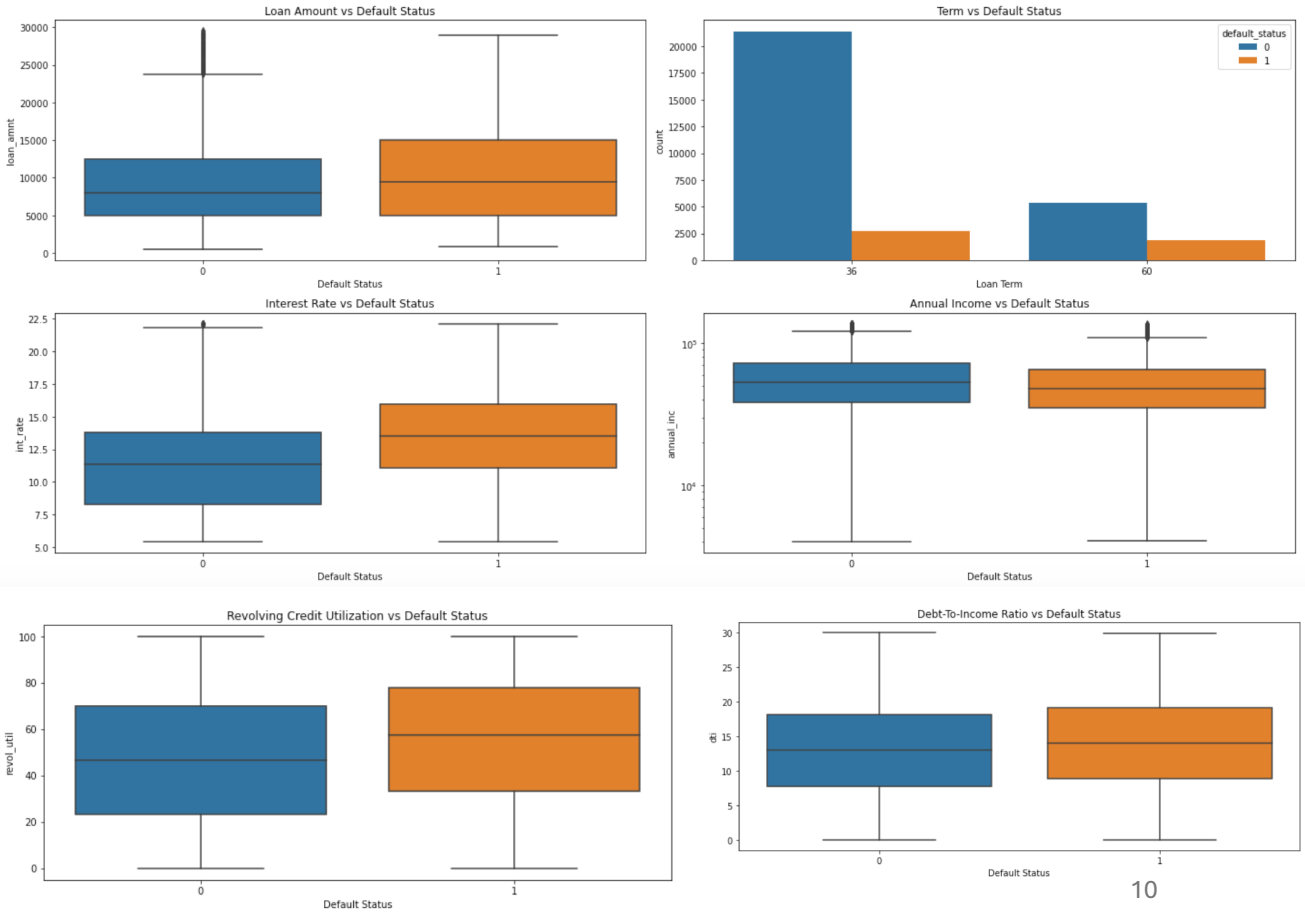




Bivariate/Multivariate Analysis .. (1/5)

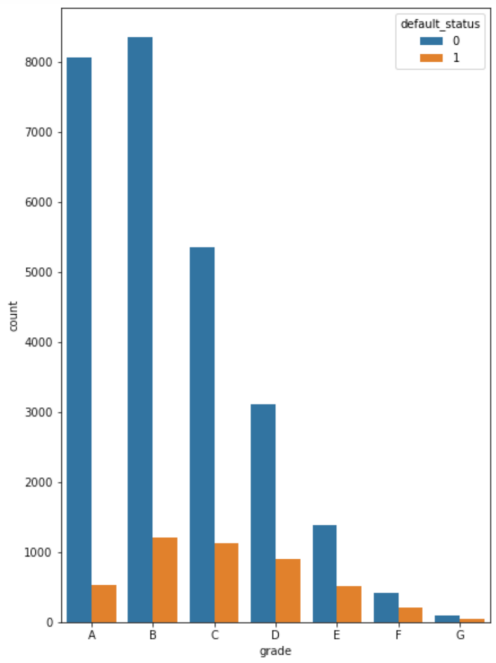
We computed a new column 'default_status' with 1 indicating default and 0 as fully paid. This column is used in context with other variables

