Lending Club Case Study

Prachi Goliwadekar

Rathnagiri Nagarajan

General Info

During the EPGP course, this case study serves as an assignment aimed at comprehending the given scenario, applying EDA techniques acquired throughout the coursework, and offering insights into data patterns to address business objectives.

Background of Lending Club

- Lending Club is a consumer finance company which specializes in lending various types of loans to urban customers.
- The company could face the following issues if it doesn’t evaluate the loan applicant diligently:
  - Credit Loss:
    - Common financial loss for lending companies
    - Arises from lending to 'risky' applicants
    - Occurs when borrowers fail to pay or default on their loans
  - Defaulters ('Charged-Off' Customers):
    - Borrowers who cause the largest amount of loss to lenders
    - Organization aims to identify and stay away from such borrowers
    - Goal is to minimize credit loss and financial risks

Objective

- The objective is to use EDA techniques against past loan applicant's data to:
  - Determine critical factors that will Improve ability to identify High-risk applicants and take precautionary actions
  - Cut down credit loss by identifying High Risk applicants driving attributes (or driver variables) that may lead to loan default
  - Improve Business portfolio and risk assessment

EDA Techniques used in evaluating case study

- Analyze, clean and prepare data for statistical analysis
- Univariate analysis
- Unordered Categorical Variable Analysis
- Ordered Categorical Variable Analysis
- Derived Variable Analysis
- Bivariate Analysis
- Correlation Analysis

Analyze Dataset and Clean data

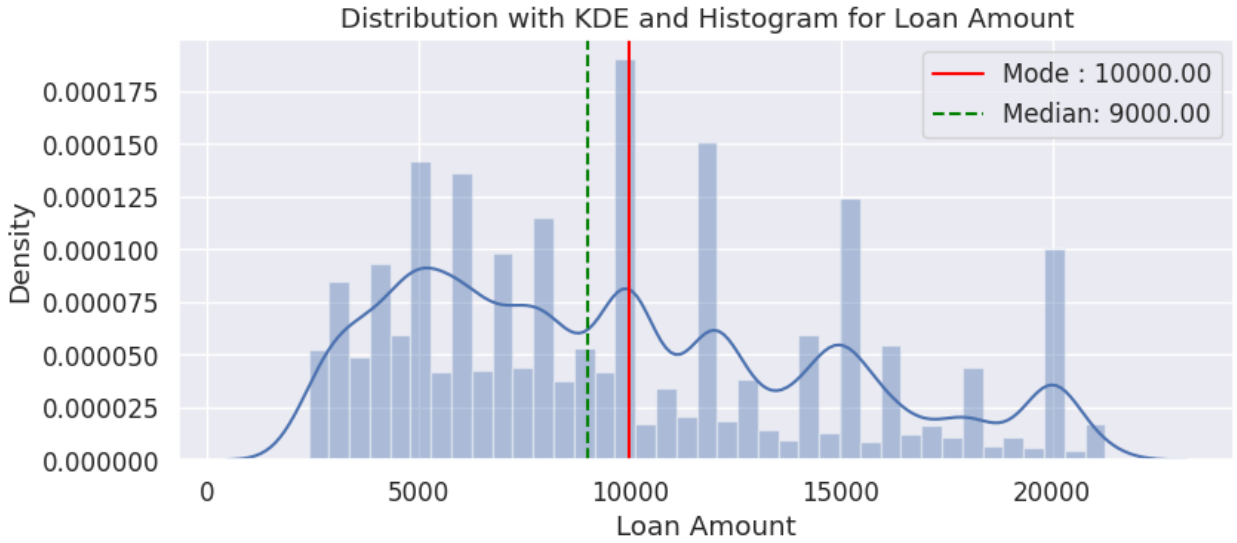
- Following methods were employed to prepare dataset for analysis
  - Find and drop the columns with all Null Values
  - Drop columns with few values or not be useful for analysis
  - Exclude rows that has outliers
  - Clean and Correct the datatype of columns
    - by removing trailing characters like “%” and converting it into relevant datatypes
    - split date string column into day, month and year columns
    - Rounding off columns to 2 decimal places
  - Consider only the rows with loan_status != “Current”, the reason being they are already paying customers
  - Drop text columns that may not be useful

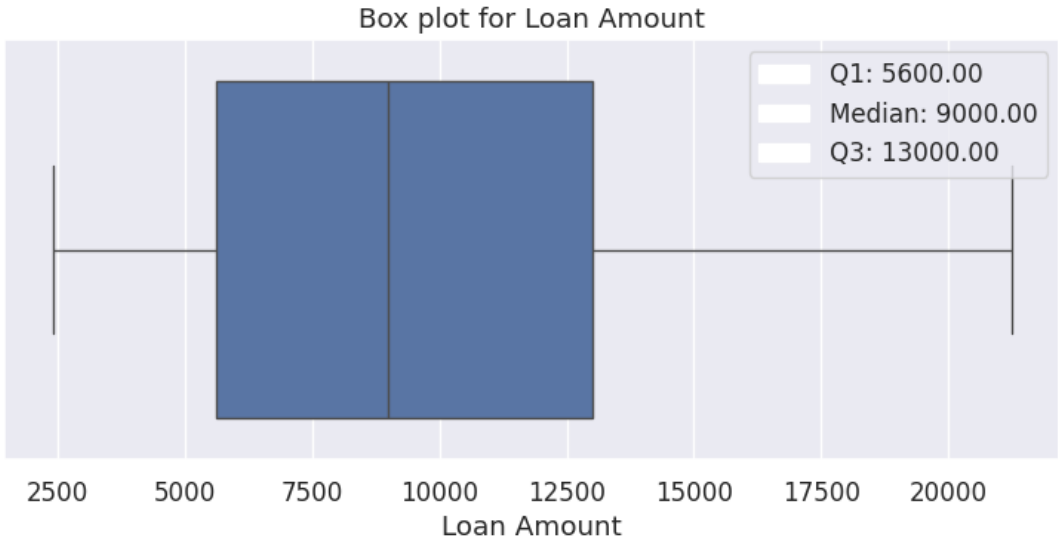
Analyze Dataset Contd..,

- Create and Add derived columns especially the following for histograms bins (buckets)
  - Loans Issue month bins as follows: Q1, Q2, Q3, Q4
  - Debt to Income ratio bucket: Very Low, Low, Moderate, High and Very High
  - Loan Amount bucket:0-5k, 5K – 10K, 10K-15K, 15K-above

Univariate Analysis

- The following variables were considered for Univariate Analysis:
  - Loan amount
  - Funded amount
  - Investors funded amount
  - Interest Rate
  - Installment amount
  - Annual Income of applicants
  - Debt to Income Ratio

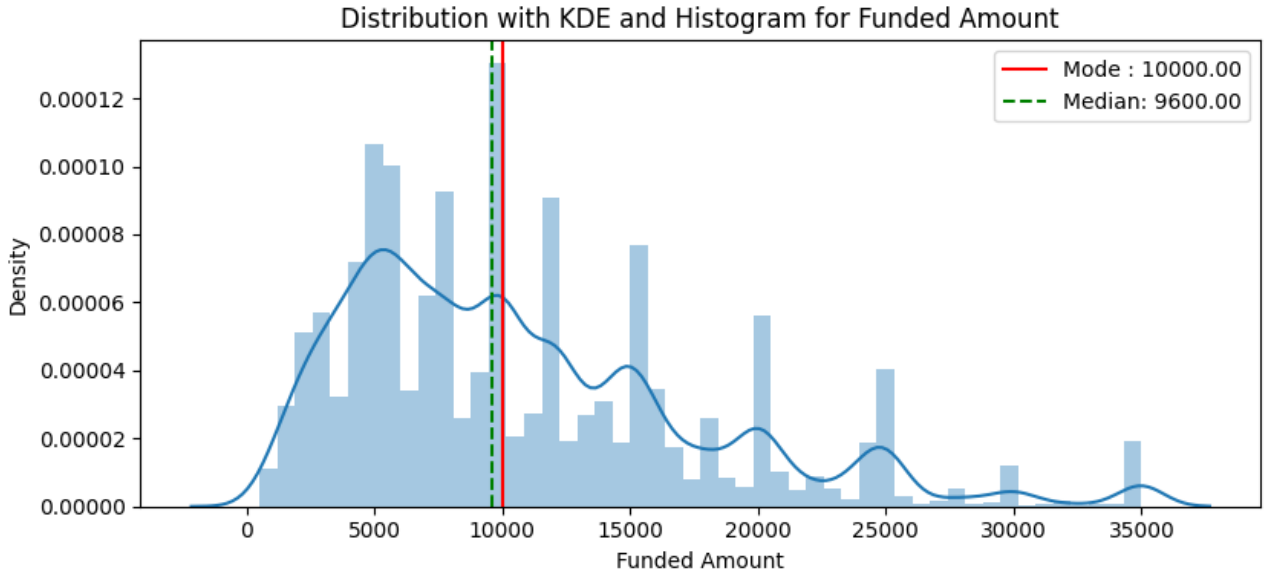


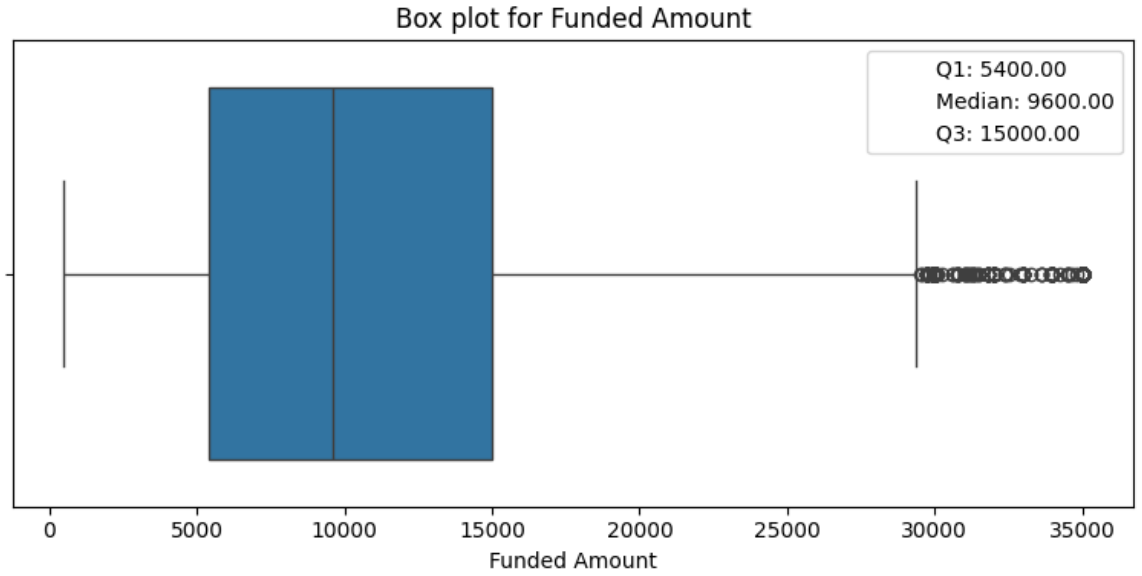


** The loan amount has been treated to remove outliers and accordingly the quantiles bottom 5% and top 90% removed*

Univariate Analysis - Loan amount

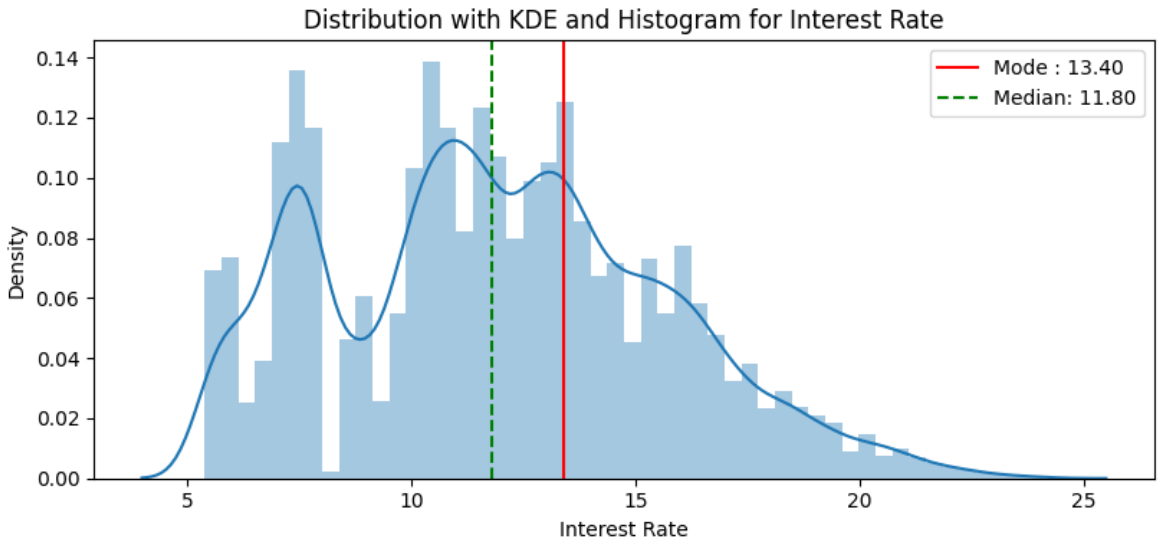
- The KDE curve indicates there is multi-modal clusters, With the peak of the KDE line is at 5000 and further smaller peak at 10000 and so on, indicating the most common loan amount is around the modal clusters of 5,000 and 10000.
- The median is 10000, signifying that half of the loans are below 10000 and half are above.
- While the box plot indicates that most of the loan amount is in range of 5600 and 13000, the difference between the smaller modes from modal clusters 5000 or 10000 and median from box plot confirms that the plot is rightward skew, with the "tail" of higher loan amounts greater than 15000, pulling the median towards higher value despite the majority number of loans provided being smaller.

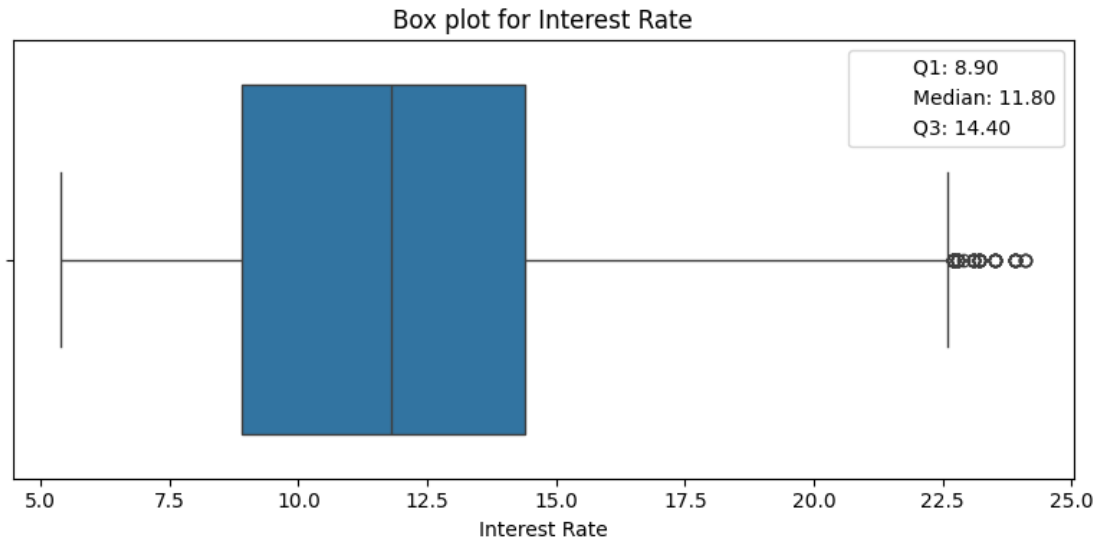




Univariate Analysis – Funded Amount

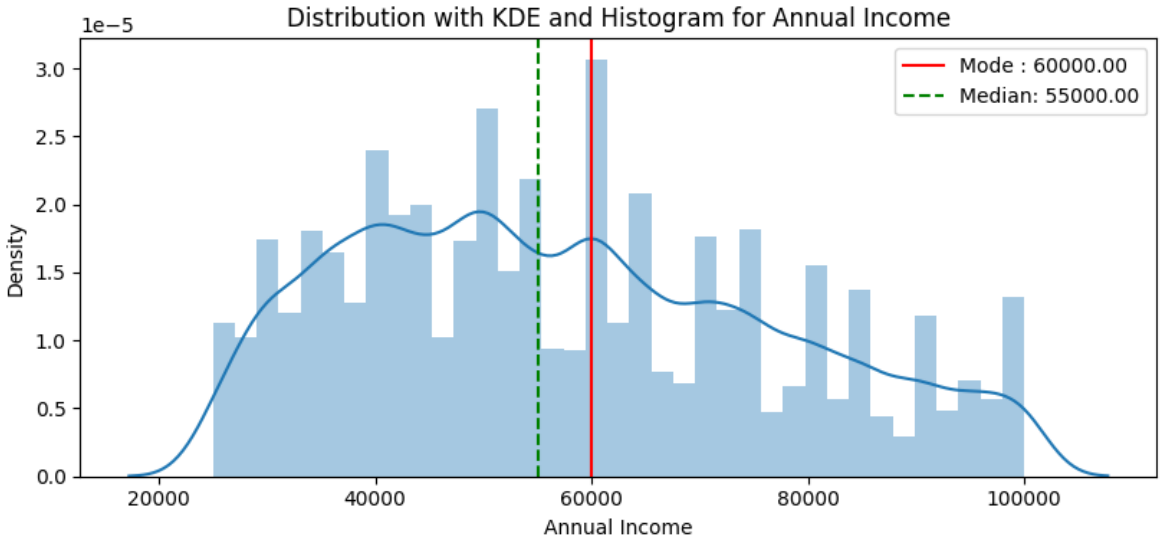
- The KDE curve indicates there is multi-modal clusters, With the peak of the KDE line is at 5000 and further smaller peaks at 10000 and so on, indicating the most common loan funded amount is around the modal clusters of 5,000 and 10000.
- The median is 9600, signifying that half of the loans are below 10000 and half are above.
- While the box plot indicates that most of the loan amount is in range of 5000 and 15000, the difference between the smaller mode 5000 and median from box plot confirms that the plot is rightward skew, with the "tail" of higher loan funded amounts greater than 15000, pulling the median towards higher value despite the majority number of loans funded being smaller.

