# Lending Club Case Study

## Exploratory Data Analysis

Ammula Sivani
Rentala Somanadh

# List of contents

- Problem Statement
- Data Cleaning
- Derived columns
- Data Conversion
- Univariate Analysis
- Bivariate Analysis
- Correlation
- Conclusion

# Problem Statement

- We work for a consumer finance marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.
- It specialises 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.
- Like most other lending companies, *lending loans to 'risky'* applicants is the largest source of financial loss *(called credit loss)*. The 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'*.
- The core objective of the exercise is to **help the company minimise the credit loss**. There are two potential sources of **credit loss** are:
- Applicant **likely to repay the loan**, such an applicant will bring in profit to the company with interest rates.** Rejecting such applicants will result in loss of business**.
- Applicant **not likely to repay** the loan, i.e. and will potentially default, then approving the loan may lead to a financial loss* for the company

# Data Cleaning

- The loan.csv file has 39717 rows and 111 columns
- The files does not have any duplicate files
- It had 1140 rows with loan status as current which are deleted as they are irrelevant to use case, post which it had 38577 rows and 111 columns
- Dropped all the columns with more than 65%  null values in them, leading to 38577 rows and 55 columns
- Dropped removing id, URL and member_id rows because they are unique values as mentioned in the data dictionary and we cannot draw any correlations
- Dropped zip code row as we are considering the state the applicant can live and both of them serve the same purpose
- Dropped emp_title, title and desc rows because they cannot be used in analysis because they very wide spread/ very explanatory
- Dropped all the below rows "earliest_cr_line","inq_last_6mths","last_pymnt_d","application_type","out_prncp","out_prncp_inv","total_rec_prncp","total_rec_int","total_rec_late_fee",'collection_recovery_fee',"last_pymnt_amnt","last_credit_pull_d","recoveries","delinq_2yrs",'total_pymnt','total_pymnt_inv', 'funded_amnt', 'funded_amnt_inv' as these values are calculated post loan approval and do aid in making decision if loan must be approved or not

# Data Cleaning

- Post removal of all the above columns, we are left with 30 columns
- Dropped Pymnt_plan', 'initial_list_status' ,'collections_12_mths_ex_med', 'policy_code', 'acc_now_delinq', 'chargeoff_within_12_mths', ‘delinq_amnt', 'tax_liens’ columns as they have one single value and do not act as differentiators. Post this we are left 22 rows
- Emp_length column had 1033 rows with Null values in it, which were dropped. Post which we were left with 37544.
- Pub_rec_bankruptcies had 697 rows with null values which were dropped.
- Revol_util had 47 rows with null values in it which were dropped.
- Post the above rows drop we are left with 36800 rows.

# Derived Columns

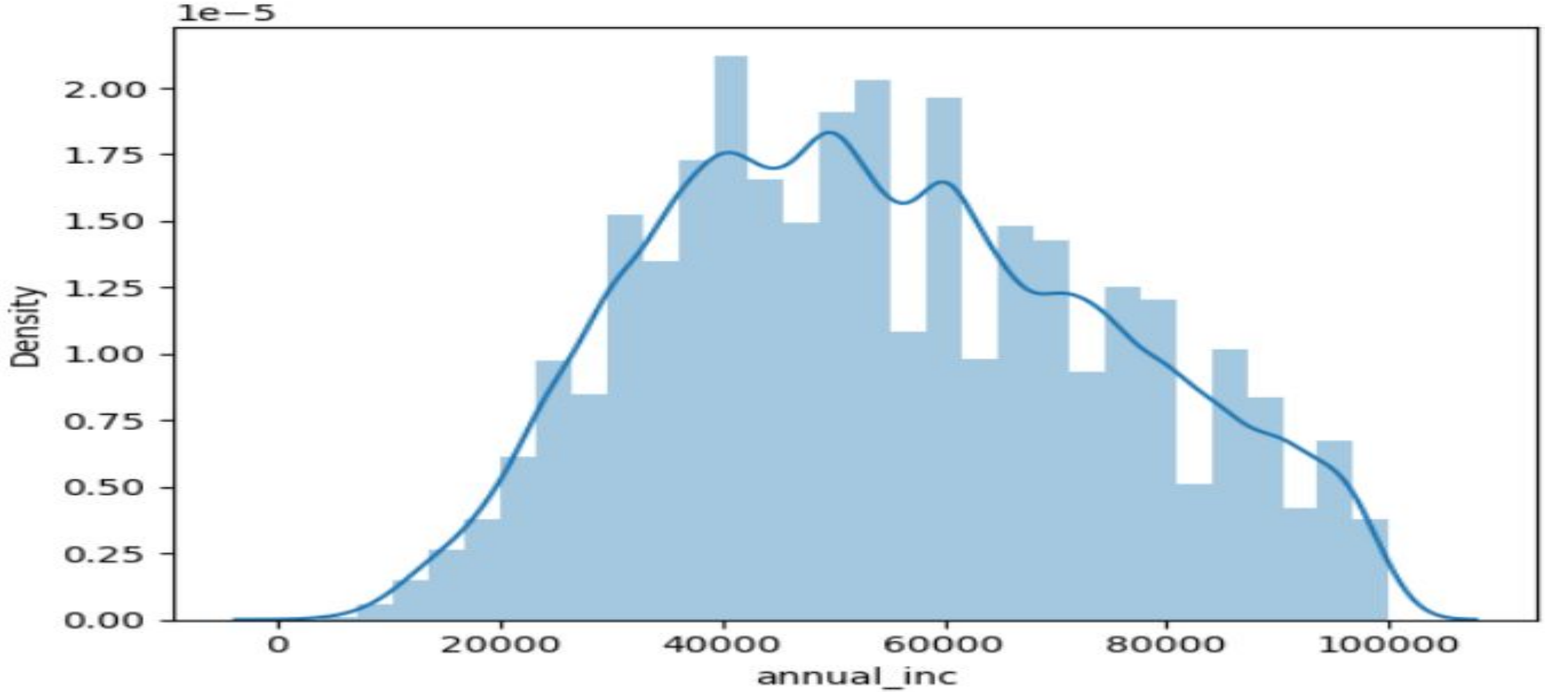
- From Issue_d column, 2 columns issue_d_year and issue_d_month were derived where issue_d_year represents loan issued year and issue_d_month represents loan issued month.
- Post this we are left with 36800 rows and 23 columns

# Data Conversion

- The row Term was of object and had the string “ month” in it. Post removal of the word “month” it was converted into int type.
- The columns had Int_rate, revol_util were of object type as they had % in it. Post removal of that, they were converted into float type.

