Gramener Case Study

Submitted by:

Purva Sharma

Shashank Shekhar

Vidya Kurada

Objective

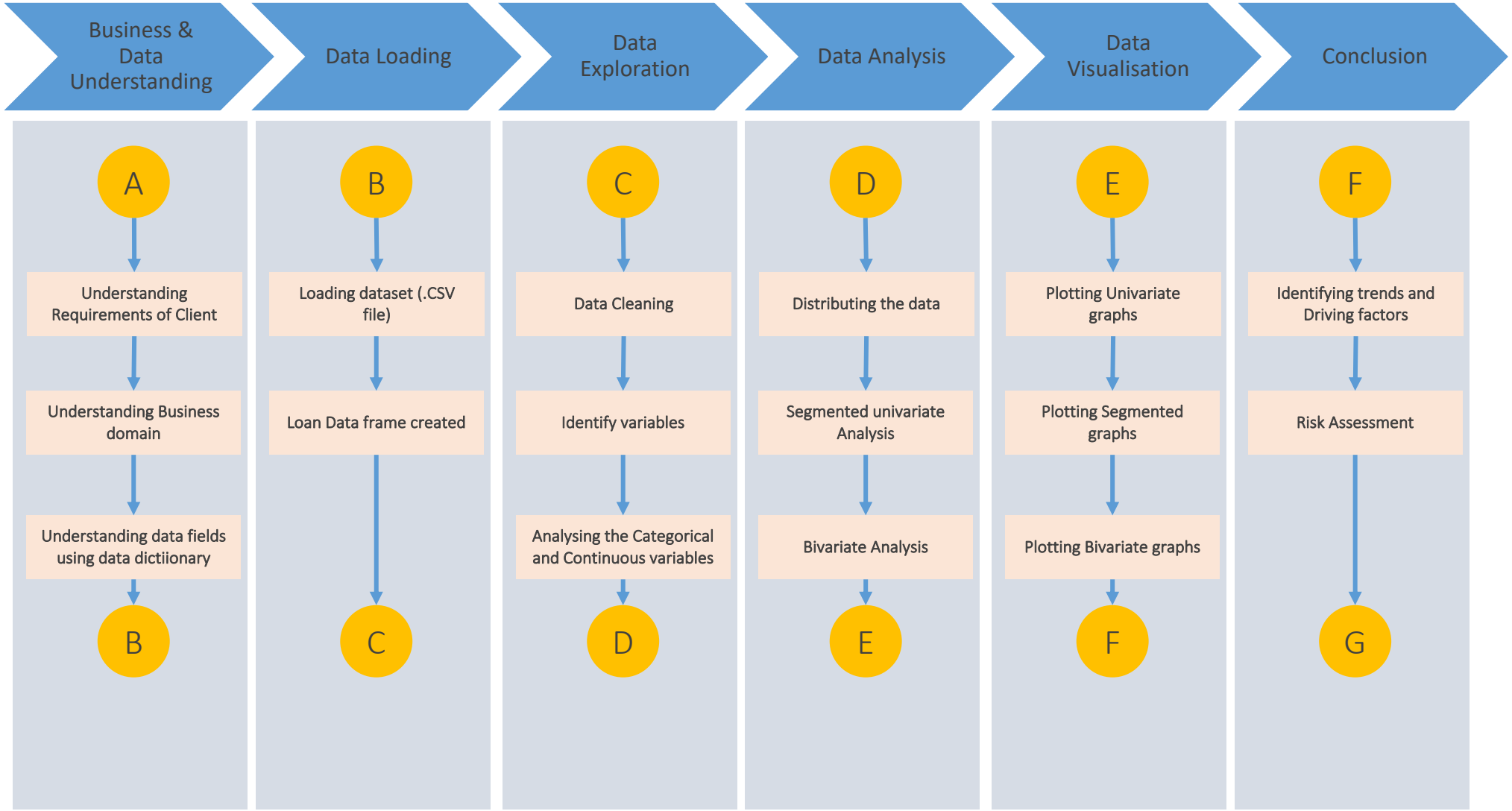
Purpose of this exercise is to:

1. Identify and analyse the patterns which indicate if a person is likely to default a loan.
2. Understanding the driving factors (or driver variables) behind loan default
3. Utilise the captured knowledge for company's portfolio and risk assessment.

Problem approach :

- Business and data understanding
- Data Loading
- Data Cleaning
- Data Analysis
- Data Visualization
- Recommendation's

Problem Solving Methodology



Business Understanding

- Understanding the requirement of the client.
- Understanding the data provided.
- Getting hold of the domain using the data dictionary provided.
- Checking the Risk Factors in the given dataset.

Data Understanding

Varaible Table	
Categorical Variables	Continuous Variable
Home_Ownership	Loan_Amt
Verification_Status	Funded_amt
Emp_Length	Funded_amnt_inv
Term	Int_Rate
Addr_State	Installment
Grade	Annual_Inc
Sub - Grade	Dti
Purpose	Revol_bal
Delinq_2yrs	Inq_last_6mths
Pub_rec	Revol_bal

We divided the fields into Categorical and Continuous variables

Income Classification	
Annual Income	Zone
>= 70000	High
>= 35000 and < 70000	Medium
>= 10000 and < 35000	Low
>= 1 and < 10000	Poor

Interest Rate Classification	
Interest Rate	Zone
>=16	High
>= 8 and < 16	Medium
>= 8 and < 16	Low
> 0 and i< 8	Not Known

Loan amount	
Loan amount	Zone
>= 23000	Very High
> 15000 and <= 23000	High
> 5000 and l<= 15000	Medium
> 0 and <= 5000	Low
Others	Not Known

We categorized numeric values for our analysis .

Exploratory Data Analysis

Step 1: Data Cleaning

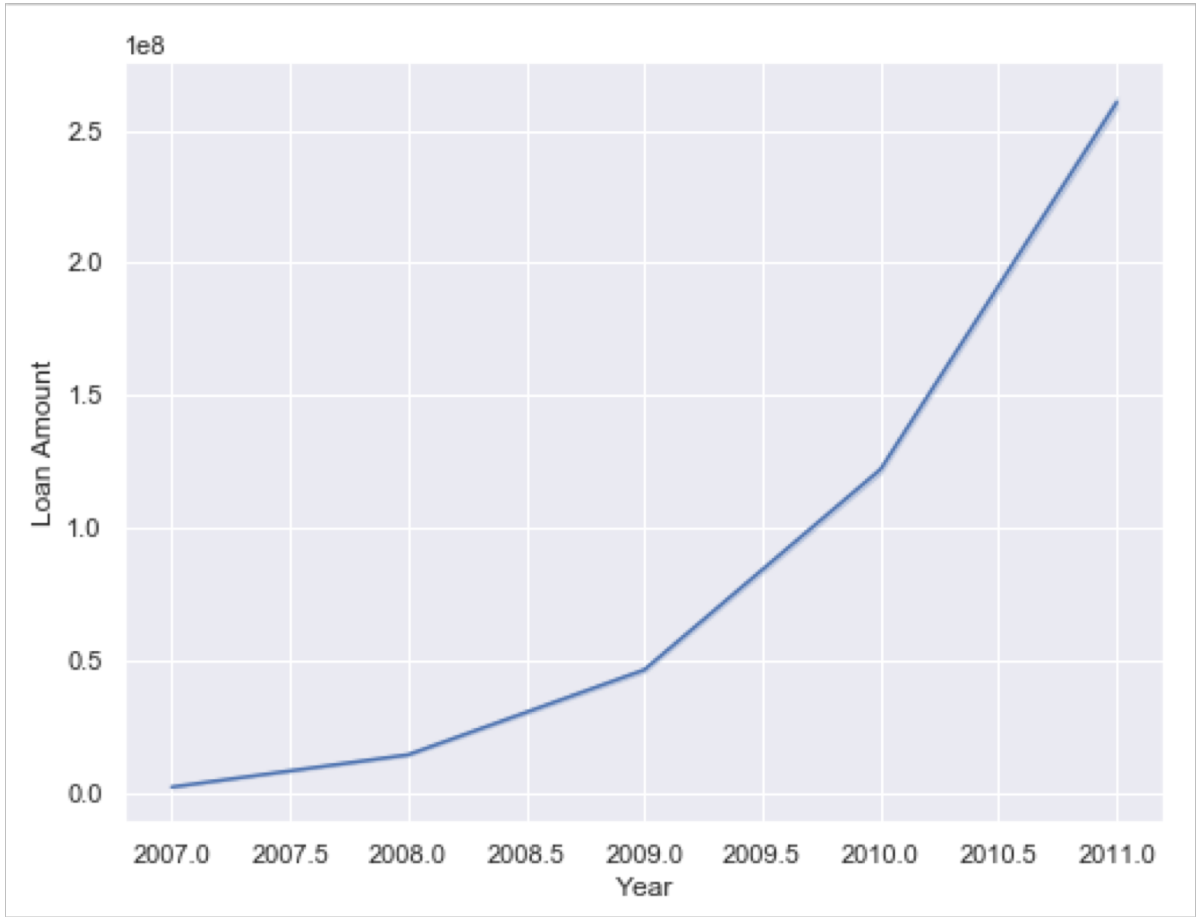
We follow the below steps to clean and understand the data:

1. First we loaded the data in the main data frame and analysed its attributes.
2. After this we started cleaning the data:
    - Removed the unwanted columns
    - Removing the Empty columns
    - Removing the rows with large number of null values
    - Removing those columns which has only 1 unique value
    - Converting object type column into numeric
    - Removing the unwanted characters from the columns
3. After cleaning , we are ready for the exploration and analysis
4. In our data frames we are only talking about **Fully paid and charged off** request. We are not taking Current accounts as this will not provide any useful information about the Loan payment and risk assessment.

Company Growth

Growth Curve of the company from past 5 year:

- The graph clearly shows that the company has a steep growth curve.
- It entered a billion-dollar market by the Year 2011.



