Lending Club Case Study

Satya Krishna Vasista E

Gautam Singh

ML-AI C50

UpGrad Executive PG for Machine Learning and AI

Overview

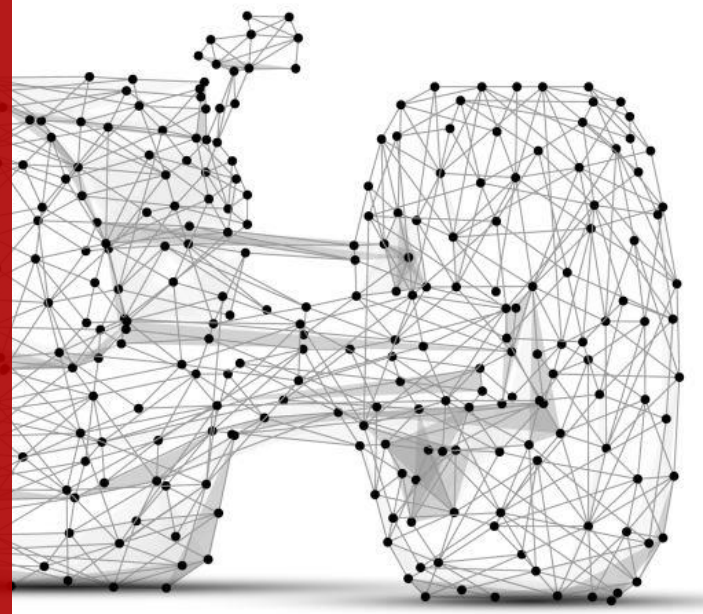
Lending Club is a lending platform that helps customers lend money at an interest rate based on banking standards.

This case study is to gain knowledge on understanding the risk analytics in banking and financial services and how data can be used to minimise the risk of losing money while lending to the customers.

Aim

Analysing the past lending data, 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 utilise this knowledge for its portfolio and risk assessment. aka, Identify the key factors that will help identify a potential defaulter.



About Us

S K Vasista Eranki

A Full-stack developer with 12+ years of experience, currently working on automating deployments and maintenance on Azure Cloud. Experience on product development for MSCM domain, and cloud infrastructure and configuration management tools.

Gautam Singh

Back-End Engineer with 5+ years of experience developing and maintaining scalable, high-performance, and secure server-side applications. Experience in various domains like Health, FMCG, Real Estate, and Social Media platforms.

Initial Analysis

- The Data contains 38717 Records and 111 rows which correspond to a Loan Data set, which has loans catered to different purposes.
- ~50% of the data has null values or columns with only 1 value. These columns may not be useful for the analysis.
- The Data Column ‘loan_status’ attributes the Status of the Loan. The value ‘Charged Off’ means its defaulted. This column is required to filter out the defaulted data and perform analysis.
- The Data has few demographic information which can be made use of.
- The data has many columns that correspond to the current on-going loan information (such as outstanding principal amount etc..)
- The Data has information about some of the codes that are determined by the Grade and Sub-Grade. These can be considered as a Lending Club’s categorization of loans. A specific analysis based on this can give insightful results.

Steps

Data Cleaning	Dealing with Nulls & NA
	Filling in Missing Information
Data Conversions & Extraction	Converting the columns that have Numeric Data
	Converting the columns that represent DateTime
	Extract Month, Year from the required Date Time columns
Outliers Data Treatment	Discard Data containing the Outliers as it may impact the current analysis
Data Analysis	Univariate Analysis on the data
	Bivariate Analysis on the data
	Binning Continuous Data to identify Patterns
	Comparing various Data

Data Cleaning – Nulls & NA

- Initial Analysis has ~52% of data with Nulls or NA values.
- Upon further Analysis, we observe Most of the Columns have only 1 value (either it be NA/NULL/a valid value). These columns are not useful for analysis. They can be discarded.

- The columns related to person identification (id, member_id, emp_title, title, URL, desc, etc) is assumed that they may not add any relevance to the analysis. Hence, they are discarded.
- Also, some columns related to the current loan information, like last due date etc. are also assumed to be non-relevant as the analysis is not person centric.
- The columns pub_rec_bankruptcies, tax_liens have mostly ~1 non-null information. As we cannot fill in with any information, it is assumed & discarded.

Data Cleaning –Missing Information

- Following columns have missing information.

- The Column **funded_amnt_inv**
  - *which represents the loan amount funded by the investor has some values=0; which doesn't make sense. Also those records cannot be sampled/assumed. Hence, its better to discard them.*

- The Column **emp_length**
  - *which represents the employment tenure has some missing information. We can take the most frequently observed value in the emp_length column (mode) and applying there. (The column emp_length is a categorical column. Hence, considering the most occurred value).*

Data Conversions & Extraction

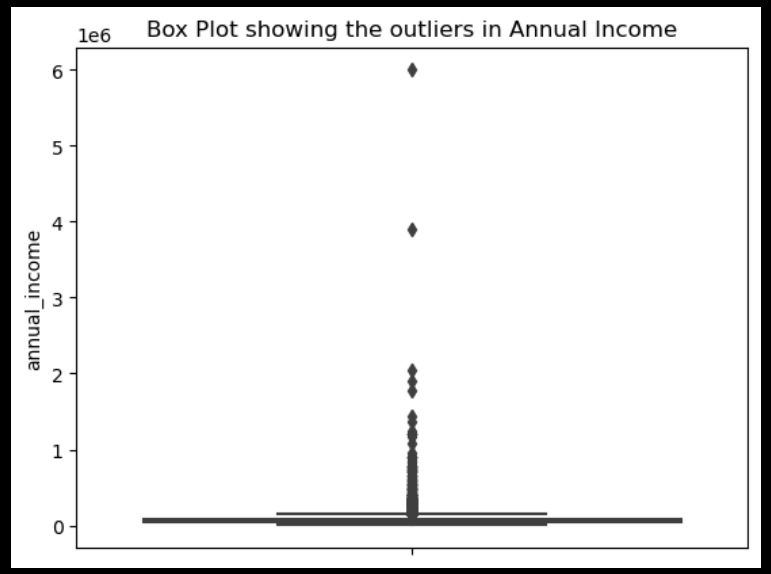
- Following columns require data conversions.
- The Column **issue_d**
  - *represents the loan issued date is identified as a string data. The values are converted to python date time. The Month and Year information is extracted.*
- The Column **int_rate**
  - *which represents the loan interest rate. It has % postfixed for every value.*
- The Column **term**
  - *which represents the loan term in months. It has months postfixed for every value which is removed.*
- The Column **emp_length**
  - *which represents the employment tenure in years. It has year/years postfixed for every value which is removed.*

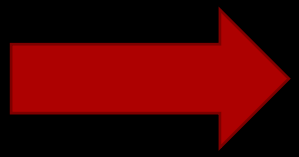
Outliers Data Treatment

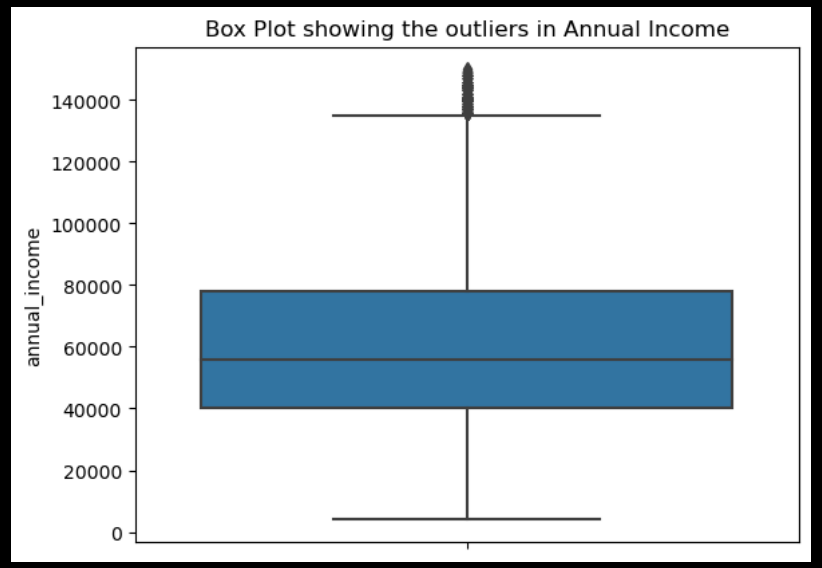
The column ***annual_income*** has very high range of values.

These values can cause shift the analysis data.

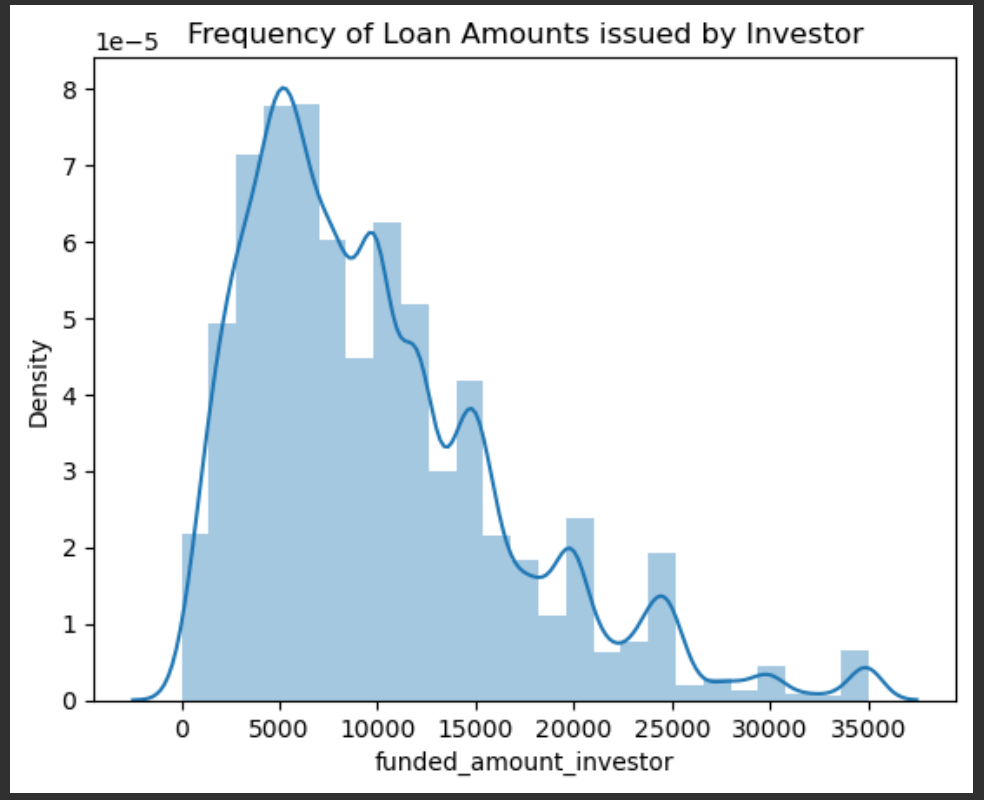
After ignoring the ***annual_income*** values >150000, we see the data is good for analysis.

