LENDING CLUB CASE STUDY

Krishna Sai Sangaraju

GuruKeerthana Gaddam

(ML C63)

LENDING CLUB CASE STUDY

Problem Statement

Lending Club, a leading online marketplace for consumer loans, faces a crucial dilemma in its loan approval process. They need to make smart decisions to minimize financial losses, especially those caused by loans to high-risk borrowers. These losses, known as credit losses, occur when borrowers default on their loans (fail to repay). Borrowers categorized as "charged off" contribute most significantly to these losses.

This analysis aims to help Lending Club reduce credit losses by addressing two challenges:

Avoiding risky borrowers: Conversely, approving loans for borrowers unlikely to repay can lead to significant financial losses for Lending Club.

Identifying Default Risk with Exploratory Data Analysis (EDA)

Our approach leverages Exploratory Data Analysis (EDA) using the provided dataset. This analysis aims to pinpoint the key factors (driver variables) that strongly indicate a borrower's risk of defaulting on a loan. By understanding these factors, Lending Club can improve its loan portfolio management and risk assessment strategies.

In essence, Lending Club wants to know: What characteristics make a borrower more likely to default? This knowledge will empower them to make informed decisions about loan approvals and minimize credit losses.

DATA UNDERSTANDING

Loan Status: This key attribute indicates the outcome of past loans. Here's what each value means:

•**Fully Paid:** Borrowers who successfully repaid the entire loan amount (principal + interest).

•**Charged-Off:** Borrowers who defaulted on their loans.

•**Current (Excluded):** Loans still being repaid (not defaulted yet). These are excluded for this analysis.

Decision Matrix: There are two main loan outcomes:

•**Loan Accepted:**

- **Fully Paid:** Borrowers repaid the loan in full.
- **Charged-Off:** Borrowers defaulted on the loan.

•**Loan Rejected:** Applications that didn't meet our criteria, so this data is not available.

Key Predictors: These attributes are crucial for predicting loan approval and potential defaults. Note that some may be excluded due to missing data.

Customer Demographics:

•**Annual Income (annual_inc):** Higher income suggests a better chance of loan repayment.

•**Home Ownership (home_ownership):** Owning a home increases loan approval likelihood (collateral).

•**Employment Length (emp_length):** Longer employment indicates financial stability and higher approval chances.

•**Debt-to-Income Ratio (dti):** Lower DTI reflects a higher ability to repay the loan.

•**State (addr_state):** Location data can reveal trends in delinquency or default rates.

DATA UNDERSTANDING

Loan Characteristics:

- **Loan Amount (loan_amt):** Amount of money requested by the borrower.
- **Grade (grade):** Creditworthiness rating associated with the loan's risk level.
- **Term (term):** Duration of the loan (months).
- **Loan Date (issue_d):** Date the loan was approved.
- **Purpose (purpose):** Reason for the loan (debt consolidation, home improvement, etc.).
- **Verification Status (verification_status):** Whether the borrower's information is verified.
- **Interest Rate (int_rate):** Annual interest rate charged on the loan.
- **Installment (installment):** Regular monthly payment amount.
- **Public Records (public_rec):** Number of negative public records impacting loan risk.
- **Public Records Bankruptcy (public_rec_bankruptcy):** Number of local bankruptcy records. (Higher = Lower Approval Chance)

APPROACH

Loading Data: During dataset loading, variables with mixed data types were converted accordingly.

Handling Null Values: Dropped columns with null values, and imputed columns for null percent less percents and remove rows if it has less than 1%

Unique Value Check: Removed 9 columns with only a single unique value.

Duplicate Rows Check: No duplicates found.

Dropping Records and Columns: Removed records where loan_status = ‘Current’ as it does not help us for our analysis.

Fix rows and columns: For int_percent removed ‘%’ at last and renamed to int_rate_percent

Derived Metrics: Created ‘issue_year’, ‘issue_month’ and few range columns to convert numerical to categorical data(funded_amnt, dti, installment, annual_inc, int_rate_percent)

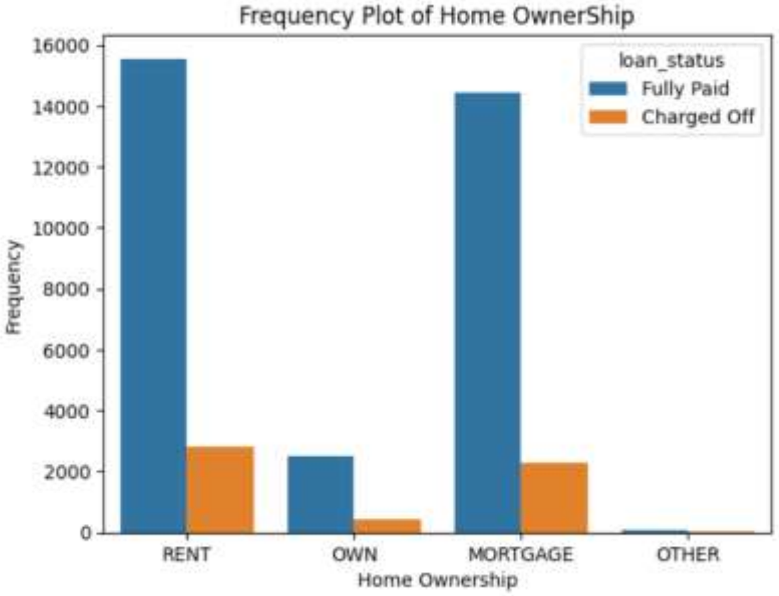
Outliners Treatment: Checked continuous columns and found there are outliers in annual_inc and remove rows which has value greater than its 0.99 quantile

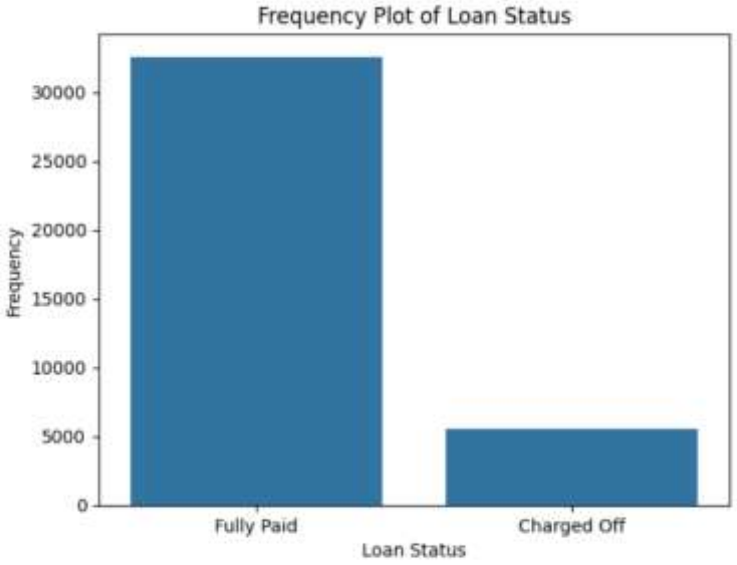
After cleaning, the dataset contained 38,481 rows and 31 columns(range columns of 5 and issue_month and issue_year added).

