Lending Club Case Study

BY

Dr. Parul Shah & Dhanshree Vyas

Flow of the Presentation

- Problem statement and the analysis approach briefly.
- Results of univariate, bivariate analysis along with data visualisation.
- Summary on the most important results in the presentation.

Problem Statement

- Lending loans to ‘risky’ applicants is the largest source of financial loss .
- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.
- Identification of such applicants using EDA is the aim of this case study.
- The goal is to identify and understand the **driving factors** behind loan default.

Approach used for EDA Step-wise

- Data Cleaning
  - Remove rows and columns with no data or invalid data
  - Remove columns with constant data or redundant data
  - Remove columns with all unique entries
  - Relabel data and convert them to correct types
- Perform Univariate analysis to get initial insights into data and decide which attributes need further analysis to achieve the goal.
- Perform Bi-variate analysis to establish their relation with risky-loan.
- With the help of data visualisation, derive important insights and identify attributes that affects risk status.

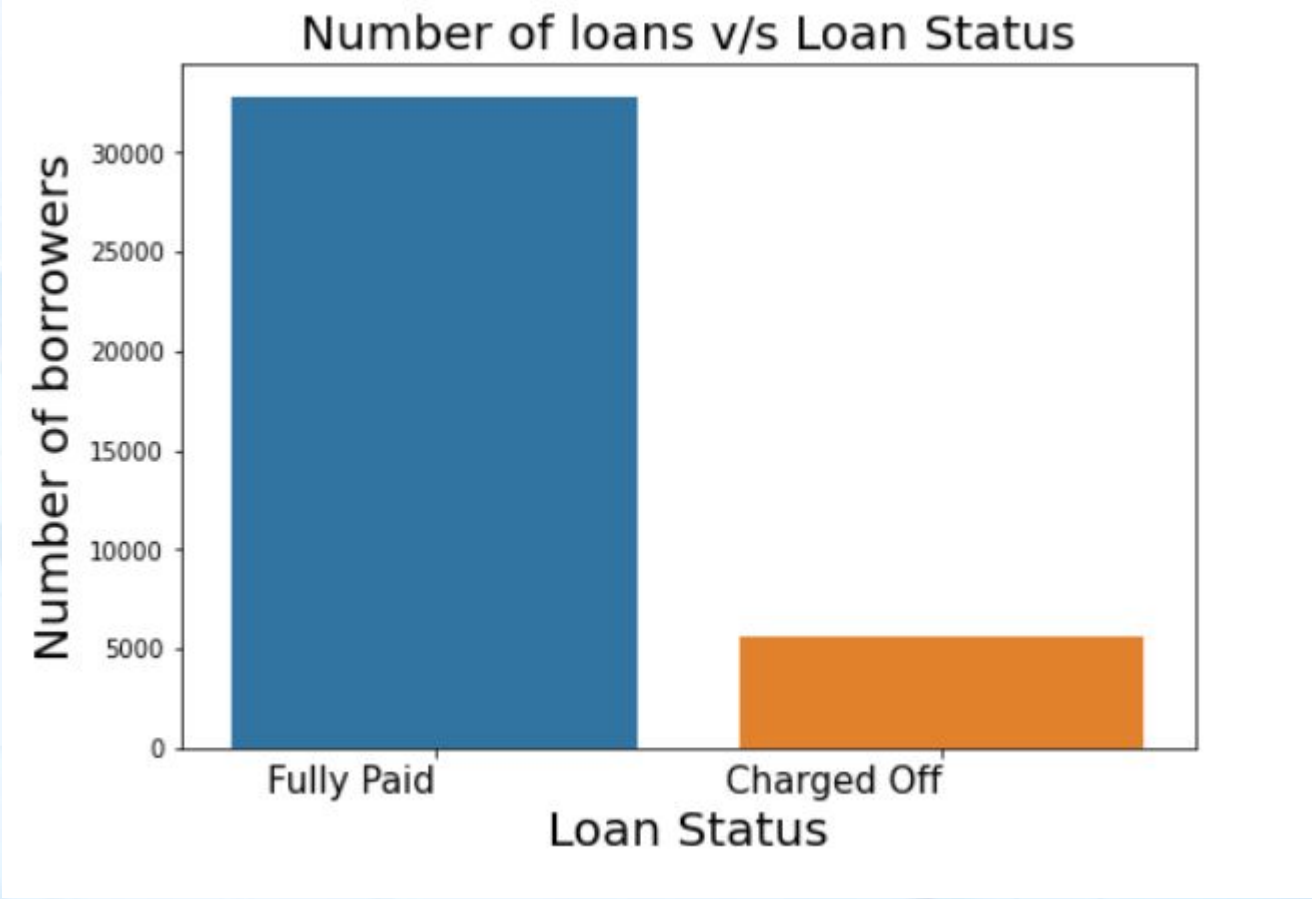
Step 1 : Data Cleaning

- Original data had 111 columns and 39717 rows
- After applying all the data cleaning mentioned, we were left with 39 attributes (columns) and data of 38448 borrowers
- Some assumptions made:
  - Borrower data for whom funded amount bby investor = 0, does not have any information for learning
  - Borrowers with loan status ‘current’ is neither defaulter nor fully-paid, so no learning in tat data too

Step 2: Univariate Analysis

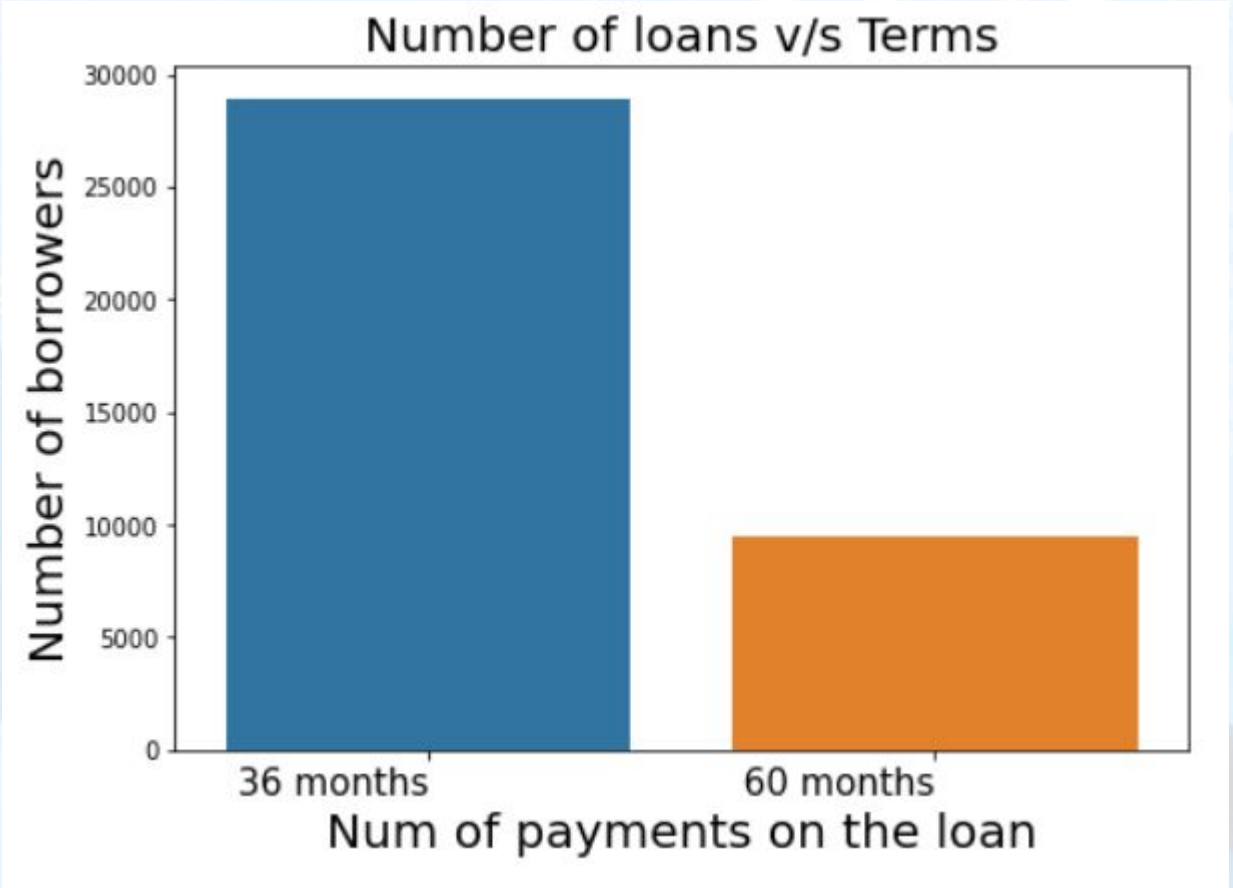
Insights on Loan Status

While there is no major insights here,
it is good to know that number of fully paid loans is significantly higher than num of defaulting loans.