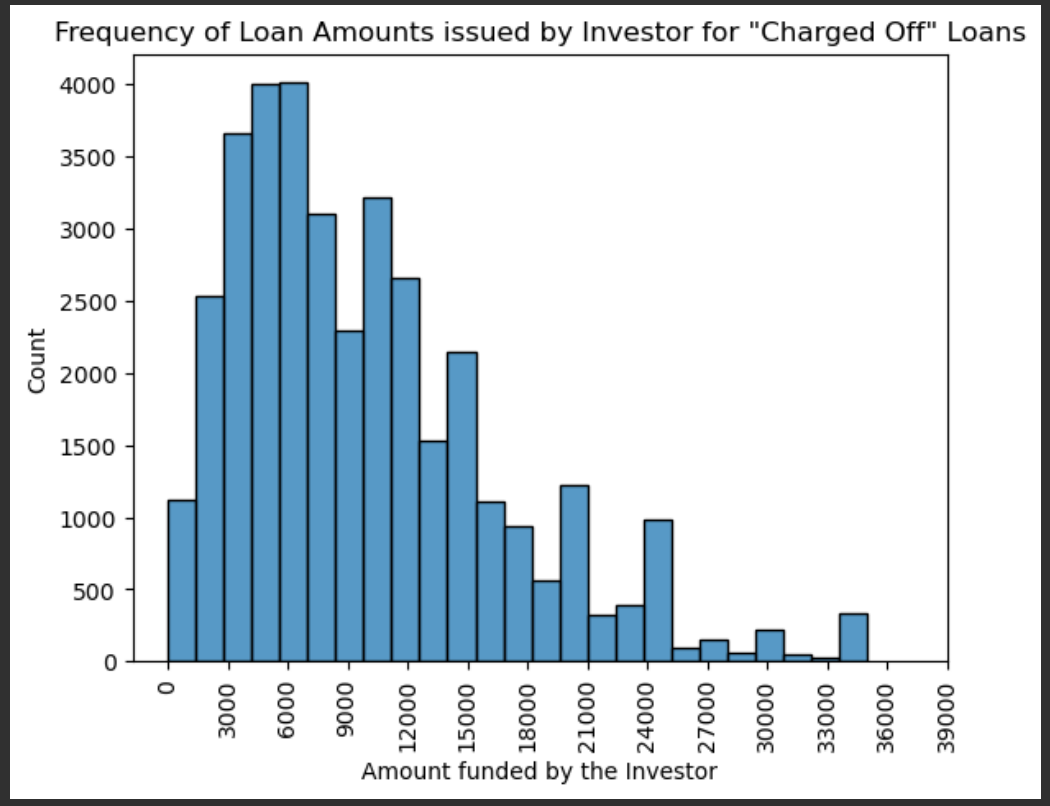






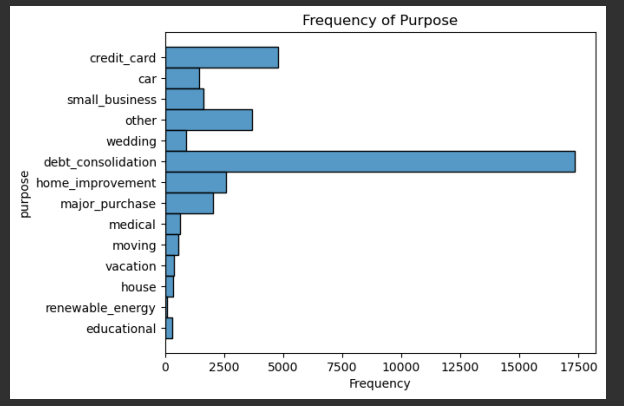
Data Analysis – Loan Amounts funded by Investor

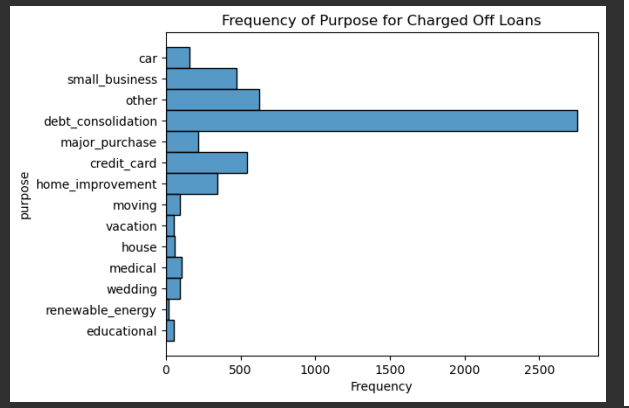


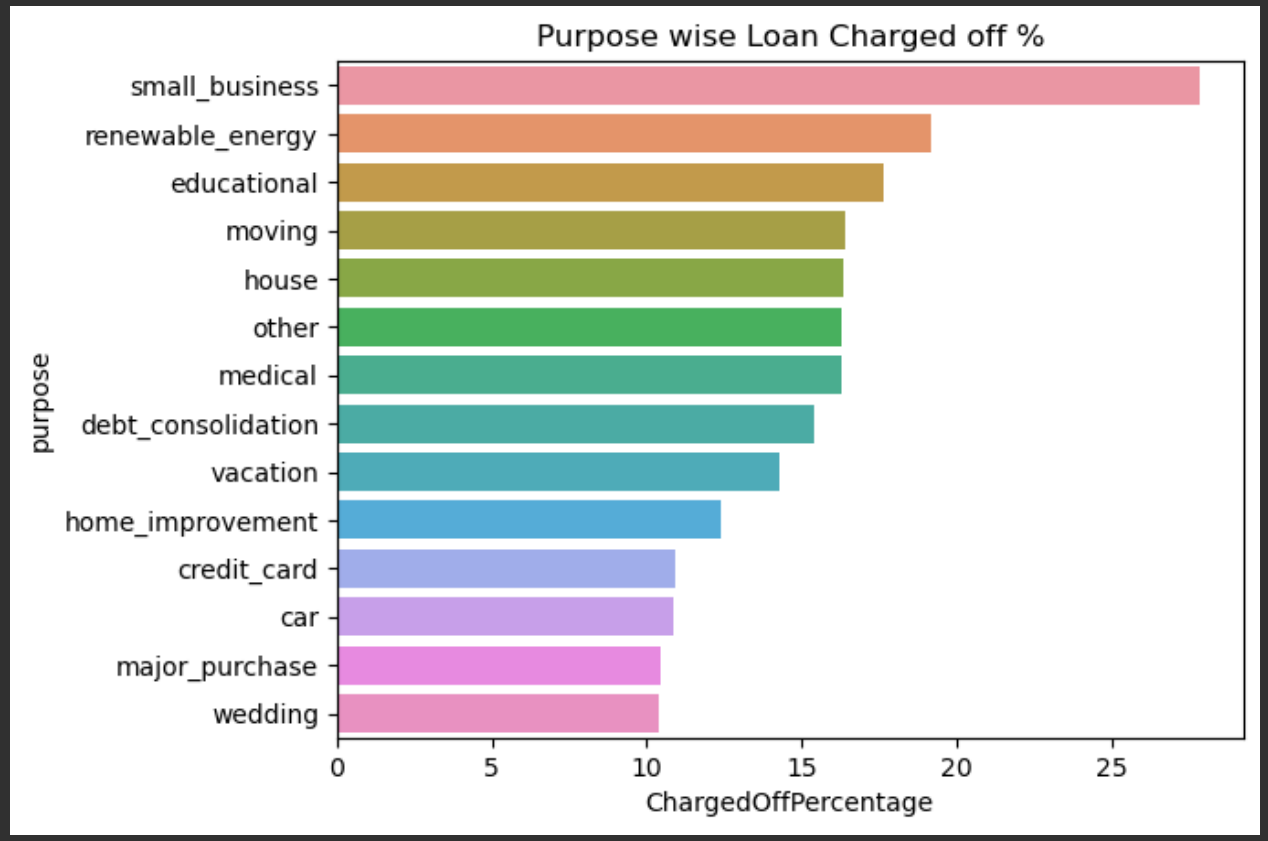


Most of the loans funded are with an investor fund between <10000. And loans funded between 3000-6000 have higher chances of Charged Off

Data Analysis – Purpose



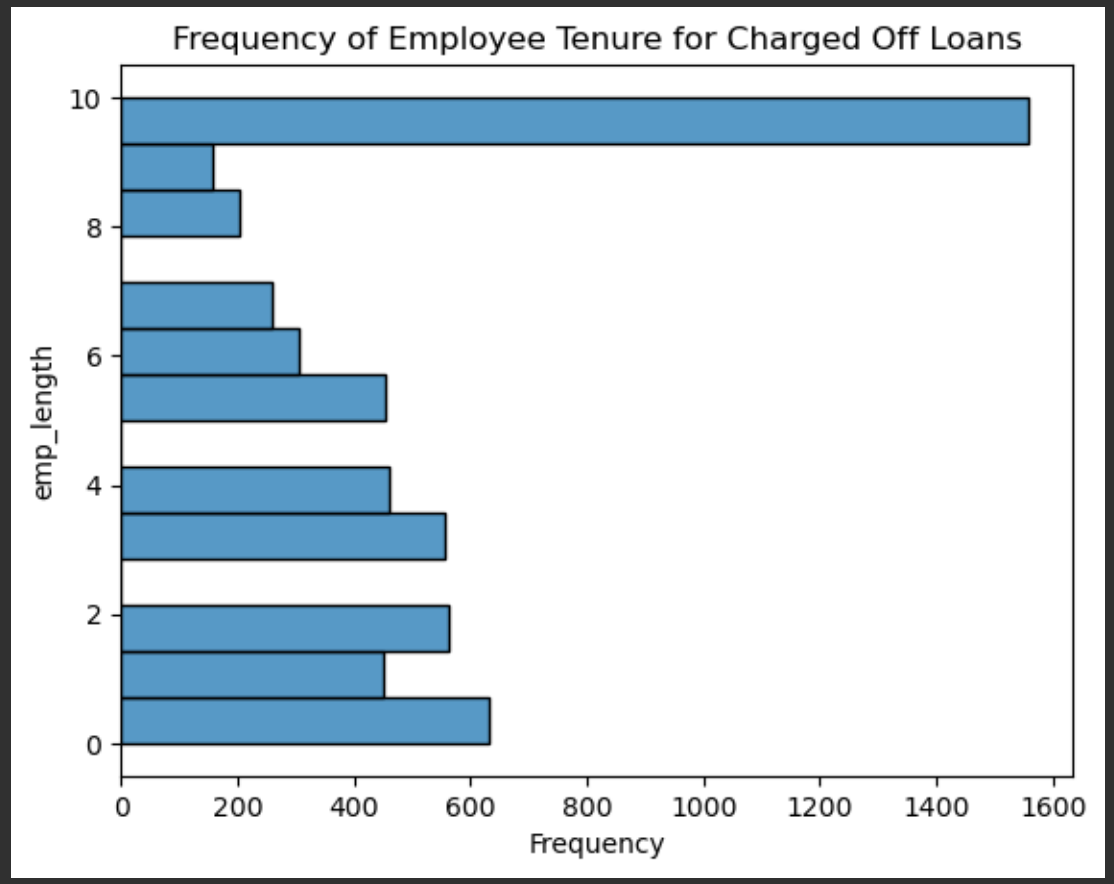




Most Loans are issued against the Purpose of **Debt Consolidation.**

However, more % of loans are Charged Off against the Purpose '**Small Business**'.

Data Analysis – Employee Tenure

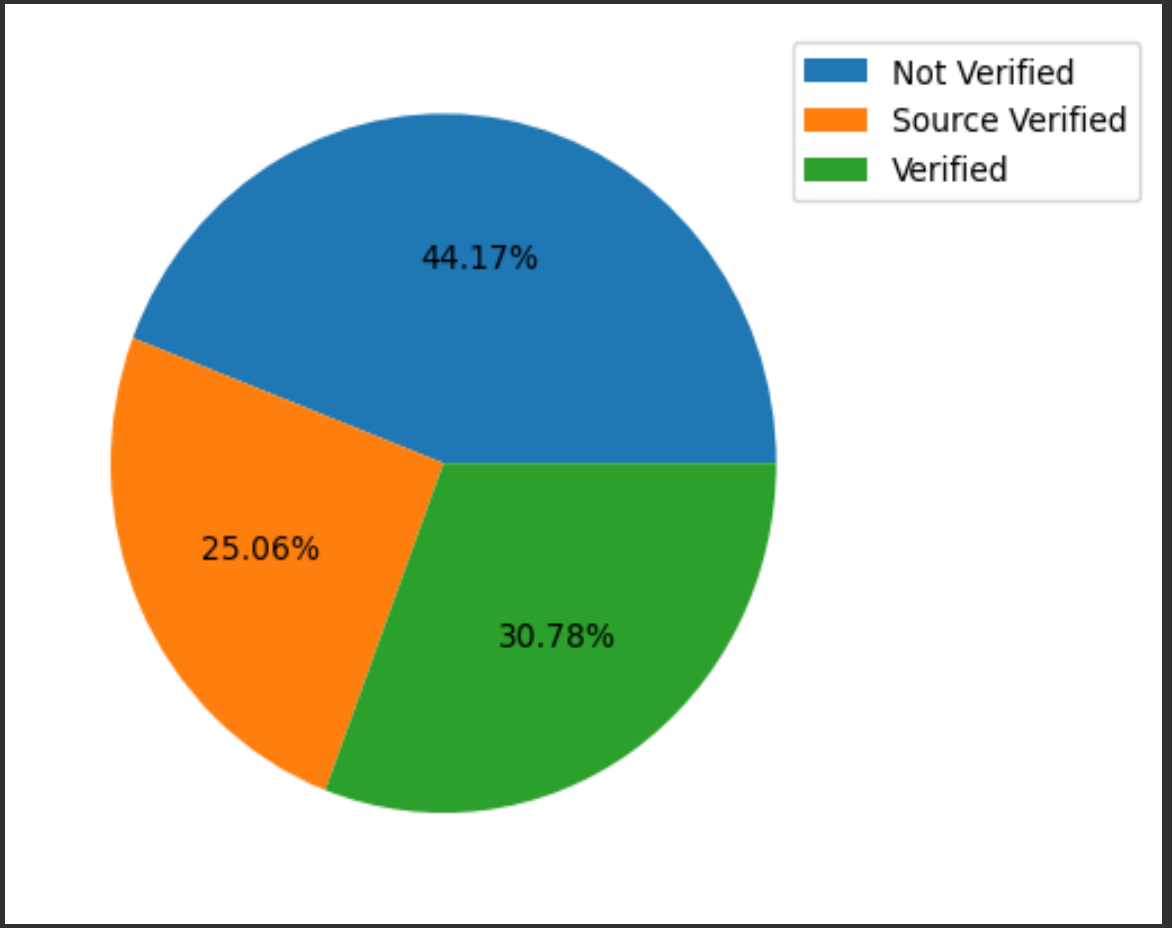


Loans taken by people with employment tenure **10 or more years** are more likely to be Charged Off.