UNIVARIATE ANALYSIS

Unordered Categorical Variables :

- loan_status
- home_ownership
- purpose
- addr_state

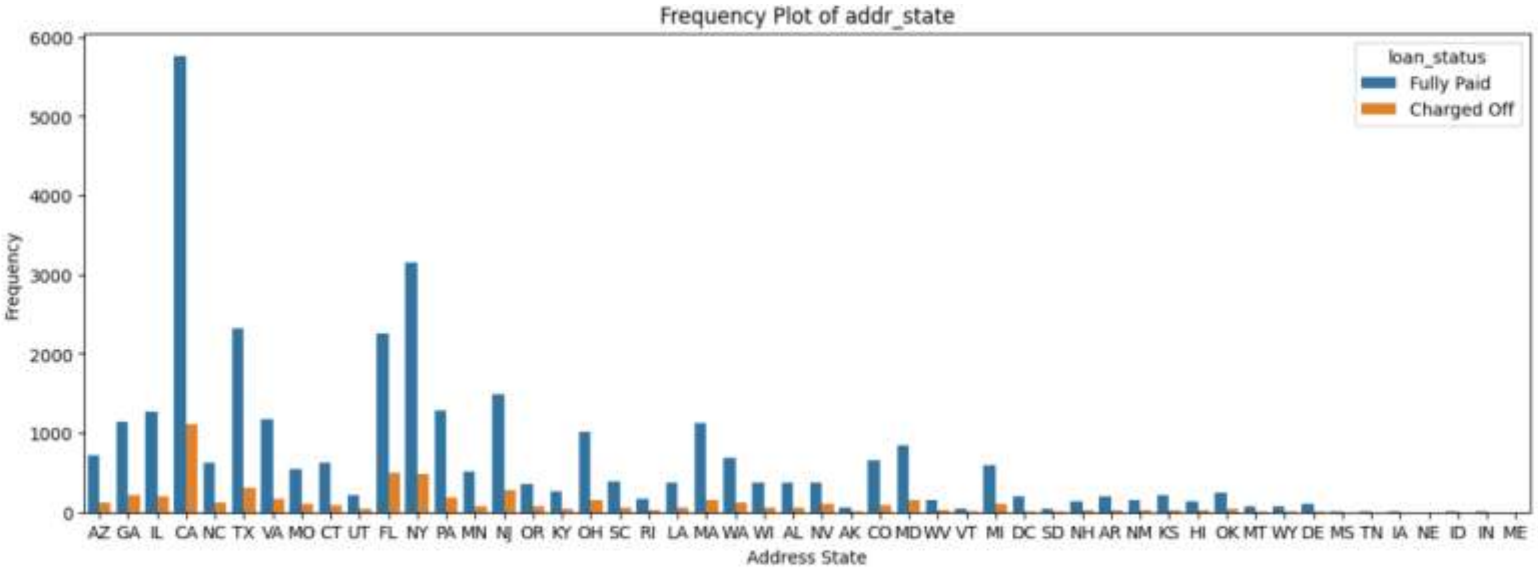




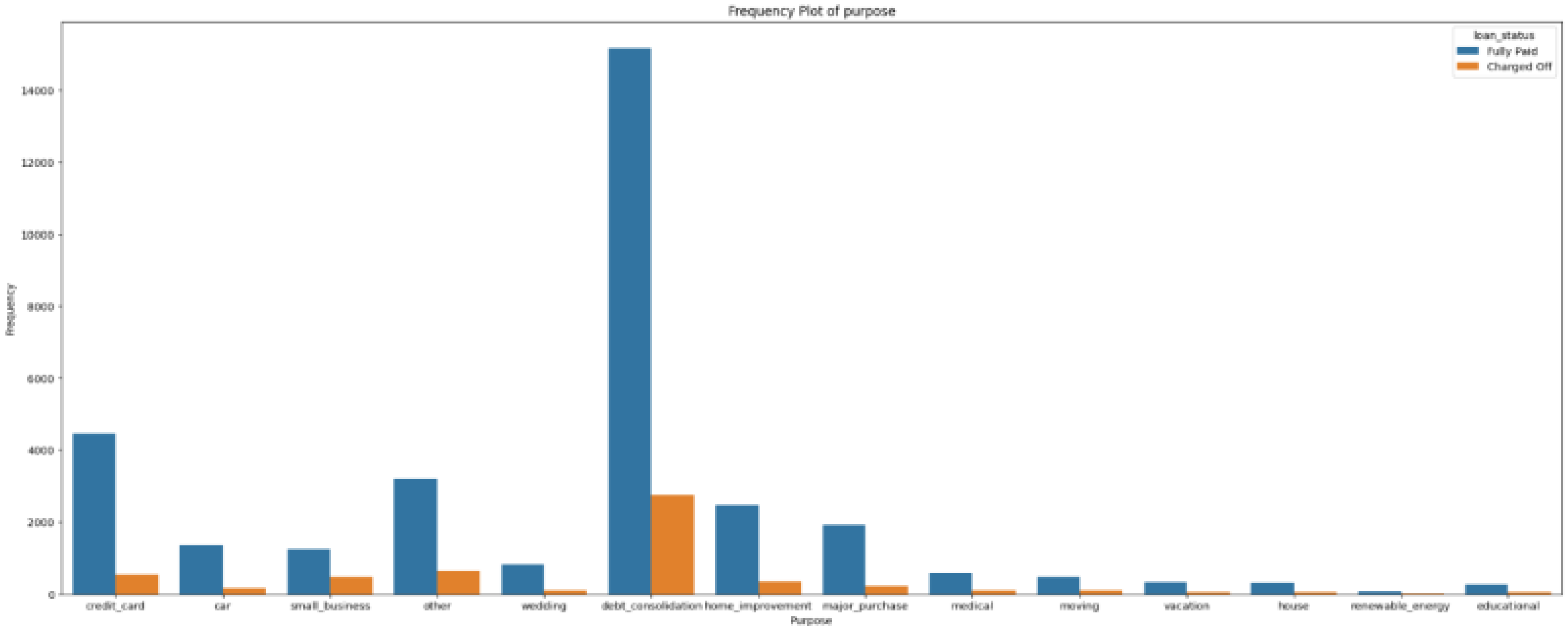
Applicants with Own house are likely to payback than RENT and MORTGAGE Home_ownership. RENT ,MORTGAGE can be used for further analysis

Approximately 14.6% of loans are defaulted ,thus focussing on Charged Off gives us insights in taking decision

UNIVARIATE ANALYSIS



California (CA), Florida (FL) and New York (NY) have highest rate of defaulters. addr-state may not be a major contributor for defauters but it can add value while taking optimal decisions.