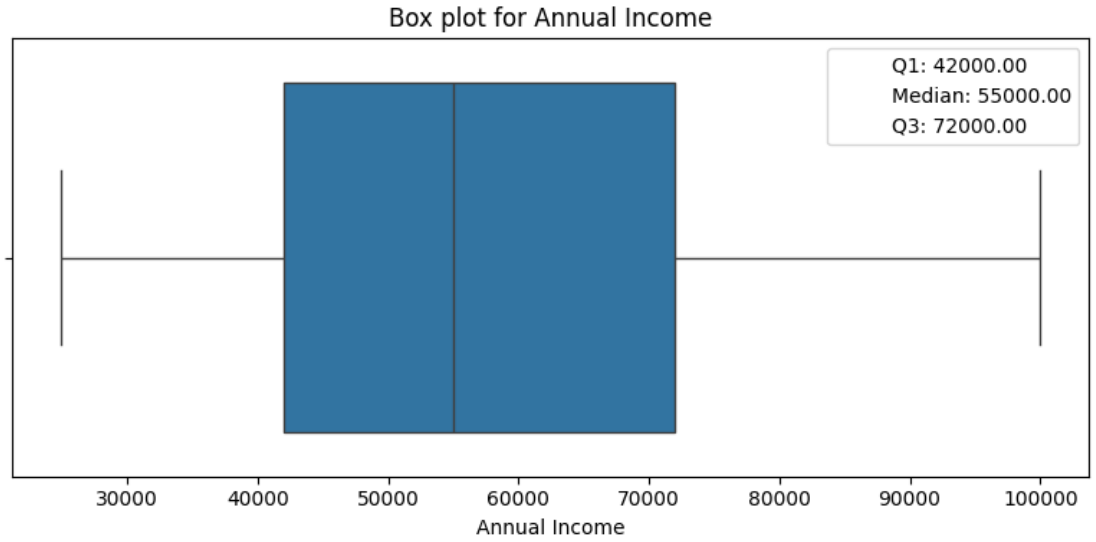




Univariate Analysis – Interest Rate

- The KDE curve indicates there is multi-modal clusters, With the peak of the KDE line is at 10.8 and further smaller peaks at 7.5, 13.40 and so on, indicating the most common Interest rate is around the modal clusters of 10.8 and 13.40.
- The median is 11.8, signifying that half of the Interest rates are below 11.8 and half are above.
- While the box plot indicates that most of the Interest rates is in range of 8.90 and 14.40, the difference between the smaller mode 7.5 or 10.8 or 13.40 and median from box plot confirms that the plot is rightward skew, with the "tail" of higher *Interest rate amounts greater than 14.0, pulling the median towards higher value despite the majority number of Interest rates being smaller.

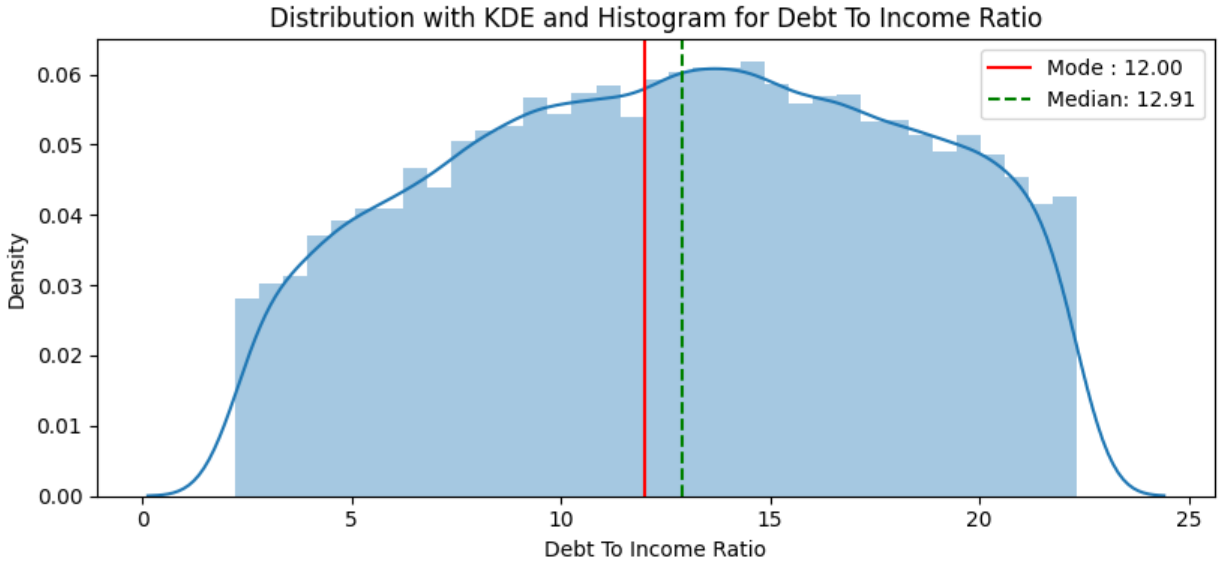


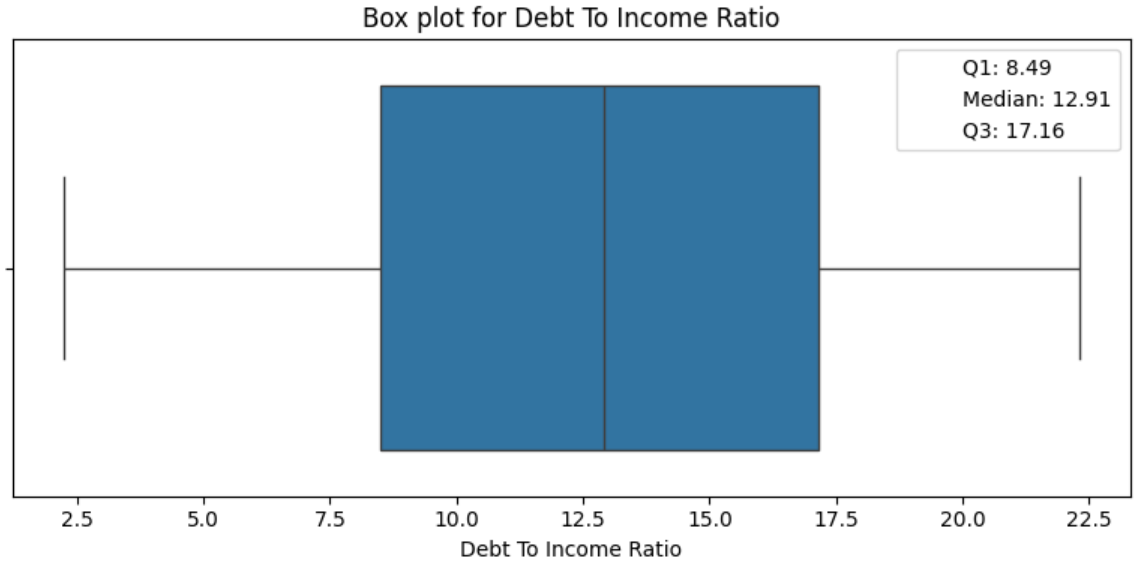


** Due to large number of outliers the quantiles of bottom 5% and top 85% were excluded from the analysis*

Univariate Analysis – Annual Income

- Before removal of outliers, Most of the annual income was between 5 lacs to 10 lacs and therefore This column required major outlier treatment. So, the Annual income corresponding to bottom 5% and top 15% were removed and only data between 5% to 85% were considered.
- The KDE curve indicates there is multi-modal clusters, With the peak of the KDE line is at 50000 and further smaller peaks at 40000 and 60000 and so on, indicating the most common Annual Income is around the modal clusters of 40000 and 60000.
- The median is 55000, signifying that half of the Annual Income are below 55k and half are above.
- While the box plot indicates that most of the Annual Income is in range of 42000 and 72000, the difference between the smaller mode values of 50000 or 40000 or 60000 and median from box plot confirms that the plot is rightward skew, with the "tail" of higher *Annual Income* amounts greater than 72000, pulling the median towards higher value despite the majority number of Annual Income being smaller.





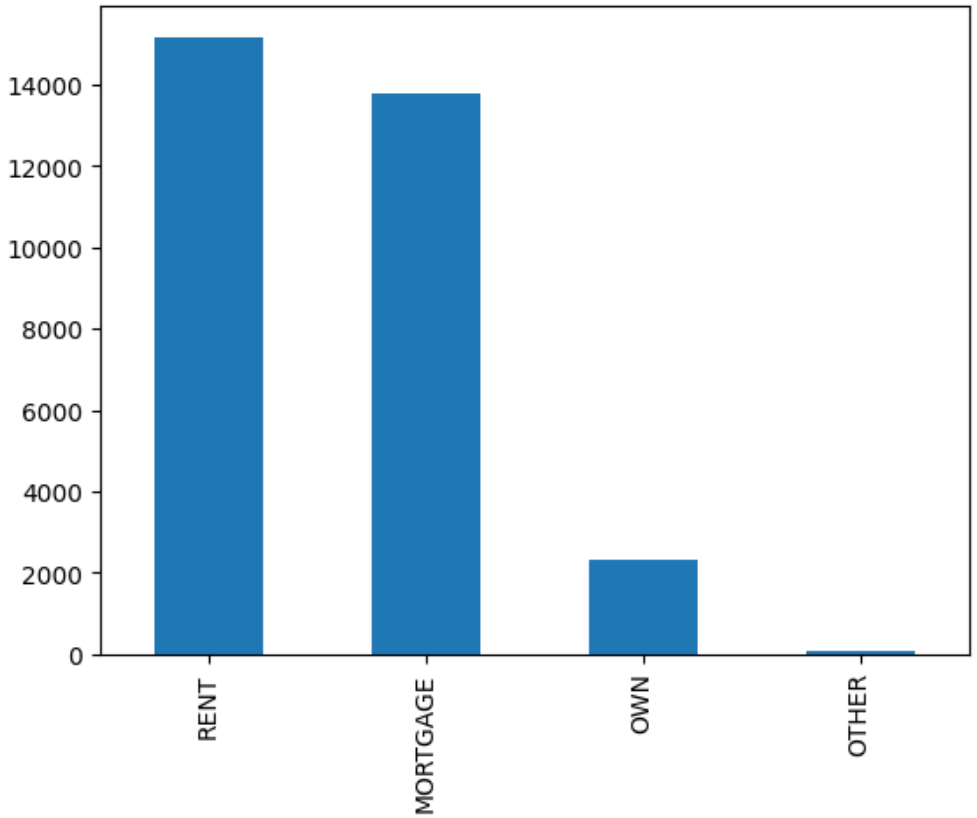
** Due to large number of outliers the bottom 5% and top 10% were excluded from the analysis*

Univariate Analysis – Debt to Income (DTI) Ratio

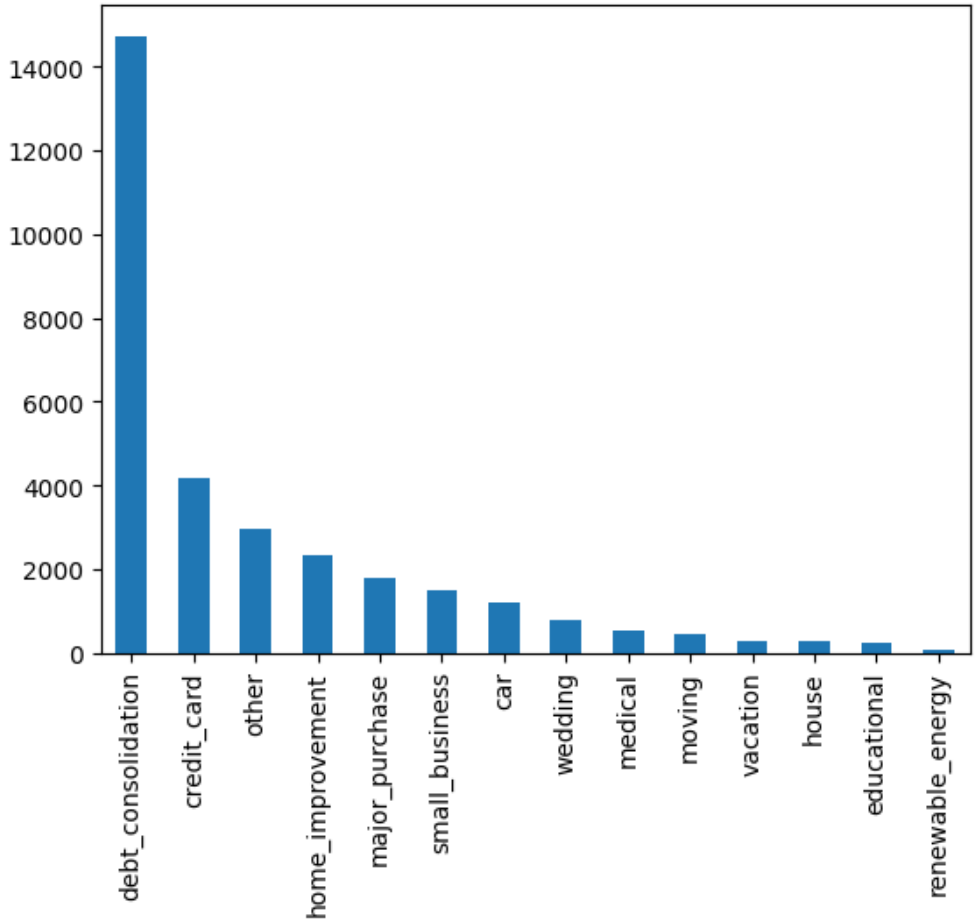
- The KDE graph and histogram plot now shows a central tendency and single modal with most of the applicant debt to income ratio were in the range of 8.49% and 17% approximately.
- The single modal cluster seems to be concentrated around the 12% which shows most loans are being given to applicants with Low DTI (0-15%).
- This suggests most applicants have a strong financial position and a lower to medium risk of defaulting on debt obligations