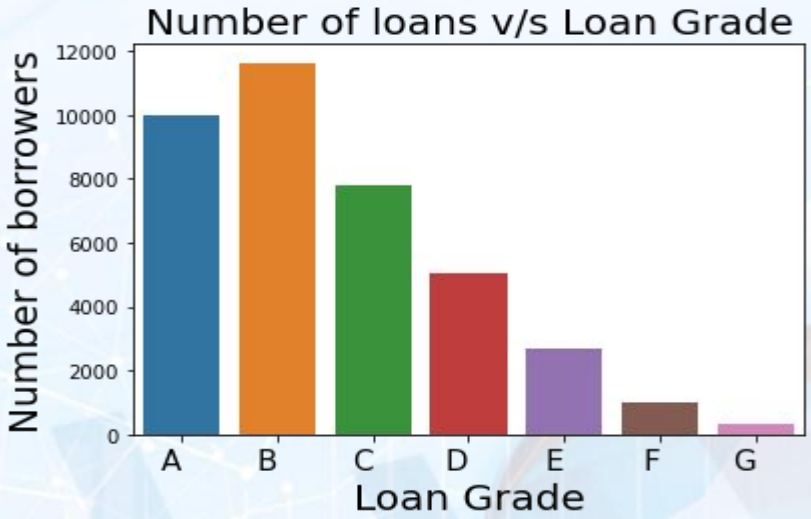


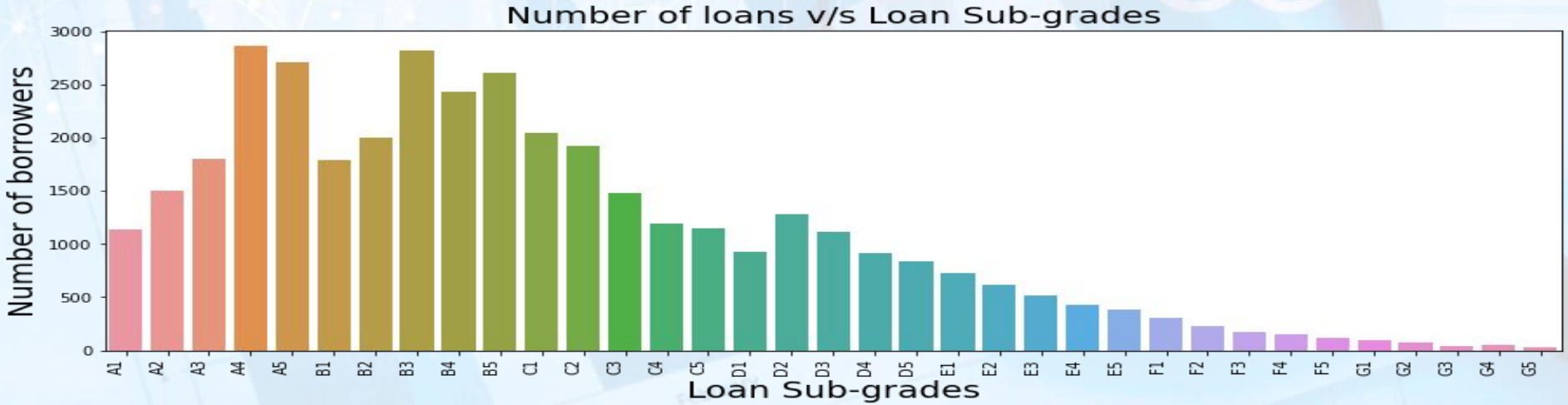
Insights on Number of Payments (Terms)

It is clear that most loans have 36 months term, so we must dig deeper and learn more about Loan status of those who have 36 months terms.



Loan Grades and Sub-Grades



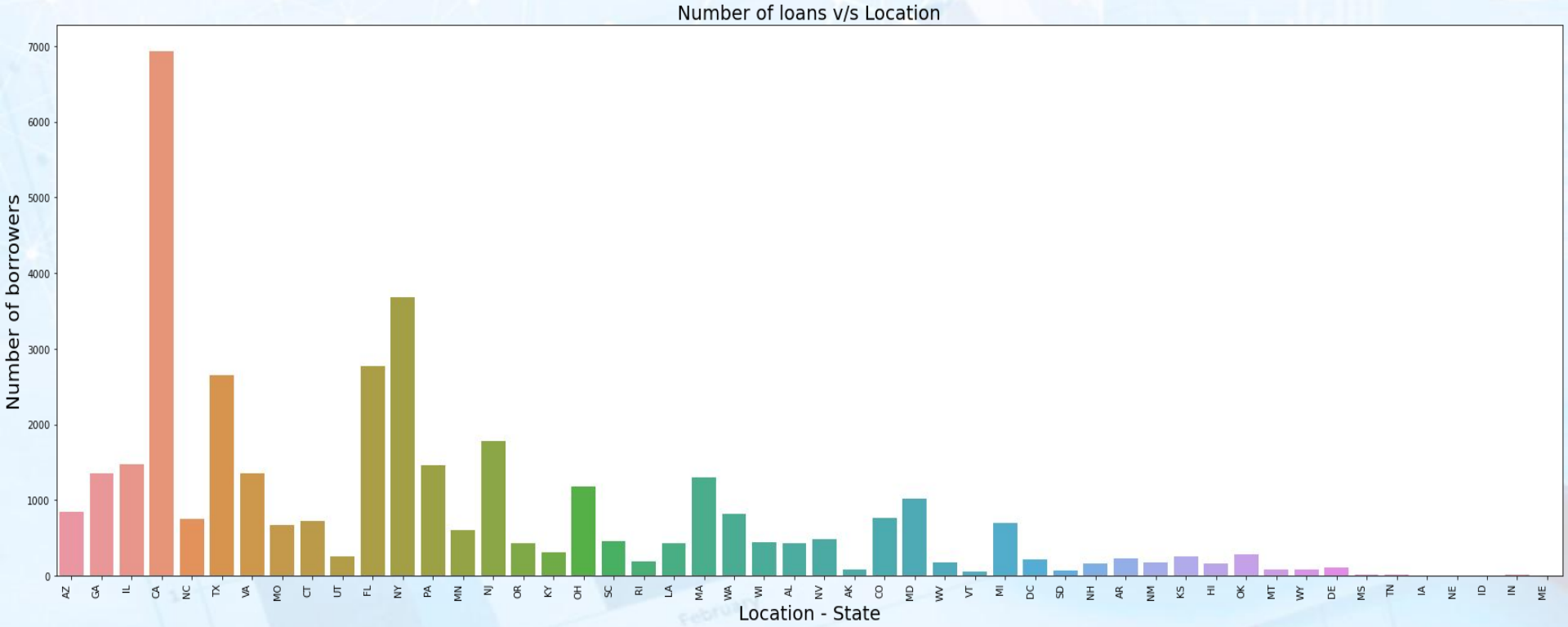


Insights on Loan Grades and Sub-Grades

- **Loan Grade:** Almost 1/3rd of total loans are of grade B and 1/4th are of Grade A, 1/5t is of Grade C. So together, Grade A, B, C and D is covering majority loans. We need to analyse them further to ensure these after all 'low-risk' loans.
- **Loan Sub-Grade:** This is in sync with Grade. We need to dig deeper, especially for sub-grade A4, A5, B3, B4, B5 and ensure these are 'low-risk' loans.