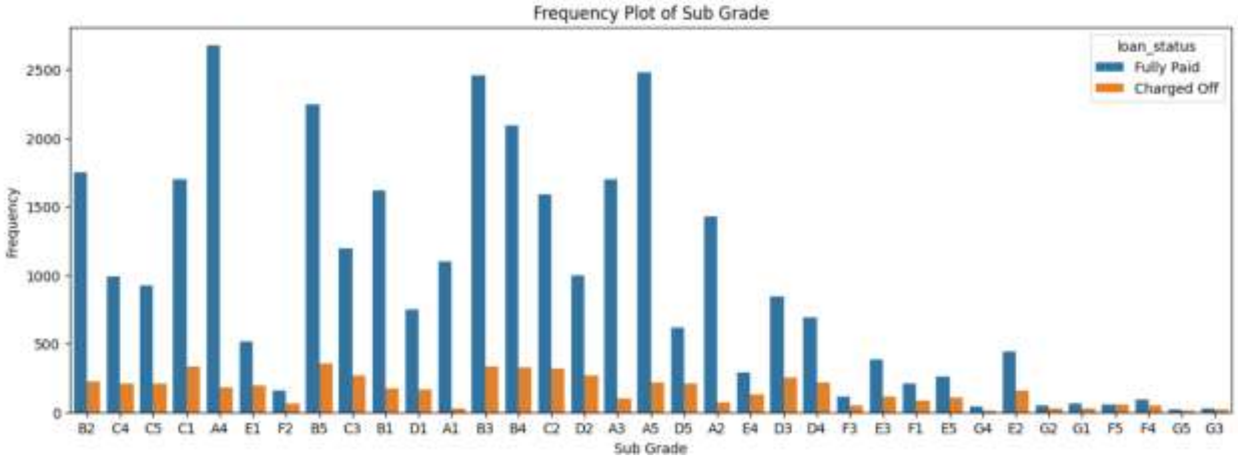
Majority of people are acquiring loans for below purposes

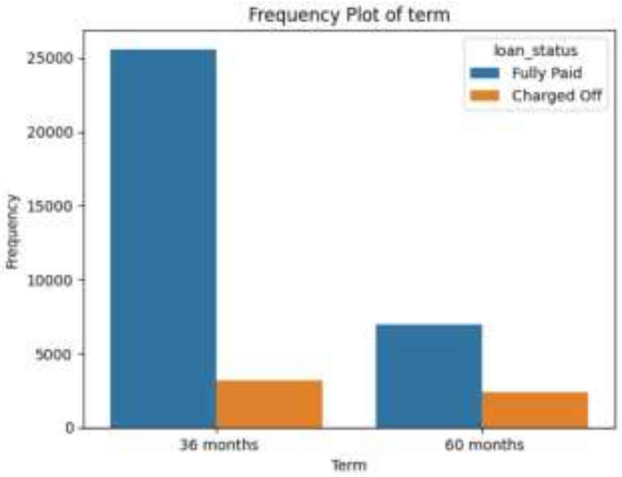
To clear other debts (debt_consolidation, credit card) which is why they are not able to manage finances and loan getting charged off. To start up a small business loan is acquired and used as investment. In both the above reasons risk factor is high due to unpredictable income

UNIVARIATE ANALYSIS

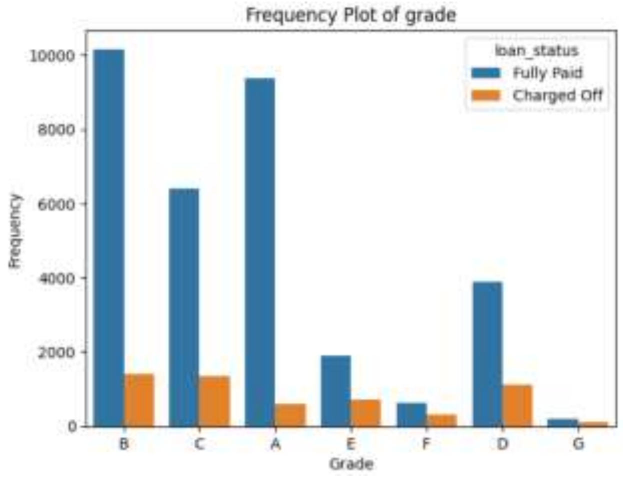
Ordered Categorical Variables

- term
- grade
- sub_grade
- emp_length
- issue_year
- issue_month





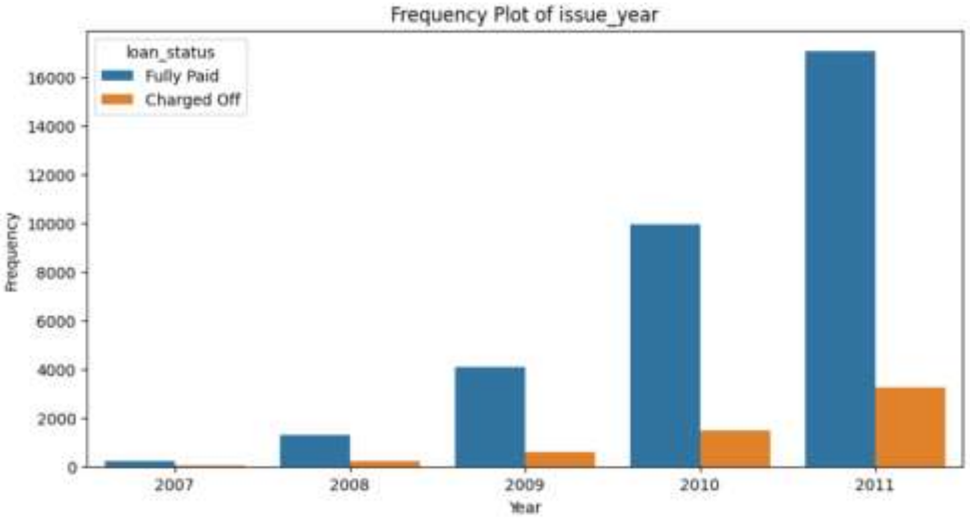
Borrowers with Charged Off loans are more likely to have sub-grades B3, B4, B5 (within Grade B), C1, C2, C3 (within Grade C), and D2, D3 (within Grade D) from Univariate analysis



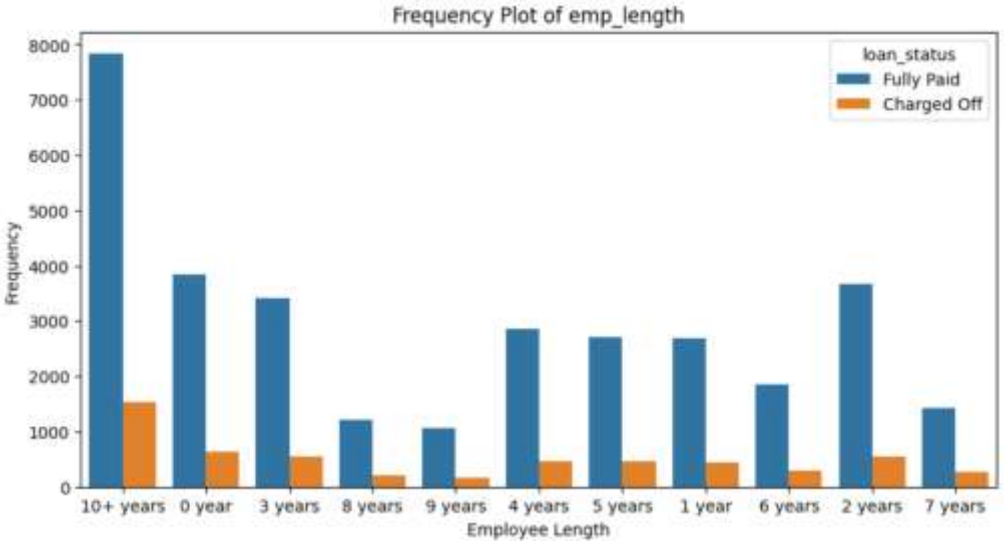
Borrowers with Charged Off loans are more likely to have chosen a 36-month loan term over 60 months from Univariate Analysis.

Borrowers with Charged Off loans are more likely to have received loan grades of B, C (or) D based on Univariate Analysis.

