iiit-b

UpGrad

Lending Club Case Study

SUBMISSION

Group Members:
Shantnu Kumar Singh
Shilpa

Lending Club Case Study

Background: Lending Club (LC), headquartered in San Francisco, California, is the world's largest peer-to-peer lending platform. It was the first to register its offerings as securities with the U.S. Securities and Exchange Commission (SEC) and introduced loan trading on a secondary market.

How It Works: Customers seeking loans can easily apply online at LendingClub.com. Lending Club utilizes data and technology to assess borrowers, determine appropriate interest rates, and manage loan servicing. Qualified applicants receive loan offers without affecting their credit scores. Investors can then choose loans to invest in based on their individual risk tolerance, investment goals, and time horizon.

Objective: The goal is to identify key variables that strongly indicate the likelihood of loan default. This will assist Lending Club in making more informed decisions about loan approval or rejection.

Loan Default Prediction Analysis: Key Insights and Methodology

Problem Statement: Identify key variables that predict potential loan defaults, helping Lending Club decide on loan approvals or rejections.

Data Summary: Overview of dataset including key features, missing values, and basic statistics.

Data Cleaning: Process of handling missing, inconsistent, or erroneous data to ensure data quality.

Data Conversions vs Derived Columns:

- **Conversions:** Transforming data types or formats.
- **Derived Columns:** Creating new variables based on existing data for better insights.

Dropping/Imputing Rows:

- **Dropping:** Removing rows with missing or irrelevant data.
- **Imputing:** Filling missing values using techniques like mean, median, or mode.

Outliers: Identifying and handling data points that deviate significantly from the rest of the dataset.

Univariate Analysis: Examining individual variables to understand their distribution and properties.

Bivariate Analysis: Exploring relationships between two variables, often using scatter plots or cross-tabulations.

Correlations: Measuring the strength and direction of relationships between different variables.

Conclusions: Summarize findings, key insights from the analysis, and their impact on loan default predictions.

Problem Statement

Problem:

- You work for a consumer finance company specializing in offering various loans to urban customers. Upon receiving a loan application, the company must decide whether to approve or reject it based on the applicant's profile. Two major risks arise from this decision:
  - **Lost Business:** If the applicant is likely to repay, rejecting the loan results in lost business for the company.
  - **Financial Loss:** If the applicant is likely to default, approving the loan could lead to financial losses.

Objective:

- Use Exploratory Data Analysis (EDA) to understand how consumer and loan attributes affect the likelihood of loan default.

Constraints:

- When a person applies for a loan, the company can take one of two actions:
  1. **Loan Accepted:** If approved, one of the following occurs:
     - **Fully Paid:** The applicant has repaid the loan completely (principal + interest).
     - **Current:** The applicant is still making payments, meaning the loan is not yet fully repaid and can't be labeled as "defaulted."
     - **Charged-Off:** The applicant has defaulted, failing to pay installments for an extended period.
  2. **Loan Rejected:** The company rejects the application due to the candidate not meeting the requirements. Since the loan was not granted, no default information is available for these cases.

Loan Default Prediction Analysis: Key Insights and Methodology

Data Summary

- The dataset contains **39,717 rows** and **111 columns** from the file `Loan.csv`.
- The attributes are categorized into two types:
  - **Loan Attributes**: Details about the loan, such as amount, interest rate, and loan status.
  - **Customer Attributes**: Information about the applicant, such as income, credit history, and employment details.

Loan Default Prediction Analysis: Key Insights and Methodology

Data Cleaning

- **No headers, footers, or summary/total rows** were found, and no duplicate rows were detected.
- **Removed 1,140 rows** where `loan_status ='current'`, as they do not contribute to the analysis.
- **55 columns** with all null or blank values were deleted since they don't participate in the analysis.
- Dropped **'url'** and **'member_id'** (unique identifiers), keeping only **'id'** for potential future analysis.
- Dropped **'desc'** and **'title'**, which are text/description-based and not relevant to the analysis.
- Analysis was limited to the **Group** level, so the **Sub-group** columns were removed.
- Based on domain knowledge, **21 behavioral data columns** (captured post-loan approval) were deleted, as they are not available during the loan approval process.
- **8 columns** with values that were always **1** (unique) were dropped.
- Two columns with **more than 50% missing values** were also removed.

After the data cleaning, we are left with **38,577 rows** and **20 columns**.

Loan Default Prediction Analysis: Key Insights and Methodology

Data Conversions vs Derived Columns

- Additional string value has been trimmed from ‘term’ column and has been converted to int data types.

- ‘int_rate’ has been converted from string to int. Additional ‘%’ has been trimmed.

- Column ‘loan_funded_amnt’ and ‘funded_amnt’ converted to float.

- loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', ‘dti' 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 using for further analysis.

- ‘loan_amnt_b’, ‘annual_inc_b’, ‘int_rate_b, and ‘dti_b’ derived columns(multiple bucket kind of data from continuous data) has been created for better analysis.

Loan Default Prediction Analysis: Key Insights and Methodology

Dropping/Inputting the rows

• ‘emp_length’ 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 exits for numeric data 'loan_amnt', 'funded_amnt', ‘funded_amnt_inv','int_rate', 'installment', ‘annual_inc'.

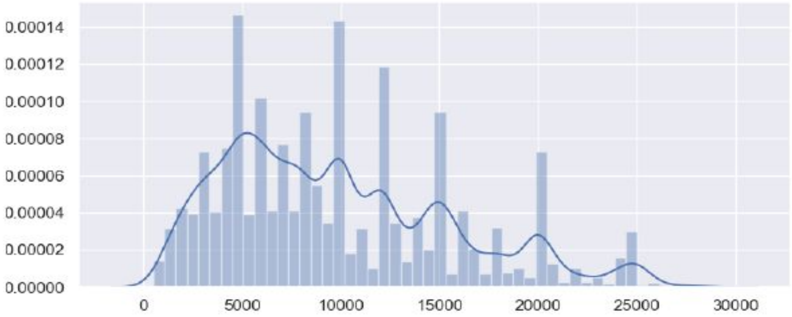
• Outliers treatment has been done for above fields using quantile mechanism.

