Lending Club Case Study

Exploratory Data Analysis

Analysis Done By -

1. MRUDHUL KOMMANA
2. NEERAJ KUMAR BHOLA

Content

Problem Statement

Loan Application Process

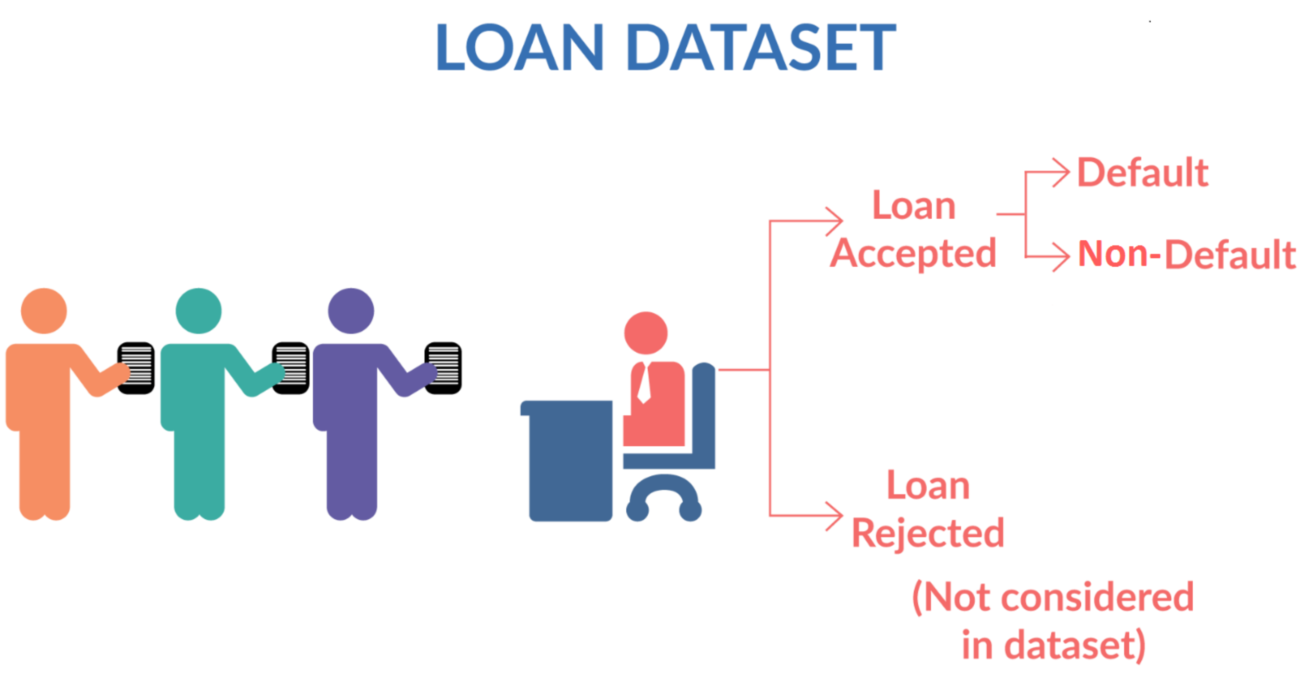
Lending Club, a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. So, as to minimize the business loss and maximize the profit.

Key Risks in Loan Approval

1. ***Loss of Business***:
   - **Opportunity Cost**: If a credit worthy applicant is rejected, the company may miss out on potential future revenue from that customer.
   - **Reputation Damage**: A high rejection rate could harm the company's reputation, leading to fewer loan applications and customer dissatisfaction.

2. ***Financial Loss***:
   - **Default Cost:** In the event of a default, the company incurs costs associated with collection efforts, legal proceedings, and potential write-offs.
   - **Increased Risk Premium:** A higher default rate may lead to increased borrowing costs for the company, as lenders perceive it as riskier.

Understanding the Dataset



Loan Outcomes & Data Limitations

Understanding Loan Outcomes and Data Limitations

Key Loan Outcomes:

1. **Fully Paid:** The applicant has successfully repaid the loan in full, indicating a positive repayment behavior.
2. **Current:** The loan is still being actively repaid, showing ongoing commitment from the applicant.
3. **Charged-Off:** Applicant has defaulted on the loan due to non-payment, which is a significant risk for lenders.
4. **Rejected:** The loan application was denied by the company.

Data Limitations:

Rejected Loans: The dataset only includes information on approved loans, as there is no transactional history for rejected applications.

Understanding the Business Objectives

Key Business Challenges and Opportunities:

Credit Risk Management: The company faces a significant financial risk due to loan defaults.

Data-Driven Decision Making: Leveraging data analytics can help identify early warning signs of default and improve risk assessment.

Portfolio Optimization: By understanding the factors that drive default, the company can optimize its loan portfolio to minimize risk.

Data Summary

Data Summary

- “loan.csv” file containing 39717 rows and 111 columns was provided for the analysis.
- There are two types of attributes
  - Loan Attribute and
  - Customer attributes.

Data Cleaning

- **Observations -**
- No header, footers, summary or rows numbers were found in the dataset.
- no duplicates rows found.
- 1140 rows present of loan_status=‘current’ which has been deleted as *loan_status =‘current’* is not required for analysis.
- 55 columns have all the rows values as *“null / blank”* and doesn’t participate in analyze has been removed.
- *‘url’ and ‘member_id’* is unique in nature and has been deleted. Have considered ‘id’ for further analysis.
- *‘desc’ and ‘title’* contains text/description values and doesn’t participate has been dropped from analysis.
- Limiting our analysis to *‘Group’* level only hence sub-group has been dropped.
- Using domain knowledge, behavioral data is captured and hence will not available during the loan approval and is not considered in this analysis.
- 21 behavioral data columns has deleted.
- 8 columns *whose values were 1*, since this is a unique value it has been dropped from analysis.
- There were *two columns* which is having more that 50% of data as “NA” has been removed.
- After completing all the data cleaning process, we are left with **38577 rows and 20 columns** for our analysis.

