Lending Club Case Study:

Aparna Bindage
Kajal Kankariya

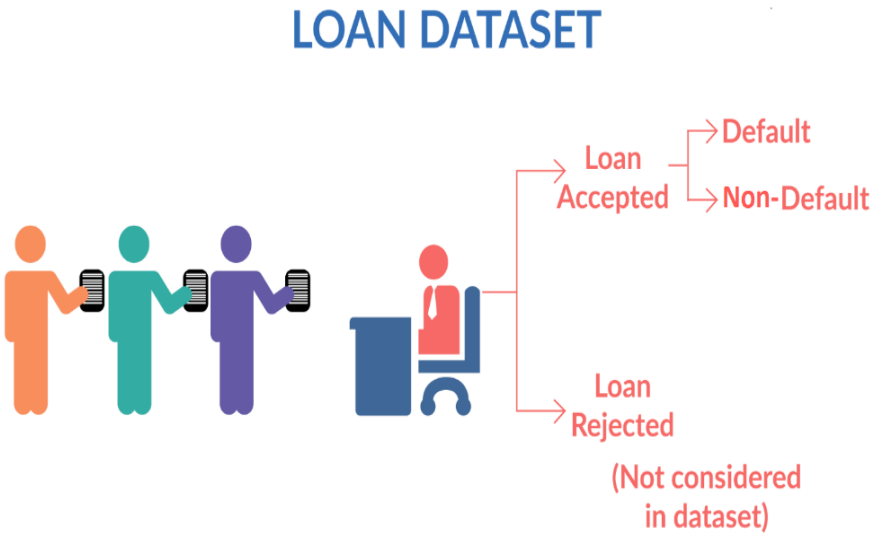
Agenda

- Problem Statement
- Steps of EDA
- Data Understanding
- Data Cleaning
- Missing Value Treatment and Metrics
- Outlier detection
- Data Analysis
- Univariate Analysis
- Bivariate Analysis
- Recommendations and Insights

Problem Statement

When the company receives a loan application, the company has to make a decision 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 company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.



Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

Steps In EDA

There are four major parts that are needed to be done for this case study:

1. Data understanding
2. Data cleaning
3. Data Analysis
4. Recommendations

Data Understanding

1. Remove rows with loan status as Current: We want to analyze if new applicant can default or not, so we need to check the behaviour of fully paid and charged off records to analyze

2. Drop columns –
    a. Missing values more than 90%
    b. Customer behaviour features – these features are not available at the time of loan application
    c. Not useful for analysis (url, zip_code, desc, title)
        i. url only has id which is different
        ii. zip_code has only first 3 values in dataset
        iii. desc column has many unique values which are string
        iv. title – many different values ranging from numeric to characters
    d. All records having same value

Data Cleaning

1. Remove unwanted characters from values.
   - Remove "months" and “%” string so that it can be converted to proper datatype
   - Emp_length should have values between 0-10 so replace <1 with 0 and 10+ with 10.

2. Fix incorrect data type
   - Convert term column values to int datatype
   - Convert int_rate column values to float datatype
   - Convert annual_inc column values to int datatype
   - First change issue_d column to datetime format

3. Outlier Detection

Outliers are seen in continuous variables, we have only treated outliers in annual_inc, as they were clearly visible.

Missing Value Treatment and Metrics

- In “emp_length” imputation with mode value cannot be done because the null values can also mean employees with 0 experience

- Similarly, “emp_title” cannot be imputed because it is not relevant to assign titles to null values as there are many unique values in this column