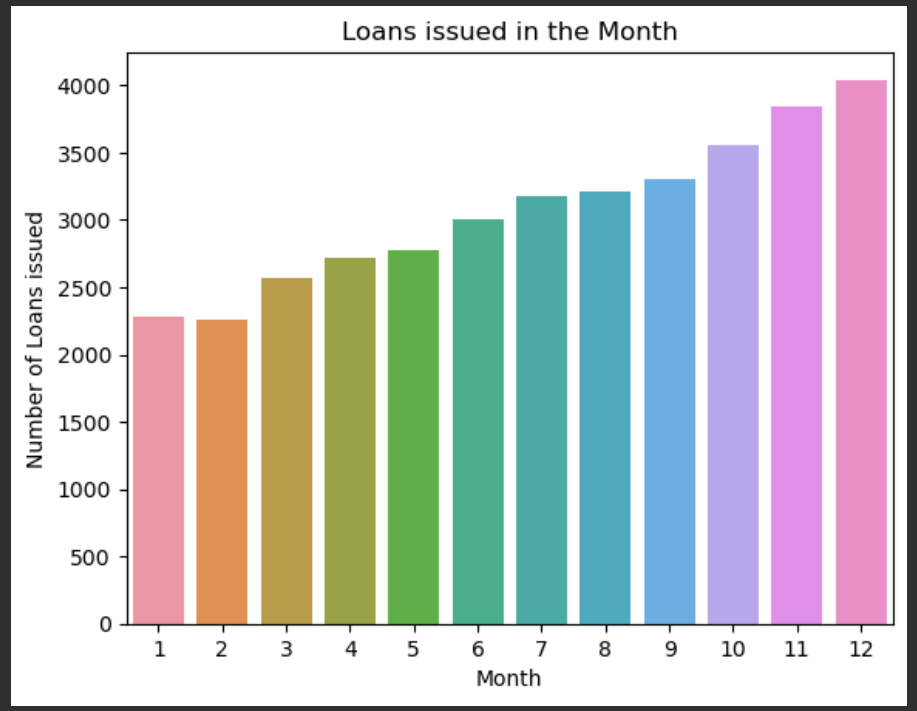
Data Analysis – Verification Status

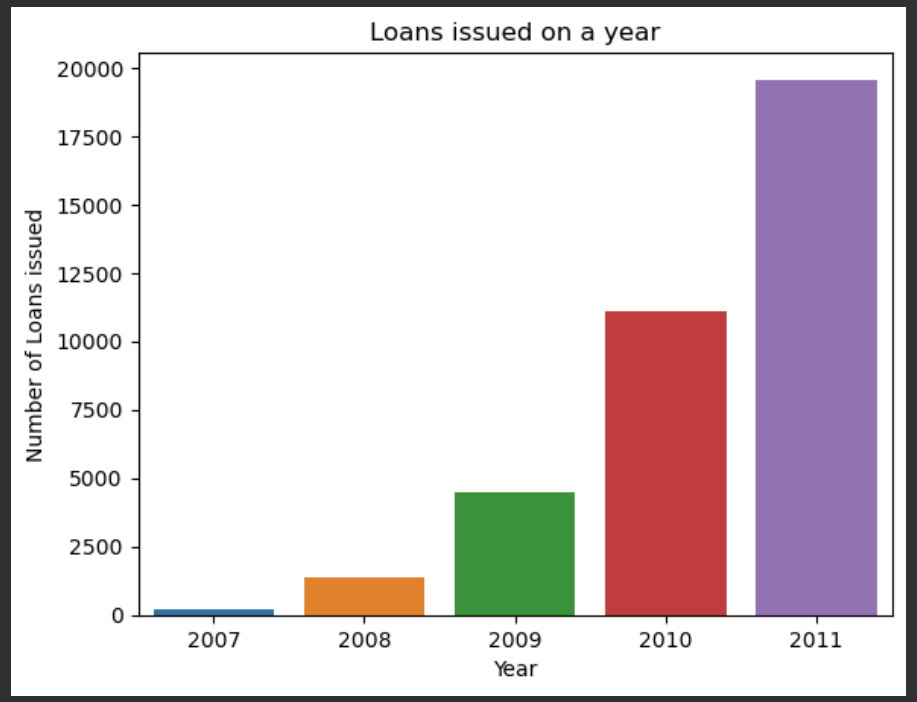
Almost 44% of the loans are Not Verified.

This could be a reason for the defaulters.



Data Analysis – Issue Month & Year



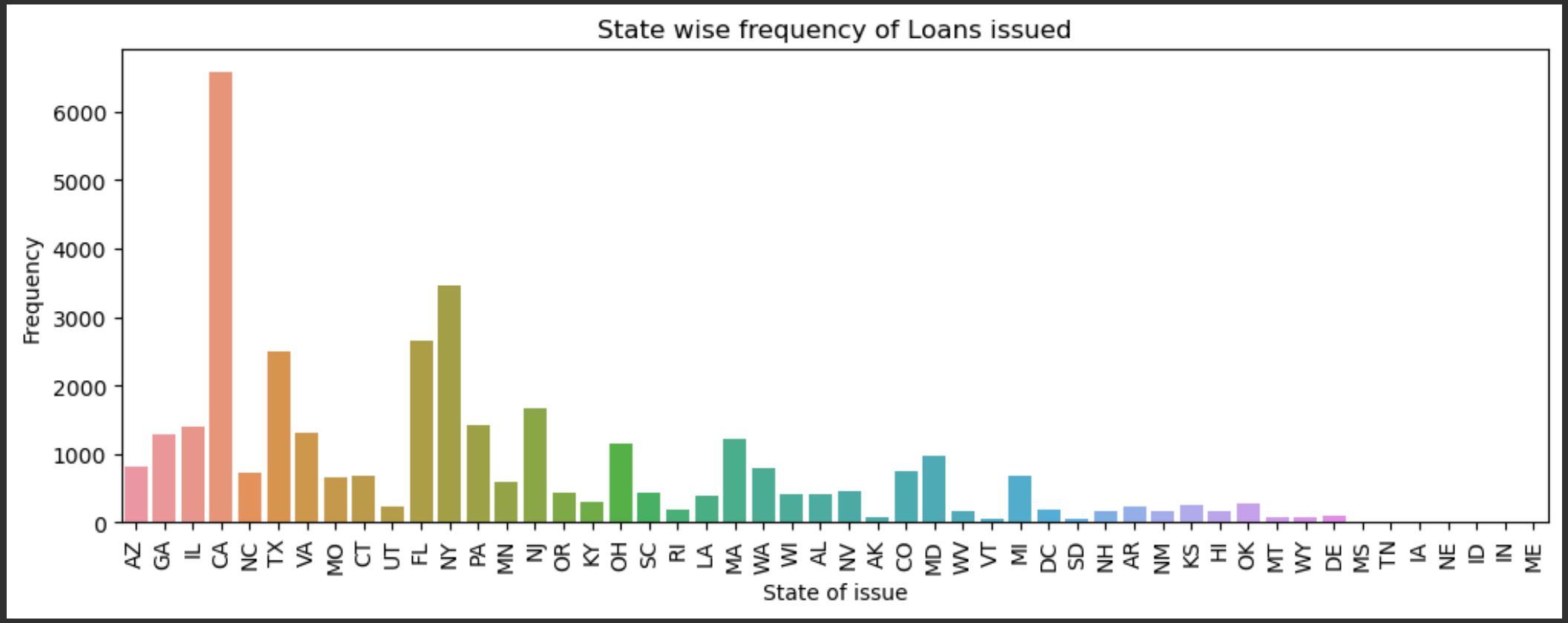


We observe there is a consistent increase of loans issued YOY (Year-on-year).

We observe most of the loans are taken during the year 2011.

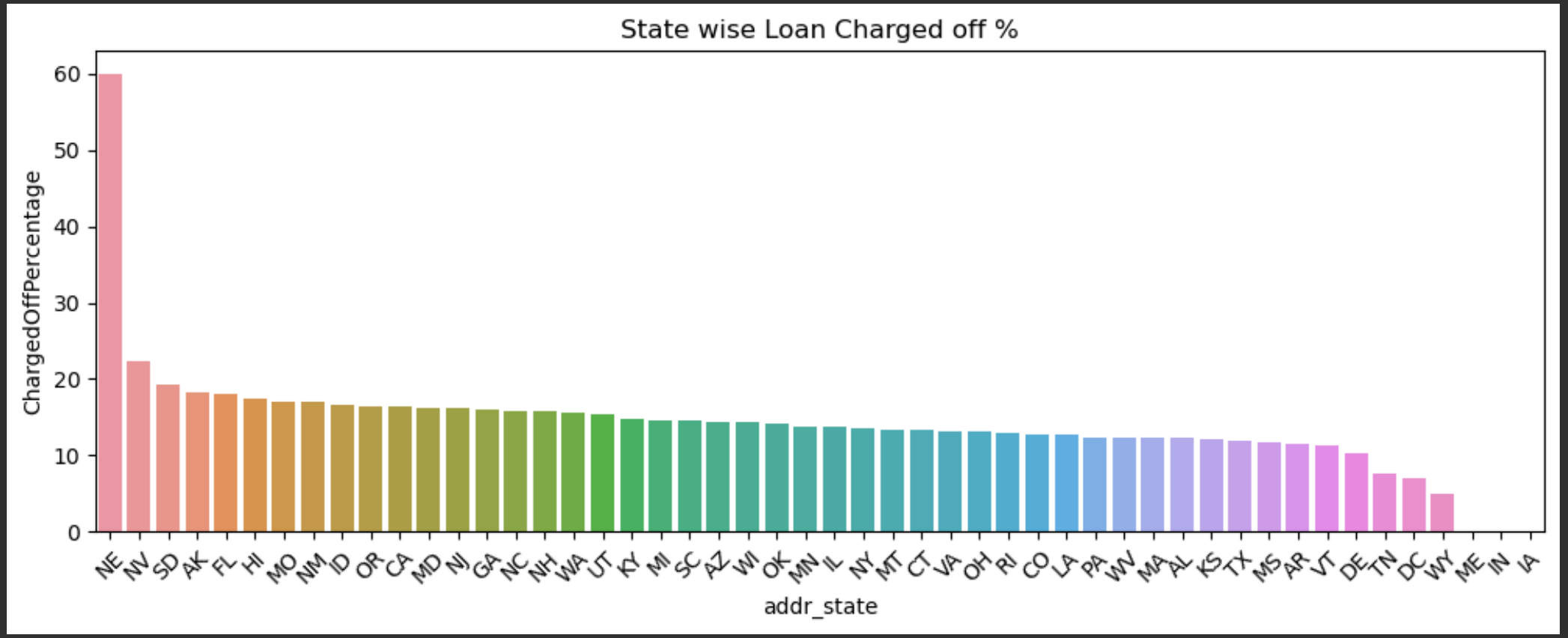
Also, most of the loans are issued during the December.

Data Analysis – State



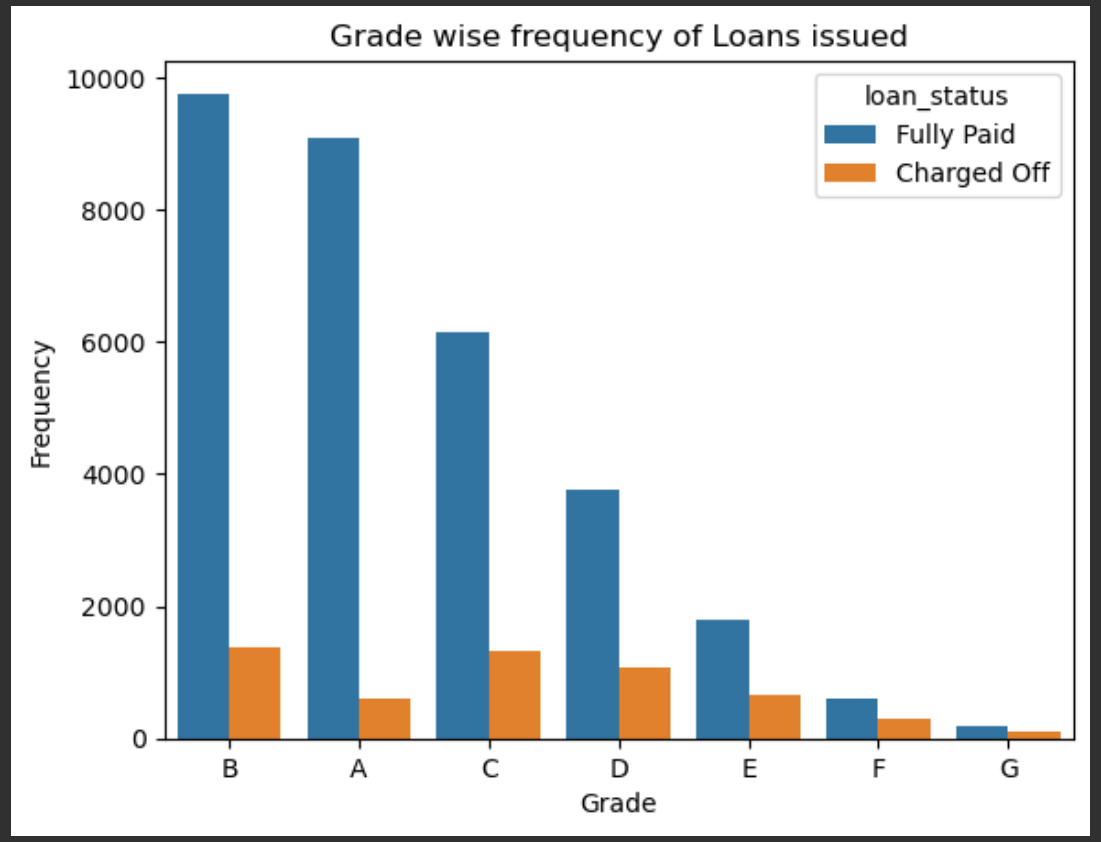
The state CA has abnormally high amount of loans issued.

