Lending Patterns

EDA case study to recognize patterns of loan defaulters

By: Snehal Jadhav and Shrey Jain

Context

- There is consumer finance company which specializes in lending various types of loans to urban customers.
- We are provided with historical data set of loan applicants and whether they ‘defaulted’ or not.

Goal of analysis

- The goal is to recognize patterns that suggest whether a person is likely to default or not.
- This information can be used to take actions such as denying the loan, reducing the loan amount, or offering loans to risky applicants at a higher interest rate.

Tech stack used

- Python 3.9
- Jupyter notebook
- Libraries:
  - Pandas
  - Numpy
  - Seaborn
  - Matplotlib

Data description

- Historical data consists of:
  - 111 columns
  - 39717 rows
- Data type of column values:
  - float64
  - int64
  - object
- Dataset contains both numerical and categorical variable

Data exploration and cleaning

Columns with NaN values

- 54 columns with all null values, which we have removed
- These columns were having no impact on analysis

Columns with only single value

- Total 9 columns with only having single value
- These aare also dropped because of no significant value for analysis

Metadata and description columns

- Columns denoting ID, addresses, title, URL, detail description
- Total of 8 columns found
- These columns were also dropped as this data is not required for the analysis goal of finding patterns on defaulters

Data exploration and cleaning

Post lending columns

- Columns which denotes data values which were taken after loan approval
- As the analysis is to find pattern to recognize defaulters before loan approval, we can drop these columns as well
- Total of 10 such columns were dropped

Missing values

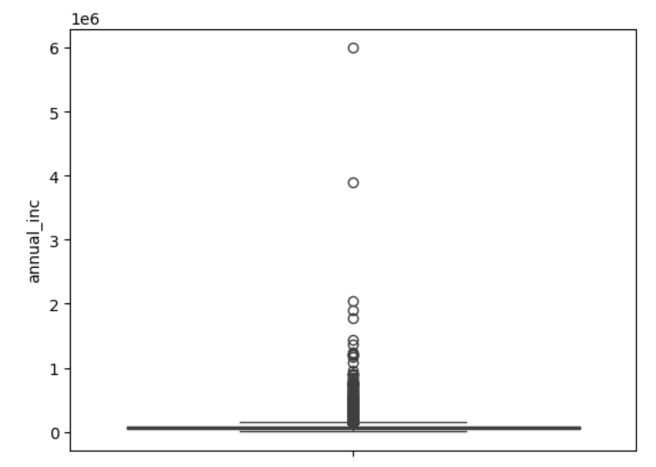
- Column like emp_length is imputed with mode value as missing values were significant low in number
- Revol_util column had only 50 rows with null values, as this number is far low compared to 39K rows, it was much optimistic to drop these rows

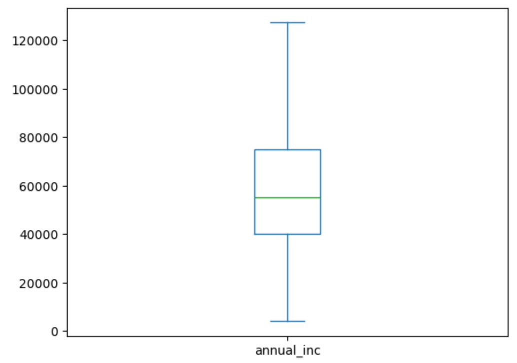
Date columns

- Created a separate column based on issue_d column which is issue_year, this column have only the year of loan issue

Outlier

- Column annual_inc have good number of outliers
- This column required outlier treatment
- The graph shown here denotes column with outlier an after outlier treatment



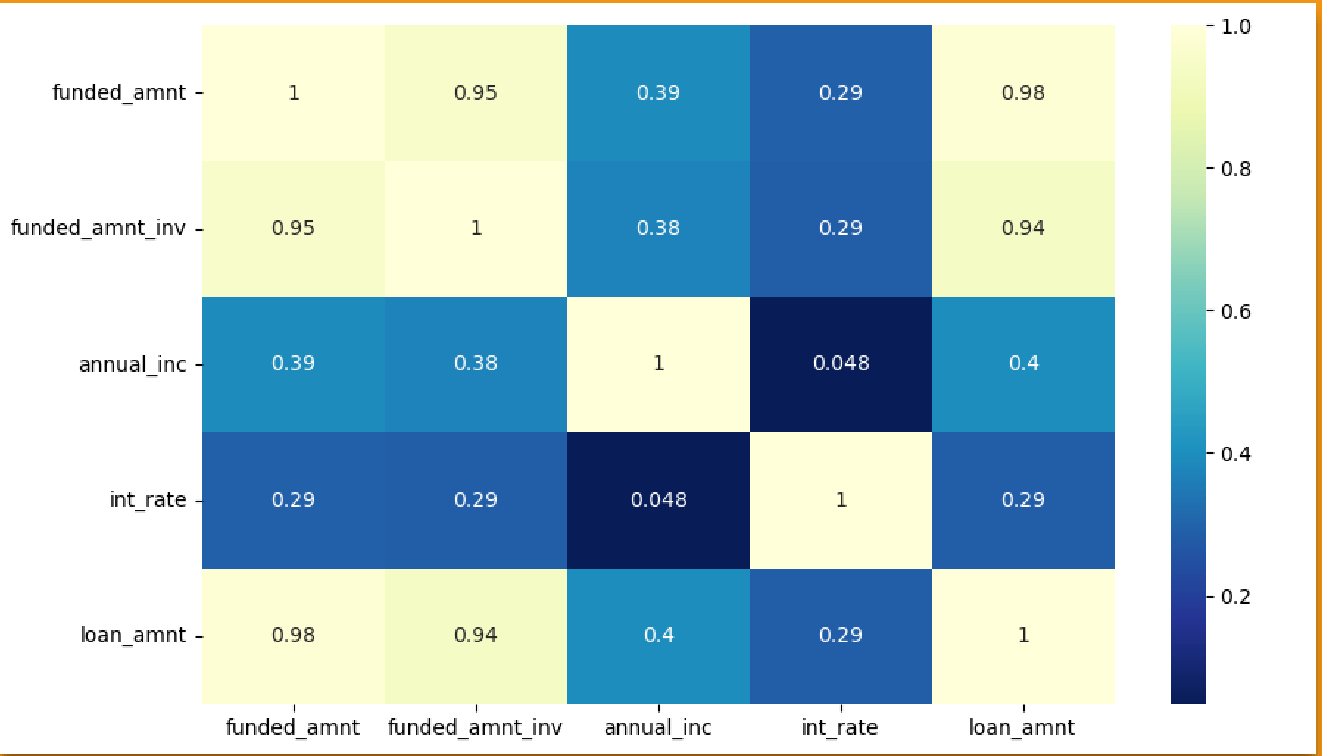


Multivariate Analysis

To get correlation among the coumns

Correlation between numerical columns

- fund_amt and fund_amt_inv is almost same as Loan Amount w.r.t Annual Income
- We can drop fund_amt and fund_amt_inv