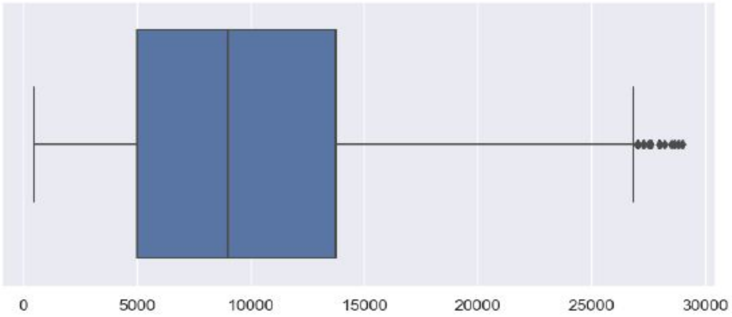
Univariate Analysis

Loan Amount:

Observations:

1. Most of the loan amount applied was in the range of 5k-14k.
2. Max Loan amount applied was ~27k.



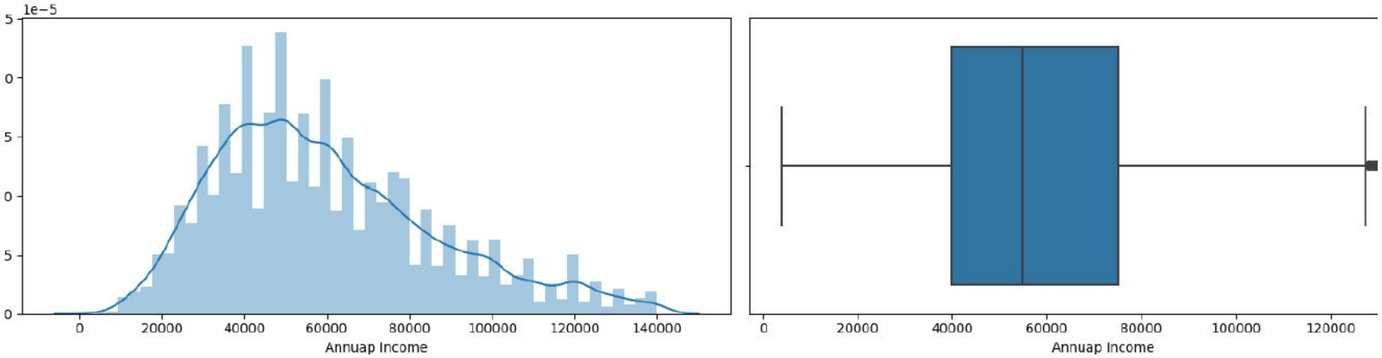


Univariate Analysis

Loan Amount:

Observations:

1. The Annual income of most if applicants lies between 40k-75k.
2. Average Annual Income is : 59883.0

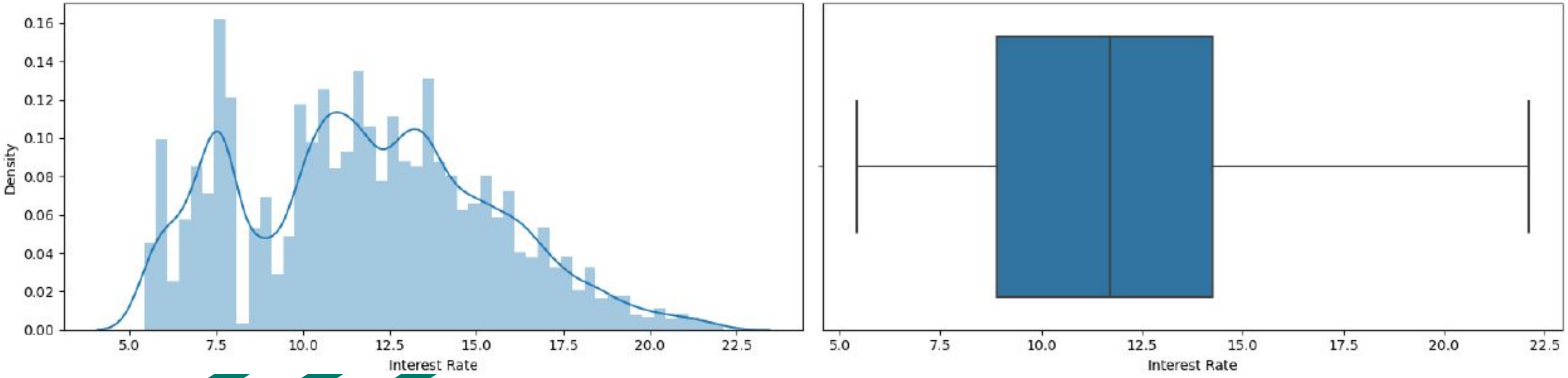


Univariate Analysis

Interest Rate:

Observations:

1. Most of the applicant's rate of interest is between in the range of 8%-14%
2. Average Rate of interest of rate is 11.7 %