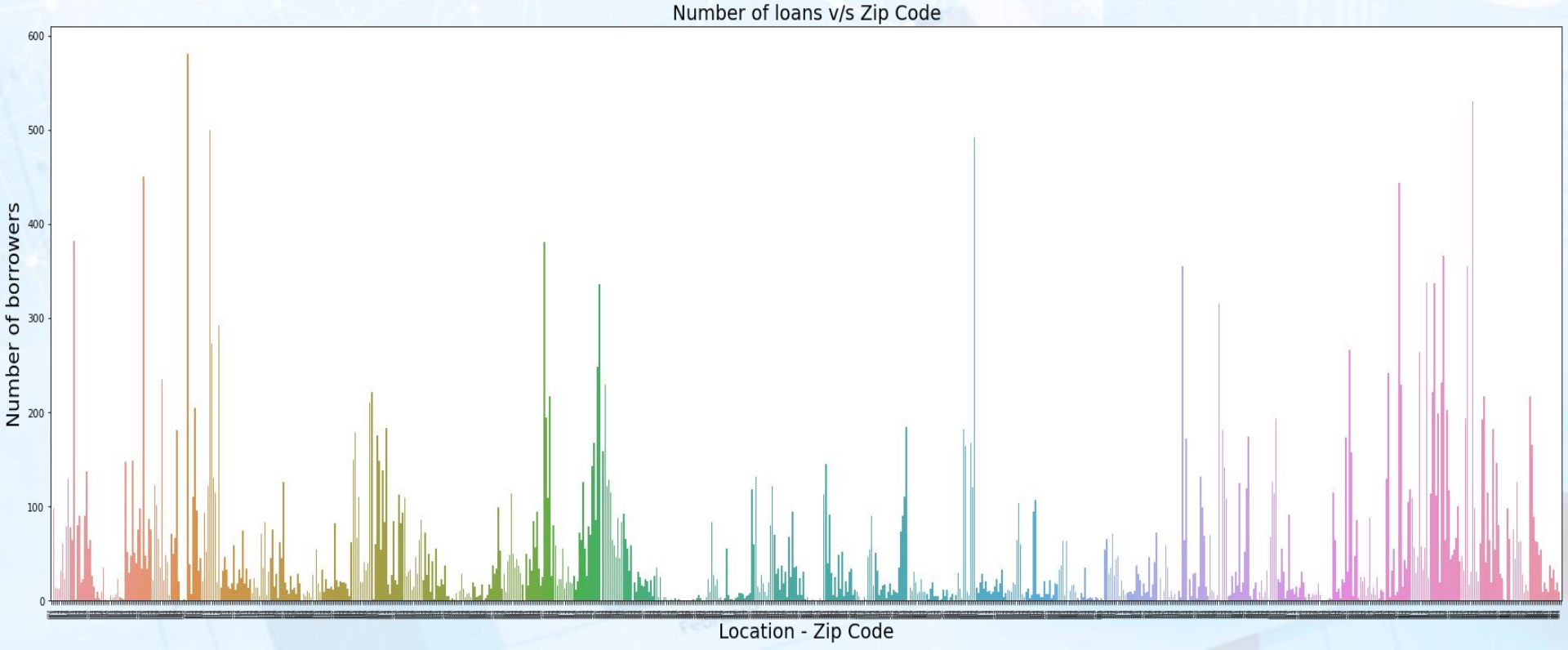
Insights on Location

- **State:** Some of the states significantly larger number of loans compared to others and hence need to study them further. e.g CA, NY, FL, TX.

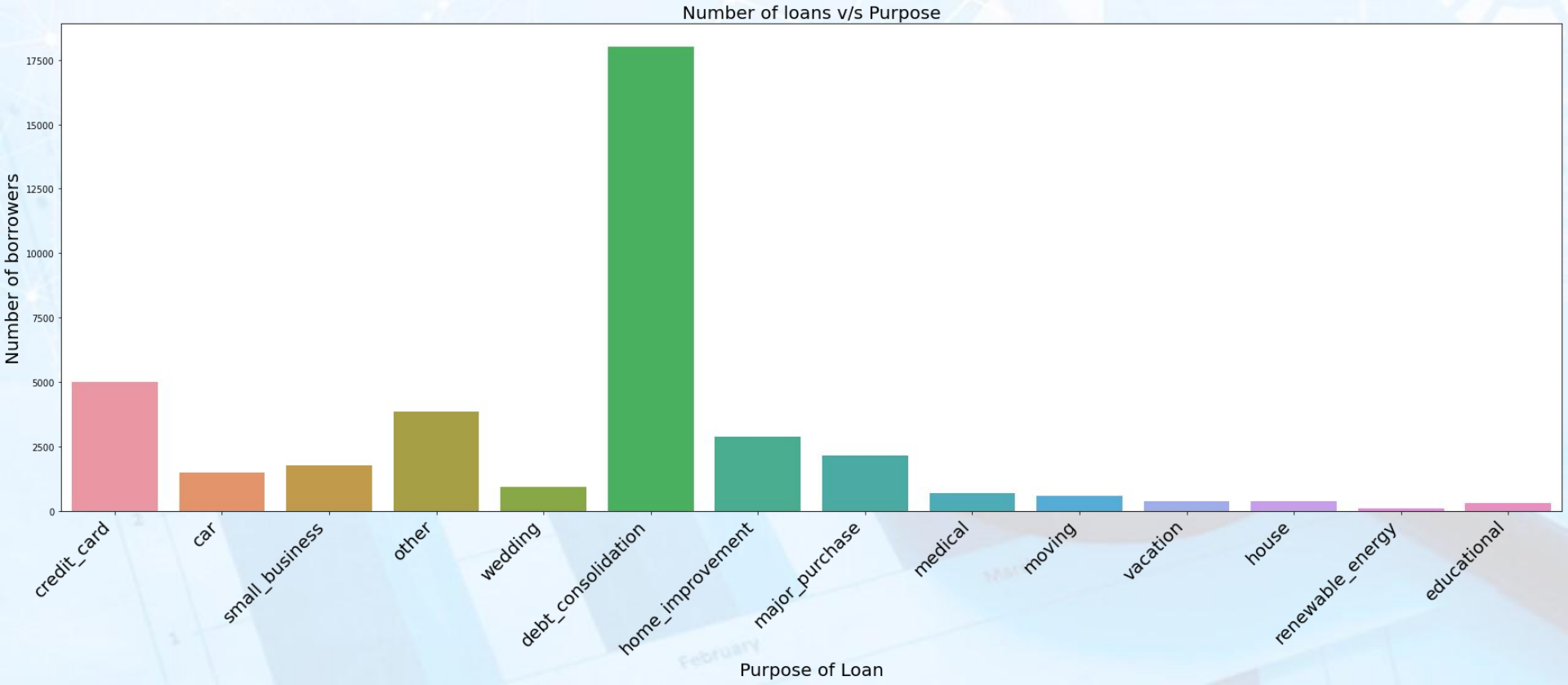


Insights on Location

- **Zip Code:** Zip also show some tall spikes, those zip codes need further analysis. (Note: It is difficult to read zip code on this graph, but we can find out those spikes by further inspection.



Purpose of taking Loan

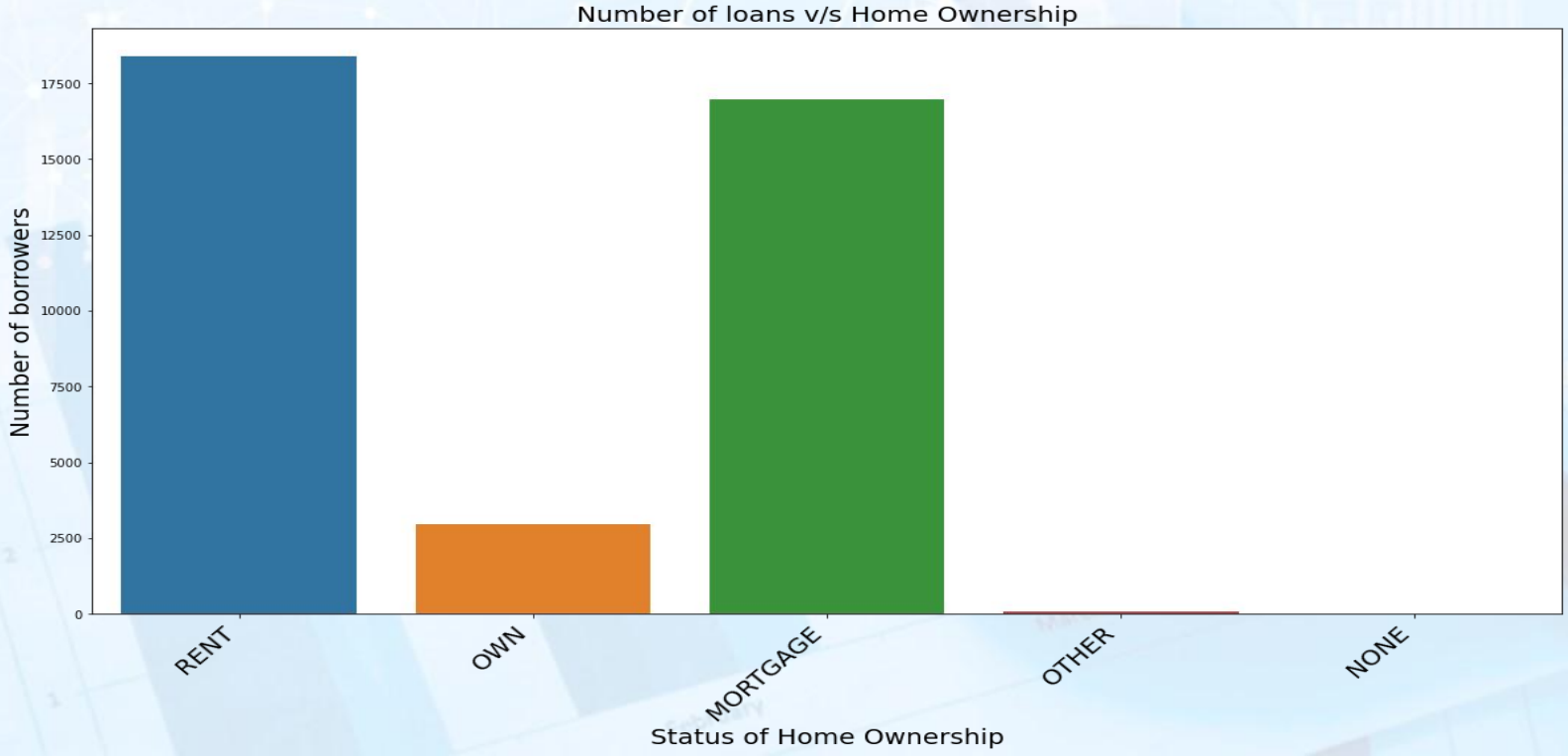


Insights on Purpose of taking Loan

- A huge number of loan is taken for debt consolidation and credit card payments. Both these category fall into not-secured loans and hence it is very important to know that these loans are 'low-risk'. This needs further analysis.
- "other” also shows good number of loans, as there is no more information about theses loans, maybe we can check their grades and sub-grades to understand them better.