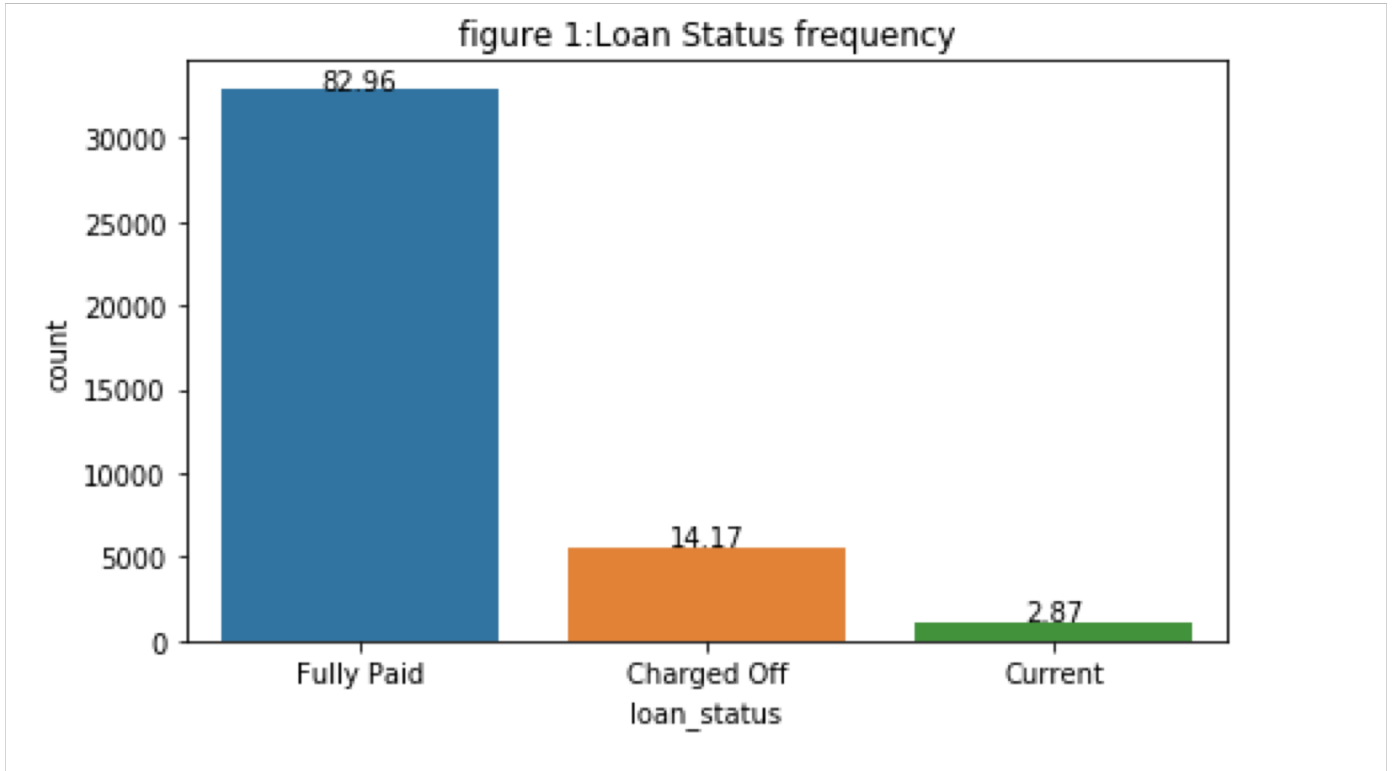
Data Distribution

Checking the Loan Status overall the data

Based on the dataset 82.96% of the applicants have already fully paid the loan amount.

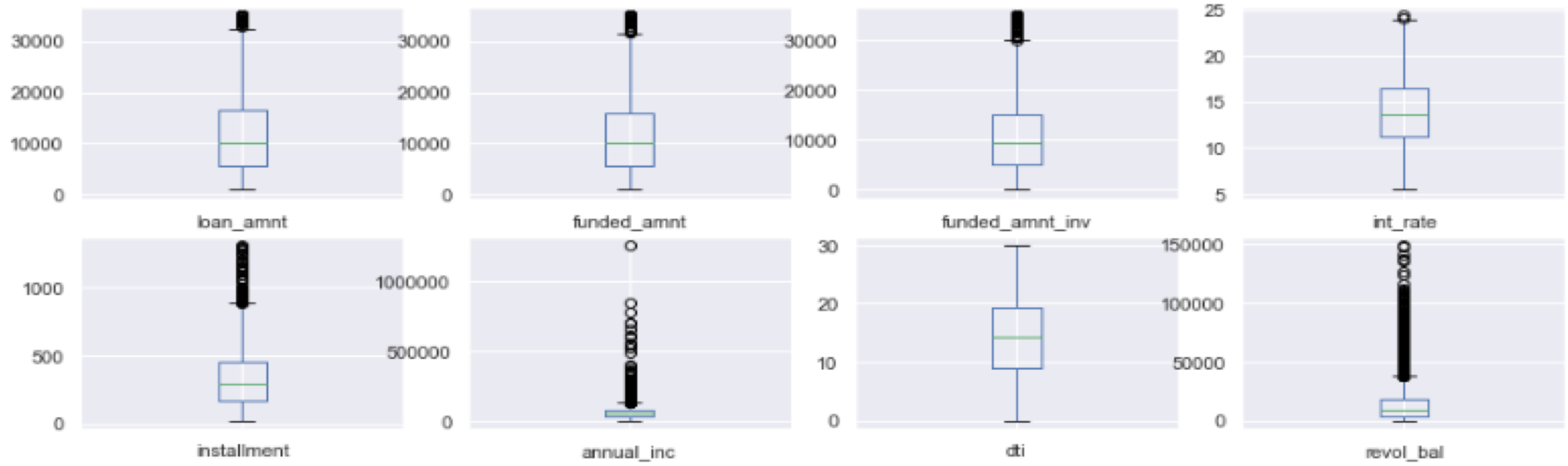
14.17% are under charged off and only 2% are currently paying.

Hence, we will be using **Fully Paid and Charged Off loan candidates** for our prediction and analysis.



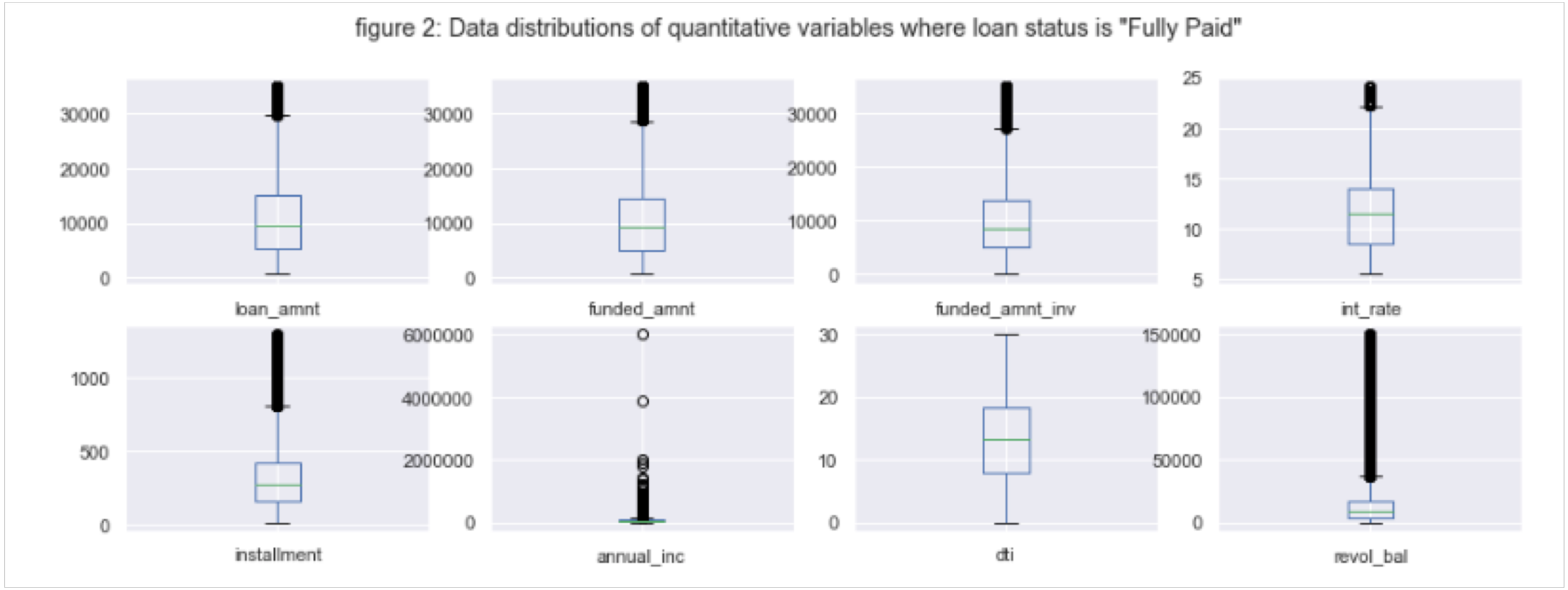
Quantitative variable analysis – for Charged Off Records

figure 1: Data distributions of quantitative variables where loan status is "Charged Off"



- Large number of outliers are visible in installment, annual income and revol_bal. These loans effect the mean in an extreme way. For Example: An installment unpaid worth > 30,000 USD will increase the company non payable asset (NPA) more than having unpaid installments of 20 borrowers at mode level < 1500 USD.
- To re emphasis, annual income looks largely skewed in distribution.
- It is the same for the ***Fully Paid records*** as well.