Decision making attributes

Numerical columns

- annual_inc
- int_rate
- loan_amnt

Categorical columns

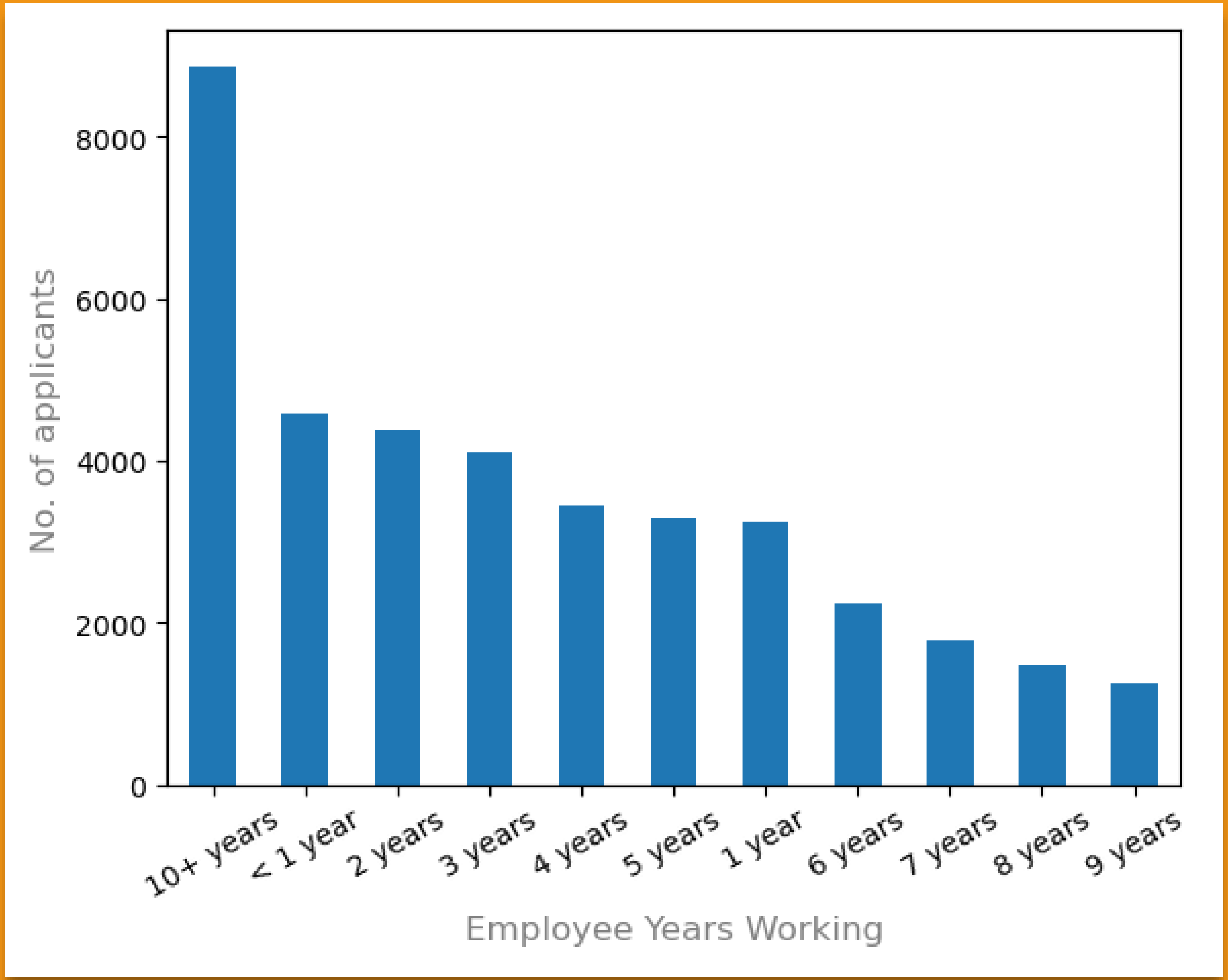
- grade
- sub_grade
- emp_length
- verification_status
- home_ownership
- purpose
- loan_status
- term
- Target