Data Conversions vs Derived Columns

- ✓ Additional string value has been trimmed from *'term'* and *'int_rate'* column and has been converted **to int** data types.
- ✓ Column *'loan_funded_amnt'* and *'funded_amnt'* converted **to float**.
- ✓ *'lineament', 'funded_amnt', 'funded_amnt_inv', 'int_rate', 'data'* columns valued rounded off **to two decimal points**.
- ✓ *issue_d* has been converted **to datatype**.
- ✓ Creating a derived columns for *'issue_year'* and *'issue_month '* from *'issue_d'* which will be used for further analysis.
- ✓ *'lineament_b', 'annual_inc_b', 'int_rate_b,* and *'data_b'* derived columns (multiple **bucket** kind of data from continuous data) has been created for better analysis.

Dropping / Imputing the rows

- *'emp_lenght'* and *'pub_rec_bankruptcies'* contains 2.67% and 1.80% of rows as null, which is very small percentage of data which we can drop it.
- Total % of rows deleted: **4.48%**
- Outliers exists for numeric data *'lineament', 'funded_amnt', 'funded_amnt_inv','int_rate', 'installment', 'annual_inc'*.
- Outliers treatment has been done for above fields using quantile mechanism.

Univariate Analysis

EDA - Univariate Analysis:

Descriptive Statistics: Calculate mean, median, mode, standard deviation, and other summary statistics for numerical variables.

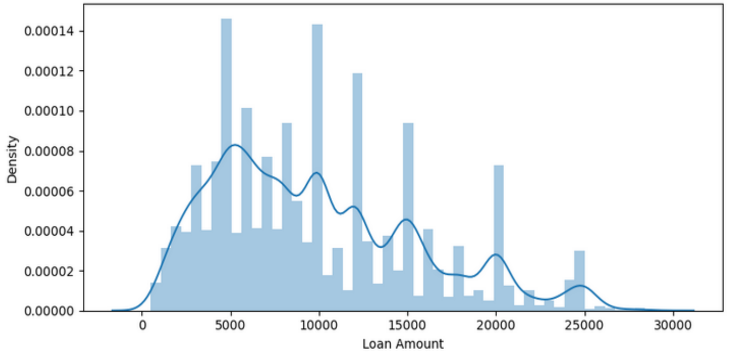
Frequency Distributions: Analyze the distribution of categorical variables.

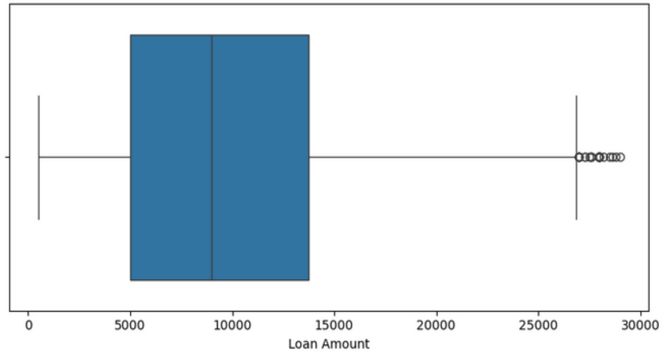
Visualization: Use histograms, box plots, and bar charts to visualize the data.

Univariate Analysis

Loan Amount

```
plot_digram(loan_df, 'loan_amnt')
# describe the Loan Amount
print(loan_df['loan_amnt'].describe())
plt.show()
```






```
count    33191.000000
mean      9820.838480
std       5809.600807
min        500.000000
25%       5000.000000
50%       9000.000000
75%      13750.000000
max      29000.000000
Name: loan_amnt, dtype: float64
```

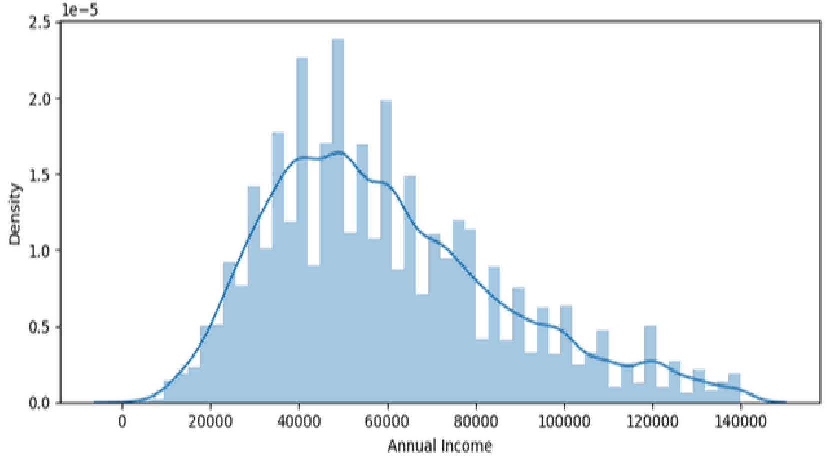
Observation of Loan Amount:

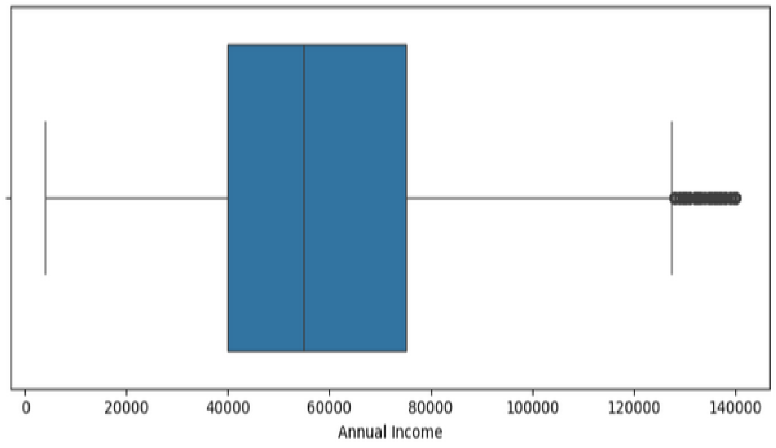
Most of the loan amount applied was in the range of 5k-14k.

Max Loan amount applied was ~29k.