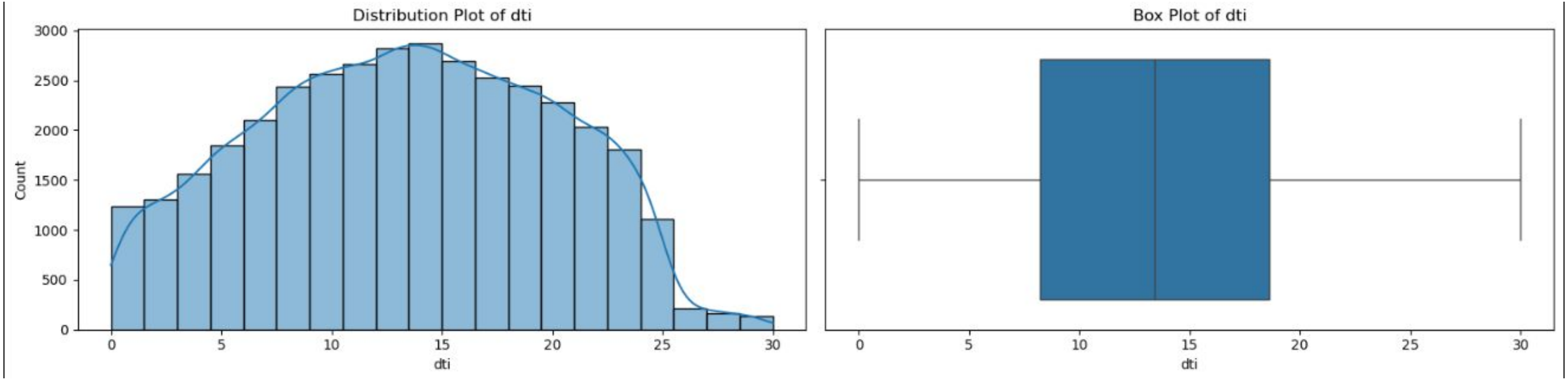
# Univariate Analysis

# Univariate Analysis of Numerical variables



Annual Income mostly lied in the range of 40-75k

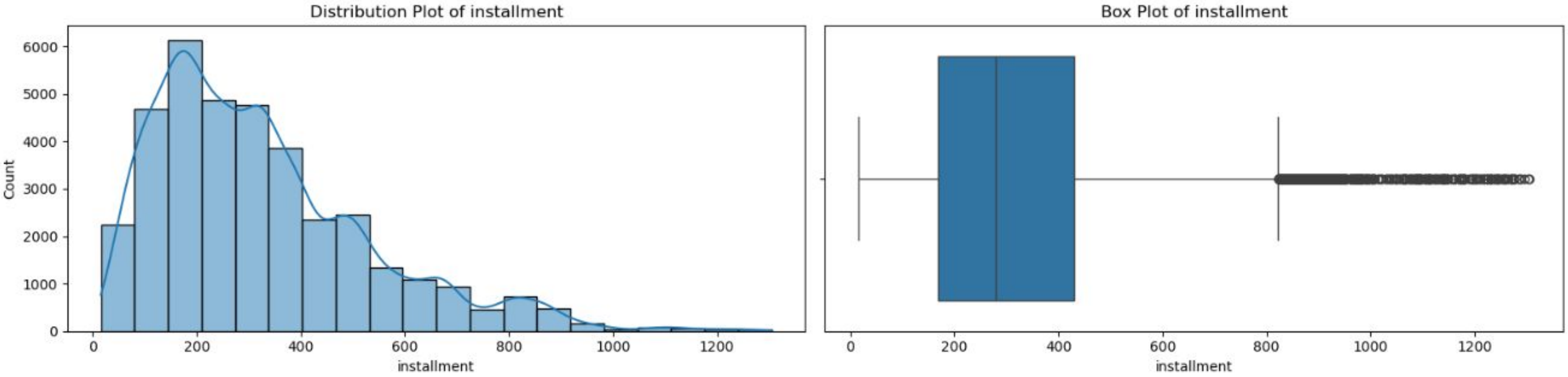
# Univariate Analysis of Numerical variables


Distribution Plot of dti

Box Plot of dti

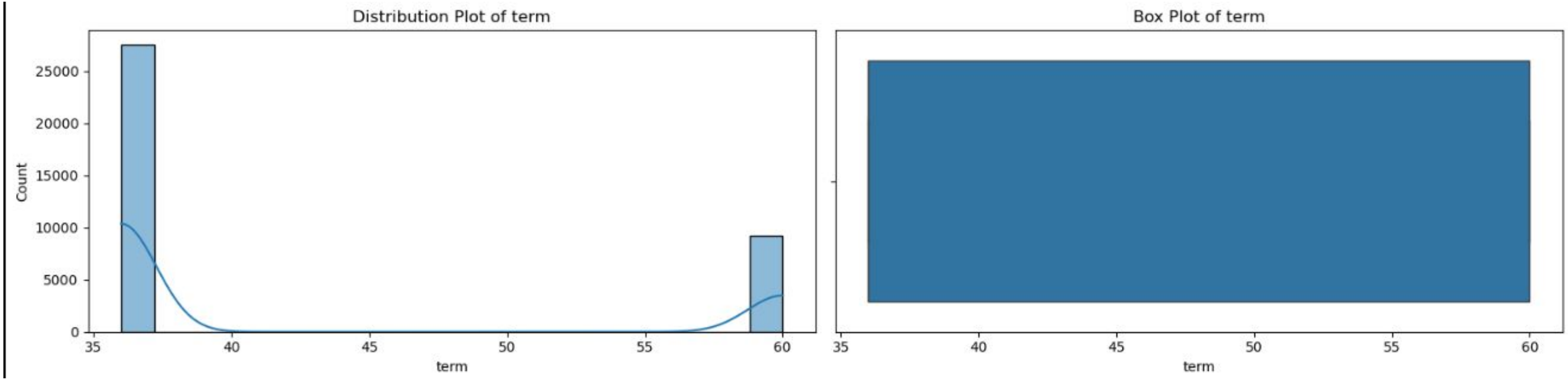
- Most of the DTI ranges from 0 to 20 and ,max being at 30

# Univariate Analysis of Numerical variables



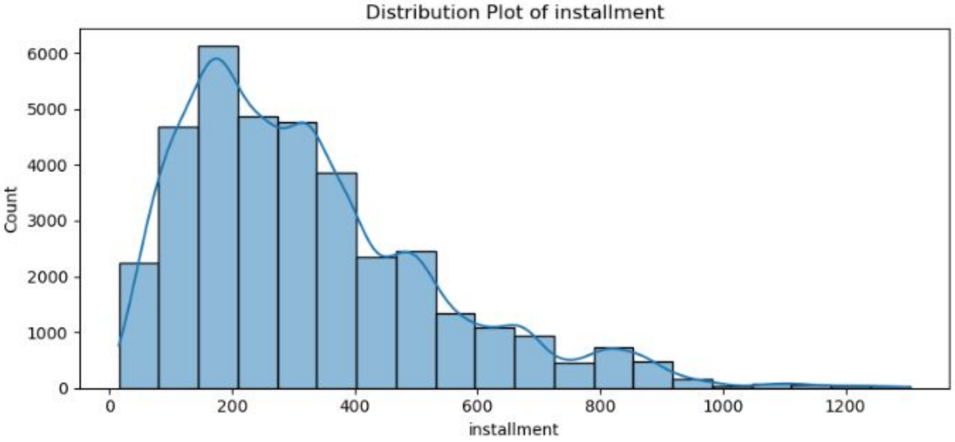
Installments range from 20 to 400 and maximum being at 700