Univariate Analysis

To check column wise behavior and outlier detection

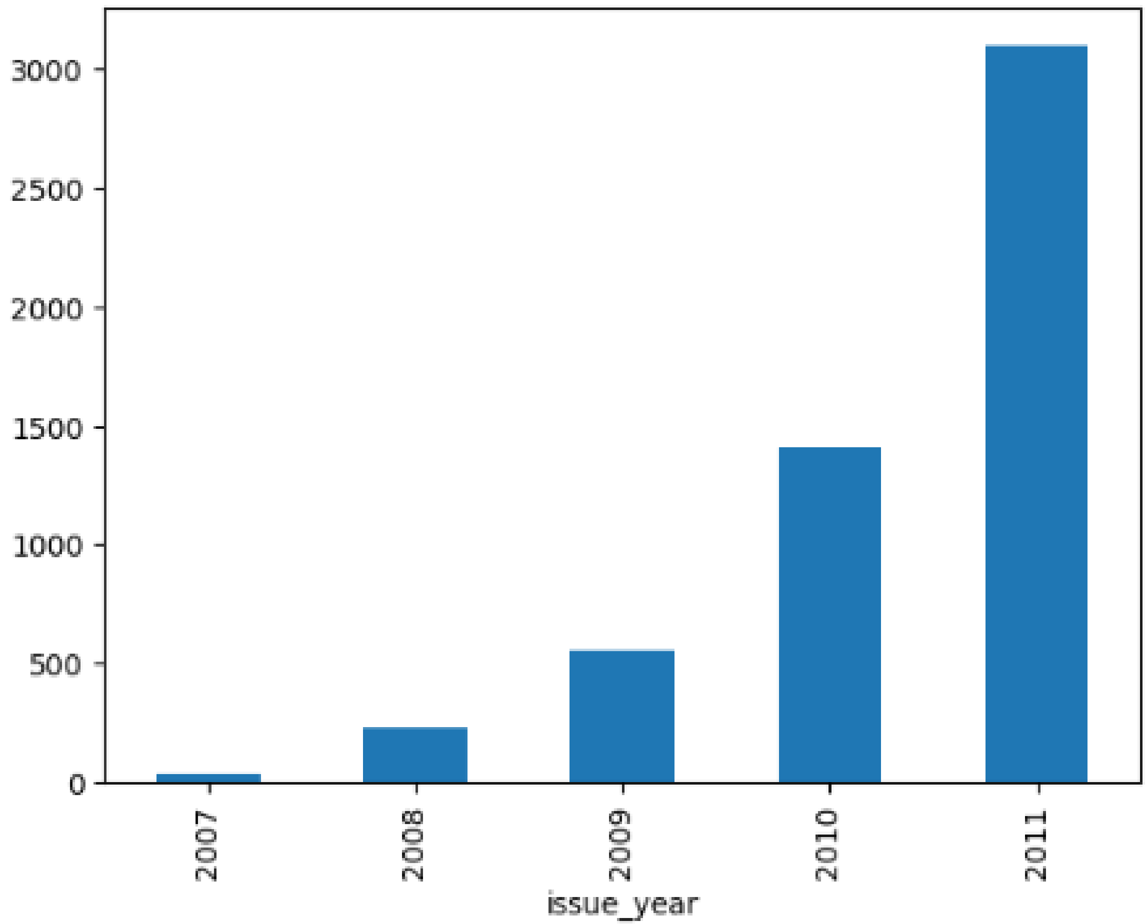
Loan applicants vs Working year

- Applicants with >10 years have availed maximum number of loans



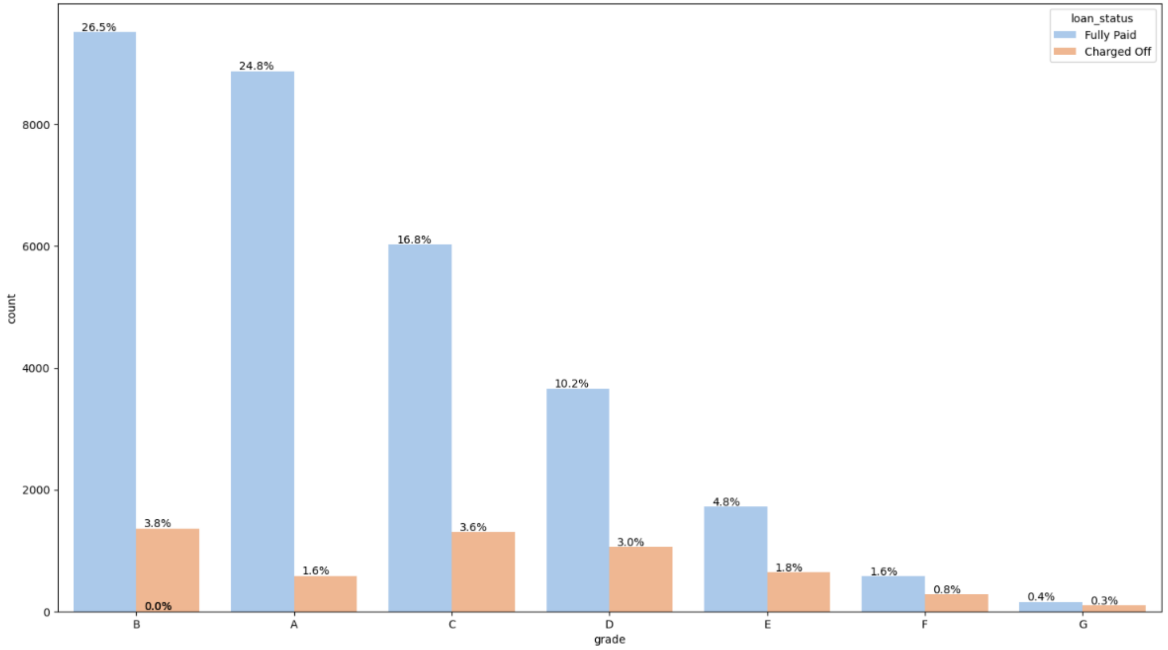
Year wise analysis

- Number of defaulters are more in 2011
- Possibility some mass event happened like layoffs or recession



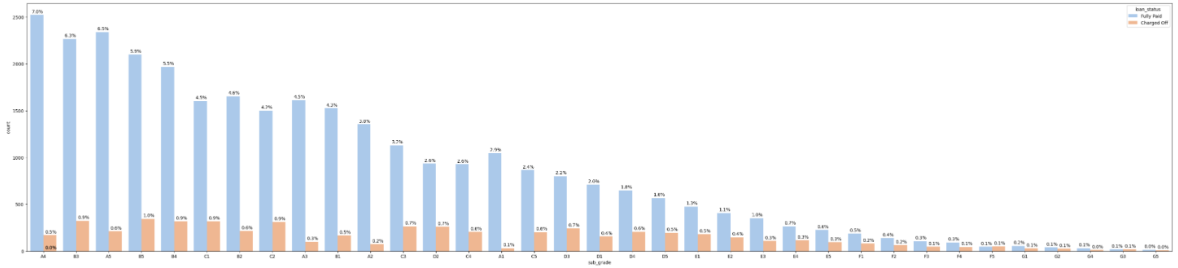
Segmented Analysis: grade

- Comparison of number of Fully paid and charged off
- The following chart denotes grade wise data
- We can see Grade A,B,C,D are having significant large number of non-defaulters than with grades E,F,G