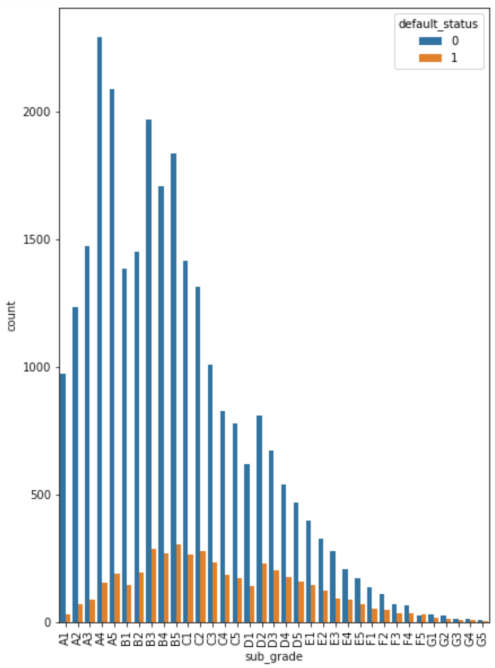
- Higher loan amounts seem to be associated with defaults, though the difference is not high
- Loans with a 60-month term appear more likely to default compared to those with a 36-month term.
- Loans with higher interest rates have a stronger association with defaults.
- Borrowers with lower annual incomes show a higher tendency to default, while higher-income borrowers are less likely to default but it is also a weak trend.
- Defaulters have a much higher median and IQR for credit utilization compared to non-defaulters i.e. defaulters use a higher percentage of their available credit. This could be an important flag for the consumer lending company as borrowers with higher revolving utilization may be a predictor of financial instability
- Defaulters have a higher DTI (Debt-To-Income Ratio) median compared to fully paid borrowers

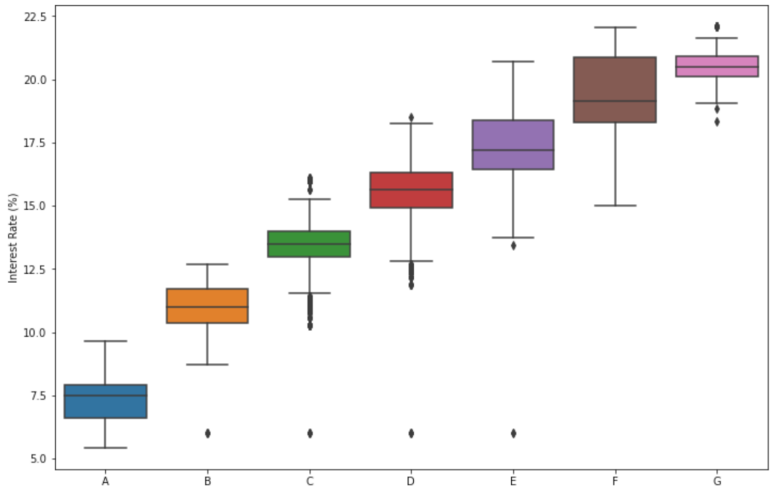


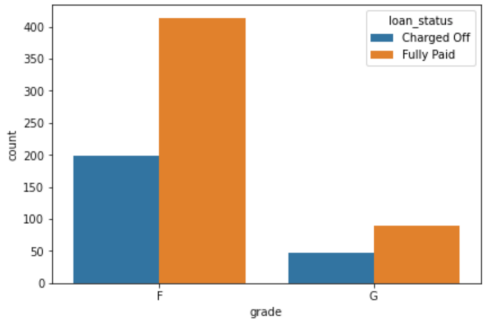
Bivariate/Multivariate Analysis .. (2/5)

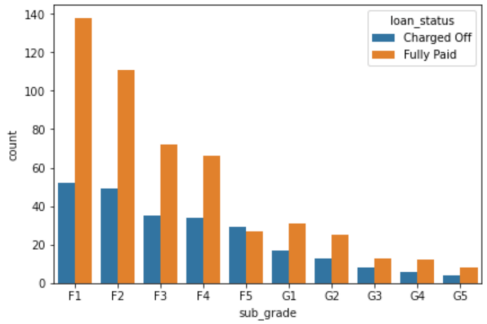
- Higher loan grades namely A and B have lower interest rates and lower loan grades from C to G have much higher interest rates
- It is obvious that Loan with Grades F and G are less likely to be paid off and highly likely to default.