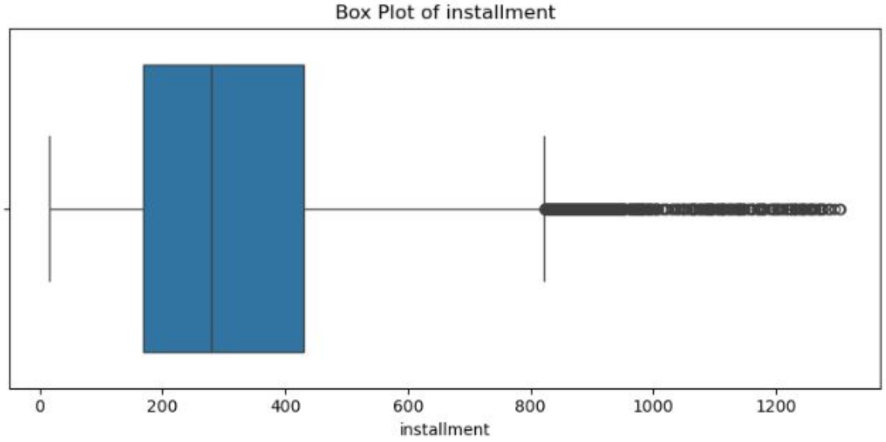
# Univariate Analysis of Numerical variables



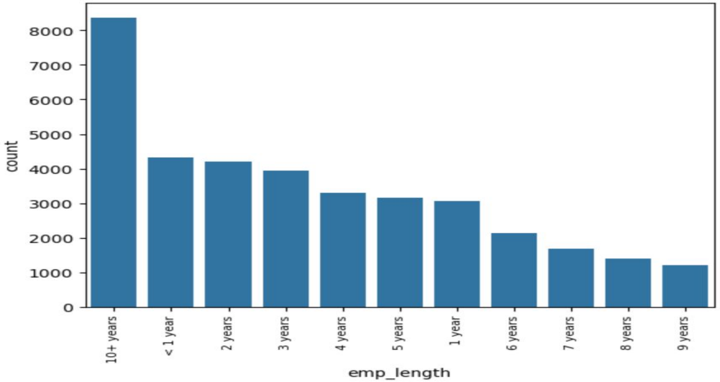
All the people stuck to either 36 or 60 months loan pay off term

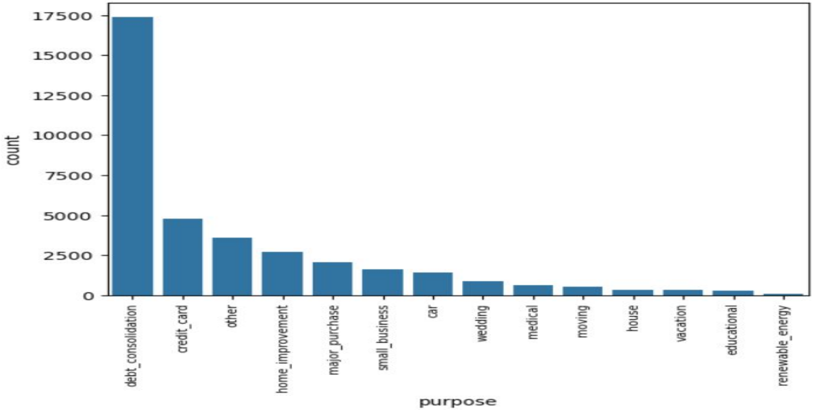
# Univariate Analysis of Numerical variables





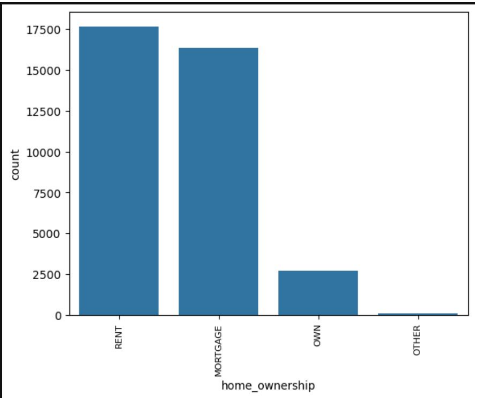
# Univariate Analysis of Categorical variables

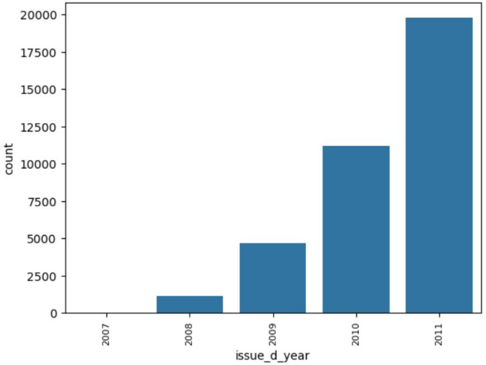


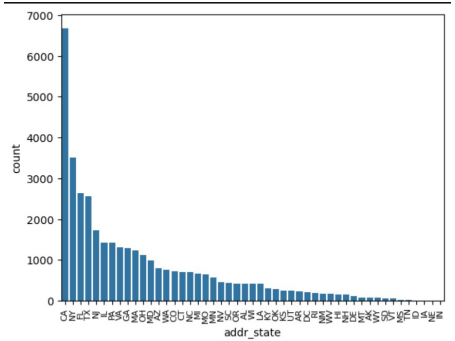


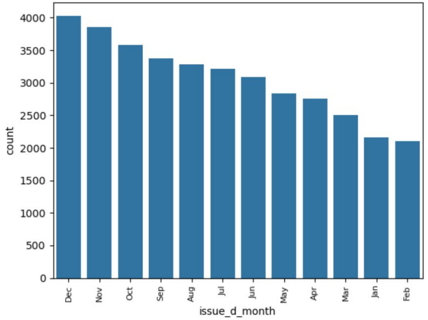
- Most of the people who applied for the loan have been working for more than 10 years
- Majority of them have applied loan for debt consolidation.