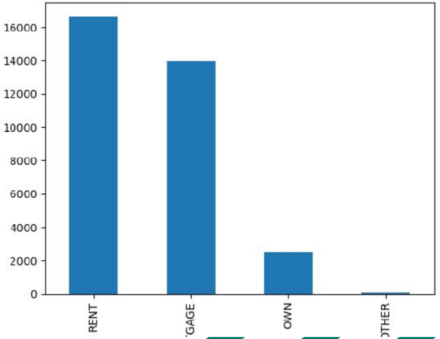


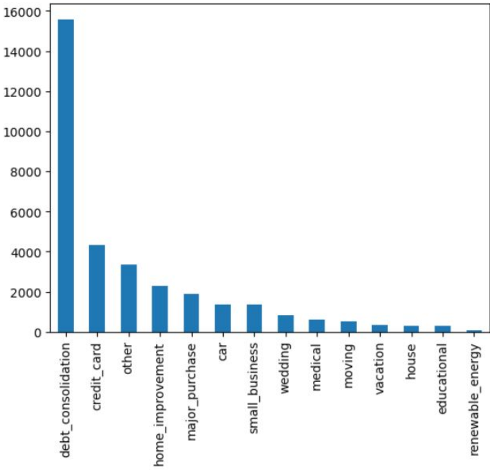
Univariate Analysis

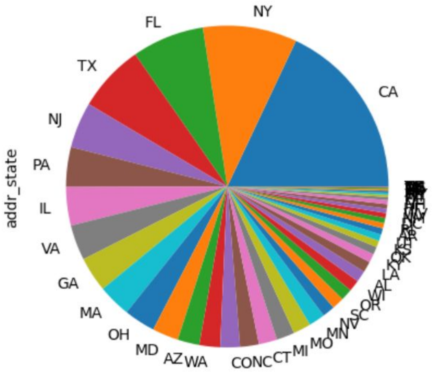
Unordered & Ordered Categorical Variable:

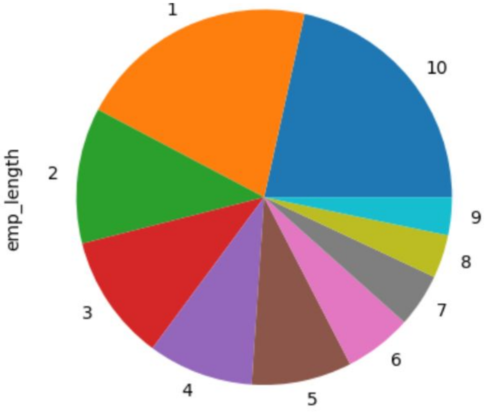
Analysis:

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









Bivariate Analysis

Interest Rate vs Charged off

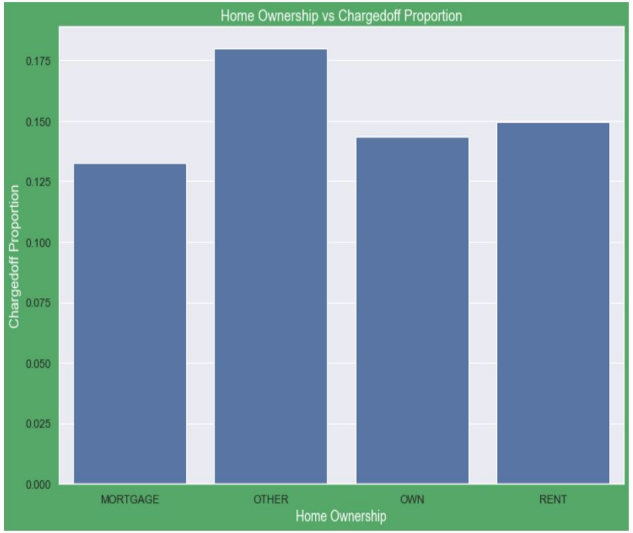
Observations:

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

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

- Charged off proportion is increasing with higher interest rates.

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



Bivariate Analysis

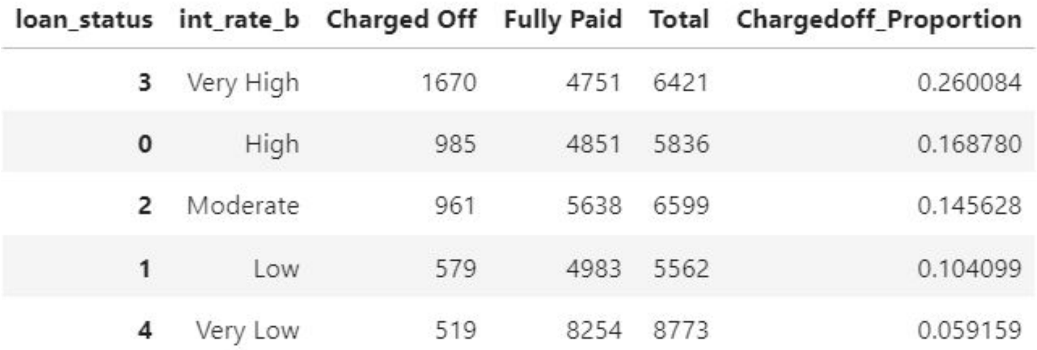
Annual income vs Charged Off

Observations:

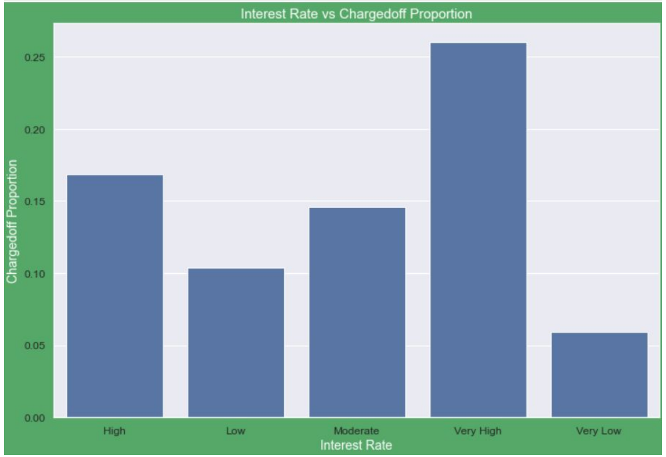
- Income range 80000+ has less chances of charged off

.• Income range 0-20000 has high chances of charged off.

