Lending Club Case Study

VINAY KUMAR SAROHA

VISHVESH DEOBHANKAR

Overview

- **Lending Club**: A consumer finance marketplace connecting borrowers with investors.
- **Loan Approval**: Decisions based on applicant profiles.
- **Credit Loss**: Biggest risk from "risky" borrowers who default, leading to financial loss.
- **Key Challenge**: Minimize credit loss by:
  - Avoiding rejection of reliable applicants (business loss).
  - Preventing approval of risky loans (financial loss).

Dataset Analysis Overview

Key Attribute: Loan Status

- **Fully-Paid**: Loan repaid
- **Charged-Off**: Loan defaulted
- **Current**: In-progress (ignored)

Important Features

- **Customer Demographics**: Income, Home Ownership, Employment Length, Debt to Income, State
- **Loan Attributes**: Loan Amount, Grade, Term, Purpose, Interest Rate, Installment, Public Records

Ignored Data: Post-loan behavior and overly granular data.

Data Cleaning

1. **Rows to Drop**
    1. Rows where *loan_status = Current* will be dropped.
    2. Duplicate rows will be removed.

2. **Columns to Drop**
    1. Columns with only NA values.
    2. Columns with constant or zero values.
    3. Columns where more than 65% of data is missing.
    4. Redundant columns like *id, member_id, emp_title, desc, title, url*.
    5. Customer behavior columns that don't contribute to loan approval prediction.

3. **Data Type Conversion**
    1. Convert columns like *loan_amnt, funded_amnt, int_rate* to the appropriate data types.

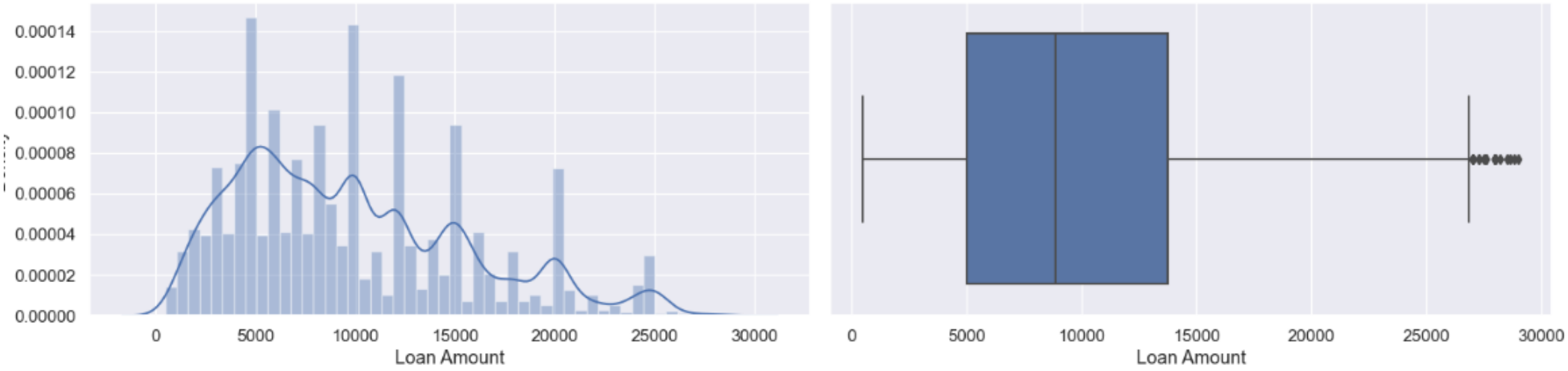
4. **New Columns Added**
    1. *verification_status_n* based on verification hierarchy.
    2. *issue_y* and *issue_m* extracted from *issue_d*.

