LENDING CLUB CASE STUDY

Niraj Ram S & Nishu Kumari

Lending Club Case Study

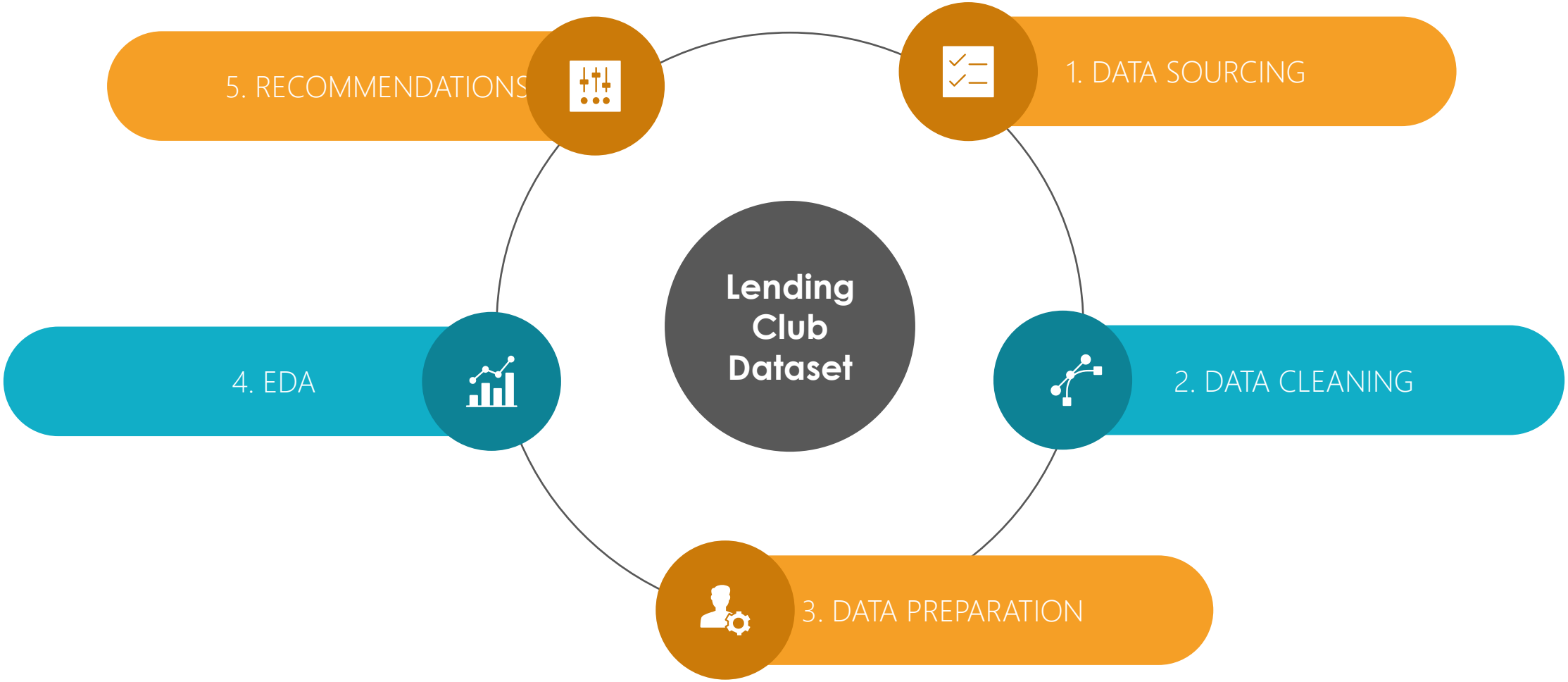
BASIC UNDERSTANDING

- You work for a consumer finance company which specializes in lending various types of loans to urban customers. 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 data given below 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, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.
- When a person applies for a loan, there are two types of decisions that could be taken by the company:
- Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
    - 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
- Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Business Objectives

- This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.
- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.
- In other words, 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.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

Problem Solving Methodology



Data Cleaning

ANALYZE THE DATASET

1. Analyze the Shape of the Dataset
2. Understand info about the dataset
3. Understand the columns present
4. See few rows of the dataset

Rows : 39717
Columns : 111

REMOVE DUPLICATE ROWS

There are no duplicate rows in the dataset

Rows : 39717
Columns : 111

REMOVE EMPTY ROWS & COLUMNS

1. Remove Empty Rows
2. Remove Empty Columns – (54)

Rows : 39717
Columns : 57

REMOVE UNWANTED COLUMNS

1. Remove columns with max empty values
2. Remove columns with all unique or all same values
3. Remove non significant columns

Rows : 39717
Columns : 20

REMOVE ROWS WITH MISSING KEY COLUMN VALUES

1. Analyze rows with missing values for key columns
2. If key columns are missing remove those rows

Rows : 36539
Columns : 20

Data Preparation

CORRECT DATATYPES

All columns Datatype are objects now, change it to required Datatypes

Rows : 36539
Columns : 26

MODIFY DATA FORMAT AS NEEDED

There are several columns like emp_length, term etc where some formatting is needed, analyze each columns and format to get the required data

Rows : 36539
Columns : 26

ADD REQUIRED DERIVED COLUMNS

1. Add any new columns derived from existing columns
2. Add new bucket / bin columns categorizing serios of values as needed