# Univariate Analysis of Categorical variables





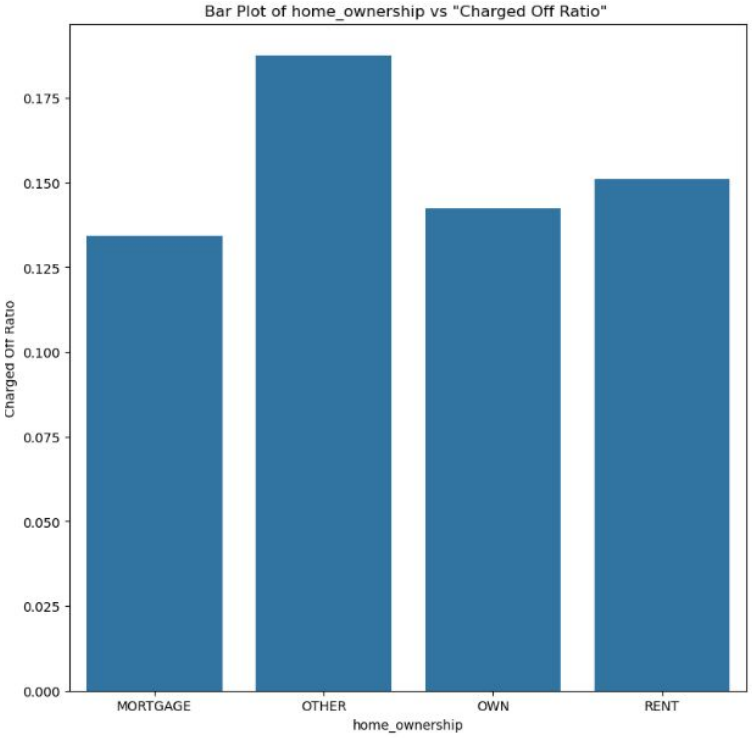


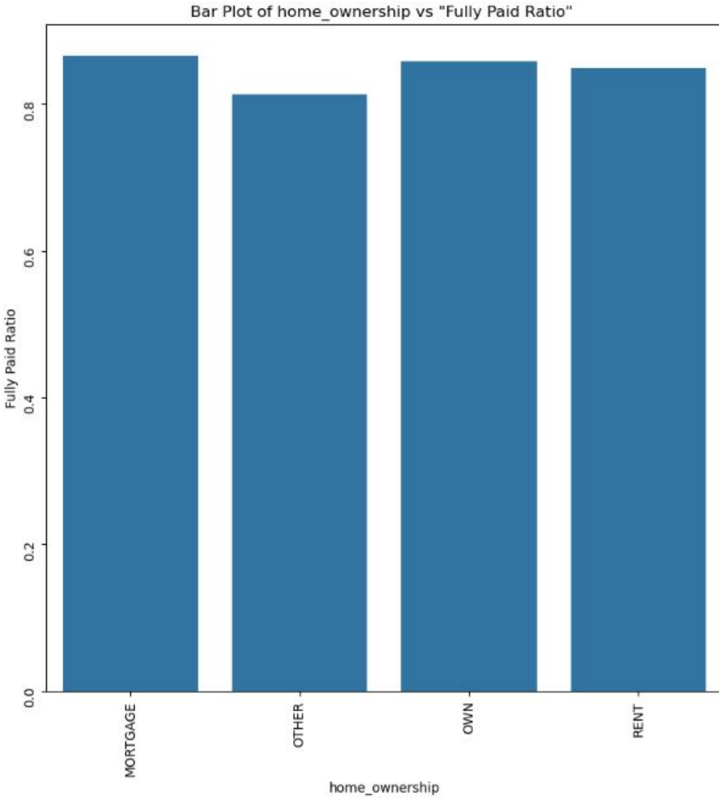


- Most of the people who applied for the loan live in a rented house
- Most of the loans were approved in the month December and most of the loans were approved in the year 2011
- Loans were mostly approved for the people living in California

# Bivariate Analysis

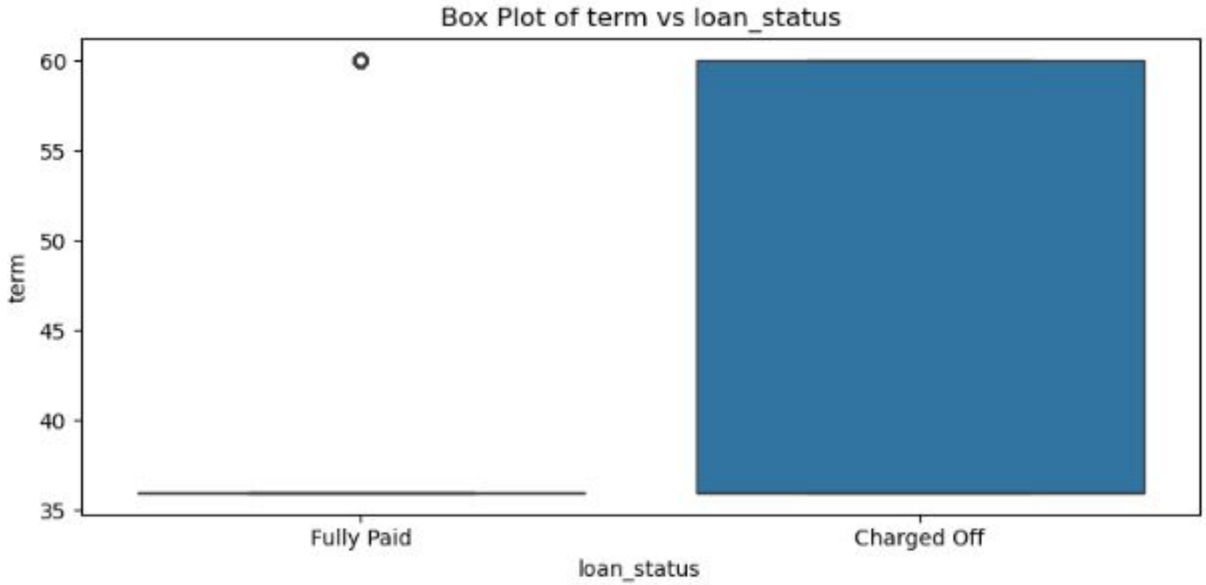
# Home Ownership vs Charge Off & Fully Paid Proportions