Approach

- Data Reduction
  - Drop rows/columns
- Data Transformation
  - Convert Data Types
- Missing Values
  - Identify & Impute
- Outlier Treatment
  - Detect & Handle Outliers

Quantitative Variable Analysis

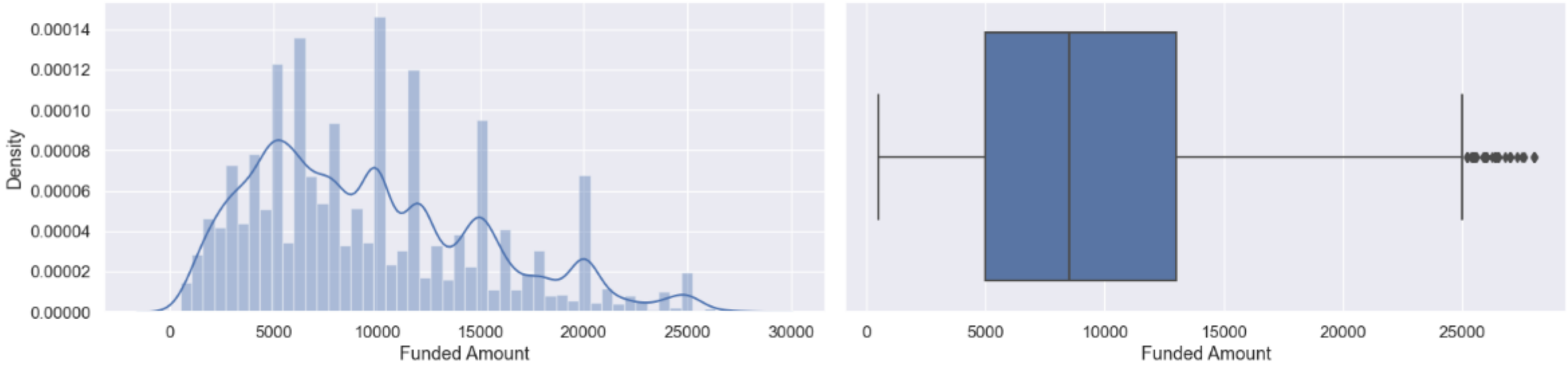
```
univariate_analysis(loan, 'loan_amnt')
```



Majority of the loan_amount is in the range of 5K to 14K

funded_amnt

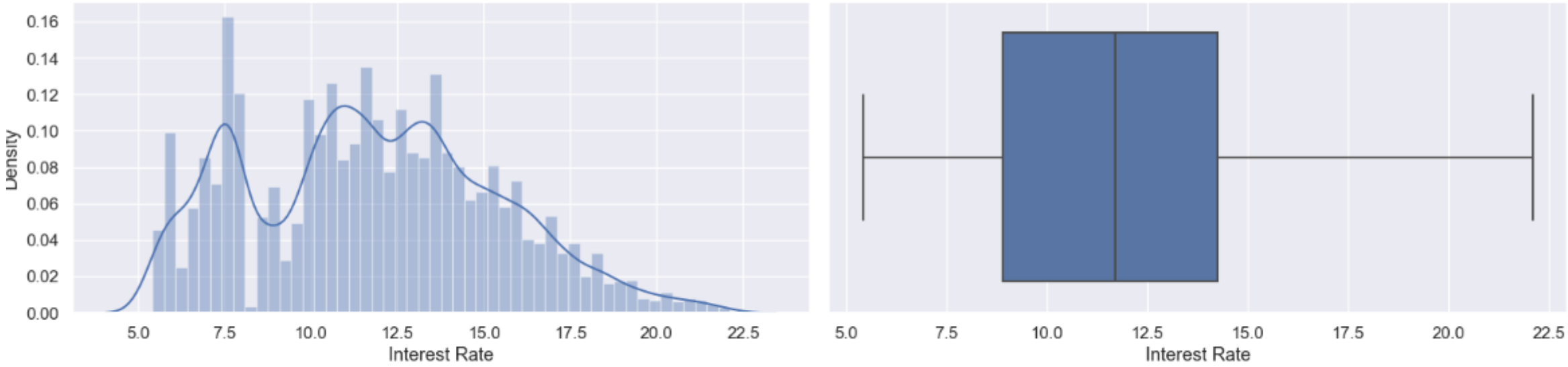
```
univariate_analysis(loan, 'funded_amnt')
```