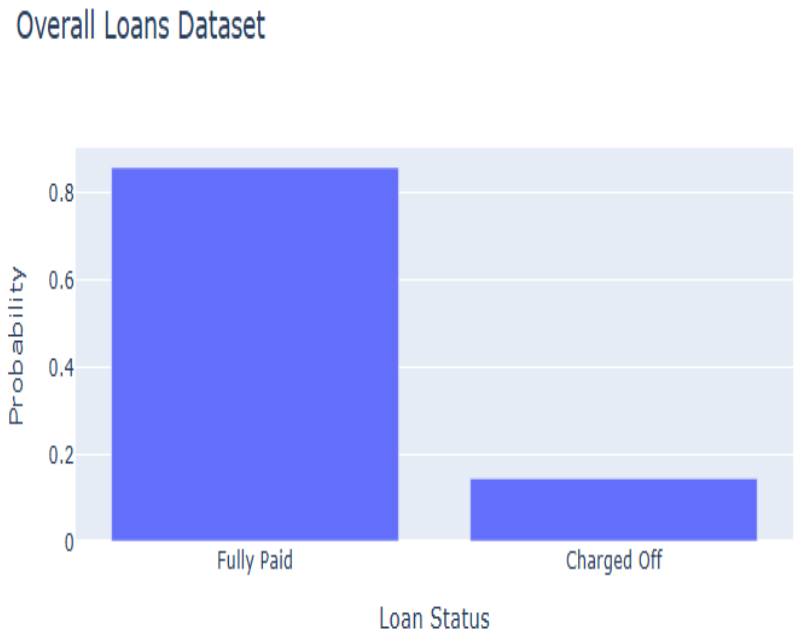
Rows : 36539
Columns : 26

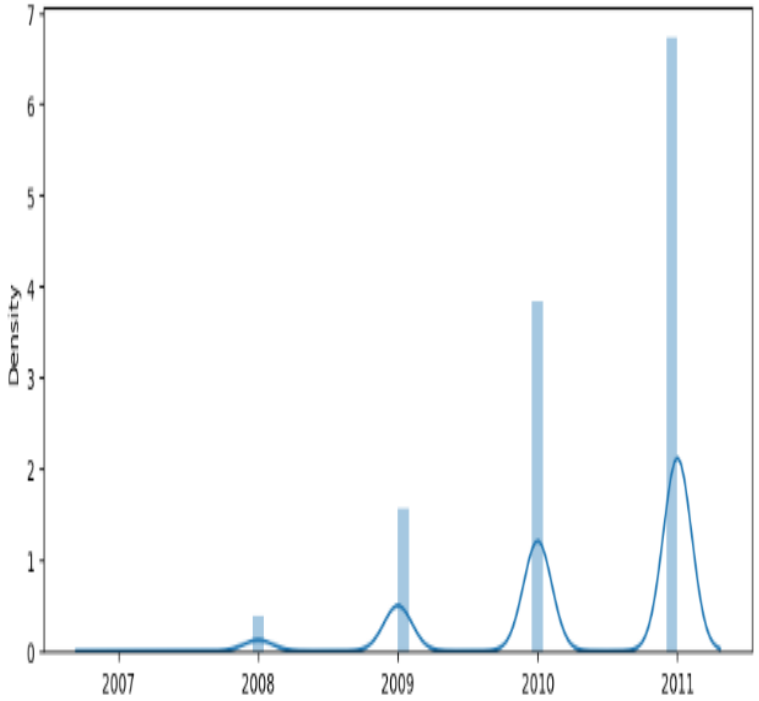
REMOVE OUTLIERS

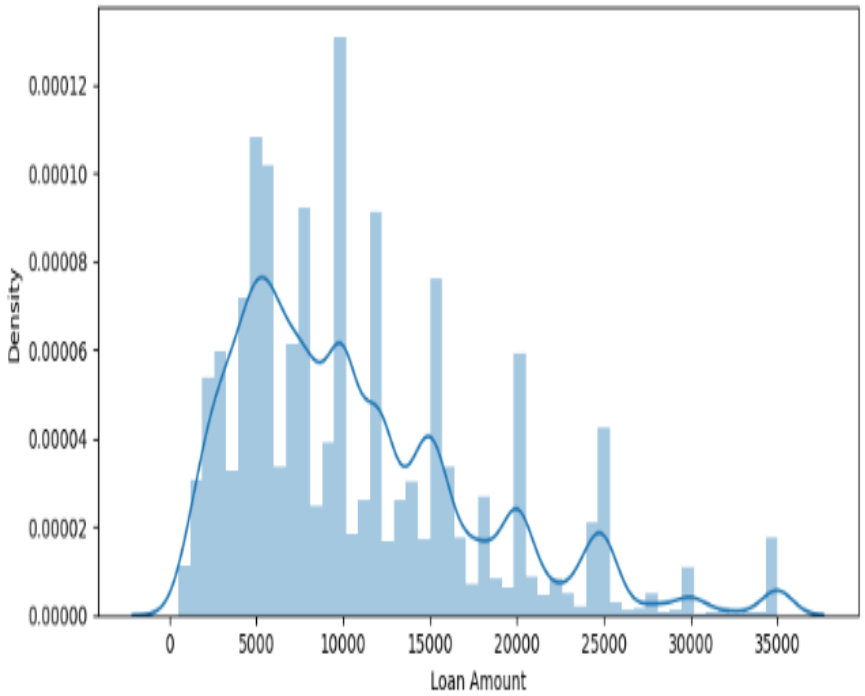
Analyze all the numerical columns using Boxplot and remove any extreme outliers from the dataset

Rows : 33608
Columns : 26

UNIVARIATE ANALYSIS







CHARGED OFF LOAN %

14%

- 14+% of the Loans are Charged Off from our Dataset
- 1 in 7 loans are defaulted

MAXIMUM LOANS

2011

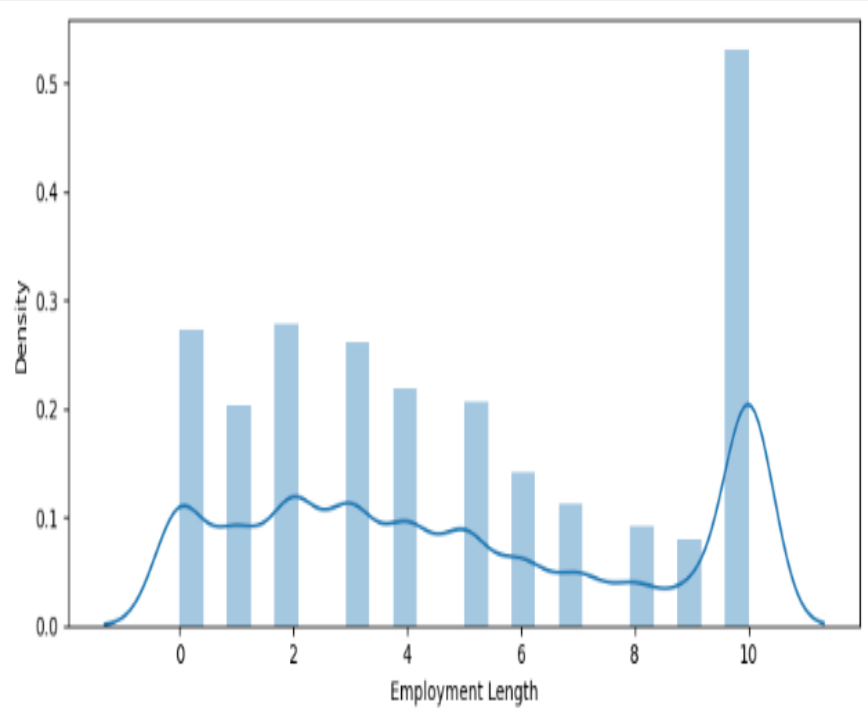
Number of loans given is increasing year on year with 2011 (We have data available till then) being the year with most number of loans given