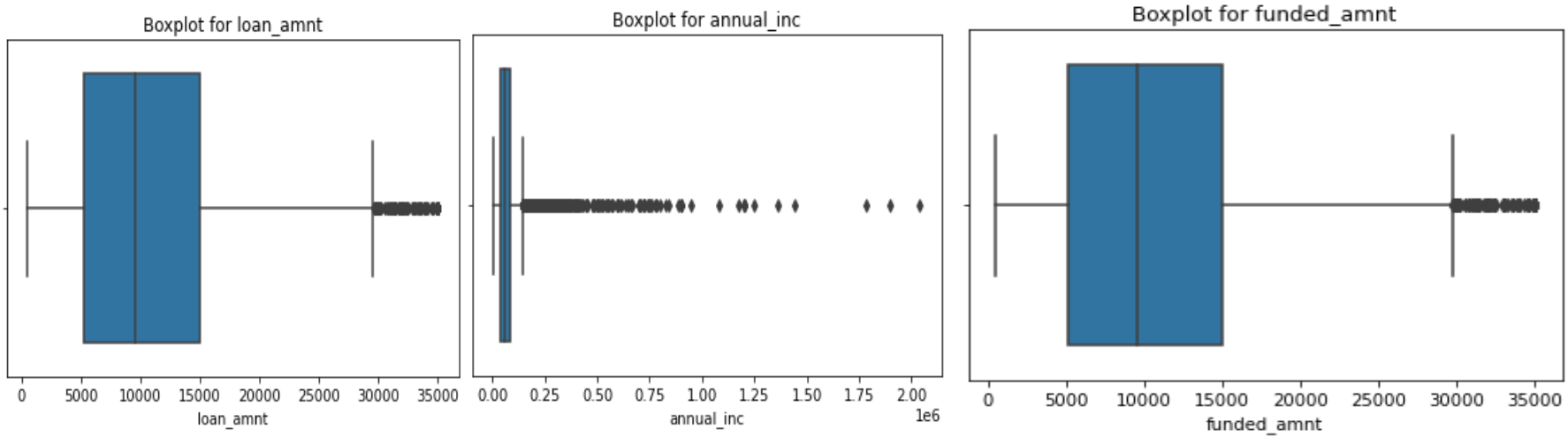
- **Type Driven Metrics** –

  1. To convert “annual_inc” into bins. This is because it will be easy to analyze large amount of data after converting into bins
  2. To convert “dti” values into bins. This is because it will become easy to visualize dti against some other parameter

- **Data Driven Metrics** – Extract loan issue month & year from date column

Outlier Detection

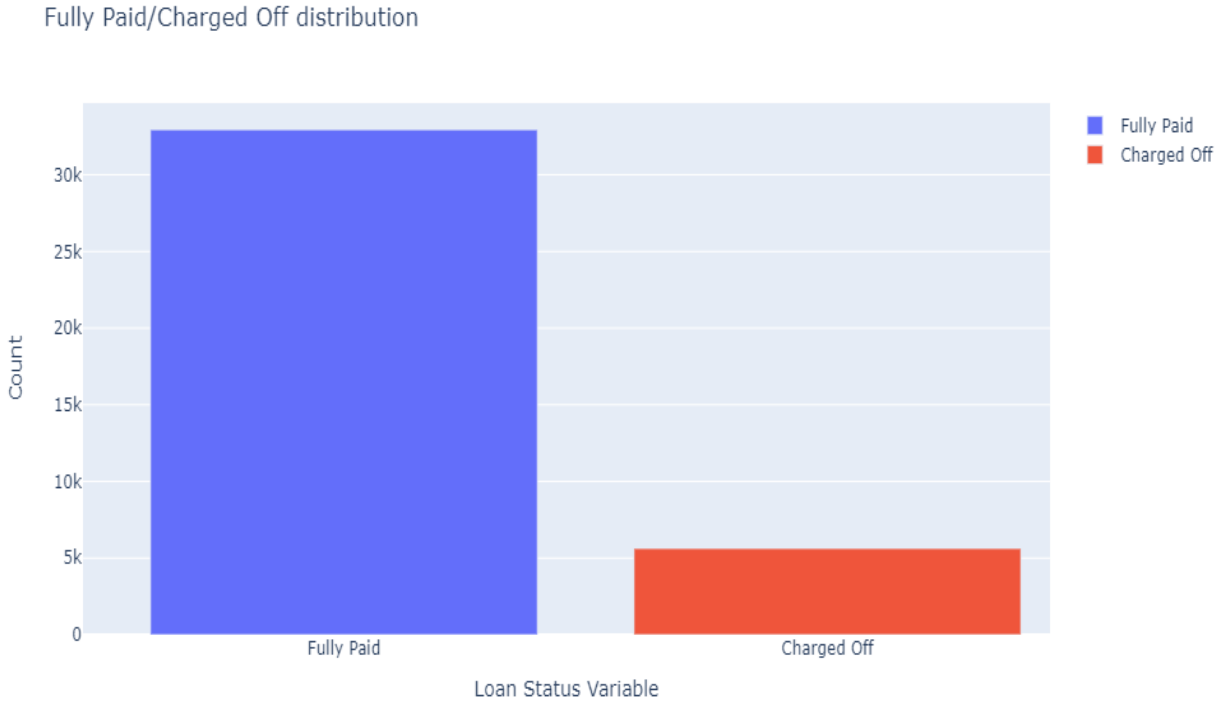
Visualize all outliers for continuous variables



