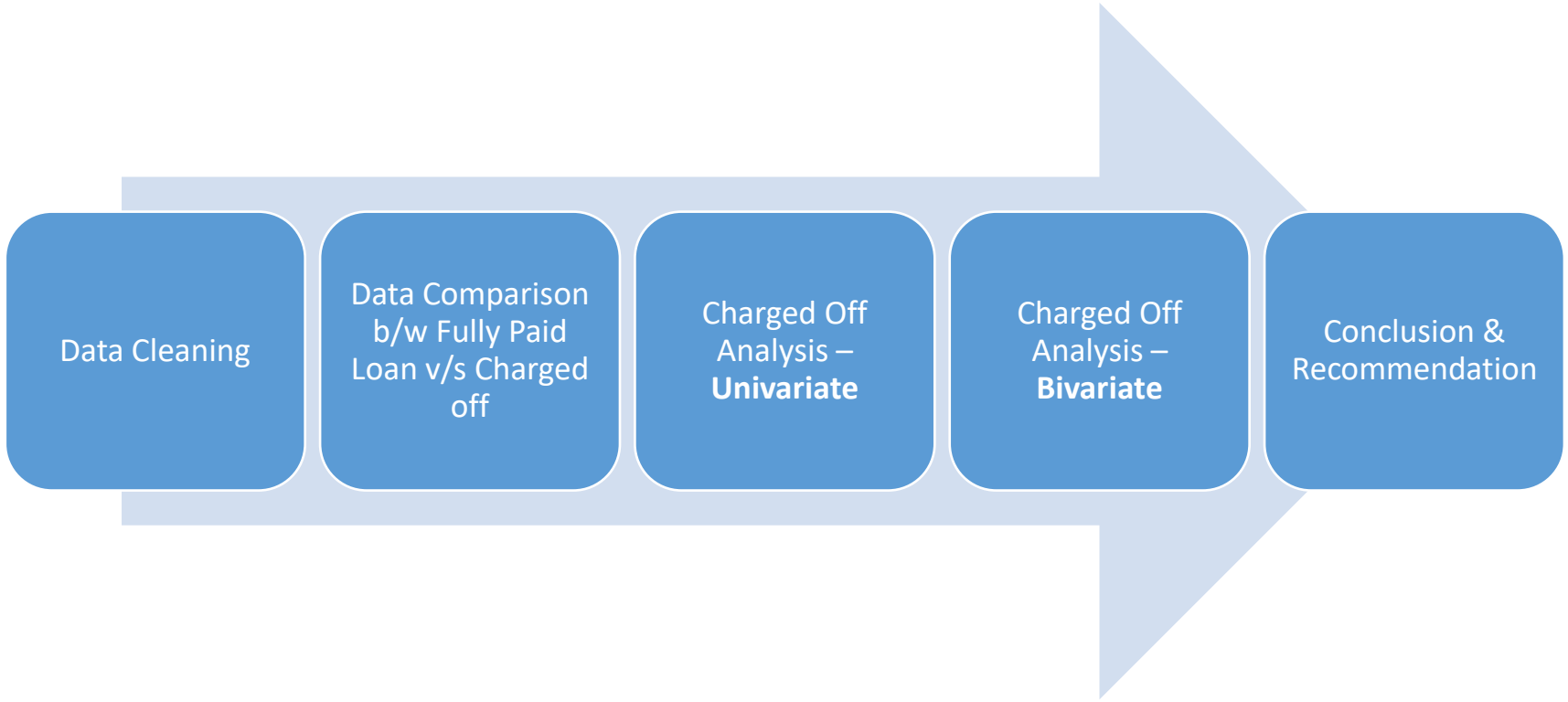
# Gramener Case Study

## Report

Group Members:
1.Kamal Kumar Kalra
2.Rishi Narang
3.Nitu Sinha
4.Neha Sharma

# Gramener Case Study

**Objective:** Dataset provided by in the Gramener case study contains the information about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. Further, understand the ‘Driving Factors’ or ‘Driver Variables’ behind Loan Default phenomena.



# Analysis Overview



**Step 1: Data Importing**

The “Loan.csv” dataset provided to us for the Gramener case study had 39717 rows & 111 columns to analyse.

**Step 2: Data Cleaning**

It was imperative to reduce the structure of the dataset for an effective study of the data. The steps we followed here have been mentioned below:

- In the ‘Loan’ dataframe: We checked for the duplicates in the dataset with ‘id’ column as the basis. There were no duplicates and all the 39717 rows were unique.
- Further we checked for the Nulls counts & percentages in the columns & on the basis of same, dropped 58 columns to make a new dataframe named. This reduced our dataset to 53 columns from 111 columns.
- We studied these 53 columns and analysed their significance through the data in these columns and also referred to the data dictionary to decipher the details. On the basis of our study and outside research, we dropped 13 more columns to create another dataframe which had 40 columns & 39717 rows.
- In the dataframe few columns required format corrections:
    - We inserted a new column 'interest_rate_percent' to get the data in the string formatted column ‘int_rate’ to float format.

# Analysis Overview



- Also, converted the formats for columns - interest_rate, issue_d_year, last_pymnt_d_year, earliest_cr_line_year, last_credit_pull_d_year – from string to datetime formats. Further, we extracted year from these columns & created 4 new columns to store the same.
- Inserted a new column ‘emp_tenure’ that contained the string formatted data of ‘emp_length’ in float format.
- Further, keeping in mind the Univariate analysis that we had planned to conduct, we inserted few columns that CATEGORISED few of the existing columns.
  - ‘emp_levels’ categorised the ‘emp_tenure’ into - Junior, Executive, Mid Level, Senior Level
  - 'interest_rate_levels' categorised the 'interest_rate_percent' into - Low_rate, Medium_rate, High_rate

**Step 3: Data Analysis**

We have conducted our analysis in the following order:

A) Overall analysis (Fully Paid + Charged Off)

B) Charged off analysis

B.1) Univariate Analysis

B.2) Bivariate Analysis

# Analysis Overview



**A) Overall Analysis (Fully Paid + Charged Off)**

Here, we have studied Fully paid loans (No Default loans) & Charged off loans (Default loans) together. This helped us in understanding the factors that were present in Fully paid loans but were missing in the Charged off loans. And hence, if these factors are considered before giving loans, then the situation of Defaulted loans will improve to a considerable extent for the Lending Club.

*Note: We have not studied the 'Current loans' as these are continuing loans and will not be a good study for this kind of analysis.*

# Analysis Overview



**For Overall analysis, we have employed the following techniques for smooth analysis:**

**1) Bin creation**: Bins have been created for the following variables: (the criteria of bin creation has been mentioned ahead with the plots)

- Annual Income: Categories - Very Low, Low, Average, Above Average, Medium, High, Very High
- Interest Rate: Categories- Low rate, Medium rate, High rate
- DTI (Debt to Income ratio): Low, Medium, High, Very High
- 'emp_levels' categorised the 'emp_tenure' into - Junior, Executive, Mid Level, Senior Level

**2) Ratio Analysis:** between the counts of Fully Paid & Charged off while studying the various variables.

- E.g., while studying the Annual Income Level factor, we calculated the ratio of - the count of Low Level Annual Income bin for the Fully Paid loans and the count of Low Level Annual Income bin for the Charged off loans. (Ratio (Fully Paid/Charged off)).
- This helped us in doing a minute study of all the variables that we studied. Also, the counts of No Default Loans were very high in comparison to the Default Loans. Ratio analysis helped us to accommodate this difference in counts as well.

**3) Correlation Analysis between various variables** – as Annual Income, Loan Amount, Interest Rate Percent, Instalment of Loans, DTI (loan to income ratio)

# Analysis Overview



**B) Charged Off  Analysis**

This segment solely concentrates on the loans that are in “Default”. We have done univariate and bivariate analysis for this section.

Here also we have employed various techniques for analysis which have been mentioned below.

**1) Bin creation:** Bins have been created for the following (the criteria of bin creation has been mentioned ahead with the plots)-

- Annual Income: Categories - Very Low, Low, Average, Above Average, Medium, High, Very High
- Interest Rate: Categories- Low rate, Medium rate, High rate
- DTI (Debt to Income ratio): Low, Medium, High, Very High
- ‘emp_levels’ categorised the ‘emp_tenure’ into - Junior, Executive, Mid Level, Senior Level

**2) Bar Graph based analysis:** We have extensively used bar graphs & histograms to study the Charged off data segment.

# Analysis Overview



**B) Charged Off  Analysis**

This segment solely concentrates on the loans that are in “Default”. We have done univariate and bivariate analysis for this section.

Here also we have employed various techniques for analysis which have been mentioned below.

**1) Bin creation:** Bins have been created for the following (the criteria of bin creation has been mentioned ahead with the plots)-

- Annual Income: Categories - Very Low, Low, Average, Above Average, Medium, High, Very High
- Interest Rate: Categories- Low rate, Medium rate, High rate
- DTI (Debt to Income ratio): Low, Medium, High, Very High
- ‘emp_levels’ categorised the ‘emp_tenure’ into - Junior, Executive, Mid Level, Senior Level

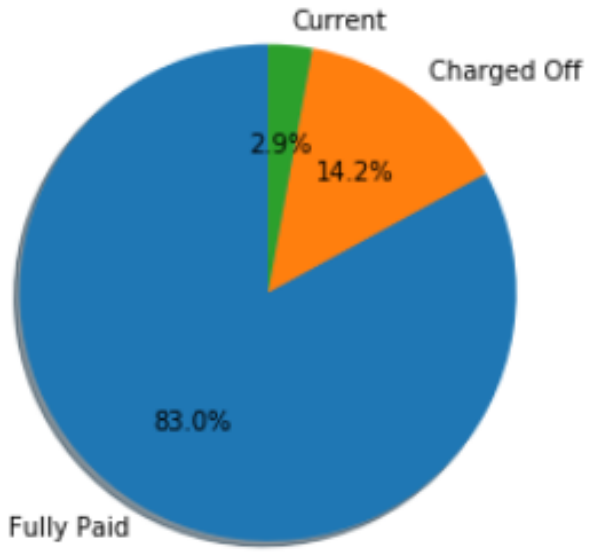
**2) Bar Graph based analysis:** We have extensively used bar graphs & histograms to study the Charged off data segment.

# Key Driver variables identified for analysis

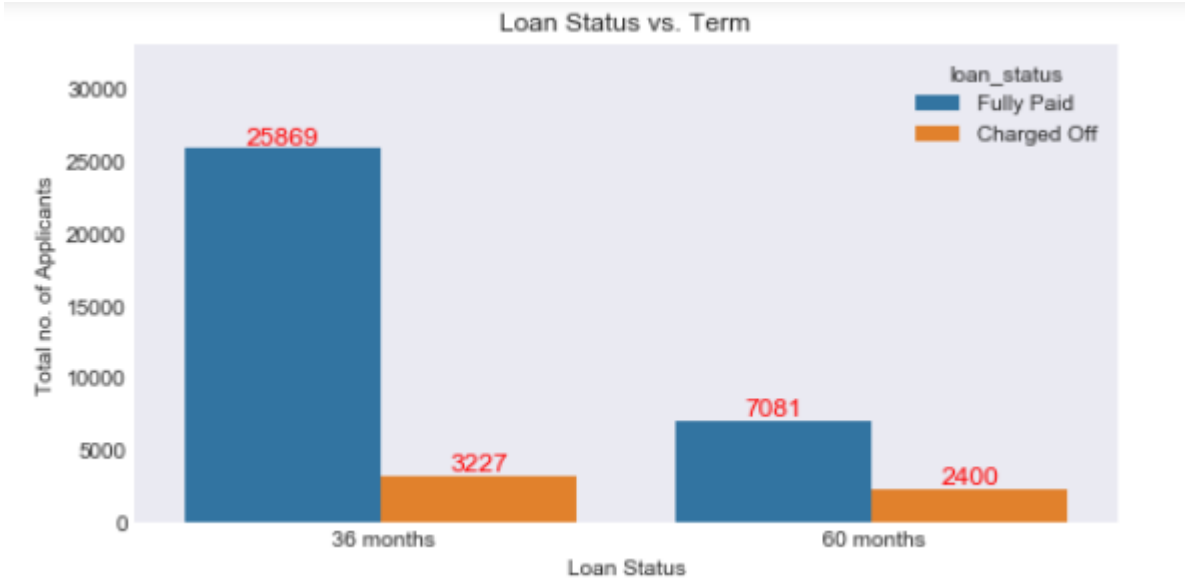
- Following are the key driver variables identified for the charged off ( defaulter) analysis:
  - Address state of application (addr_str)
  - Employment Length of applicant (emp_length) -Various levels are derived for this “emp_length” variable
  - Home owner status (home_ownership)
  - Annual Income of applicants (annual_inc) - Various levels are derived for this “annual_inc” variable
  - Interest Rate (int_rate) - Various levels are derived for this “annual_inc” variable
  - Verification Status (varification_status)
  - Purpose of loan (purpos)
  - dti
  - Loan term distribution (term)
  - Grade (grade) and Sub grade (sub_grade)
  - Delinq (delinq_2yrs)
- Outlier Treatment: Since there are only few outliers for annual_inc parameter, we have not removed them for the analysis. Instead, we have created various segments (bins) to overcome this.