Majority of the funded_amnt is in the range of 5K to 13K

int_rate

```
univariate_analysis(loan, 'int_rate')
```



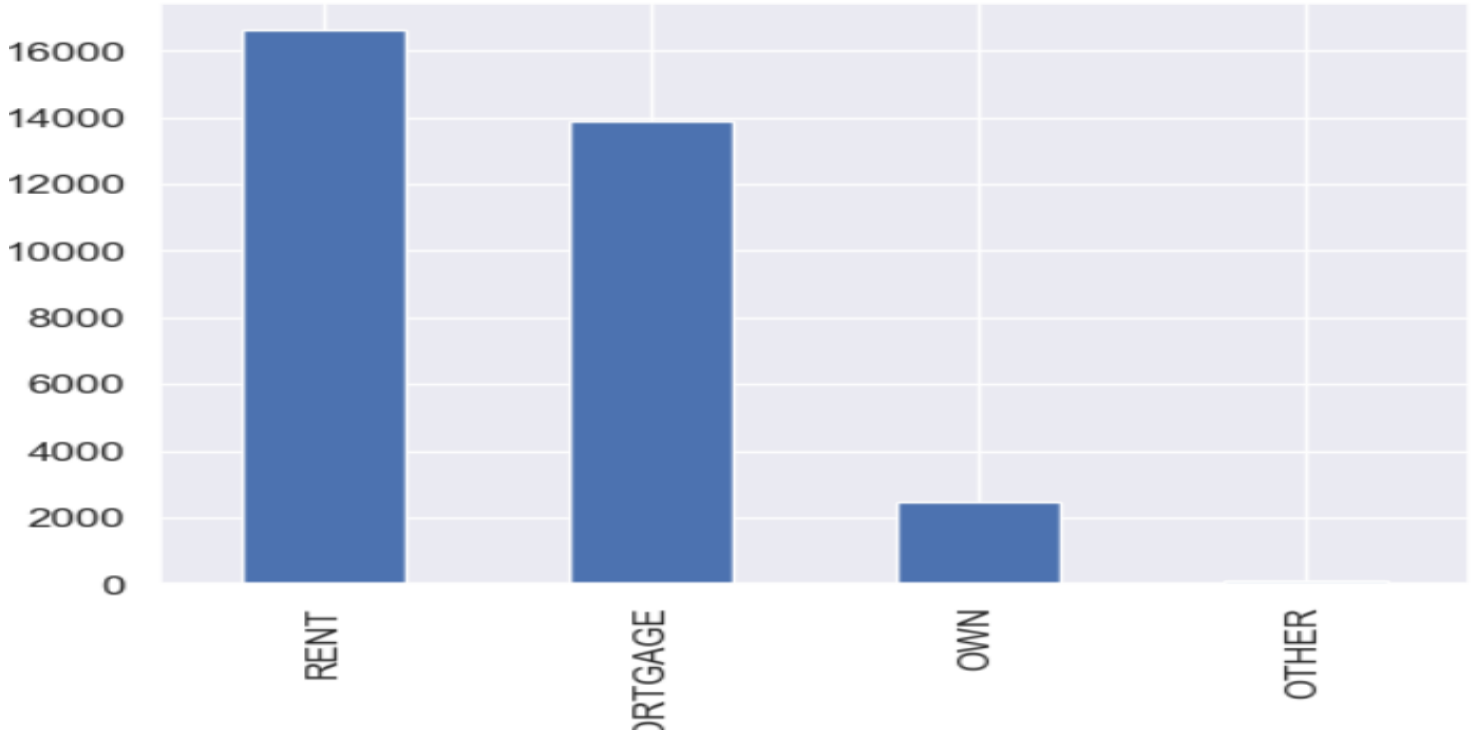
Majority of the interest rate is in the range of 5% to 16% going at the max to 22%