- Notice that with increase in annual income charged off proportion got decreased.



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



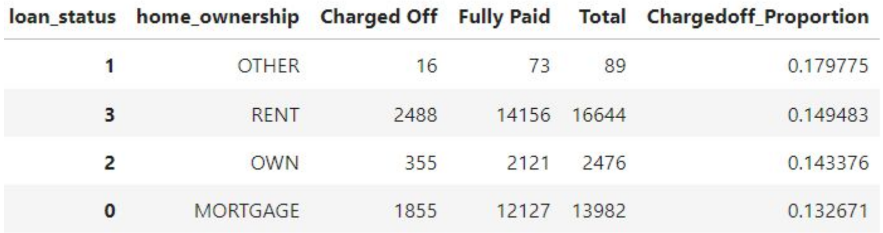
Bivariate Analysis

Home Ownership vs Charged off

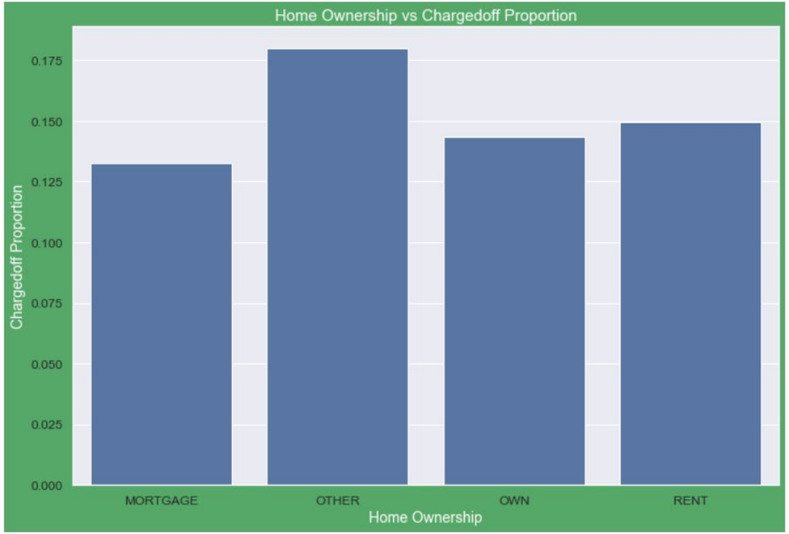
Observations:

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

- From the graph even shows high chances of charged off. Proportions, but data available is very limited compared to other points



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

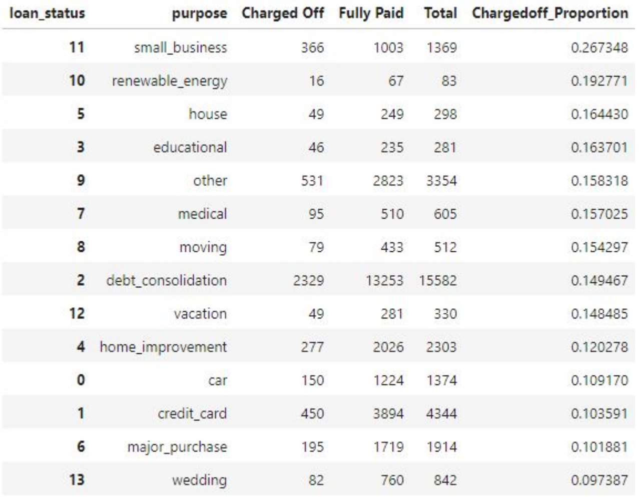


Bivariate Analysis

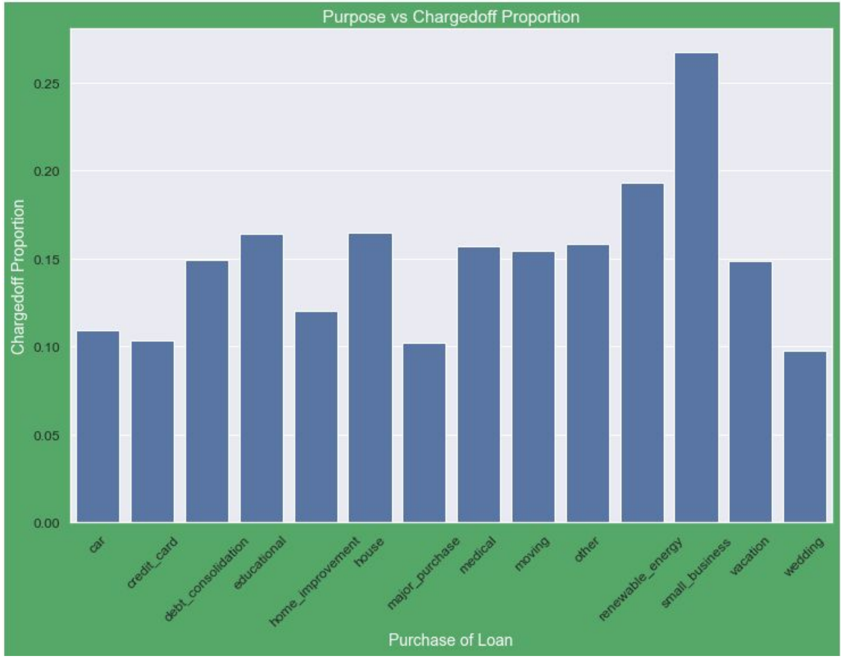
Purpose vs Charged Off

Observations:

- Those applicants who is having home loan is having low chances of loan defaults.
- Those applicants having loan for small business is having high chances for loan defaults.