Overall analysis (Fully Paid + Charged Off)

# Plots – Overall Loan & Loan Status Vs Term




**Plot 1**

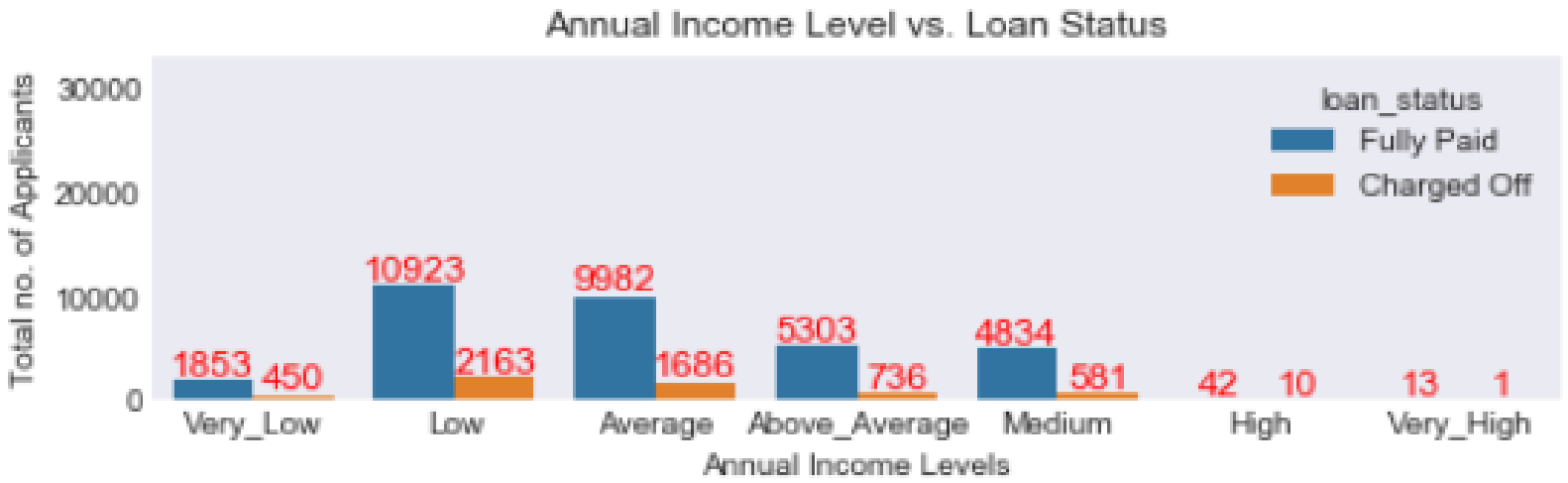



**Plot 2**

• Plot 1 depicts overall distribution of loan status. 83% population has paid loan fully , 2.9% are still in process of paying loans and 14.2% are the charged off (defaulters) who have not paid their loans.

• Plot 2 depicts overall loan status Vs Term. From the plot 2 is evident that the loan granting for the period (term) of 36 month is better than the 60months. **Thus granting loan for longer period is problematic.**

# Plots – Annual Income Level Vs Loan Status




**Plot 3**

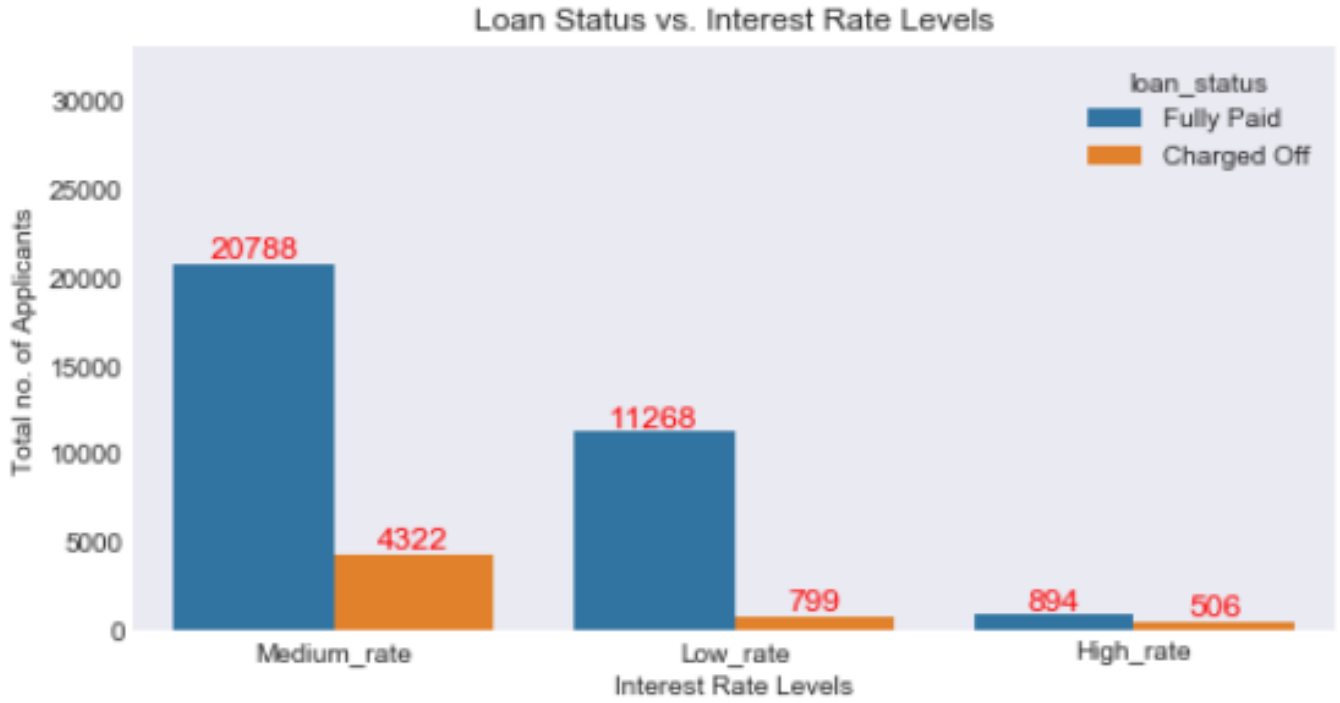
**Annual Income Slab**

| | |
|---|---|
| Very Low | : < 25000 |
| Low | : 25000 to 50000 |
| Average | : 50000 to 75000 |
| Above Average | : 75000 to 100000 |
| Medium | : 100000 to 500000 |
| High | : 500000 to 1000000 |
| Very High | : > 1000000 |

- Plot 3 depicts Annual Income Level vs Loan Status. The ratio of Medium income slab is a problematic one as the ratio of this slab is more as compared to others. However, ratio of Very High income slab is more but this is due to the reason that it consist some outlier data.
- With increase in income, instances of fraud (defaults) are less

| Annual Income Level | Ratio (Fully Paid/Charged off) |
|---|---|
| Very Low | 4.117778 |
| Low | 5.049931 |
| Average | 5.920522 |
| Above Average | 7.205163 |
| Medium | 8.320138 |
| High | 4.2 |
| Very High | 13 |

# Plots – Loan Status Vs Interest Rate Levels




**Plot 4**

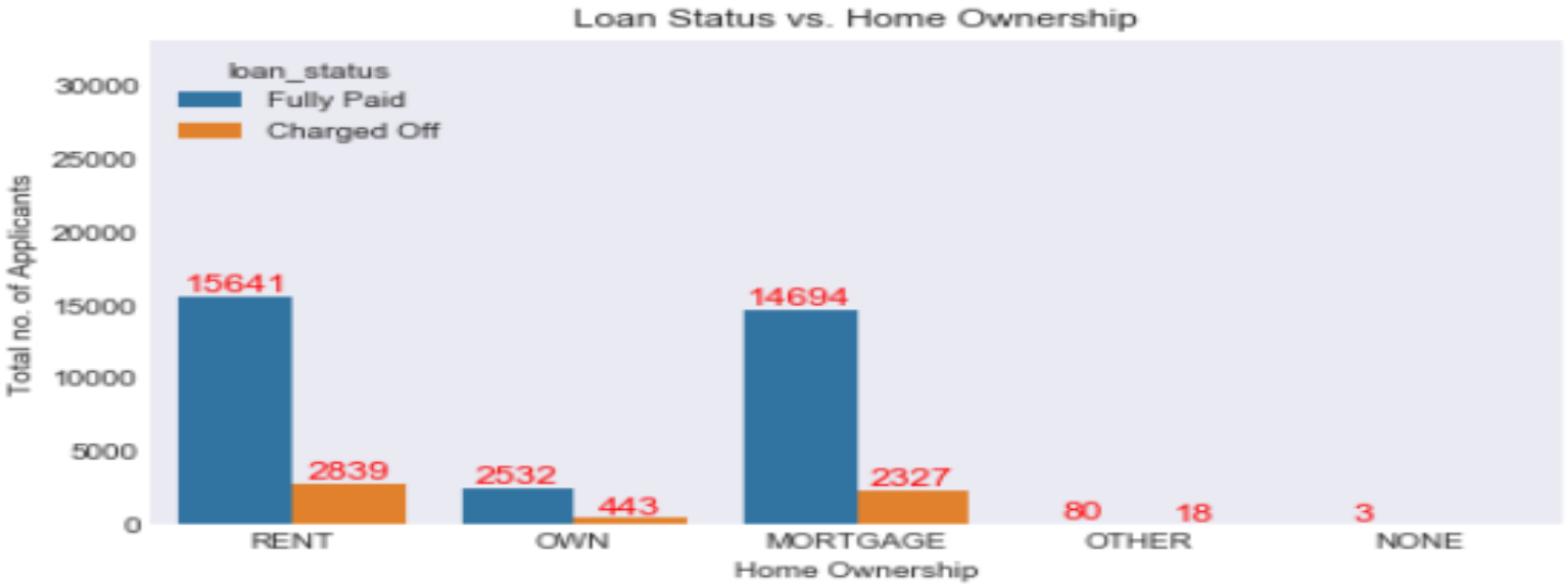
**Interest Rate Slab (%)**

| | |
|---|---|
| Low Rate | : < 10 |
| Medium Rate | : 10 to 19 |
| High Rate | : >20 |

• Plot 4 depicts that the ratio of fully paid loan vs charged off is high for low rates as compared to medium rate and high rate.

| Interest Rate Levels | Ratio (Fully Paid/Charged off) |
|---|---|
| Medium Rate | 4.80981 |
| Low Rate | 14.10263 |
| High Rate | 1.766798 |

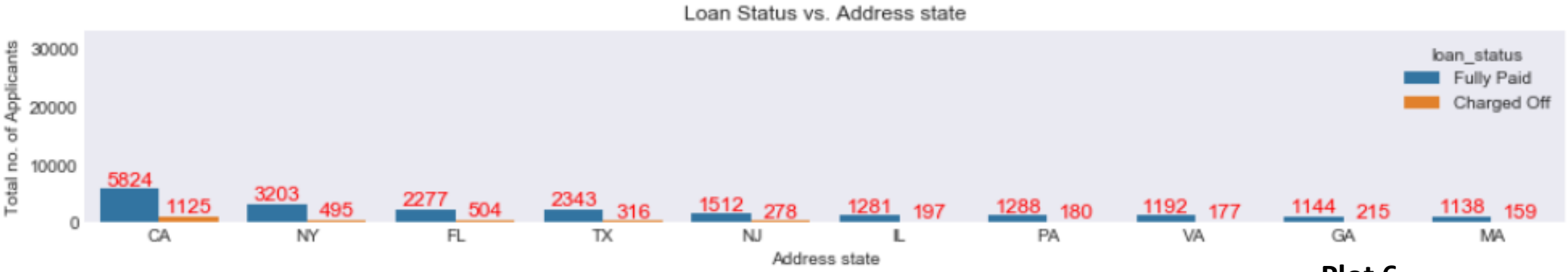
# Plot – Loan Status Vs Home Ownership



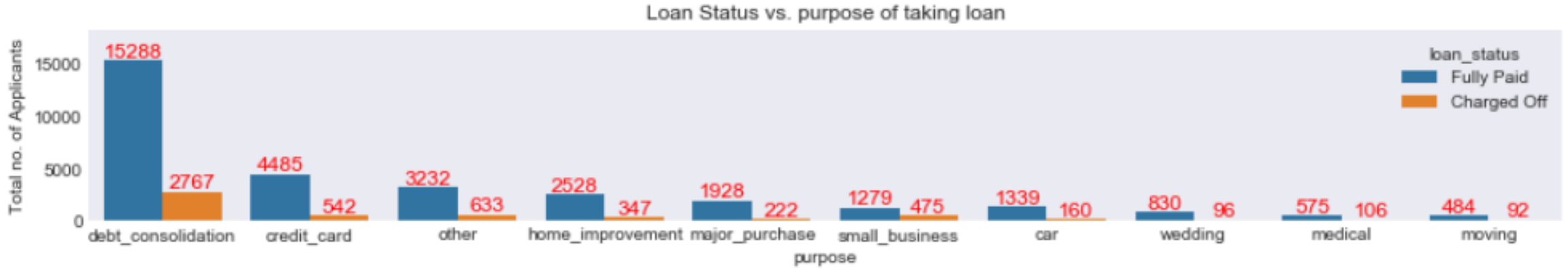

**Plot 5**

• Plot 5 depicts that the ratio of fully paid loan vs charged off is high for Mortgage followed by Rent . Hence, ratio wise these two areas are problematic in terms of loan dispersed.