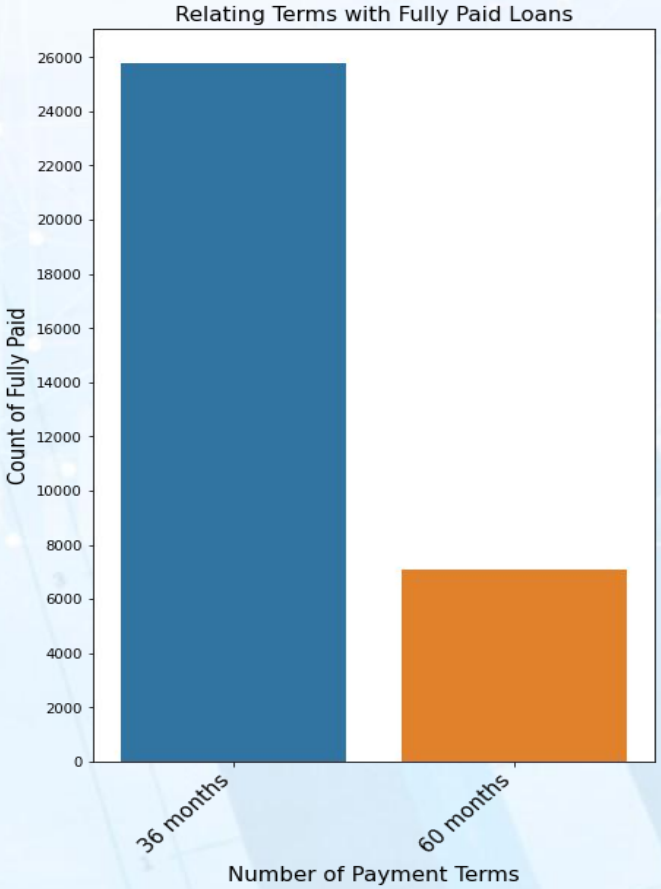
Insights on Home ownership status

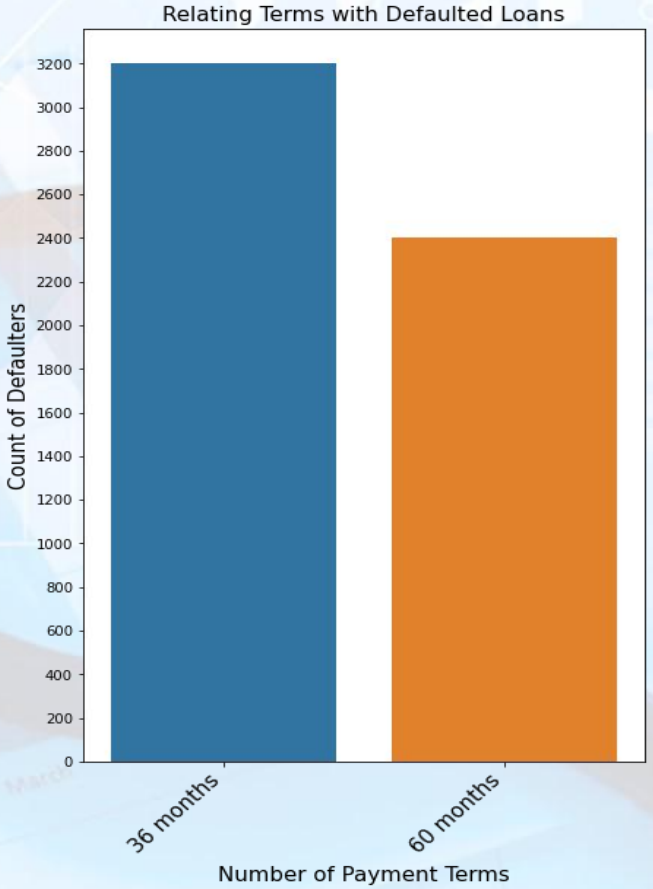
- Most loans are not-secured as a huge majority given to borrowers who either do not own a home or have already mortgaged it. NOT a good scenario and definitely need further investigations.



Step 3: Bivariate Analysis

Insights on Relation of Terms with Loans

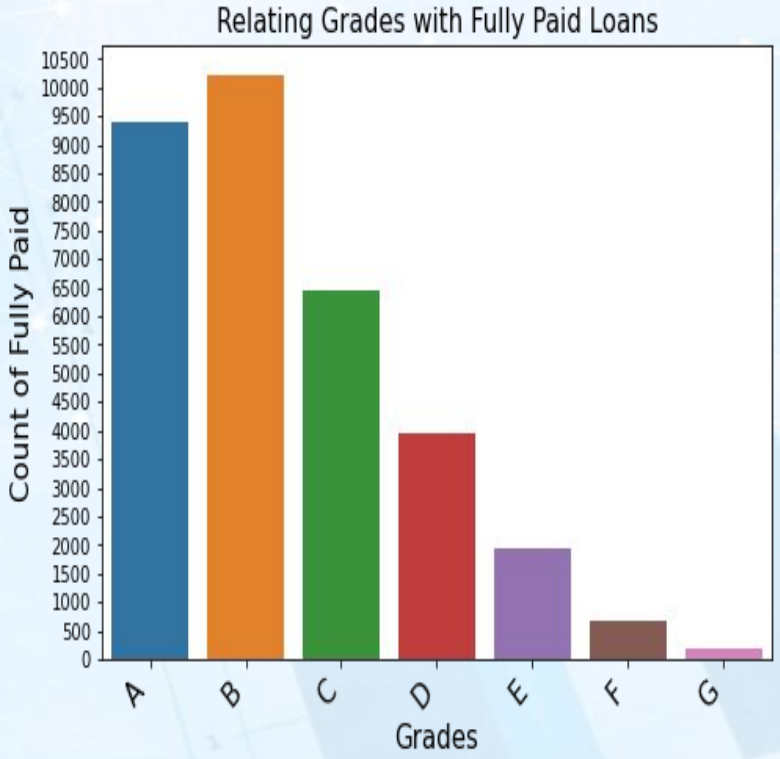


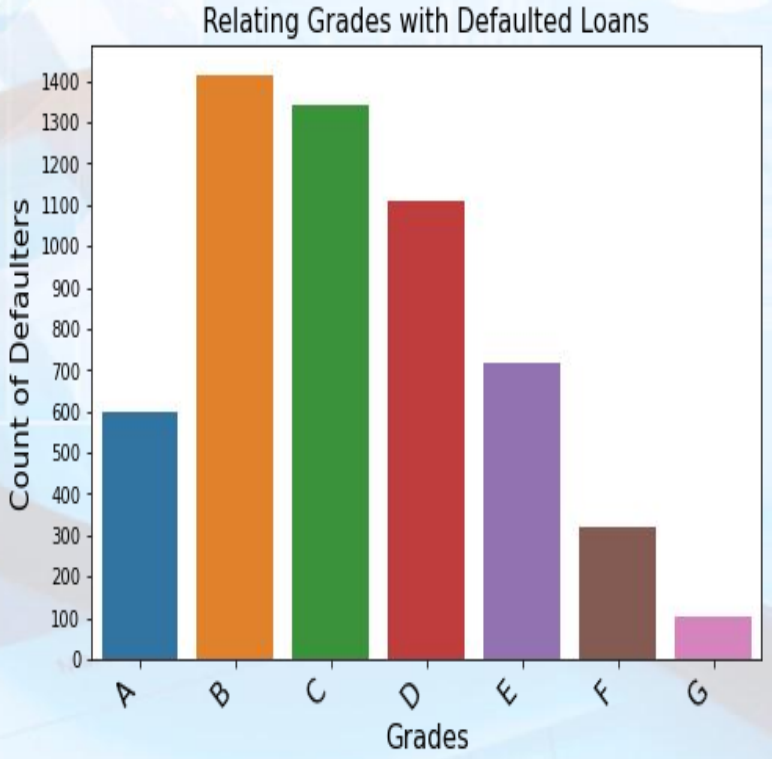


Insights on Relation of Terms with Loans

- **Fully Paid Loans:** 36 months term loans are more likely to be fully paid compare to 60 months term loan.
- **Defaulted Loans:** 36 months term loans more likely to be defaulted but 60 months term loan is also having high chances of getting defaulted.