Univariate Analysis

Annual Income





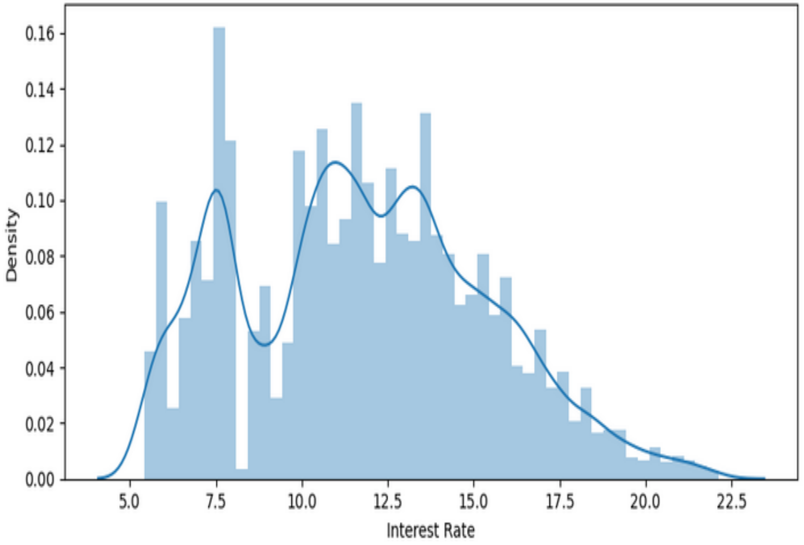

```
count     33191.000000
mean      59883.284700
std       26916.857415
min        4000.000000
25%       40000.000000
50%       55000.000000
75%       75000.000000
max      140000.000000
Name: annual_inc, dtype: float64
```

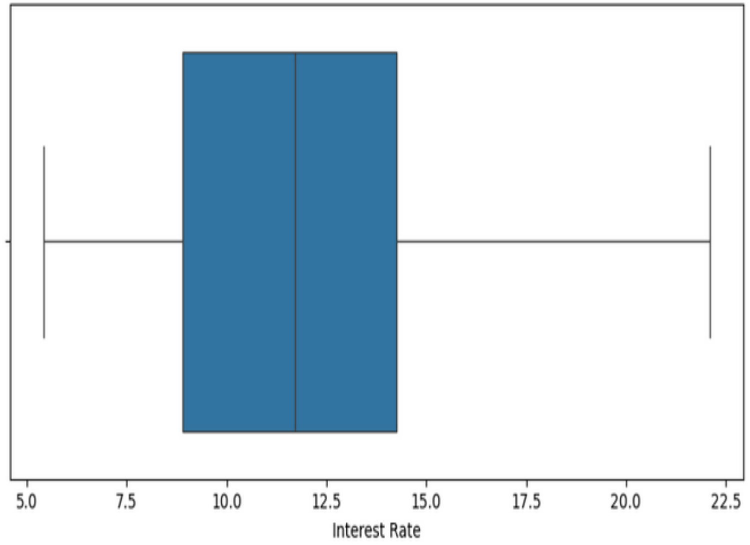
```
# The Annual income of most if applicants lies between 40k-75k.
print("Average Annual Income is :", round(loan_df['annual_inc'].median(),2))
```

```
Average Annual Income is : 55000.0
```

Univariate Analysis

Interest Rate



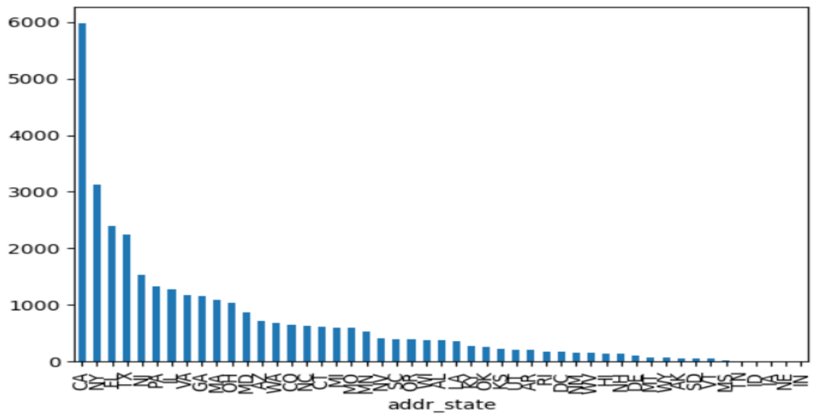


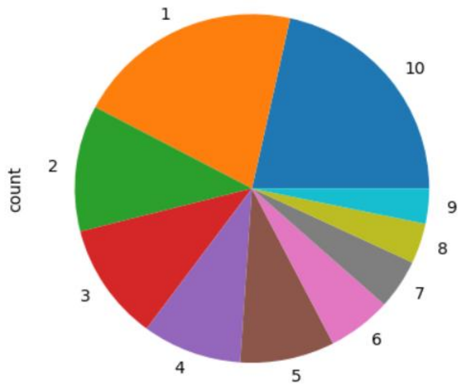

```
count    33191.000000
mean        11.782783
std          3.591944
min          5.420000
25%          8.900000
50%         11.710000
75%         14.260000
max         22.110000
Name: int_rate, dtype: float64
```

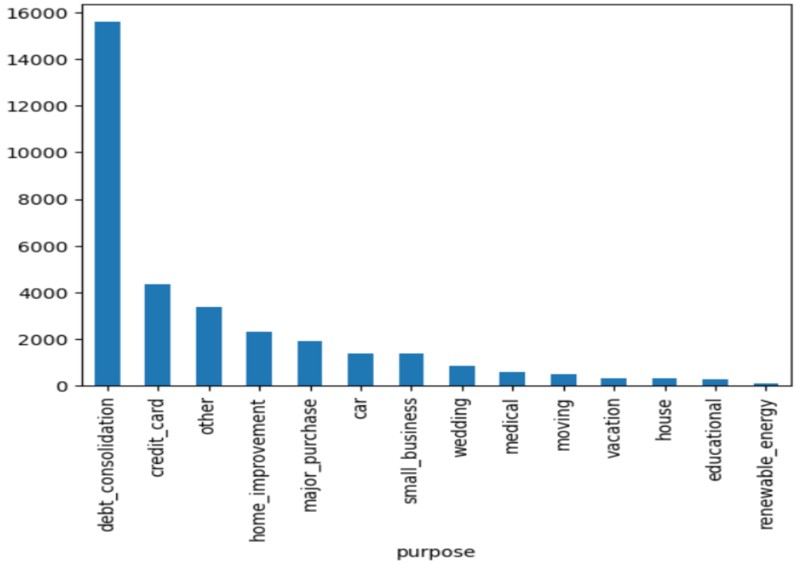
- Observation of Rate of Interest
- Most of the applicant's rate of interest is between in the range of 8%-14%.