Unordered Categorical Variable Analysis

home_ownership

```
loan['home_ownership'].value_counts().plot.bar()
```

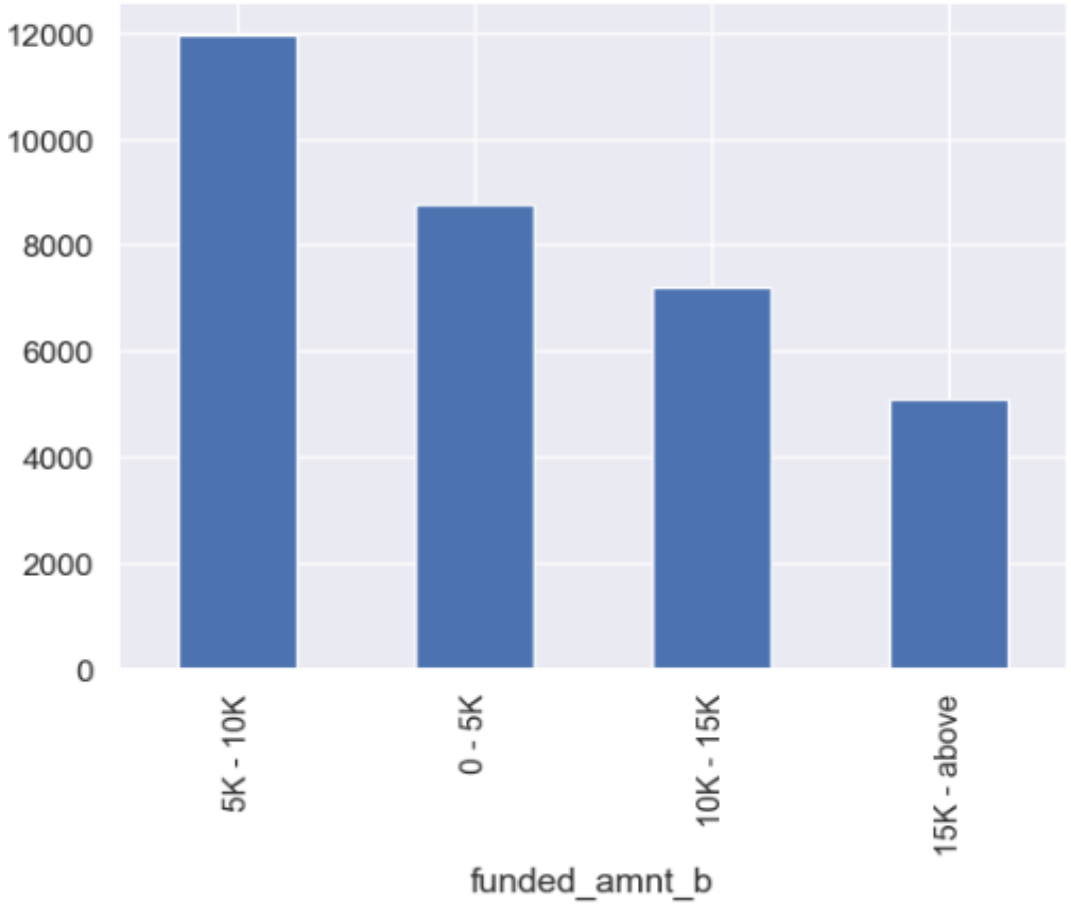
```
<Axes: xlabel='home_ownership'>
```



funded_amnt_b

```
loan['funded_amnt_b'].value_counts().plot.bar()
```

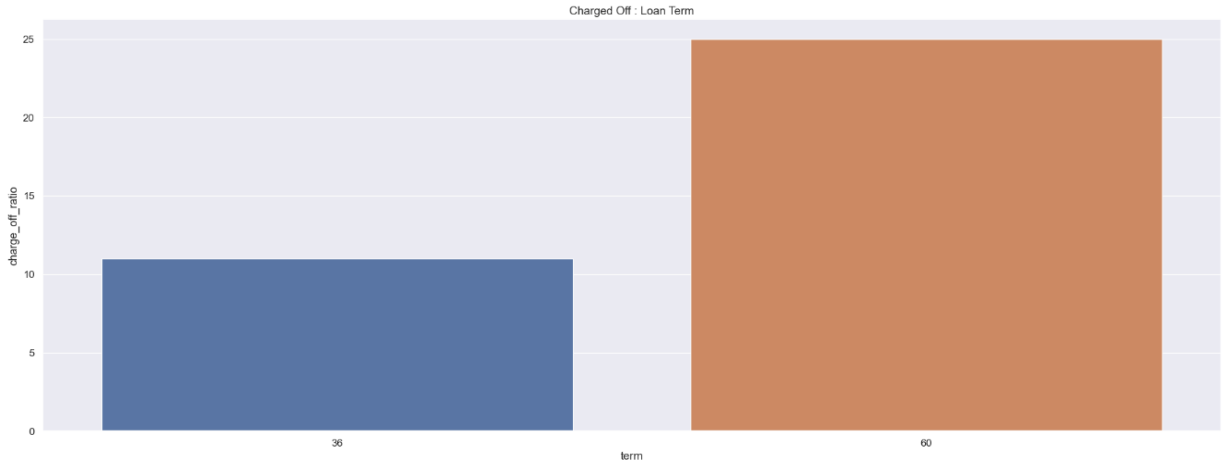
```
<Axes: xlabel='funded_amnt_b'>