Insights on relation of Grades with Loan status



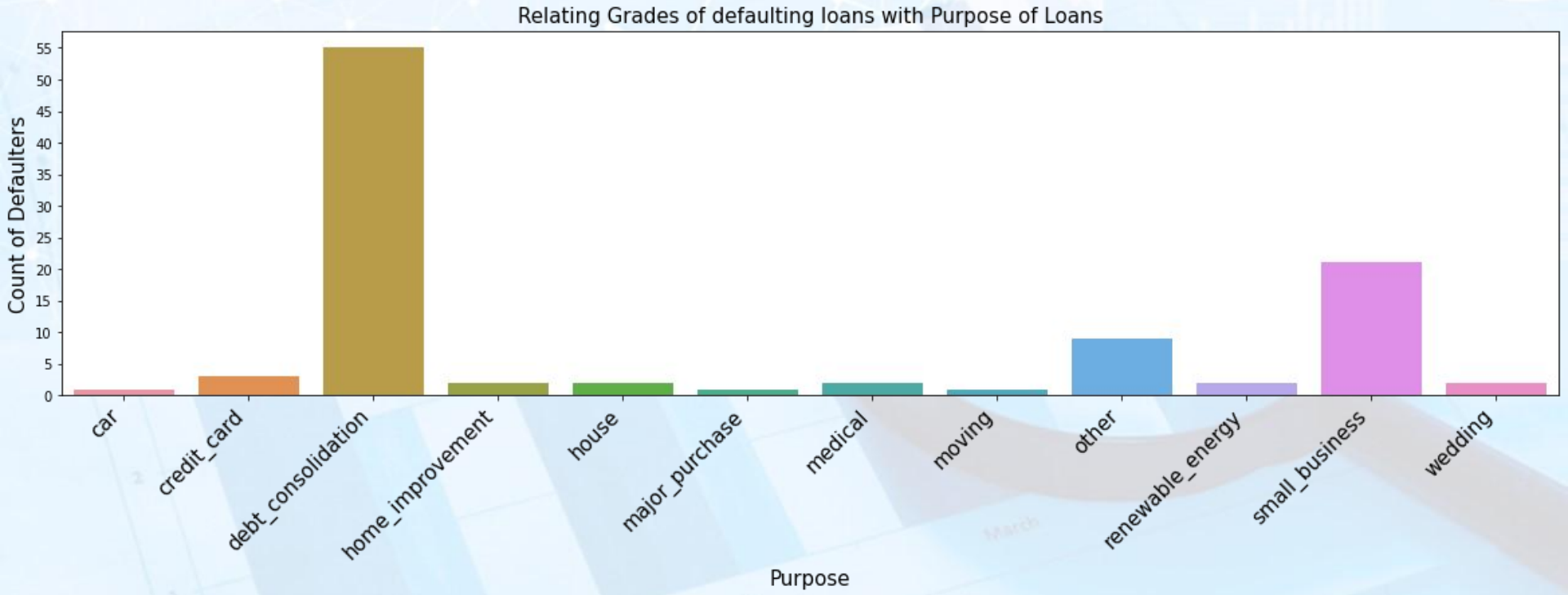


Insights on relation of Grades with Loan status

- First bar chart indicates that Grade A & B makes almost 50% of total fully paid loans. This is evident in the box plot too. If we examine second bar chart (of defaulters), Grade A makes less than half of Grade B and C. This shows that Grade A is one of the most safest and 'low-risk' loan.
- From bar charts we can see that ratio of defaulters to fully paid in grade B is approximately 0.14. Box plots shows that 25% of all defaulters are under Grade B. This makes Grade B also relatively a low-risk loan.
- 75% of fully paid loans are under Grade A, B & C, so Grade C is a moderate-risk loan
- Above Grade C is what needs more attention.

- **Box plot shows, maximum Grade of Fully paid is Grade F (not considering the outliers), which makes Grade G a very high-risk loan, almost sure shot defaulters.**

Insights on relation of Grades G with Purpose

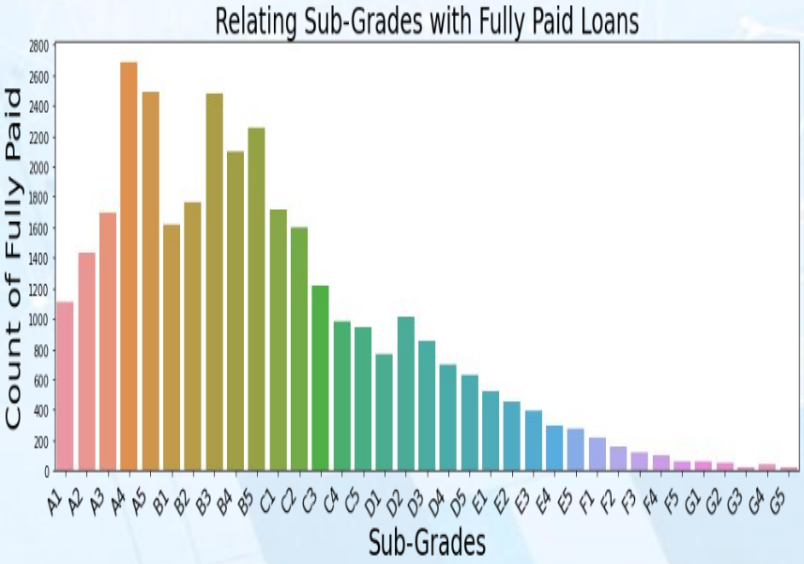


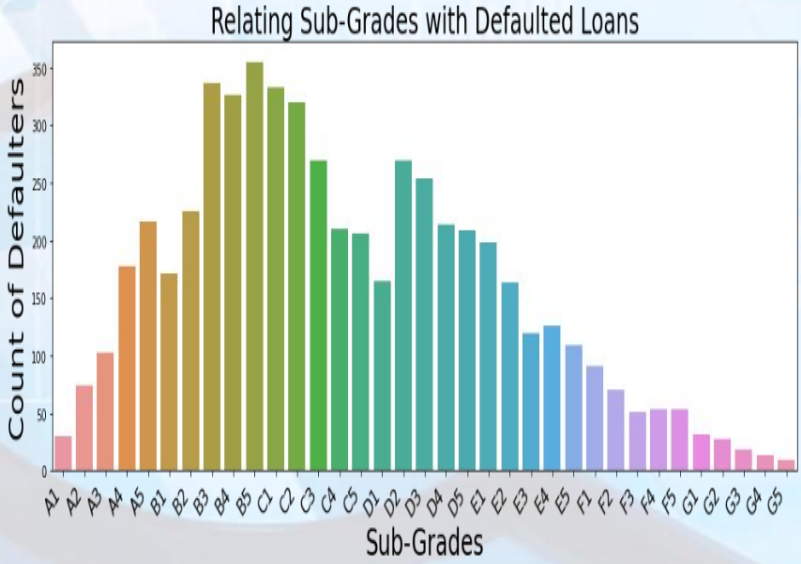
Insights on relation of Grades G with Purpose

- The bar chart clearly indicates that **debt_consolidation loans are high-risk loans, more likely to default.**

- Small_business and 'other' too are relatively high-risk and must be looked into properly before sanctioning the loan.