loan_status	purpose	Charged Off	Fully Paid	Total	Chargedoff_Proportion
11	small_business	366	1003	1369	0.267348
10	renewable_energy	16	67	83	0.192771
5	house	49	249	298	0.164430
3	educational	46	235	281	0.163701
9	other	531	2823	3354	0.158318
7	medical	95	510	605	0.157025
8	moving	79	433	512	0.154297
2	debt_consolidation	2329	13253	15582	0.149467
12	vacation	49	281	330	0.148485
4	home_improvement	277	2026	2303	0.120278
0	car	150	1224	1374	0.109170
1	credit_card	450	3894	4344	0.103591
6	major_purchase	195	1719	1914	0.101881
13	wedding	82	760	842	0.097387

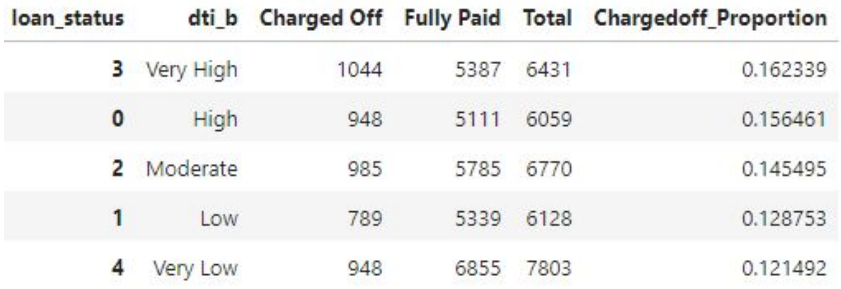


Bivariate Analysis

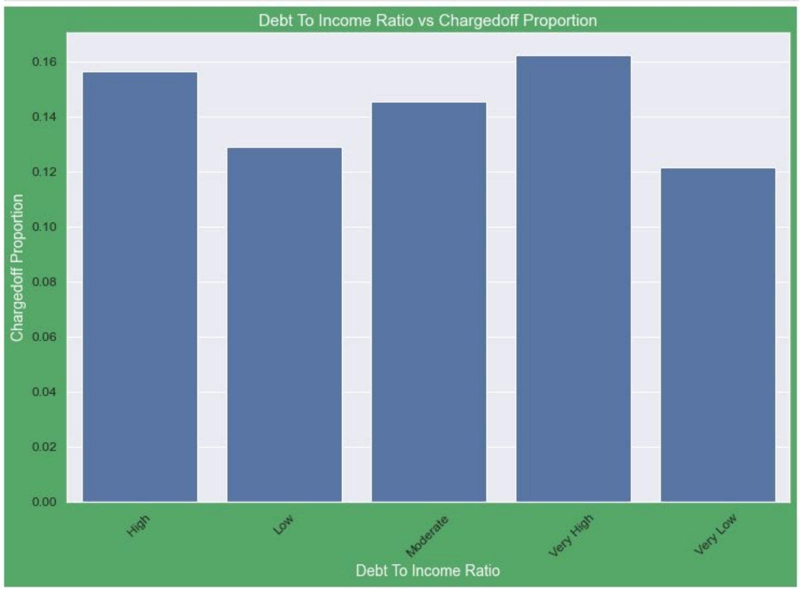
DTI Vs Charged off

Observations:

- High DTI value having high risk of defaults
- Lower the DTO having low chances loan defaults



loan_status	dti_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
3	Very High	1044	5387	6431	0.162339
0	High	948	5111	6059	0.156461
2	Moderate	985	5785	6770	0.145495
1	Low	789	5339	6128	0.128753
4	Very Low	948	6855	7803	0.121492

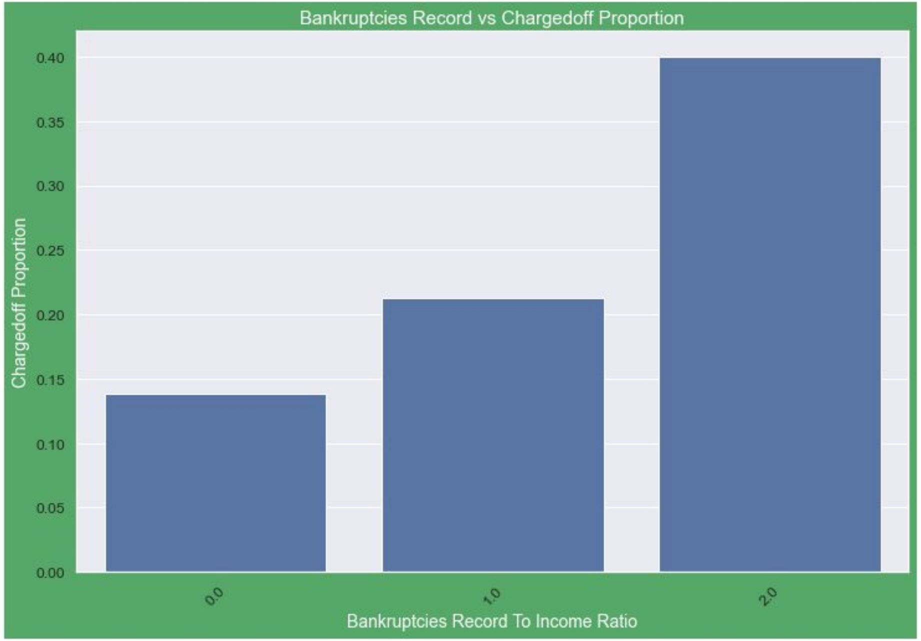


Bivariate Analysis

Bankruptcies Record vs Charged off

Observations:

- Bankruptcies Record with 2 is having high impact on loan defaults
- Bankruptcies Record with 0 is low impact on loan defaults
- Lower the Bankruptcies lower the risk.