```

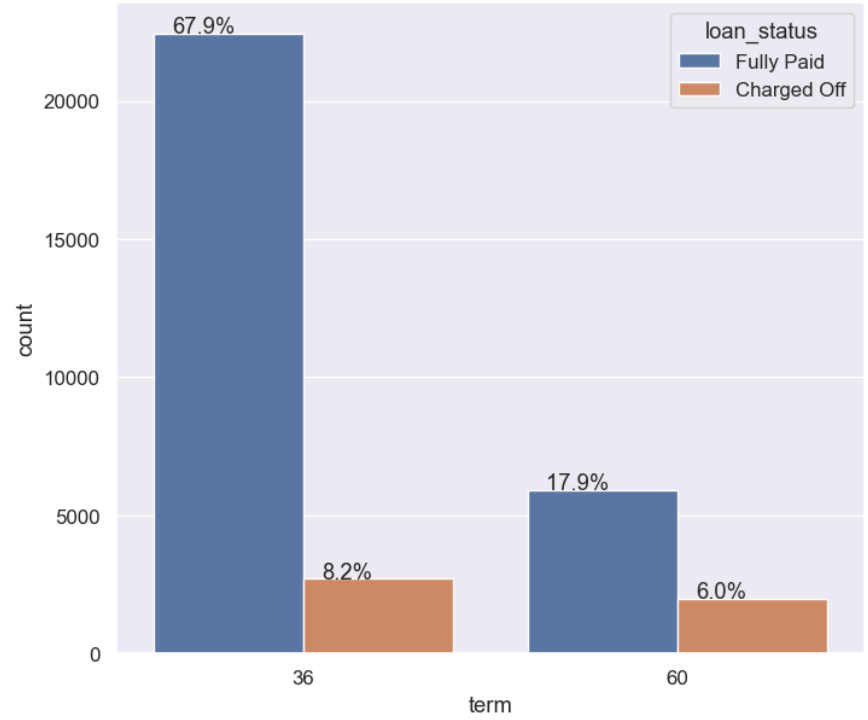


Highest funded amount applications fall in the range of 5k to 10k

Bivariate Analysis

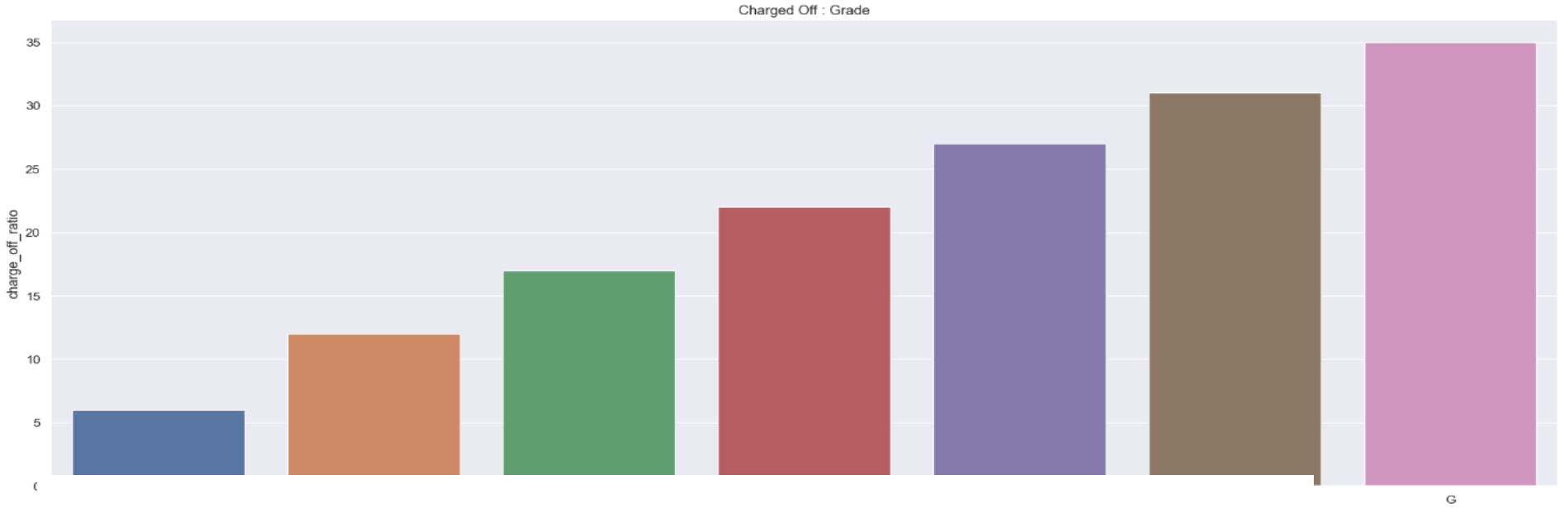
- The volume of loans are in the category of term = 36
- The overall percentage of Charge Off's is slightly higher in term = 36 (8%) as compared to term=60 (6%)