Unordered Categorical Variable Analysis

Majority of the homeowner status are in status of RENT and MORTGAGE

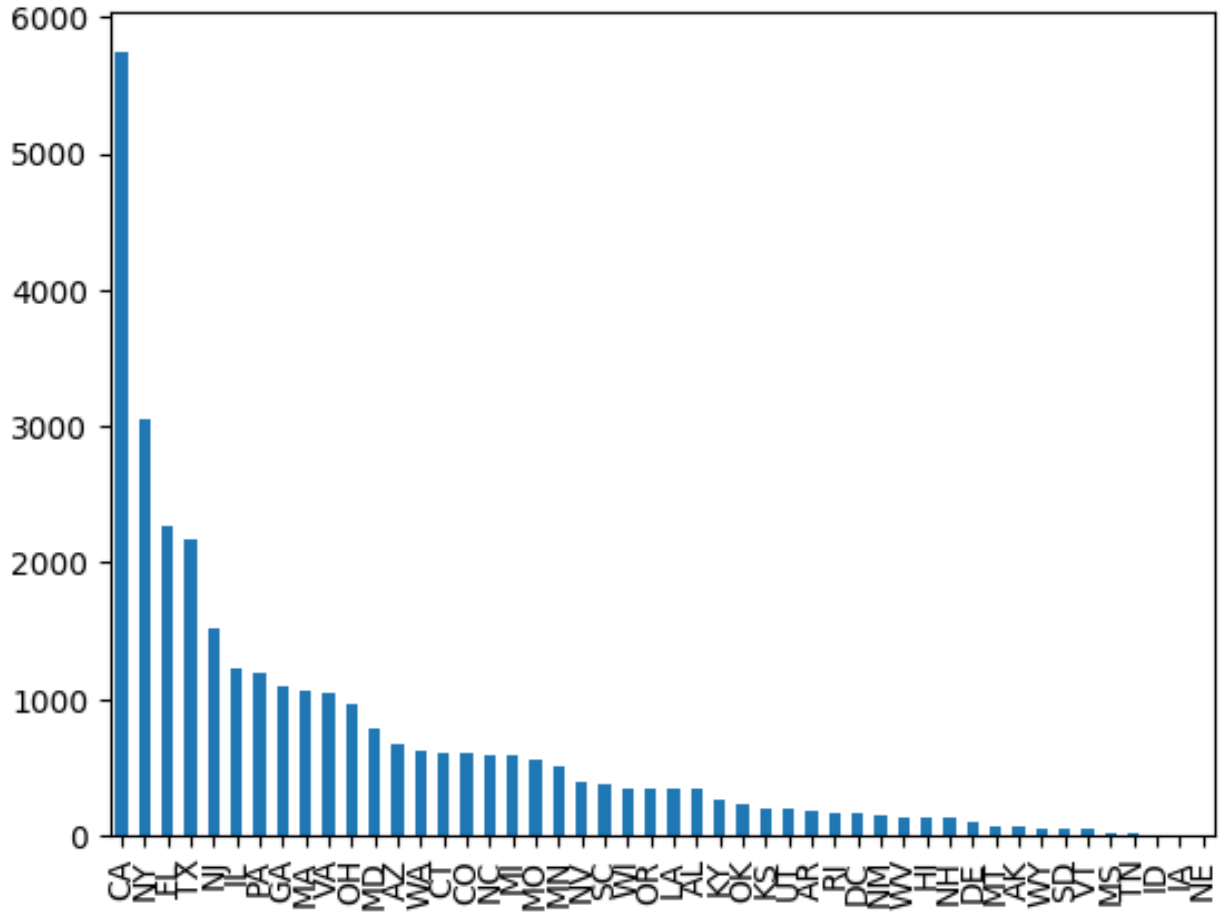


Majority of the purposes of loan is debt consolidation



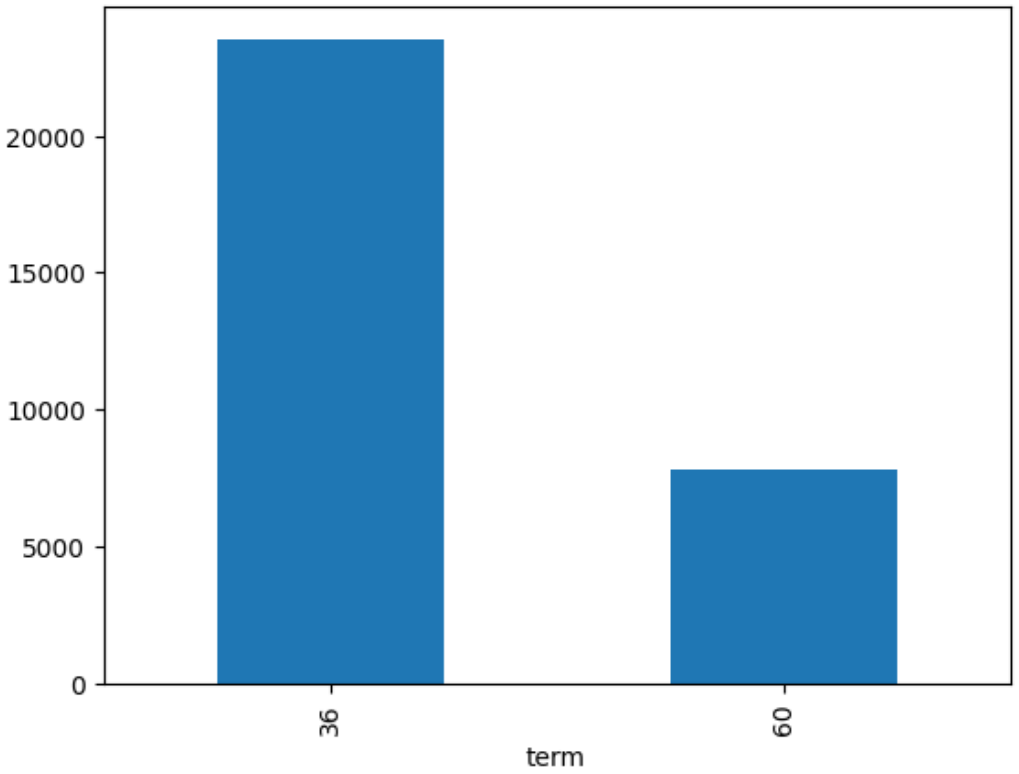
Unordered Categorical Variable Analysis

CA State has the maximum amount of loan applications

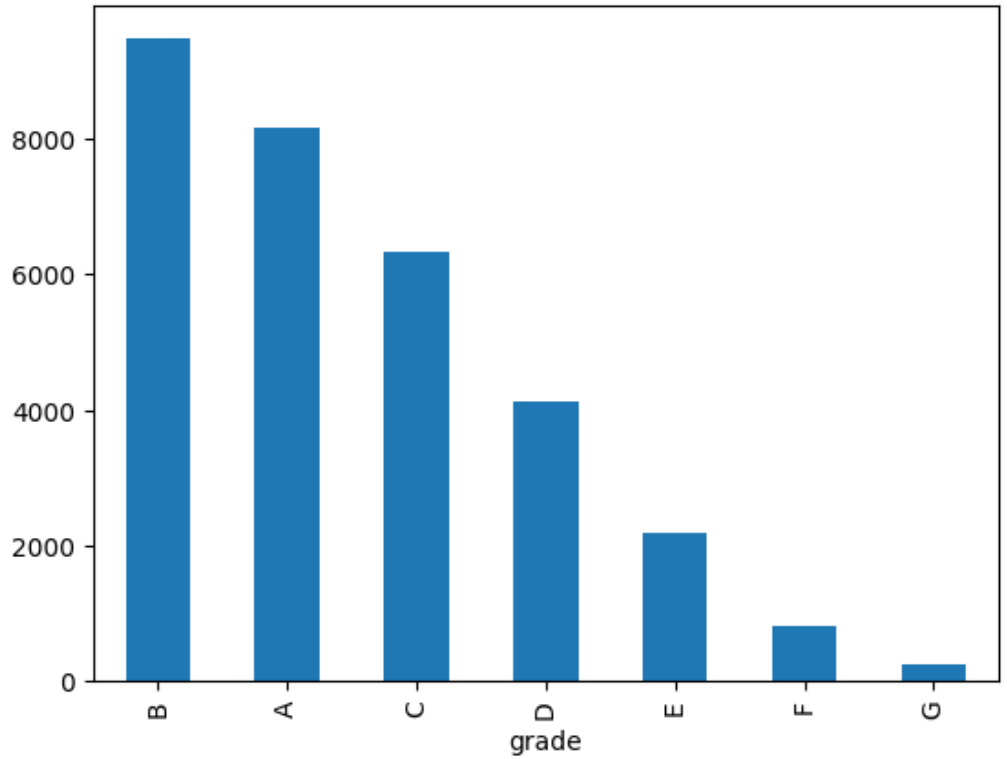


Ordered Categorical Variable Analysis

Most of the loans are of 36 months duration

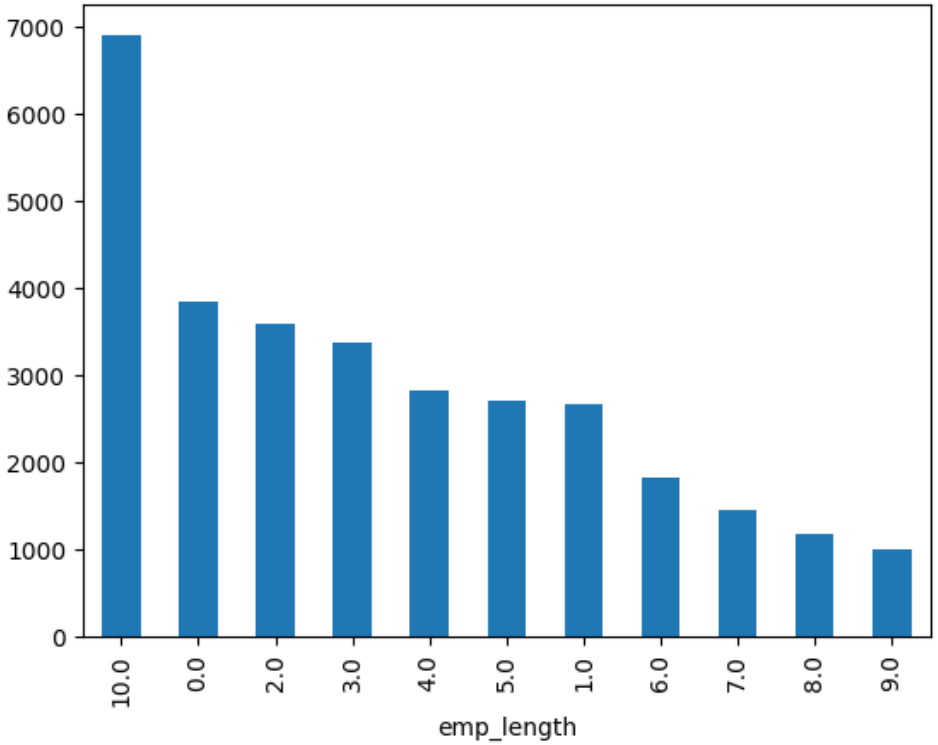


Most of the loans are of Grade B

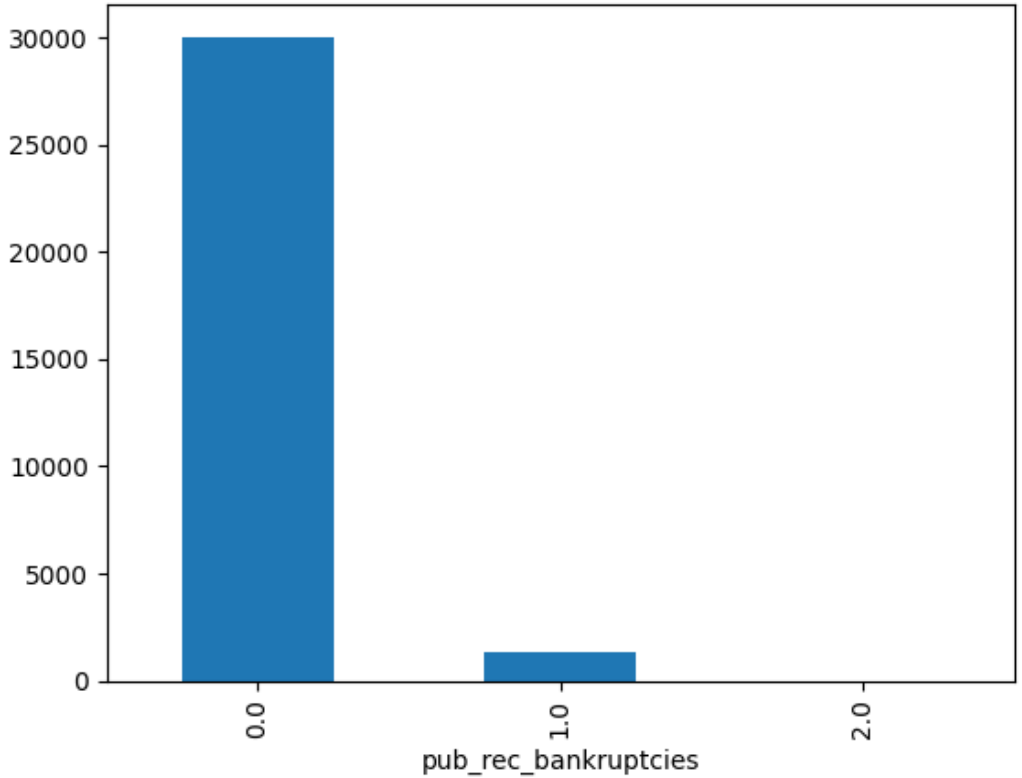


Ordered Categorical Variable Analysis

Majority of the employment length of the customers are 10+ years and then in the range of 0-2 years

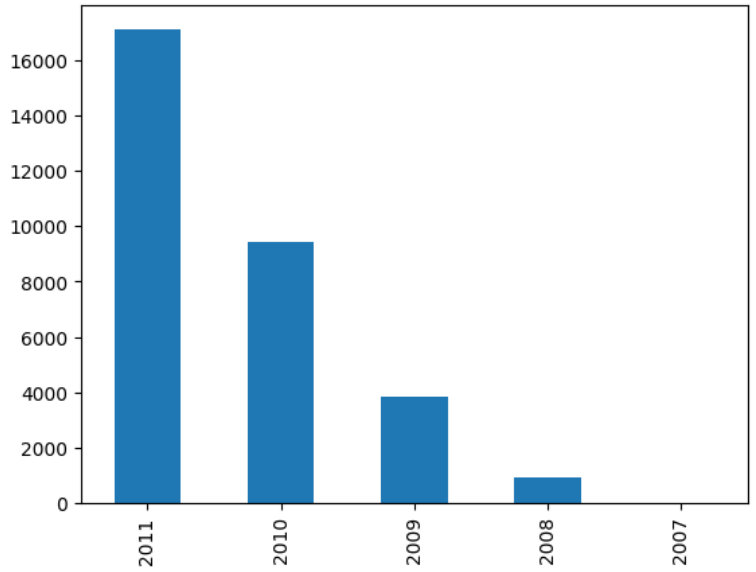


Majority of the loan applicants are in the category of not having a public record of bankruptcies

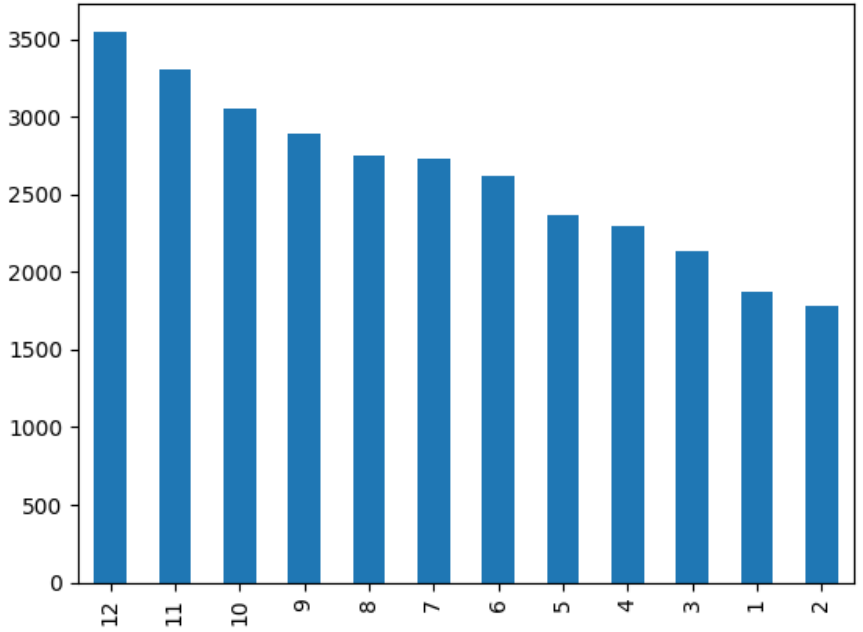


Derived Variable Analysis

Loan application count year over year



Loan application counts highest in year end

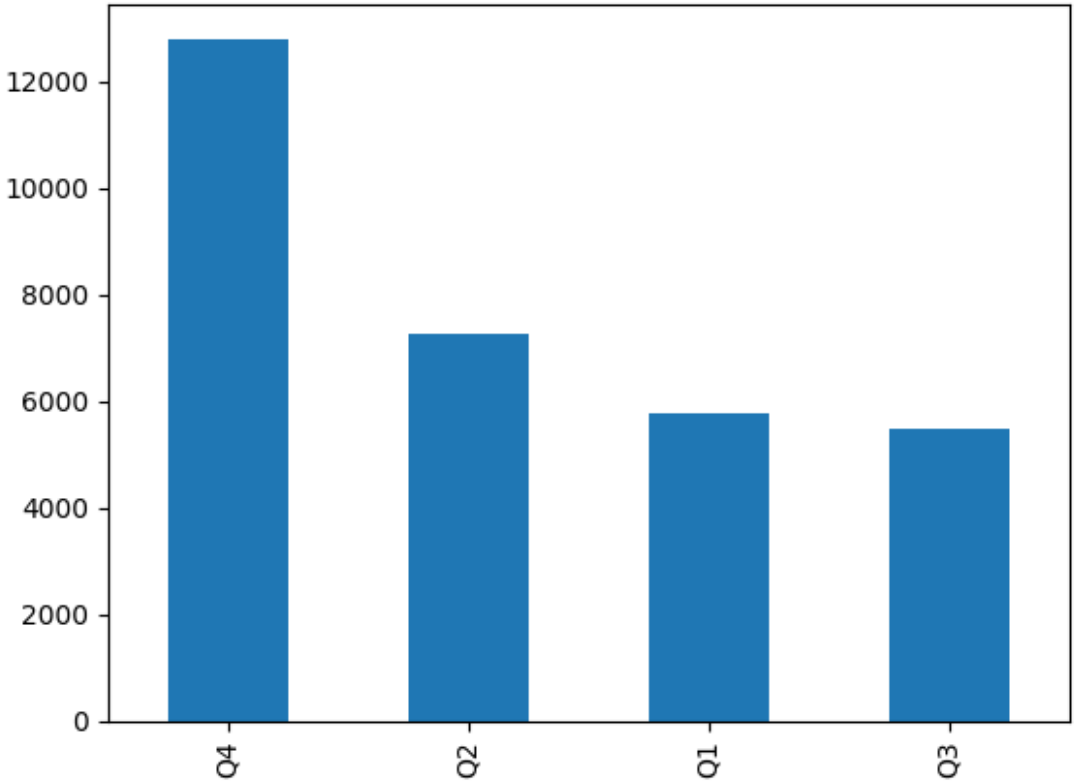


The lowest loans application count are in the month of Jan/Feb/March and highest counts are in 10/11/12.