# Plots – Loan Status Vs (Address State or Purpose)




**Plot 6**



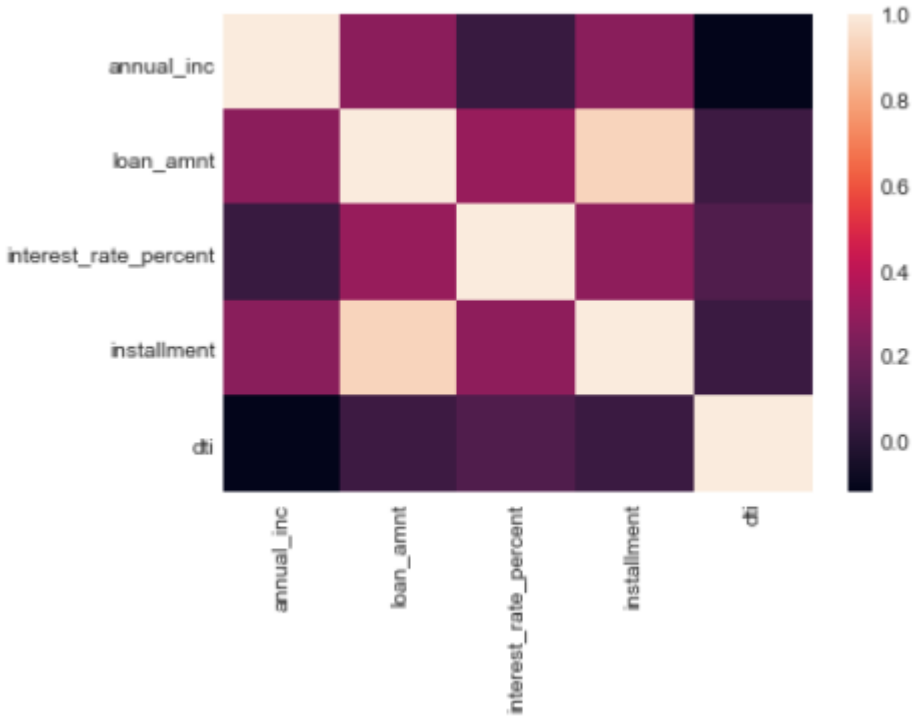

**Plot 7**

- Plot 6  The maximum defaulters are from CA.
- Plot 7 depicts that the applicants are taking loan for the purpose of debt_consolidation in both fully paid and charged off categories.

# Plots – Correlation Plots

| | annual_inc | loan_amnt | interest_rate_percent | installment | dti |
|---|---|---|---|---|---|
| annual_inc | 1 | 0.27 | 0.049 | 0.27 | -0.12 |
| loan_amnt | 0.27 | 1 | 0.3 | 0.93 | 0.062 |
| interest_rate_percent | 0.049 | 0.3 | 1 | 0.28 | 0.11 |
| installment | 0.27 | 0.93 | 0.28 | 1 | 0.052 |
| dti | -0.12 | 0.062 | 0.11 | 0.052 | 1 |

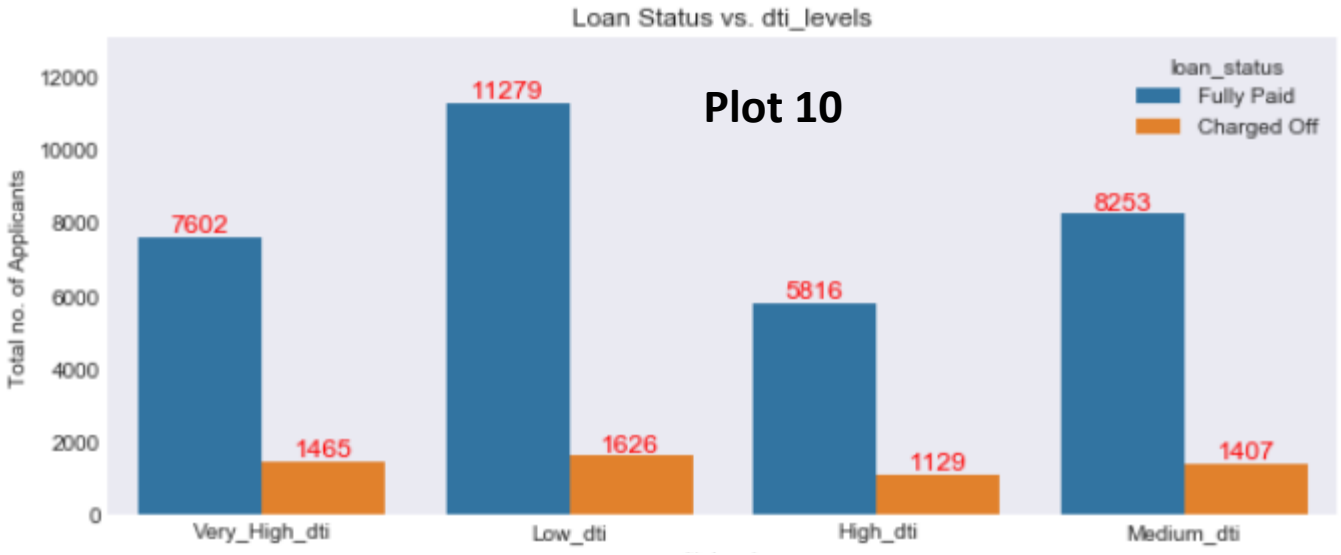
**Plot 8**




**Plot 9**

- Plot 8 depicts that there is high correlation between instalment and loan amount.
- Plot 8 also depicts that there is very low correlation between  dti vs other parameters.

# Plots – (Loan Status or Annual Inc Lever )Vs dti levels

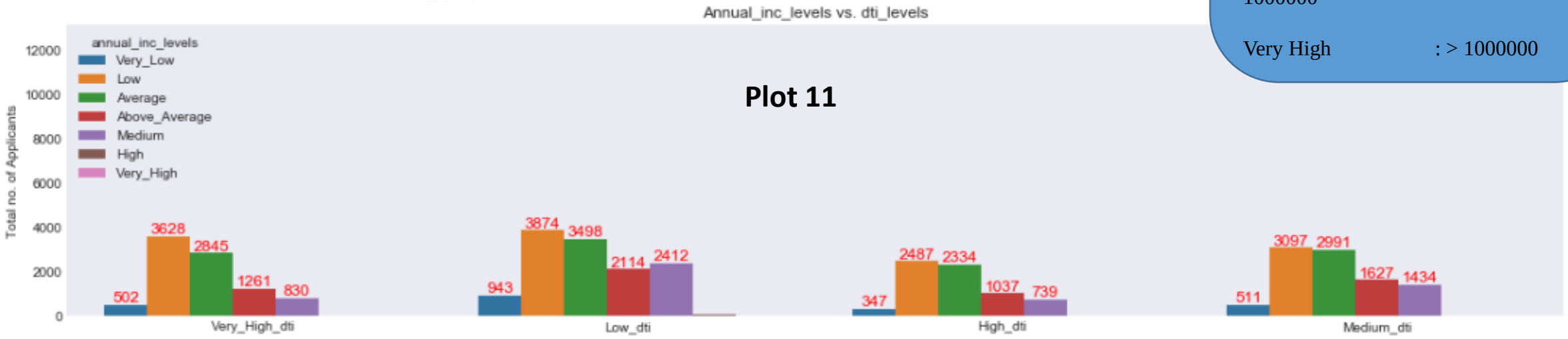



**dti Slab (number)**

| | |
|---|---|
| Low | : 0 to 9 |
| Medium | : 10 to 15 |
| High | : 16 to 20 |
| Very High | : >20 |

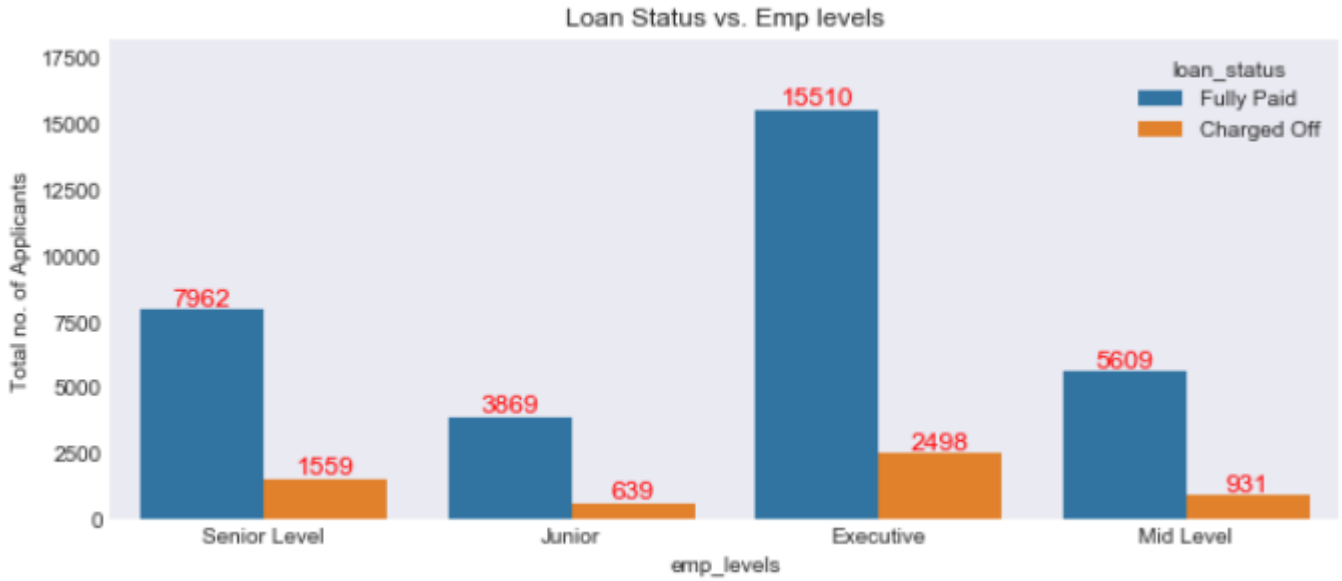
**Annual Income Slab**

| | |
|---|---|
| Very Low | : < 25000 |
| Low | : 25000 to 50000 |
| Average | : 50000 to 75000 |
| Above Average | : 75000 to 100000 |
| Medium | : 100000 to 500000 |
| High | : 500000 to 1000000 |
| Very High | : > 1000000 |




- Plot 10 depicts that applicants having very high dti have higher chance of being charged off
- Plot 11 applicants having low and average income have very high dti

# Plots – Loan Status Vs Employment Levels




**Plot 12**

**Employment length**

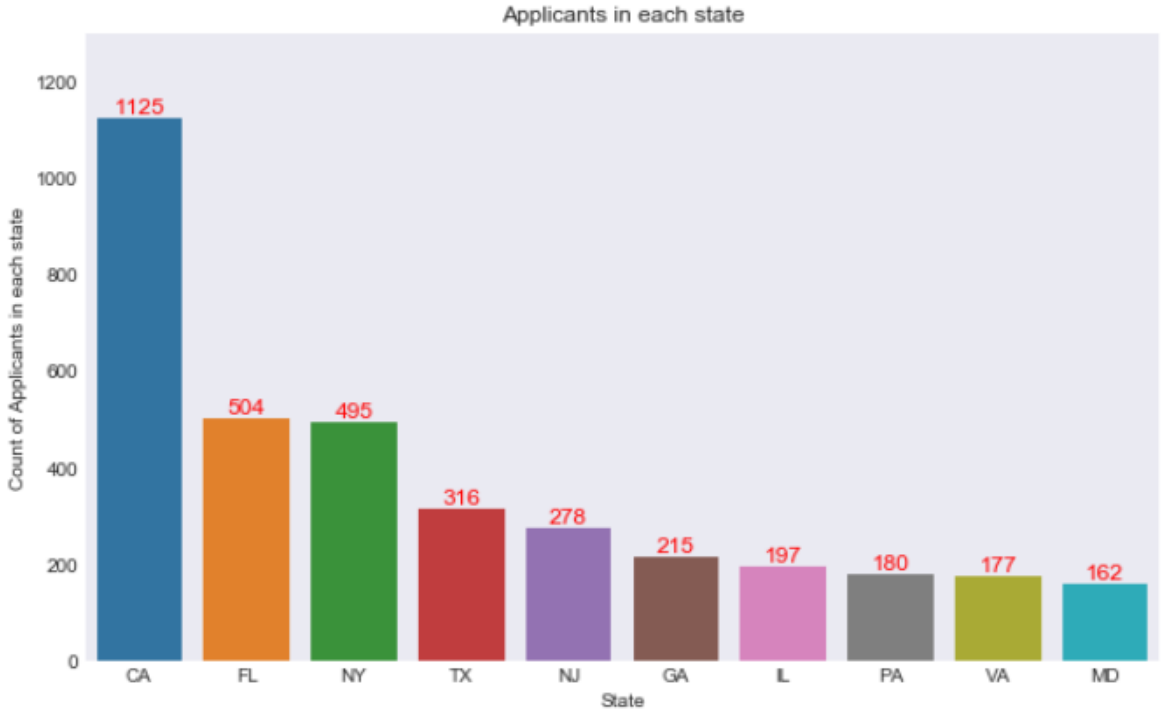
| | |
|---|---|
| Junior | : <1 |
| Executive | : 1 to 5 |
| Mid Level | : 6 to 9 |
| Senior Level | : >10 |

• Plot 12 depicts that ratio of fully paid loan vs charged off is high for applicants working in executive level followed by junior level.