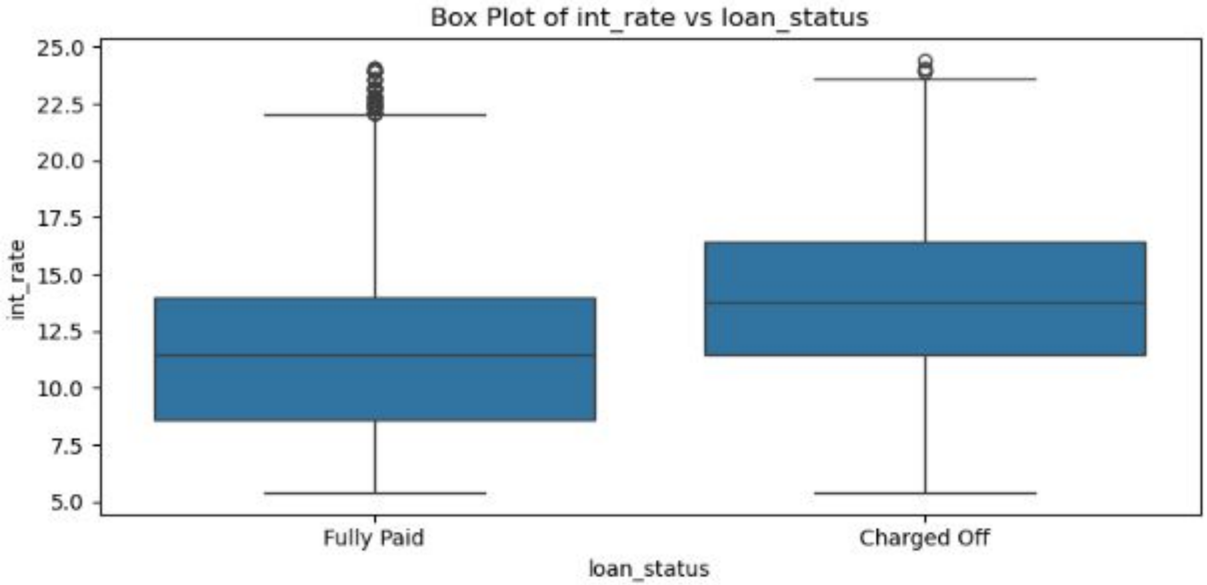
# Term vs Loan Status

- Customers opting for higher terms are likely to Charge-off

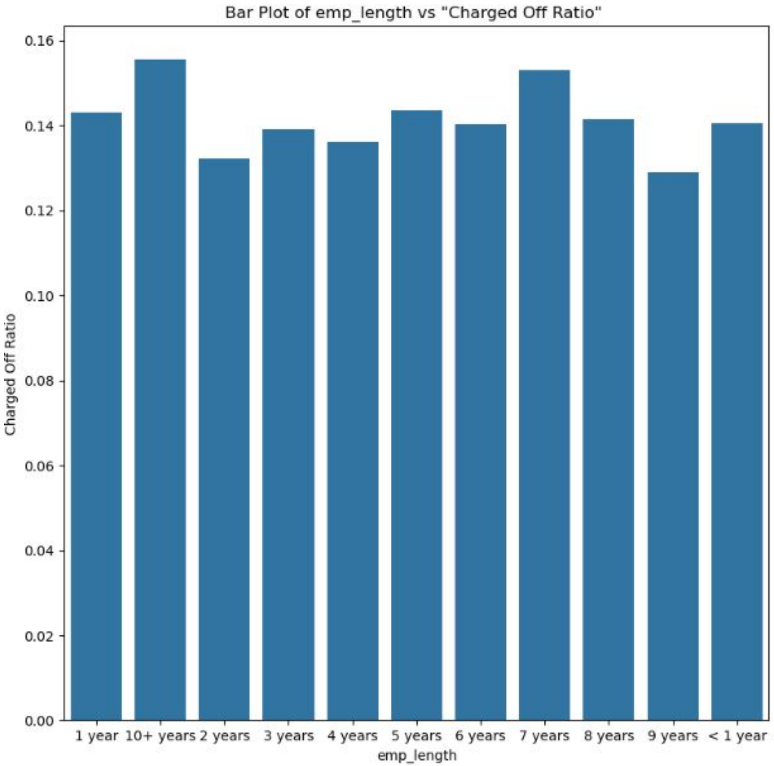


# Interest Rate vs Loan Status

- Customers getting burdened by higher rate of interests and charging off.

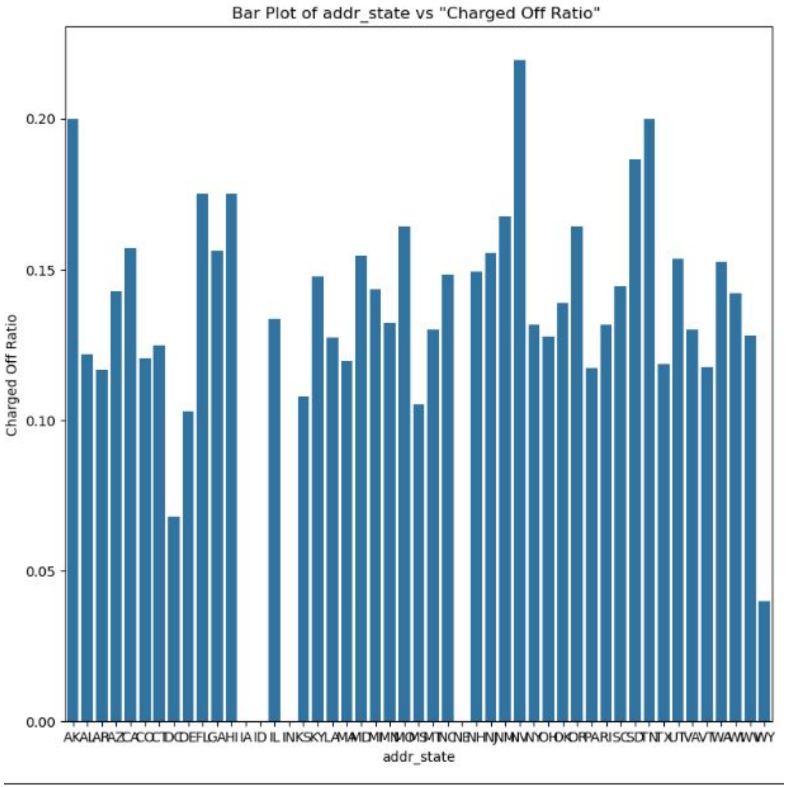


# Employee Length vs Charge Off Ratio



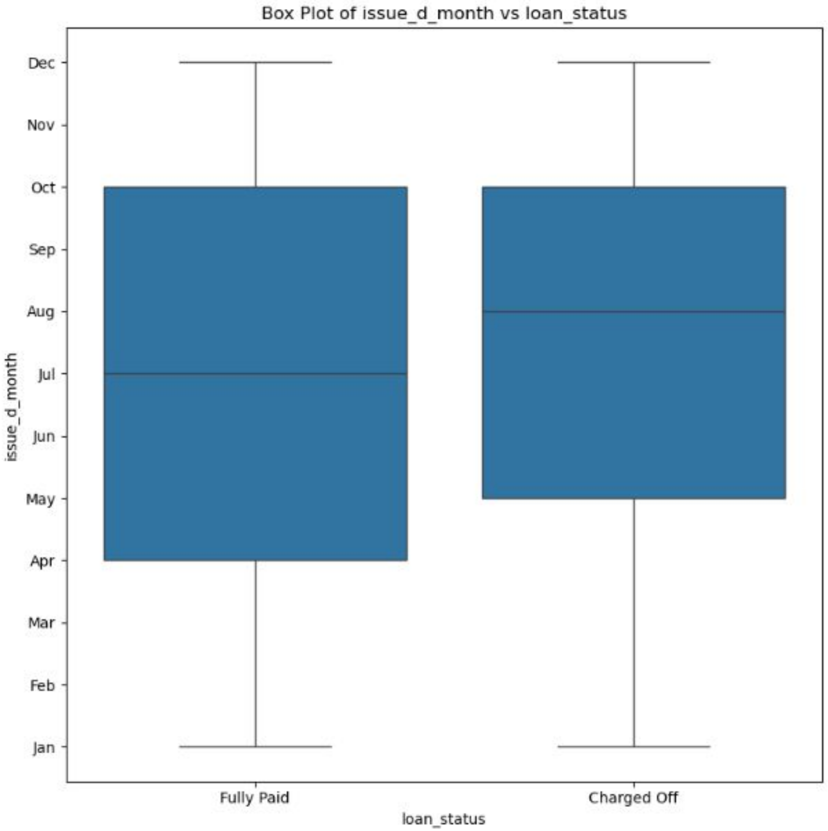
- Employee length or experience not much impacting Charge off ratio
- In all experience categories customers are charging off significantly with similar ratio.