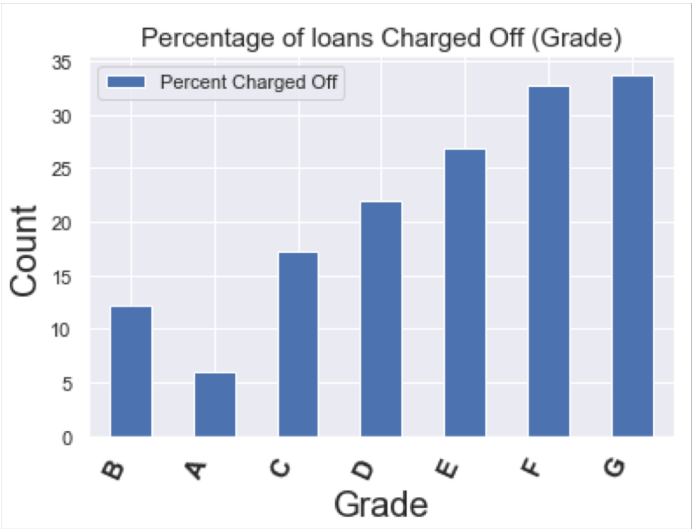
Quantitative variable analysis – for Fully Paid Records



Segmented Univariate Analysis for Categorical variables - Grade , Term

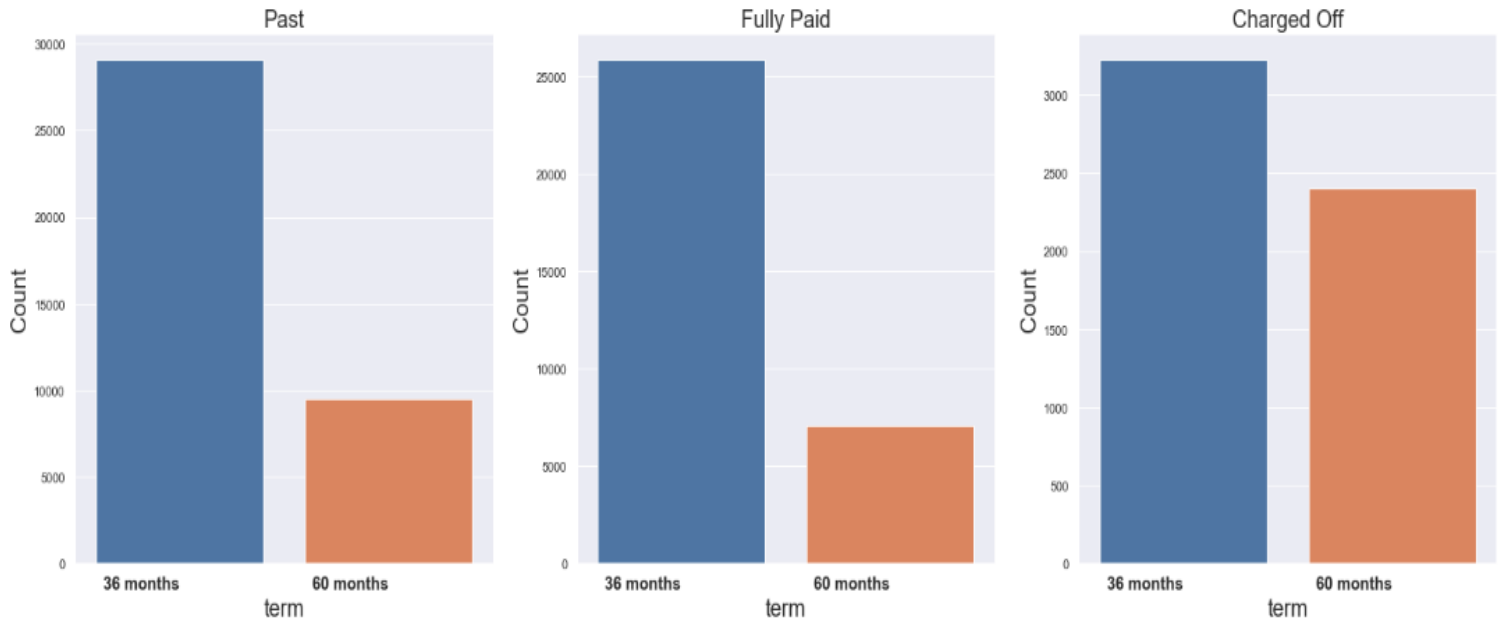
1. We are comparing 3 segments for analysis:
   - 1. Total sanctioned loans (Charged off + Fully Paid)
   - 2. Charged Off
   - 3. Fully Paid
2. While Analyzing it seemed evident that (Plots 1-3):
   - **B,C and D largely contribute to charged off loans**
3. Now what is less evident is that, **number of loans given in category "B" are as high as total number of loans given.**
   - If you see the **percentage of loans "Charged Off" to total loans given, it is only 12%**
4. Where as **grades "C,D,E,F,G" has highest number of charged off loans to the loans given.**
   - We neglect "E,F,G" loans since they contribute very less.
   - Comparing **percent loans charged off in "B,C,D" w.r.t loans sanctioned in that grade, "C,D" are more.**
   - **B is better performing** in getting fully paid but still cannot ignore with **12% loans charged off.**

Grade	Total loans	Percent loan	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
B	11675	30.26	87.79	12.21	0.14
A	10045	26.04	94.01	5.99	0.06
C	7834	20.31	82.81	17.19	0.21
D	5085	13.18	78.01	21.99	0.28
E	2663	6.9	73.15	26.85	0.37
F	976	2.53	67.32	32.68	0.49
G	299	0.78	66.22	33.78	0.51

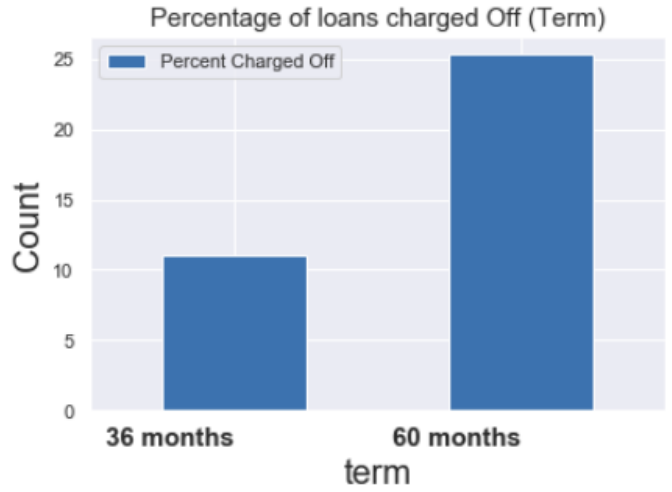


Segmented Univariate Analysis for Categorical variables - Term

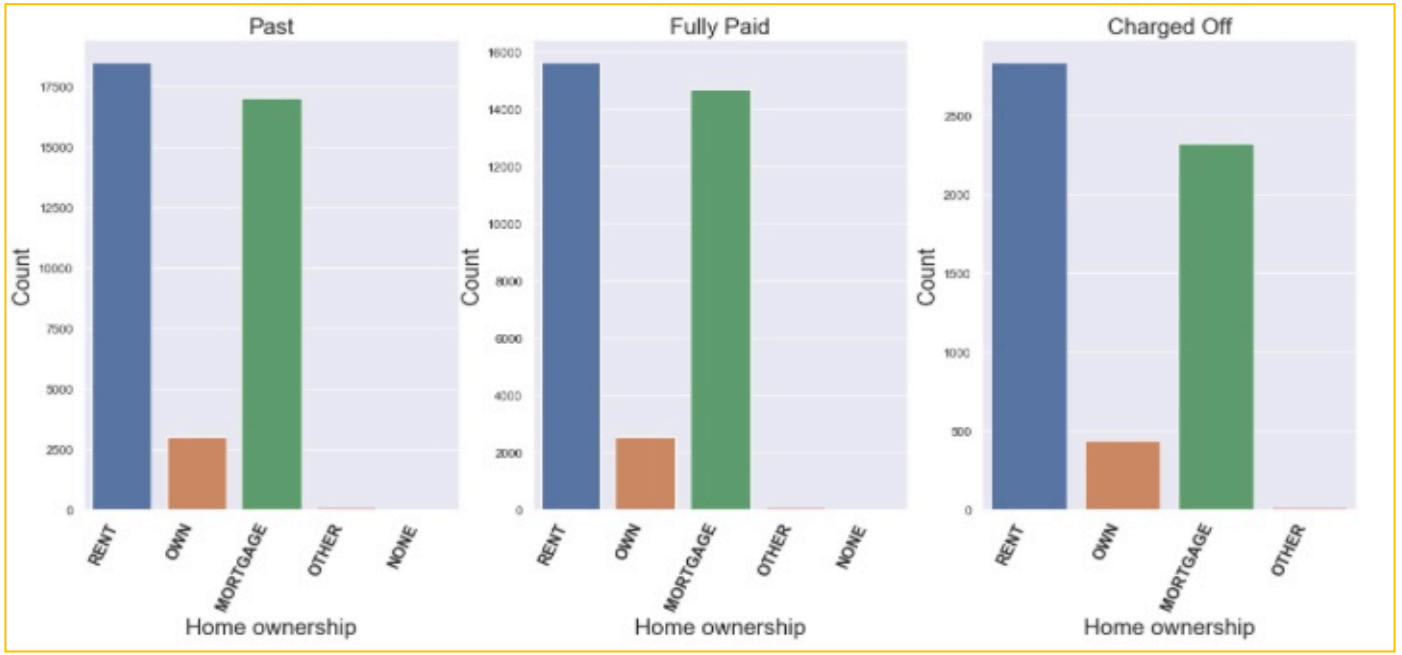
1. We are comparing 3 segments for analysis:
   - 1. Total sanctioned loans (Charged off + Fully Paid)
   - 2. Charged Off
   - 3. Fully Paid
2. Similarly to Grades, **more loans are charged of in 36 months term category.**
3. But if you take the **ratio of number of loans charged off to the total number of loans sanctioned.**
   - **Only 11% loans charged off** in the 36 months term
4. Now, check the **ratio of charged off loans to sanctioned loans in 60 months category, this percentage is as high as 25%.**
5. Probably this **might be the reason why LC is giving less number of 60 months loans**
6. We can by far conclude, it is **highly likely that 60 month term loans charge off more than 36 months loans**



Term	Total loans given	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
36 months	29096	75.42	88.91	11.09	0.12
60 months	9481	24.58	74.69	25.31	0.34



Segmented Univariate Analysis for Categorical variables - Home ownership



1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid

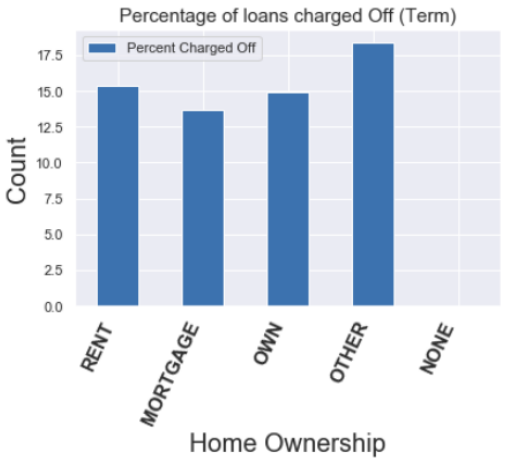
2. Check from the table that, approx. **92% loans are sanctioned in "Rent" and "Mortgage" Category**

3. And as expected, **these 2 categories have more charged off loans but not exceeding 15% of total loans given in each category.**

4. Plot – 4 is little **deceptive** saying that, **"Others" have more charged off loans but there almost <100 loans sanctioned in total (negligible).**
   - Others can be omitted from analysis