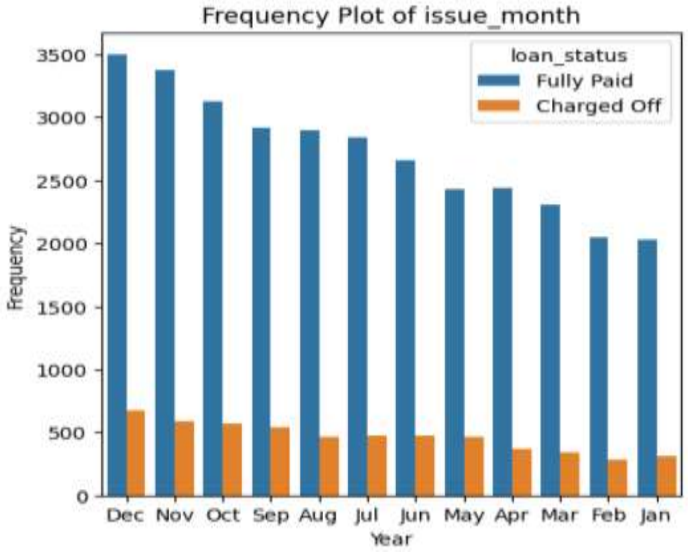
UNIVARIATE ANALYSIS



Charged Off seemed to super hike in 2011 which may have multiple reasons like economic crisis/recession.



Borrowers with Charged Off loans are more likely to fall into two employment length categories: very long tenure (10+ years) or very short tenure (less than a year) from Univariate Analysis.

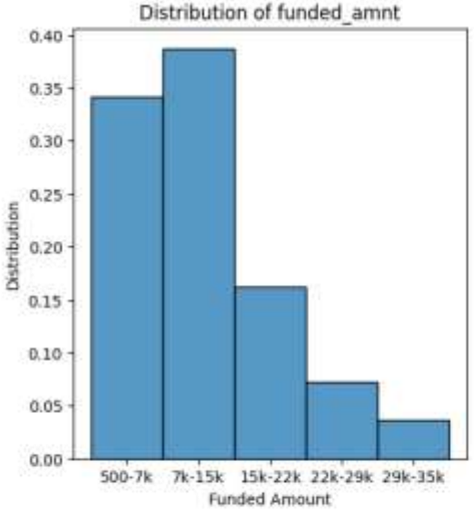


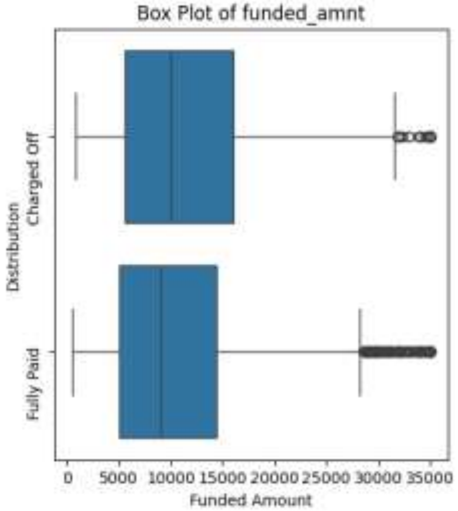
Borrowers with Charged Off loans are more likely to have received their loans in 3rd quarter of the year based on Univariate Analysis.

UNIVARIATE ANALYSIS

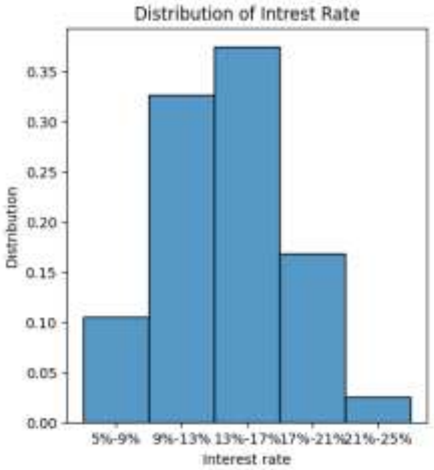
Quantitative Variables

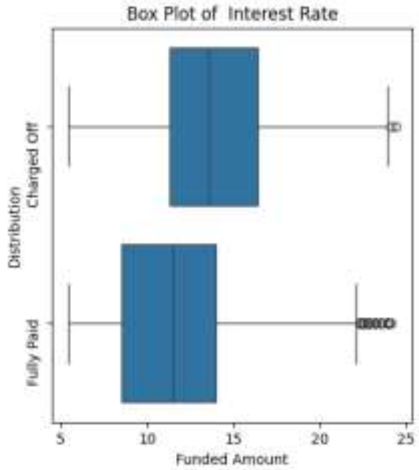
- funded_amnt
- int_rate_percent
- installment
- annual_inc
- dti

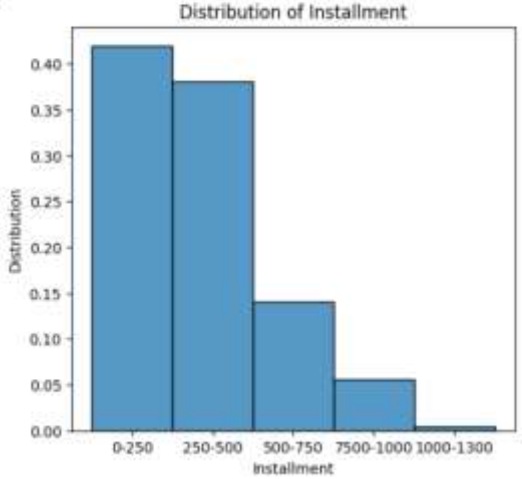


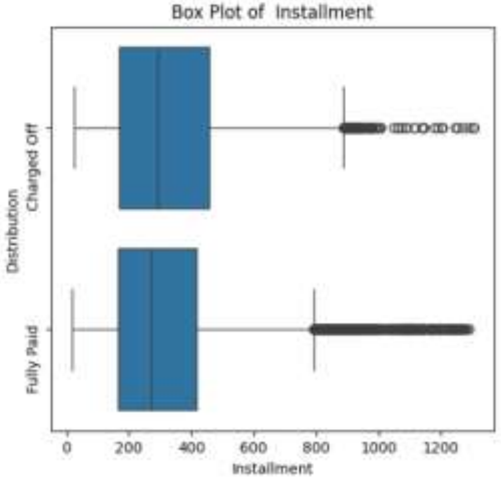


For Charged-off applicants in the analysis have a median funded amount of 10,000, with the middle 50% of applicants ranging between 5,500 and 16,000. Majority are found in range of 7k-15k