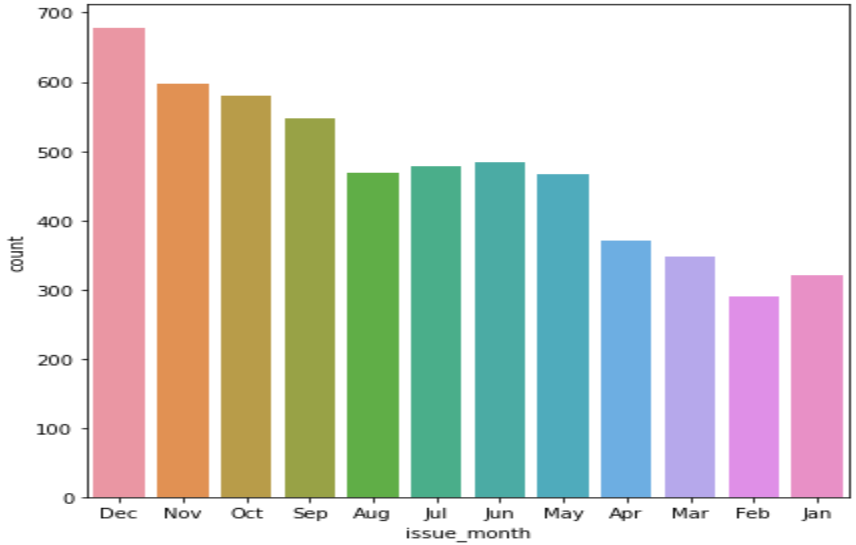
Insights

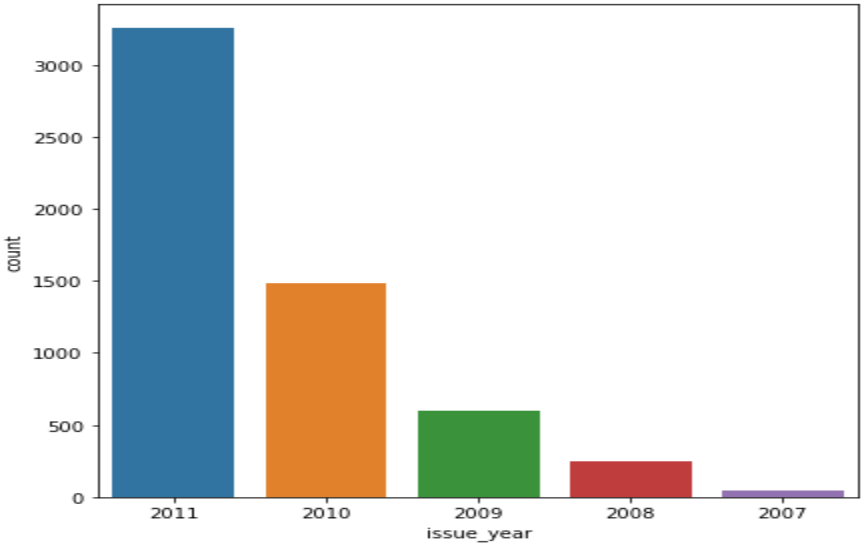
1. Most of the variables have values beyond upper fence.
2. “annual_inc” has clearly 2 outliers with income 60L and 39L. Here we have removed these outliers.

Data Analysis - Univariate Analysis



Insights:
Fully paid applicants are more in number as compared to charged off





Insights

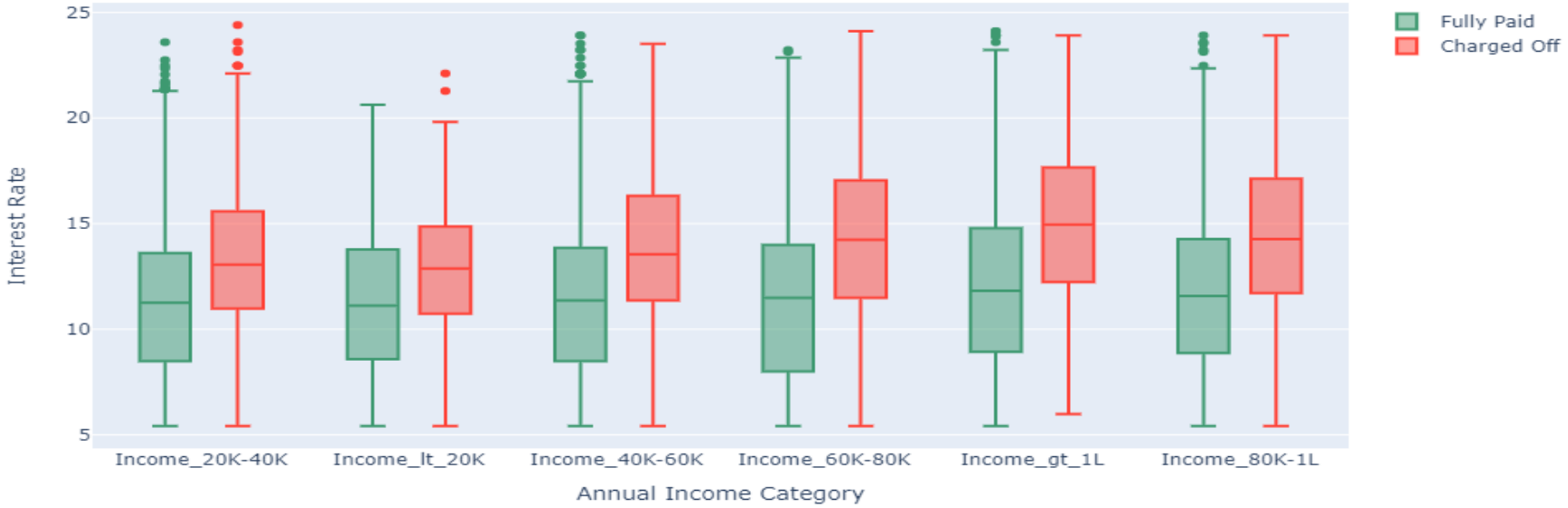
1.Many applicant's loan is issued at the end of year. Increasing pattern can be seen in graph, over the months