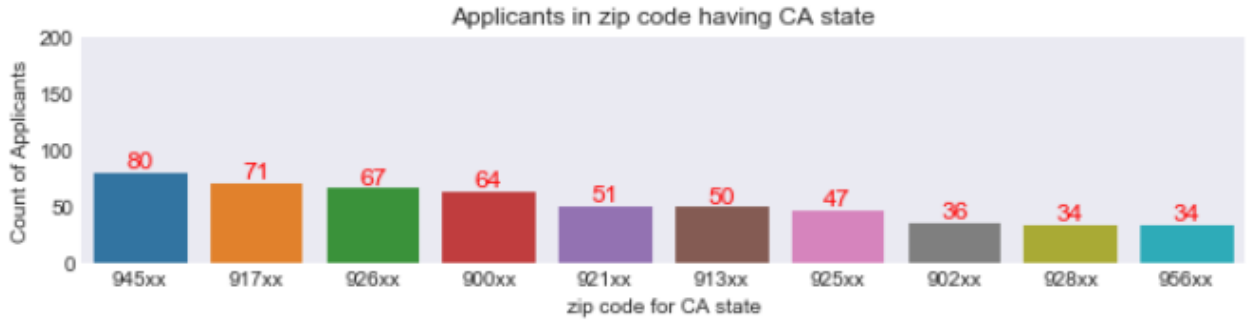
# Charged Off Analysis

# Univariate Analysis

# Plots – Top 10 States Vs Count of Applicants




**Plot 13**

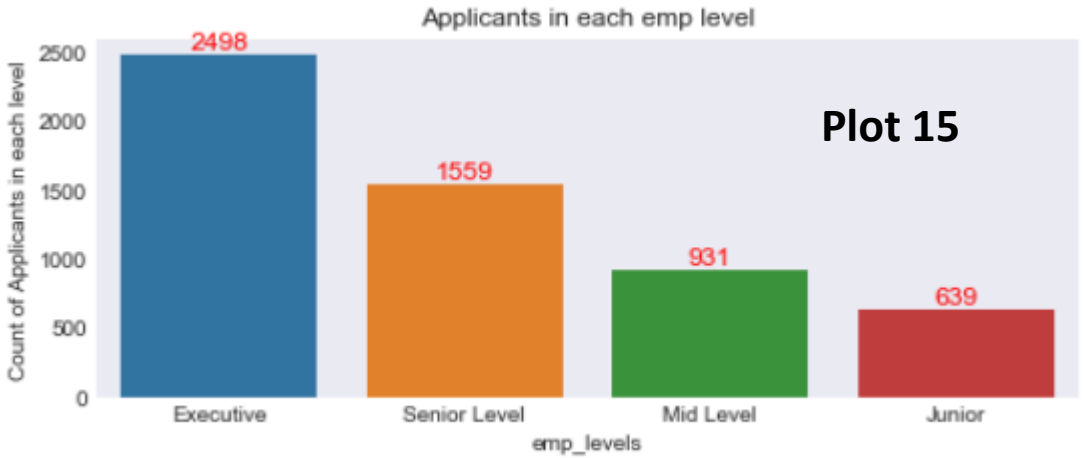



**Plot 14**

• Plot 13 depicts that applicants living in CA (California) have the highest number of defaulter followed by FL (Florida) and NY (New York)

•Plot 14 depicts that applicants living in 945xx have the highest number of defaulter followed by 917xx and 926xx.

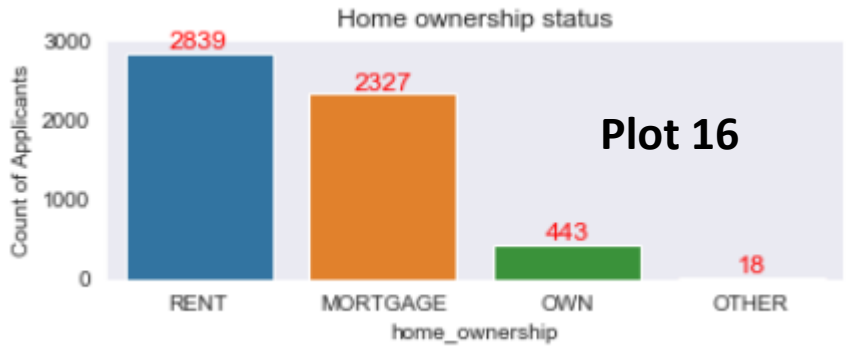
# Plots – Employment Level and Home Ownership status




**Plot 15**

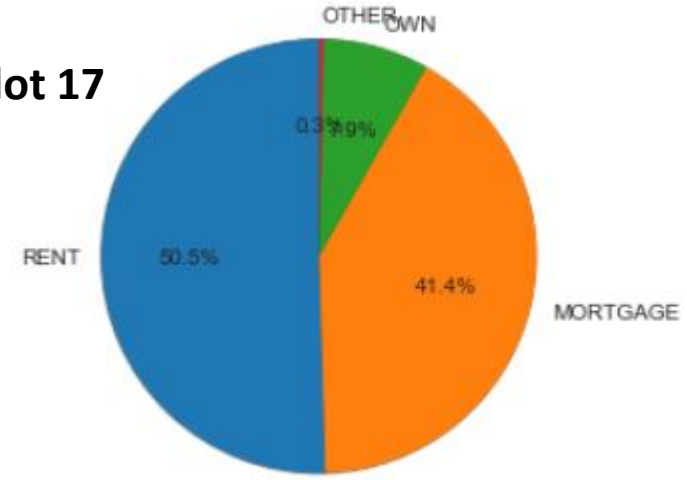
**Employment length**

| | |
|---|---|
| Junior | : <1 |
| Executive | : 1 to 5 |
| Mid Lever | : 6 to 9 |
| Senior Level | : >10 |



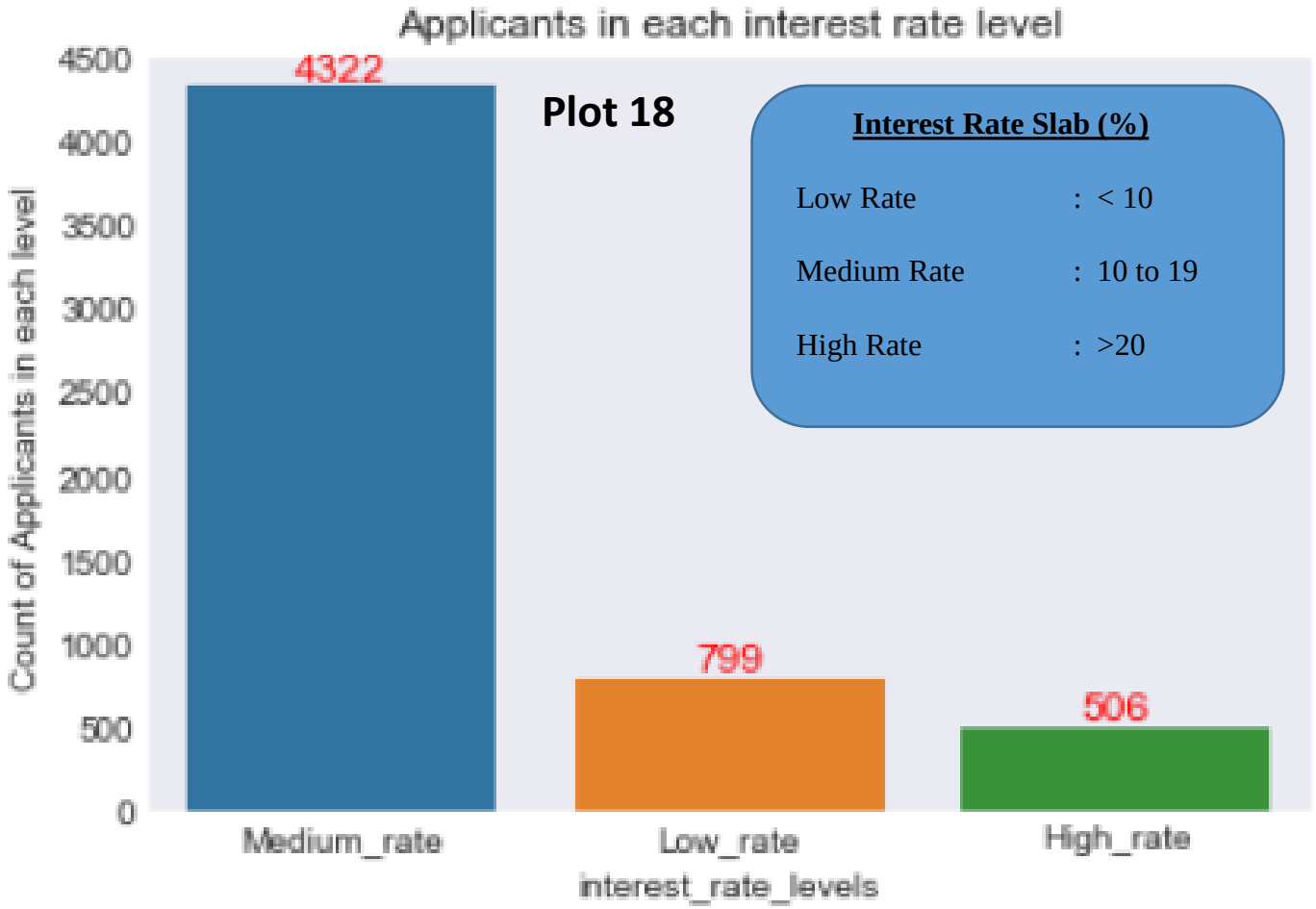

**Plot 16**

**Plot 17**




- Plot 15 depicts that executive level applicants have higher charged off number followed by senior level applicants.
- Plots 16 and 17 depict that applicants having rented and mortgage status have highest default counts with ratio of 50 .5% and 41.4 percent respectively.

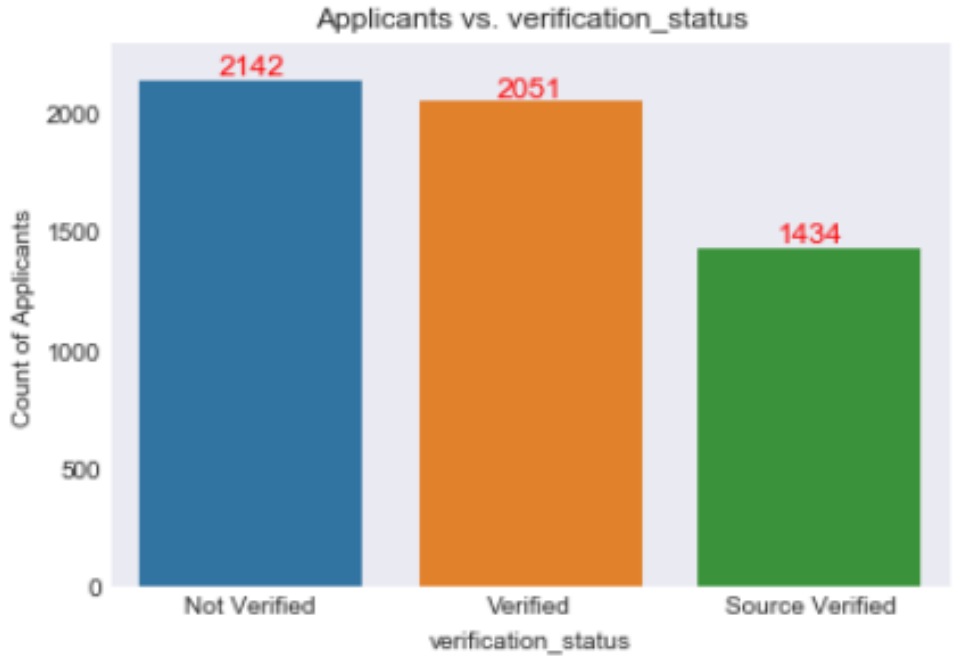
# Plots – Interest Rate and Verification status