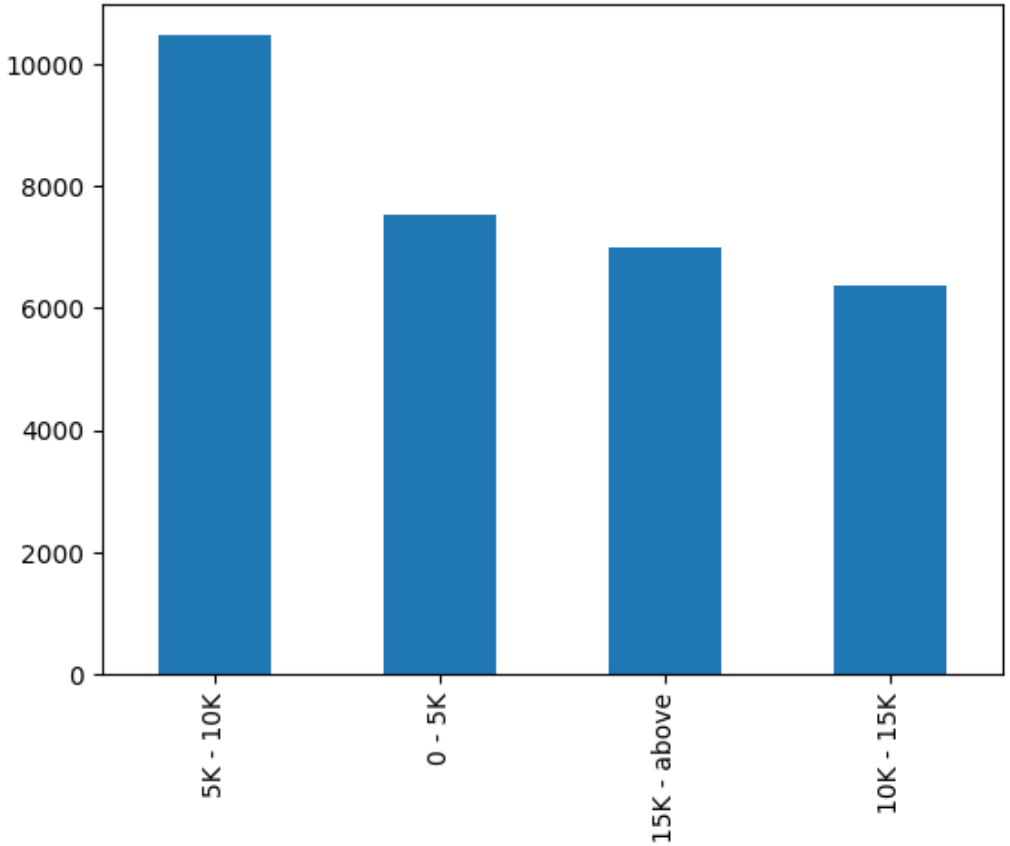
- Possibly because by year ends people face the financial challenges
- Possibly because of festive seasons
- Possibly because they are consolidating debt by year end

Derived Variable Analysis

Highest loan application volume in Quarter 4 of a year

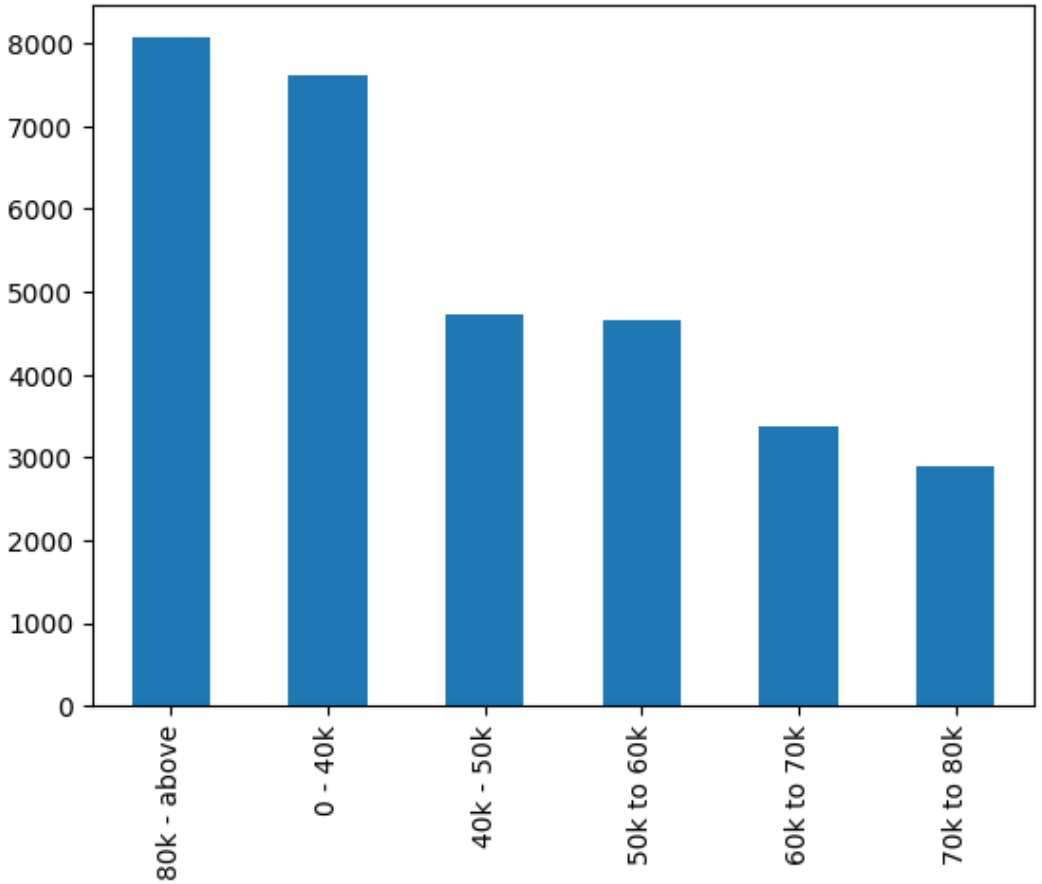


Highest loan amount applications fall in the range of 5k to 10k

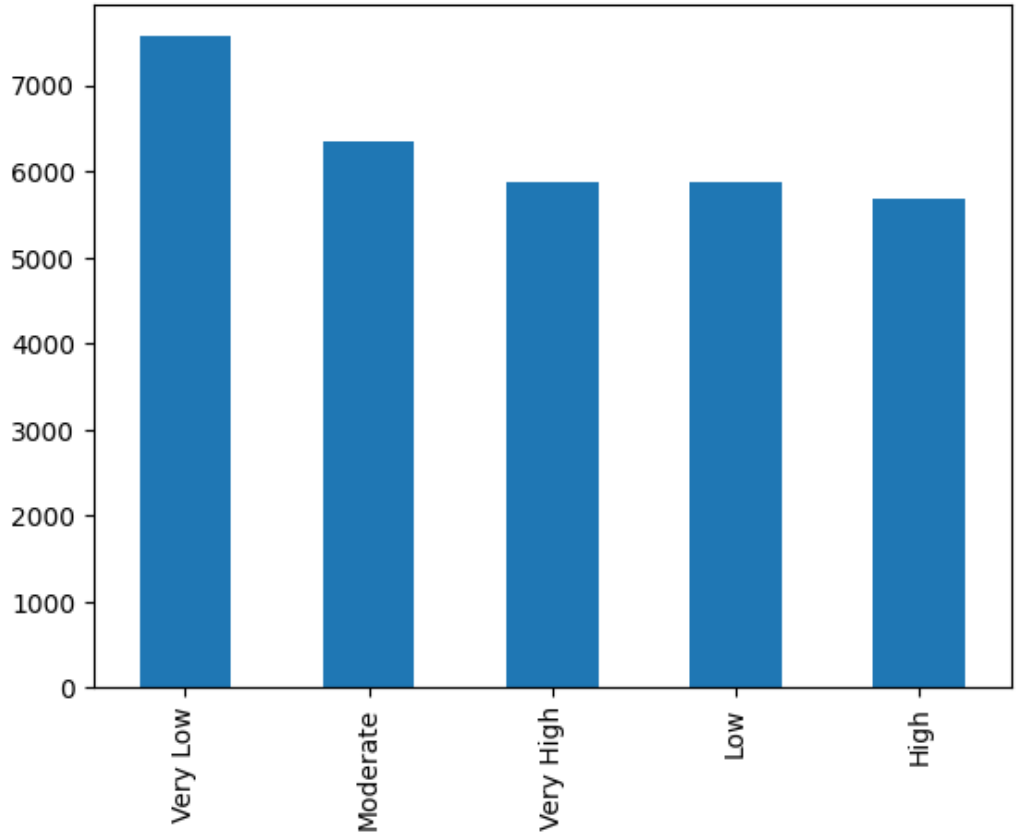


Derived Variable Analysis

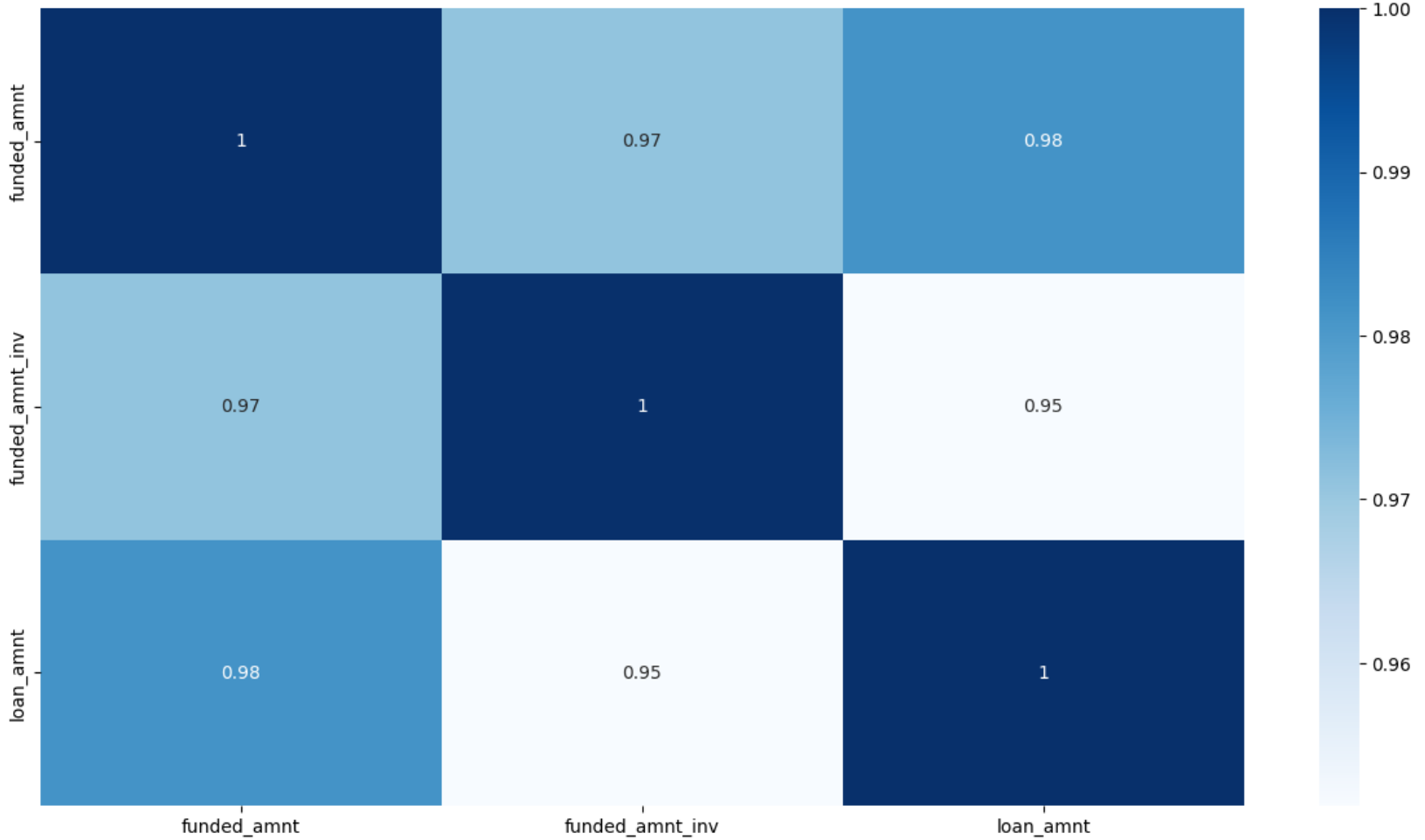
Highest loan applicant income is 80k and above



Most applicants have Very low or Moderate DTI ratio indicating financially stable applicants