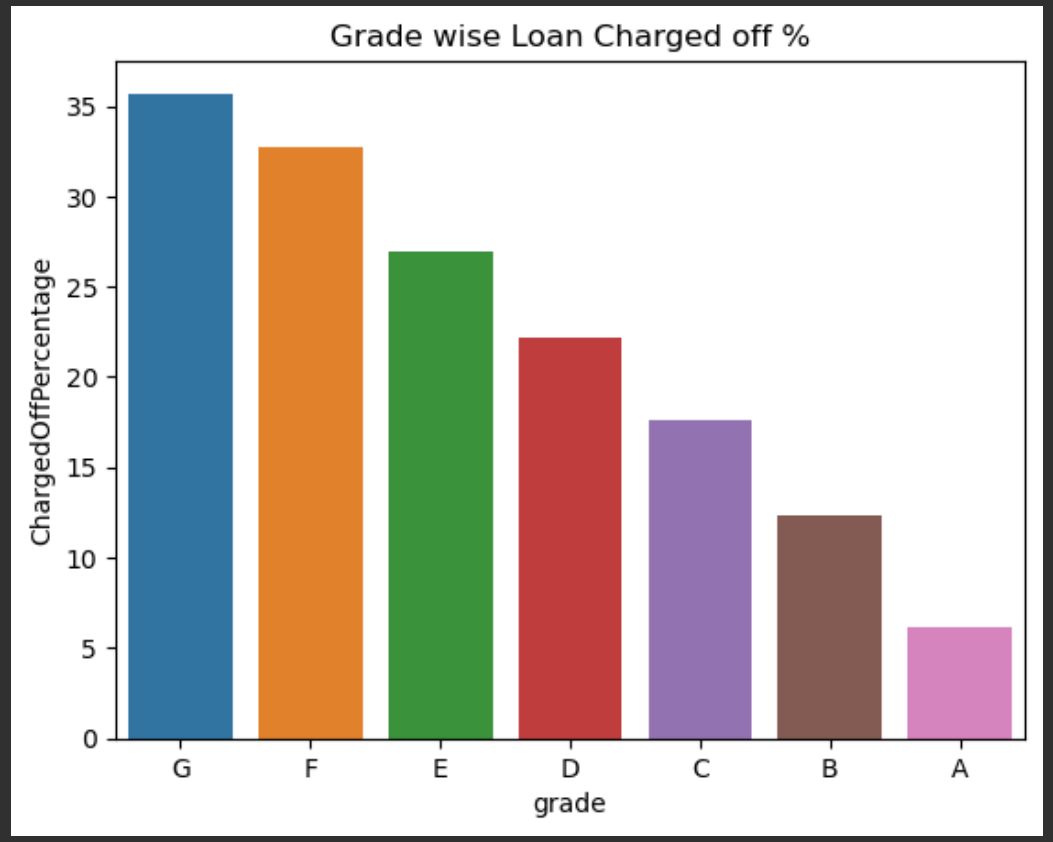
Data Analysis – State (contd.)



We Observe that the State: NE has more % of loans charged off

Data Analysis – Grade

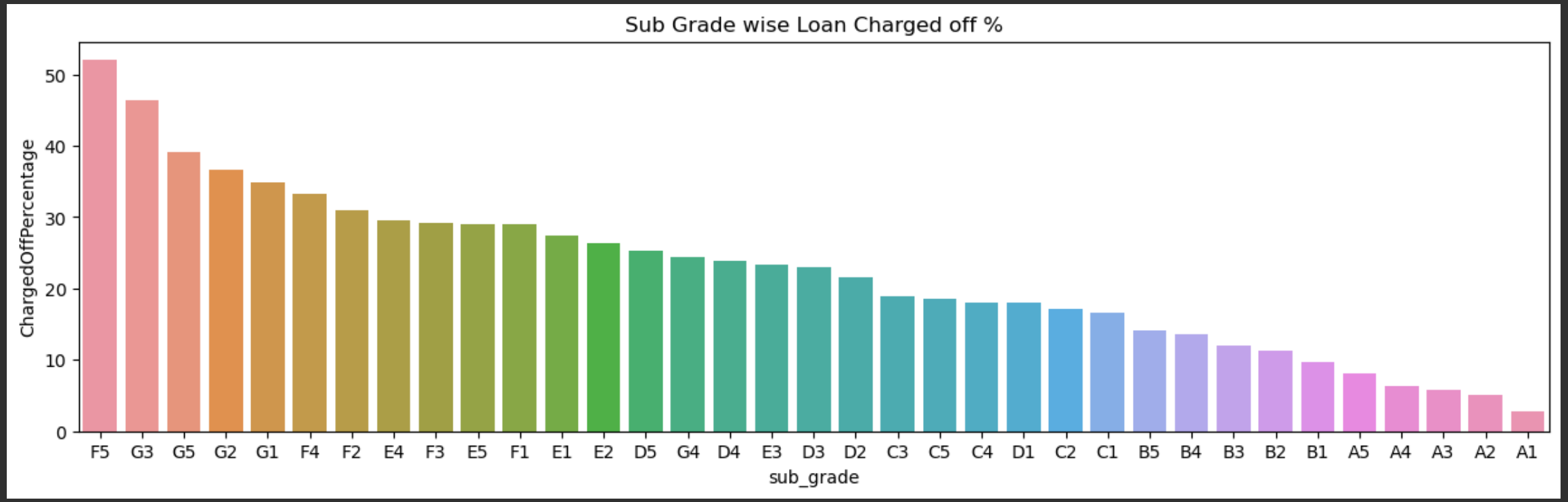




Most of the loans issued are under grade B;

While more % of loans are charged off under the grade G.

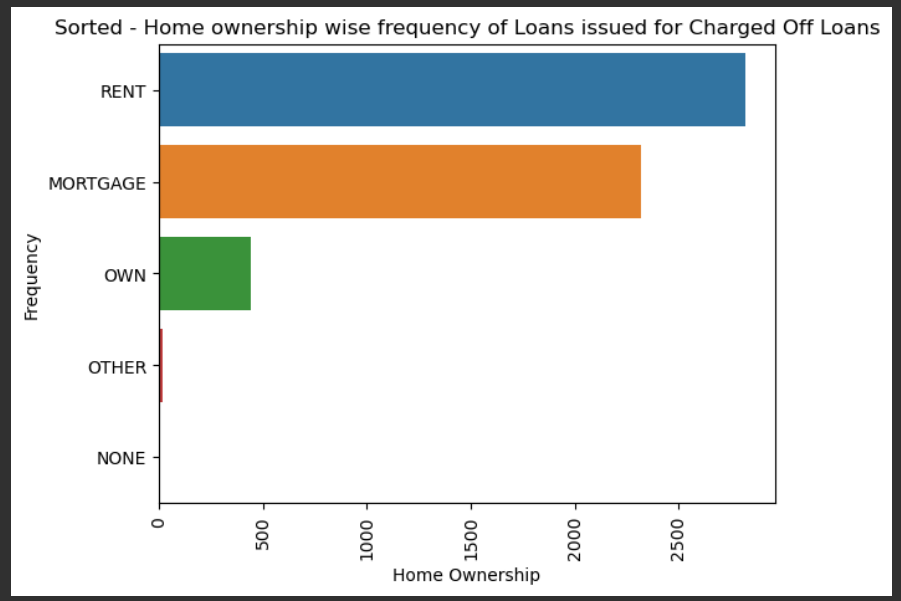
Data Analysis – Sub-Grade

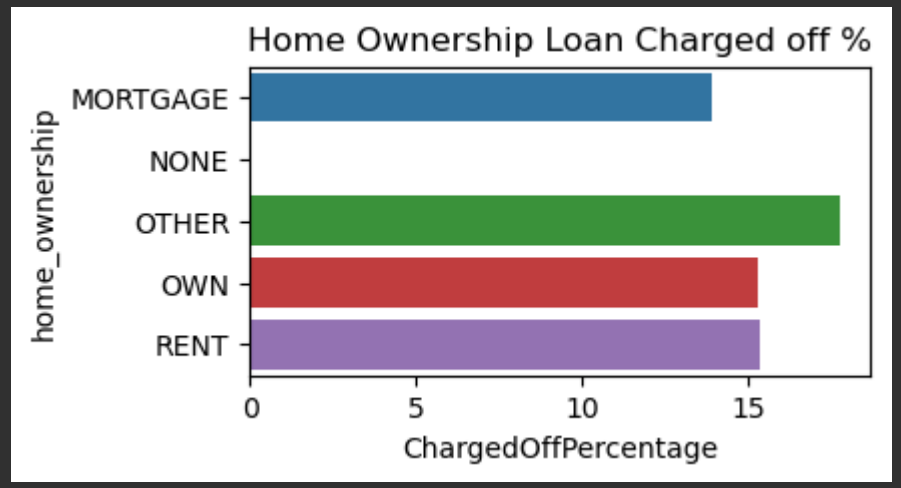


The sub-Grades F5 has more % of loans Charged Off.

The Sub-Grades G3, G5, G2, G1 seems to be at the top of the loans being Charged Off.

Data Analysis – Home Ownership

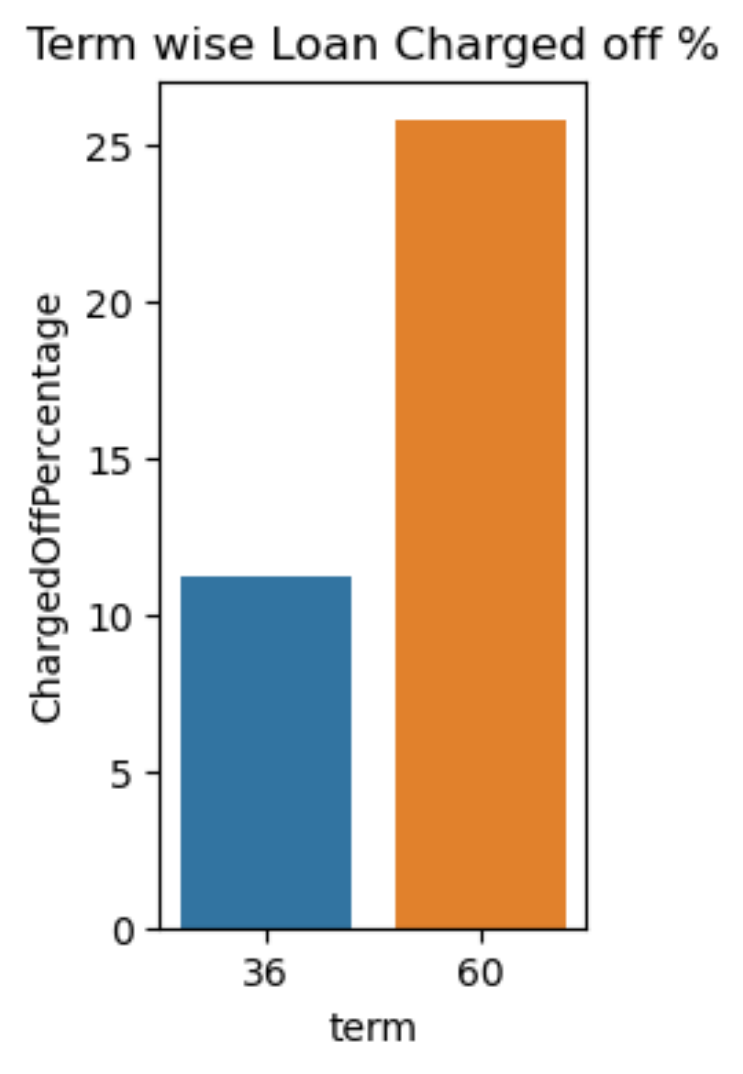




More loans are given to the people who have **home Ownership – Rent** and are also charged off more.