5. We did not see any specific risk involved giving loans to other categories which should be alarmed.

Home Ownership	Total loans given	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
RENT	18480	47.9	84.64	15.36	0.18
MORTGAGE	17021	44.12	86.33	13.67	0.16
OWN	2975	7.71	85.11	14.89	0.17
OTHER	98	0.25	81.63	18.37	0.22
NONE	3	0.01	100		

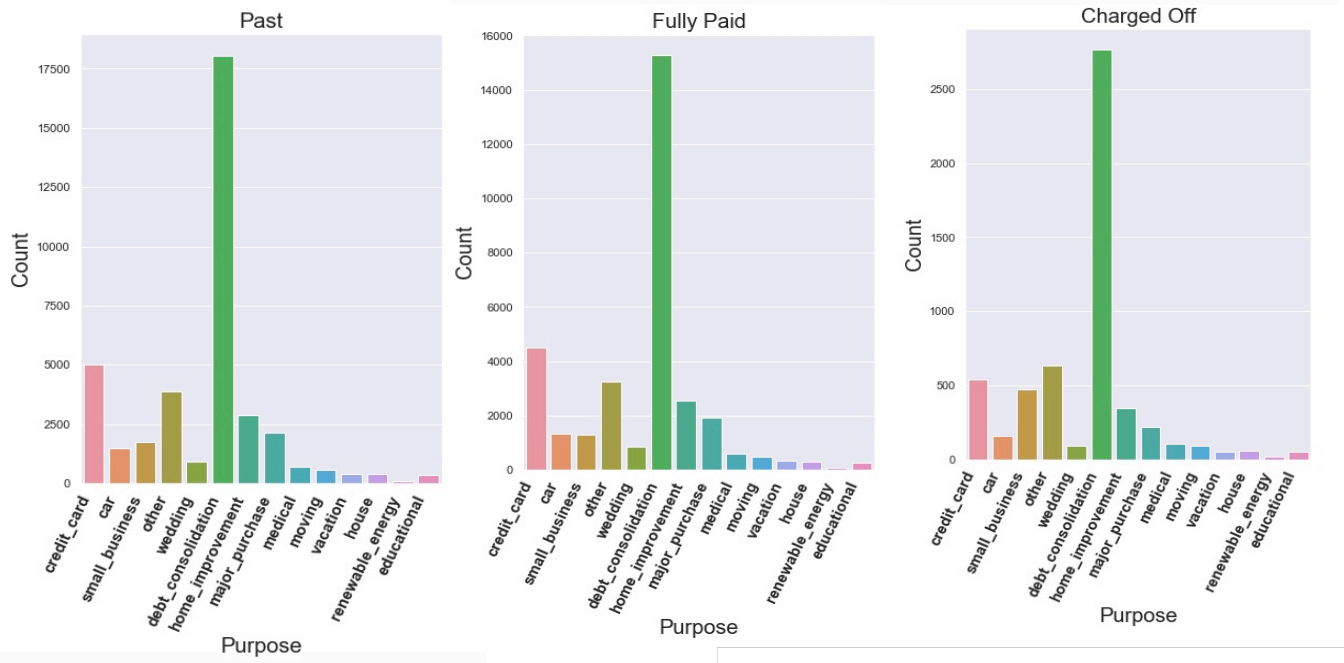


Segmented Univariate Analysis for Categorical variables - Purpose

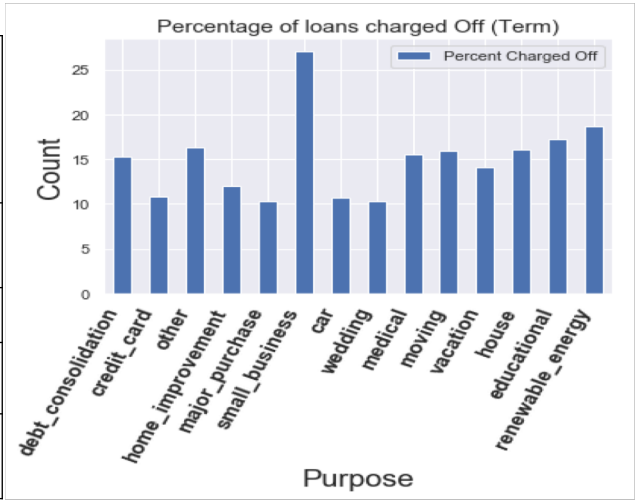
1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid

2. In purpose, we consider looking at only 4 primary categories.
   - Debt Consolidation
   - Credit Card
   - Other
   - Home improvement
3. These **4 constitute to 75% of the loans sanctioned**.
4. **"Debt Consolidation"** contributes to charged off loans but **largely due to the volume of loans given, holding to 15% of the loans sanctioned getting charged off.**
5. But more importantly, **"other" purpose category loans are charged off at 16%.**

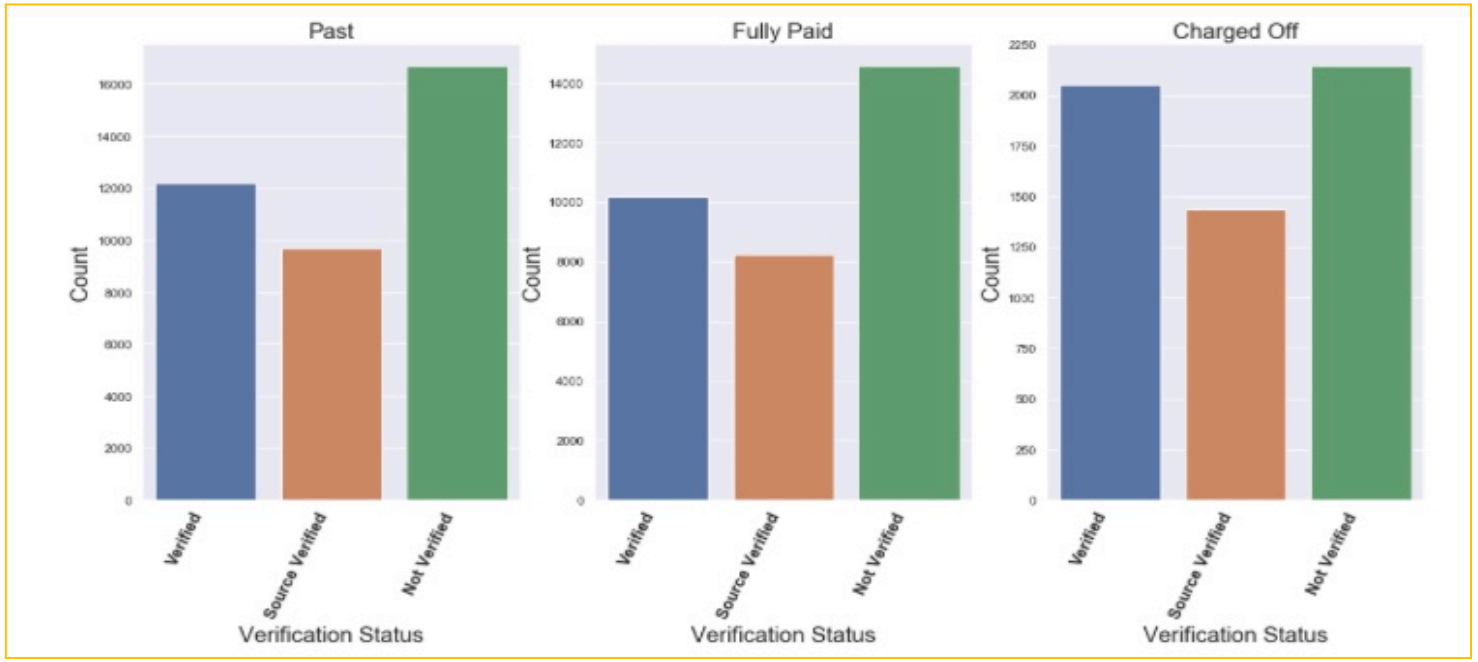
It's advisable that LC slows down on debt consolidation and charged off loans.



Purpose	Total loans given	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
debt_consolidation	18055	46.8	84.67	15.33	0.18
credit_card	5027	13.03	89.22	10.78	0.12
other	3865	10.02	83.62	16.38	0.2
home_improvement	2875	7.45	87.93	12.07	0.14



Segmented Univariate Analysis for Categorical variables - Verification Status



1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid

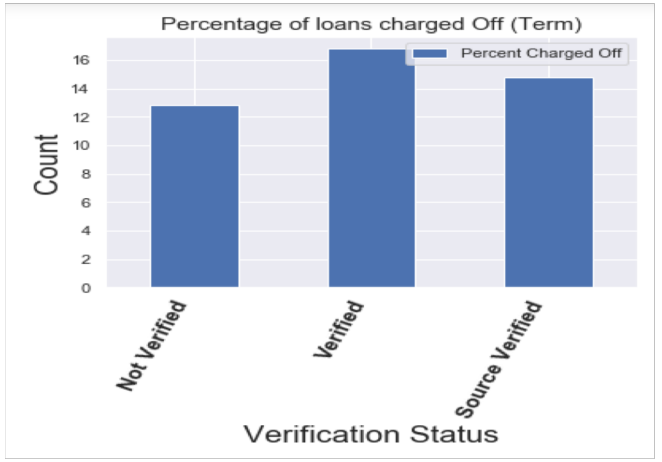
2. Its pretty self explanatory that **charged off loans mostly are not verified.**

3. Plot- 4 with percentages counter intuits that by saying **verified loans have more charged off to sanctioned percent at 17%.**

4. **What might be the reason for being or not being verified ?**
   - Repetitive Customers are not verified
   - Verified customers are new

5. At this stage, nothing can be safely said, we will see **some bivariate analysis to support or counter act the data here.**

Verification Status	Total loans given	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
Not Verified	16694	43.27	87.17	12.83	0.15
Verified	12206	31.64	83.2	16.8	0.2
Source Verified	9677	25.08	85.18	14.82	0.17