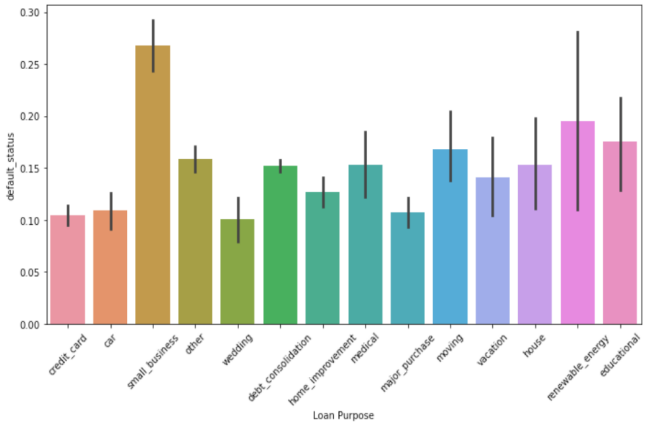


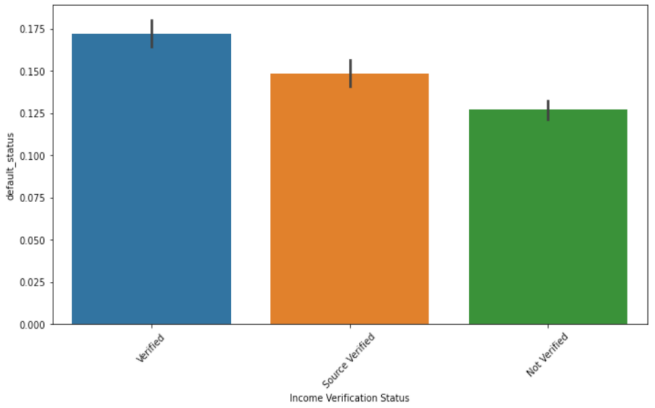


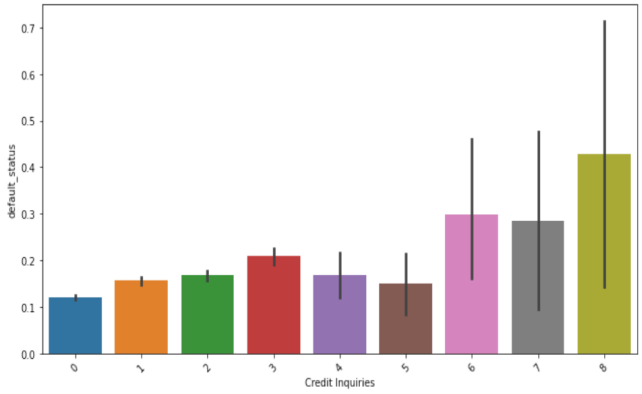


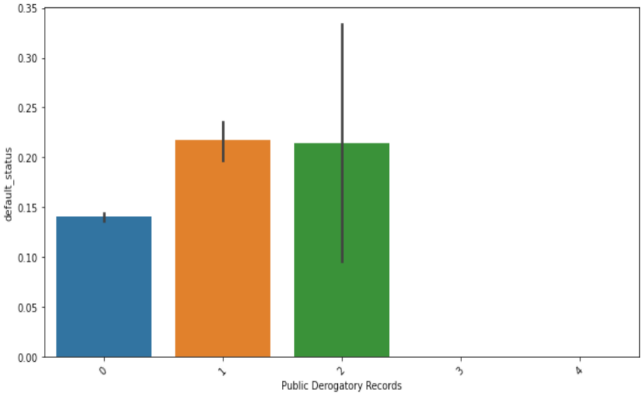
Bivariate/Multivariate Analysis .. (3/5)

- It appears that loans for small business defaulted the most. It may be a risky category for first time investors and chances of business failures may be high
- It is surprising to see that borrowers with verified incomes defaulted more
- It does look like in general as number of inquiries increase the chances of default behavior increase as well with the pattern peaking at 7 inquiries in last 6 months
- It looks like 1-2 derogatory records do contribute to default behaviour