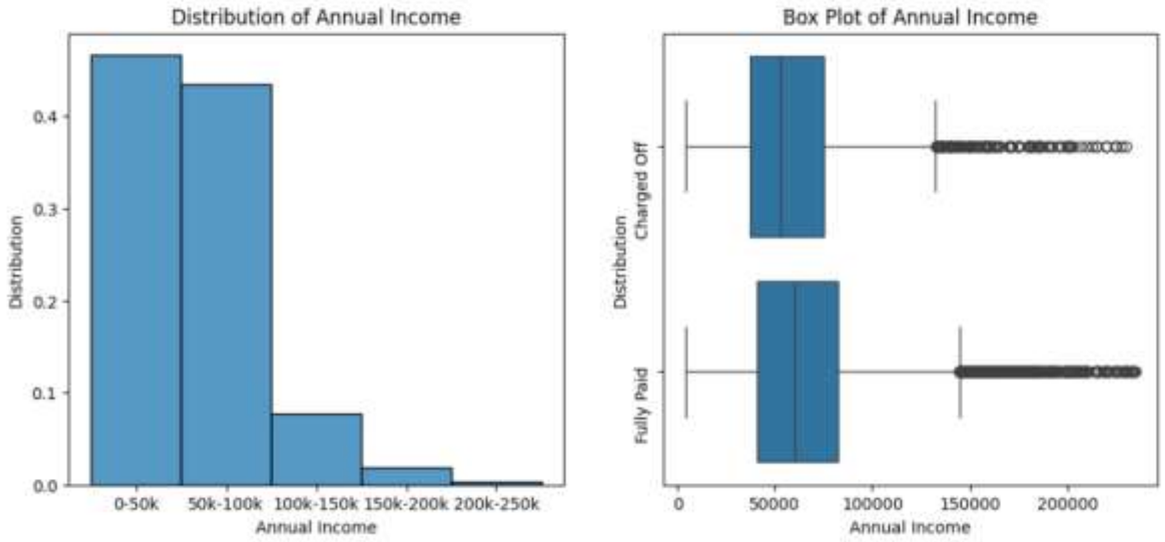


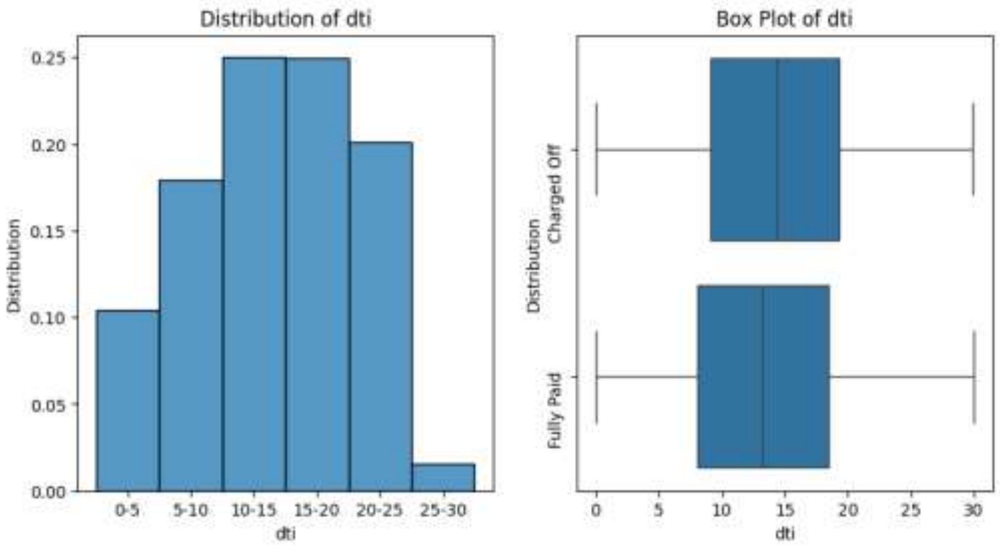


For charged-off applicants, the average interest rate is 13.57%. Additionally, the analysis shows that at least 25% of applicants had interest rates below 11.28%, while 75% had rates at or below 16.40%. Majority are found in range of 13%-17%

For charged-off applicants, installment show a central tendency of 292.04 . The analysis also indicates that at least 25% of applicants had installments below 168.45 and at least 75% had installments not exceeding 454.38. Majority are found in range of 0-250

UNIVARIATE ANALYSIS





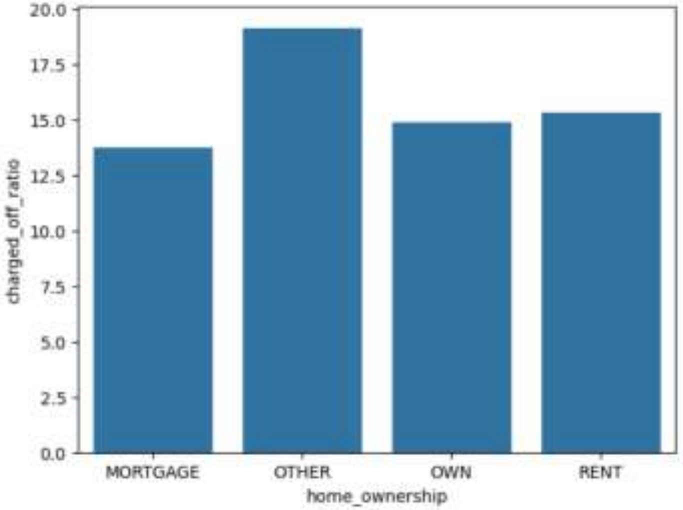
For charged-off applicants, Annual Income show a central tendency of 52,800 . The analysis also indicates that at least 25% of applicants had Annual Income below 37,000 and at least 75% had income not exceeding 74,879. Majority are found in range of 0-50k

For charged-off applicants, dti show a central tendency of 14.34 . The analysis also indicates that at least 25% of applicants had dti below 9.13 and at least 75% had rates not exceeding 19.31. Majority are found in range of 10-15

BIVARIATE ANALYSIS

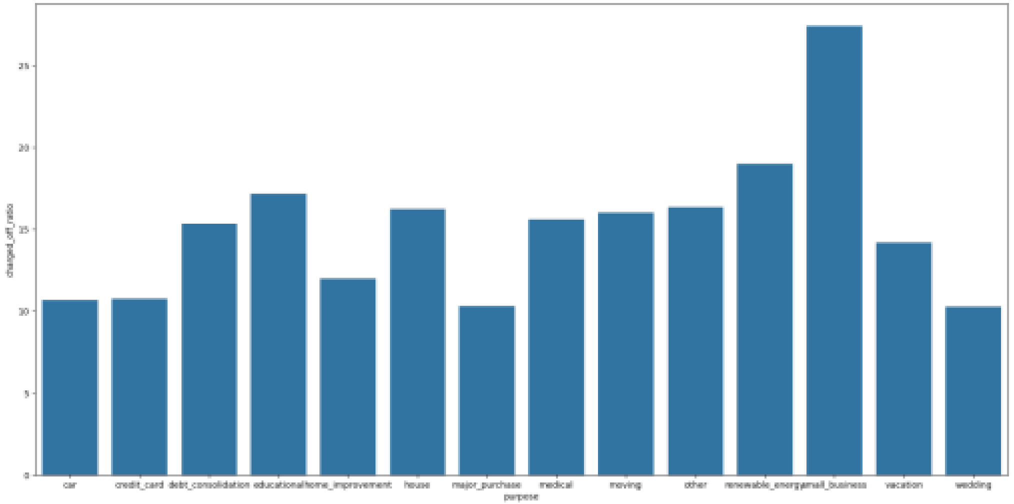
Categorical Variables:

- home_ownership
- purpose
- addr_state
- term
- grade
- sub_grade
- emp_length
- issue_year
- Issue_month
- funded_amnt_range
- int_rate_percent_range
- annual_inc_range
- Installment_range
- dti_range