Derived Variable Analysis – Identify Key correlations

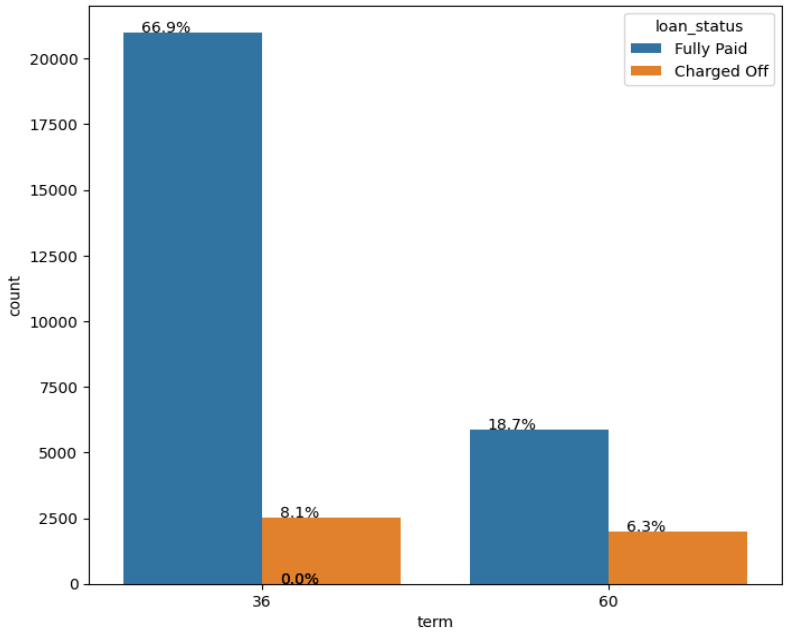


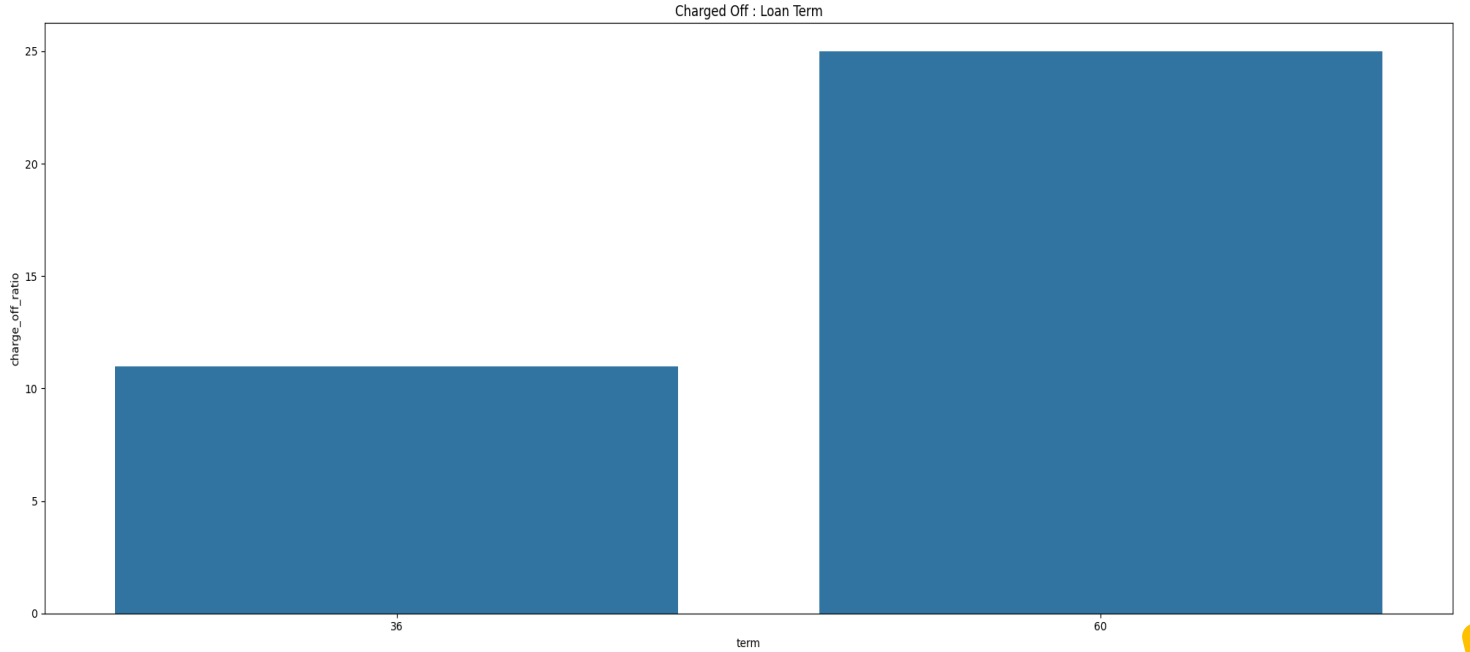
Loan Amount, investors funded amount and the funded amount all have high correlation – indicating most people get funded their loan amount

Bivariate Analysis

- A Statistical Method used to analyze the relationship between two variables.
- We Analyze the “Charges off” loan status against different relevant attributes of loan dataset.

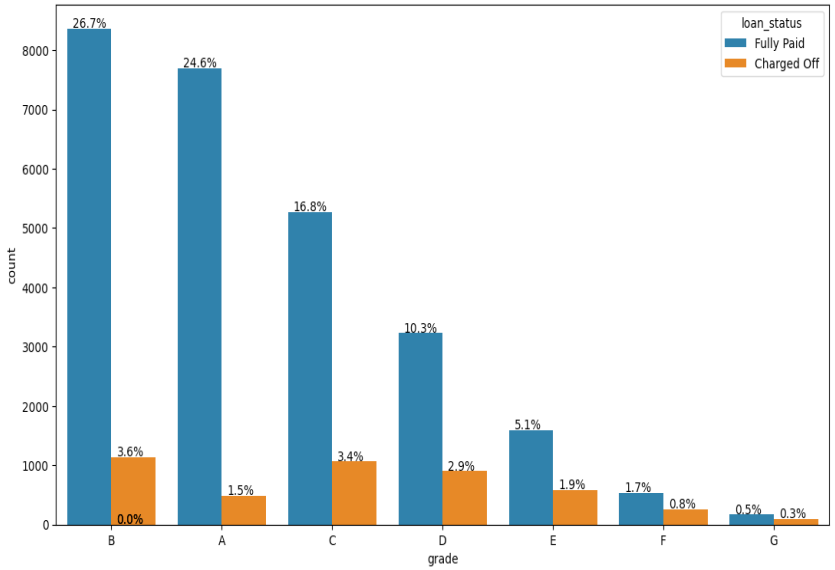
Bivariate Analysis – Term Limit Vs Loan Status

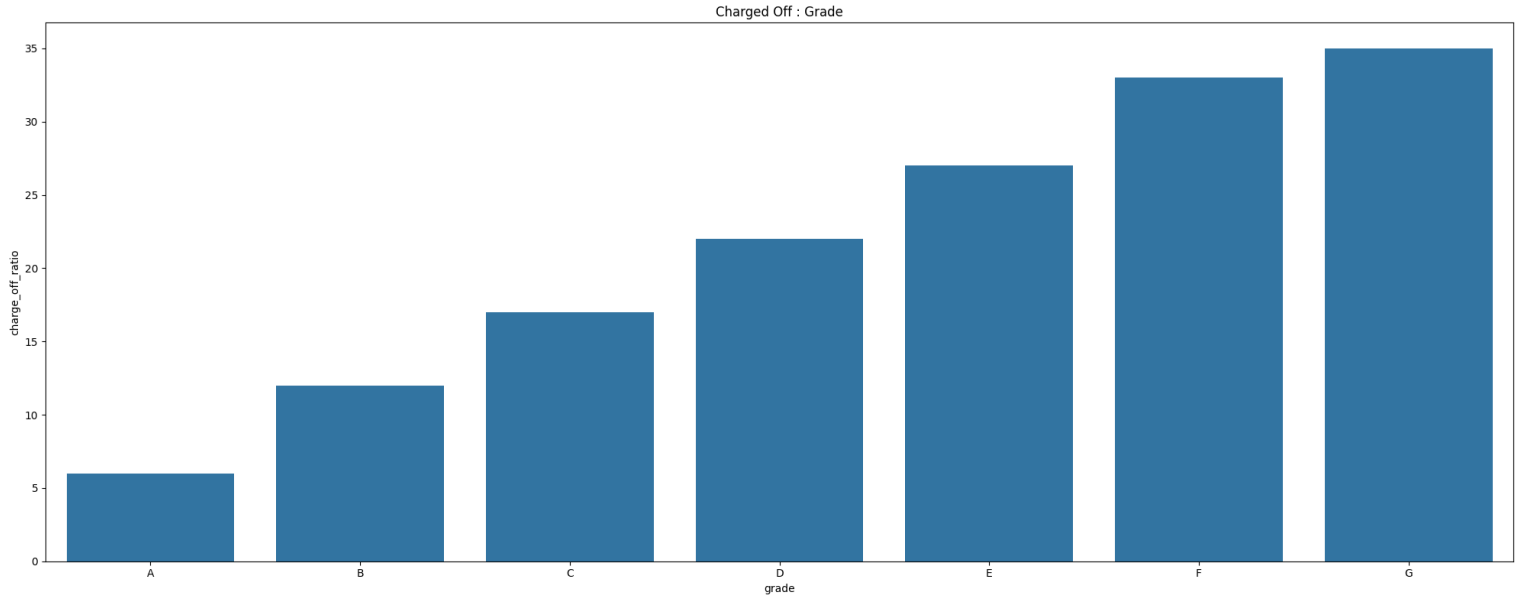




- The high volume of loans are in the category of term = 36
- The overall percentage of volume of Charge Off's is slightly higher in term = 36 (8%) as compared to term=60 (6%)
- If we calculate the ratio of Charge Off's within a category
  - Charge Offs ratio is for the term=60 is 25% which is much higher than term=36 (10%)
  - term=60 is the loan applications which require more scrutiny
  - Most of the applicants with term=60 potentially will have high Charge Offs

Bivariate Analysis – LC Grade Vs Loan Status



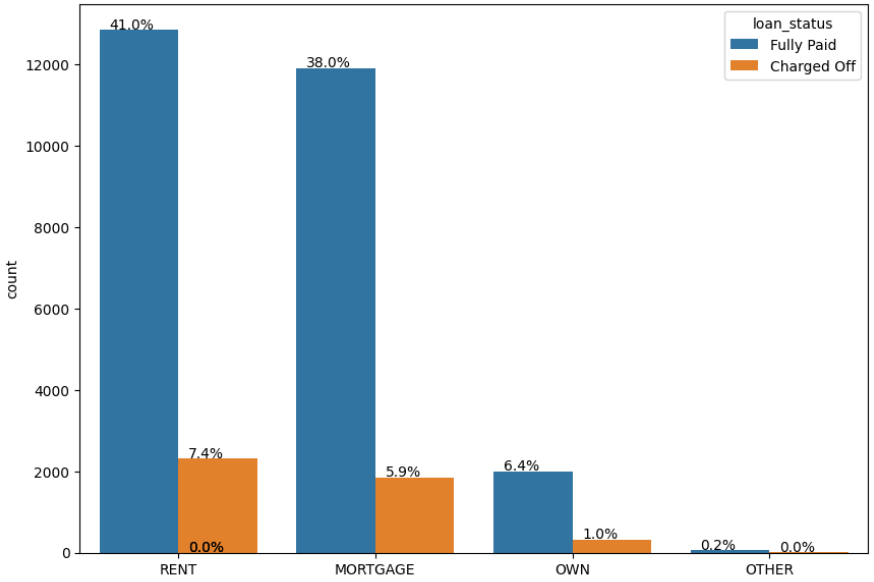


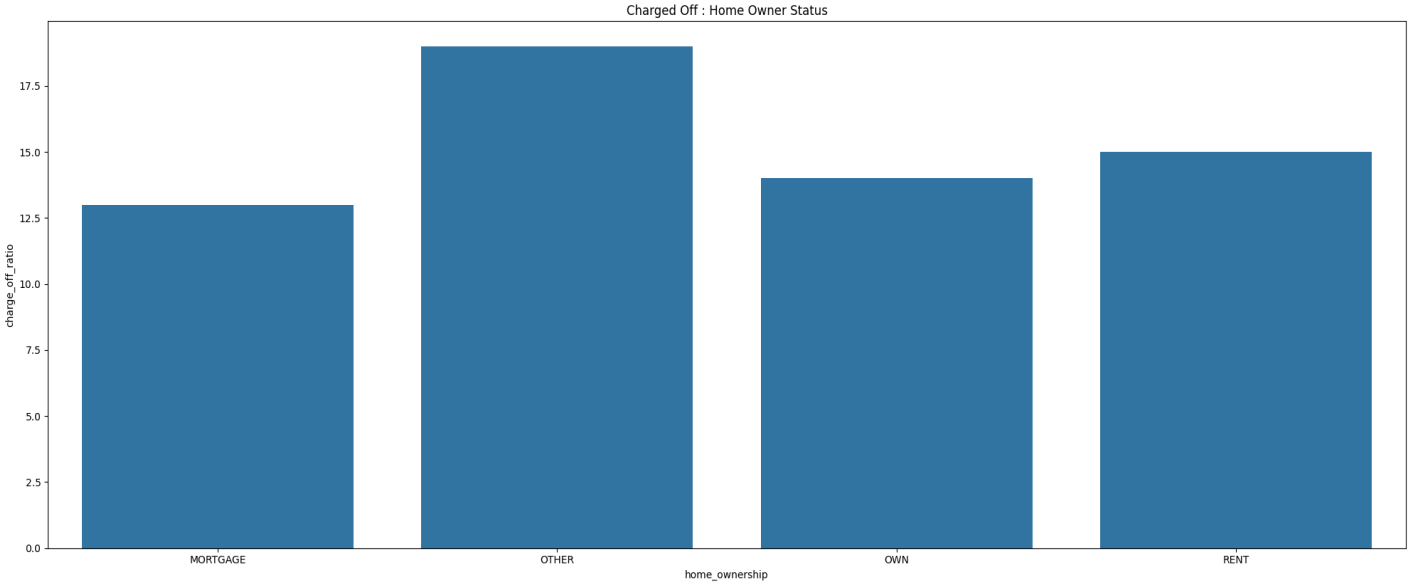
- Most of the loan volume is in grade=B
- Highest percentage of overall Charge Offs are in grade B (3.7%) and C(3.6%)
- If we analyze the Charge Off Ratio to loan status within a category
  - The highest percentage of Charge Offs are in the grade=G
  - Highest cluster of Charge Offs are in the grades G,F (> 30%)
  - The volume of Grade G is extremely low 158 thus it does not contribute to overall risk significantly

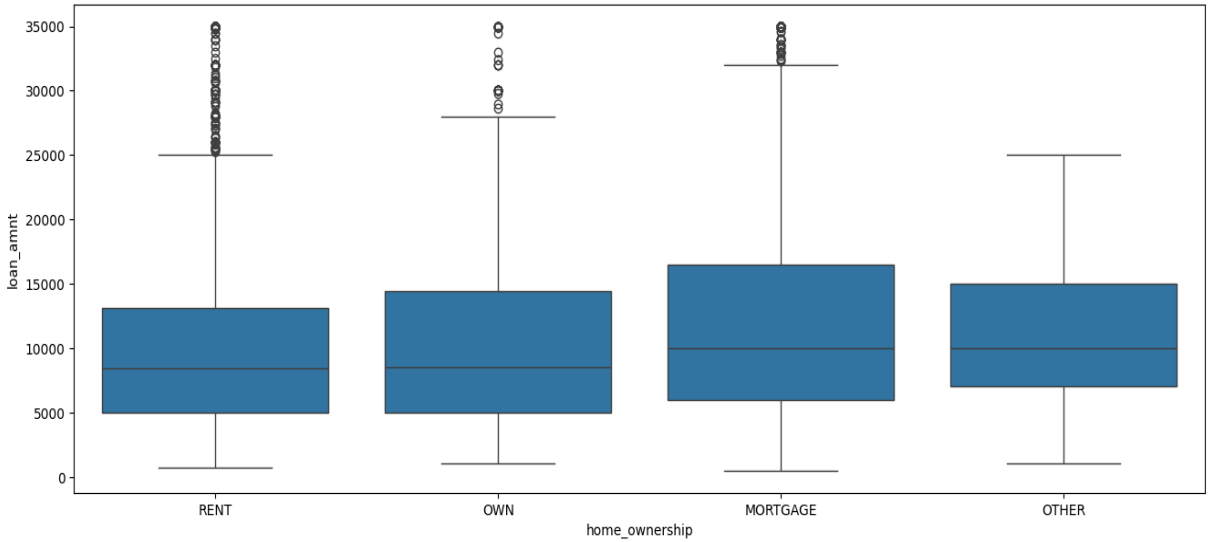
Inferences

- Highest risk of charge offs are in the grades of B and C in terms of volume
- But, In terms of ratio percentages, Grade "F" and "G" have very high chances of charged off. Grade "A" has very less chances of charged off.
- Probability of charged off is increasing from "A" to "G"