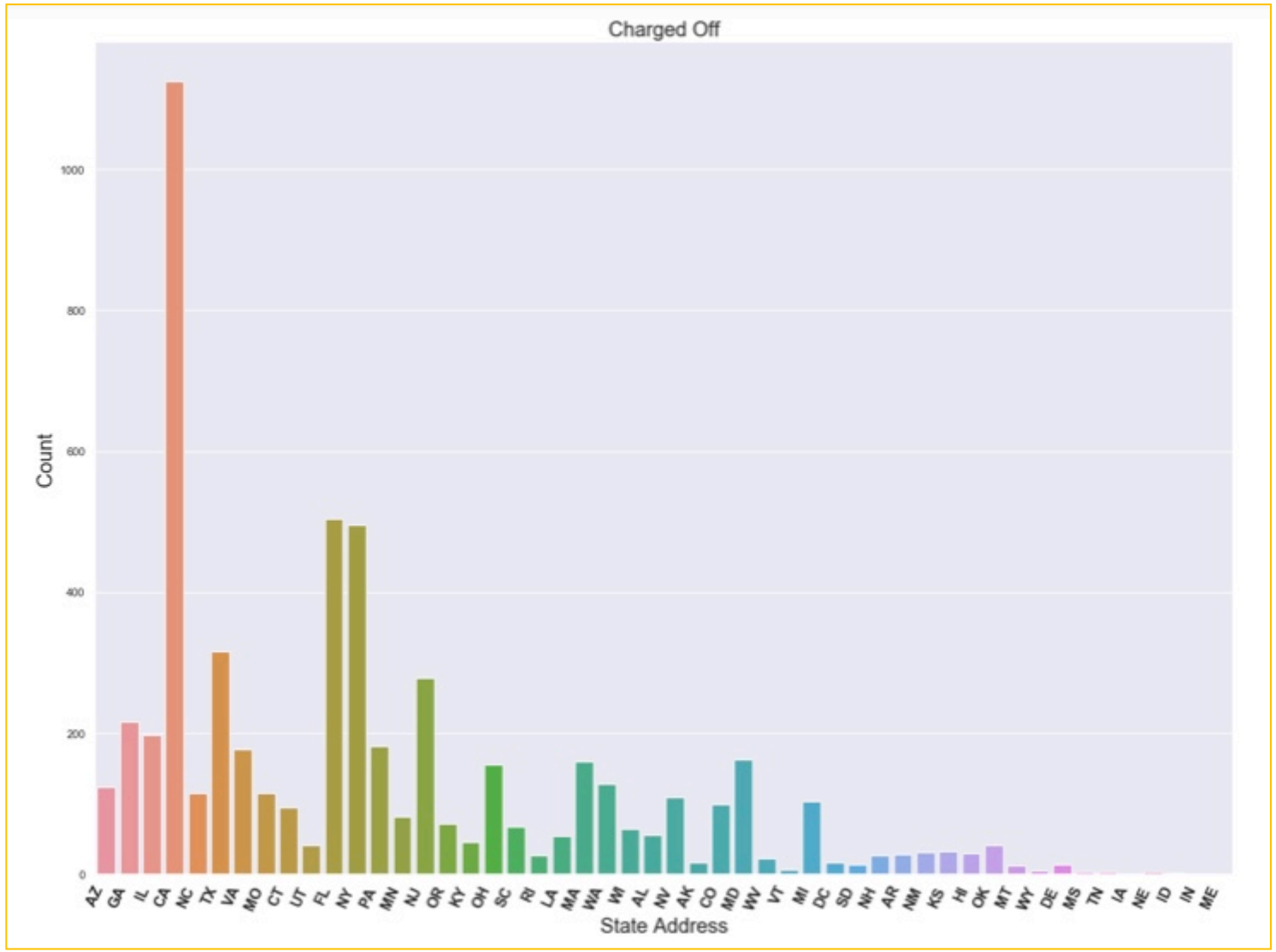


Segmented Univariate Analysis for Categorical variables: State Address for Fully Paid Users

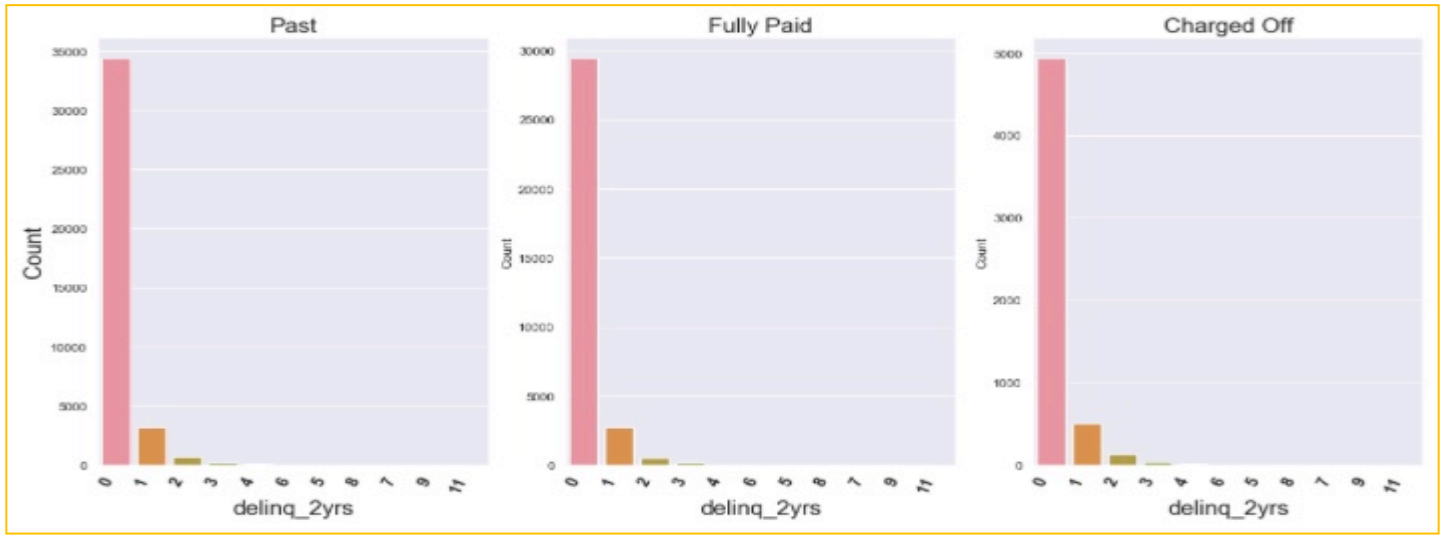
1. We are comparing 3 segments for analysis:
    1. Total sanctioned loans (Charged off + Fully Paid)
    2. Charged Off
    3. Fully Paid

2. We can clearly see that state CA(California) has highest number of Charged Off .
Since CA is LC headquarters and more loans are issued there, the dominance is merely because of the volumes.

*Note: We limited the graphs to only Charged Off case because others look predominantly same.



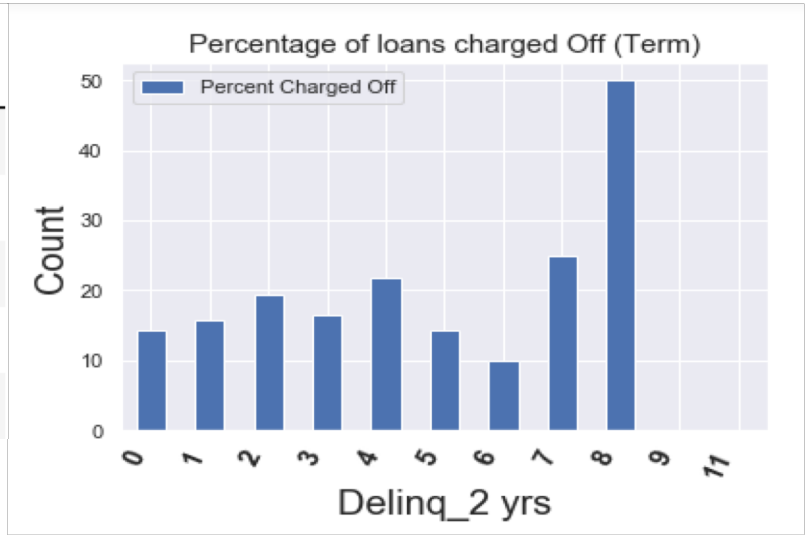
Segmented Univariate Analysis for Categorical variables: 30 days delinquent for last 2 Years



1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid

2. The graph depicts that 90% of the times, applicants with 0 delinquency record are given the loans. So, the plots are showing mere volume influence.
3. But in the overall percentage graph, people with any delinquent history are risky which is correct also.
4. Even LC sanctioned loans to people largely with one delinquency (approx 9%). We strongly feel even it is riskier.

Total loans given in each delinq_2yrs	Percent loan given in each delinq_2yrs	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
34386.0	89.14	85.65	14.35	0.17
3207.0	8.31	84.19	15.81	0.19
673.0	1.74	80.68	19.32	0.24
212.0	0.55	83.49	16.51	0.20
60.0	0.16	78.33	21.67	0.28