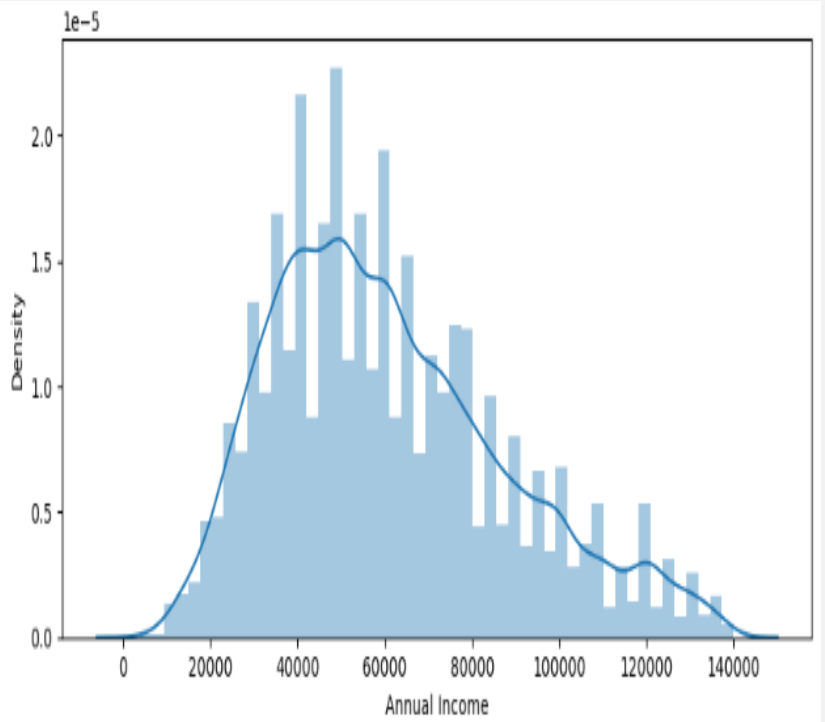
AVERAGE LOAN AMOUNT

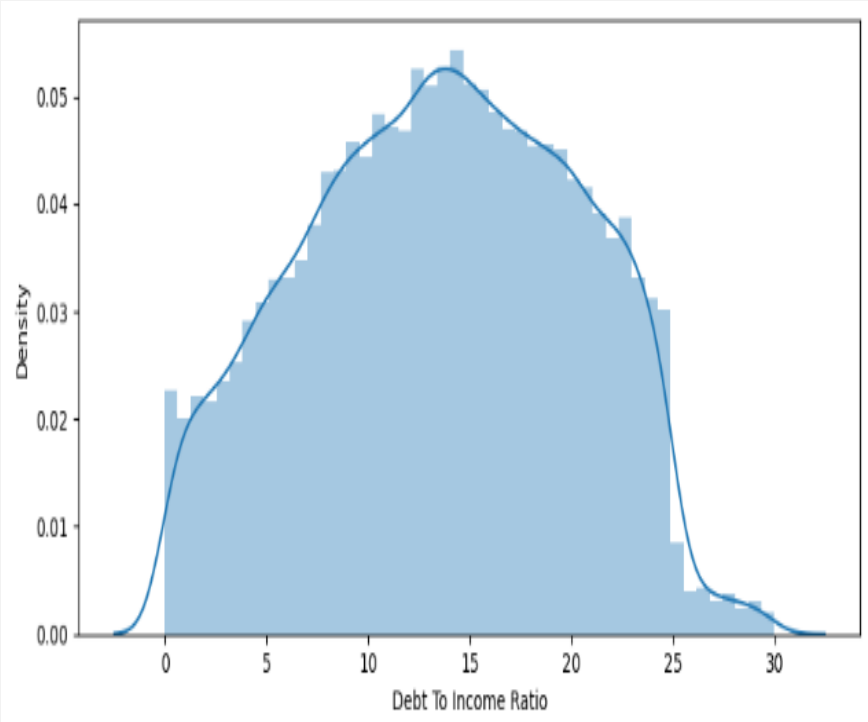
$ 10,000

- Most of the loan amount is between 5K to 15K with average around 10K
- Funded Amount and Funded Amount by Investors is similar to Loan Amount value and has similar distribution

UNIVARIATE ANALYSIS







EMPLOYEE TENURE

10+

- For most of the records employee tenure more than 10 years of experience
- 10+ years has been rounded off to 10 years

ANNUAL INCOME

60K

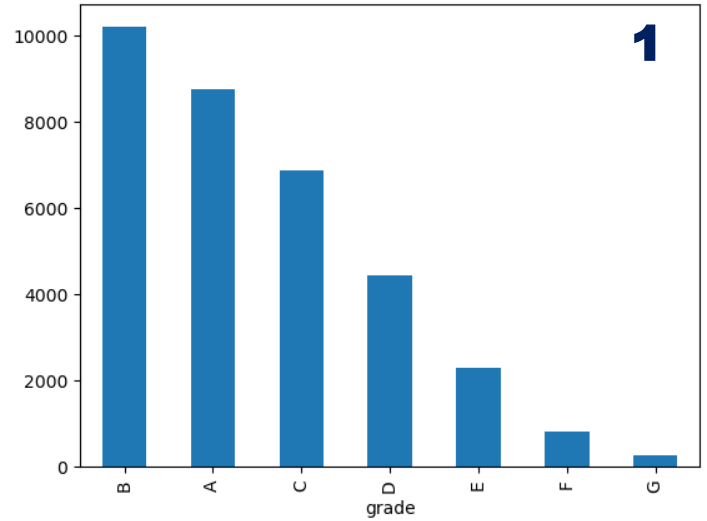
Mean Annual income is around 60K with most of the income ranging between 40K and 78K. This distribution is left skewed normal distribution, hence we can say that majority of the borrowers have very low annual income compared to rest