Top 3 Charged Off Rates by Home Ownership

- OTHER = 18.75
- RENT = 15.35
- OWN = 14.93



Top 3 Charged Off Rates by Purpose

- small_business = 27.39
- renewable_energy = 19.00
- educational = 17.03

Top 5 Charged Off Rates by Address State

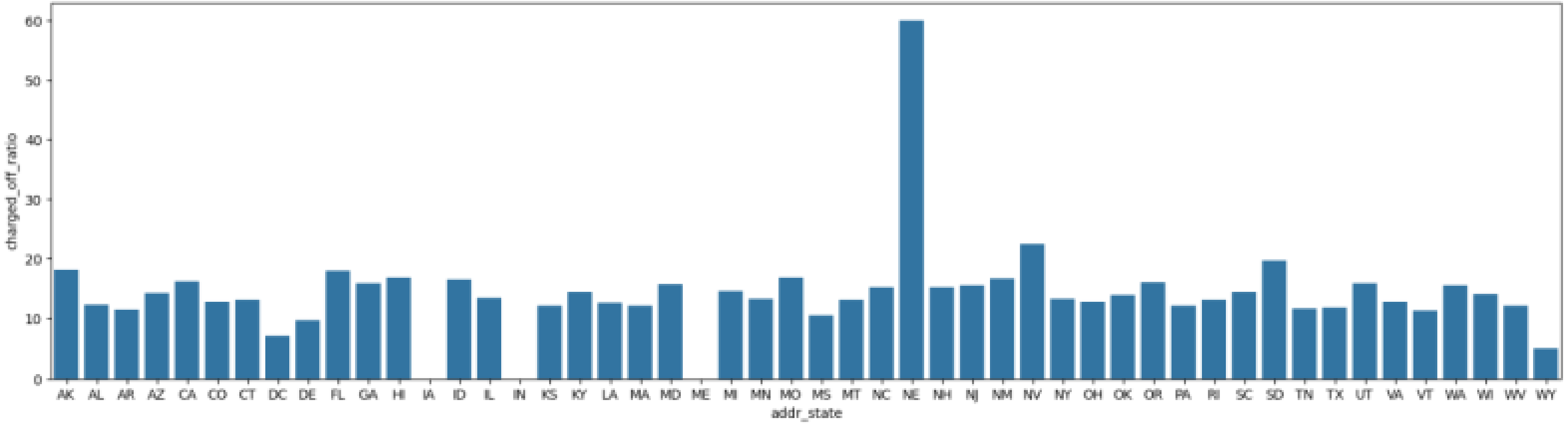
•NE (Nebraska) = 60.00 (Because it has total 5 loans out of them 3 are charged off)

•NV (Nevada) = 22.53

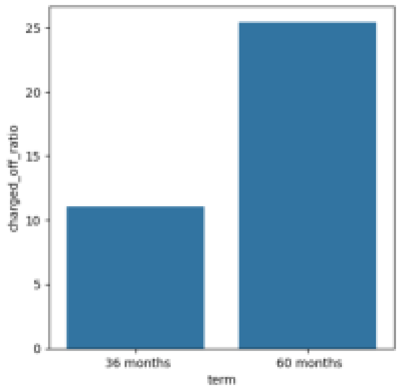
•SD (South Dakota) = 19.35

•AK (Alaska) = 18.18

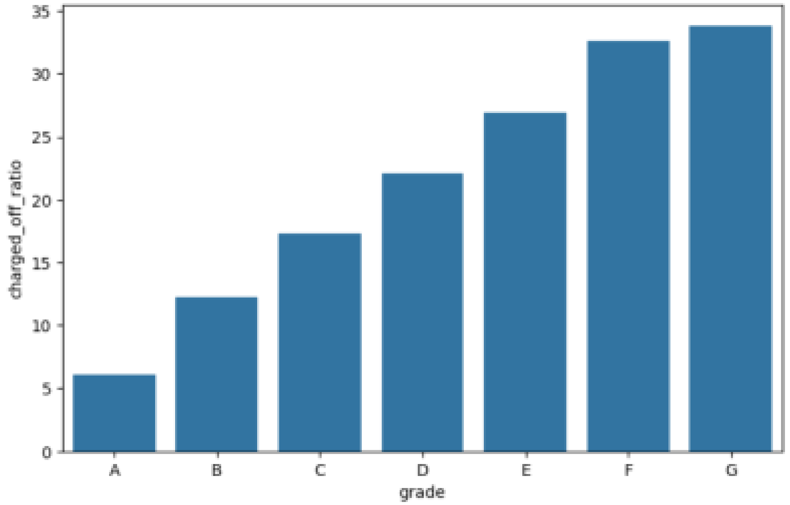
•FL (Florida) = 18.08



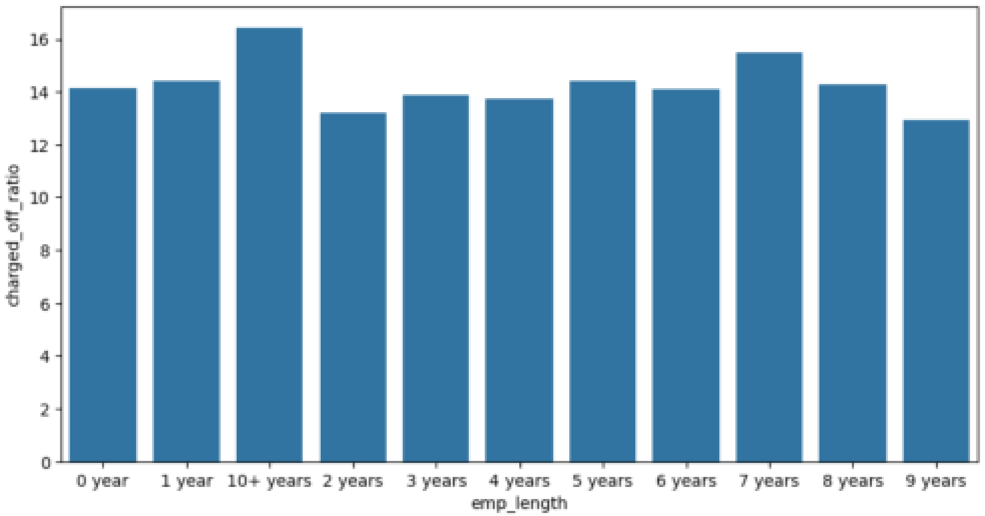
BIVARIATE ANALYSIS



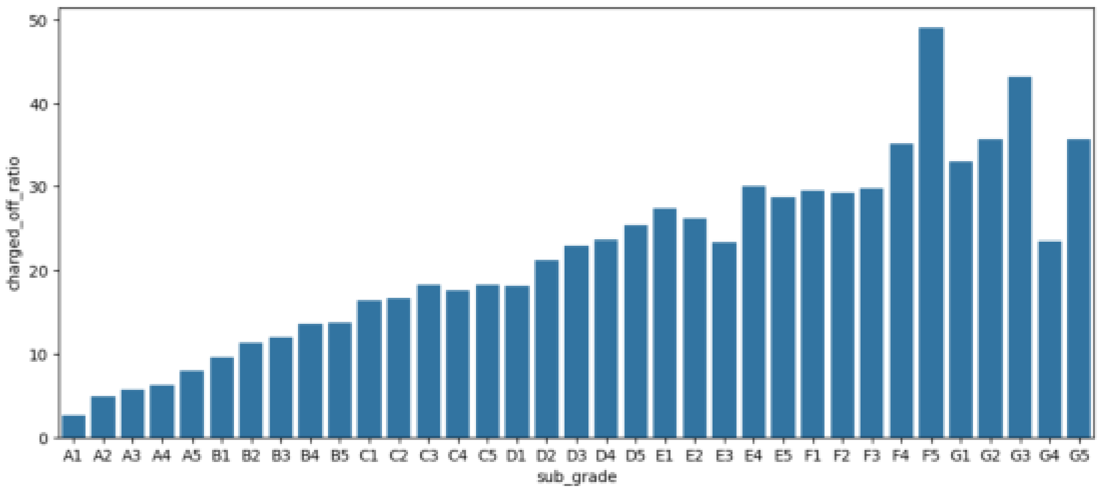
People who took loan for larger term/tenure have higher Charged offs like 60 months(charged off rate 22.70)



Highest grade being A and ranging so on. Grade plays a crucial role in defaulting process because lower the grade higher the risk of defaulting in future.