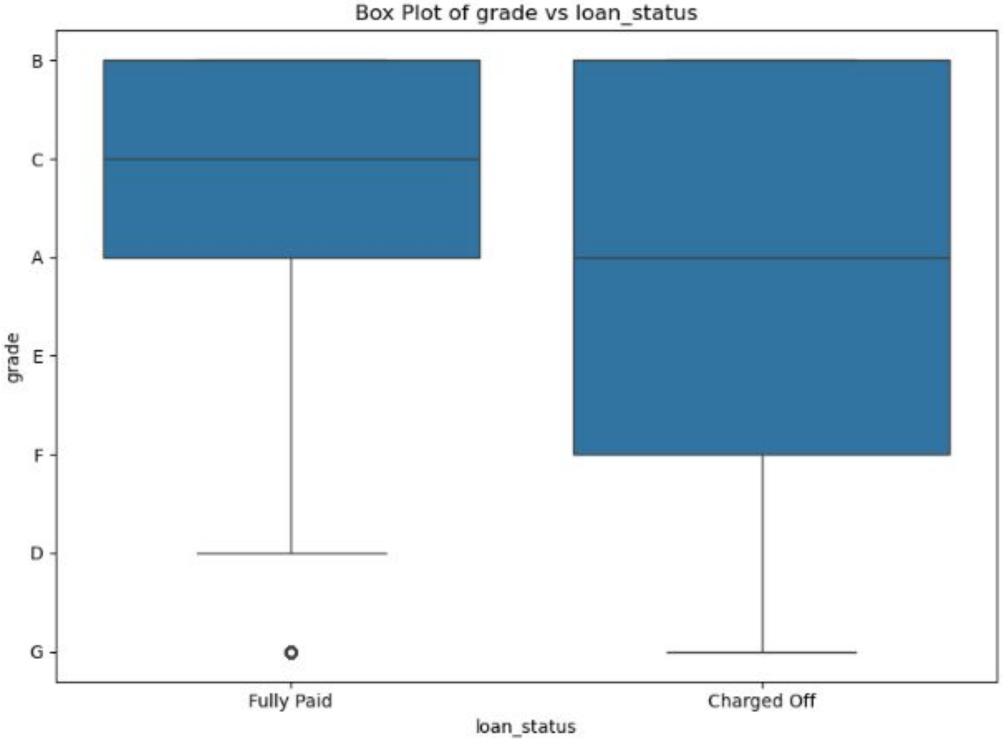
# Impact of State Economic Situation vs Charge Off Ratio



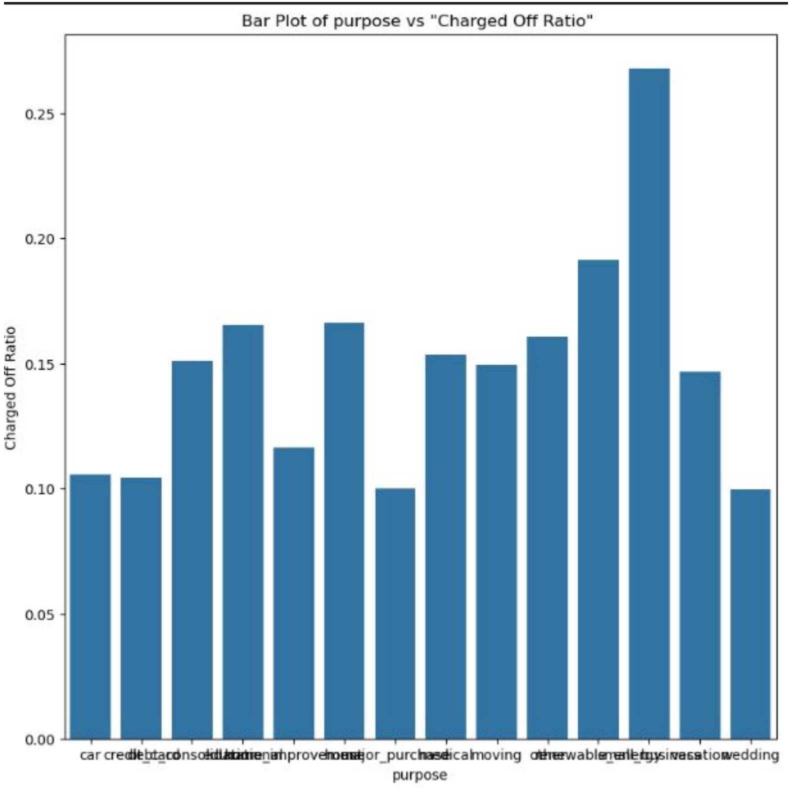
- Based on State’s economic situation, customers need to pay the installments and affect overall loan payment - either Charge off/ Fully Paid

# Categorical Parameters Highly Influencing Loan Status





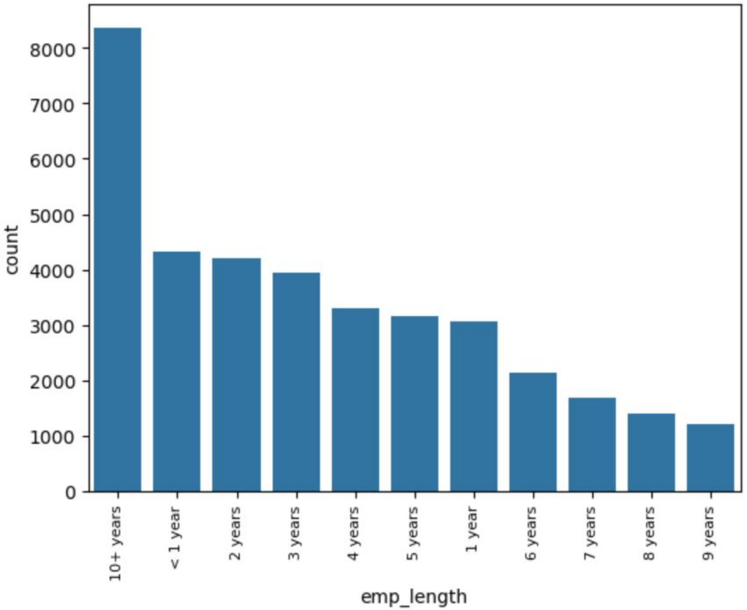
# Purpose Parameter - Determining the charge off