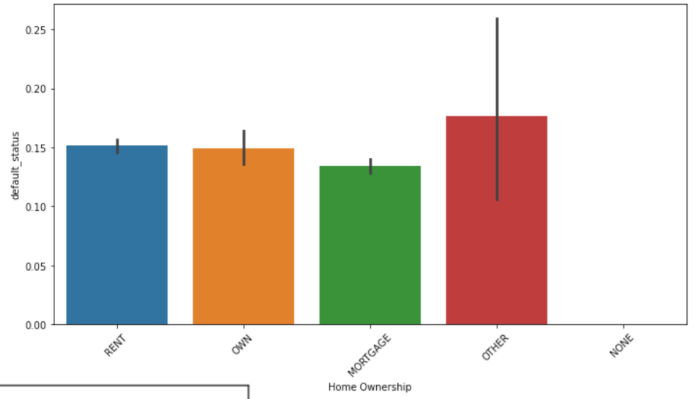


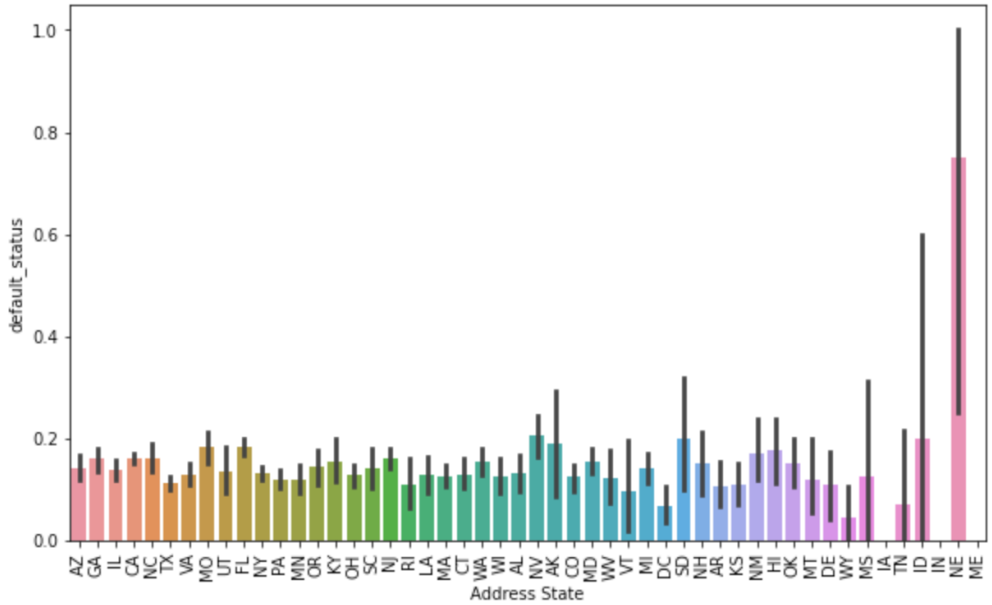




Bivariate/Multivariate Analysis .. (4/5)

- It seems like the 'other' and 'rent' categories tend to have more defaults. Not entirely sure what other means in the dataset. It could indicate the borrower to be a dependent.
- It looks like borrowers from state 'NE' or Nebraska are most likely to default. Some more risky states are likely to be 'ID' (Idaho), 'NV' (Nevada), 'SD' (South Dakota)