**Plot 18**

**Interest Rate Slab (%)**

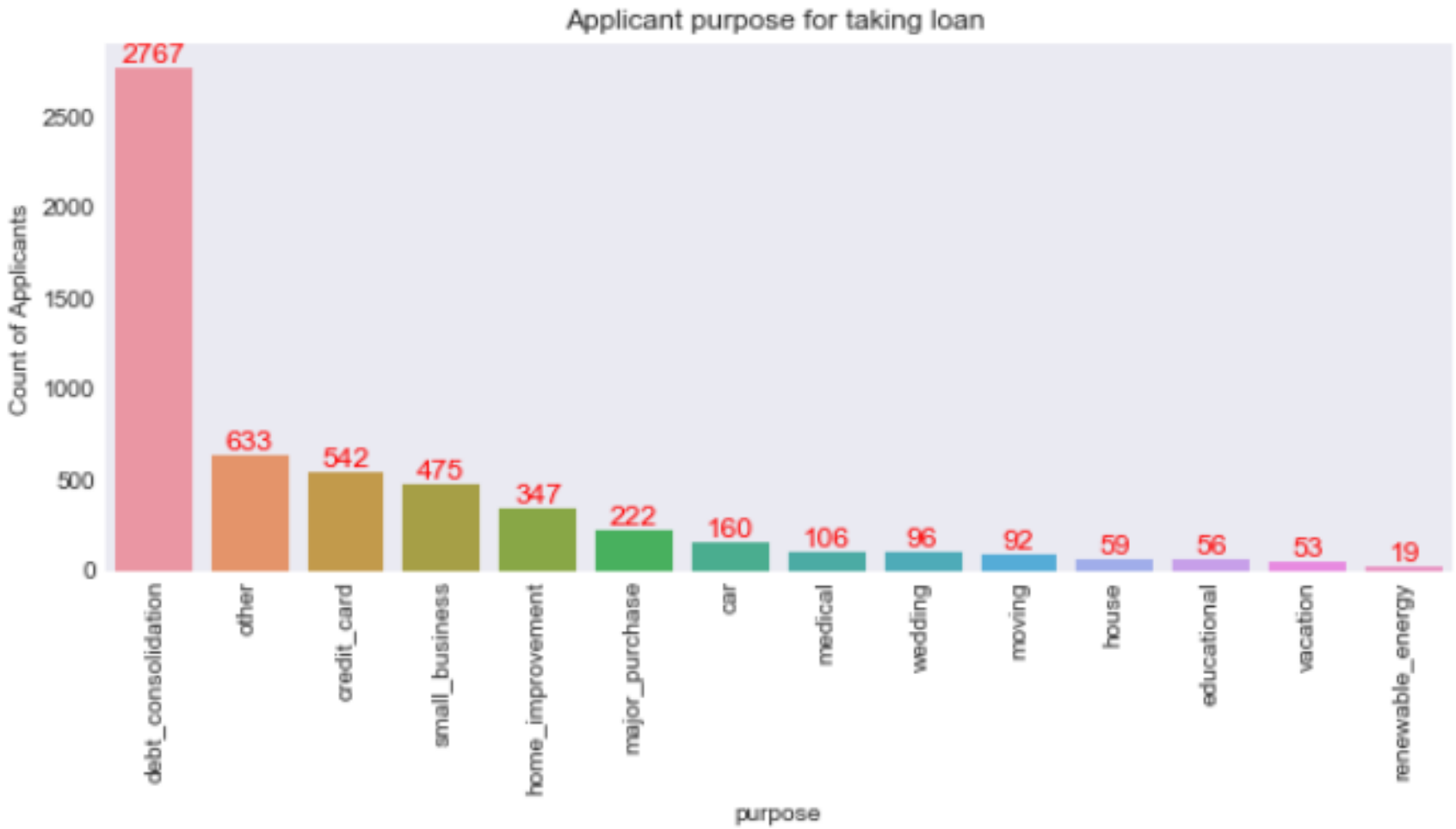
| | |
|---|---|
| Low Rate | : < 10 |
| Medium Rate | : 10 to 19 |
| High Rate | : >20 |




**Plot 19**

- Plot 18 depicts that applicants received interest rate of 10 to 19 (Medium rate) have higher count of defaulter (charged off)
- Plots 19 depicts that Non Verified applicants have higher chance of being as a defaulter.

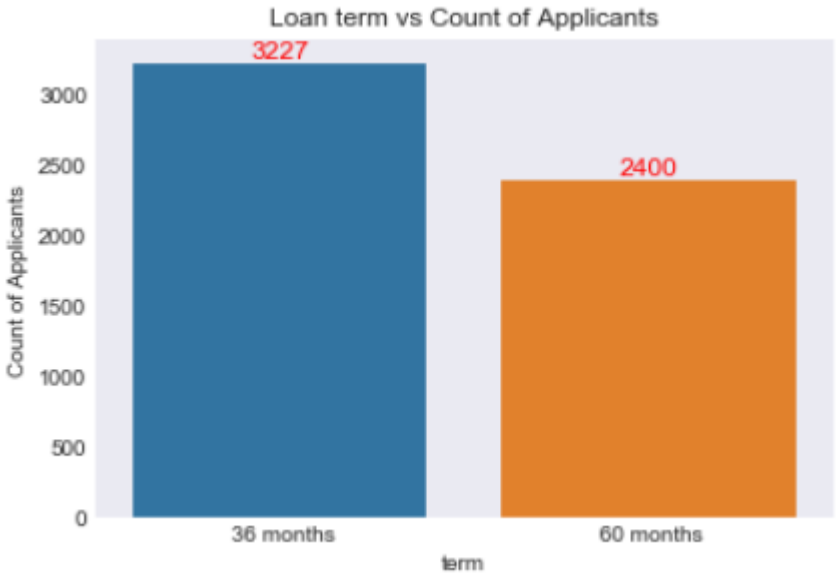
# Plots – Purpose of Loan vs Applicant Count



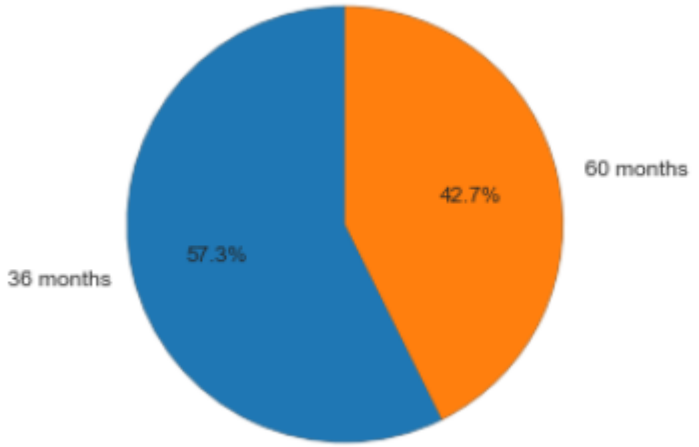

**Plot 20**

• Plot 20 depicts that main problematic purpose of loan among the candidates is debt_consolidation category following other and credit_card.

# Plots – Loan term distribution



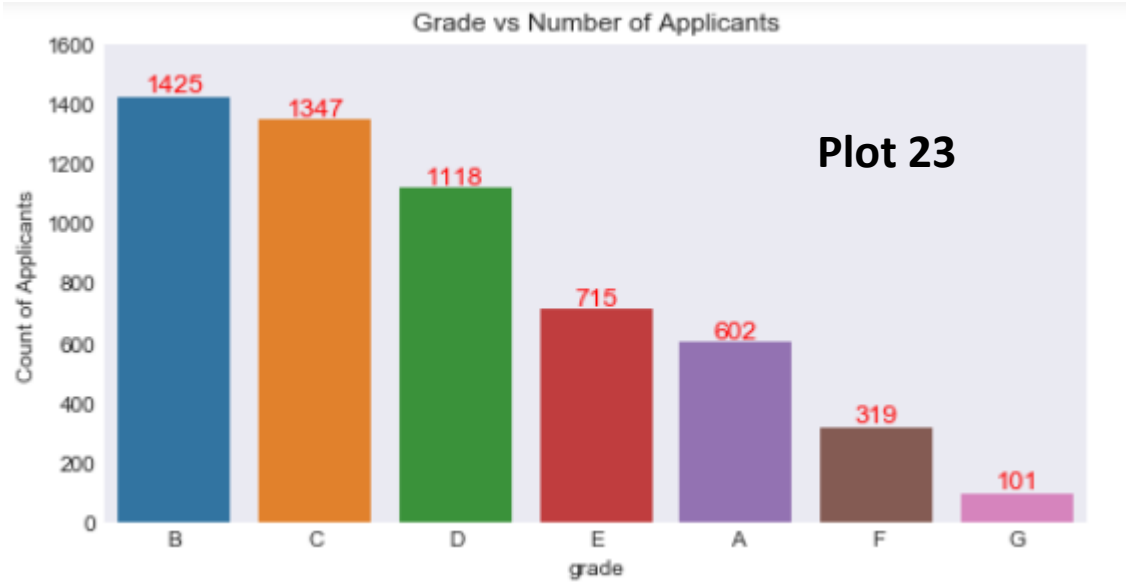

**Plot 21**



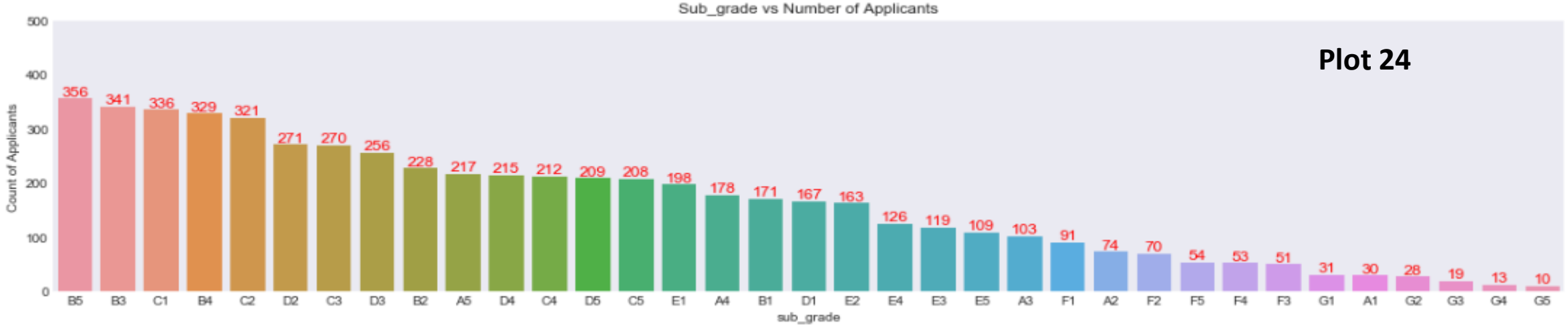

**Plot 22**

- Plots 21 and 22 depicts that loan term of 36 months have higher charged off rate i.e. 57.3%

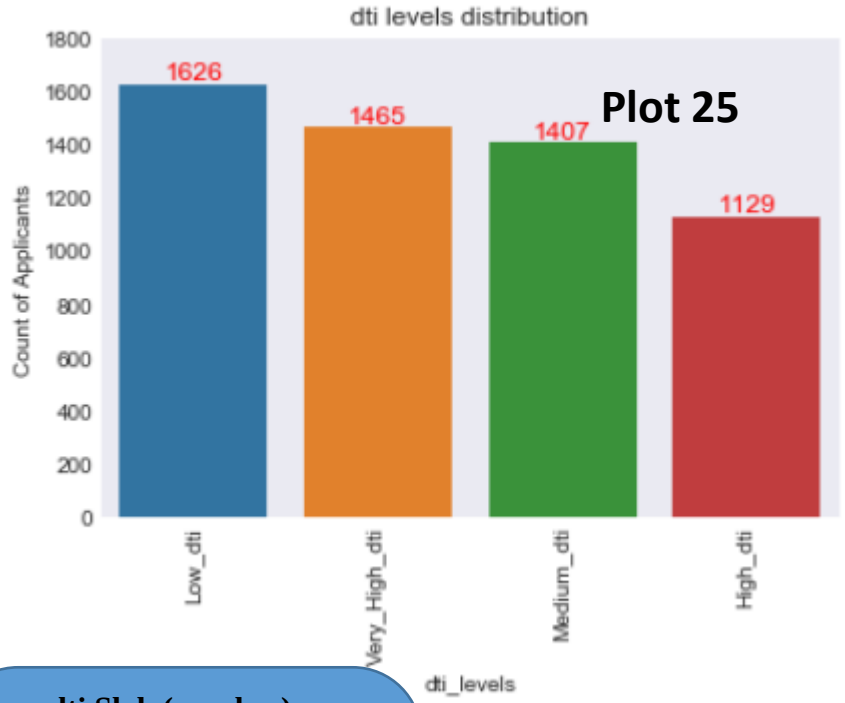
# Plots – Grade and Sub-Grade Distribution