- If we calculate the ratio of Charge Off's within a category
    - **Charge Off**s ratio is for the term=60 is 25% which is much higher than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





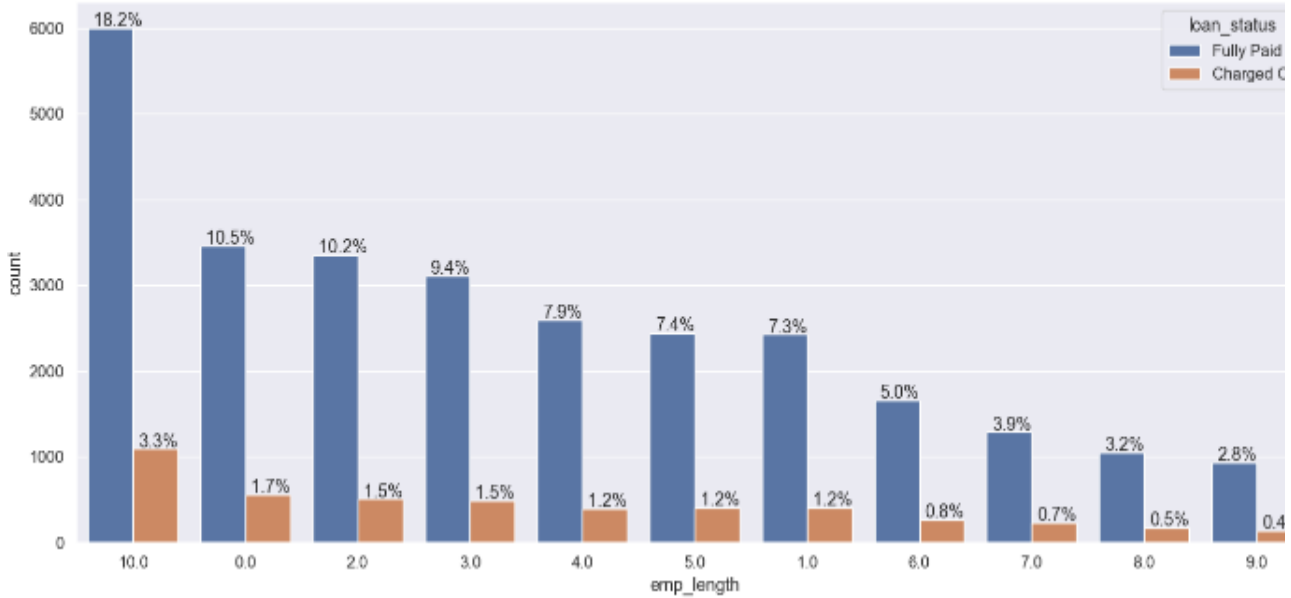
```
## The ratio of Charge Offs within the category total
ratio_wise_plot(loan, 'grade')
```