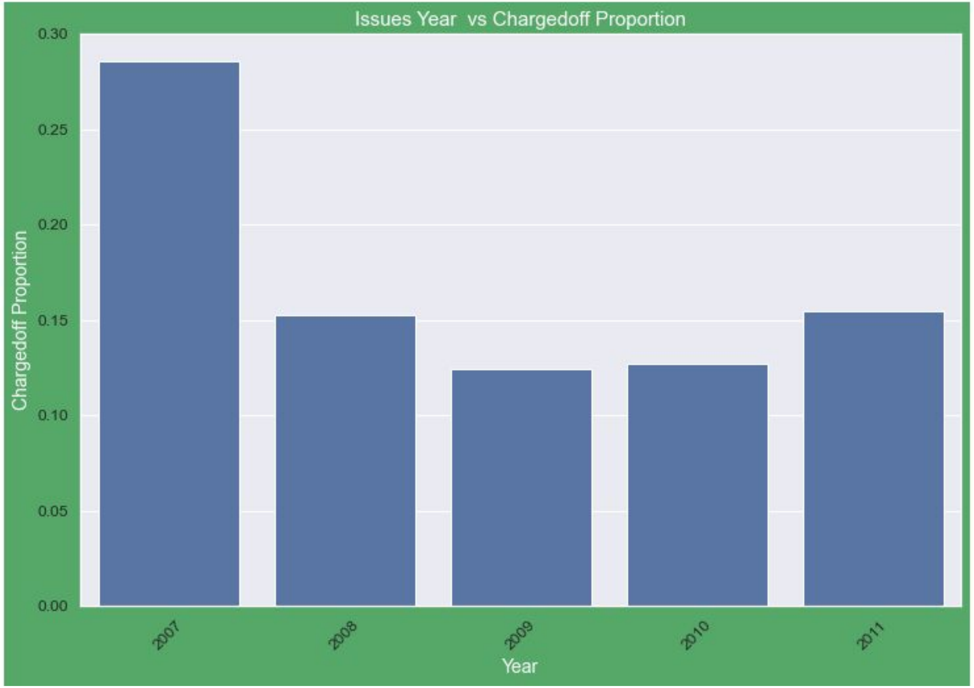


Bivariate Analysis

Issue Year vs Charged off

Observations:

- Year 2007 is highest loan defaults.
- 2009 is having lowest loan defaults.

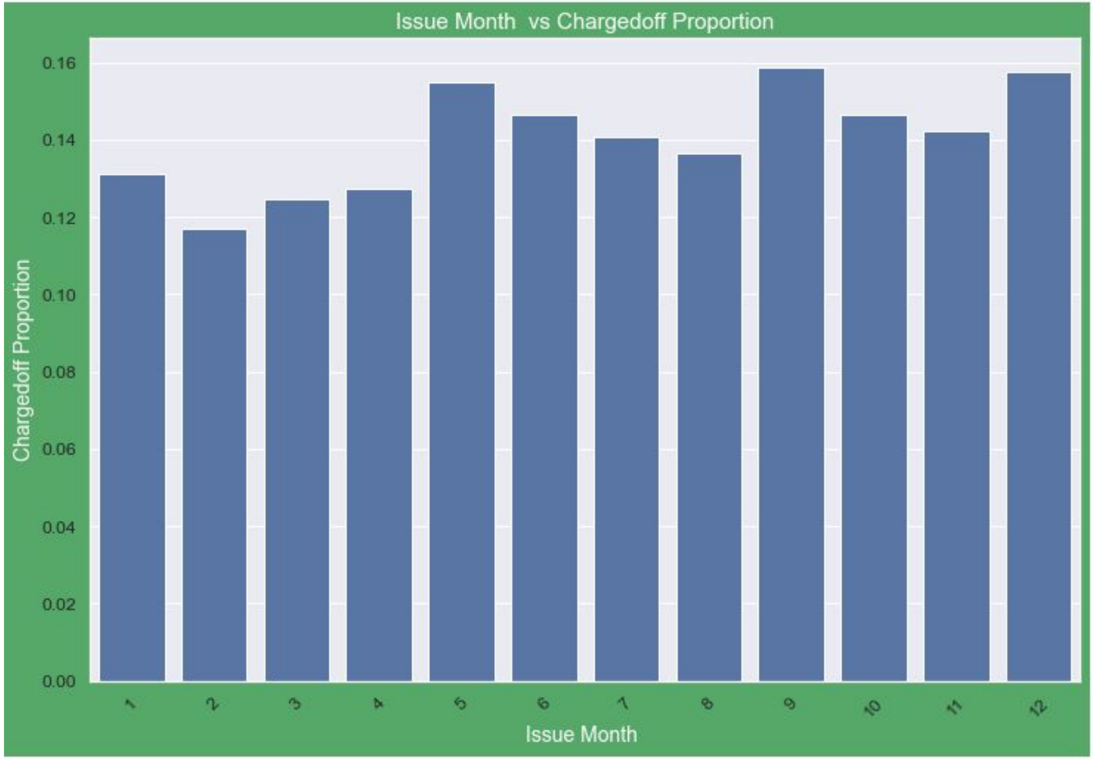


Bivariate Analysis

Issue Month Vs Charged off

Observations:

- Those loan has been issued in May, September and December is having high number of loan defaults
- Those loan has been issued in month of February is having high number of loan defaults
- Majority of loan defaults coming from applicants whose loan has been approved from September-to December