2.Agents can try to engage with lenders at the end of year

3.Loan issued in the year 2011 were also as compared to other years

Bivariate Analysis

- Here we choose two or more features to understand the defaulter category



Insights

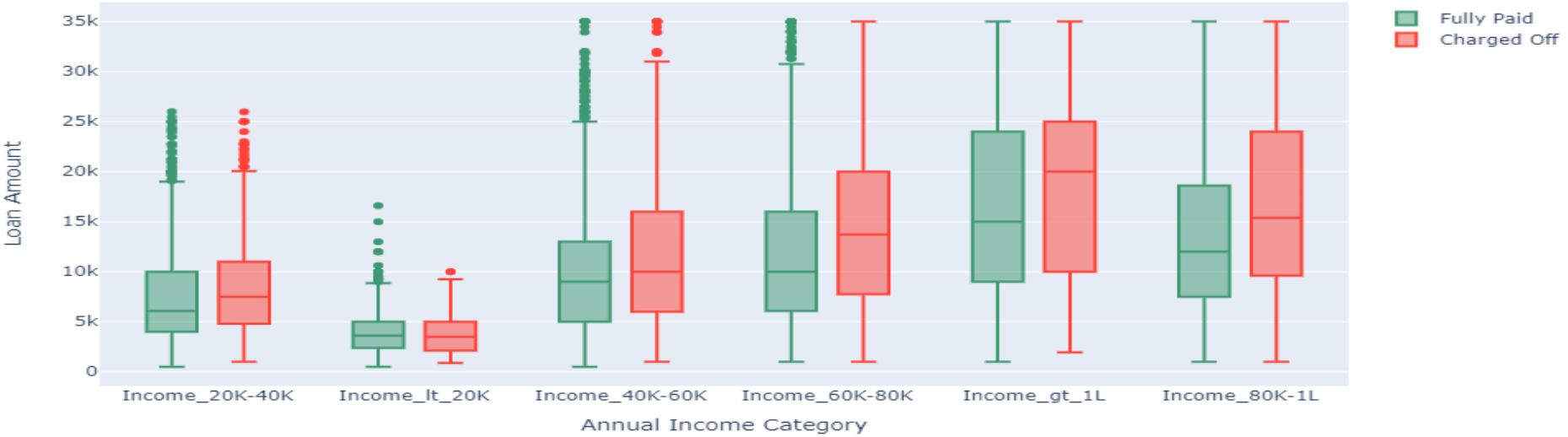
1.For each annual income category if the interest rate goes beyond threshold, it is likely that applicant will default. Ex- Income greater than 1L if interest rate is above 15, chances are there that applicant will default

2.For each annual income category, the median values for fully paid is less than charged off.

3.For each annual income category, the interest rate is around 11 for fully paid and 14 for charged off

Recommendations

1.Interest rate beyond certain threshold for specific income category will result in defaulter category, so it is recommended that interest rate should not be more than threshold