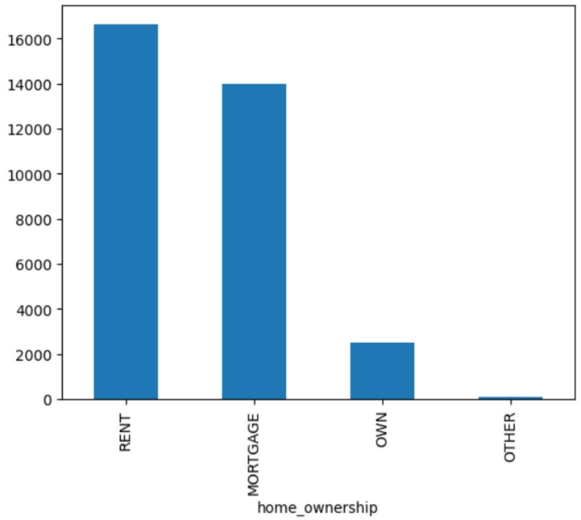
Average Rate of interest of rate is 11.71%

Unordered & Ordered Categorical Variable Analysis







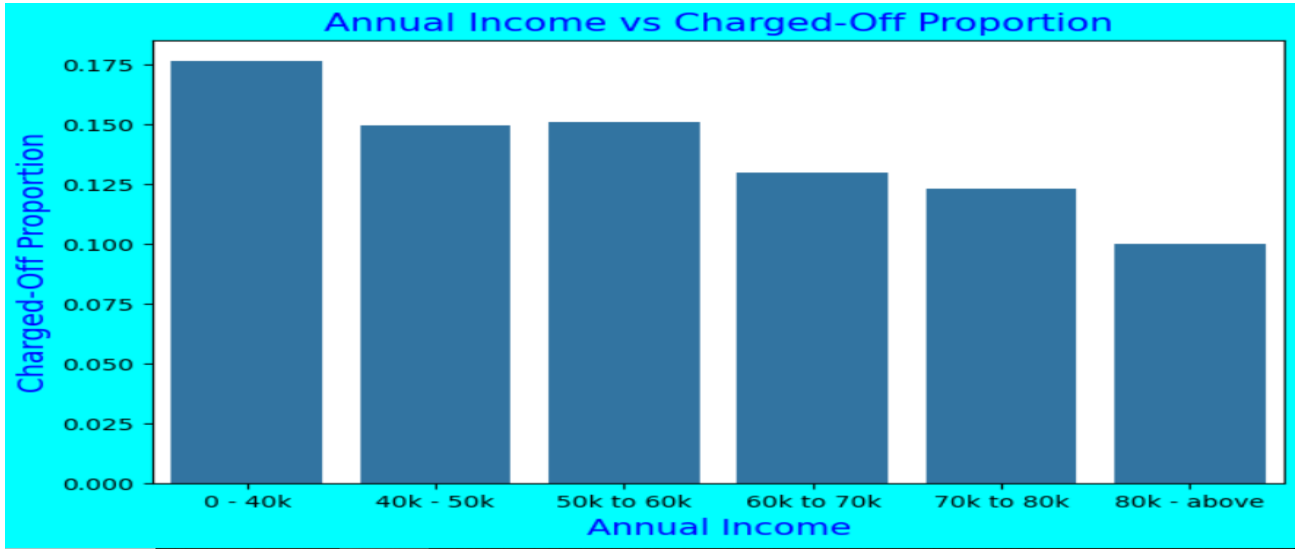


__Observations__:

1. Most of the Loan applicants are from CA, NY and FL (States).
2. Most of the applications are having 10+ yrs of Exp.
3. Most of the loan applicants are for debt_consolidations.
4. Majority of loan applicants are either living on Rent or on Mortgage.

Bivariate Analysis

Annual Income vs Charged Off



loan_status	int_rate_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
3	Very High	1670	4751	6421	0.260084
0	High	985	4851	5836	0.168780
2	Moderate	961	5638	6599	0.145628
1	Low	579	4983	5562	0.104099
4	Very Low	519	8254	8773	0.059159

Observations:

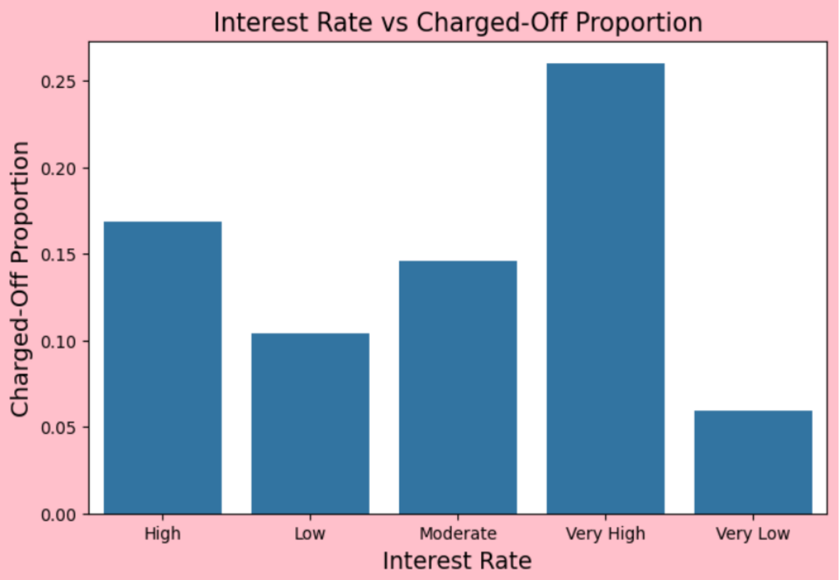
1. Income range 80k & above have less chances of charged off.