Insights on relation of Sub-Grades with Loan status

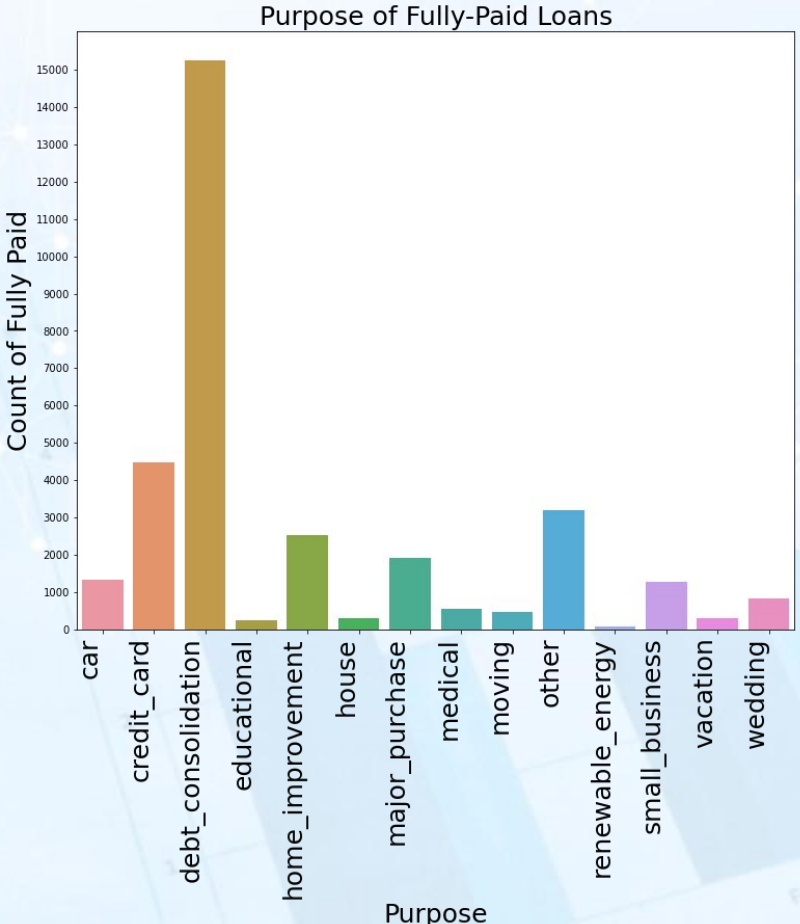


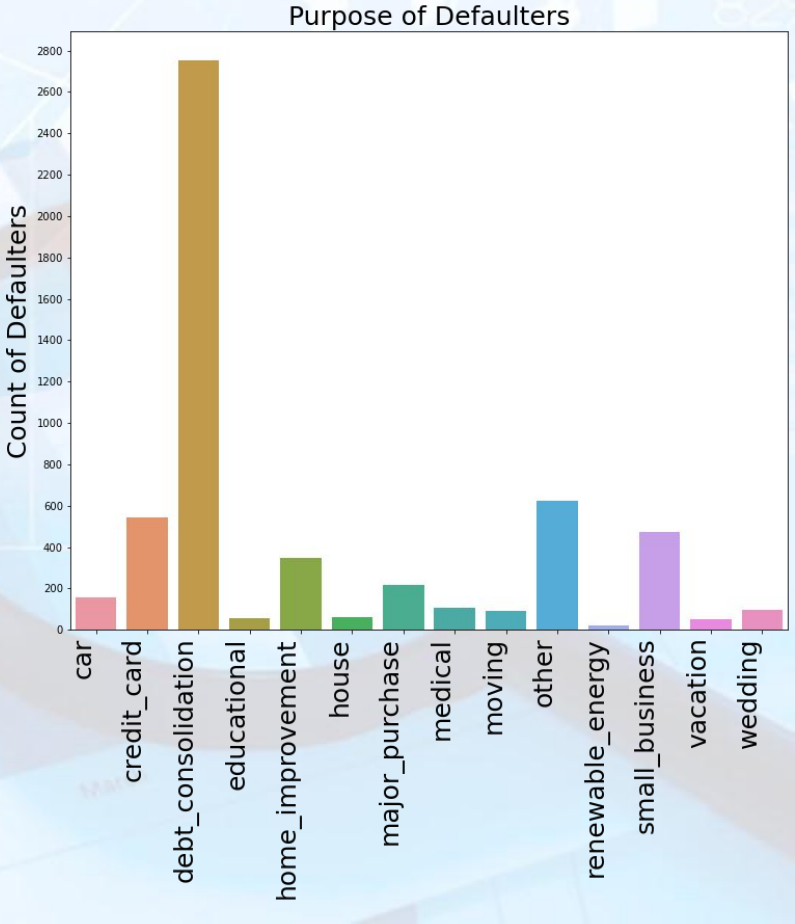


Insights on relation of Sub-Grades with Loan status

- Once again it shows that Sub-Grades of A are less likely to default and hence they are some of the most low-risk grades.
- 75% of fully paid loans are below sub_grade C3 (Rank 13), so these grades are relatively low-risk loans.
- 50% of defaulters are between sub_grade B4 to D4 (rank 8 to 18). A little more scrutiny needed for these grades before sanctioning the loan.
- **Grade F3 (rank 27) and above are highly likely to default.**
- **SUB-Grades of G once again proves to be the riskiest sub-grades.**

Effect of Purpose of Loan on defaulting



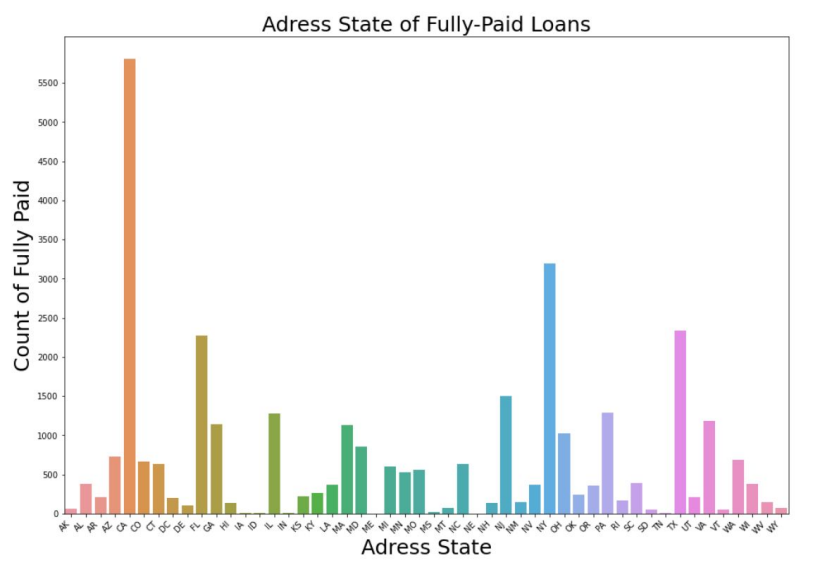


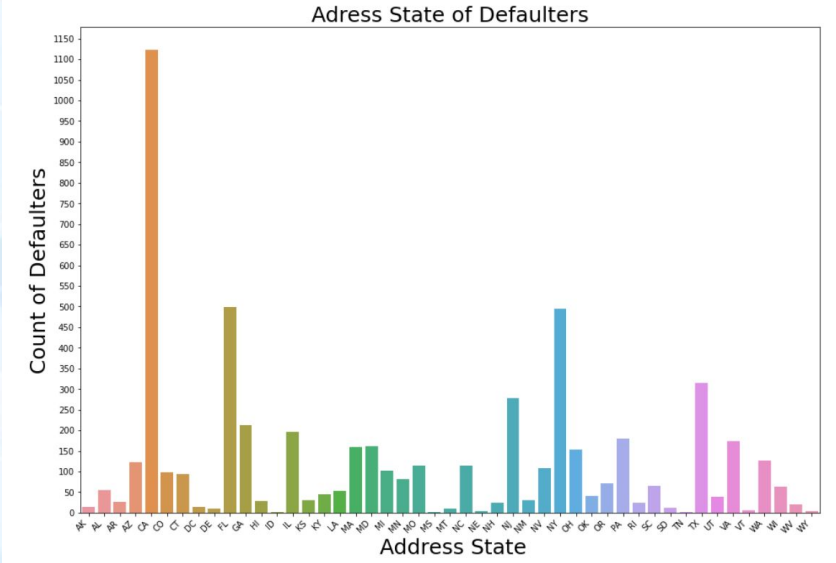
Insights on effect of Purpose of Loan on defaulting

- Once again it shows that **loans taken for debt_consolidation is highly likely to default and hence they are high-risk. Process to sanction them should be tightened.**

- Same should be done for 'Other', 'Credit_card' and 'small_business' categories. The too look like high-risk loans.