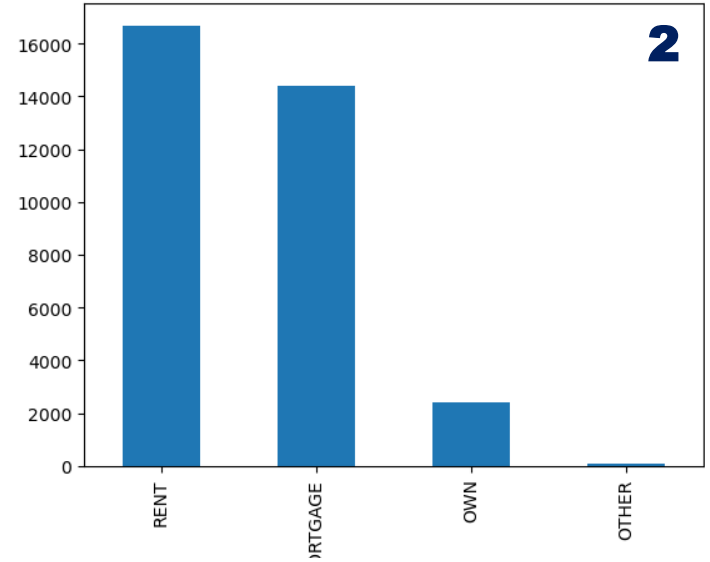
AVERAGE DTI

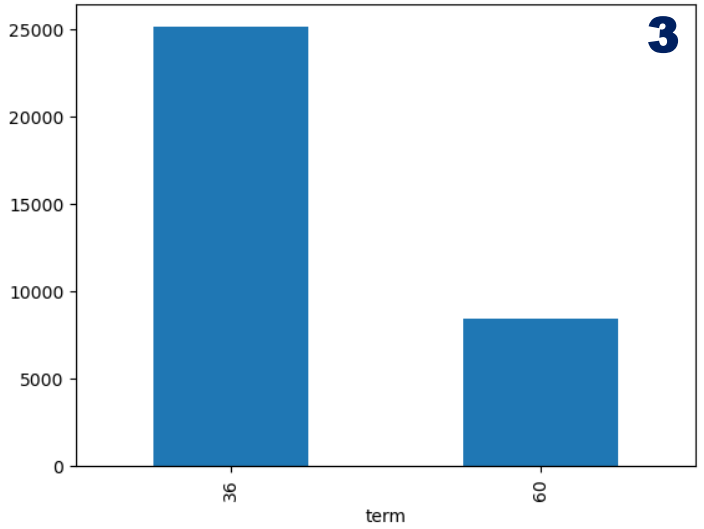
13.57

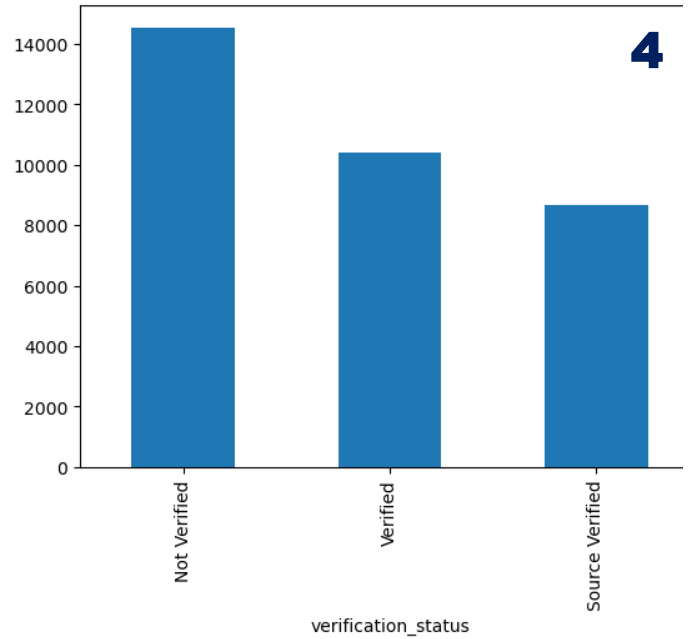
For the loans given Debt to income ratio is around 15, with a mean of 13.57 with just few values beyond the 75 percentile of 18.82