2. Income range of 0 to 40k have high chances of charged-off.

3. Notice that with increase in annual income charged off proposition got decreased.

Bivariate Analysis

Interest Rate vs Charged off:



loan_status	home_ownership	Charged Off	Fully Paid	Total	Chargedoff_Proportion
1	OTHER	16	73	89	0.179775
3	RENT	2488	14156	16644	0.149483
2	OWN	355	2121	2476	0.143376
0	MORTGAGE	1855	12127	13982	0.132671

Observations:

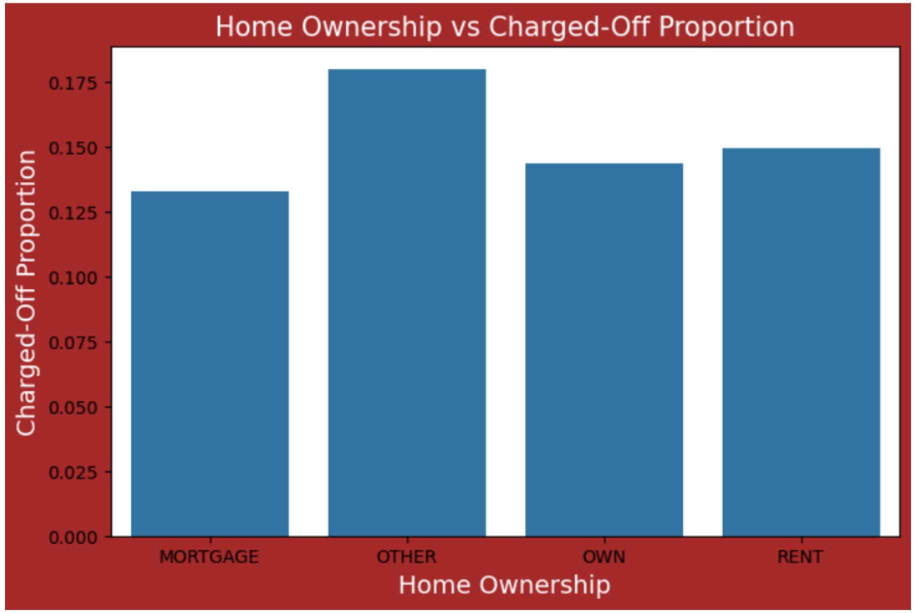
1. Interest rate less than 10% or very low has very less chances of charged off. Interest rates are starting from minimum 5 %.

2. Interest rate more than 16% or very high has good chances of charged off as compared to other category interest rates.

3. Charged off proportion is increasing with higher interest rates.

Bivariate Analysis

Home Ownership vs Charged off



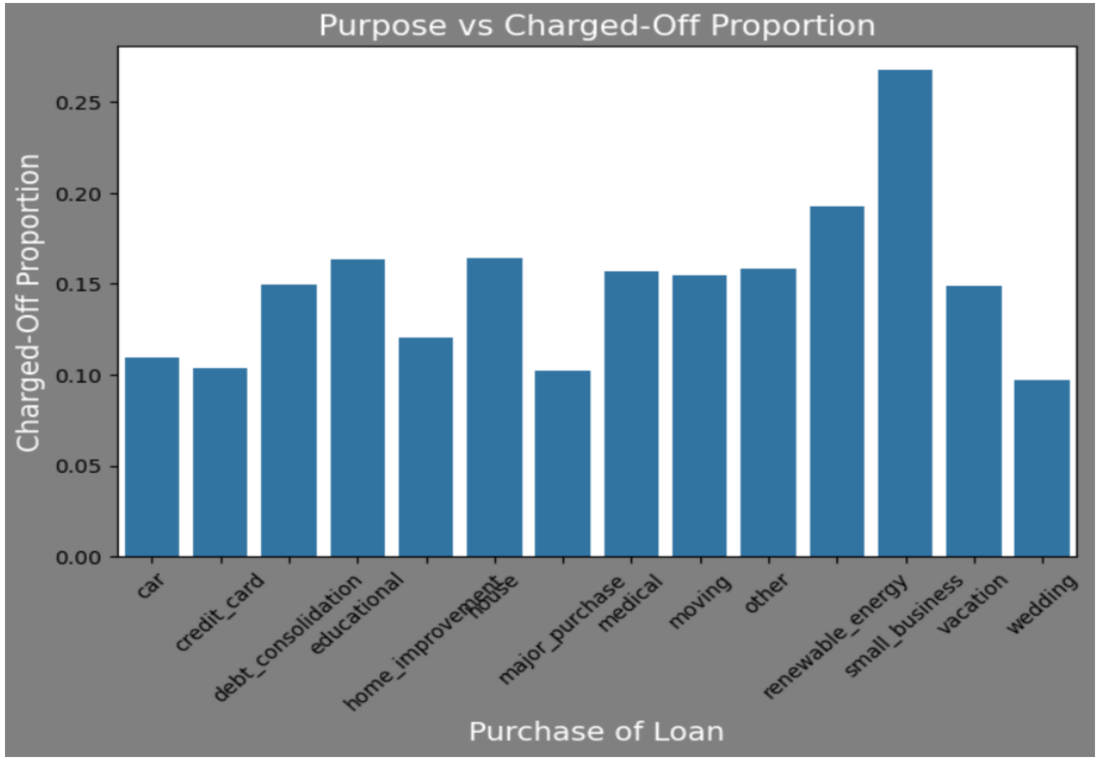
loan_status	home_ownership	Charged Off	Fully Paid	Total	Chargedoff_Proportion
1	OTHER	16	73	89	0.179775
3	RENT	2488	14156	16644	0.149483
2	OWN	355	2121	2476	0.143376
0	MORTGAGE	1855	12127	13982	0.132671

Observations:

Those who are not owning the home is having high chances of loan defaulter.