On Further analysis, the loans with **home_ownership** mentioned as OTHER have a higher rate of Charged Off.

Data Analysis – Term



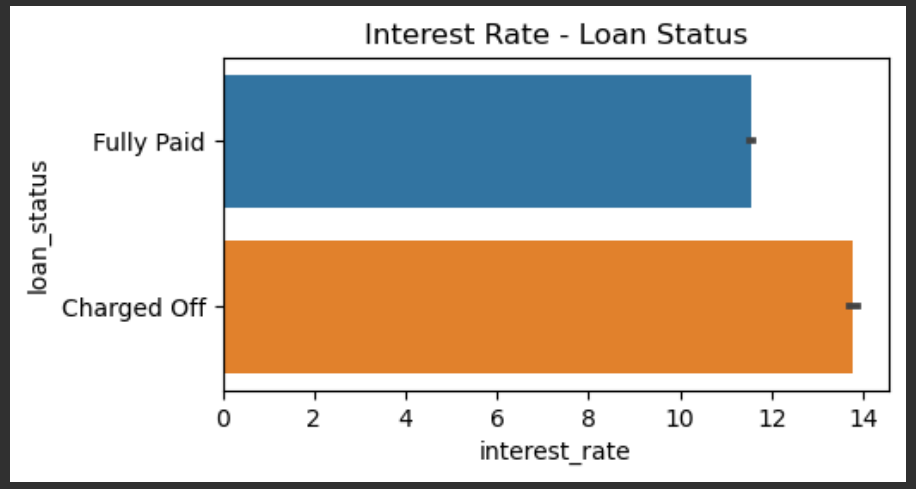
60 months tenured loans show a higher Charged Off %.

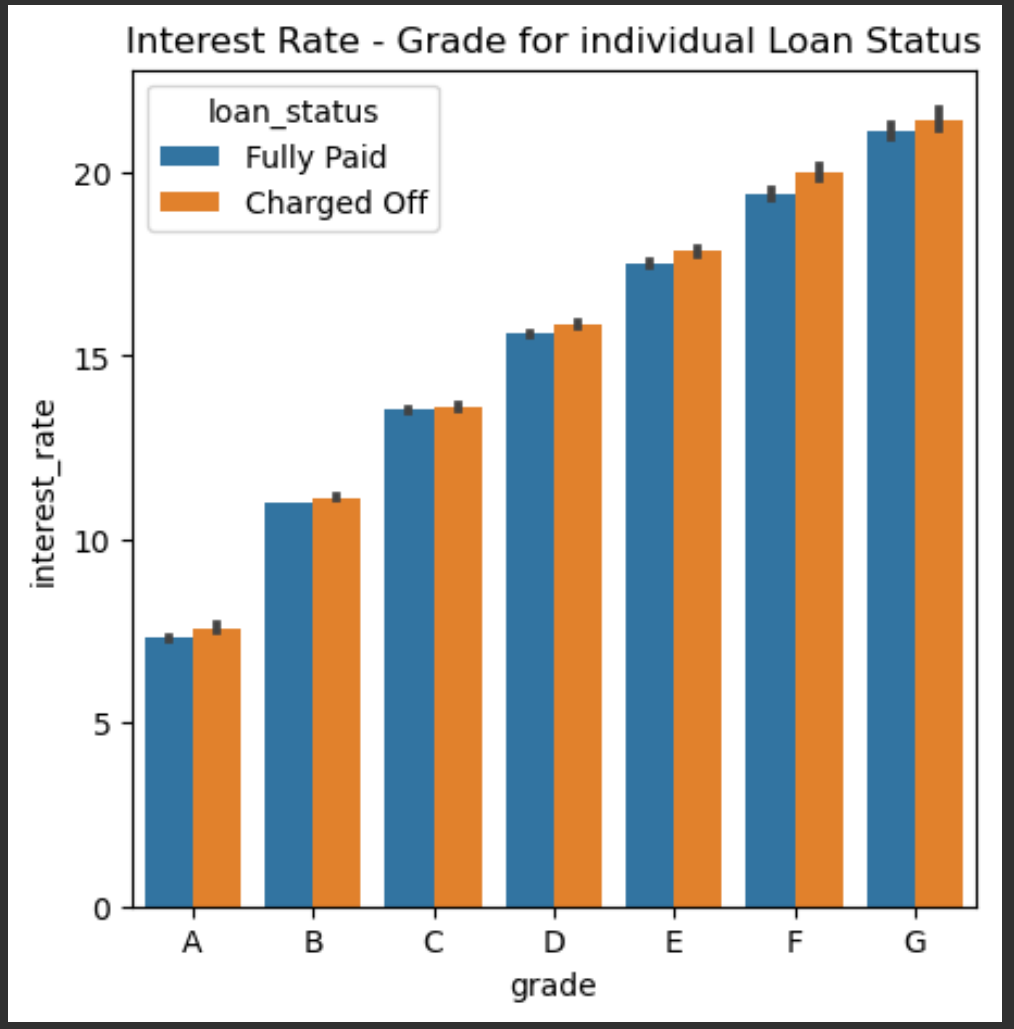
Avoid higher tenured loans.

Data Analysis – interest rate

On Analysis, the Loans are getting Charged off for Higher Interest Rates.

Higher grades also mean higher interest rates.





Summary of Analysis

Investor funded Amount range between **3000-6000** are having a higher chances of getting Charged Off.

Purpose attribute may be a factor for the loans getting Charged Off.

People with **employment 10 or more years** are more likely to be Charged Off.

Almost **44%** of the loans are **not verified**. This could be a reason for the defaulters.

It is observed of a pattern on the **Month** of the loan issued.

Loans issued to the **State NE** has more % of loans charged off

Also, the loans issued under **sub-Grades F5, G3, G5, G2, G1** are prone to be Charged Off.

Loans issued to home ownership is **OTHER** tend to have their loans Charged Off more.

Also, home ownership as **RENT** tend to take more loans and also Charged Off.

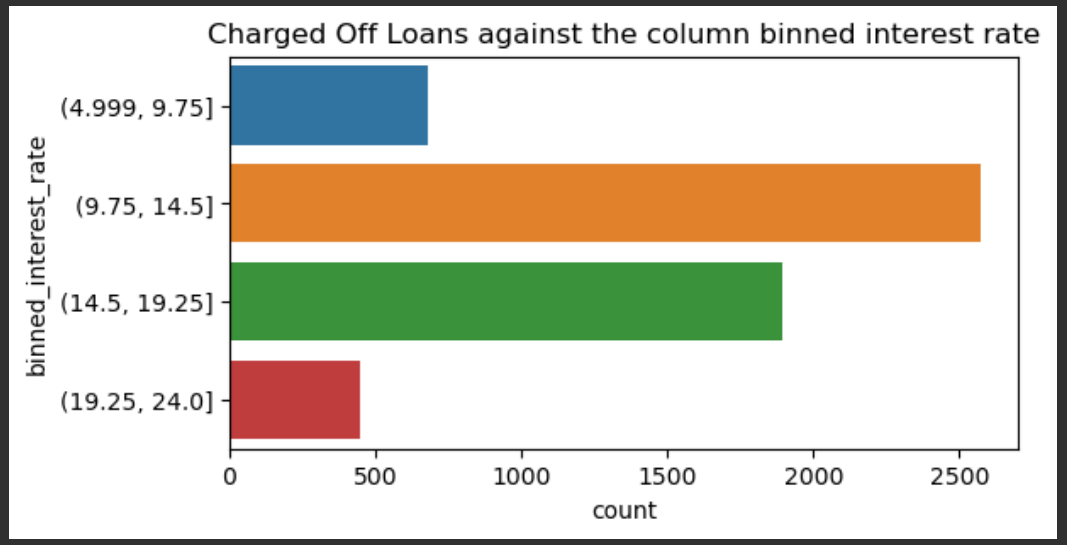
Loans with a Tenure of **60 Months** have Charged off more.

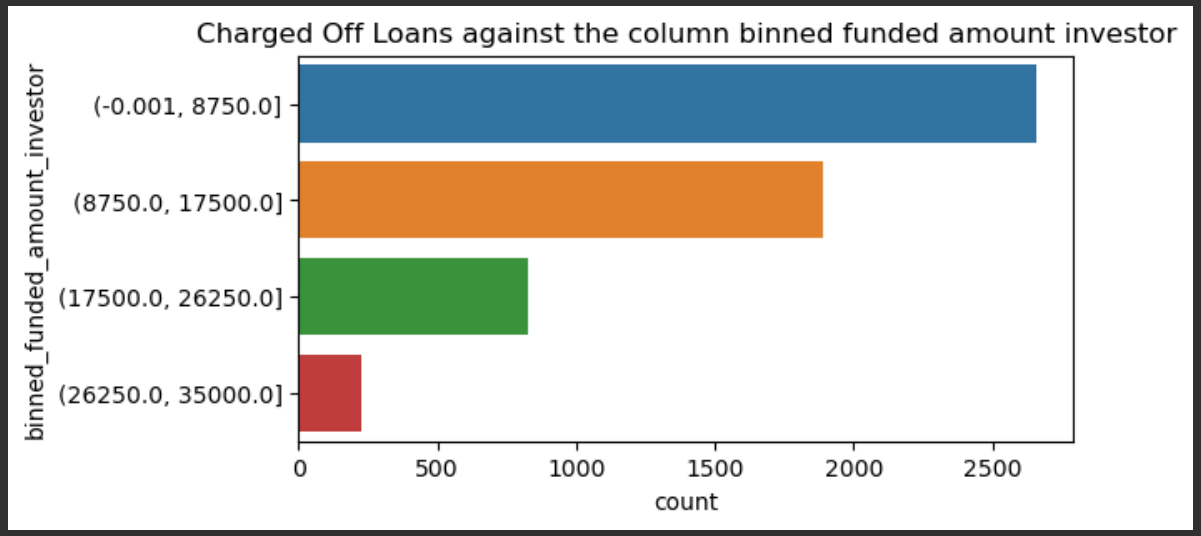
Interest rates are increasing for each Grade. Higher interest rates show strong correlation of getting Charged Off

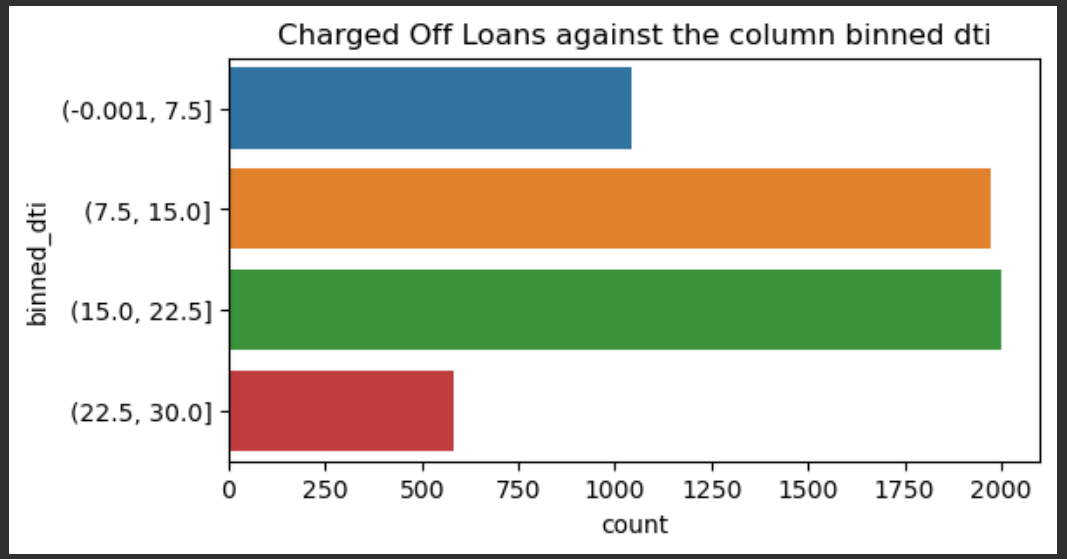
Binning of Data

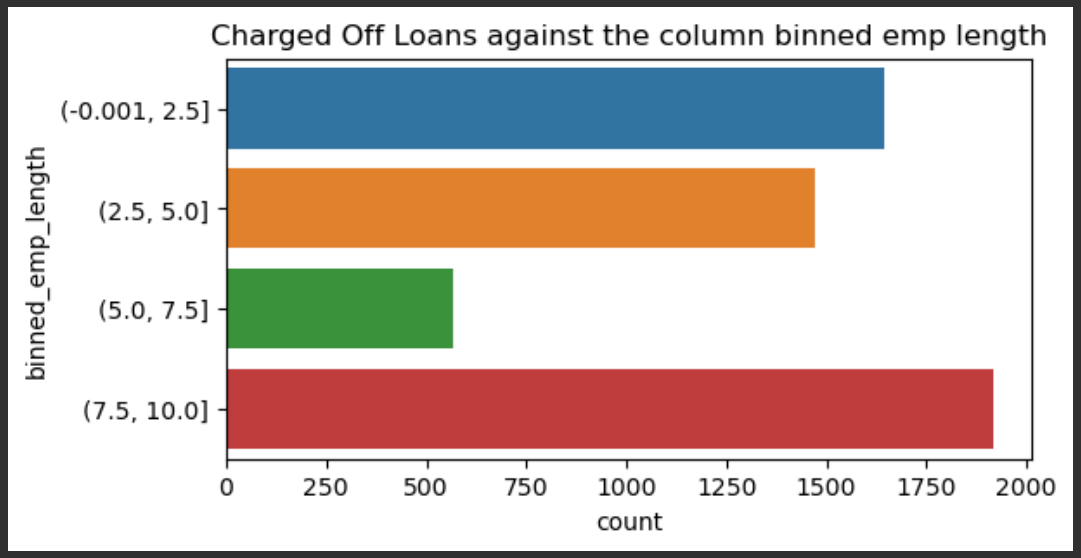
- Binning is a concept of dividing the continuous variables into bins so as each bin may act as a category.
- Following attributes are categorized into 5 bins:
  - annual_income
  - funded_amount_investor
  - installment
  - total_received_late_fee
  - dti
  - interest_rate
  - emp_length
  - inq_last_6mths
  - public_derogatory_records