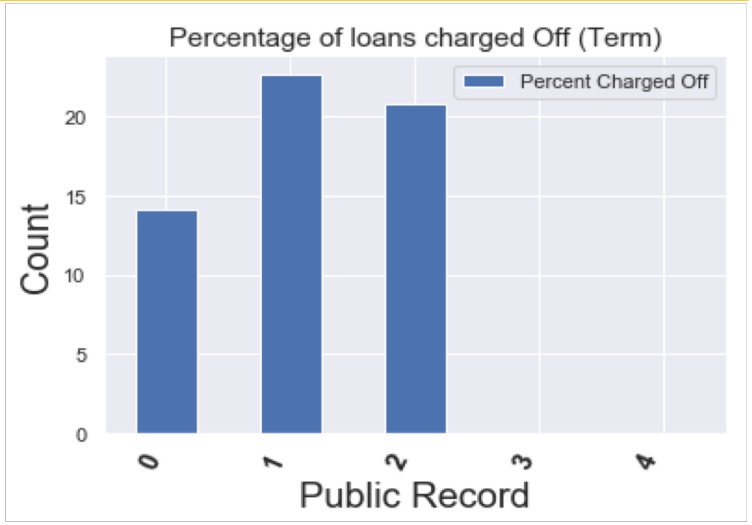


Segmented Univariate Analysis for Categorical variables: Public Records

1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid

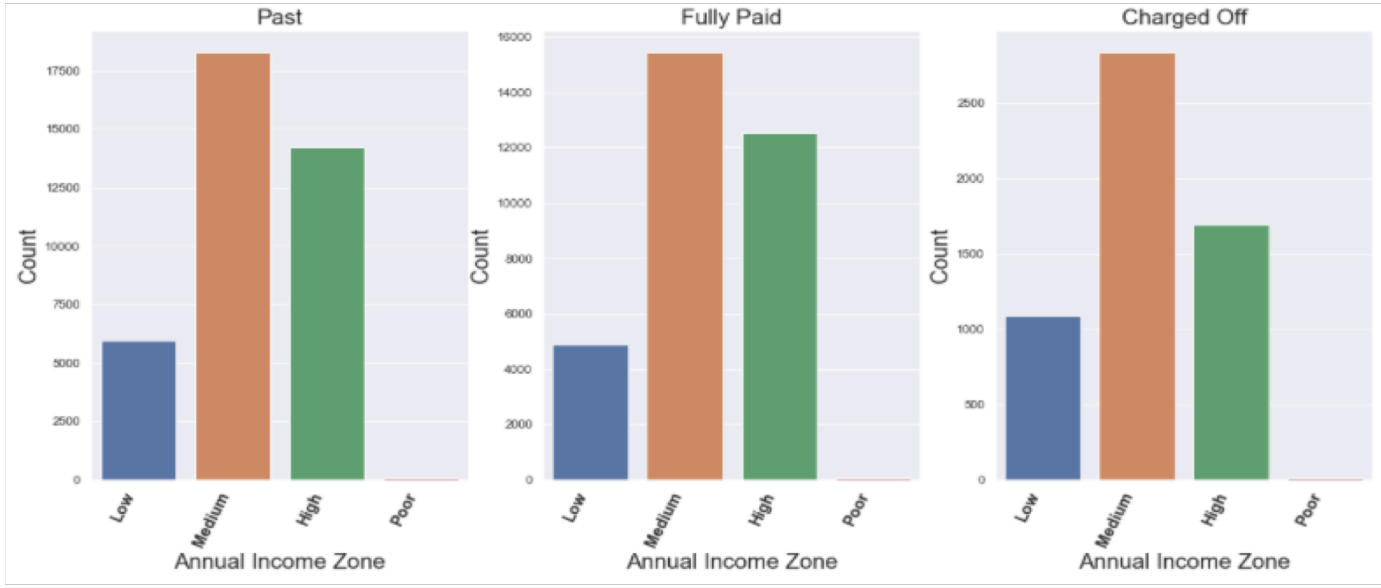
2. Public records includes bankruptcy, judgments, liens, lawsuits, and foreclosures. Basically it's a **legal liability**. The **candidates with high number of Public records are not favourable.**
3. If we go by the % vs ratio graph, we can see that company approves loan to only those applicants where public record count is less like 1 or 2. Here, such charged off cases are 22.70%

Total loans given for Pub_record	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
36507	94.63	85.87	14.13	0.16
2013	5.22	77.30	22.70	0.29
48	0.12	79.17	20.83	0.26
7	0.02	100	NAN	NAN
2	0.01	100	NAN	NAN

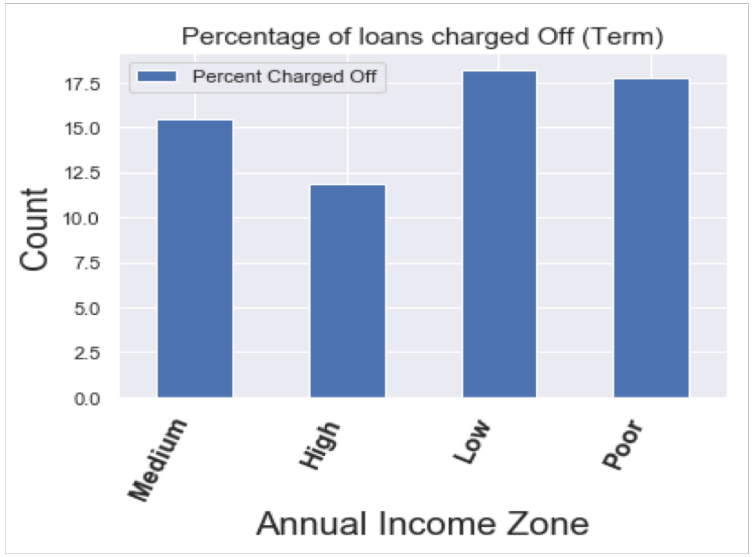


Segmented Univariate Analysis for Quantitative variables

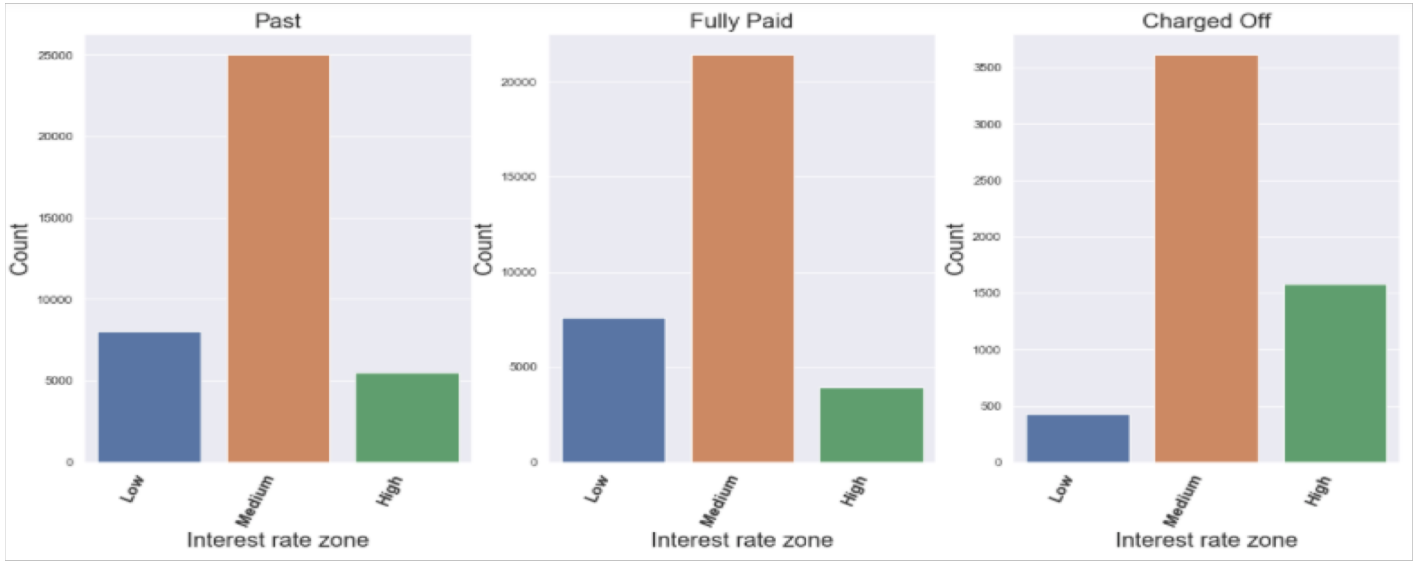
1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid
2. Poor income groups are riskier: No. of charged off loans might be high but total amount may not be.
3. Medium income groups are risky but becoming unavoidable.



Annual Income	Total loans given	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
Medium	18290	47.41	84.50	15.50	0.18
High	14234	36.90	88.13	11.87	0.13
Low	5974	15.49	81.79	18.21	0.22
Poor	79	0.20	83.28	17.72	0.22

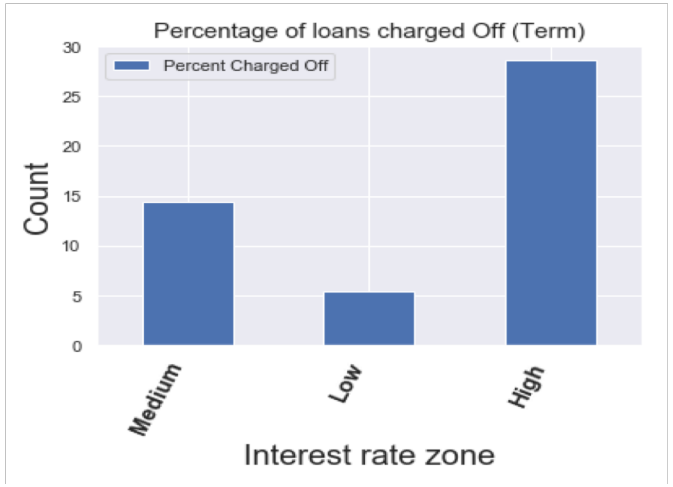


Segmented Univariate Analysis for Quantitative variables

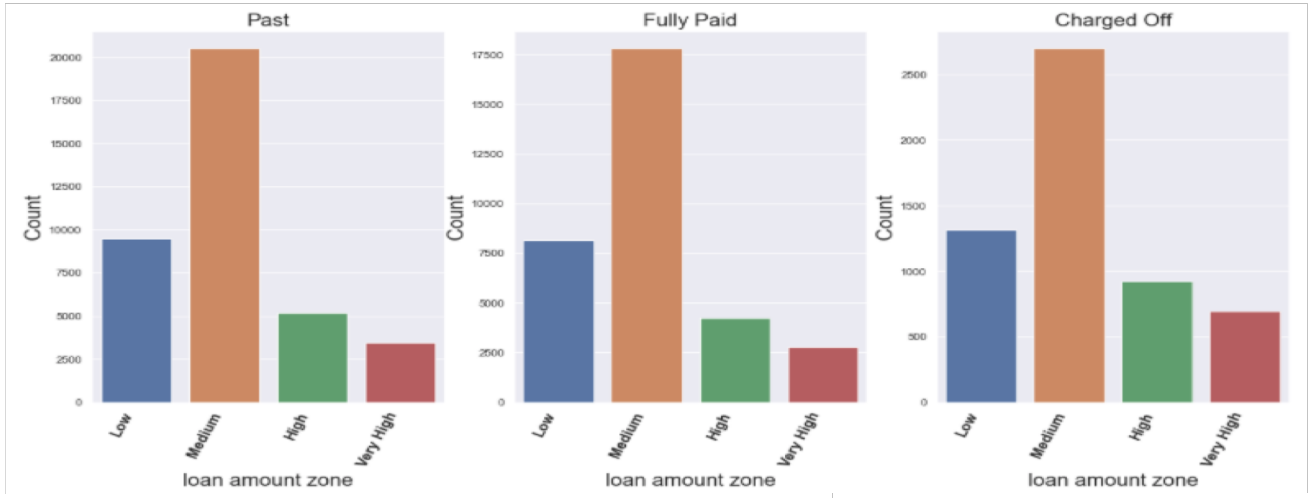


1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid
2. Plots 1-3 show high charged off loans in medium range. But they are because of volumes. 64% of loans LC sanctions typically belong to this category. It is intuitive with more loans comes more defaulters.
3. Plot-4 shows that the percentage of loans defaulted w.r.t sanctioned are more in high interest rate category. This phenomena is understandable.