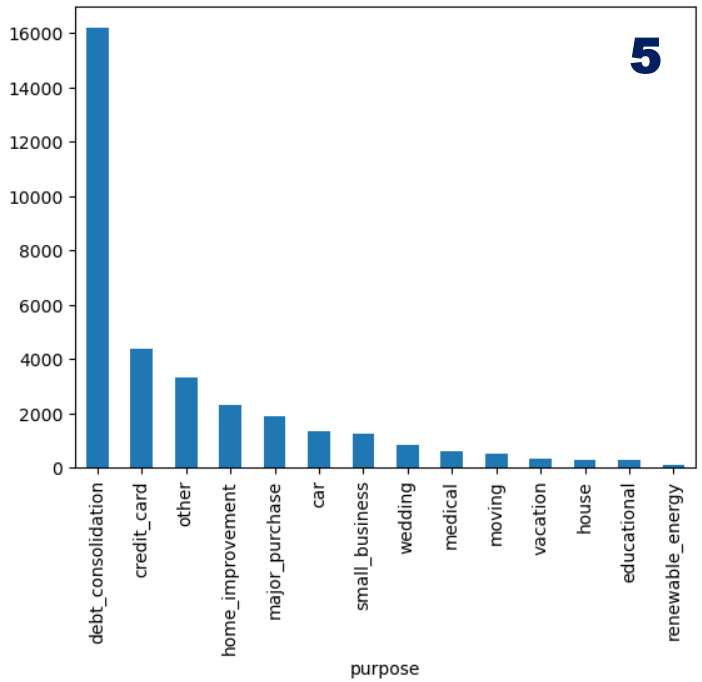
UNIVARIATE ANALYSIS

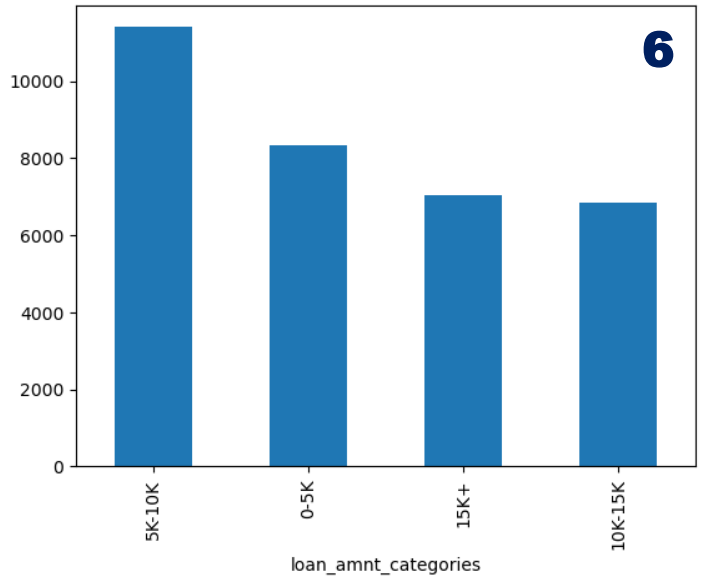








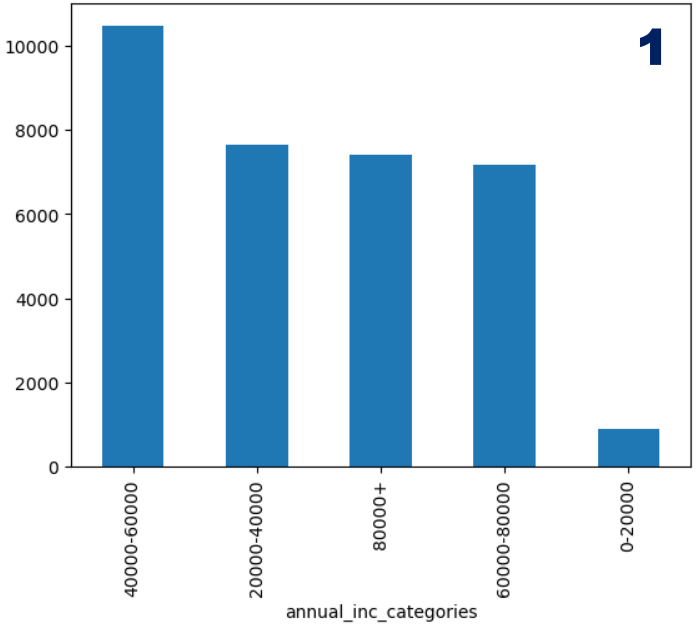


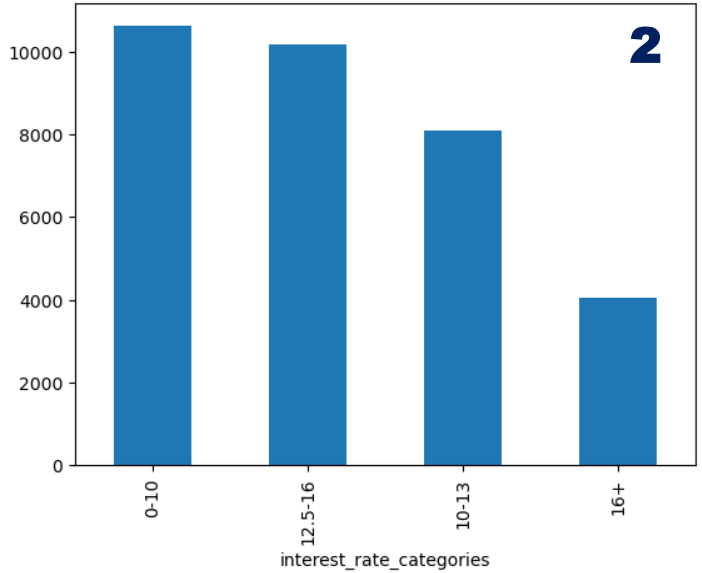


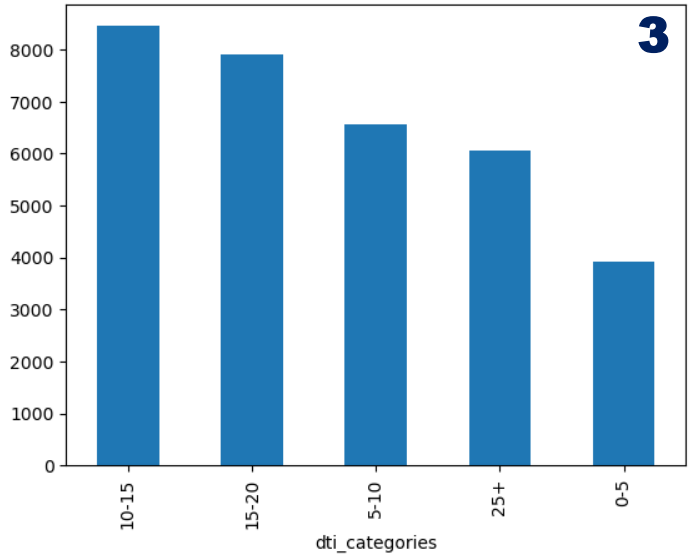
INFERENCES

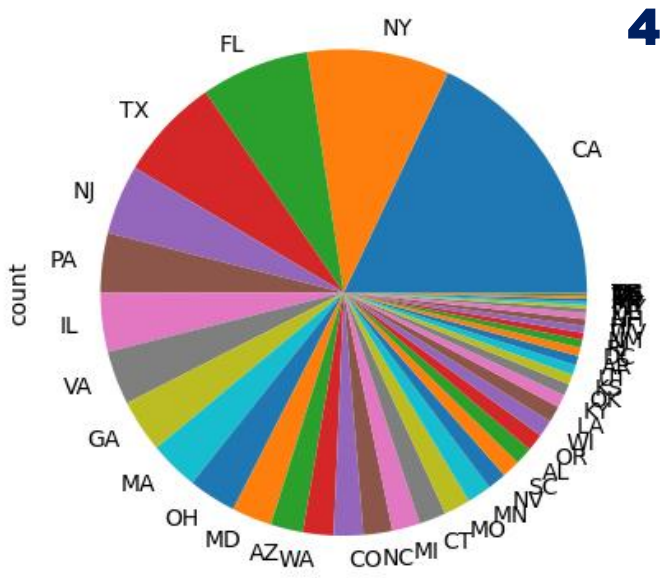
1. Most of the loans are of B category followed by A and C. Number of Risky loans F and G are quite less followed by medium risky loans
2. Most of the loan applicants are either on Rent or Mortgage with very few Owning houses or others, need to further evaluate which category defaults the most
3. Most of the loans have Tenure of 36 months. There are only 2 tenure 36 and 60 months
4. Indicates that for good number of loans Income is not verified by LC, with very few Source verified. May be a driving factor for default loans?
5. Most of the loans are related to debt_consolidation followed by Credit_card. Hence Debt Consolidaiton is the popular loan category
6. Most of the loans are between 0-10K, when compared to loans above 10K

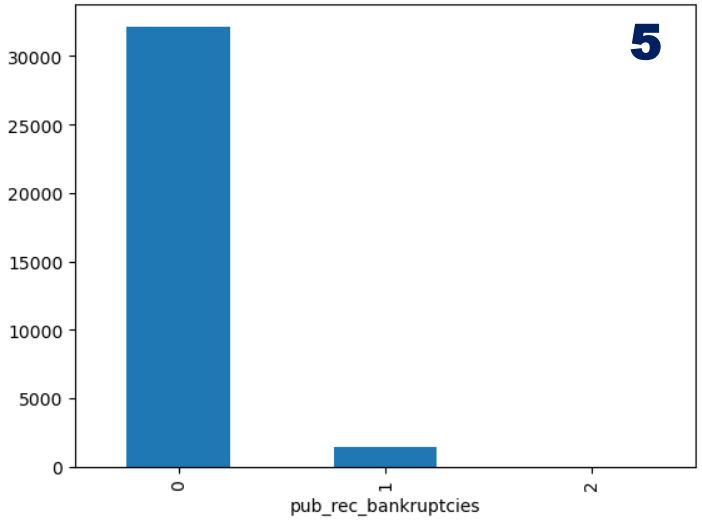
UNIVARIATE ANALYSIS











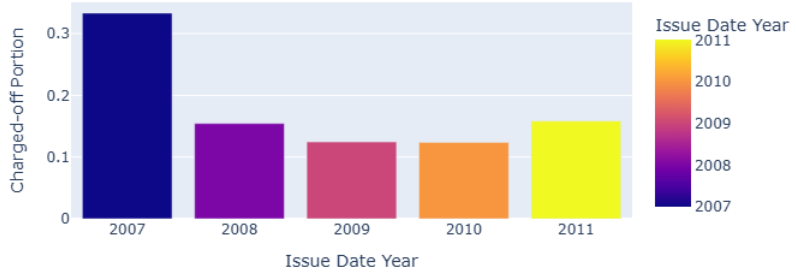
INFERENCES

1. Most of the people who have applied for loans have annual income between 40K to 60K
2. Most of the Loans have interest between 0 to 10 with few having beyond 16+
3. Most of the loans have Debt to Income Ratio of 10-15, followed by 15-20 with least number of borrowers having DTI ratio of 0-5
4. Most of the loans are from California, almost 50% of the loans are from major urban cities like California, NewYork, Florida, Texas, New-Jersey
5. The majority of borrowers have no record of Public Recorded Bankruptcy

BI-VARIATE ANALYSIS

Issue Date Year vs Charged-off Portion

1

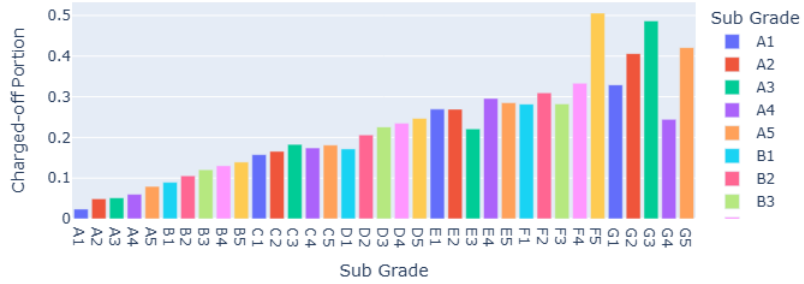


Grade vs Charged-off Portion

2

Sub Grade vs Charged-off Portion

3

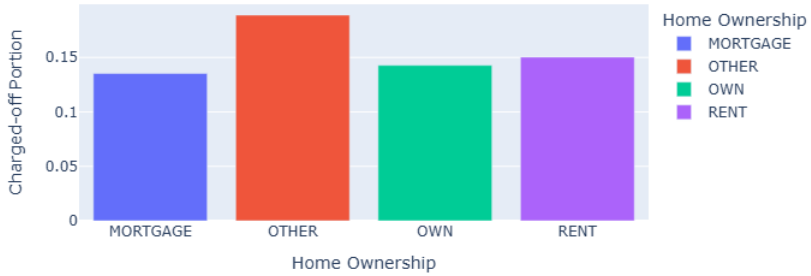


Employee Tenure vs Charged-off Portion

4

Home Ownership vs Charged-off Portion

5



Annual Income Categories vs Charged-off Portion

6

INFERENCES

1. 2007 has maximum percentage of Charged off loans, with 2010 and 2009 having the lowest. However 2007 just had around 6 loans out of which 2 are charged off
2. As expected, Grade A has lowest number of charged off portion which steadily increases as grades range from A to G
3. As expected, Sub Grades of A has lowest number of charged off portion which steadily increases as grades range from A to G with exception being F5 category