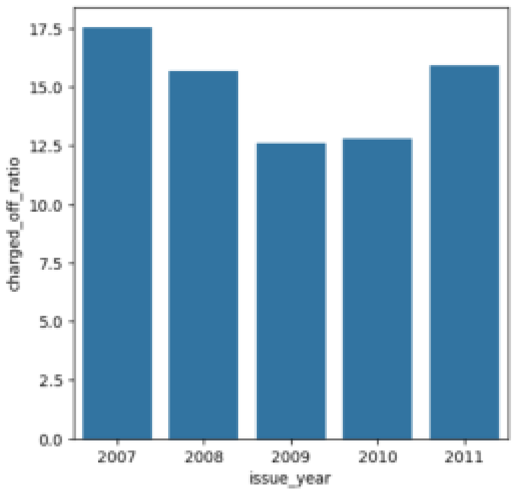


Majorly more the emp_length higher the ratio of getting charged off. Here are the top 3 => 10+ years, 7 years, 1 year



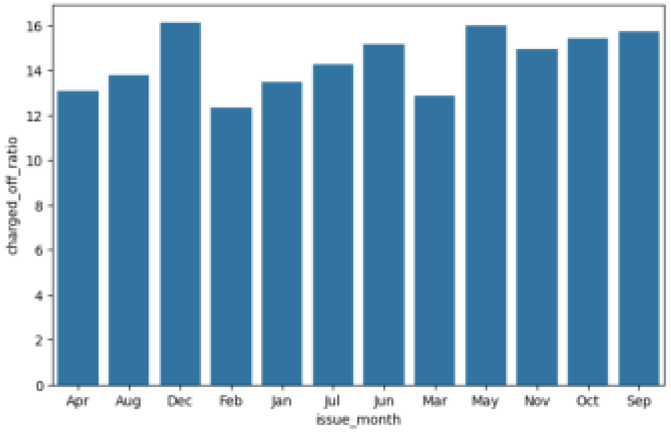
Highest sub-grade being A1, A2, A3, B1 and ranging so on.

•Sub-Grade can also play a crucial role in defaulting process because lower the sub-grade higher the risk of defaulting in future.

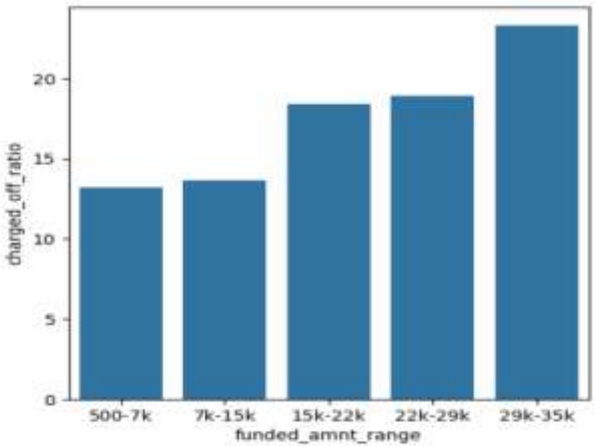


Examining the plots, we observe that the charged-off ratio decreased from 2007 to 2009, then gradually increased from 2009 to 2011. This does not significantly affect the overall charged-off rate.

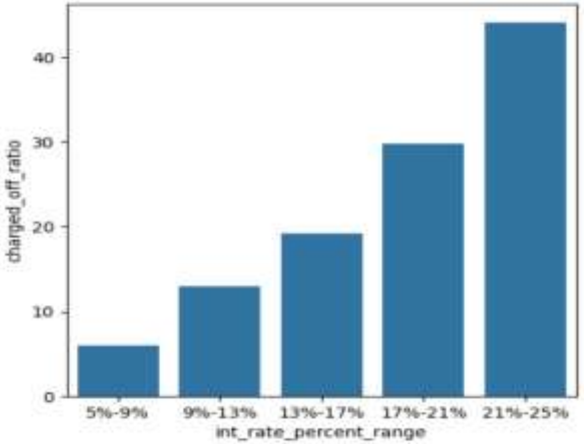
BIVARIATE ANALYSIS



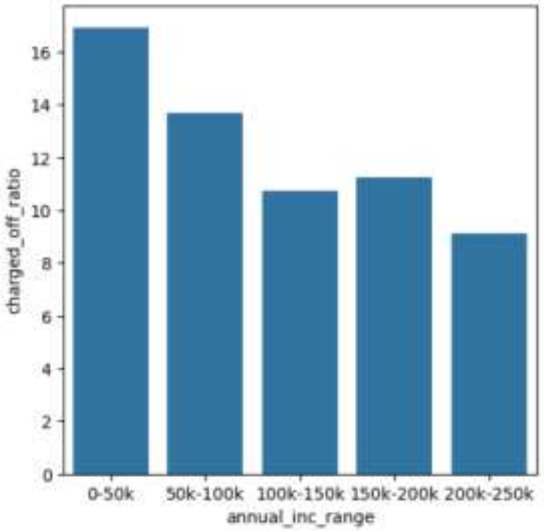
From above we can observe there is high charged off ratio in Dec, May, Sept.



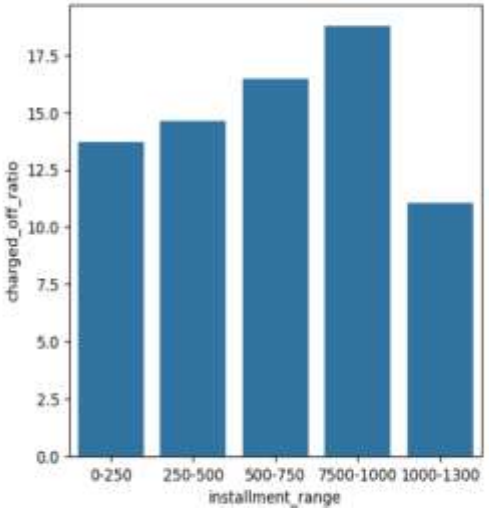
As the installment increases, the Charged Off Rate also increases.



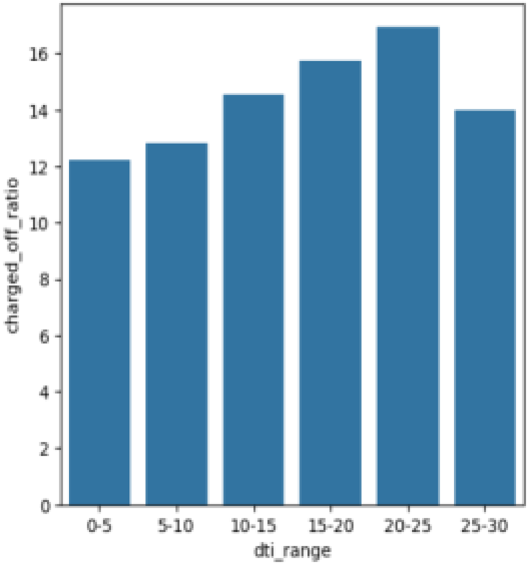
Loans with highest interest rates are going to be defaulted more.