Pyt

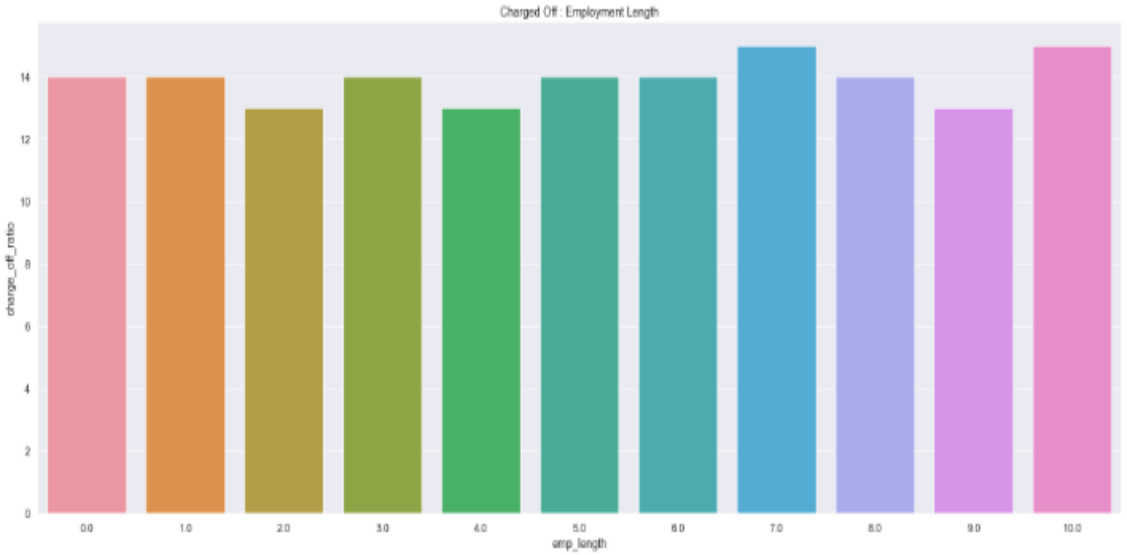


- The Majority of *loan volume is in grade=B*
- Highest percentage of overall Charge Offs are in grade B (3.7%) and C(3.6%)
- If we analyse the Charge Off Ratio within a category
    - The highest percentage of **Charge Off**s are in the *grade=G*
    - Highest cluster of **Charge Off**s are in the grades 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 off's are in the grades of B and C
    - Grade "F" and "G" have very high chances of charged off. The columes are low
    - Grade "A" has very less chances of charged off.
    - Probablity of charged off is increasing from "A" to "G"

```
# Overall ratio of Charge Offs against the total
series_plot(loan, 'emp_length', 'loan_status')
```