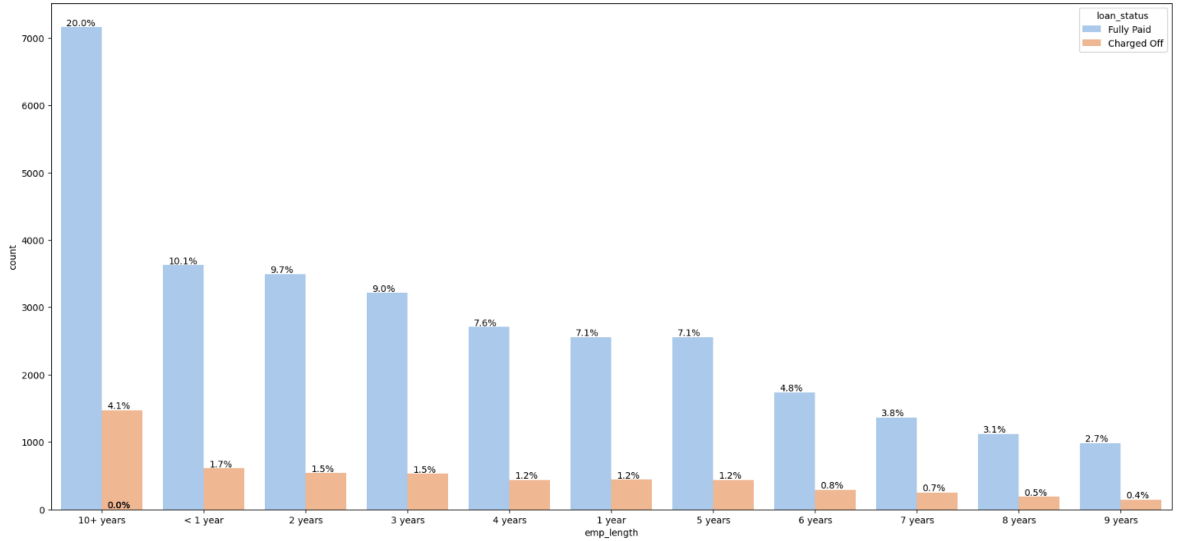
Segmented Analysis: sub_grade

- Comparison of number of Fully paid and charged off
- The following chart denotes sub_grade wise data



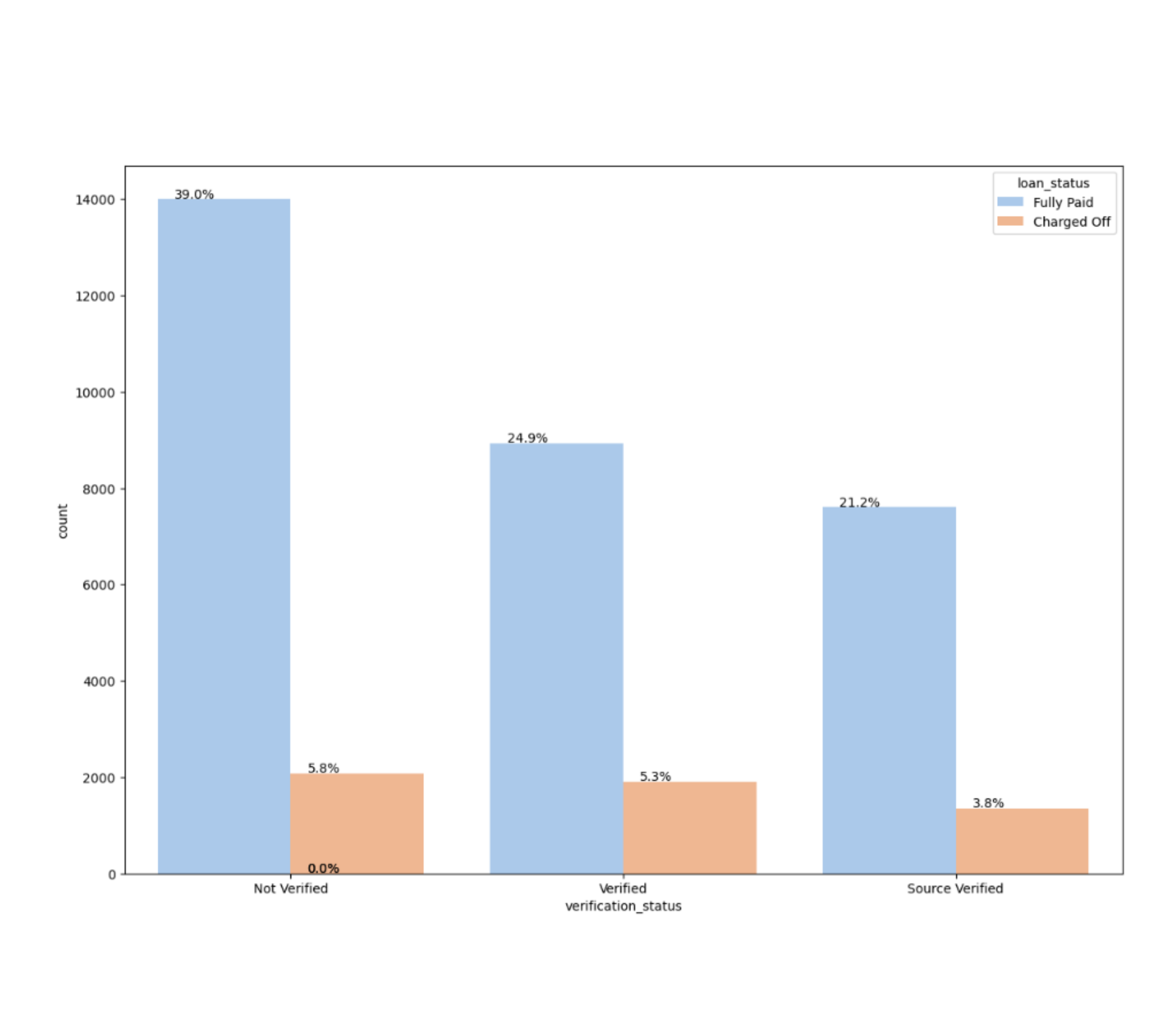
Segmented Analysis: emp_length

- Comparison of number of Fully paid and charged off
- The following chart denotes emp_length wise data
- We can see that employees with >10 years of employment have greater numbe of loans an large number of defaulters as well