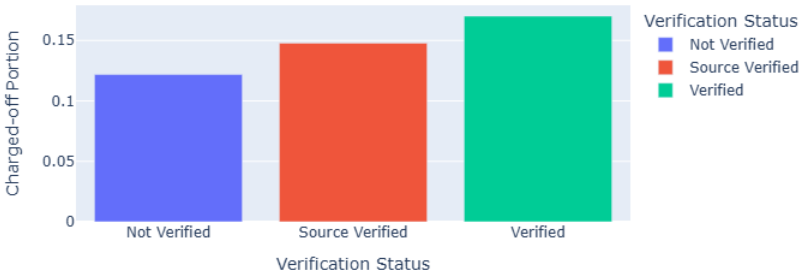
INFERENCES

4. Employee Tenure has not much significance with Charged off loans
5. Home ownership has no impact on default of loan
6. Lower Annual income people tend to get default more than higher income people. As we can see reduction of charged off loans as Annual income increases

BI-VARIATE ANALYSIS

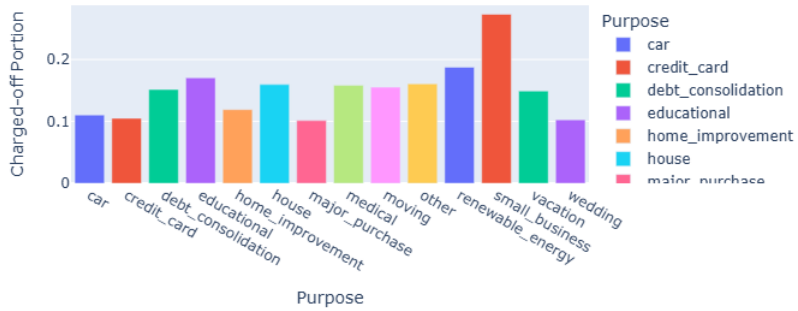
Verification Status vs Charged-off Portion

1



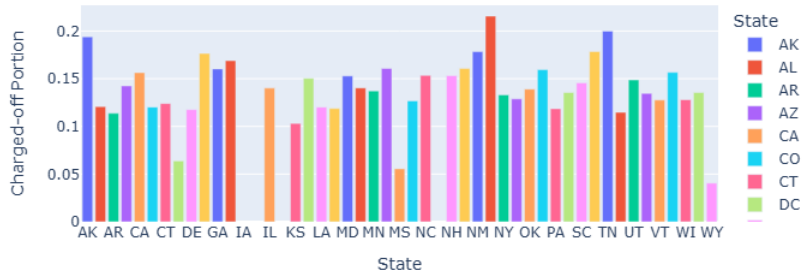
Purpose vs Charged-off Portion

2



State vs Charged-off Portion

3

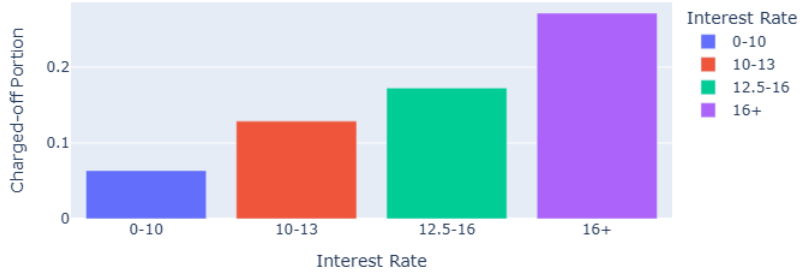


Dept to Income Ratio vs Charged-off Portion

4

Interest Rate vs Charged-off Portion

5

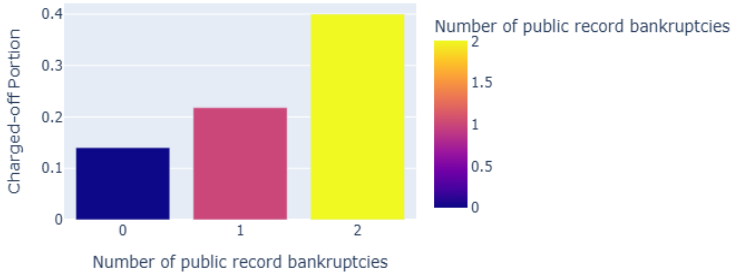


Loan Amount vs Charged-off Portion

6

Number of public record bankruptcies vs Charged-off Portion

7



INFERENCES

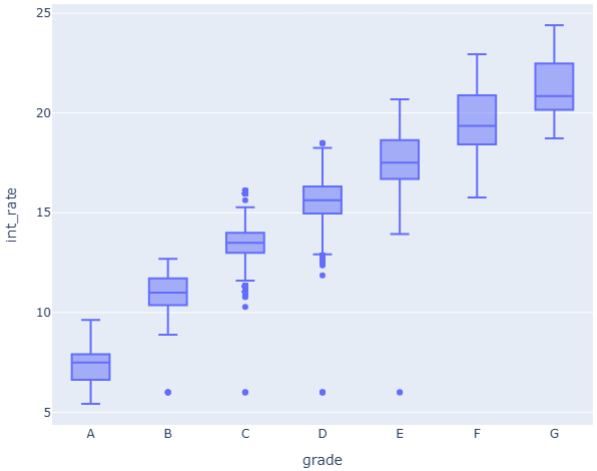
1. Surprisingly, Verified Loans have higher Default Rate
2. When it comes to Purpose of Loan, Small business loans are quite risky and has maximum percentage of Charged off portion with major purchase being the lowest
3. Urban Cities have relatively lower default rate

INFERENCES

4. As the DTI increases, Charged off percentage of loans also increases
5. There is a clear indication that higher the interest rate, charged off probability increases, we also have to understand in what cases interest rates will be higher
6. Higher the Loan Amount, probability of Charge off increases
7. If the public record bankruptcies increases then probability of Charge off increases as well and is very minimal when there are no public recorded bankruptcies