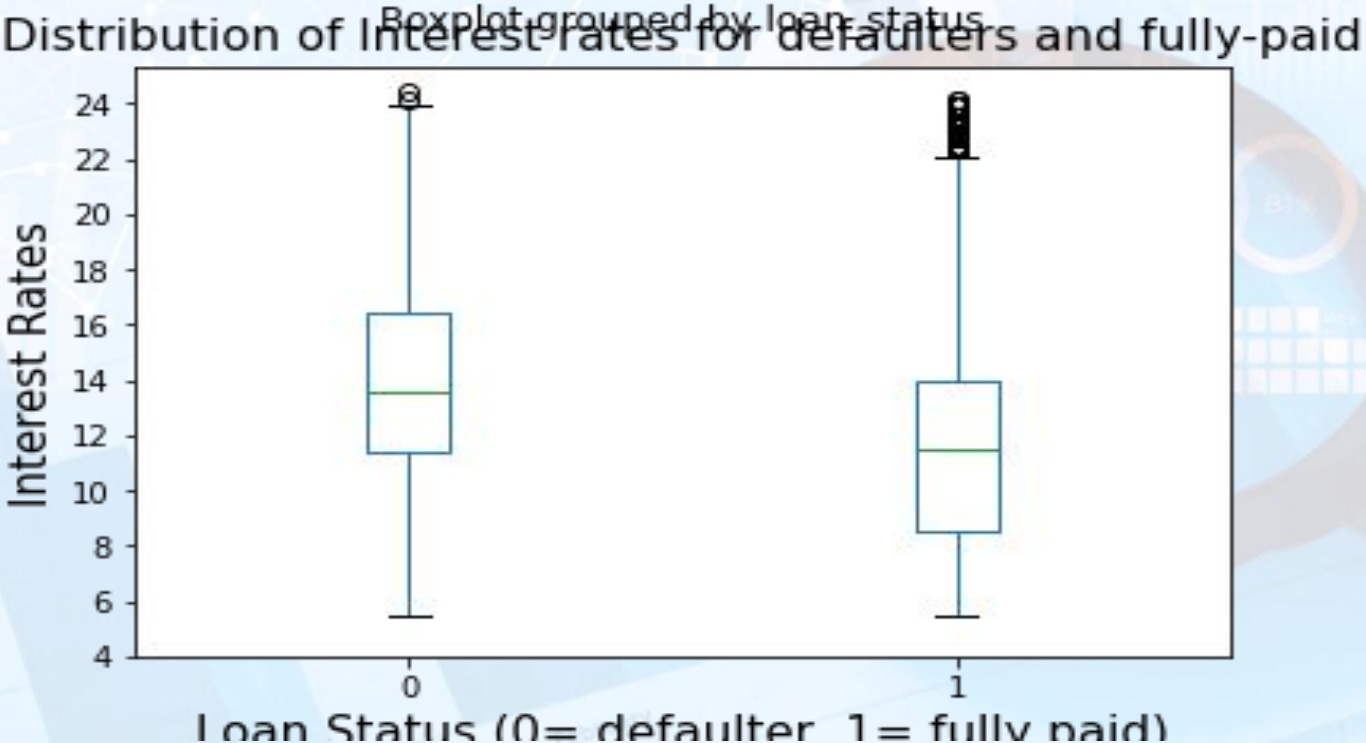
Effect of Address State on Loan Defaulting





- State of CA has huge number of loans and also largest number of defaulters. Which means scrutiny and verification process in CA must be tightened.
- States of FL, NY and TX also has similar concern on slightly lower scale

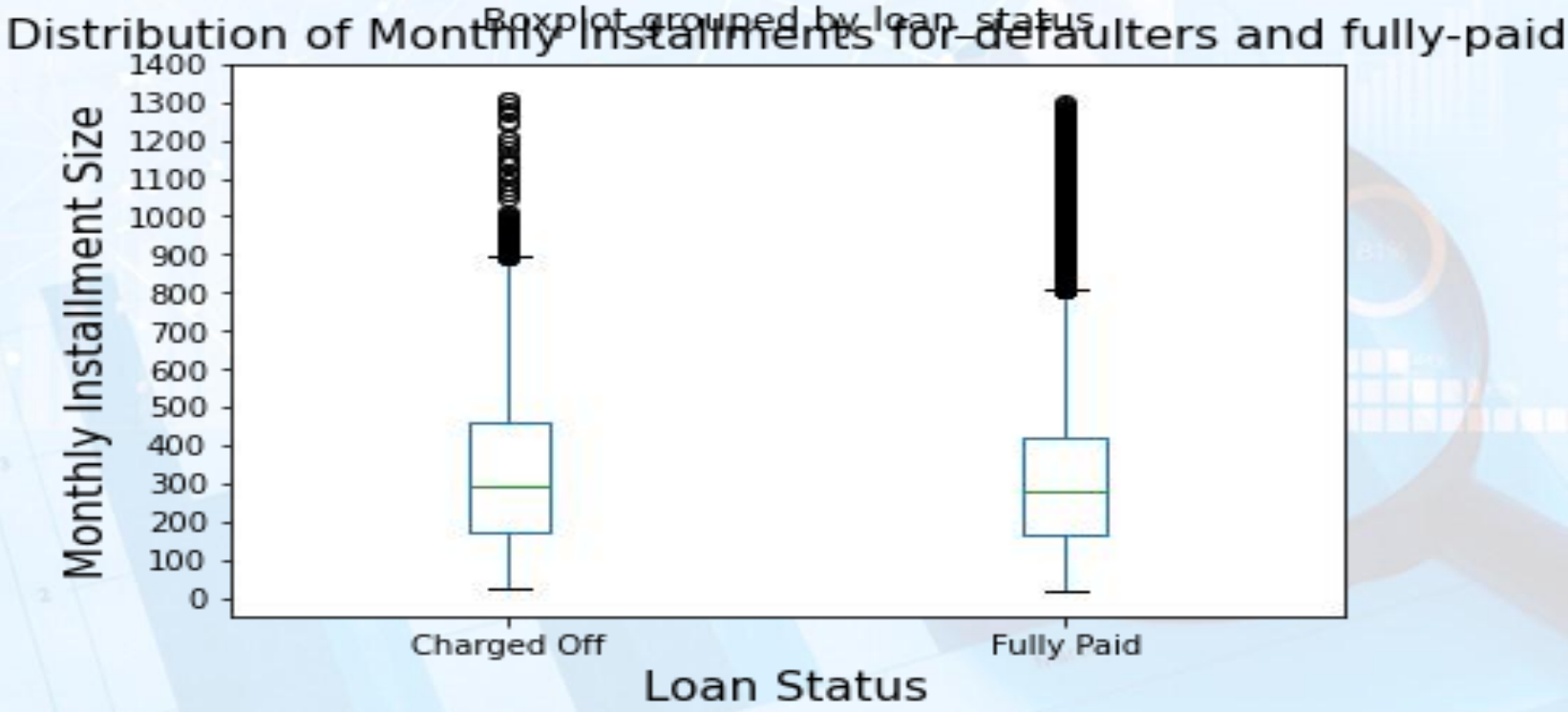
Insights on effect of Interest rates on defaulting



Insights on effect of Interest rates on defaulting

- 50% of fully paid loans have int_rate under 11 (approximately) and 50% of defaulters have int_rate between 11 to 16.
- So we can safely conclude that int_rate of < = 11 can result into low-risk loans.
- Maximum int_rate in fully-paid is 22 (not considering outliers).
- **This means loans with int_rate > 22 are high-risk and sure shot defaulters.**

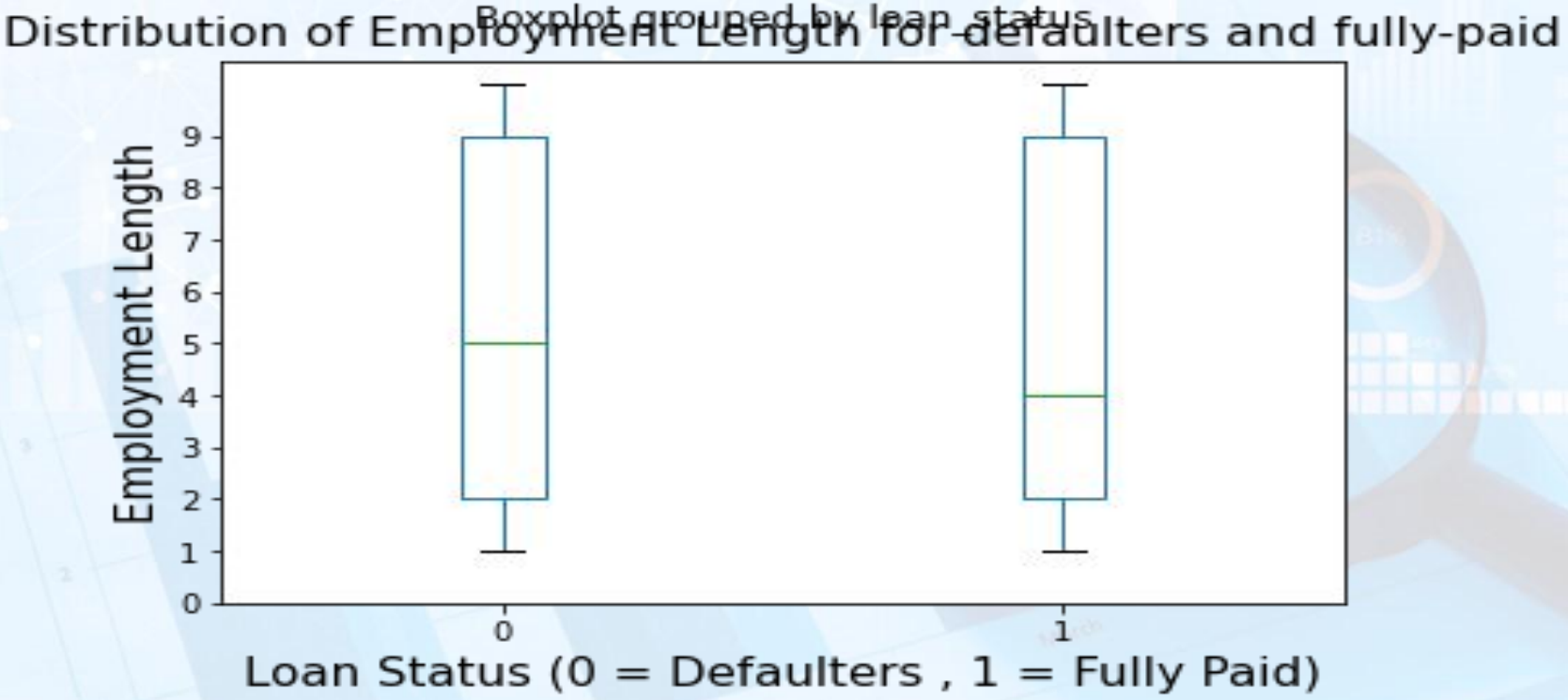
Insights on effect of Monthly Installment Size on defaulting



Insights on effect of Monthly Installment Size on defaulting

- The 50% to 75% distribution of charged-off loan is slightly larger than fully-paid loans.
- It shows that if monthly installment is larger, then chances of default is a little higher compared to smaller installments.
- Max installment of Fully paid is approximately 800 (not considering outliers).
- So installments > 800 are all high-risk loans and more likely to default.