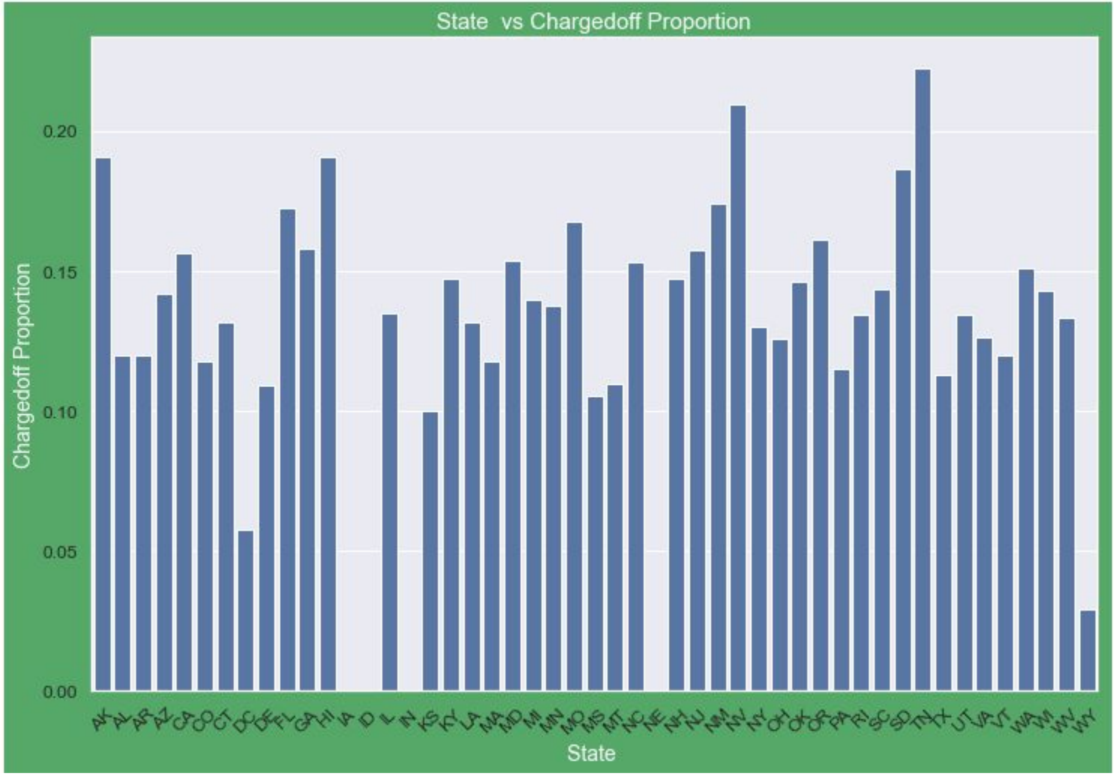


Bivariate Analysis

State vs Chargedoff

Observations:

- DE States is holding highest number of loan defaults.
- CA is having low number of loan defaults

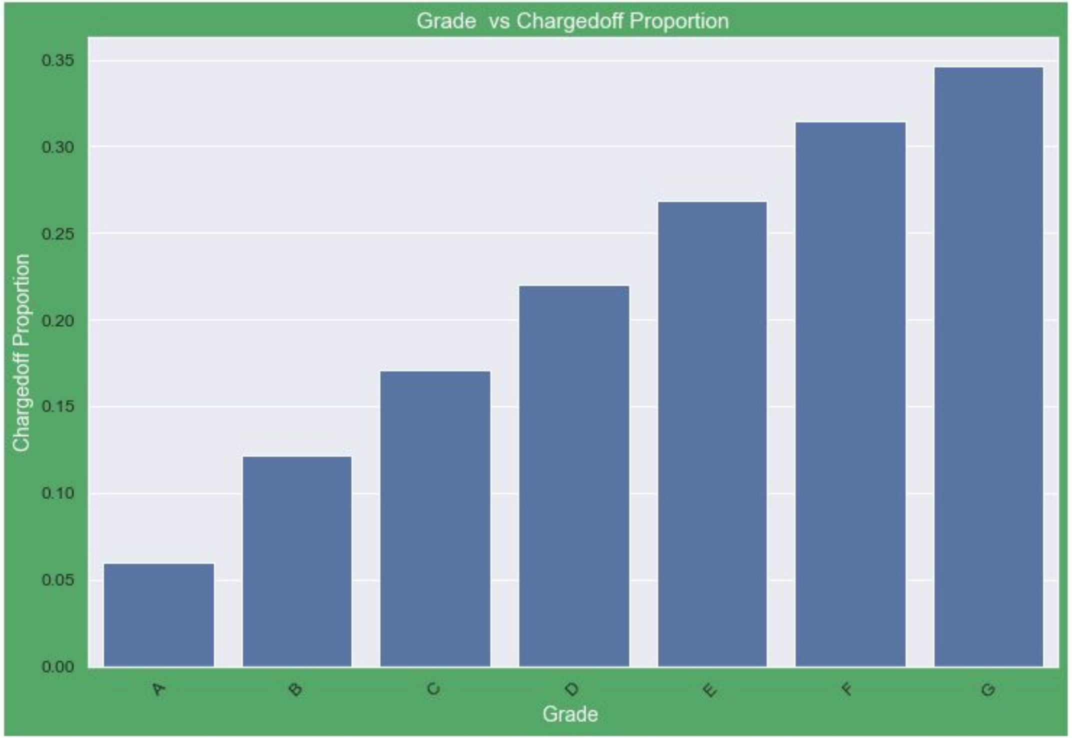


Bivariate Analysis

Grade vs ChargedOff

Observations:

- The Loan applicants with loan Grade G is having highest Loan Defaults.
- The Loan applicants with loan A is having lowest Loan Defaults.