Insights

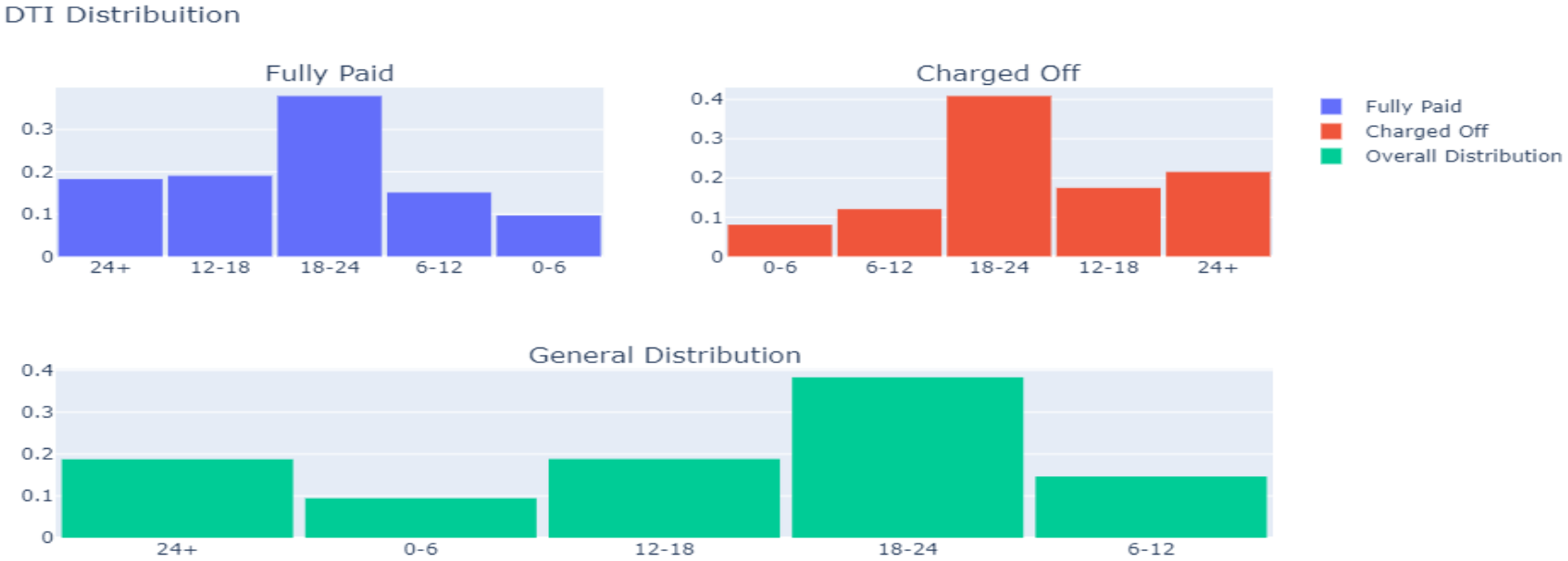
1.There is increasing correlation between income and loan amount.

2.People with more income tend to take more loan amount

3.Observation is that if people take more loan amount, then there are chances that they default. Ex - For income greater than 1L, loan amount of 15K is fully paid but loan amount of 20K is charged off

Recommendations

1.Loan amount sanctioned should not be more than a threshold for specific annual income category

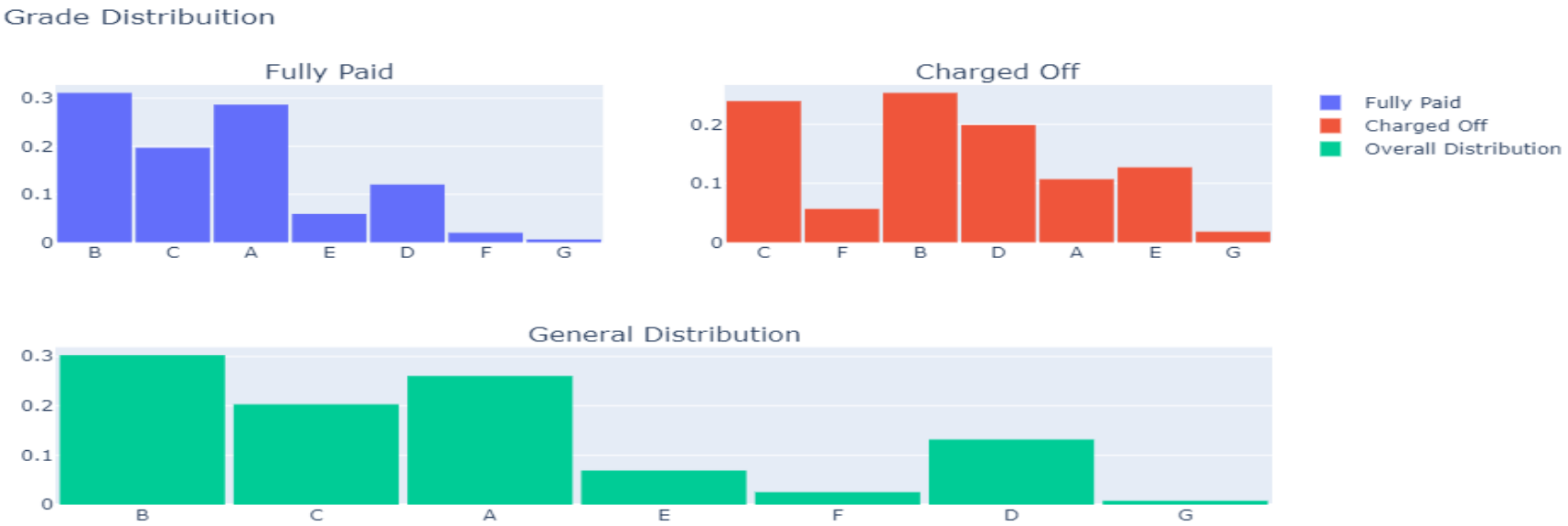


Insights

1.Applicants having dti ratio between 18-24 have chances of getting defaulted.