Bivariate Analysis – Home Ownership Vs Loan Status

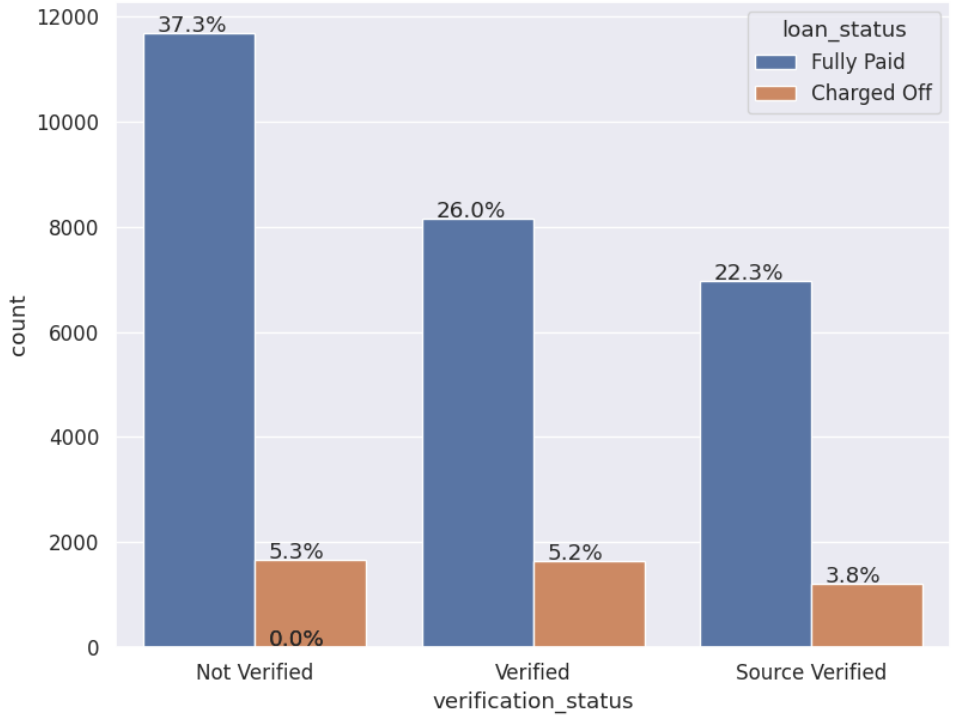


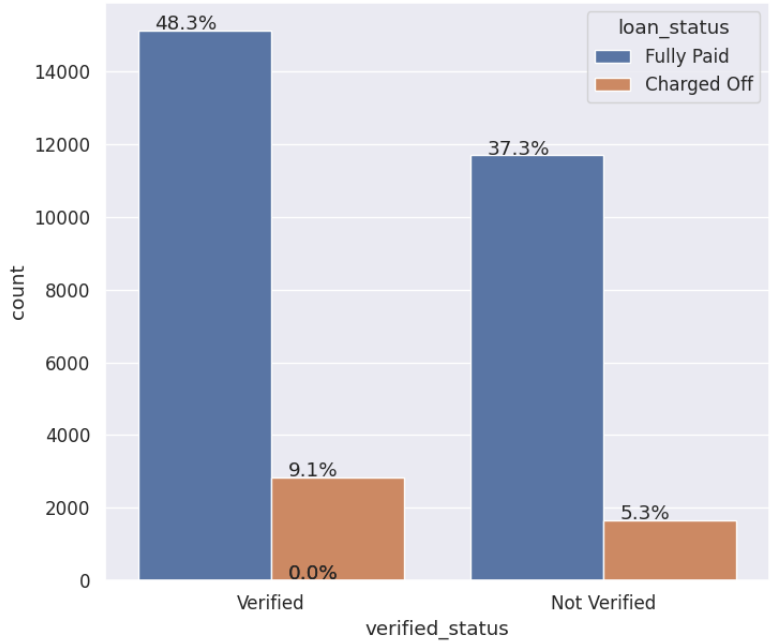


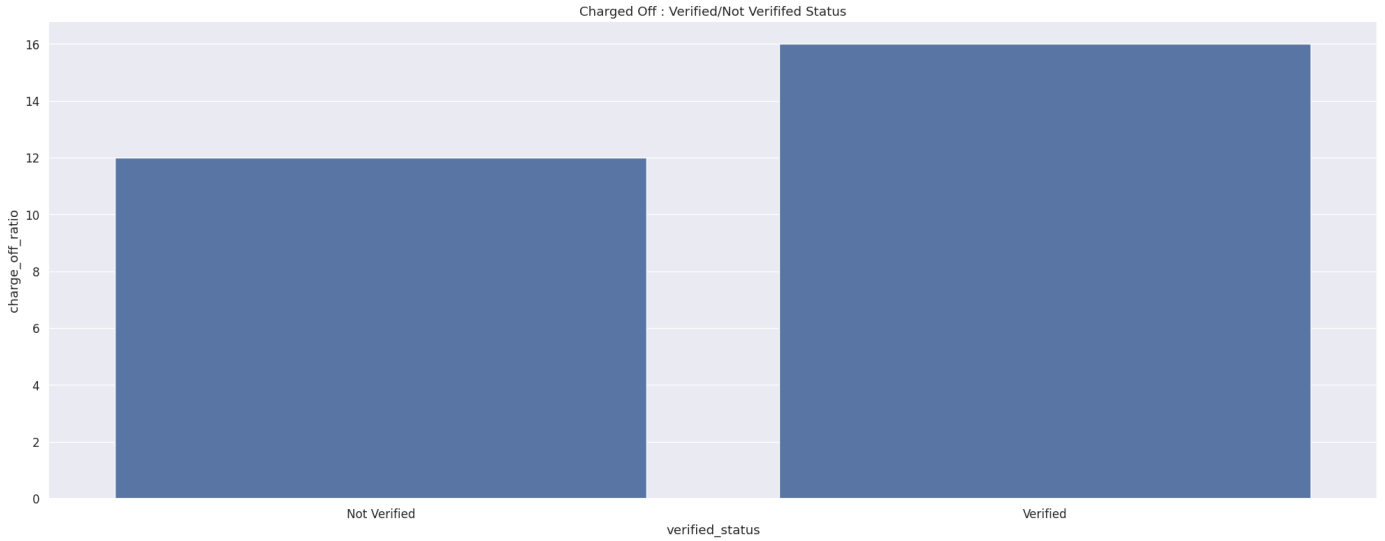


- Overall highest Charge Off numbers are in the category of RENT and MORTGAGE
- Within each homeownership category the ratio of Charge Off's is higher (ignoring Others) on Rented applicants
- The homeownership status of RENT and MORTGAGE are at the highest risk of Charge Offs

Bivariate Analysis – Verification Status Vs Loan Status

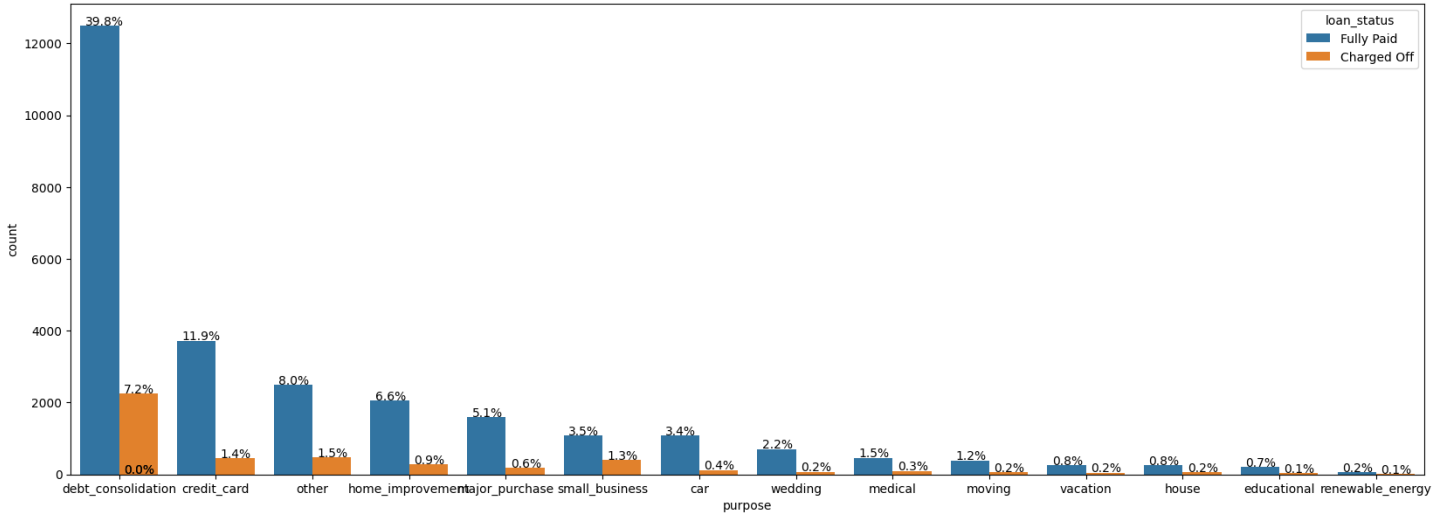


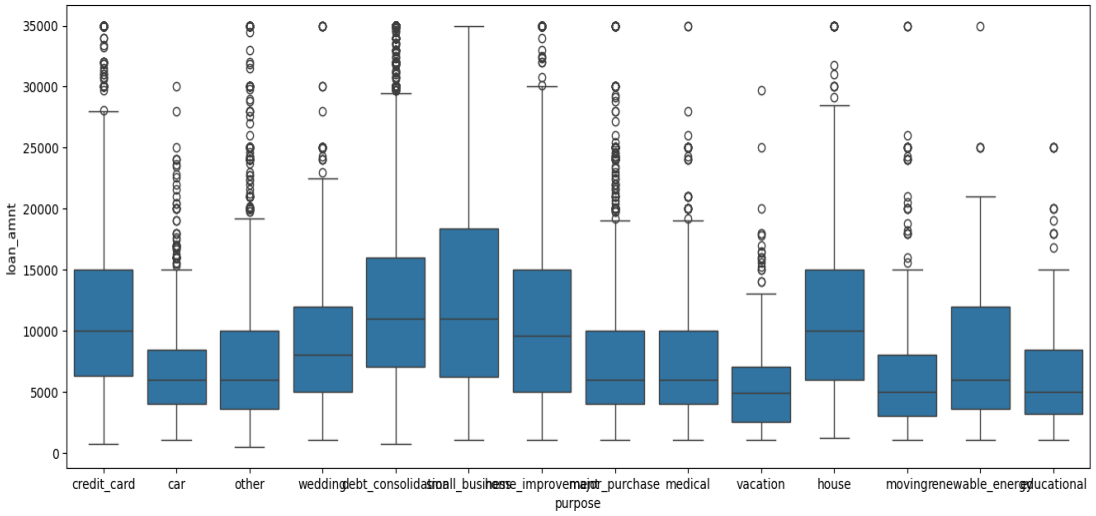


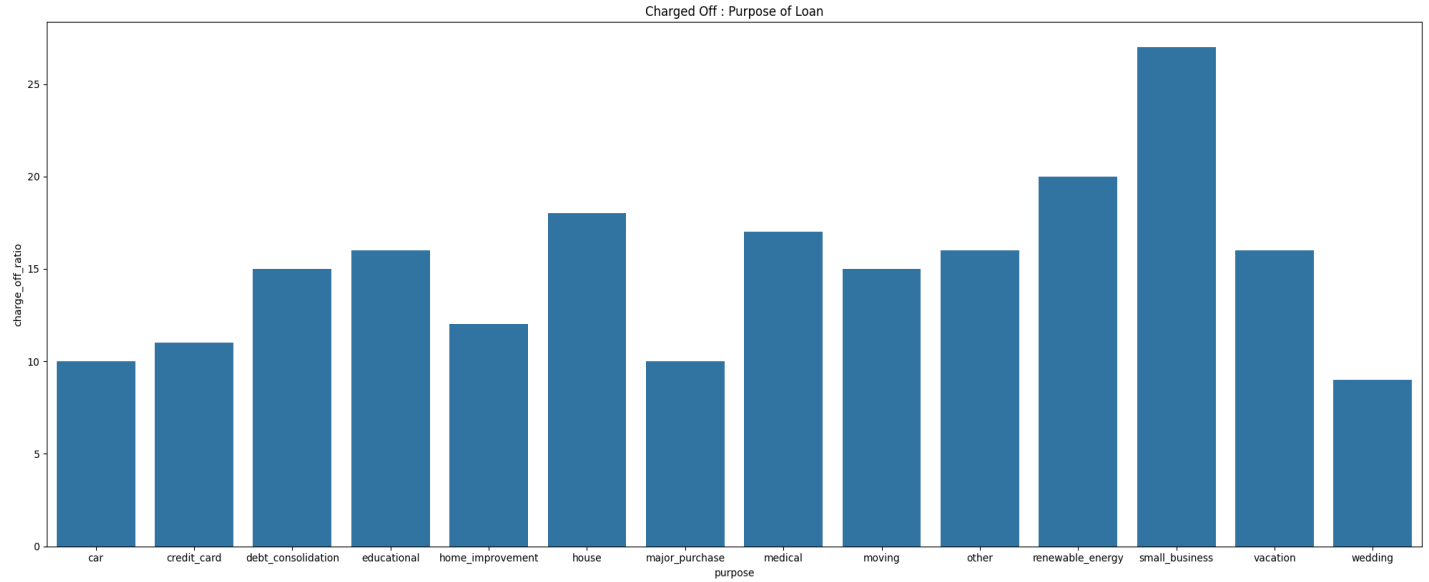


- Overall number of applications that have not been verified, when status is source verified or LC verified, the highest number is in the Status of Not Verified Status
- When Status of Source Verified and LC Verified is combined as a single Verified Status, then number of charged off customers are higher in Verified status
- However, the number of charged off customer in terms of charged off ratio are not significantly impacting with the status of being Verified or Not Verified.

Bivariate Analysis – Loan Purpose Vs Loan Status





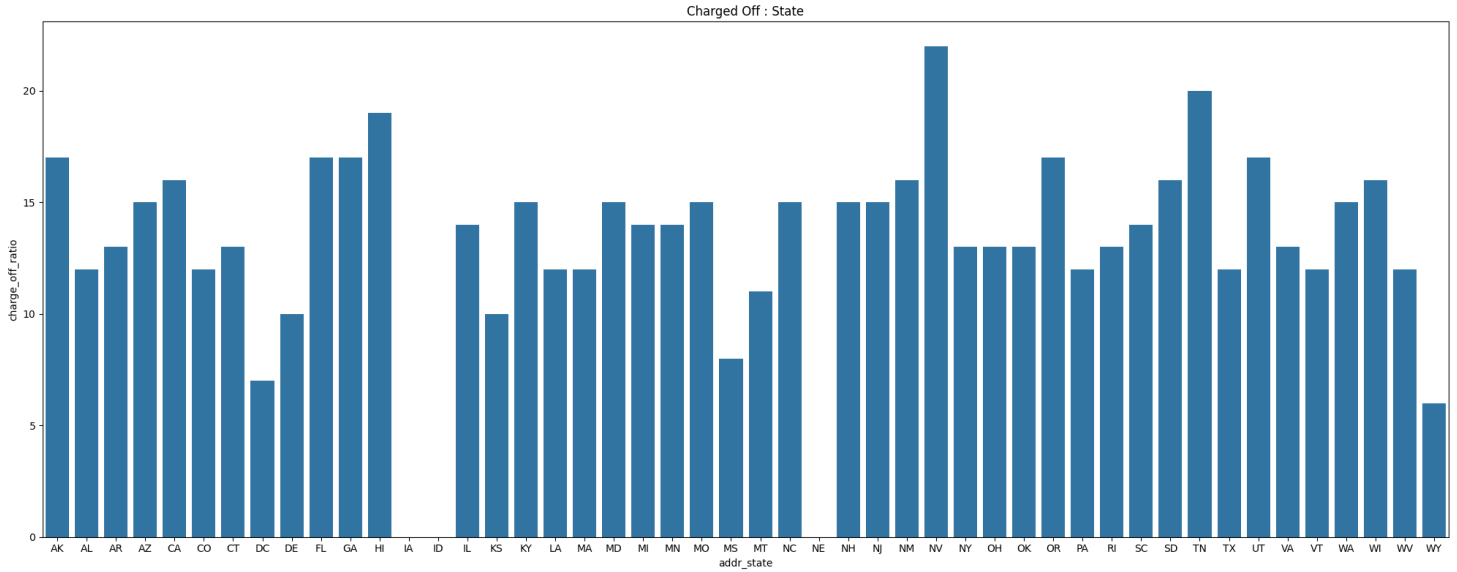


- Highest risk of Charge Offs are the category of debt consolidation
- Highest probability of Charge Offs within a category are small business but the volume is extremely low
- Highest loan amount ranges are in small business, debt consolidation and house

Inferences

- Highest risk of Charge Off's are the purpose of debt consolidation
- Small Business applicants have high chances of getting charged off.
- Renewable energy has lowest risk of Charge Off's in volume