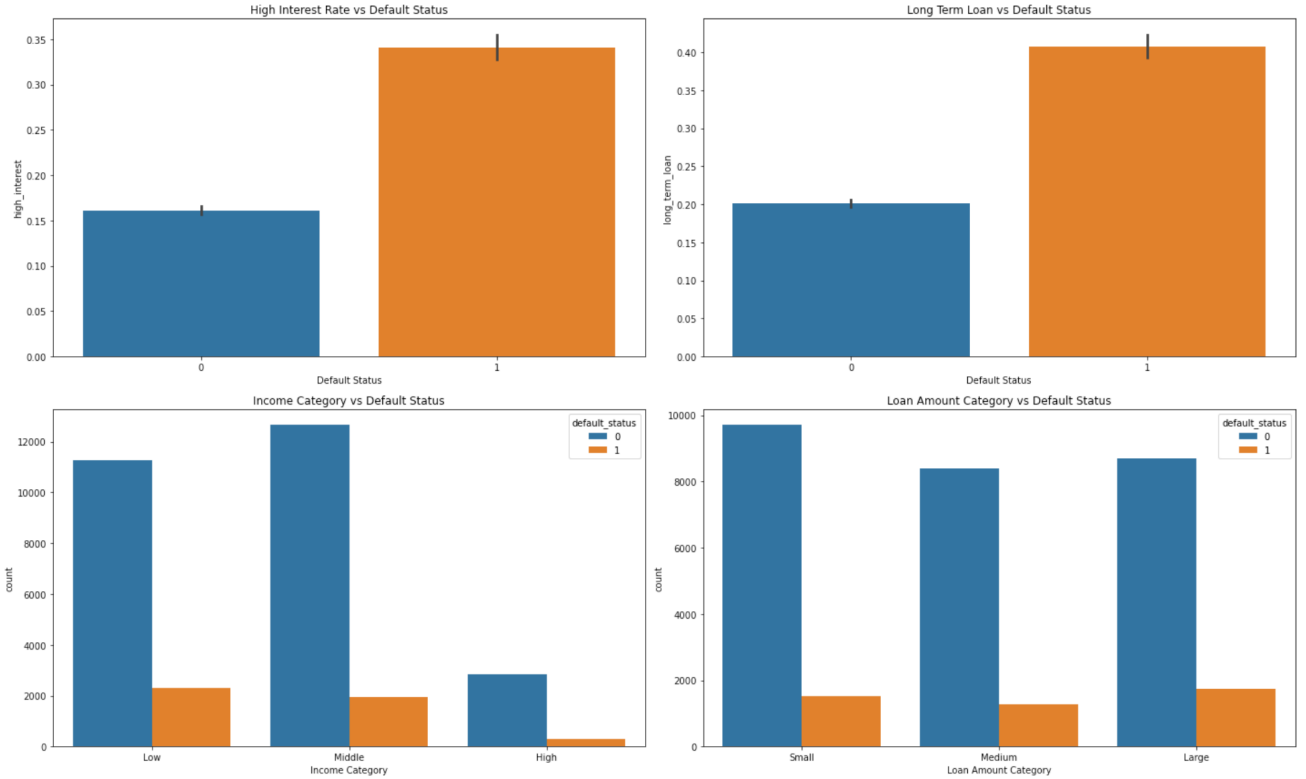




Bivariate/Multivariate Analysis .. (5/5)

We computed derived metrics like High Interest Indicator (> 15%), Long Term Loan Indicator, categorized Income into Low, Middle, High and categorized Loan Amount into Small, Medium and Large

- High Interest Rate leads to a greater number of loan defaults
- Longer Term loans (5 yrs) leads to a greater number of loan defaults
- Income Category is not a significant discriminator for defaults as majority of borrowers are concentrated on low-income category
- Larger loans tend to be riskier in terms of default though not a very strong trend



Conclusion .. (1/2)

After a long process of analysis, we will conclude on few strong indicators/patterns for loan defaults. The associated rationale is also documented

1. **Interest Rate ('int_rate')**: Borrowers with higher interest rates are more likely to default. Lenders typically charge higher interest rates to borrowers with lower creditworthiness, which is often associated with a higher probability of default. Loans with interest rates above 15% (high-interest loans) are particularly risky and show a higher correlation with defaults.
2. **Loan Term ('long_term_loan')**: Borrowers with longer loan terms (60 months) are more likely to default compared to those with shorter loan terms (36 months). Longer loans carry more risk, as the borrower has to manage payments over a longer period
3. **Loan Purpose ('purpose')**: Small Business seem to be a likely risky category as individuals investing in small business (especially if for the first time) and seeking capital for it has a higher chance of failure in their business.
4. **Revolving Utilization ('revol_util')**: High revolving credit utilization signals that the borrower is heavily reliant on available credit, which increases their risk of financial distress and default.

Conclusion .. (2/2)

5. **Debt-to-Income Ratio ('dti')**: A higher debt-to-income ratio indicates that a borrower has a significant portion of their income dedicated to debt repayment (and less money available for additional expenses or unforeseen circumstances), increasing their likelihood of default
6. **Credit Inquiries ('inq_last_6mths'):** In general, if a borrower has high number of credit inquiries over a short period of time (6 months), it shows financial strain as they are trying to arrange for credit resulting in some financial institutions looking up their credit files.
7. **Loan Amount ('loan_amnt_category'):** Larger loans carry higher risks as it would generally carry higher interest rates and installments over longer term. Unless the borrower is of sound income, it is usually hard to manage larger loans.
8. **Loan Grade and Sub-Grade ('grade', 'sub_grade'):** There is a high chance of loan defaults in Loan Grades F and G
9. **Address State ('addr_state'):** It looks like borrowers from state 'NE' or Nebraska are most likely to default. Some more risky states are likely to be 'ID' (Idaho), 'NV' (Nevada), 'SD' (South Dakota)
10. **Income Category ('income_category'):** Low Income Category borrowers are most likely to default.
11. **Derogatory Public Records or Bankruptcies ('pub_rec', 'pub_rec_bankruptcies'):** Defaulters are more likely to have either derogatory public records or public bankruptcies.