Purpose vs Charged Off

Bivariate Analysis



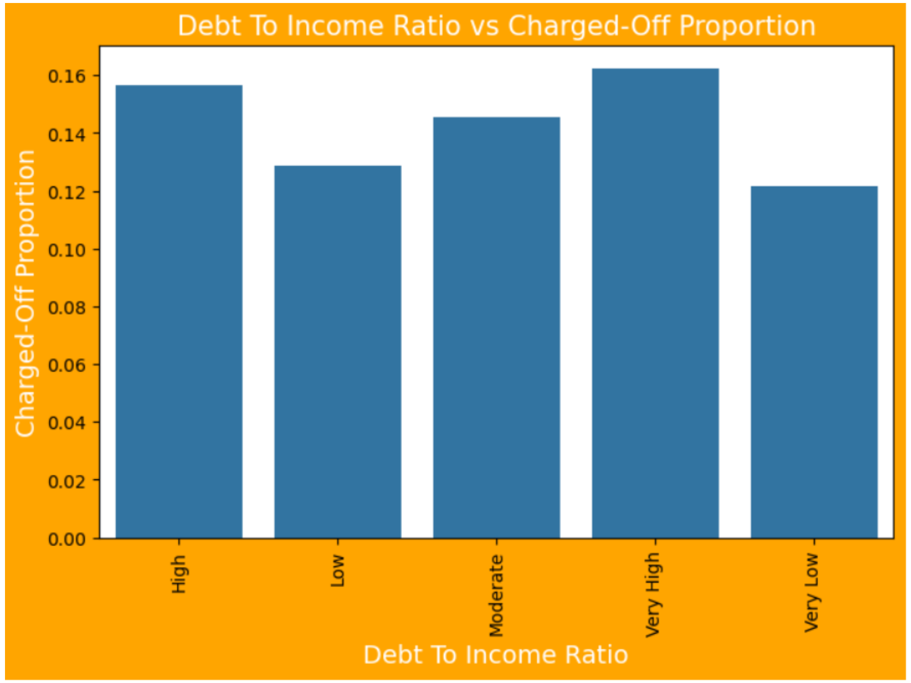
Observations:

1. Those applicants who is having home loan is having low chances of loan defaults.

2. Those applicants having loan for small bussiness is having high chances for loan defaults.

Bivariate Analysis

DTI Vs Charged off:



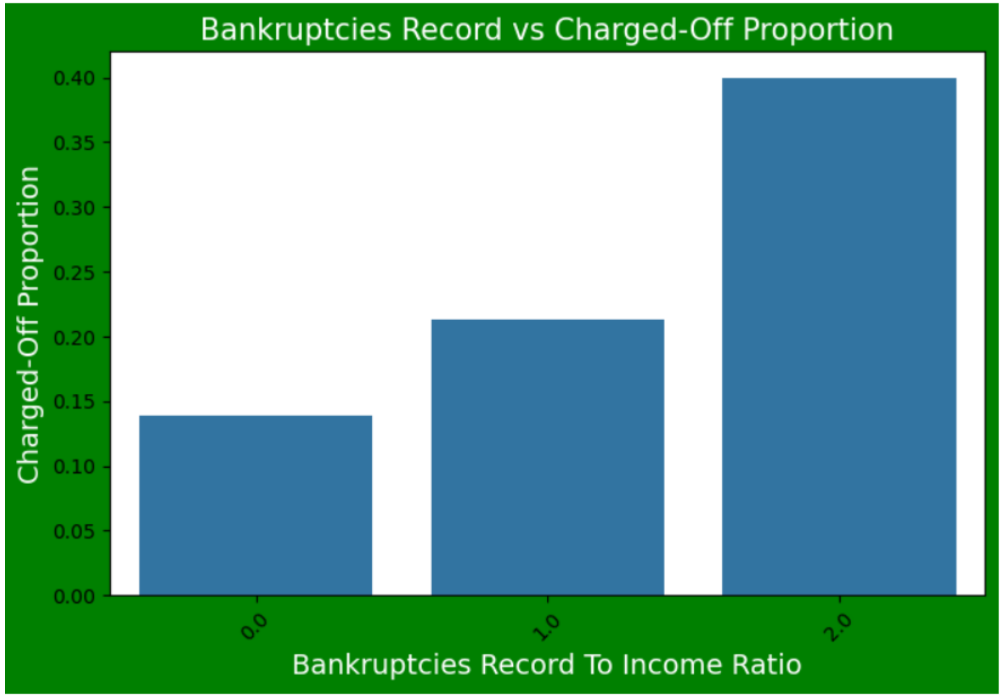
Observations:

1. High DTI value having high risk of defaults

2. Lower the DTO having low chances loan defaults

Bivariate Analysis

Bankruptcies Record vs Charged off



Observations:

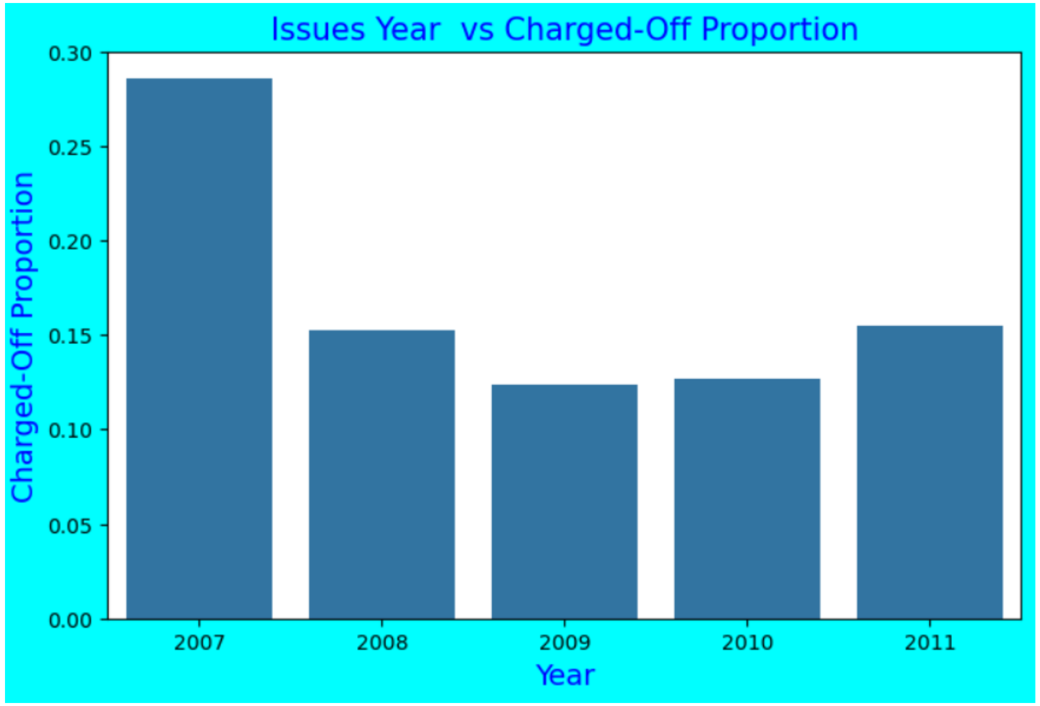
1. Bankruptcies Record with 2 is having high impact on loan defaults.