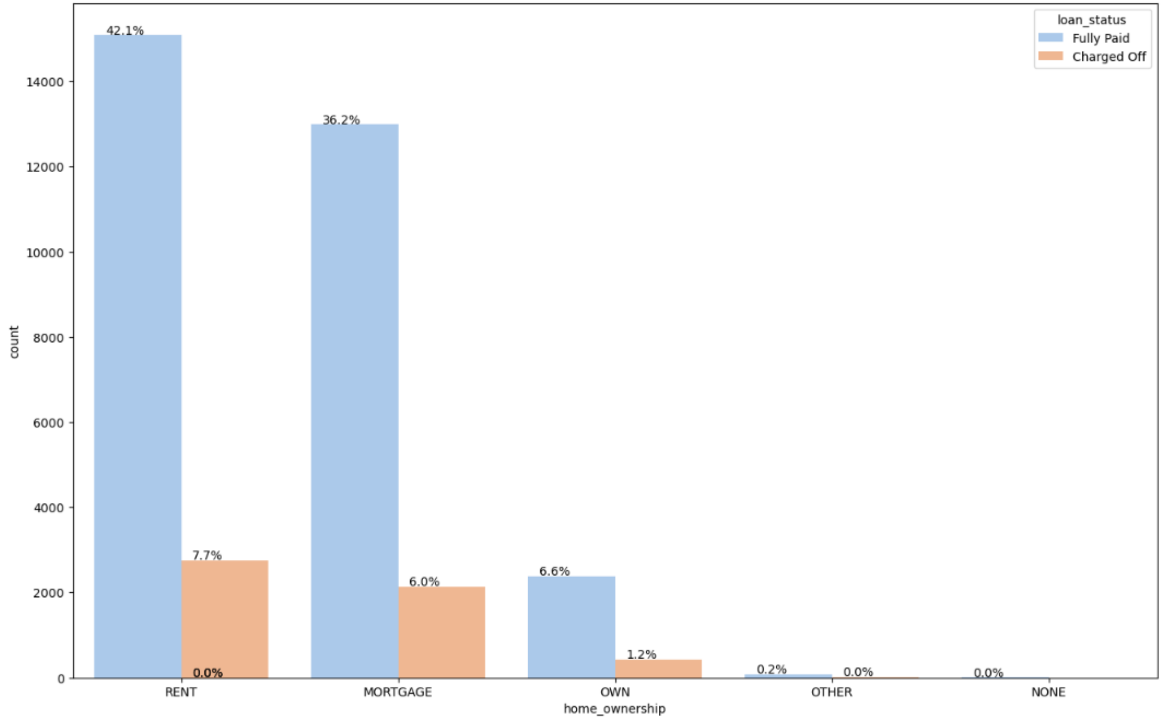
Segmented Analysis: verification_status

- Comparison of number of Fully paid and charged off
- The following chart denotes verification_status wise data
- Observation: verified and source verified have comparative large number of defaulters



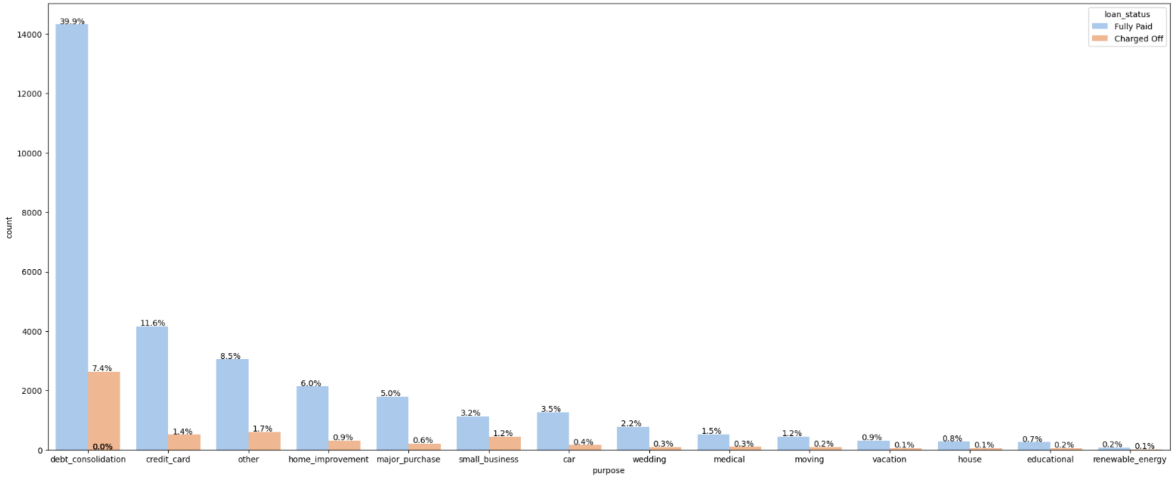
Segmented Analysis: home_ownership

- Comparison of number of Fully paid and charged off
- The following chart denotes home_ownership wise data
- Observation: People with rental or mortgaged home have higher defaulter rate