Binning of Data – Binning charged off Data

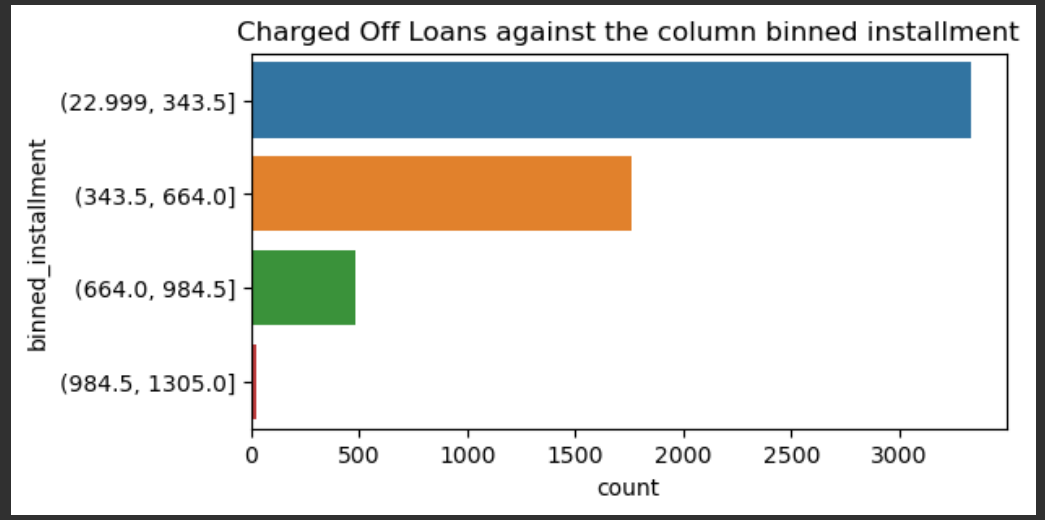


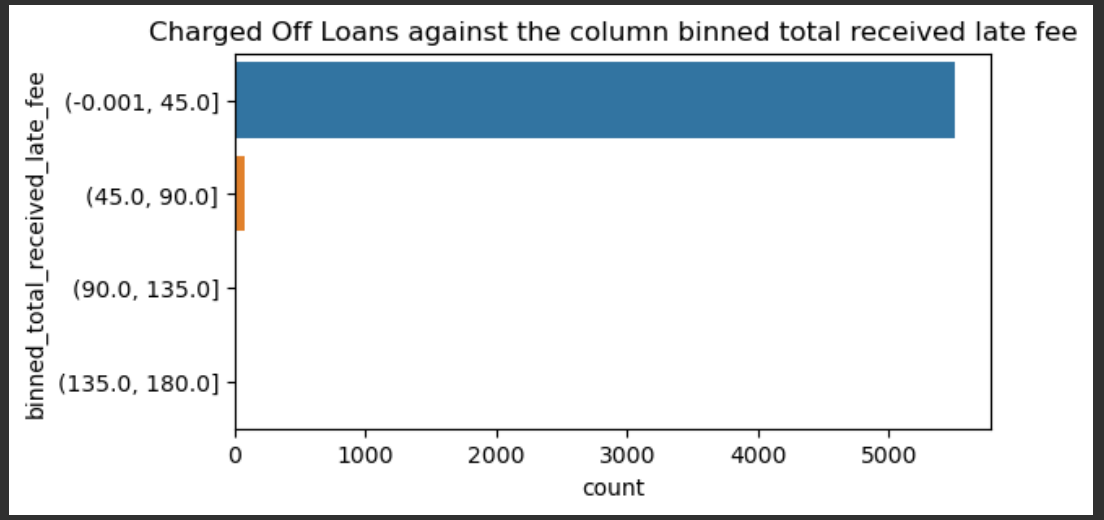


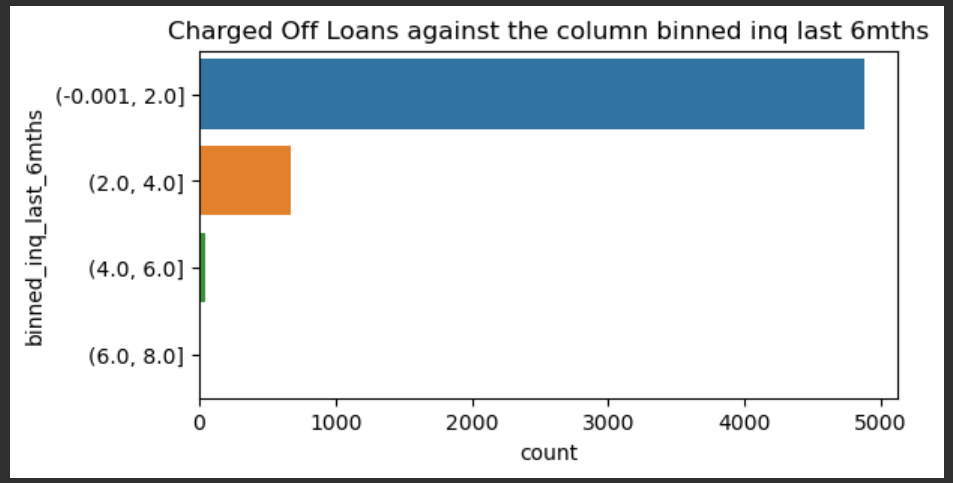


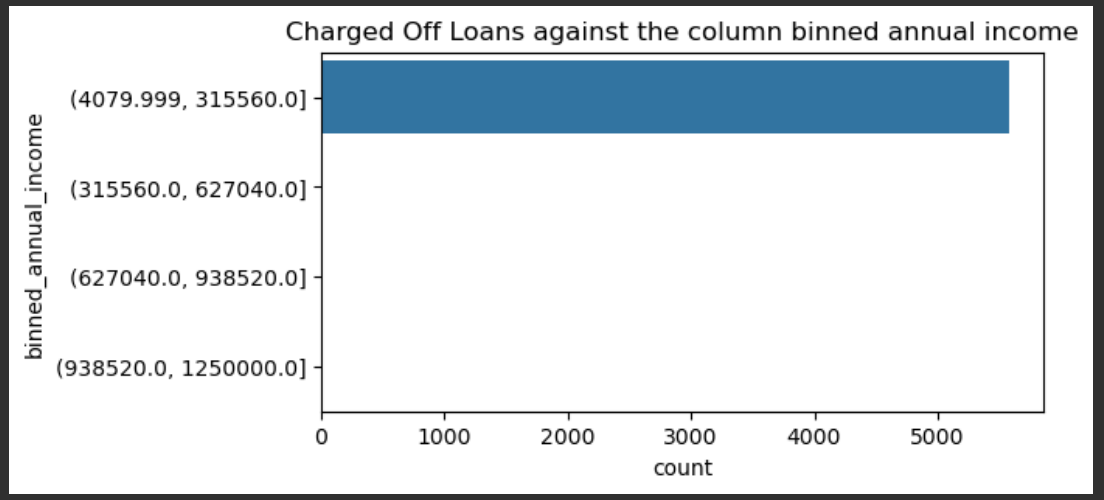


Binning of Data – Binning charged off Data





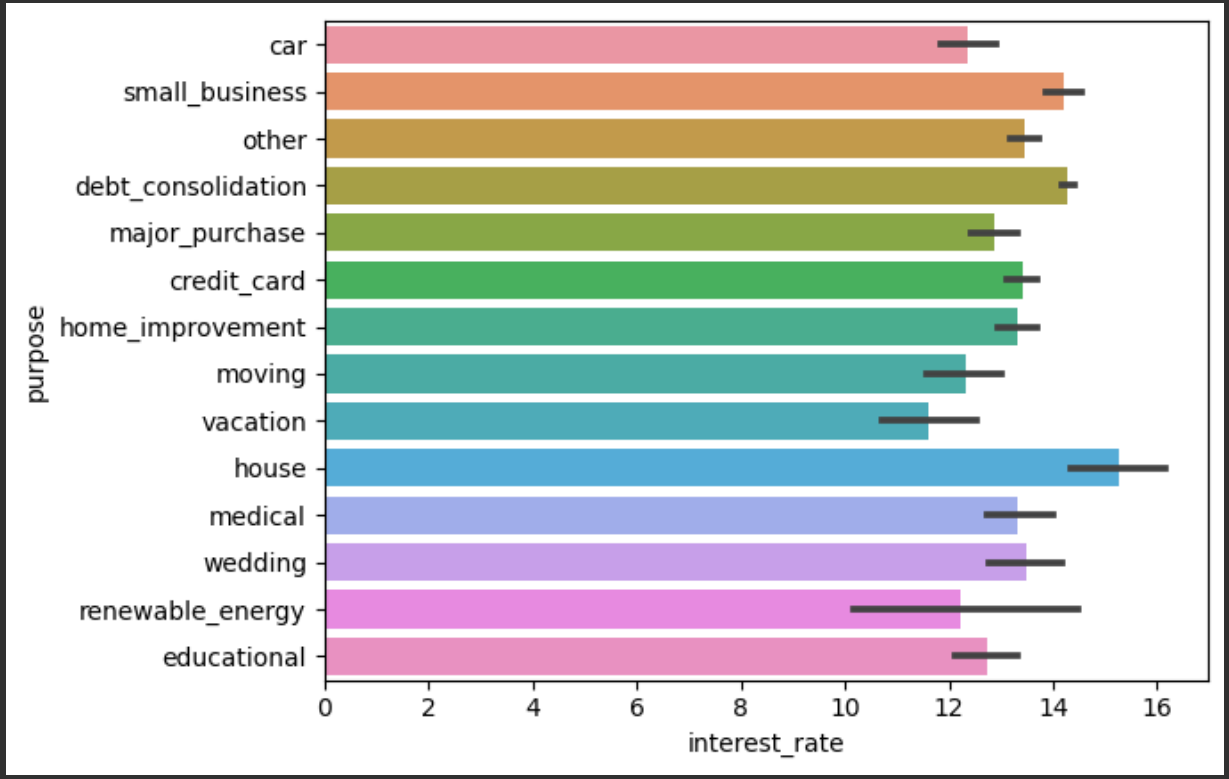




Summary of Binning data

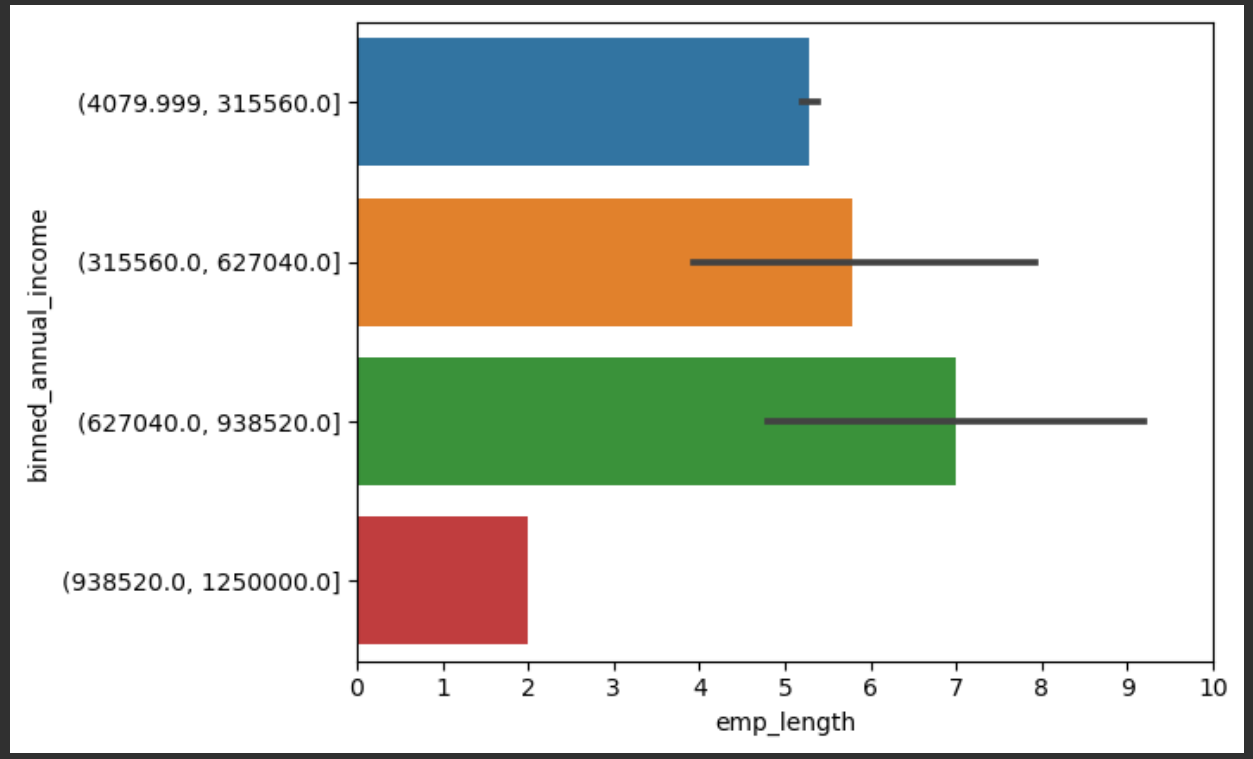
- Loans with **Investor funded amount < ~7850** are observed to be Charged Off
- Loans with **Monthly Instalments in the range of ~23 - 244** are observed to be Charged Off
- Loans with **dti in the range of ~7-22.5** are observed to be Charged Off
- Loans with **Interest Rate in the range of ~ 9.75-14.5** are observed to be Charged Off
- Loans with **1-2 inquiries in last 6 months** are observed to be Charged Off
- Loans with **at least 1 public derogatory record** are observed to be Charged Off

Multivariate Analysis (Interest Rate – Purpose)