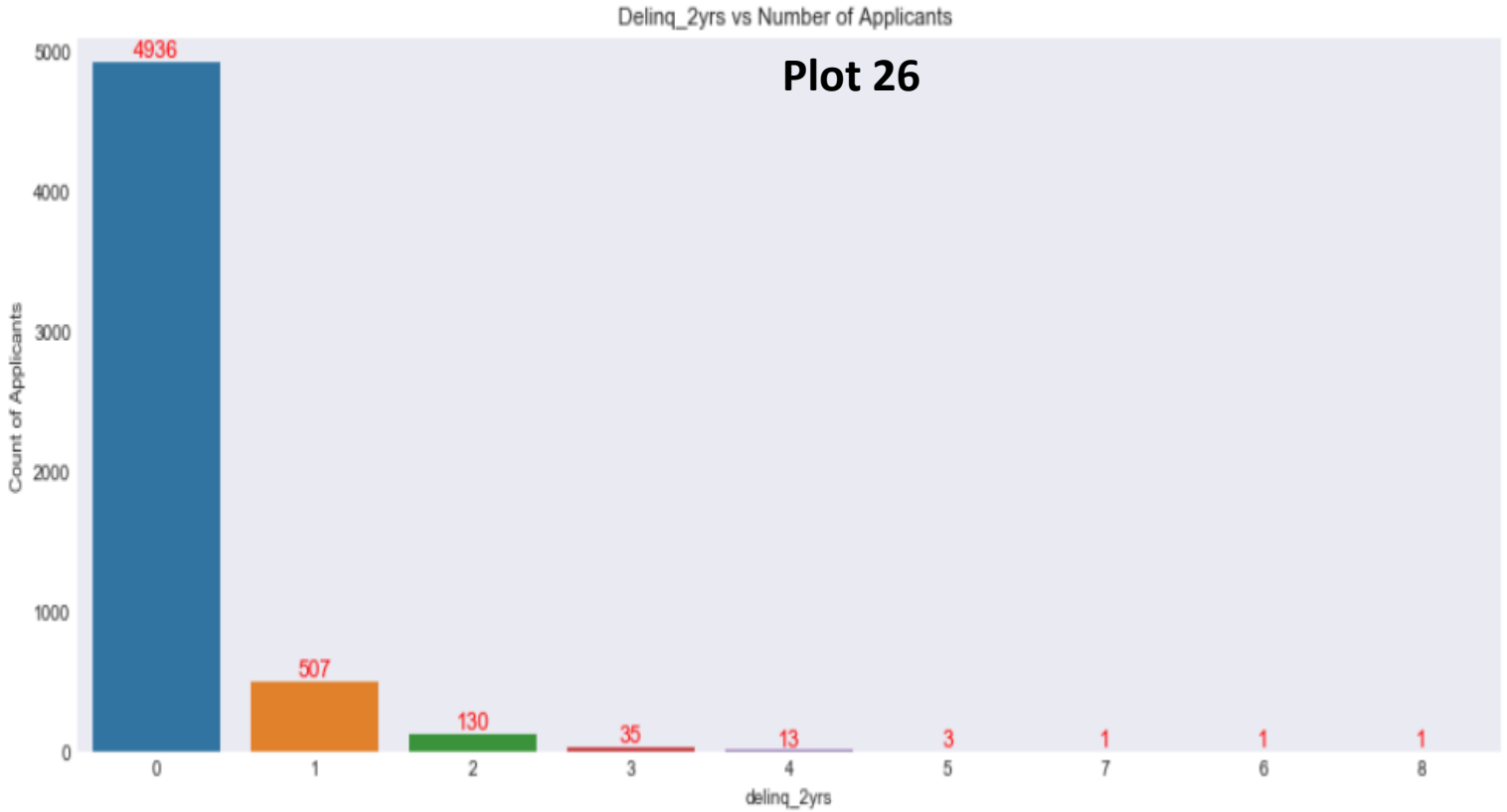



- Plots 23 depicts that Grade B and C have higher defaulter rate (Charged off)
- Plot 24 depicts that Under B Grade, B5 and B3 are the prime sub grade having highest default grade followed by C5 sub grade under C grade



# Plots – dti level Distribution and Delinq Distribution



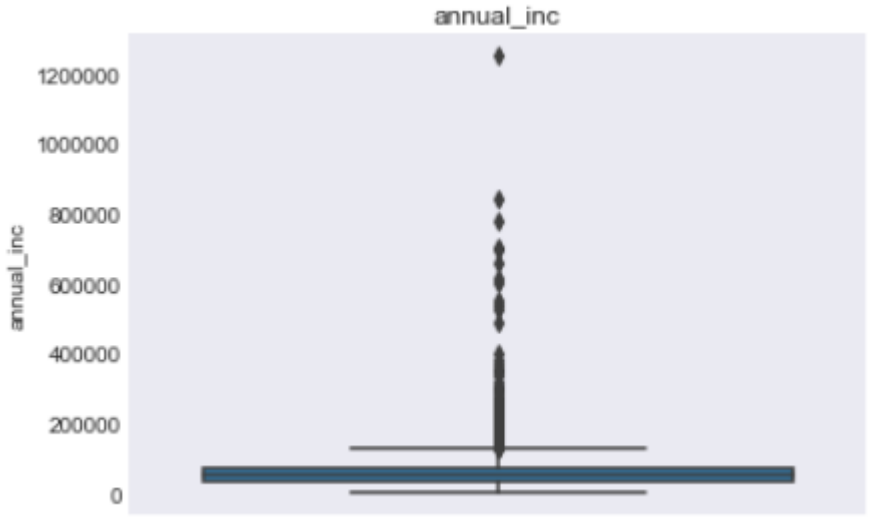




**dti Slab (number)**

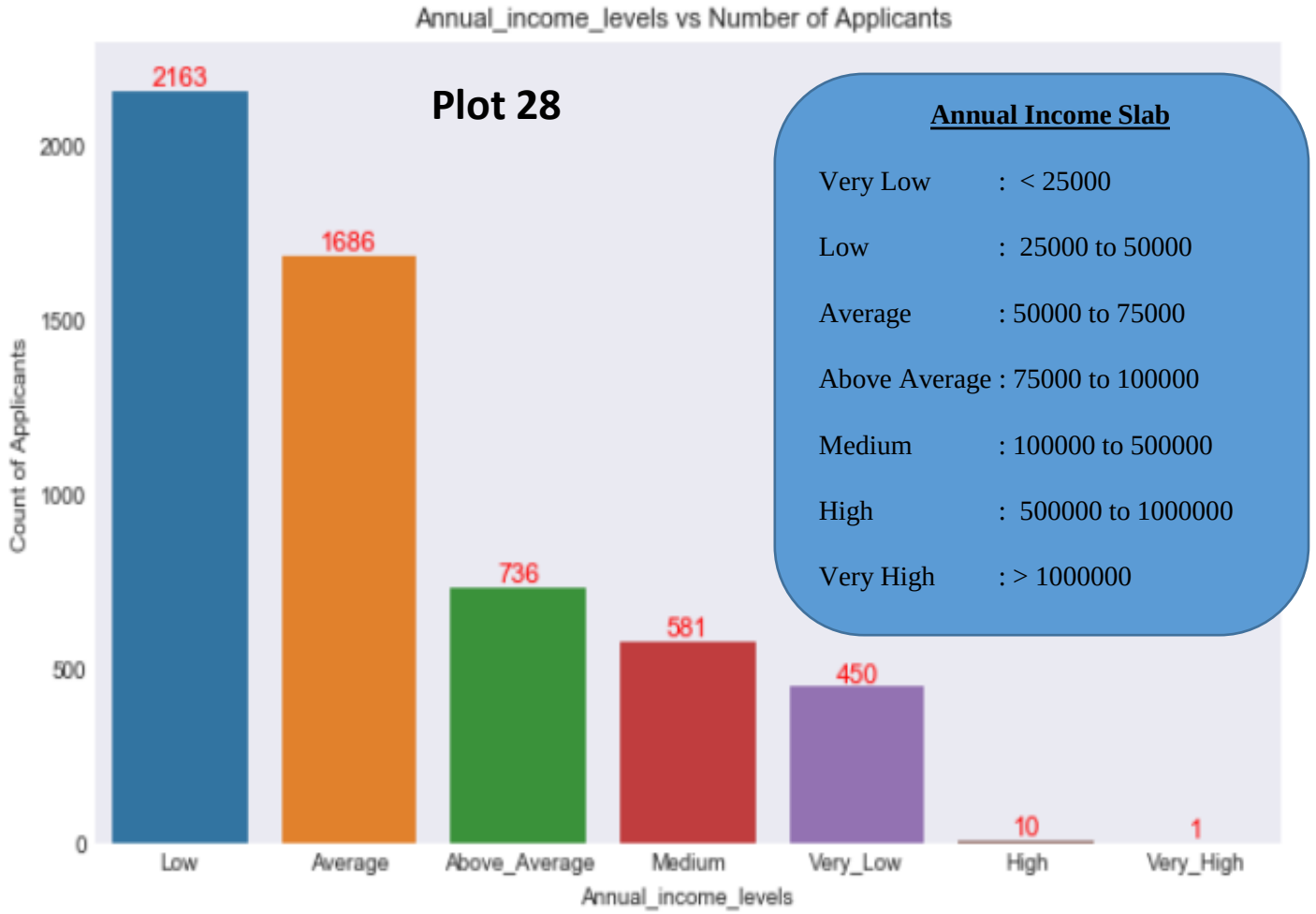
| | |
|---|---|
| Low | : 0 to 9 |
| Medium | : 10 to 15 |
| High | : 16 to 20 |
| Very High | : >20 |

- Plot 25 depicts that Low dti has higher number of defaulter (Charged off)
- Plot 26 depicts that maximum number of applicants (4936) have no criminal record.

# Plots – Annual Income Distribution and Outlier Analysis




**Plot 27**




**Plot 28**

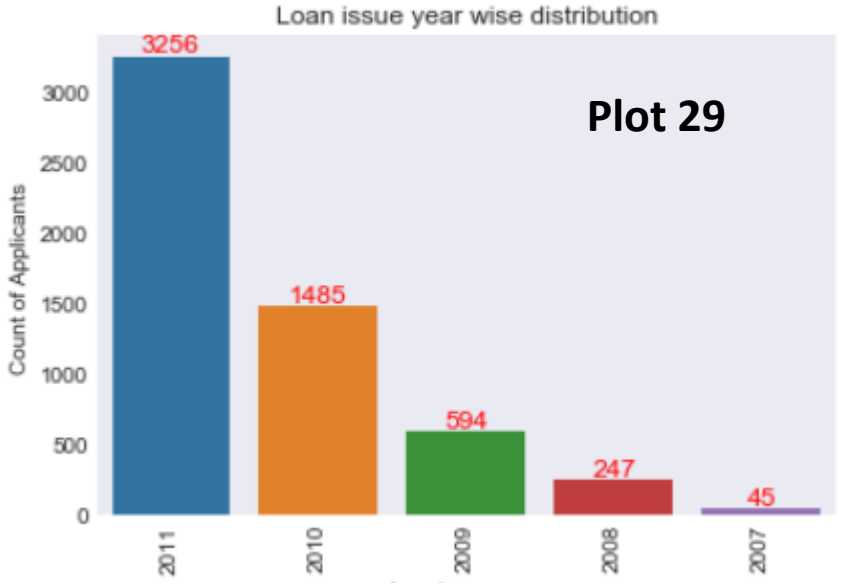
**Annual Income Slab**

| | |
|---|---|
| Very Low | : < 25000 |
| Low | : 25000 to 50000 |
| Average | : 50000 to 75000 |
| Above Average | : 75000 to 100000 |
| Medium | : 100000 to 500000 |
| High | : 500000 to 1000000 |
| Very High | : > 1000000 |

• Plot 27 depicts that there is an outlier in the charged off data set in the annual income. However, to get outlier into our analysis we have divided the annual income into slab and this will not affect our analysis.