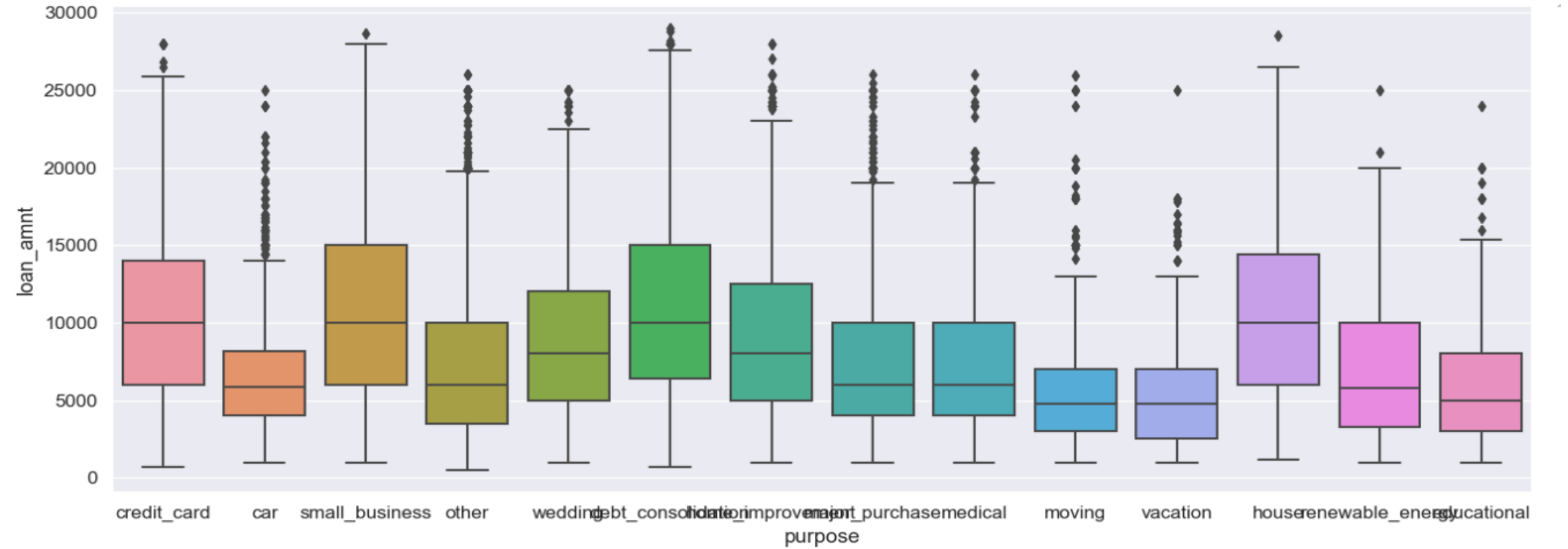



```
## The ratio of Charge Offs within the category total
ratio_wise_plot(loan, 'emp_length')
```



- Highest Charge Offs are in the employee length category of 10 Years and above
- Charge Off ratio within the categories itself are similar and inconclusive
- **Inferences**
  - Highest Charge Offs are in the employee length of 10 Years and above
  - High probablity of Charge Off's whose income range is less than 1 years
  - Ratio within the ranges are pretty much same (in conclusive)

```
plot.figure(figsize=(16,6))
sea.boxplot(y=loan.loan_amnt,x=loan.purpose)
plot.show()
```



- Highest risk of Charge Offs are the category of debt_consolidation
- Highest probablity 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

Outcome

Customer Demographics

- Majority of the loan applicants are in the range of 0 - 40K annual income
- Majority of the debt to income is in the range of 0 to 20 going at the max to 30
- Majority of the home owner 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 an public record of bankruptcies
- Majority of the employment length of the customers are 10+ years and then in the range of 0-2 years

Loan Demographics

- Highest loan amount applications fall in the range of 5k to 10k
- Majority of the interest rate is in the range of 5% to 16% going at the max to 22%
- Majority of the installment amount is in the range of 20to400
- Majority of the loan applications counts are in the term of 36 months
- Majority of loan application counts fall under the catogory of Grade B

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

Inferences

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