2. Bankruptcies Record with 0 is low impact on loan defaults.

3. Lower the Bankruptcies lower the risk.

Bivariate Analysis

Issue Year vs Charged off:



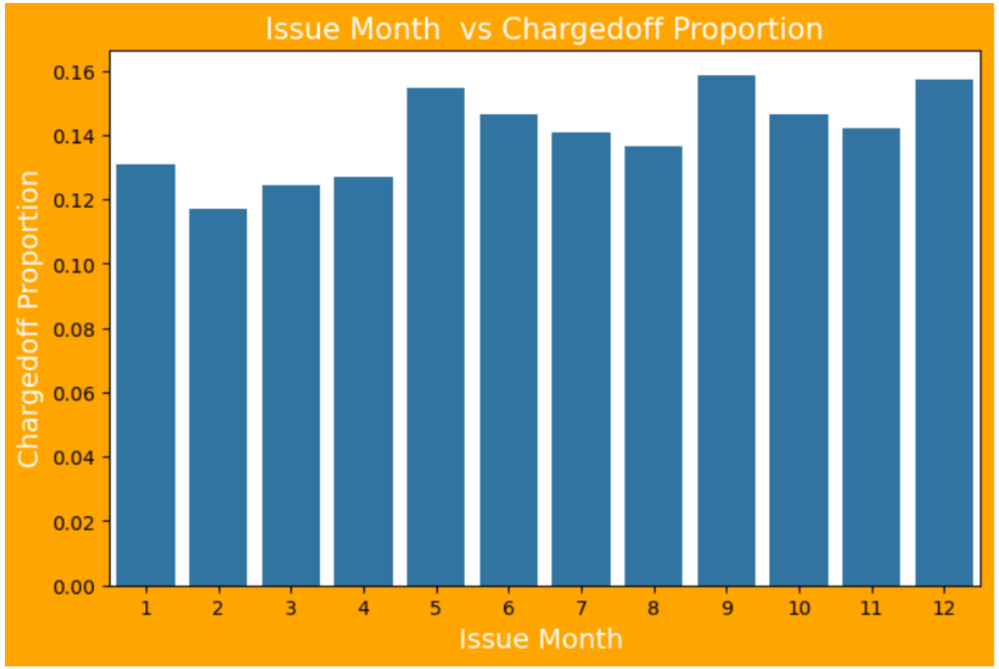
Observations:

1. Year 2007 is highest loan defaults.

2. Year 2009 is having lowest loan defaults.

Bivariate Analysis

Issue Month Vs Charged off



Observations:

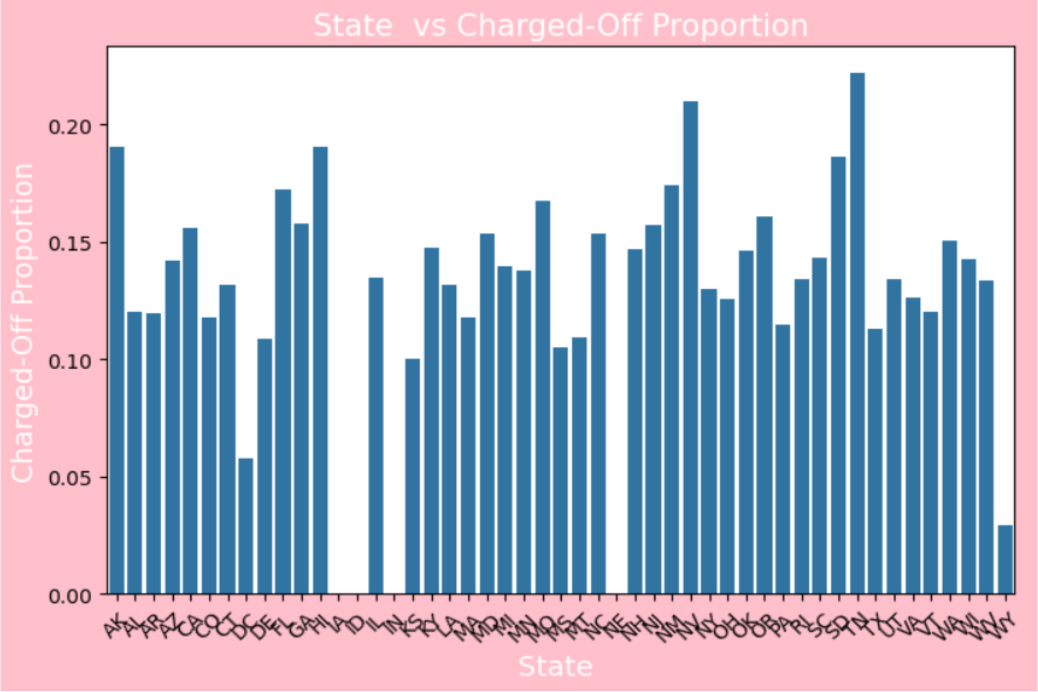
1. Those loan has been issued in May, September and December is having high number of loan defaults.

2. Those loan has been issued in month of February is having high number of loan defaults.

3. Majority of loan defaults coming from applicants whose loan has been approved from September to December.

Bivariate Analysis

State vs Charged off:



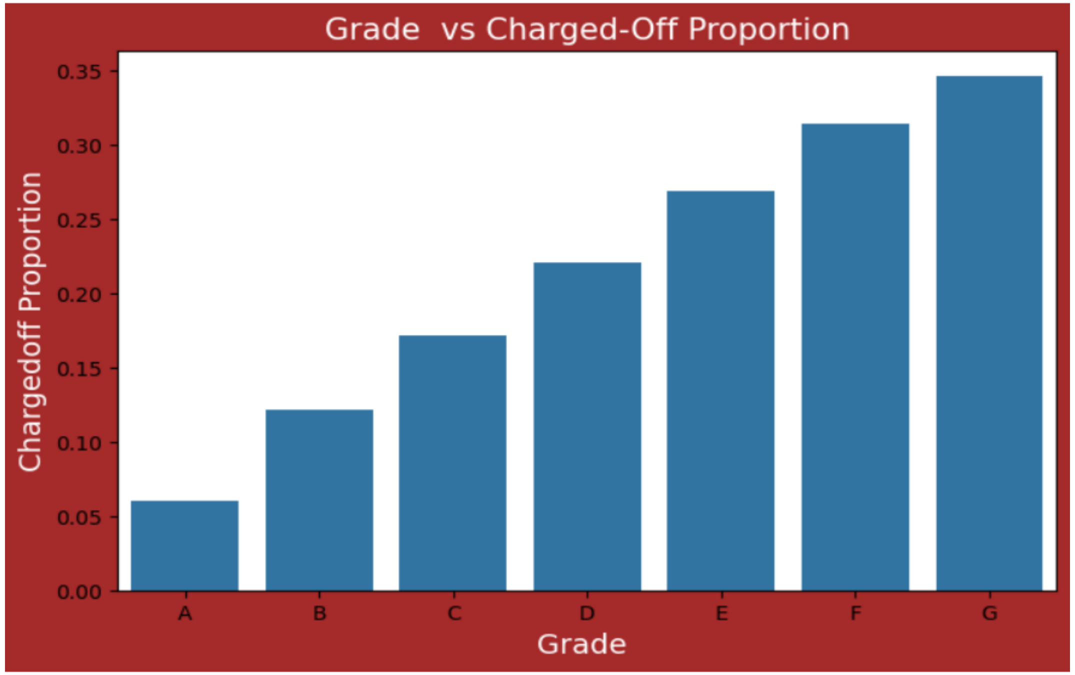
Observations:

1. DE States is holding highest number of loan defaults.

2. CA is having low number of loan defaults.

Bivariate Analysis

Grade vs Charged Off

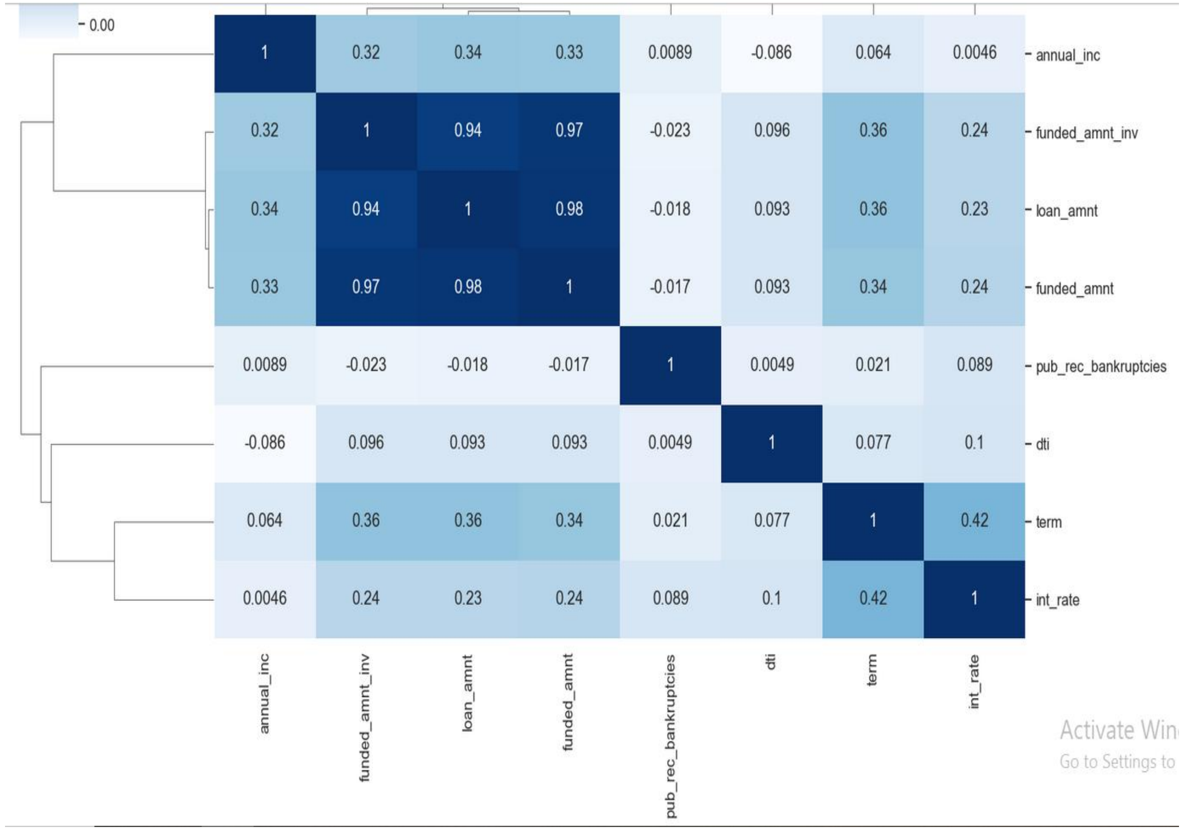


Observations:

1. The Loan applicants with loan Grade G is having highest Loan Defaults.

2. The Loan applicants with loan A is having lowest Loan Defaults.

Correlation



Negative Correlation:

1. lineament has negative correlation with lineament.
2. annual income has a negative correlation with data.

Strong Correlation:

1.term has a strong correlation with loan amount.
2.term has a strong correlation with interest rate.
3.annual income has a strong correlation with loan amount.

Final Conclusion

- Lending club should reduce the high interest loans for 60 months tenure, they are prone to loan default.
- Grades are good metric for detecting defaulters.
- Lending club should examine more information from borrowers before issuing loans to Low grade (G to A).
- Lending Club should control their number of loan issues to borrowers who are from CA, FL and NY to make profits.
- Small business loans are defaulted more. Lending club should stop/reduce issuing the loans to them.
- Borrowers with mortgage home ownership are taking higher loans and defaulting the approved loans.
- Lending club should stop giving loans to this category when loan amount requested is more than 12000.
- People with more number of public derogatory records are having more chance of filing a bankruptcy.
- Lending club should make sure there are no public derogatory records for borrower.

Thank You !

- Please feel free to contact us on below github ids.

- Mrudhul Kommana – mrudhulk
- Neeraj Kumar Bhola - NeerajBhola21