•Plot 28 depicts that charged off count is very large in case of low income applicants followed by Average income applicants

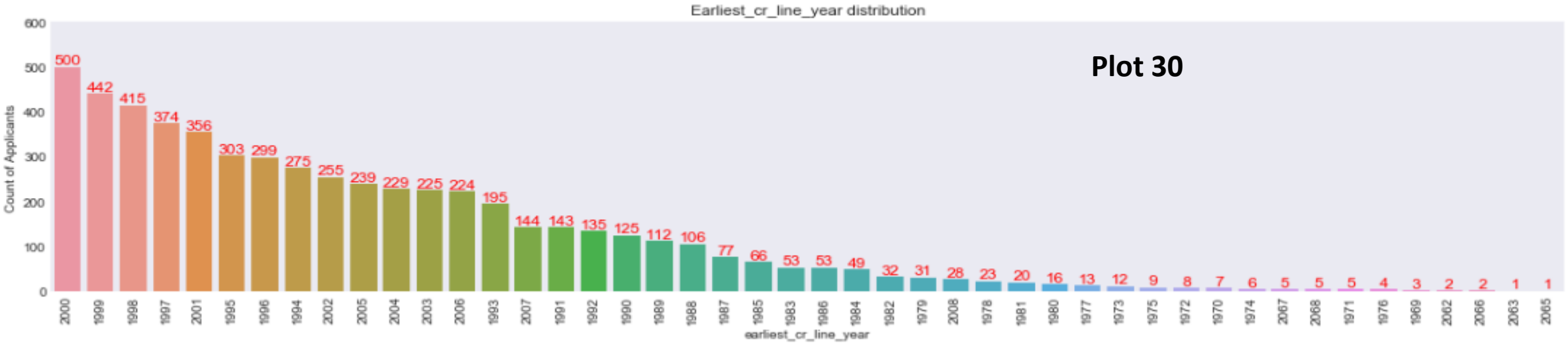
# Plots – Loan Issue Year and Credit Line Year




**Plot 29**

• Plot 29 depicts that 2011 is the high loan dispersed year in the last five years (2007-2011) and has highest defaulter (charged off)

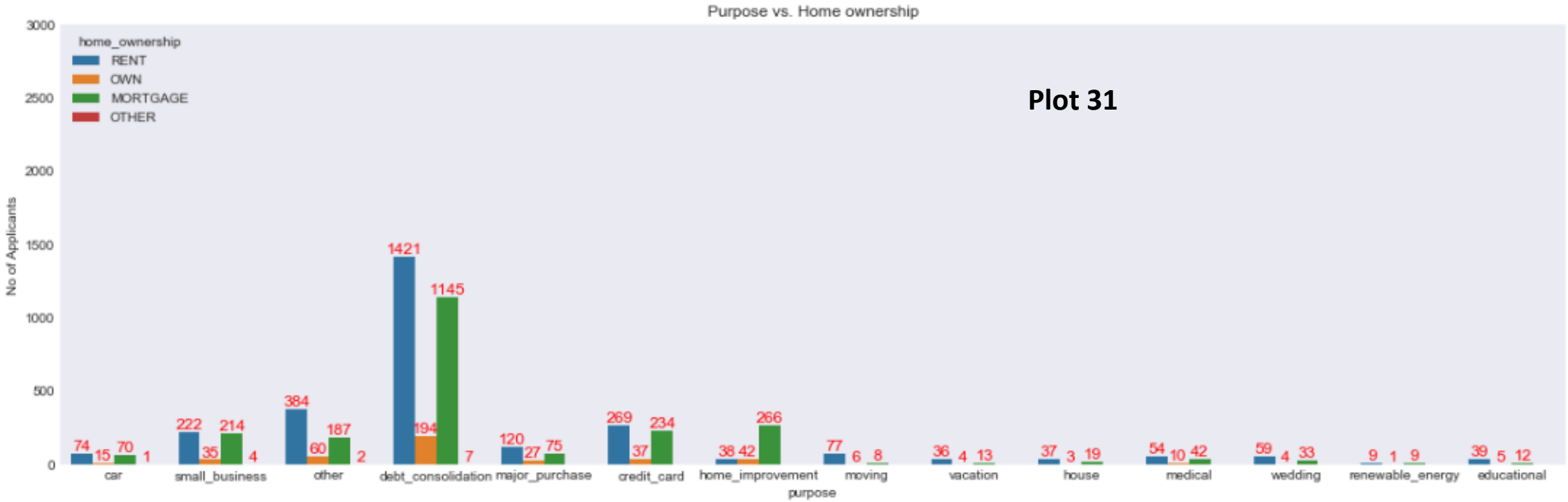
• Plot 28 depicts that from the years 1997 to 2000 maximum defaulter count has been seen.




**Plot 30**

# Bivariate Analysis

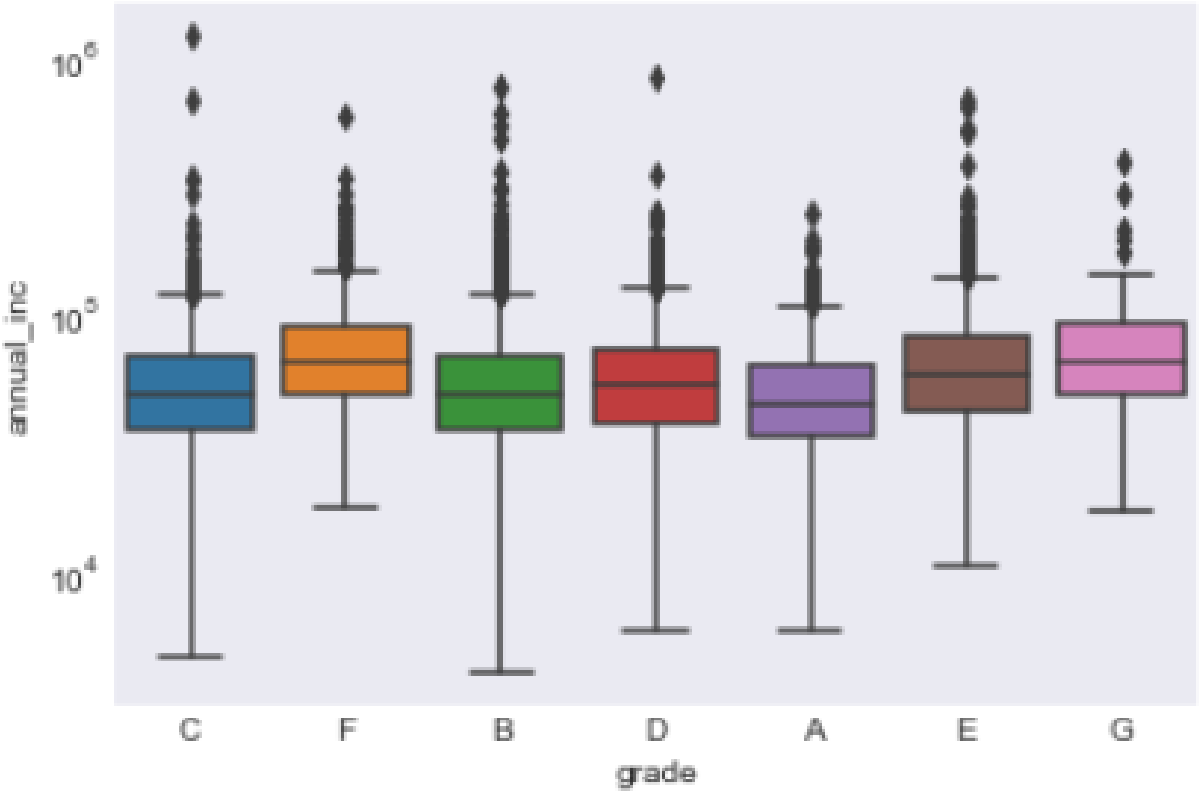
# Plots – Purpose vs Home Ownership




**Plot 31**

• Plot 31 depicts that  most of the applicants getting defaulter (charged off) by taking loan in “dept_cosolidation” especially rented applicants followed by Mortgage applicants.

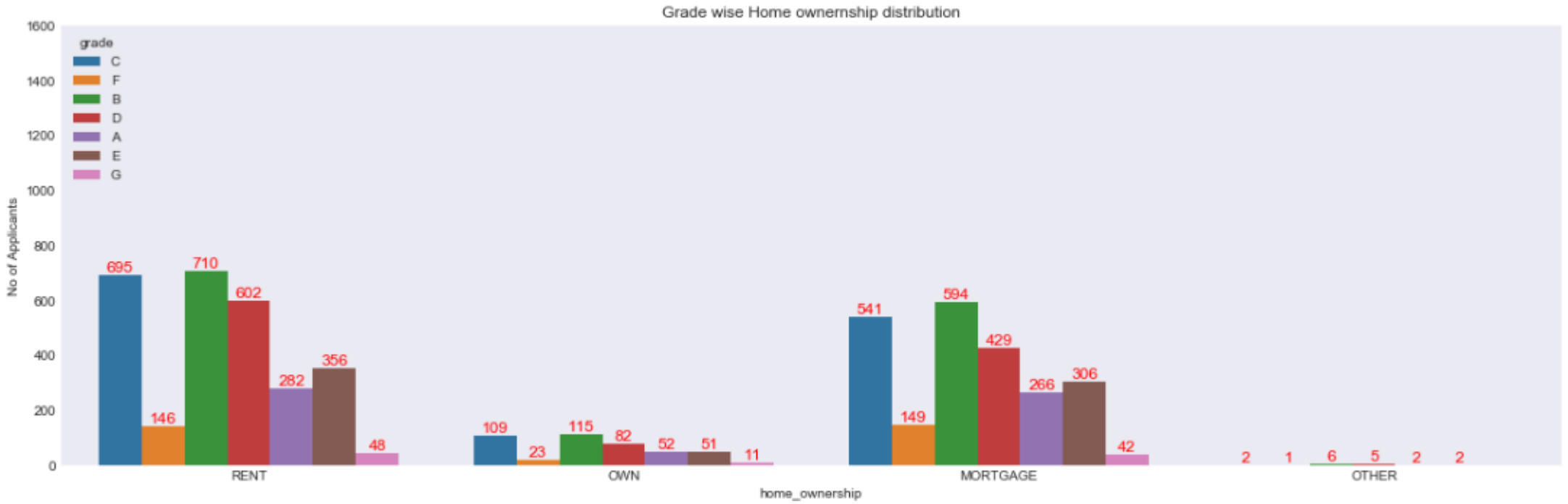
# Plots – Annual Income Vs Grade Distribution




**Plot 32**

- Plot 32 depicts that  Grade F has maximum annual income followed by Grade G & Grade E
- Outlier is maximum in Grade C

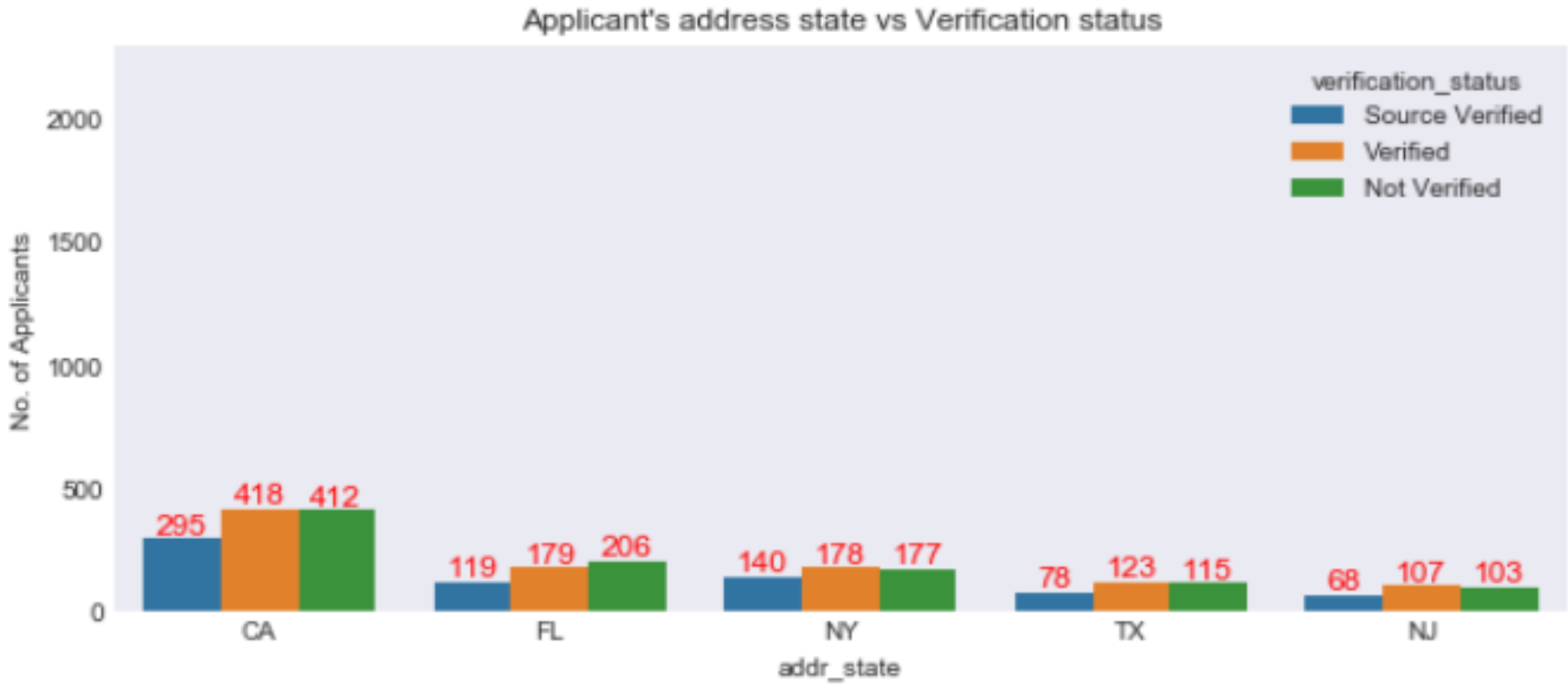
# Plots – Home Ownership Vs Grade Distribution




**Plot 33**

- Plot 33 depicts that most of the defaulter are from Grade C and Grade B in both Rent and Mortgage ownership categories.