Correlation

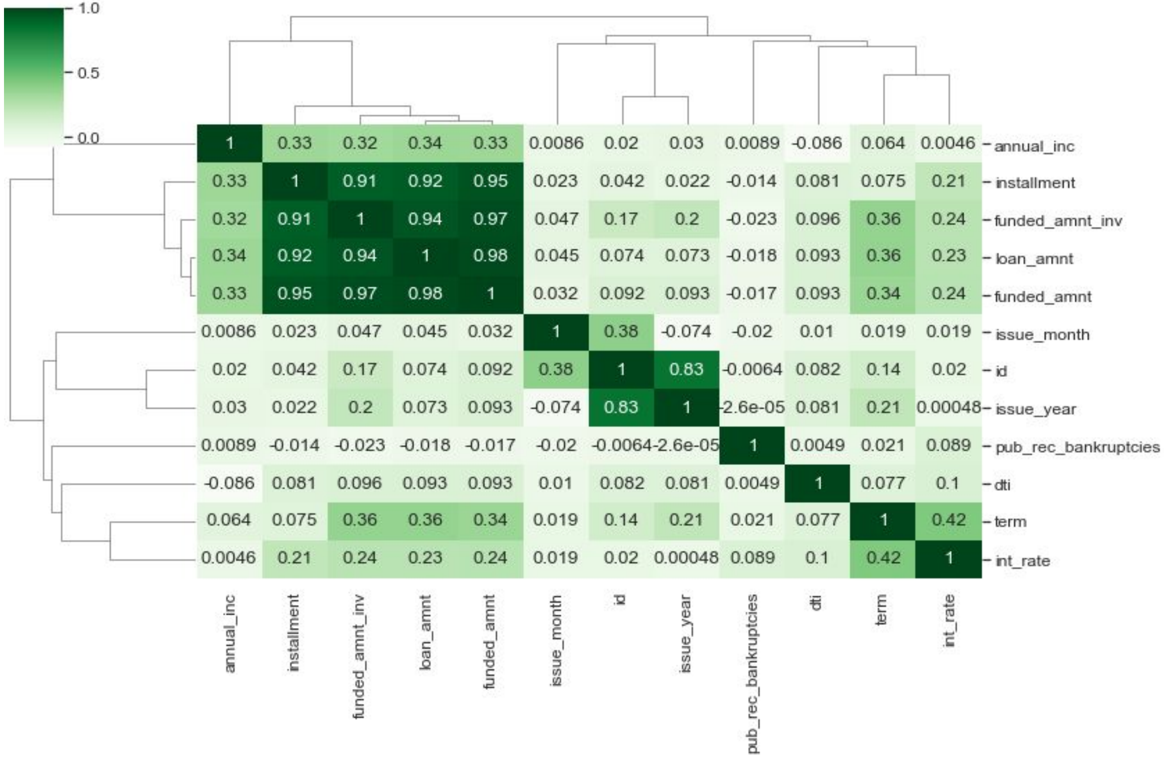
Observations:

- **Negative Correlation:**

1. Loan amount is negatively correlated with public record bankruptcies.
2. Annual income has a negative correlation with debt-to-income ratio (DTI).

- **Strong Correlation:**

1. Loan term has a strong correlation with the loan amount.
2. Loan term is strongly correlated with the interest rate.
3. Annual income shows a strong correlation with the loan amount.



	annual_inc	installment	funded_amnt_inv	loan_amnt	funded_amnt	issue_month	id	issue_year	pub_rec_bankruptcies	dti	term	int_rate
annual_inc	1	0.33	0.32	0.34	0.33	0.0086	0.02	0.03	0.0089	-0.086	0.064	0.0046
installment	0.33	1	0.91	0.92	0.95	0.023	0.042	0.022	-0.014	0.081	0.075	0.21
funded_amnt_inv	0.32	0.91	1	0.94	0.97	0.047	0.17	0.2	-0.023	0.096	0.36	0.24
loan_amnt	0.34	0.92	0.94	1	0.98	0.045	0.074	0.073	-0.018	0.093	0.36	0.23
funded_amnt	0.33	0.95	0.97	0.98	1	0.032	0.092	0.093	-0.017	0.093	0.34	0.24
issue_month	0.0086	0.023	0.047	0.045	0.032	1	0.38	-0.074	-0.02	0.01	0.019	0.019
id	0.02	0.042	0.17	0.074	0.092	0.38	1	0.83	-0.0064	0.082	0.14	0.02
issue_year	0.03	0.022	0.2	0.073	0.093	-0.074	0.83	1	-2.6e-05	0.081	0.21	0.00048
pub_rec_bankruptcies	0.0089	-0.014	-0.023	-0.018	-0.017	-0.02	-0.0064	-2.6e-05	1	0.0049	0.021	0.089
dti	-0.086	0.081	0.096	0.093	0.093	0.01	0.082	0.081	0.0049	1	0.077	0.1
term	0.064	0.075	0.36	0.36	0.34	0.019	0.14	0.21	0.021	0.077	1	0.42
int_rate	0.0046	0.21	0.24	0.23	0.24	0.019	0.02	0.00048	0.089	0.1	0.42	1

Conclusions

Key Factors Influencing Loan Defaults

- **Loan Term**: The average interest rate for defaulted applications is significantly high—12.38% for 36-month terms and 15.75% for 60-month terms.
- **Grade**: Default rates are higher among high-risk loan applicants. It is important for Lending Club (LC) to thoroughly vet these high-risk applicants.
- **Loan Amount**: The defaulter rate increases as the requested loan amount rises.
- **Annual Income**: Applicants from the 'Low' (≤45K USD) and 'Medium' (45K-90K USD) income groups have a larger share of defaulted loans.
- **Income Range**: Income ranges between 0-20,000 USD have a high likelihood of being charged off.
- **Interest Rate**: Interest rates above 16% show a greater chance of being charged off compared to lower interest rate categories.
- **Home Ownership**: Applicants who do not own a home have a higher probability of becoming loan defaulters.
- **Small Business Loans**: Applicants with loans for small businesses are more likely to default.
- **Debt-to-Income Ratio (DTI)**: A high DTI ratio is associated with a greater risk of loan default.
- **Bankruptcies**: A higher number of bankruptcies increases the chances of loan default.
- **State of Delaware (DE)**: DE has the highest number of loan defaults.