Interest Rate	Total loans given	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
Medium	25047	64.93	85.56	14.44	0.17
Low	8027	20.81	94.64	5.36	0.06
High	5503	14.26	71.31	28.69	0.40

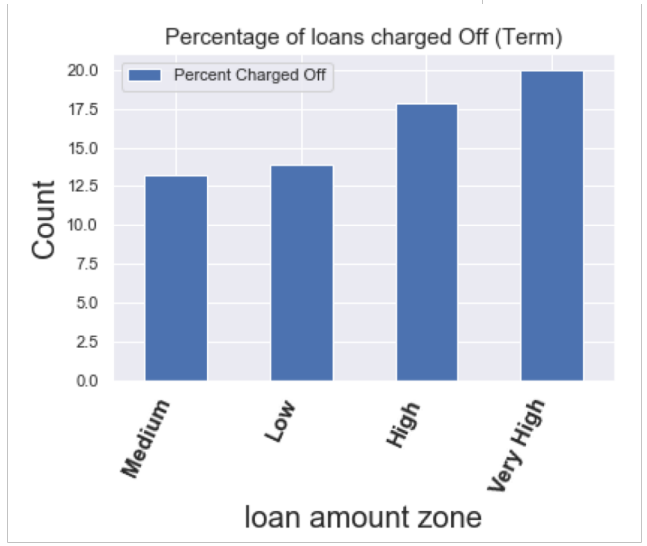


Segmented Univariate Analysis for Quantitative variables



1. We are comparing 3 segments for analysis:
   1. Total sanctioned loans (Charged off + Fully Paid)
   2. Charged Off
   3. Fully Paid
2. Plot 1-3 shows Medium and Low amount loans are charged off more. But their percentage of loans charged off to sanctioned stays at approx 26%.
3. If you refer table, in high and very high loan amount categories, loans as high as 37% are charged off. Adding to it, high loan amount means higher loss.
4. LC Should be extremely careful with high and very high loan amounts while sanctioning.

Loan Amount	Total loans given	Percent loan given	Percent Fully Paid	Percent Charged Off	Ratio of Charged Off to Paid
Medium	20485	53.10	86.83	13.17	0.15
Low	9472	24.55	86.13	13.87	0.16
High	5161	13.38	82.12	17.88	0.22
Very High	3459	79.97	20.03	0.25	79.97



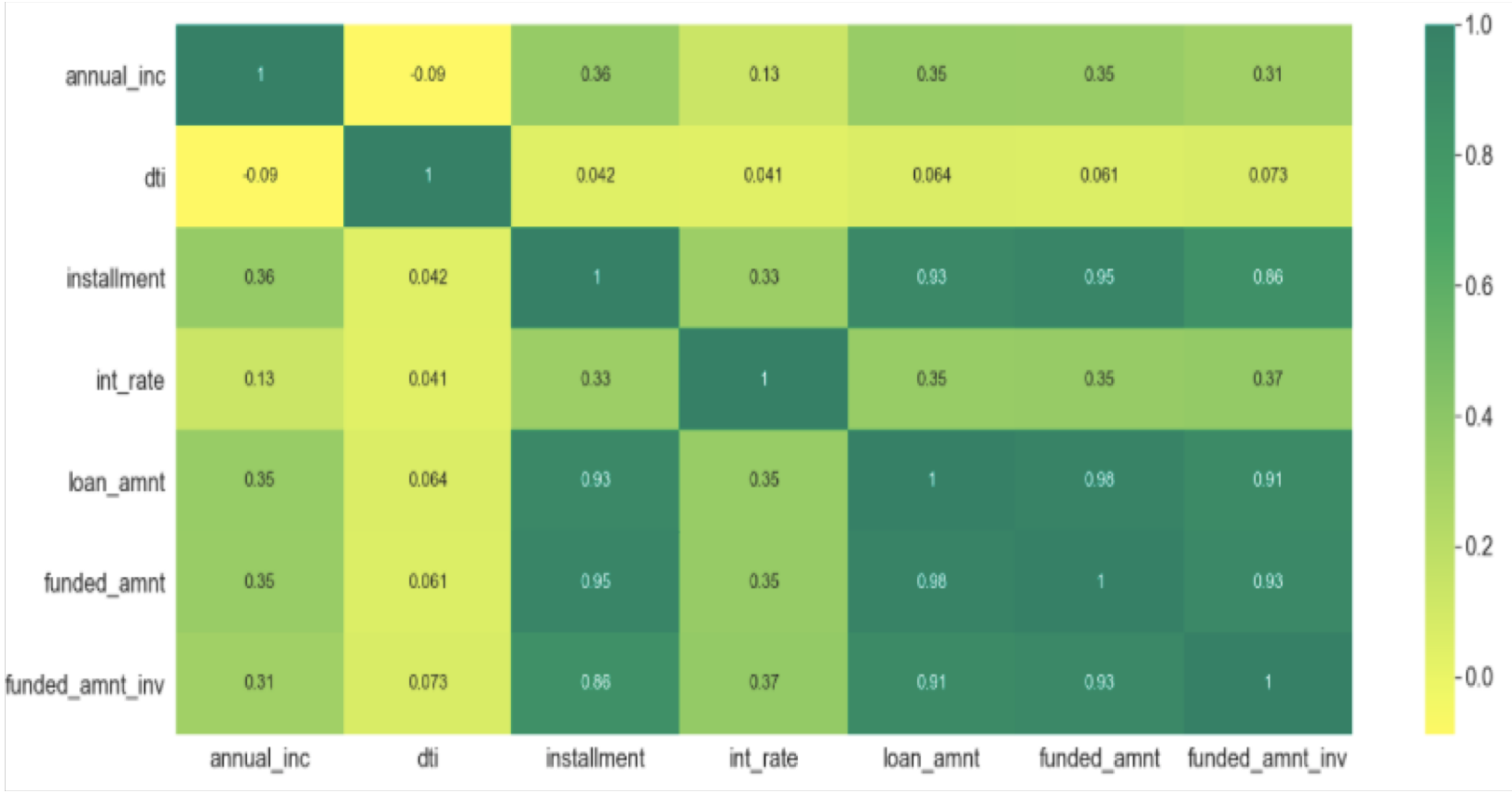
Bivariate Analysis

1. For Bivariate analysis , we limited our data with Charged Off accounts only.
2. We created the correlation between different variables and below is our observation:

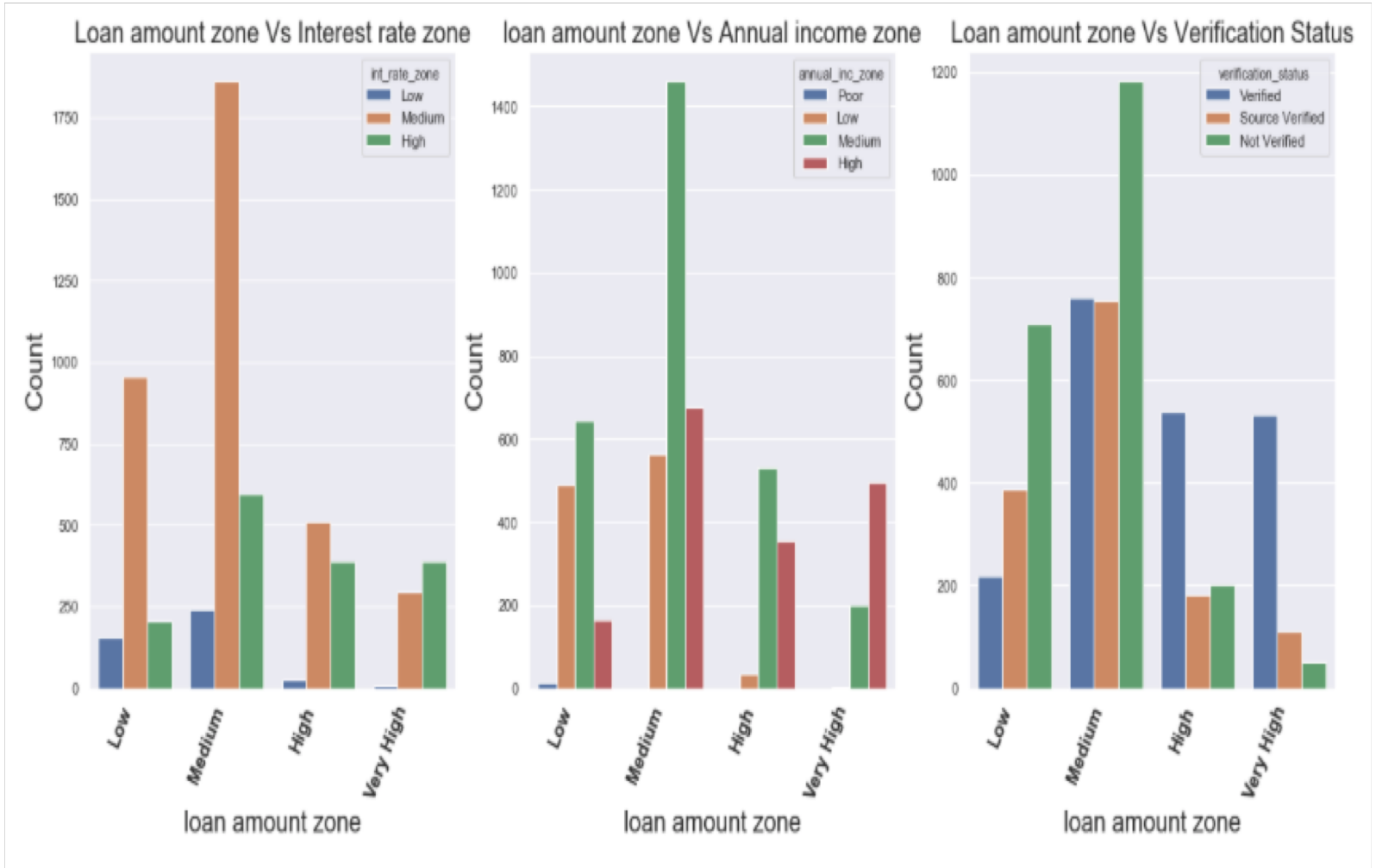
I. loan_amt as expected is highly correlated to funded_amt suggesting borrowers got what they quoted for.

II. Instalments are highly correlated to loan amounts.

To analyse further , we will create few more plots.