Recommendations

1.Before giving loan, dti value along with some other parameter should be checked



Insights

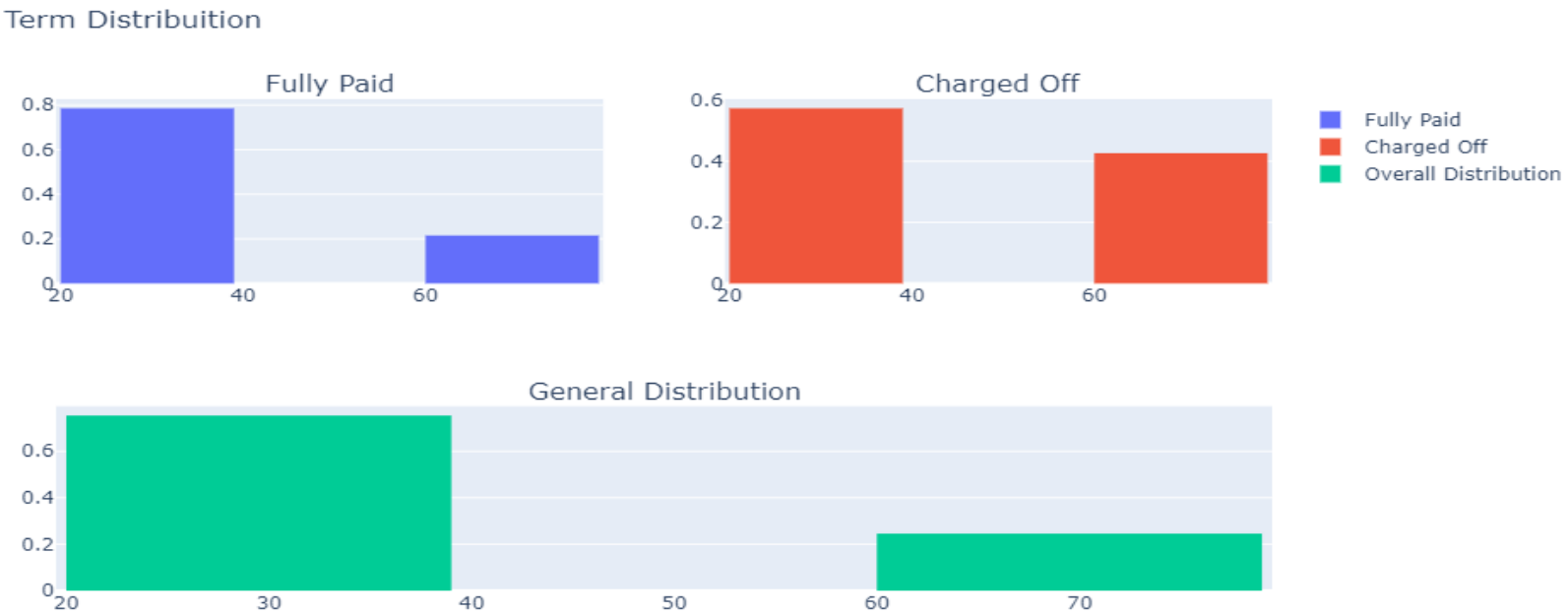
1.Applicants with grade A are less likely to default

2.Applicants having grade with B and C are more likely to default

3.There is normal distribution in data from grade A to G

Recommendation

1.Grade of applicant should be checked along with another parameter to tell if applicant can default or not