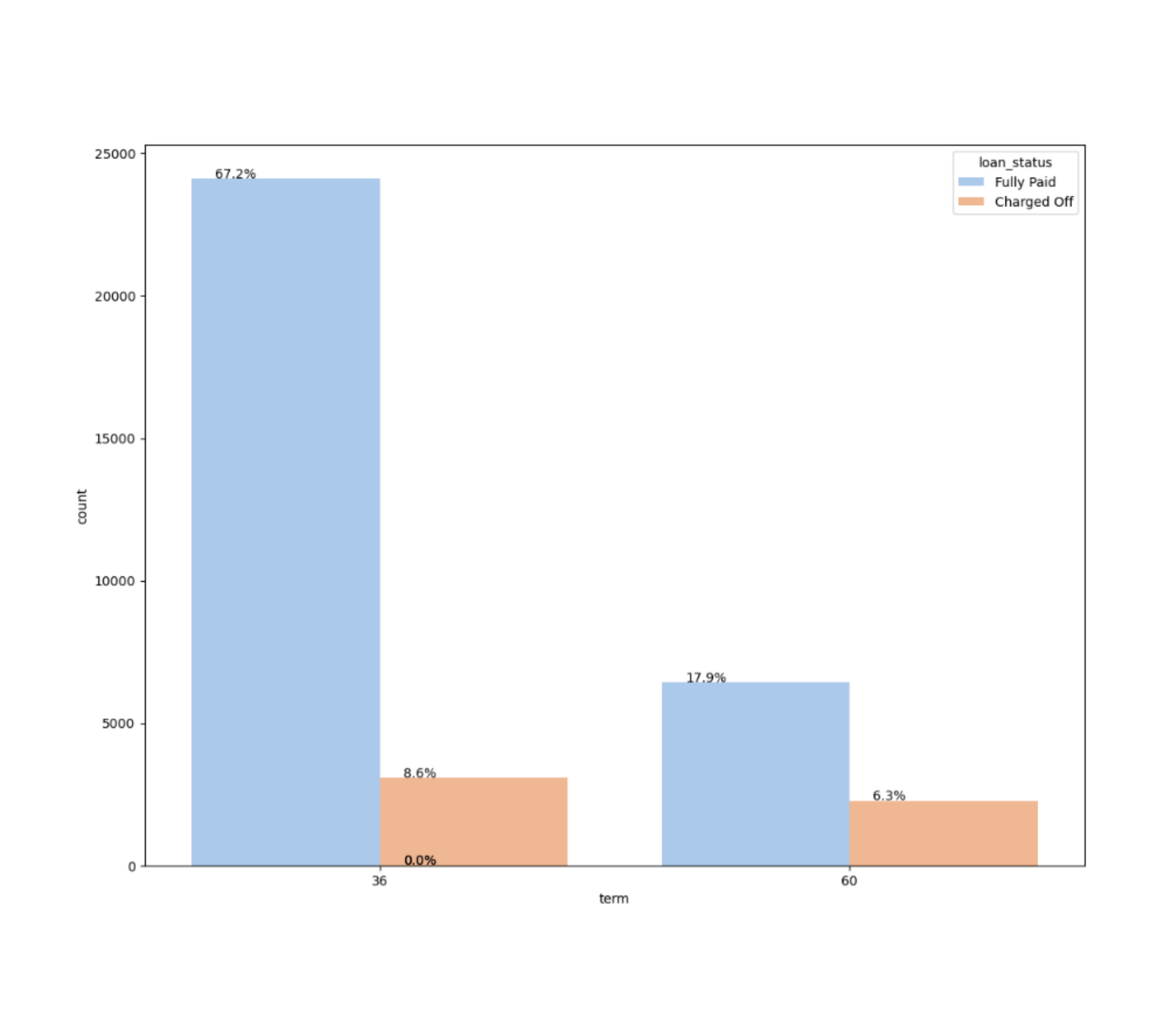
Segmented Analysis: purpose

- Comparison of number of Fully paid and charged off
- The following chart denotes purpose wise data
- Observation: No clear observation by this graph as he values are more.
- We may need a probability measure as well on this comparaing defaulters vs non-defaulters



Segmented Analysis: term

- Comparison of number of Fully paid and charged off
- The following chart denotes term wise data
- Observation: Defaulters are more when the loan term is more

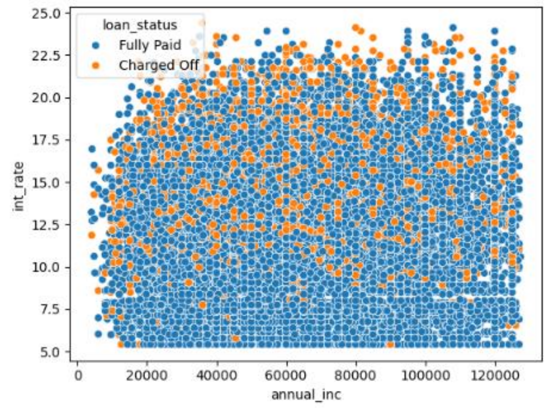


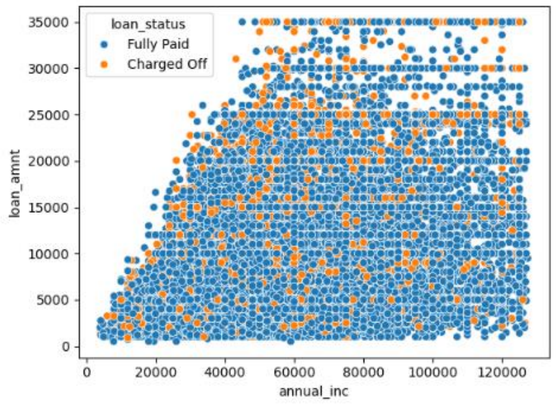
Bivariate Analysis

Observing numerical columns and categorical columns

Bivariate analysis of numerical colummns

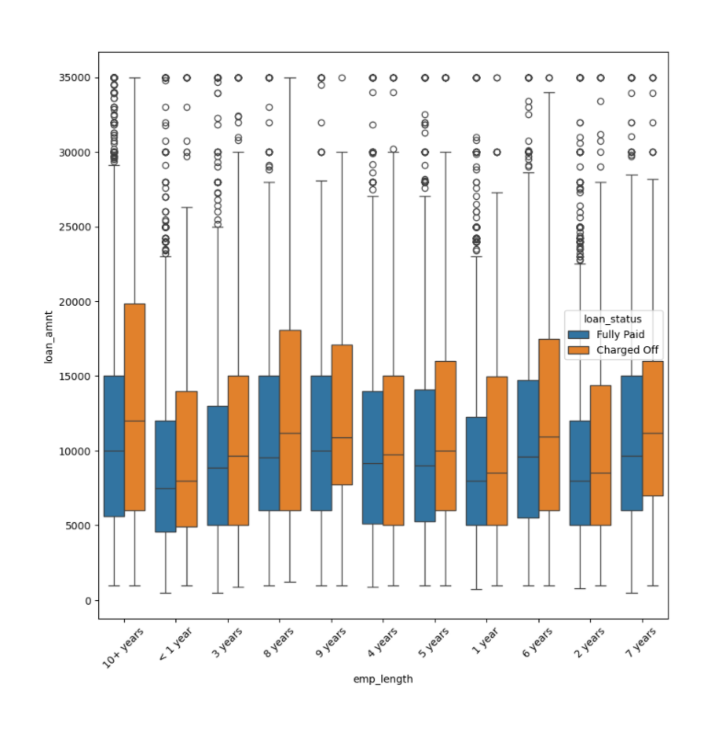
- Did not notice much difference to analyse in terms of Annual Income and Loan amount
- As the Interest Rate increases, we set a range of defaulters increasing in the range of 20000 - 60000 annual income bar