Income of loan applicant plays a vital role in loan repayment which makes it as driving factor for analysis Higher the income less likely to get defaulted.



As the Funded Amount increases, the Charge Off Rate also Increases

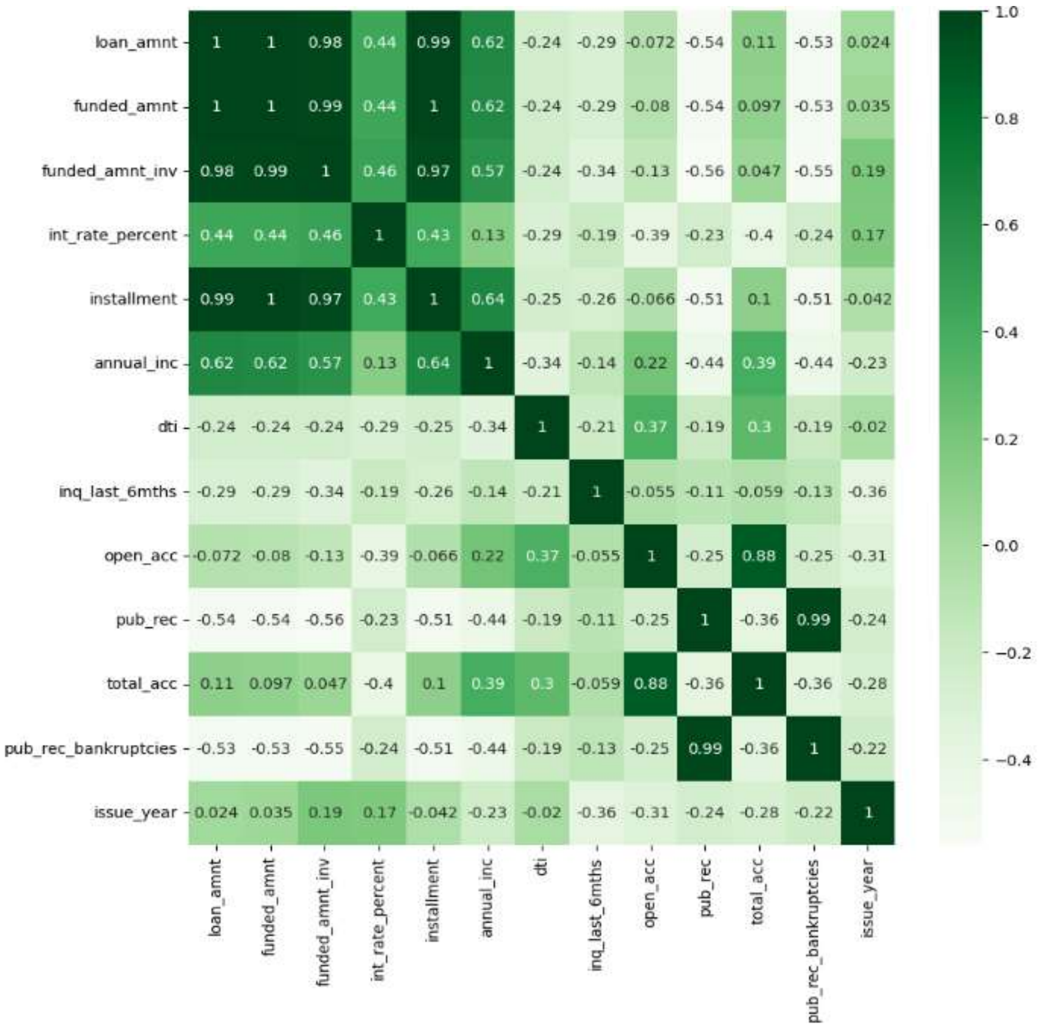


As the Debt-to-Income (DTI) value increases, the Charged Off Rate also increases. An exception is observed in the 25-30 range due to the low population data in this range

BIVARIATE ANALYSIS

Continous Variables:

- loan_amnt
- funded_amnt
- funded_amnt_inv
- int_rate_percent
- installment
- annual_inc
- dti
- inq_last_6mths
- open_acc
- pub_rec
- total_acc
- pub_rec_bankruptcies
- issue_year



	loan_amnt	funded_amnt	funded_amnt_inv	int_rate_percent	installment	annual_inc	dti	inq_last_6mths	open_acc	pub_rec	total_acc	pub_rec_bankruptcies	issue_year
loan_amnt	1	1	0.98	0.44	0.99	0.62	-0.24	-0.29	-0.072	-0.54	0.11	-0.53	0.024
funded_amnt	1	1	0.99	0.44	1	0.62	-0.24	-0.29	-0.08	-0.54	0.097	-0.53	0.035
funded_amnt_inv	0.98	0.99	1	0.46	0.97	0.57	-0.24	-0.34	-0.13	-0.56	0.047	-0.55	0.19
int_rate_percent	0.44	0.44	0.46	1	0.43	0.13	-0.29	-0.19	-0.39	-0.23	-0.4	-0.24	0.17
installment	0.99	1	0.97	0.43	1	0.64	-0.25	-0.26	-0.066	-0.51	0.1	-0.51	-0.042
annual_inc	0.62	0.62	0.57	0.13	0.64	1	-0.34	-0.14	0.22	-0.44	0.39	-0.44	-0.23
dti	-0.24	-0.24	-0.24	-0.29	-0.25	-0.34	1	-0.21	0.37	-0.19	0.3	-0.19	-0.02
inq_last_6mths	-0.29	-0.29	-0.34	-0.19	-0.26	-0.14	-0.21	1	-0.055	-0.11	-0.059	-0.13	-0.36
open_acc	-0.072	-0.08	-0.13	-0.39	-0.066	0.22	0.37	-0.055	1	-0.25	0.88	-0.25	-0.31
pub_rec	-0.54	-0.54	-0.56	-0.23	-0.51	-0.44	-0.19	-0.11	-0.25	1	-0.36	0.99	-0.24
total_acc	0.11	0.097	0.047	-0.4	0.1	0.39	0.3	-0.059	0.88	-0.36	1	-0.36	-0.28
pub_rec_bankruptcies	-0.53	-0.53	-0.55	-0.24	-0.51	-0.44	-0.19	-0.13	-0.25	0.99	-0.36	1	-0.22
issue_year	0.024	0.035	0.19	0.17	-0.042	-0.23	-0.02	-0.36	-0.31	-0.24	-0.28	-0.22	1

Correlation graph for the mentioned Continous Variables

Inference from Correlation:

- Installment, funded amount, loan amount, and funded amount by investors are highly positively correlated with each other, forming a cluster.
- Interest rate percentage is negatively correlated with total accounts.

BIVARIATE ANALYSIS

Conclusion:

Major Driving factors which can be used to predict defaulters and avoid Credit Loss

1. Purpose
2. funded_amnt
3. home_ownership
4. emp_length
5. term
6. interest_rate_percent
7. dti
8. Grade

- Loans for purpose of Small Business, renewable_energy.
- When funded amount by investor is between 29k-35k
- home_ownership results in higher chance of default, except when purpose is moving, house or renewable energy
- Charged offs are more when employement length is 10 years
- Poeple who have taken loan with longer months(60 term) with more amounts
- Loans having interest rate more than 15% having chance of defaulting.
- Higher dti(Debt-to-income) has higher chance of defaulting.
- Borrowers with least grades like E,F,G indicates high chance of defaulting