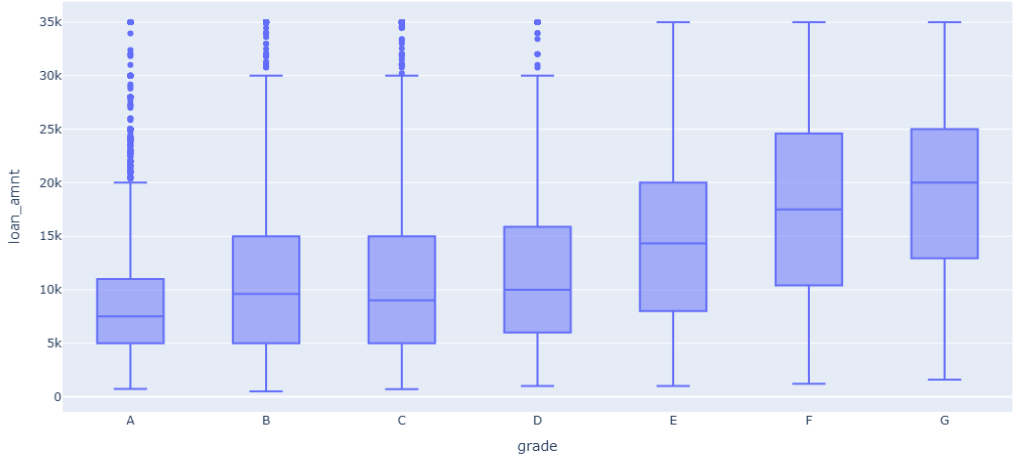
BI-VARIATE ANALYSIS

grade vs int_rate



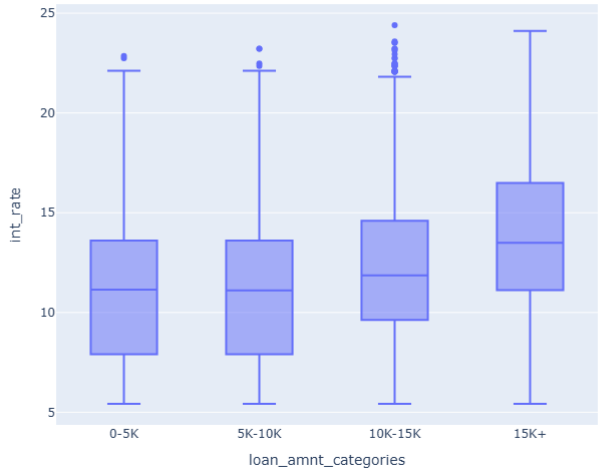
INFERENCE : Interest Rate is inversely proportional to Grade and this makes perfect sense

grade vs loan_amnt



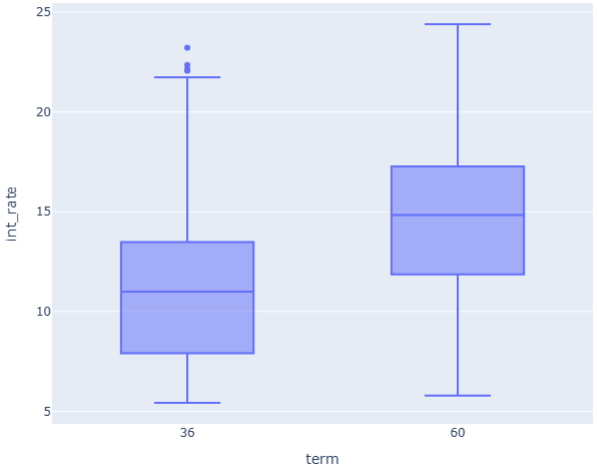
INFERENCE : Surprisingly Lower the Grade higher is the average loan amount. Which should ideally have been vice-versa

loan_amnt_categories vs int_rate



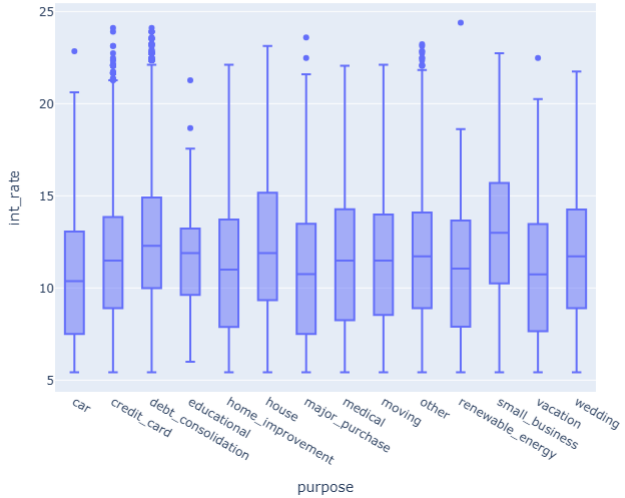
INFERENCE : As the loan amount increases, interest rate also increases

term vs int_rate



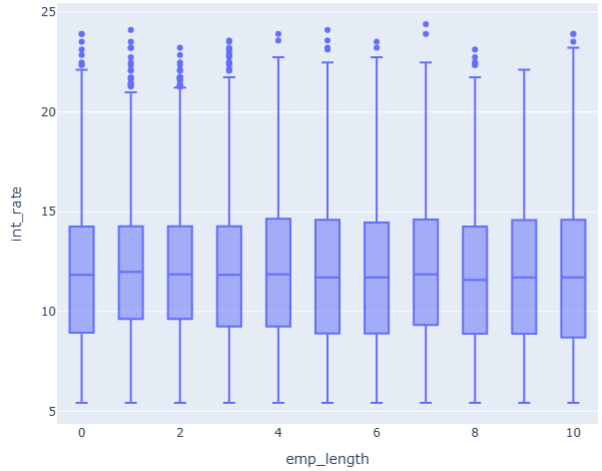
INFERENCE : As the term increases, Interest rate also increases

purpose vs int_rate



INFERENCE : Small business and debt_consolidaiton loans have higher interest rates