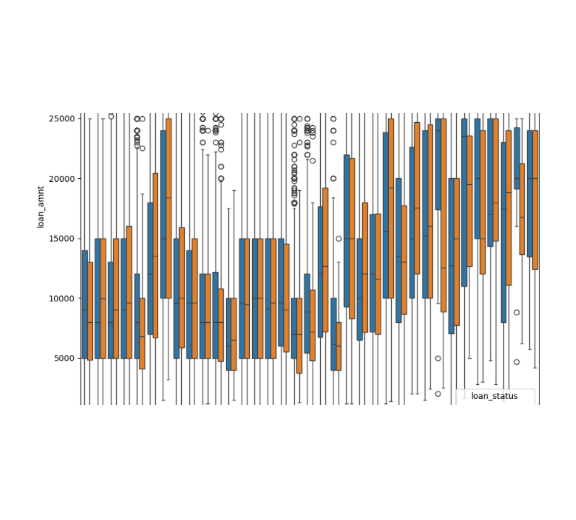


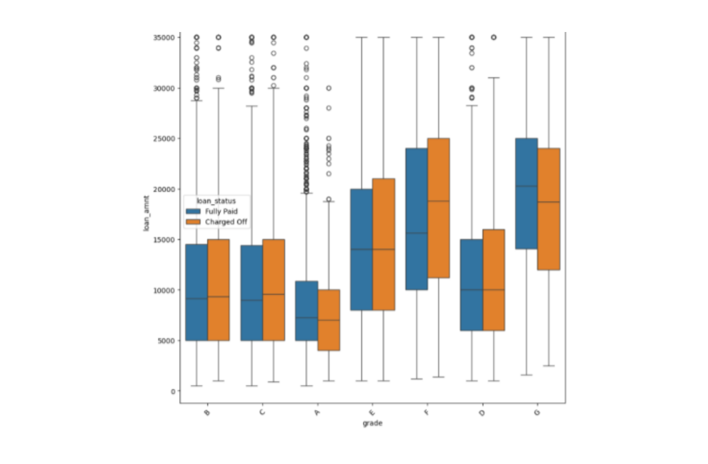


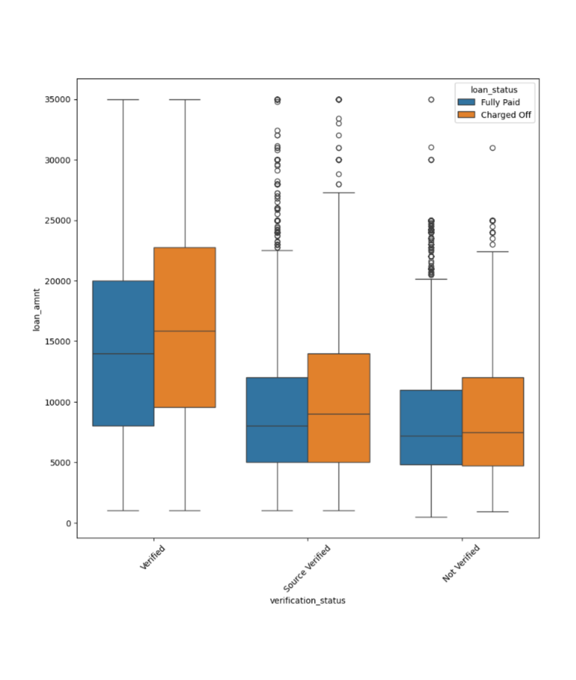
Bivariate analysis of categorical columns

- On comparing grade, sub grade, employment length and verification status with loan amount
- We found similar analysis like of univariate when comparing these columns with count of defaulters and non-defaulters
- This build us a confidence that our analysis is going in right way







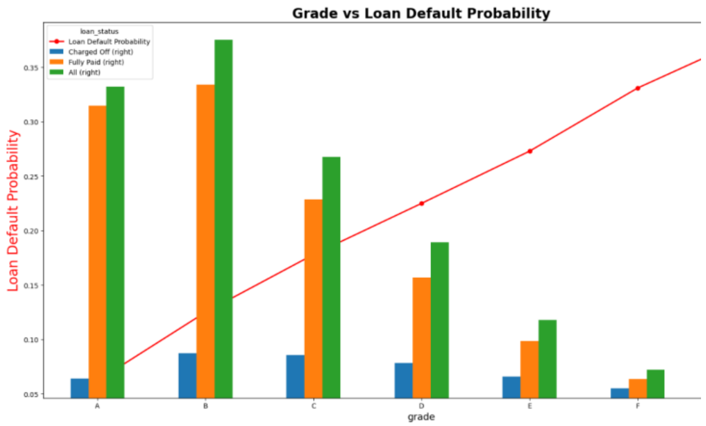


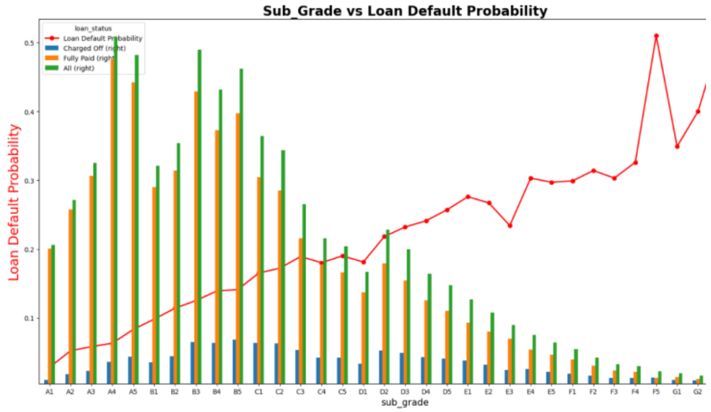
Probability of defaulting

Finding which data column have high impact/probability for being a defaulter

Grade and sub grade

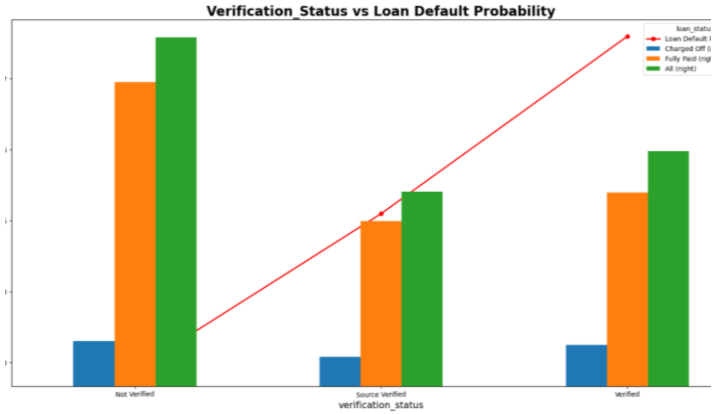
- Applicants with grade F and **G** is having higher probability of being defaulter
- Same in sub grades, Fs and Gs are having higher probability of getting defaulted