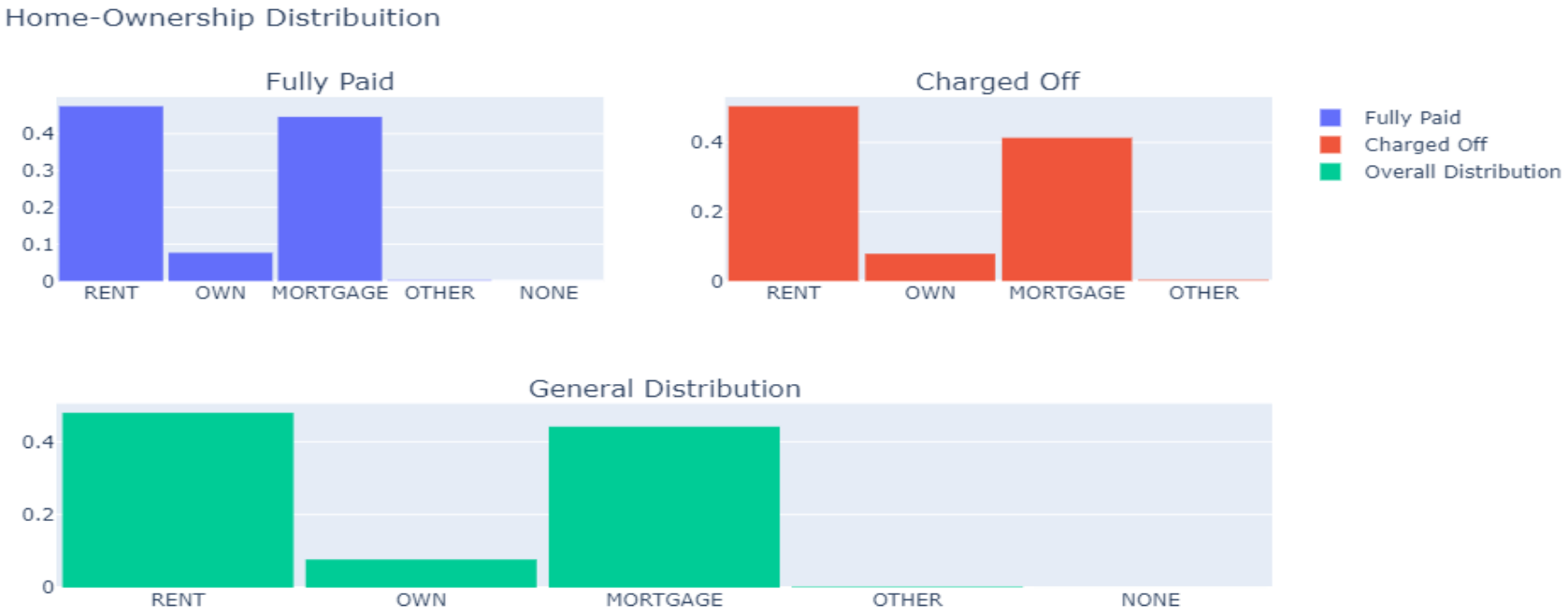


Insights

1.Applicants taking loan for 60-month terms are likely to be defaulted, when compared with fully paid.

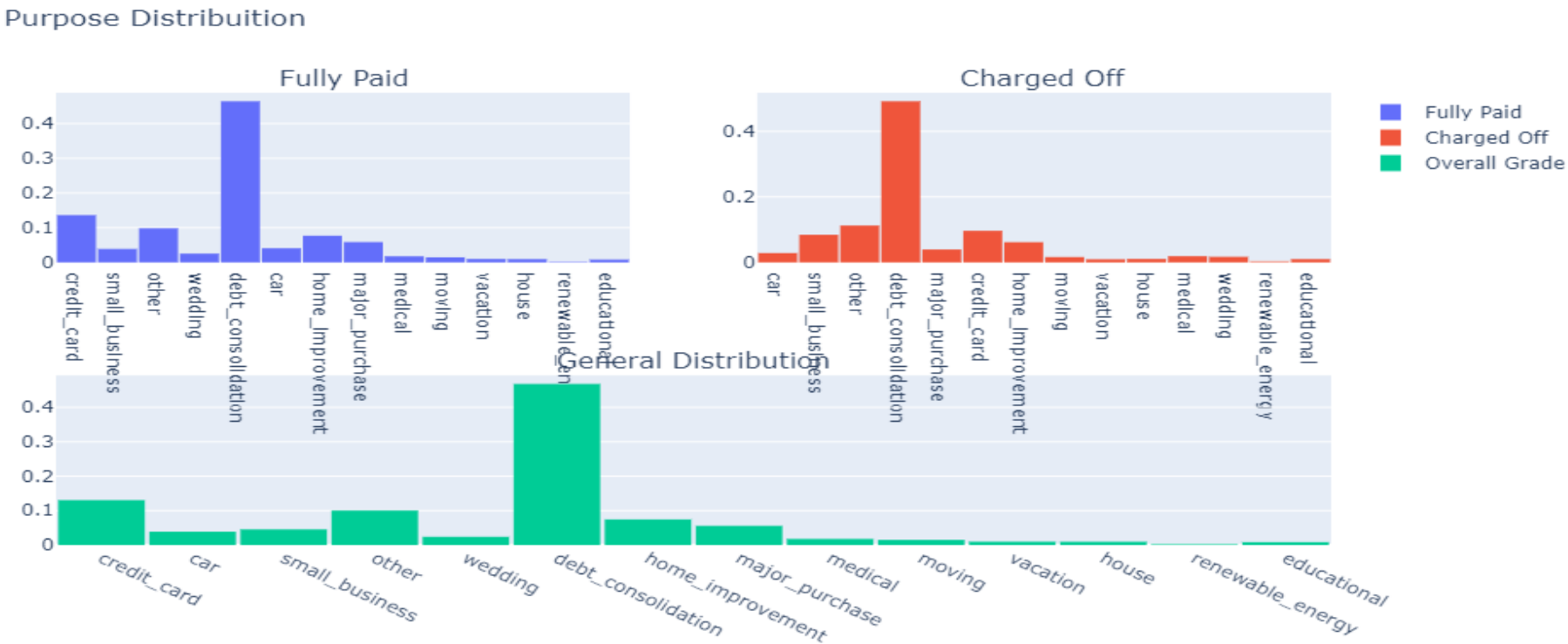
Recommendations

1.Applicants taking loan for 60 months need to verified with some other parameters as well