Bivariate Analysis:



Case1:

1. Plot-1 depicts across all income groups, medium interest rates are charged off more.
2. Plot-2 suggests something similar, Medium level annual income group charged off loans across categories.
3. Plot -3 suggests Medium range loan amounts with no verification performed poorly.

Conclusion:

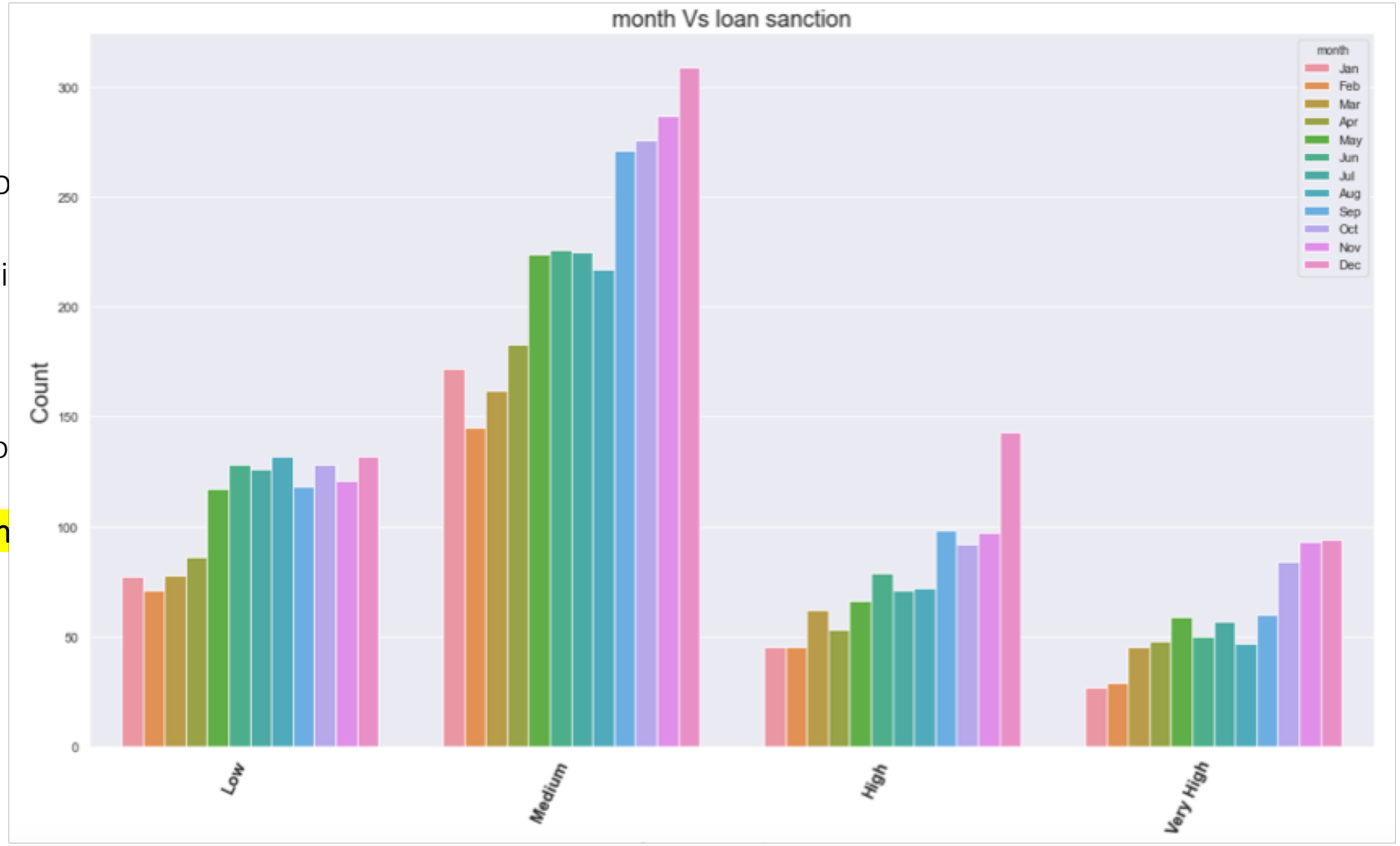
Medium range loans which are given in volumes to medium range income group got charged off more and these are the loans that are largely not verified.

From here on, LC should be more careful in verifying the mid range loans.

Bivariate Analysis:

Case2.1: Month of loan issue vs Verificatio status

1. Plot-1 is loan amount zone vs month: Thi gives a bigger picture, what are all the loans sanctioned in what all months.
2. November and December as expected always spiked in charged off loans due to holidays and financial year closing .
3. Even the high and very high amount loan sanction spiked at December.

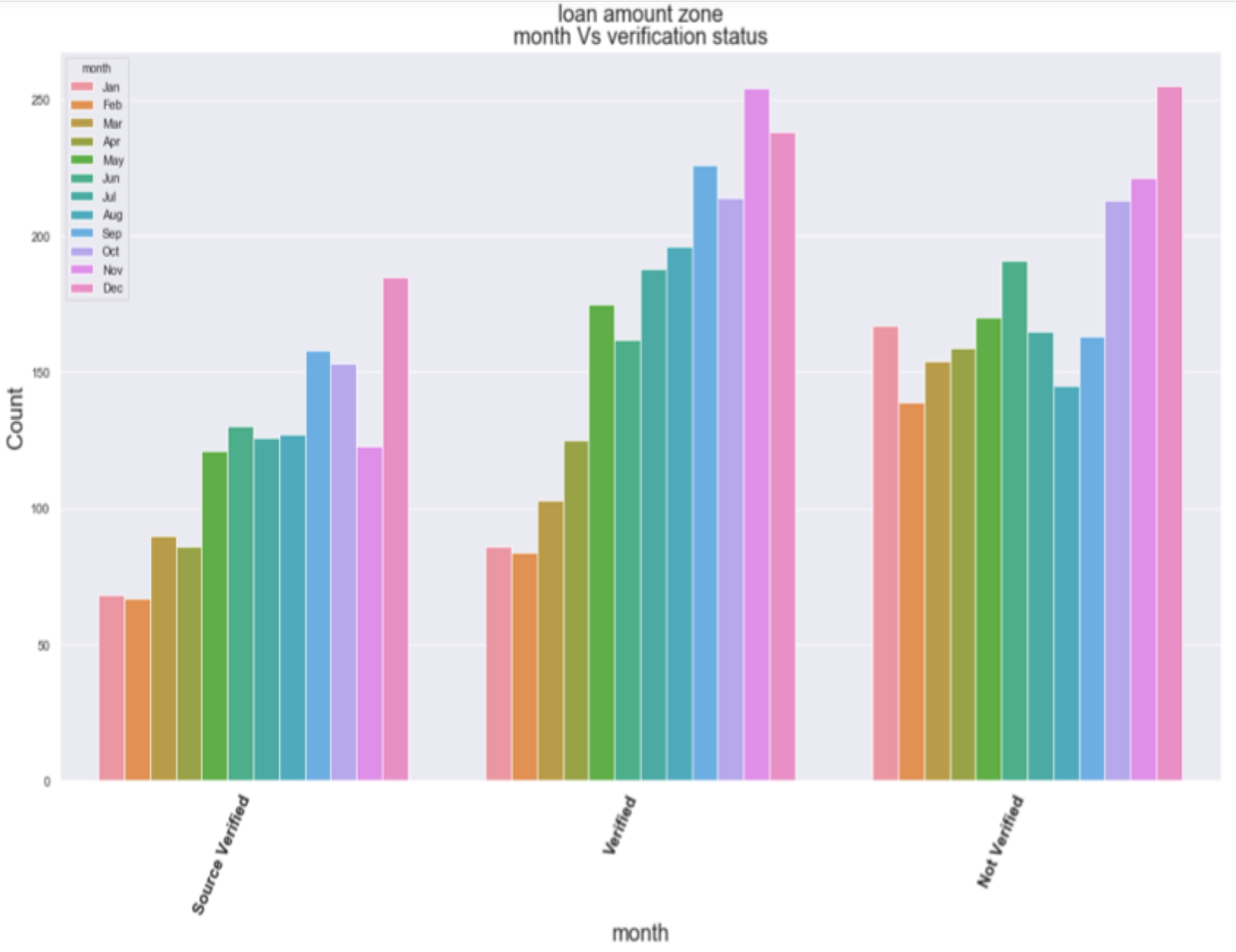


Bivariate Analysis:

Case2.2: Month of loan issue vs Verification status

1. This plot also shows similar characteristic, at the time of December, all kinds of loans spiked but particularly not verified ones.

Conclusion:
LC should either control its loan sanction or be careful about it at the Year End. They need to extra cautious during Financial Year ending , Target meeting should not be sole agenda



Conclusion

We concluded the following with our analysis:

1. Segmented univariate analysis:
    - Largely, Grade C and beyond are high risk loans
    - Longer term of 60 months are more charged off (percent wise, number charged off to total loans given)
    - Purpose with Other are highly likely to be defaulted.
    - Rented home owners are mostly defaulted.
    - Any delinquency or public record> 0 is more likely to default.
    - Mostly borrowers from state CA defaulted but that we assume is because of LC headquarters being located there and most business happens there. We feel if we have to locate one, FL is the most defaulting state
    - We found verified loans to be largely defaulted but this contradicted in the bivariate analysis.
    - From analysis of derived metrics, we get an insight on

2. Bivariate analysis:
    - Medium income borrowers with mid range loans who are not verified, defaulted.
    - Interesting insight is w.r.t financial calendar year, around November and December, not verified mid range loans sanctioned by LC defaulted.

Suggestions

Following are the suggestions that we recommend :

- Verification process needs to be air tight. All types of loan requests should be verified thoroughly before approving.
  - Verification seems to be largely compromised particularly at the financial year end.
  - Loan officers should not rush to meet the targets and should stick to company protocols strictly.
  - Company should take help of credit bureau for the borrower's profile to decide on the loan request.

- Loan requests with home ownership as either "*Rent*" or "*Mortgage*" should be scrutinized with more precision. Those should be approved only if the borrower have a reasonable income or some other means of security which can be used for recovery.

- Medium range loans were heavily sanctioned to medium range income groups. It is advisable if careful policies are written such that loan from all parties should not exceed more than 30% of borrowers income.
  This will help LC restrict giving controlled loans.

- Loan requests for paying off earlier loans should not be approved unless guaranteed repay procedures are identified.

- Last but not the least, the company should ensure that the verification authority is well trained and suitable, as we see most of the verified loans have also been defaulted raising an eyebrow over the procedure of verification.