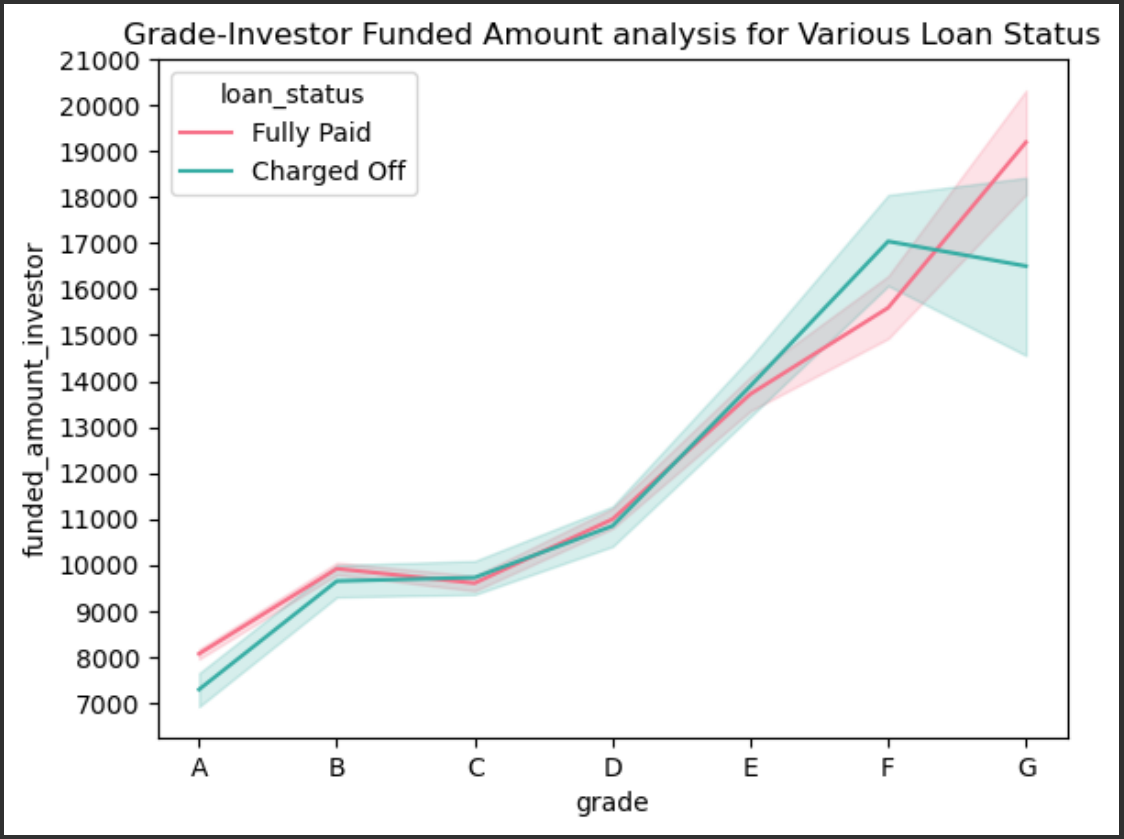
- Loans with purpose of **House** and with **higher interest rate** are having changes of getting Charged Off

Multivariate Analysis (Annual Income - emp_length)



- Loans are Charged Off when the employment is **7 years** and the annual income is in the range of **[627K-938K]**

Multivariate Analysis (funded_amount_investor- Grade)



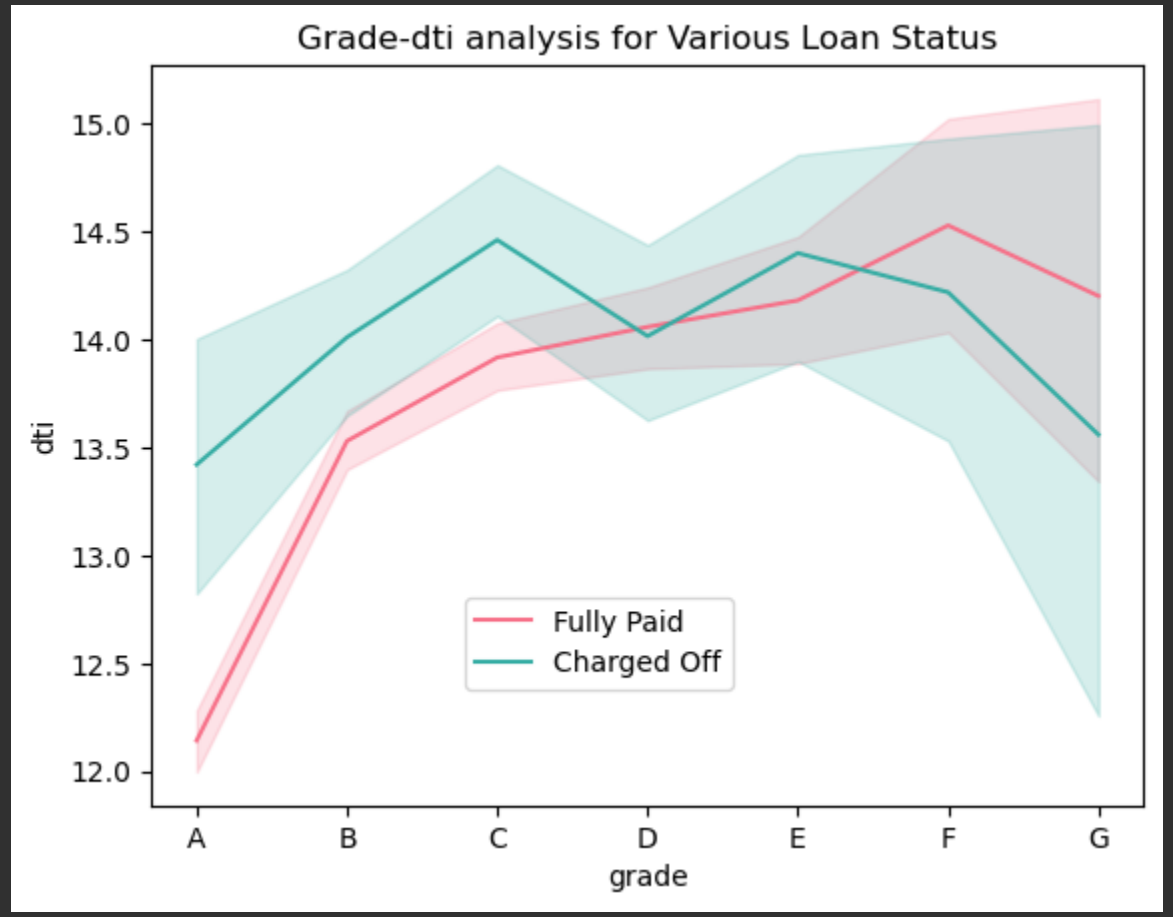
- Loans Charged Off when:
  - Grade is F and Investor funded amount > 16K.
  - Grade is G and Investor funded amount < 18K

Multivariate Analysis (Grade - dti)

- DTI → Debt to Income ratio

- Loans Charged Off when:

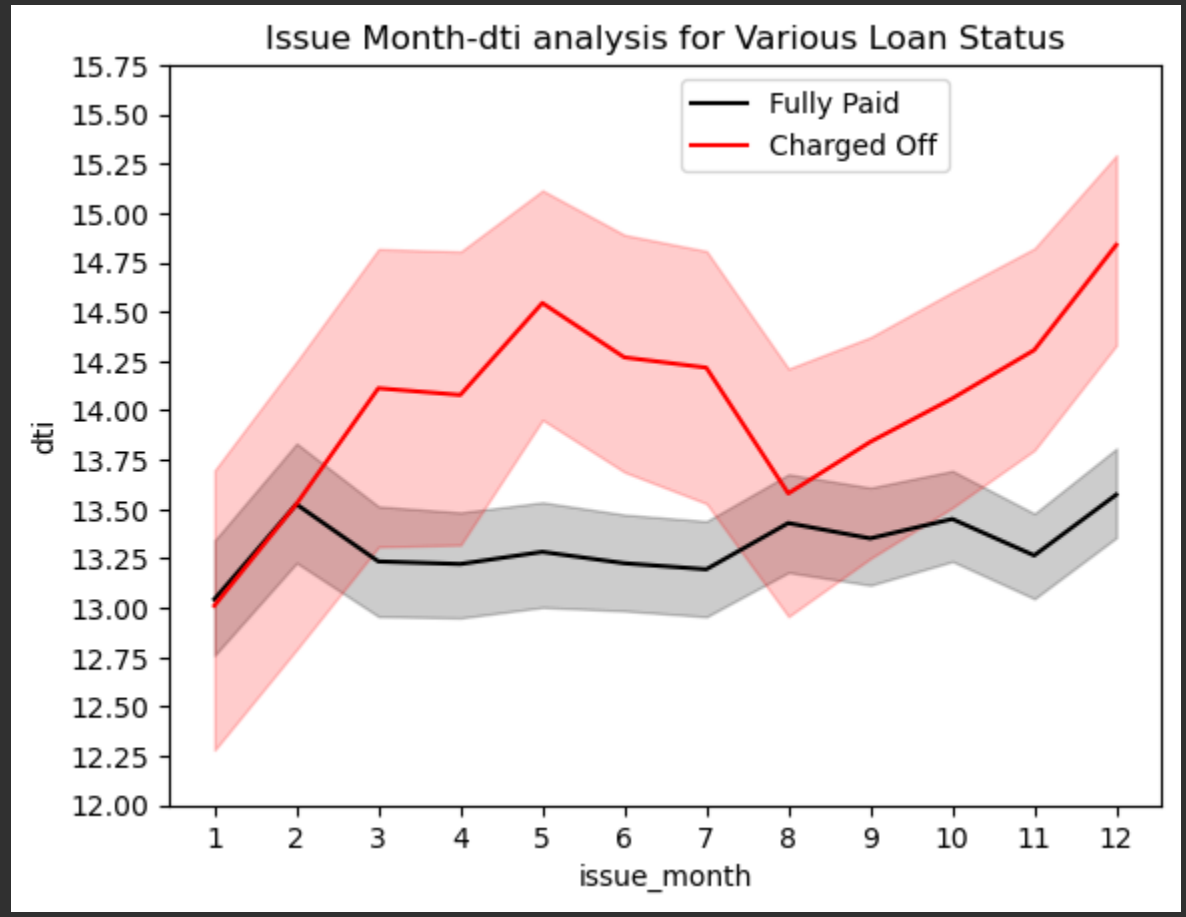
  - Grade is A with dti > 12.5
  - Grade is B with dti > ~13.75
  - Grade is C with dti > 14.
  - Grade is G and dti < 13.5.
  - Grade G covers a vast range of DTI where the Loans are Charged Off.