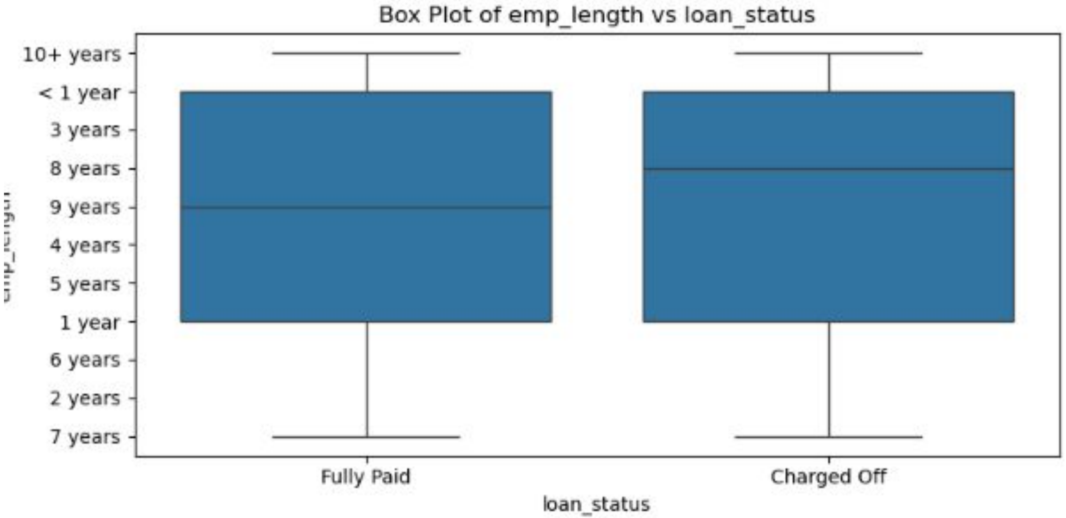


- If the lender restricts the loans where charge off is more, then it cis useful in reducing the loss to business

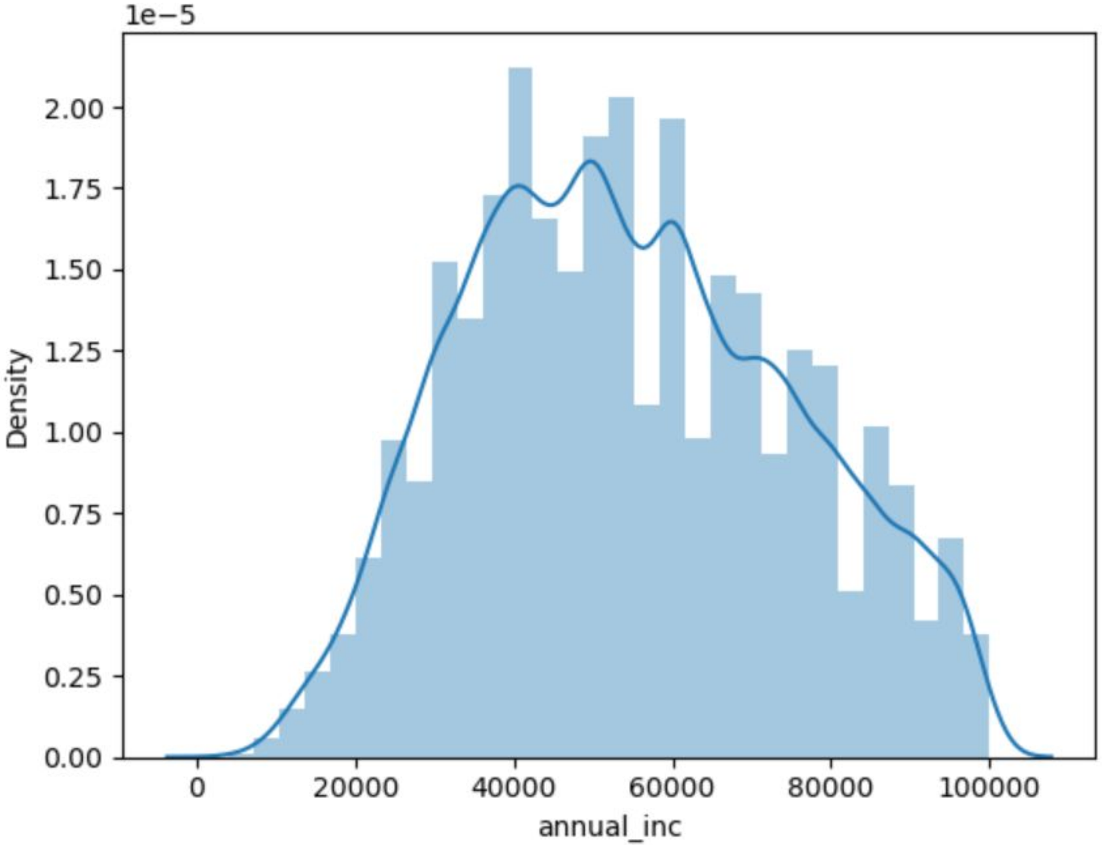
# Primary Parameter  - Customer Analysis

- Even salary and experience are directly proportional, we are not seeing much difference between, charge off and fully paid.