emp_length vs int_rate

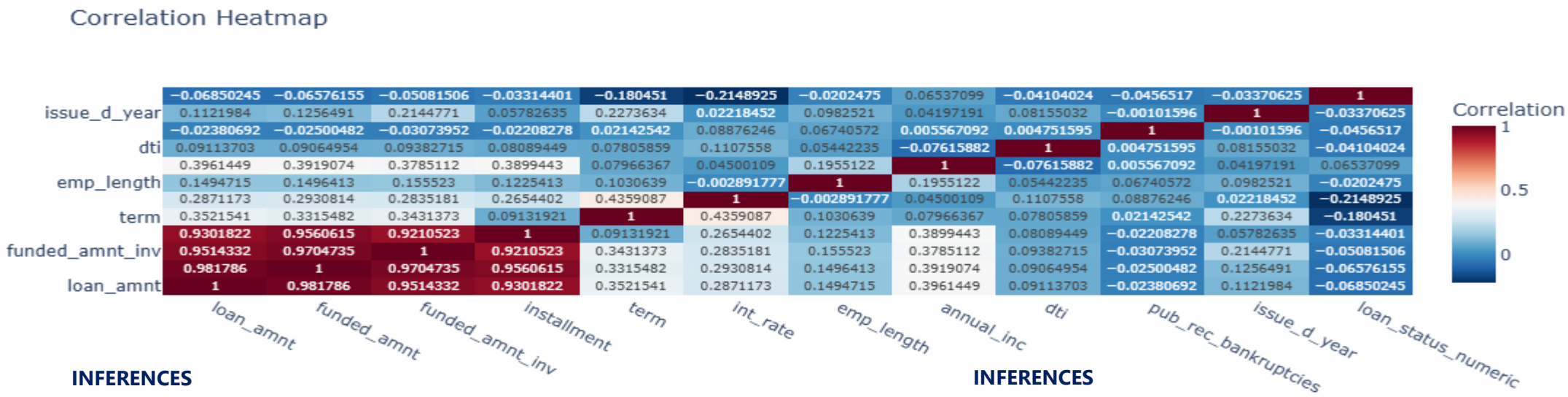


INFERENCE : Employee Tenure has very minimal to no impact on interest rate

BI-VARIATE ANALYSIS SUMMARY

- With the previous Inferences we can conclude that there are some key columns which drives the Charged Off Probability
  - Grades
  - Annual Income
  - Purpose
  - DTI
  - Interest Rate (Which in turn is decided by Grades, Annual Income, Purpose, Term)
  - Public Recorded Bankruptcies
  - Loan Amount (Very mild correlation)
- In other words, Defaults can be those who Have
  - Annual Income between <60K
  - High DTI Ratio >15
  - Grades not in A Or B
  - Public Recorded Bankruptcies
  - Higher Interest Rate >10
  - Higher Loan Amount >10K
  - Purpose of Small Business
  - Non Urban Residents

MULTI-VARIATE ANALYSIS



INFERENCES

- Basic Inferences
  - Loan Amount has very strong positive correlation with Funded Amount, Funded Amount committed by Inv and Installment
  - Loan Amount has good positive correlation with Annual Income, term
  - Number of public record bankruptcies, but it is negligible
  - Term has moderate positive correlation with Interest rate
  - DTI has moderate positive correlation with Interest Rate

INFERENCES

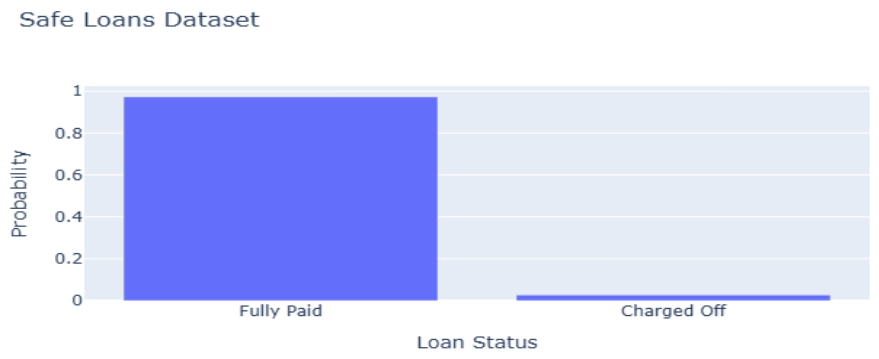
- Loan Status Inferences
  - Loan Status Numeric Value of 1 represents fully Paid and 0 represents loans that are Charged Off
  - Positive Correlation: Variables that are positively correlated with loan_status_numeric have higher values when loans are "Fully Paid". For instance, if annual_inc shows a positive correlation, it means higher annual incomes are associated with a higher likelihood of the loan being fully paid.
  - Negative Correlation: Variables that are negatively correlated with loan_status_numeric have higher values when loans are "Charged Off". For example, if int_rate shows a negative correlation, it means higher interest rates are associated with a higher likelihood of the loan being charged off.
  - From the Map above it shows
    - DTI to Loan Status correlation is negative, Borrowers with higher debt-to-income ratios are more prone to default
    - Annual Income to Loan Status correlation is positive, Borrowers with higher incomes tend to have better loan repayment performance
    - Loan Amount to Loan Status correlation is Negative, Borrowers with larger loans might find it more challenging to repay, leading to defaults
    - Interest Rate to Loan Status correlation is Negative, loans with higher interest rates might be riskier and more difficult for borrowers to repay
    - Term to Loan Status Correlation is Negative, loans with higher Term might be riskier leading to more defaults
    - Public Recorded Bankruptcies to Loan Status correlation is Negative, which indicates loans with pre recorded Bankruptcies tend to default more

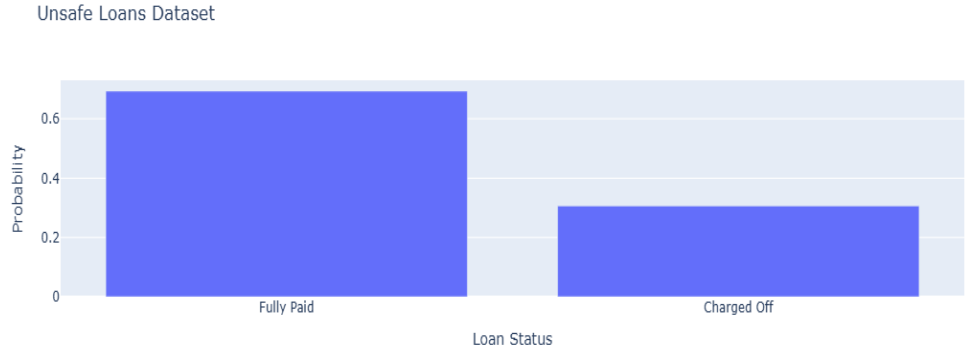
Closing Inference

SAFE LOANS DATASET

UNSAFE LOANS DATASET

PROBABILITY GRAPH





INFERENCE

CONDITIONS & INFERENCES – (1 in 33 Default)

- Conditions : DTI <15, Annual Income > 60000, Grade in A/B, Recorded Bankruptcies =0, interest rate < 10, Loan Amount <10K, State in Major urban cities (CA, NY, FL,TX,NJ,PA) and loan purpose is not for Small Business
- Inference :
  - Earlier from our Dataset we had staggering 14.5% Charged off loans
  - Where as from our Safe Loans Dataset just 3% loans are Charged Off. Which is 1 in 33 Loans

CONDITIONS & INFERENCES – (1 in 2.5 Default)

- Conditions – Grade in E, F, G, Recorded Bankruptcies, Interest rate >10, dti>15
- Default Rate will be 41% with above conditions. So, these are the major driving factors
- If more conditions are added like annual income <60K, non urban borrowers and small business borrowers then . Charged off percentage becomes 33% (1 in 3)

Closing Inference

DEFAULT LOAN PATTERNS – INDICATORS OF DEFAULT

Very High Contributors

- Lower Grades - Grades in E, F, G
- Earlier Bankruptcies - Previously Recorded Bankruptcies
- Higher Interest Rate – Interest Rate >10
- High Debt to Income Ratio - DTI > 15

High Contributors

- Lower Annual Income - <60K
- Higher Loan Amount - >15K
- Non Urban Borrowers – Not in California, NewYork, Florida, Texas, New-Jersey
- Small Business Loans

OTHER INFERENCES

Surprising Inferences - Immediate Actions

- Currently Bank gives Higher Loan Amount on an Average to Grades in G, F, E in the same order than for A, B and C, which needs to be evaluated
- Verified Loans and Source Verified Loans have higher percentage of Charged Off Loans, process of verification needs to be relooked

What is good and should be continued

- Interest Rate is Inversely proportional to the Grade
- As the Loan Amount increases Interest Rate also increases
- Small Business have Higher Interest Rates
- Interest Rate increases with Term

Thank You