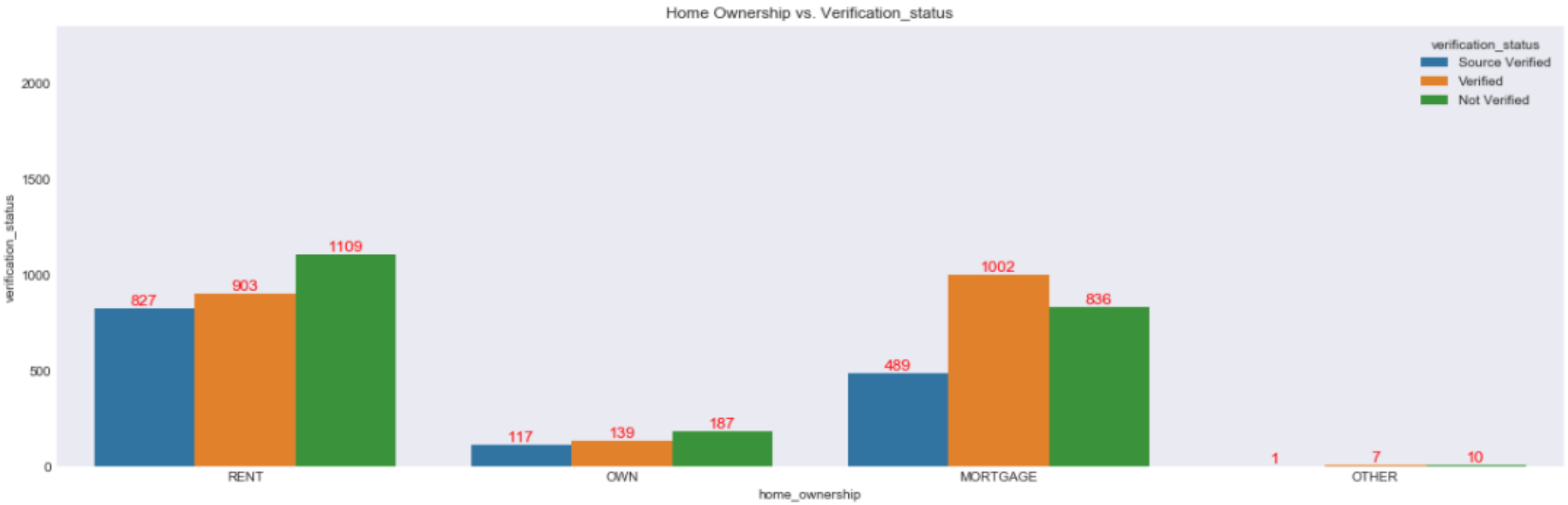
# Plots – Applicant Address Vs Verification Status




**Plot 34**

• Plot 34 depicts that  CA has most of the defaulter as applicants income are not verified. Also FL (Florida) has more number of non verified applicants then verified as a result defaulter are dominating in this state.

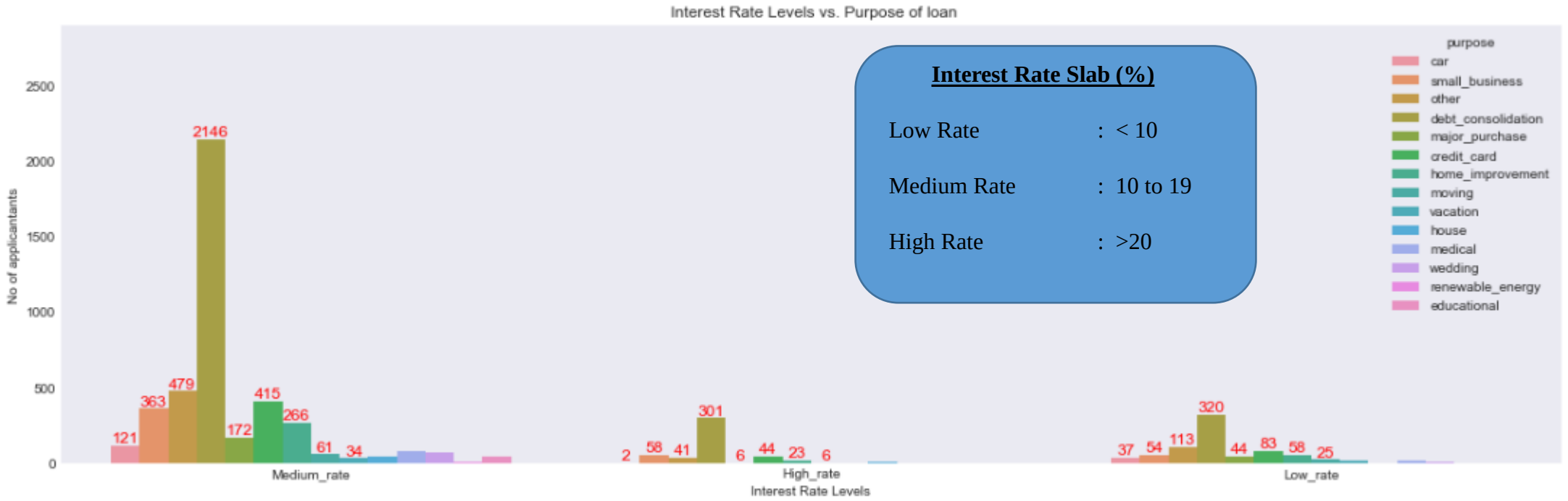
# Plots – Home Owner ship Vs Verification Status




**Plot 35**

- Plot 35 depicts that applicants that have  OWN accommodation have high count of not verified status.

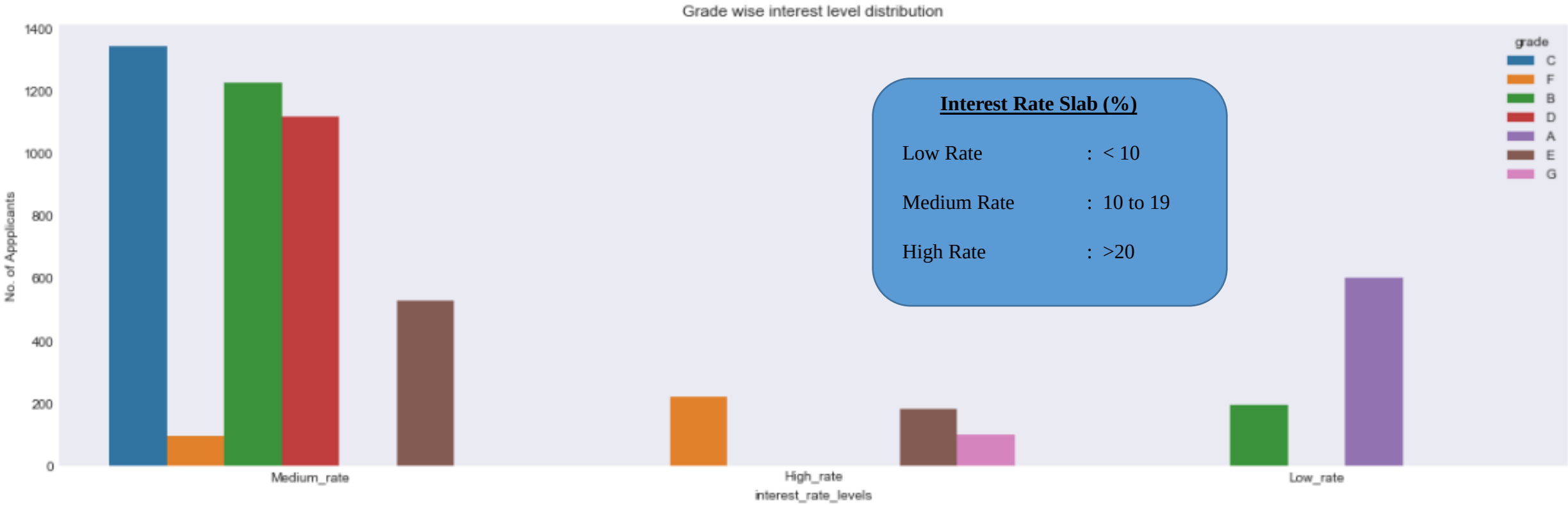
# Plots – Interest Rate Levels Vs Purpose of loan




**Plot 36**

• Plot 36 depicts that applicants are taking loan on high rate for the purpose of debt_consolidation, which make them defaulter.

# Plots – Interest Rate Levels Vs Grades



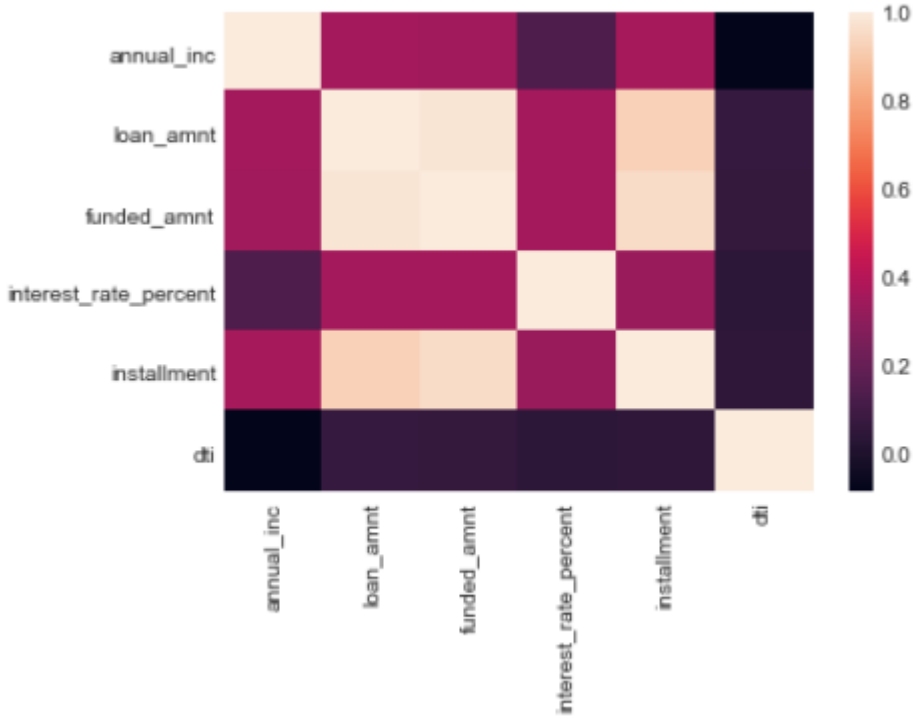

**Plot 37**

• Plot 37 depicts that Grade C applicants are defaulter (charged off) in Medium rate, Grade F applicants are defaulter (charged off) in High rate and Grade A applicants are defaulter (charged off) in Low rate.

# Plots – Correlation

| | annual_inc | loan_amnt | funded_amnt | interest_rate_percent | installment | dti |
|---|---|---|---|---|---|---|
| annual_inc | 1 | 0.35 | 0.35 | 0.13 | 0.36 | -0.09 |
| loan_amnt | 0.35 | 1 | 0.98 | 0.35 | 0.93 | 0.064 |
| funded_amnt | 0.35 | 0.98 | 1 | 0.35 | 0.95 | 0.061 |
| interest_rate_percent | 0.13 | 0.35 | 0.35 | 1 | 0.33 | 0.041 |
| installment | 0.36 | 0.93 | 0.95 | 0.33 | 1 | 0.042 |
| dti | -0.09 | 0.064 | 0.061 | 0.041 | 0.042 | 1 |

**Plot 38**




**Plot 39**

- Plot 38 & 39 depicts that loan amount and funded amount has high correlation.
- Loan amount is directly proportional to funded amount.
- Annual income is negatively correlated with dti. It means as annual income increases dti will decreases.
- The correlation between dti and interest rate percent is low.

# Conclusion and Recommendation

- Applicants living in CA (California) state have higher defaulter (charged off). In CA state 945xx has highest defaulter,
- In CA most of the applicants address are not verified.
  - **Recommendation: Should avoid providing loans in CA for applicants having un-verified status.**
- Employers in executive levels are higher charged off numbers as compared to junior level.
- Applicants having rented accommodation have highest default count.
- Applicants having own accommodation also have high rate of non- verified status.
  - **Recommendation: Verification must have be done for all the applicants irrespective of their home ownership status for reducing Charged off cased.**
- Debt_Consolidation is emerged as the prime purpose for which applicants taken loans and get defaulted
  - **Recommendation: Providing loans for debt_consolidation should be avoided and if provided then it should be provided at higher interest rate.**
- Loan term of 36 months has higher charged off rate as compared to 60 months, but if we consider fully paid case along with charged off cases then 60 months loan term is more problematic then 36 months.
  - **Recommendation: Loan should be provided with higher interest rate with shorter loan term.**
- Grade B and C have higher charged off rate. However, Grade F applicants have higher defaulter count in high interest rate.
- Defaulter in Grade B and C applicants have Rented accommodation
  - **Recommendation: Applicants verification must have been performed irrespective of the Grade system.**
- Annual income is negatively correlated with dti, it means applicants having higher income have low dti.