Multivariate Analysis (Issue Month - dti)

- DTI → Debt to Income ratio

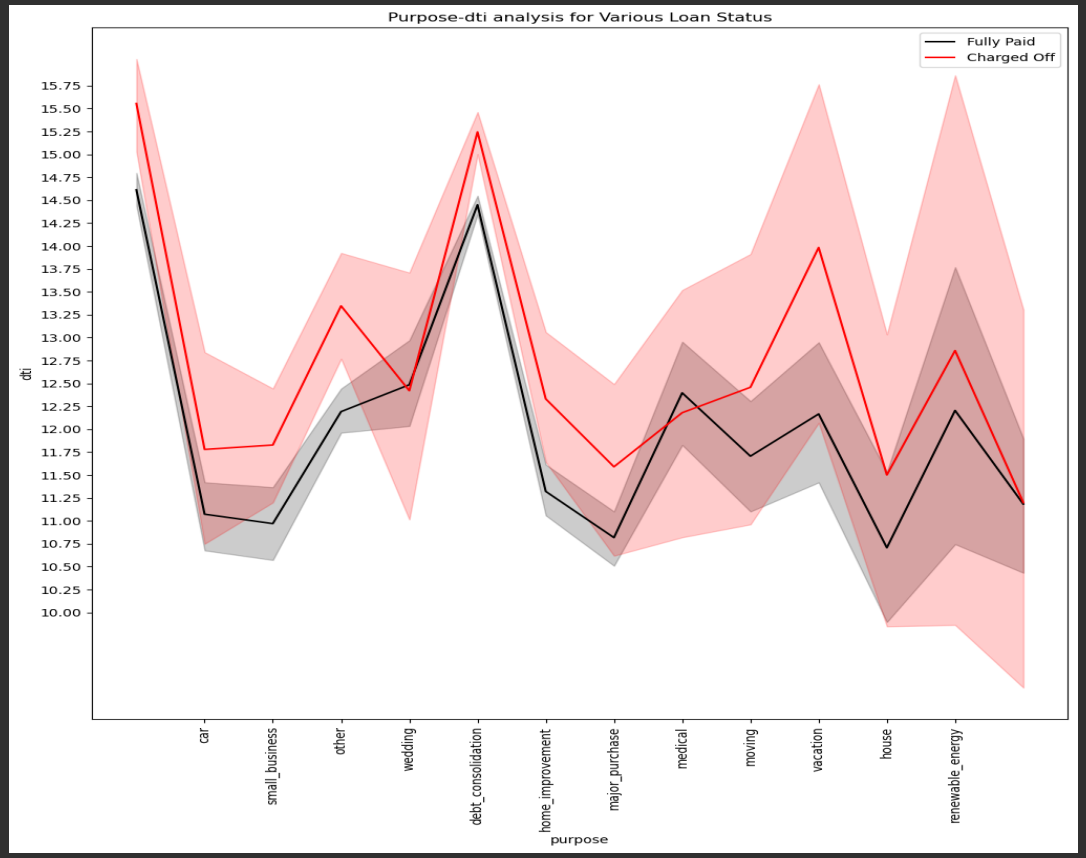
- Loans Charged Off with the dti > 13.5:
  - when issued during the Months
  - April-July
  - Oct-Dec



Multivariate Analysis (Purpose - dti)

- DTI → Debt to Income ratio

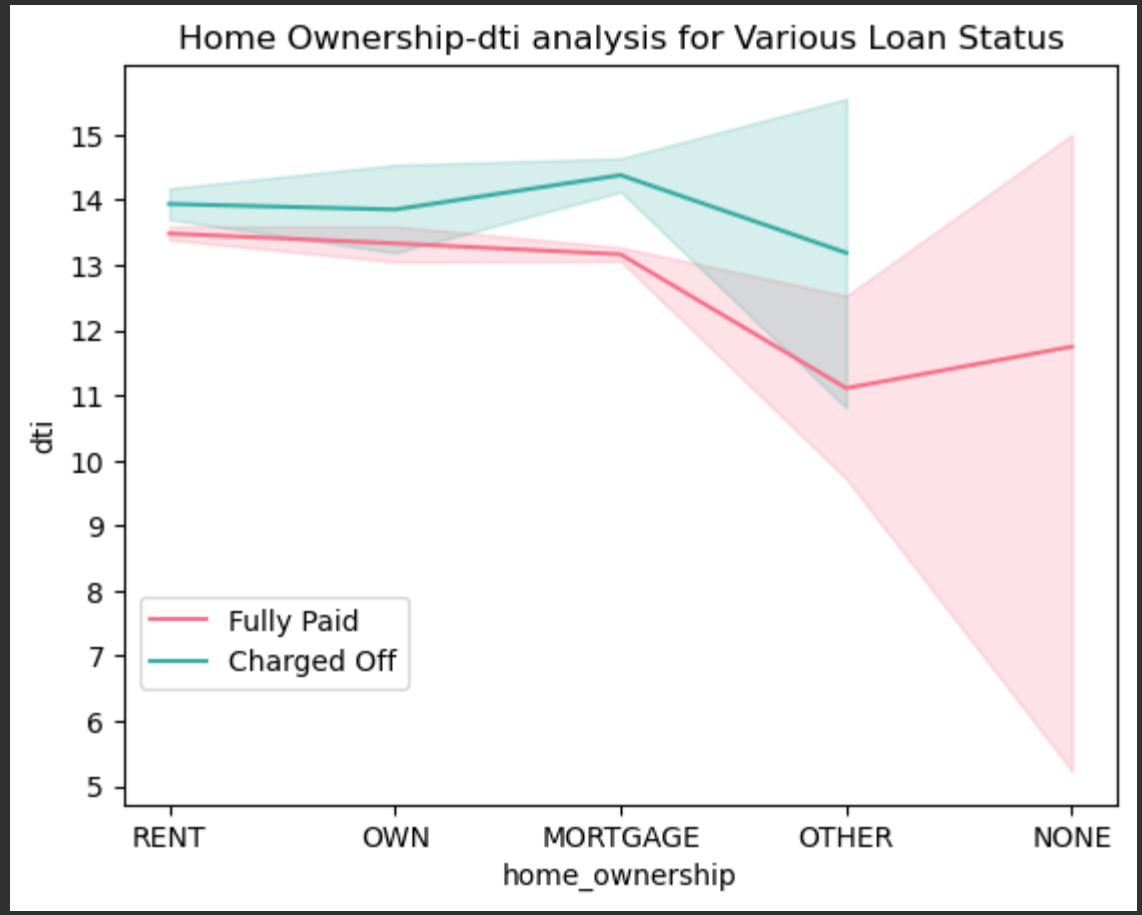
- Loans Charged Off with the following combination:
  - Small Businesses with dti > 11.50
  - Vacation with dti > 13.0
  - debt_consolidation with dti > 14.75
  - Major Purchases with dti > 11.25
  - House with dti > 11.50



Multivariate Analysis (Home Ownership - dti)

- DTI → Debt to Income ratio

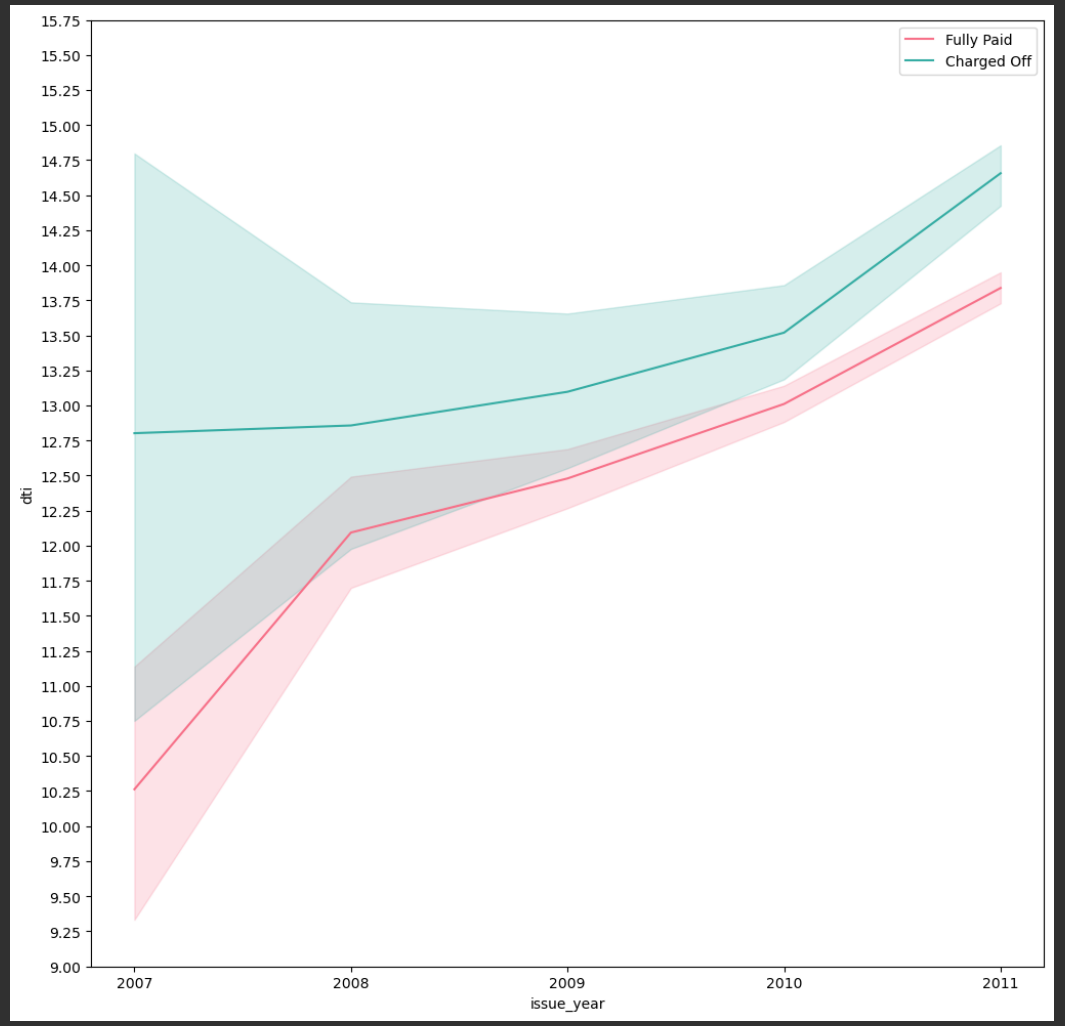
- Loans Charged Off with Home Ownership as MORTGAGE and having a DTI > 14.



Multivariate Analysis (Year- dti)

- DTI → Debt to Income ratio

- Loans issued during 2010 & 2011 with DTI> 13.25 are Charged Off



Observations

Loans with purpose of House and with **higher interest rate** are having changes of getting Charged Off

Loans are Charged Off when the **employment is 7 years** and the annual income is in the **range of [627K – 938K]**

The Loans with **Grade F** are likely to be Charged Off when the **loan_amount is > 16K**. and **Grade G** with **funded_amount_investor < 18K**

Observations – DTI analysis

- **DTI > 11.25** Purpose - Major Purchases
- **DTI >11.50** Purpose - Major Purchases, Small Business, House
- **DTI >13.0** Purpose - Major Purchases, Small Business, House, Vacation
- **DTI >13.5** Purpose - Major Purchases, Small Business, House, Vacation
  Months - Apr-Jul & Oct-Dec
- **DTI >14** Purpose - Major Purchases, Small Business, House, Vacation
  Home Ownership - Mortgage | Grade - Garde C
  Months - Apr-Jul & Oct-Dec
- **DTI >14.75** Purpose - Major Purchases, Small Business, House, Vacation,
  Debt Consolidation
  Home Ownership - Mortgage | Grade - Garde C
  Months - Apr-Jul & Oct-Dec

Conclusion

1. Loans for the Term tenure of 60 months are likely to be Charged Off.
2. Loans with higher interest rate are the key reason for Charged off. Hence, it is observed that more loans % are Charged off for the loans with higher grades.
3. Avoid loans for small businesses with DTI > 11.50
4. Avoid loans for debt consolidation with DTI > 14.75
5. Avoid loans which are not verified. Almost 44% of the loans are not verified. This could be a reason for the defaulters.
6. Avoid loans with DTI > 13.5 and issued during the months April-July & Oct-Dec.
7. Avoid loans issued for Home Ownership as MORTGAGE and DTI>14.
8. Avoid loans with type Grade C & DTI>14.
9. Avoid loans with type Grade F with loan amount > 16K & type Grade G with loan amount < 18K.
10. Avoid loans when employment tenure is 7 years & annual income is [627K-938K].
11. Avoid loans with state NE as they have highest Defaulted rate of loans