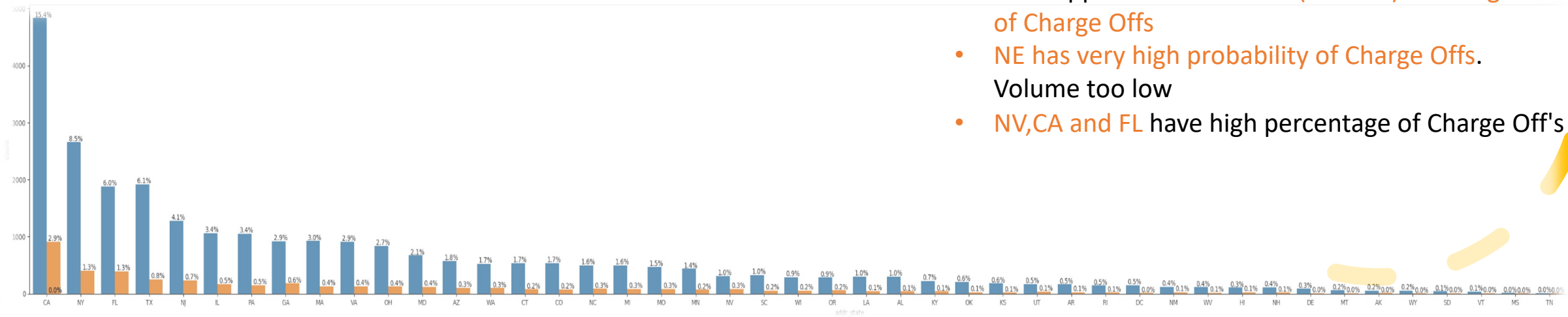
Bivariate Analysis – Address State Vs Loan Status



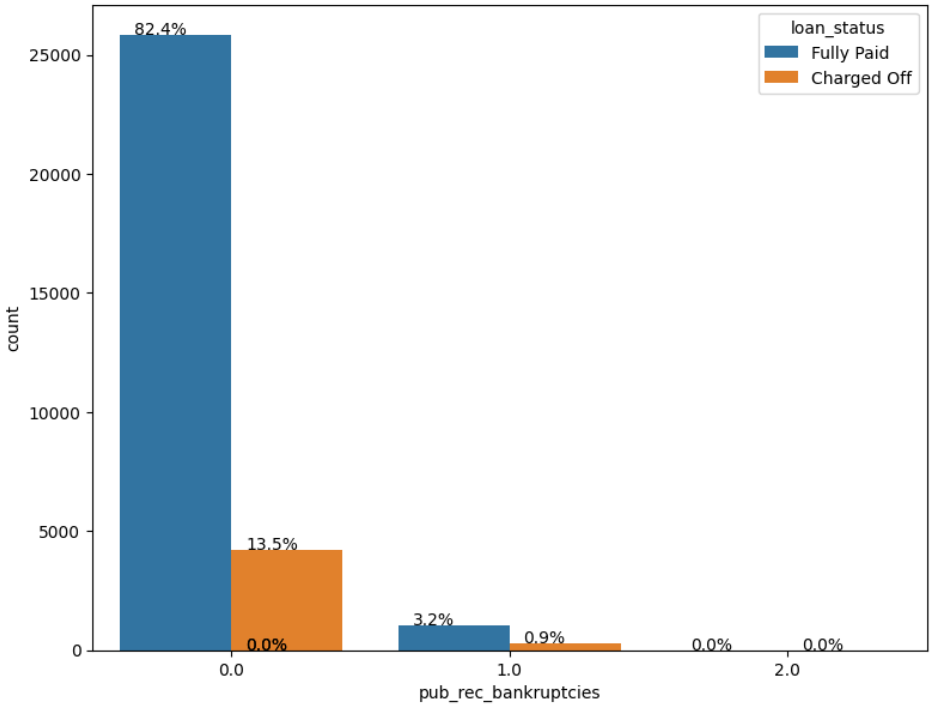
- Highest volume of loans is from CA and purely based on volumes the highest Charge Off's are from CA
- Within each state NE and NV has the highest Charge Offs
- NE has very low volume this cannot be considered
- Loan applications from NV will have high risk

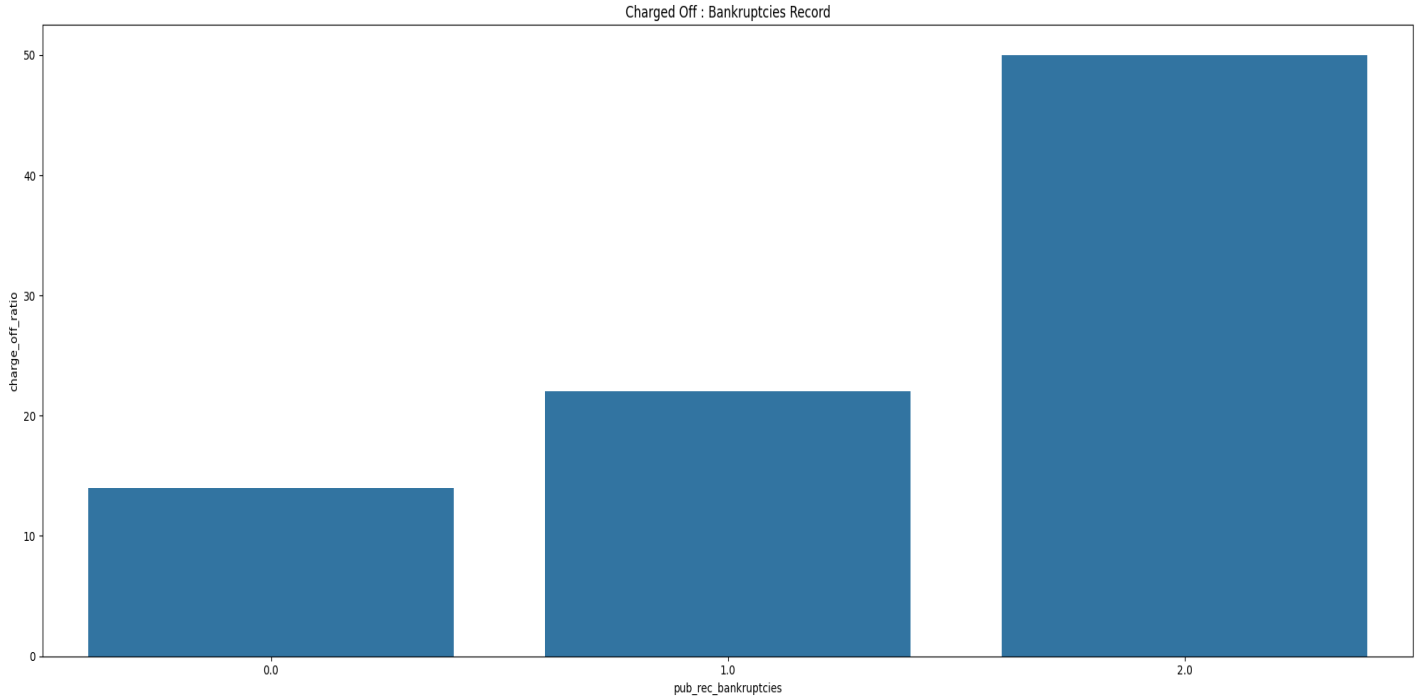
Inferences

- Loan applications from NV (Nevada) have high risk of Charge Offs
- NE has very high probability of Charge Offs. Volume too low
- NV,CA and FL have high percentage of Charge Off's



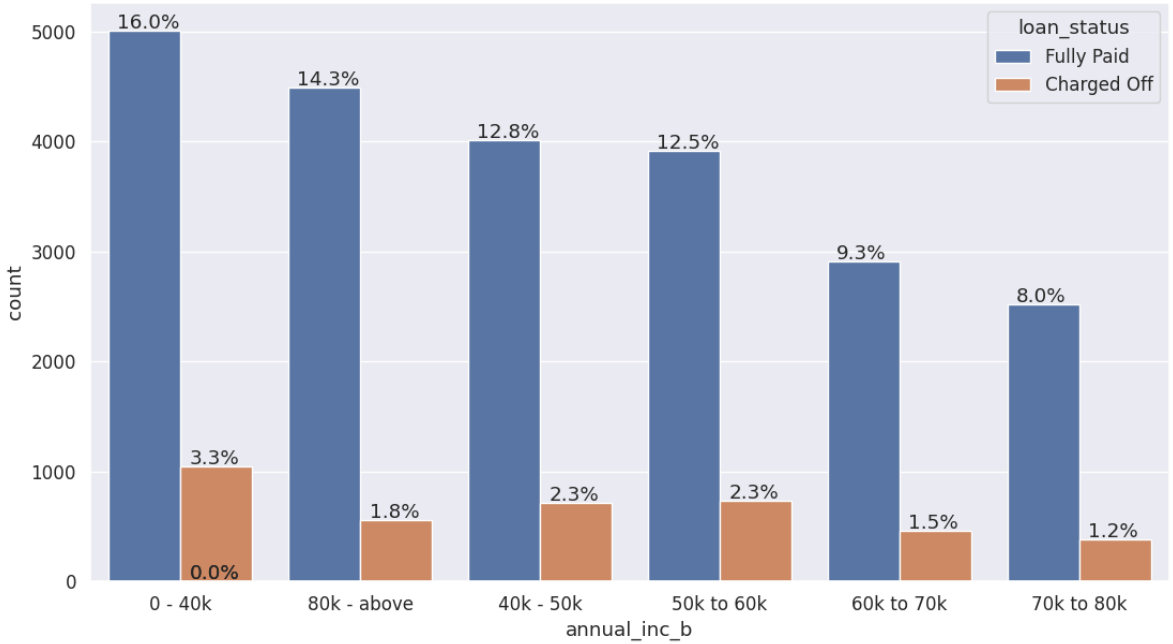
Bivariate Analysis – Public Bankruptcy Record Vs Loan Status

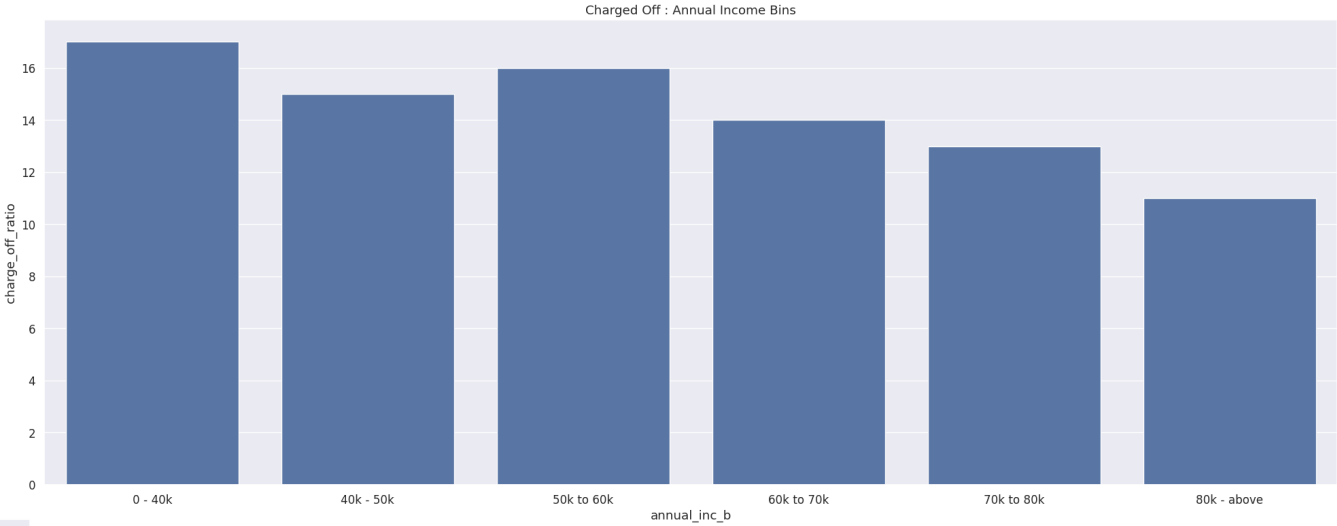


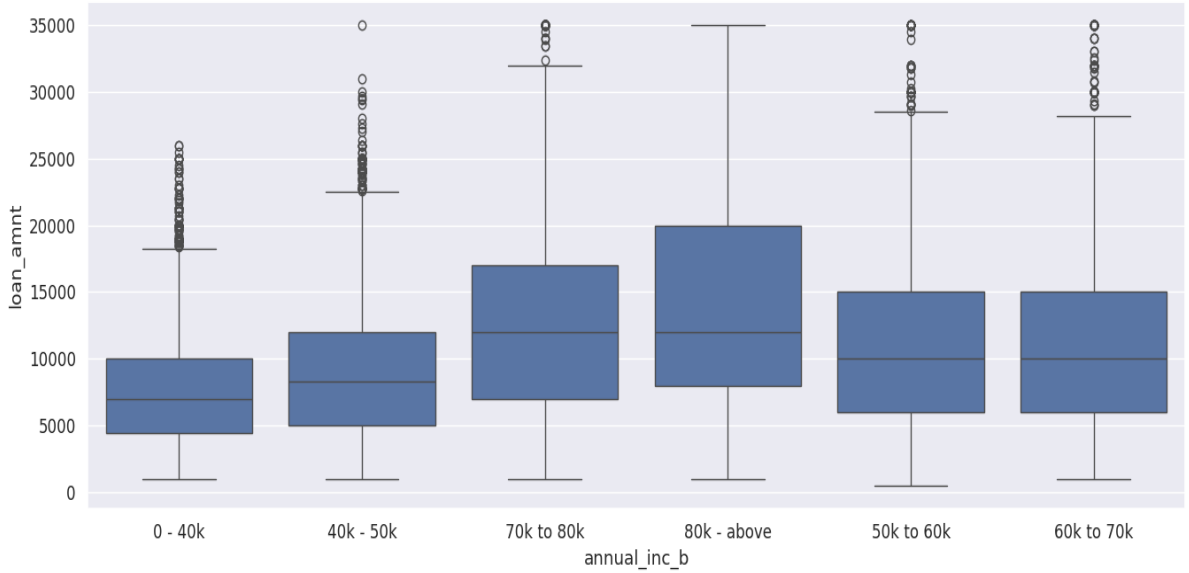


- The high volume of loans are in the category of no Bankruptcies
- The overall percentage ratio of Charge Off's is slightly higher with one bankruptcy (22%) or more as compared to no bankruptcy (14%)
- Therefore, the lender needs to be very careful with people having bankruptcy record, since they have very high chance of defaulting

Bivariate Analysis –
Annual Income Vs Loan Status







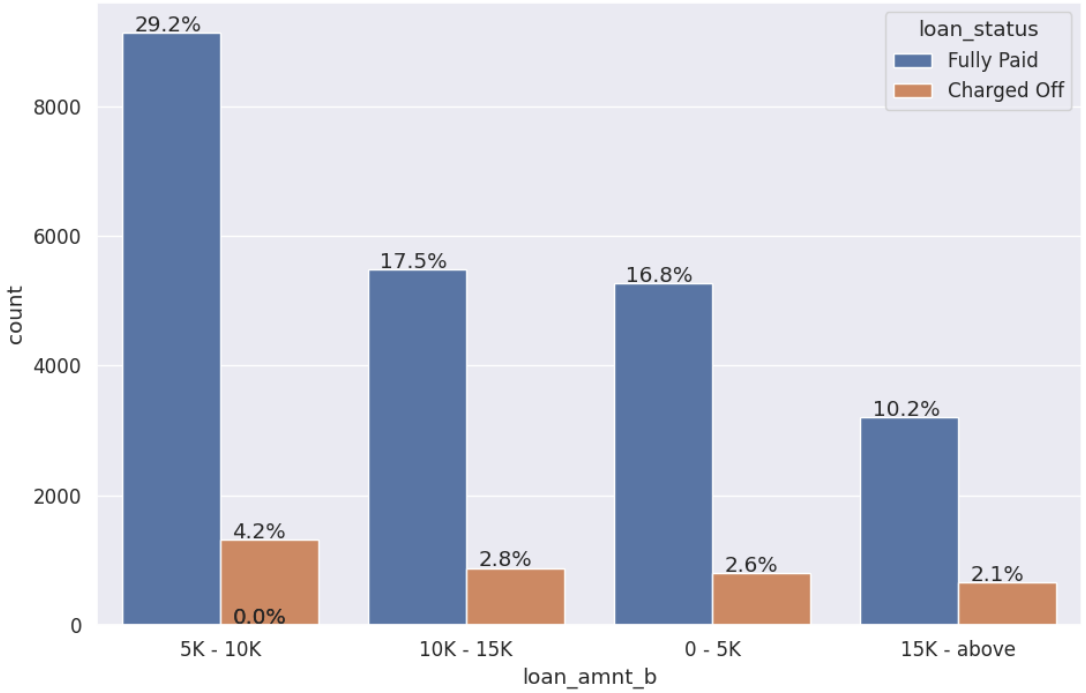
- Highest number of loans are disbursed to applicants with annual income of 40k or below
- The percentage ratio of Charge Off's is higher for applicants with annual income of 0-40K and 50-60K