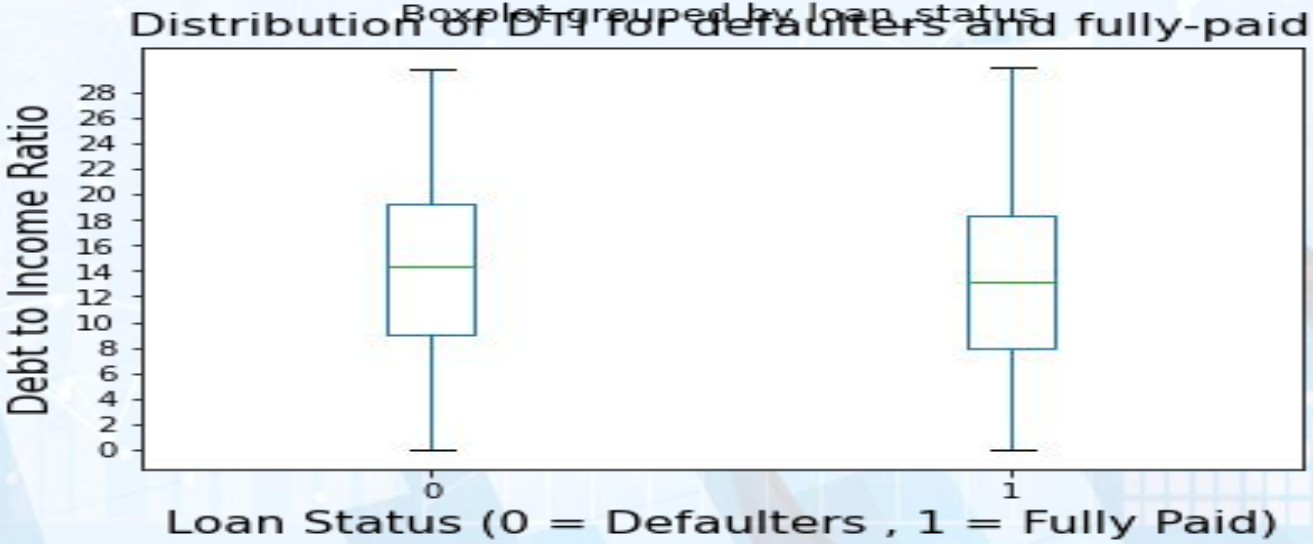
Insights on effect of employment length on risk



Insights on effect of employment length on risk

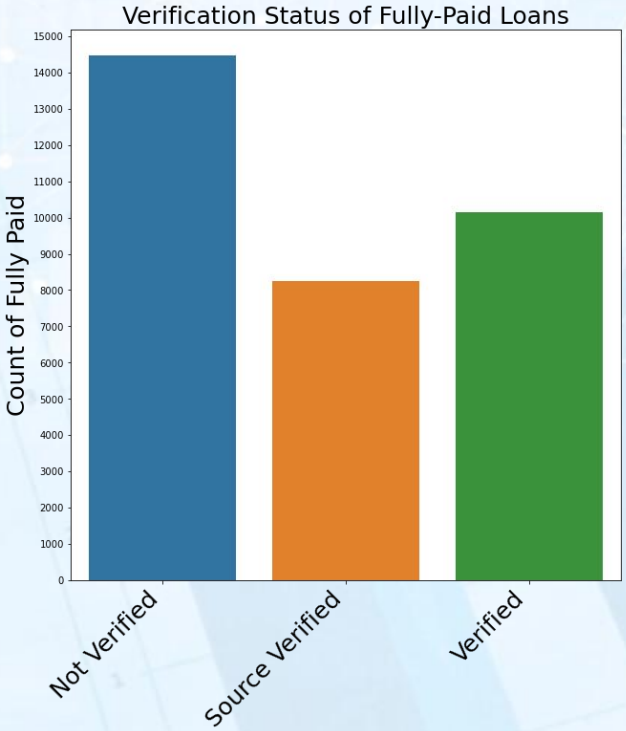
- It is difficult to draw any major insights here.
- However, median of defaulter is higher than fully paid, which kind of points that people who are longer in the job, are more likely to default. But this is a little contradicting with common logic and hence need further investigations.

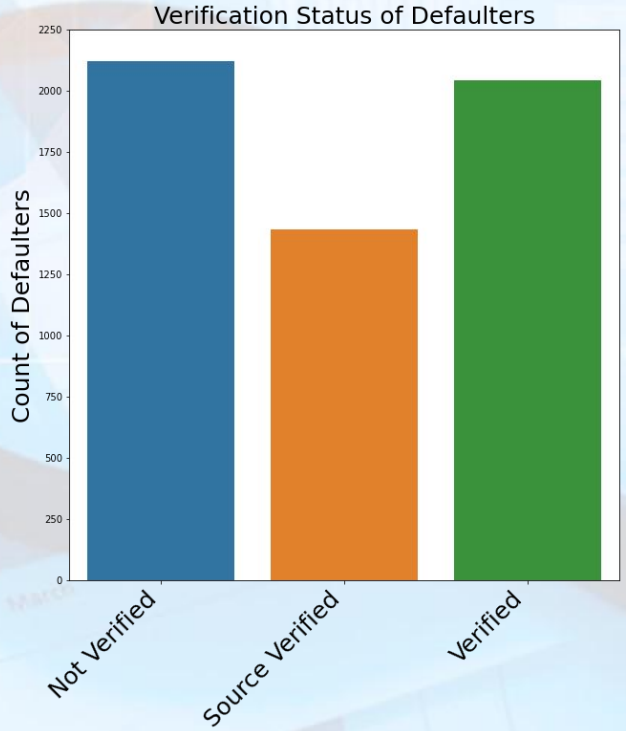
Insights on effect of DTI on risk.



- Surprisingly, no major insights on effect of DTI on risk.
- However, 75% of fully paid loans hv DTI< = 18, which means loans to borrowers with DTI > 18 are risky.

Insights on effect of Status of Verification on Risk





Insights on effect of Status of Verification on Risk

- The first plot indicates clearly that number of 'Not Verified' borrowers who have fully paid is significantly higher than 'Source Verified' or 'Verified' category.
- Second graph shows that number of 'Non-verified' defaulters is very close to 'Verified' defaulters with slight higher number.
- **This is a very serious concern. It looks like the verification process itself is faulty and needs to be evaluated and revised.**

Summary of Important Findings

- Attributes needed for a Low-Risk Loan:
  - **Grade A or B & Sub-grade - A1 to C3.**
  - 36 months term, interest rate which is less than 11, monthly installments size which is less than 400$.
- Attributes that points to a high-risk Loan:
  - **Grade G, Sub-grades F3 and above.**
  - Purpose of taking loan as debt consolidation / credit card / small business / other.
  - interest rate which is more than 22, monthly installments which is greater than 800$.

Summary of Important Findings Contd.

- Scrutiny process needs to be tightened
  - for states of CA, NY, FL and TX
  - for purpose categories debt consolidation, credit card, small business and ‘other’
- To our surprise, ratio of Fully paid / Defaulted for Mortgage is higher compared to other categories pointing to low-risk. This might be because if the house is already on mortgage, defaulting means losing it completely and so the borrowers will try their best to repay.
- **The most important finding was that non-verified loans are more likely to fully pay than the verified ones, which soul not be the case for obvious reasons.**

  The whole point of verifying borrower's information is to reduce risk. This indicates that the verification process that is followed is flawed and needs some major revision.

Acknowledgement

- Thank you Anand sir for empowering us with tools of EDA
- Thank you Sajan sir for introducing us to the world of open source repository and how we can contribute to it
- Thank you Behzad sir for strengthening our python coding skills