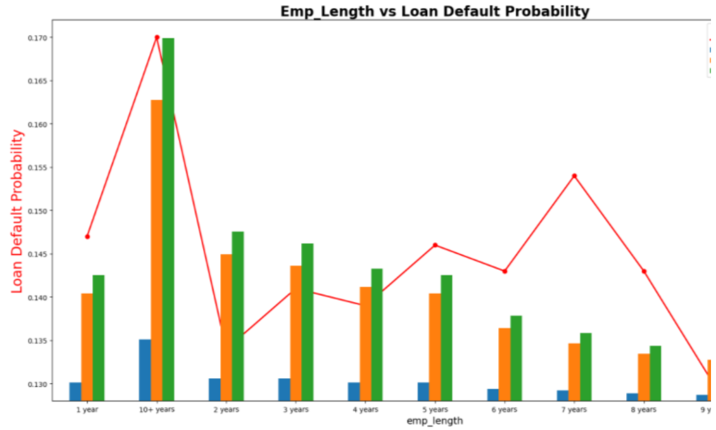




Employment length and verification status

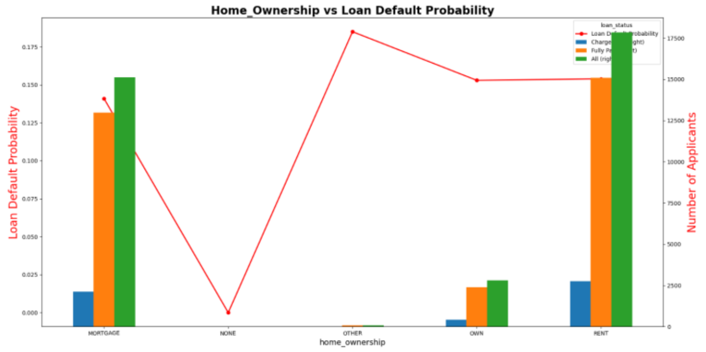
- Probability of defaulting is highest with emp length >10 years
- Also, when the verified applicants are having higher chances of being a defaulter

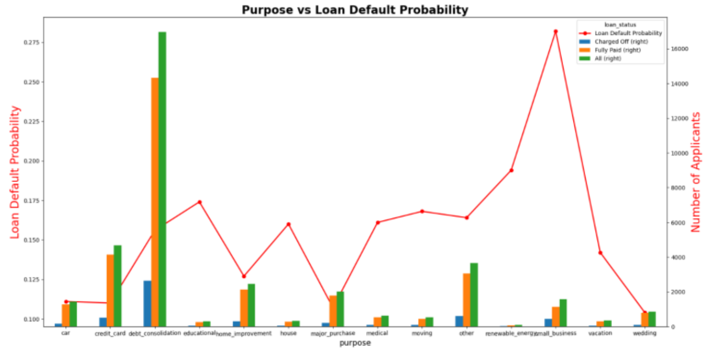




Purpose and home ownership

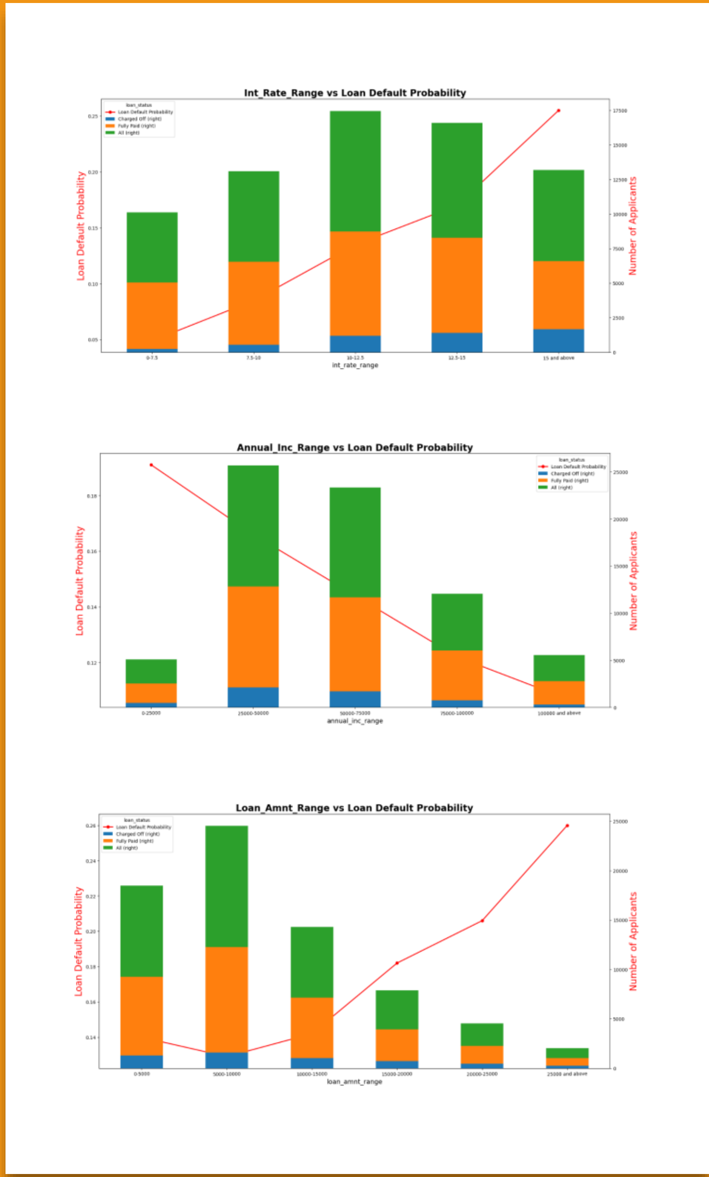
- We can observe that defaulters and probability of defaulting is more when applicant is mortagaging
- Also, when the purpose is to build small business, that time as well default rate and probability is higher





Interest rate range, loan amount and annual income

- Default rate is higher when loan amount is >25K
- Probability of being defaulter is higher when interest rate is >15%
- Applicants with <25K annual income have higher chances of being a defaulter



Final observation and patterns

Following are the patterns of applicants with probability of being a defaulter in future

- With home ownership as 'MORTGAGE'
- Having loan at interest rate of >15%
- Falling in grade F and G with sub grades in Fs and Gs
- Having employment length of > 10 years but with less annual income
- Purpose is to build small business
- Loan amount is >25000
- With annual income below 25000
- With a verified loan status

Contributors

- Snehal Amol Jadhav
- Shrey Kumar Jain