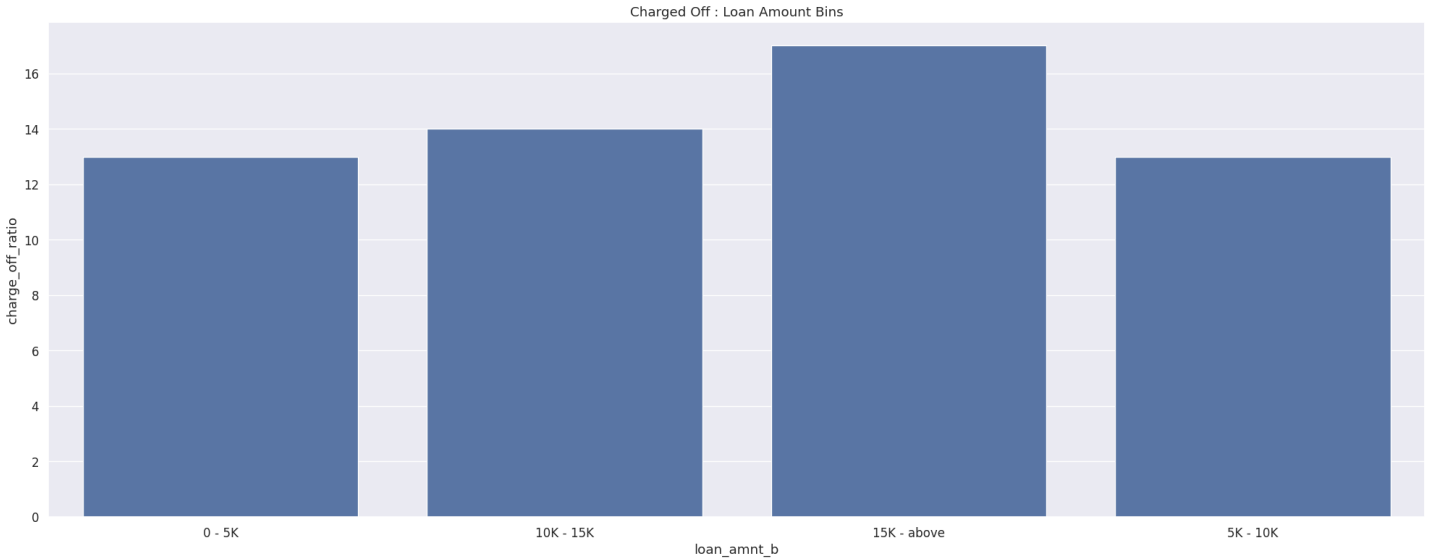
Inference

- The Lending Company should be careful to approve loans for applicants with annual income of less than 40k or even better anybody less than 60k
- Income range 80000+ has less chances of charged off
- Increase in annual income charged off proportion decreases.

** Annual Income data filtered with quantiles range between 5% and 90% to remove outliers*

Bivariate Analysis –
Loan Amount Vs Loan Status





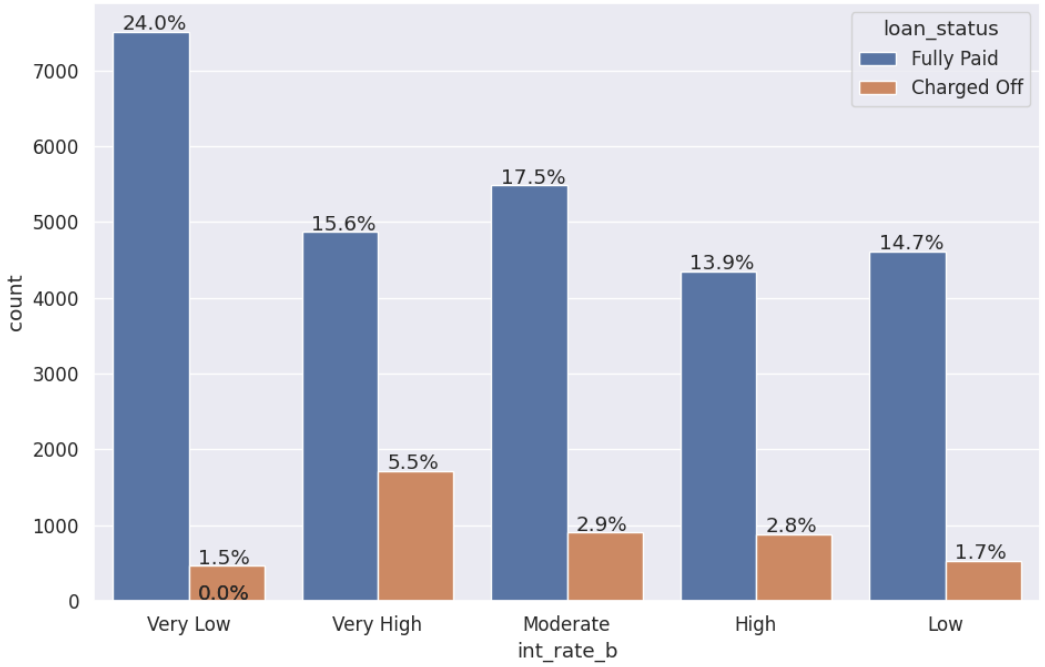
- Based on volume highest percentage of Charge Offs are in the category of 5K to 10k of loan amount
- The Charge Off ratio of all the customers within the loan amount of 15K and above is at the highest Charge Off risk

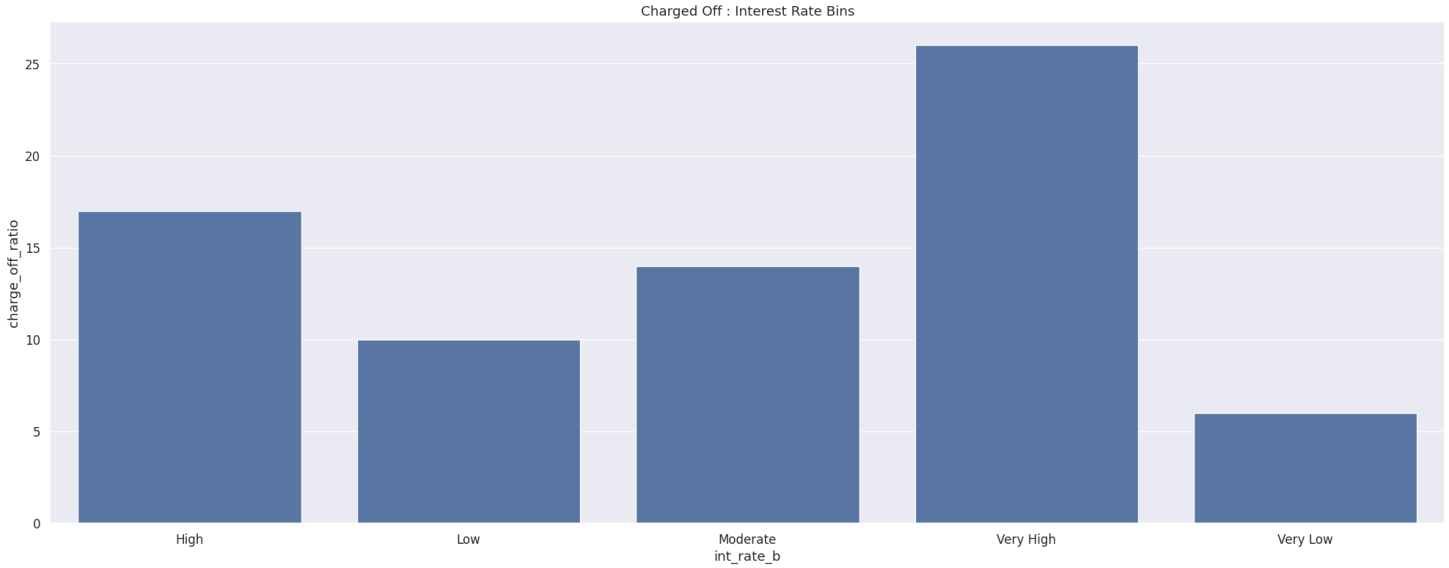
Inferences

- Charge Off risk of loan amount 15K and above is at the highest risk

** Loan Amount data filtered with quantiles range between 5% and 90% to remove outliers*

Bivariate Analysis –
Interest Rate Vs Loan Status





- Based on volume highest percentage of Charge Offs are in the category of Very High Interest Rate
- The Charge Off ratio of all the customers within the interest rates of Very high is at the highest Charge Off risk

Inferences

- Charge Off risk of Very High Interest Rate of 15% and above is at the highest risk

Correlation Analysis

Negative Correlation

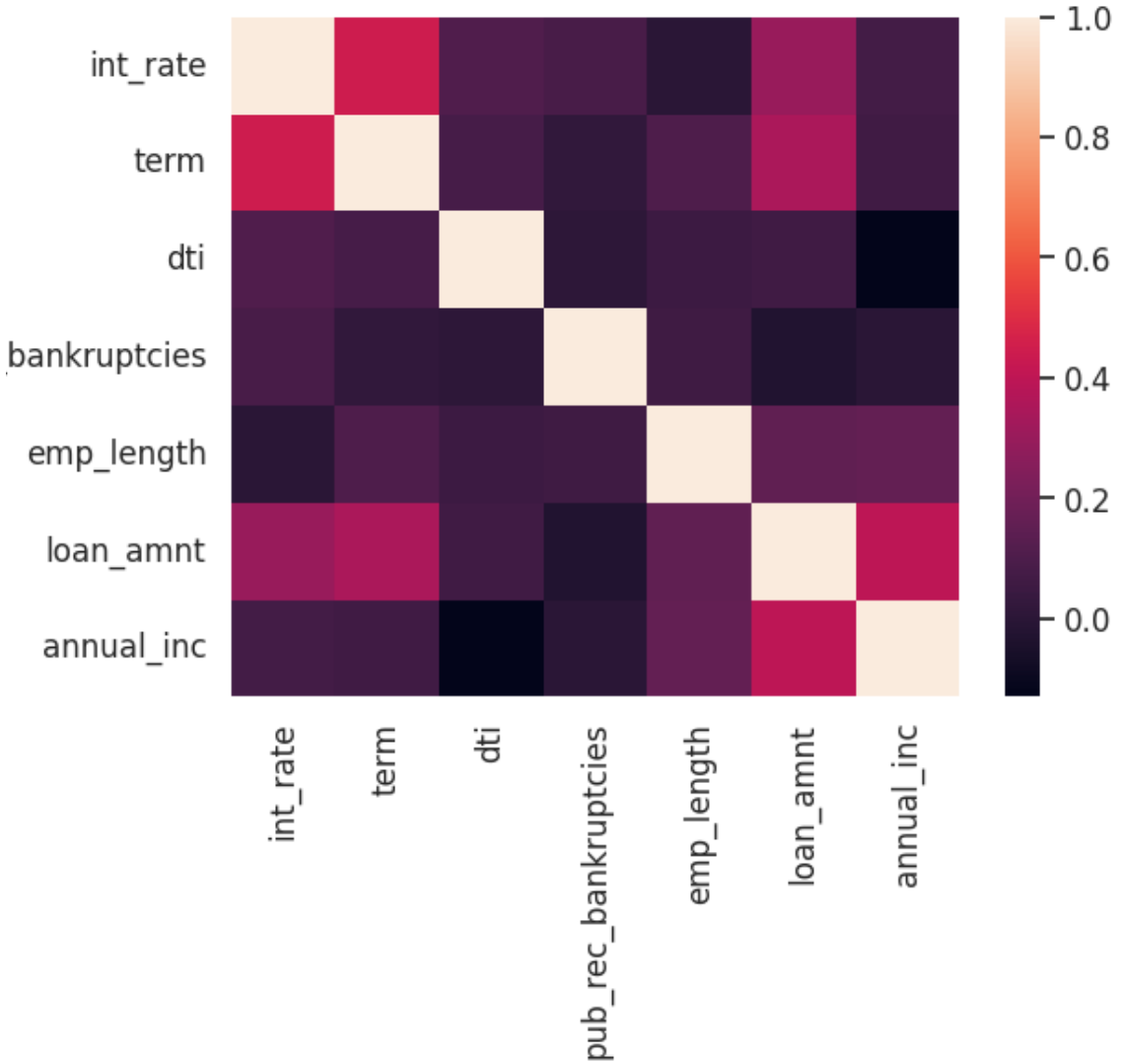
- Loan amount has negative correlation with those with public record of bankruptcies
- annual income has a negative correlation with dti

Strong Correlation

- term has a strong correlation with loan amount
- term has a strong correlation with interest rate
- annual income has a strong correlation with loan_amount

Weak Correlation

- public record of bankruptcies has weak correlation with most of the fields



Conclusion Summary

Univariate Analysis Summary - Customer Demographics

- Majority of the loan applicants are in the range of 40000- 75000 annual income
- Majority of the debt-to-income ratio is in the range of 8% to 17% indicating low (0-15% to moderate DTI (15%-20%) which suggests strong to acceptable financial position and lower risk of defaulting on debt obligations
- Majority of the homeowner status are in status of RENT and MORTGAGE
- Highest loan applications are in the category of debt consolidation
- CA (California) state has the maximum amount of loan applications
- Majority of the loan applicants are in the category of not having any public record of bankruptcies, which indicates positive financial stability, good credit history and low risk
- Majority of the employment length of the customers are 10+ years and then in the range of 0-2 years

Univariate Analysis Summary – Loan Demographics

- Highest loan amount applications fall in the range of 5k to 10k
- Majority of the interest rate is in the range of 8% to 15%
- Majority of the installment amount is in the range of 180 to 400
- Majority of the loan applications counts are in the term of 36 months
- Majority of loan application counts fall under the category of Grade B

Univariate Analysis Summary – Time Based Analysis

- Loan application counts are increasing year over year
- Highest loan application volume in Quarter 4 of every year
- Lowest loan applications are in Q1
- Possibly because by year ends people face the financial challenges
- Possibly because of festive seasons
- Possibly because they are consolidating debt by year end

Univariate Analysis Summary – Inferences

- The customer demographic data shows which segment of customers to target for highest volume of loan and not default
- Indicates more analysis is needed why other categories are not as high as other few
- Indicates the Lending Club to be prepared to handle high volume in Q4
- Indicates the Lending Club to target customers in other quarters to increase sales

Bivariate Analysis Summary – Inferences

Lending Club Company need to be aware of the following driver variables about:

1. Term Limit

Individuals seeking 60-month loan terms require more rigorous evaluation due to the increased default risk (25%) compared to those with shorter 36-month terms.

2. Grades

- Charge-off volume is highest for grades B and C.
- However, grades F and G have the highest charge-off ratios.
- Probability of charge-off increases steadily from A to G.

Bivariate Analysis Summary – Inferences

3. Applicants Homeownership

The homeownership status of RENT and MORTGAGE are at the highest risk of Charge Offs

4. Verification Status

- Charge-off rate is not significantly impacted by verification status (Verified or Not Verified), though purely in terms of volume Not Verified applicants are higher.
- Focus on other factors to improve lending decisions.
- Investigate potential flaws in the verification process.

Bivariate Analysis Summary – Inferences

5. Loan Purpose

- Debt consolidation has the highest volume of charge-offs.
- Small business loans carry a higher risk of charge-off.
- Renewable energy projects experience the lowest volume of charge-offs.

6. Address State

- Highest volume of loans is from CA and purely based on volumes the highest Charge Off's are from CA
- Loan applications from the state NV are having high risk of defaulting,
- The lending company needs to be cautious about lending to applicants from NV

Bivariate Analysis Summary – Inferences

7. Public Bankruptcy Records

The lender needs to be very careful with people having any bankruptcy record, since they have very high chance of defaulting.

8. Annual Income

- The Lending Company should be careful to approve loans for applicants with annual income of less than 40k or even better anybody less than 60k.
- Income range 80k+ has less chances of charged off
- Increase in annual income charged off proportion decreases.

Bivariate Analysis Summary – Inferences

9. Loan Amount

- For loan amount, the highest volume of charge-offs falls within the $5,000 to $10,000 category.
- However, customers with loan amounts of $15,000 and above have the highest charge-off ratio, indicating a greater risk per loan.

10. Interest Rates

- High interest rates are associated with increased charge-off risk. This could be due to affordability issues or applicants with high need regardless of interest rate.
- Loans with interest rates of 15% and above face the highest charge-off risk.