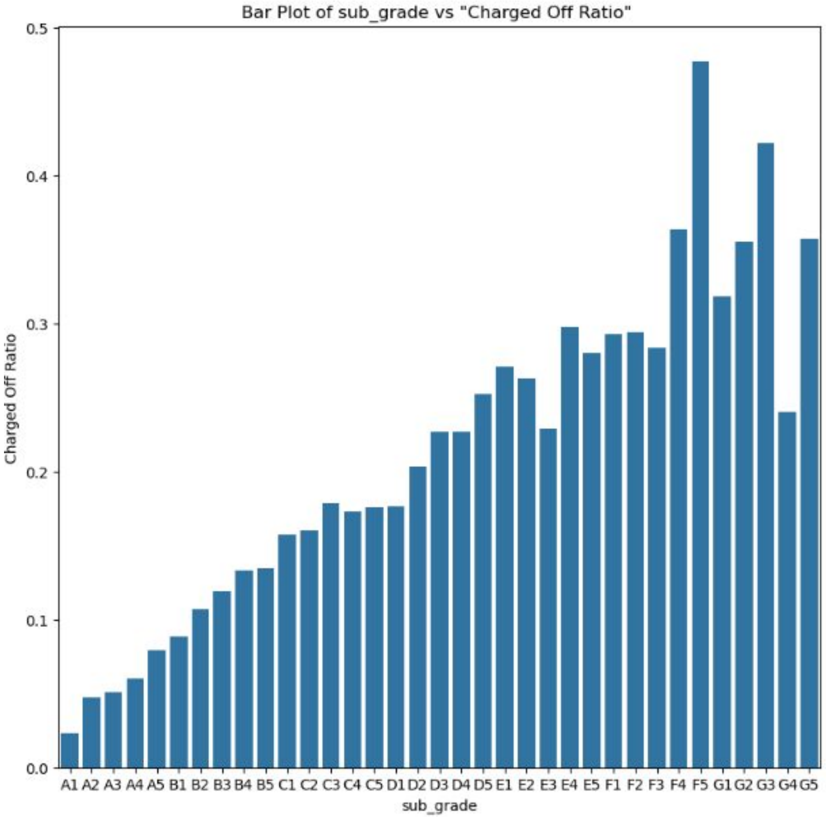




# Annual Income - A critical parameter charge off



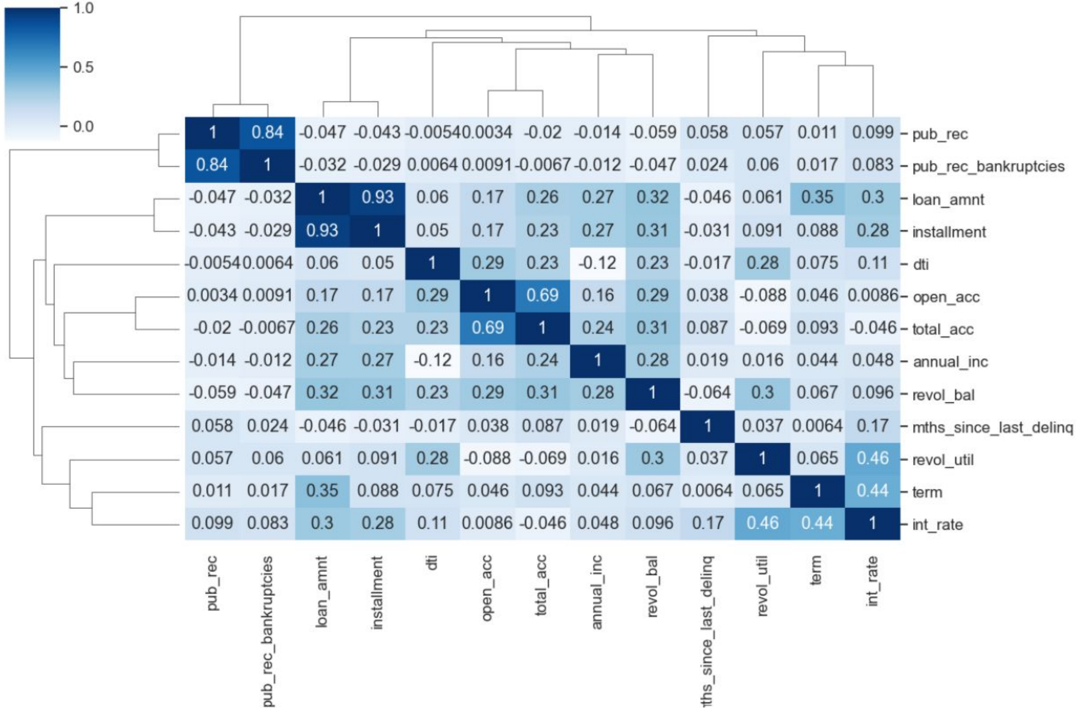
# Sub Grade - Proportionate Analysis



- Grade determination is very useful in determining customers capability to fully pay the loan.

# Correlation Matrix

- pub_rec and pub_bankruptcies are correlated, but for analysis we can keep them.
- As per definition, we see open_acc and total_acc seems to be same but here the correlation is 0.69. So, we can keep them for analysis purpose.
- All other numerical parameters are useful for determining the loan_status



| | pub_rec | pub_rec_bankruptcies | loan_amnt | installment | dti | open_acc | total_acc | annual_inc | revol_bal | ths_since_last_delinq | revol_util | term | int_rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pub_rec | 1 | 0.84 | -0.047 | -0.043 | -0.0054 | 0.0034 | -0.02 | -0.014 | -0.059 | 0.058 | 0.057 | 0.011 | 0.099 |
| pub_rec_bankruptcies | 0.84 | 1 | -0.032 | -0.029 | 0.0064 | 0.0091 | -0.0067 | -0.012 | -0.047 | 0.024 | 0.06 | 0.017 | 0.083 |
| loan_amnt | -0.047 | -0.032 | 1 | 0.93 | 0.06 | 0.17 | 0.26 | 0.27 | 0.32 | -0.046 | 0.061 | 0.35 | 0.3 |
| installment | -0.043 | -0.029 | 0.93 | 1 | 0.05 | 0.17 | 0.23 | 0.27 | 0.31 | -0.031 | 0.091 | 0.088 | 0.28 |
| dti | -0.0054 | 0.0064 | 0.06 | 0.05 | 1 | 0.29 | 0.23 | -0.12 | 0.23 | -0.017 | 0.28 | 0.075 | 0.11 |
| open_acc | 0.0034 | 0.0091 | 0.17 | 0.17 | 0.29 | 1 | 0.69 | 0.16 | 0.29 | 0.038 | -0.088 | 0.046 | 0.0086 |
| total_acc | -0.02 | -0.0067 | 0.26 | 0.23 | 0.23 | 0.69 | 1 | 0.24 | 0.31 | 0.087 | -0.069 | 0.093 | -0.046 |
| annual_inc | -0.014 | -0.012 | 0.27 | 0.27 | -0.12 | 0.16 | 0.24 | 1 | 0.28 | 0.019 | 0.016 | 0.044 | 0.048 |
| revol_bal | -0.059 | -0.047 | 0.32 | 0.31 | 0.23 | 0.29 | 0.31 | 0.28 | 1 | -0.064 | 0.3 | 0.067 | 0.096 |
| mths_since_last_delinq | 0.058 | 0.024 | -0.046 | -0.031 | -0.017 | 0.038 | 0.087 | 0.019 | -0.064 | 1 | 0.037 | 0.0064 | 0.17 |
| revol_util | 0.057 | 0.06 | 0.061 | 0.091 | 0.28 | -0.088 | -0.069 | 0.016 | 0.3 | 0.037 | 1 | 0.065 | 0.46 |
| term | 0.011 | 0.017 | 0.35 | 0.088 | 0.075 | 0.046 | 0.093 | 0.044 | 0.067 | 0.0064 | 0.065 | 1 | 0.44 |
| int_rate | 0.099 | 0.083 | 0.3 | 0.28 | 0.11 | 0.0086 | -0.046 | 0.048 | 0.096 | 0.17 | 0.46 | 0.44 | 1 |

# Conclusion

- The loan dataset is thoroughly analyzed after data understanding and manipulation.
- In Univariate analysis, significance of each and every parameter is justified.
- In Bi-variate analysis, with respect to loan status/ charge off ratio, the weight that each numerical/ categorical parameter is discussed.
- In correlation matrix, we gave a justification for the numerical parameters being used for analysis.