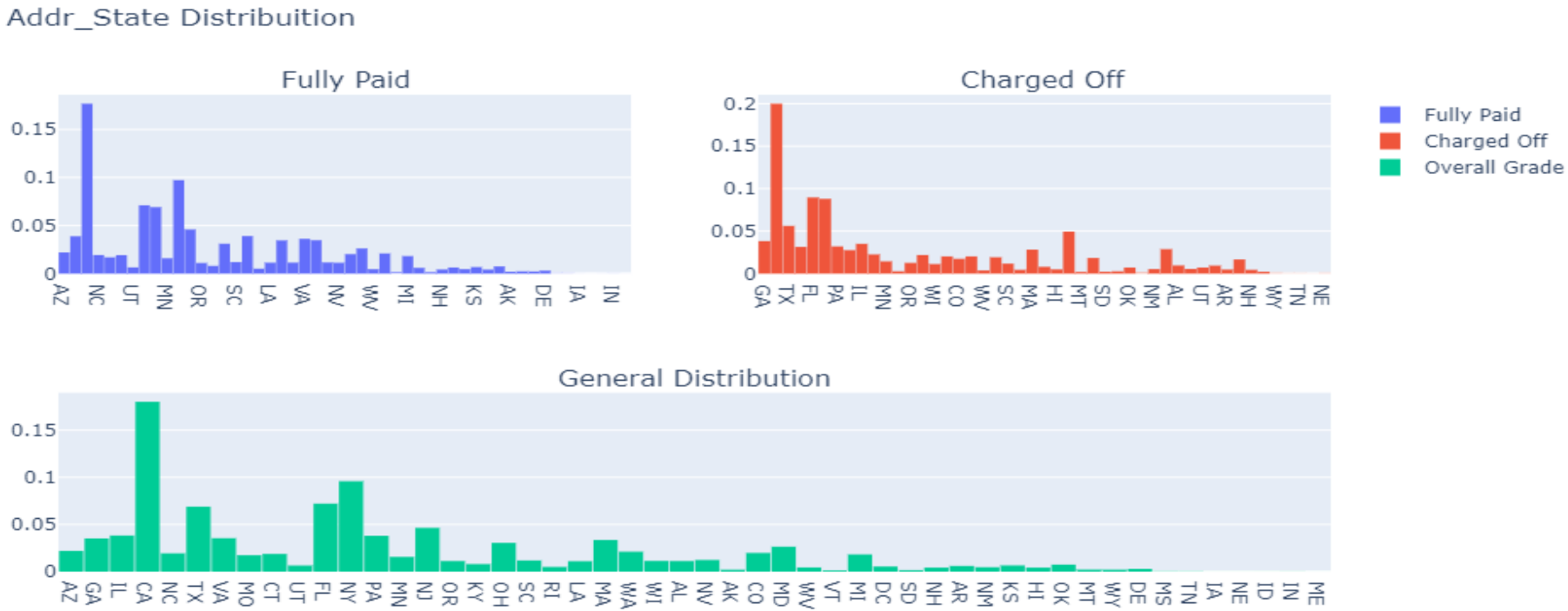
Insights

1.Applicants having “home_ownership” status with rent and mortgage are likely to default as compared to owners



Insights

1.Applicants with loan granted for debt consolidation purpose are likely to be defaulted



Insights

1.Majority of the defaulters can be seen from state CA. Then from NY and FL state.

Five Important Driver Variables -

1. Interest Rate
2. Loam Amount
3. DTI
4. Grade
5. Term

Less impacted variables -

1. Home Ownership
2. Purpose
3. Address State¶

Recommendations

- There is more probability of defaulting when:
  1. Interest rate beyond certain threshold for specific income category
     Ex- Income greater than 1L if interest rate is above 15
  2. Loan amount sanctioned should not be more than a threshold for specific annual income category
     Ex - For income greater than 1L, loan amount of 15K is fully paid but loan amount of 20K is charged off
  3. Applicants having dti ratio between 18-24
  4. Applicants having grade with B and C
  5.Applicants taking loan for 60-month terms
  6. Applicants having home ownership status with rent and mortgage

Thank You!