Exploratory Data Analysis- Gramener Case Study

An analytical approach to understand the driving factors behind loan defaults. This knowledge can be used by the companies for its portfolio and risk assessment.

Group Name:

1. Pradnya Garde
2. Prachi Agrawal
3. Bhagyashree K

Steps Followed

- Understanding the Business flow with objectives
- Understanding the problem statement
- Data Understanding
- Data Cleaning and Manipulation
- Data Analysis
  i) Univariate Analysis
  ii) Bivariate Analysis

Understanding the Data

- There are 111 variables and 39717 rows in the loan.csv dataset.
- However we see that not all variables will be helpful for our analysis.
- There are a few variables which are irrelevant for the purpose of our analysis.
- And there are a few variables with 100% null values which are irrelevant to our analysis and hence will be dropped in further process.
- The most important variable seems to be loan_status around which the complete analysis of ours will be centered.

- loan_amnt >= funded_amnt >= funded_amnt_inv
- funded_amnt = out_prncp(is non-zero for current loans only) + total_rec_prncp
- total_pymnt = total_rec_prncp + total_rec_int + total_rec_late_fee
- grades are ordered (A, B, C, D, E, F, G)
- sub_grades are ordered (1-5)
- on average higher interest rates are associated with higher grades/subgrades

Data Cleaning and Manipulation

1. Dropping null columns-
    a. 57 columns are dropped as they have greater than 70% or 100% null values.
    Now we are left with 54 columns.
2. Dropping unnecessary columns-
    a. Following columns are not required for our analysis. And hence
    are dropped.
    b. Id, member_id, url, title, recoveries, collection_recovery_fee, zip_code, addr_state
    c. Now we are left with 46 columns
3. Dropping columns with same values for all rows as they don't have any effect on our analysis.
    a. 9 columns dropped. 37 columns remaining

Data Cleaning and Manipulation

4. Dropping rows with high null percentages-
some of the rows might have greater than five NaN values. Such rows aren't of much use for the analysis and hence, should be removed.

5. Since we have to concentrate on Charged-off loans, deleting the rows that are having loan_status=Current status does not affect the analysation.

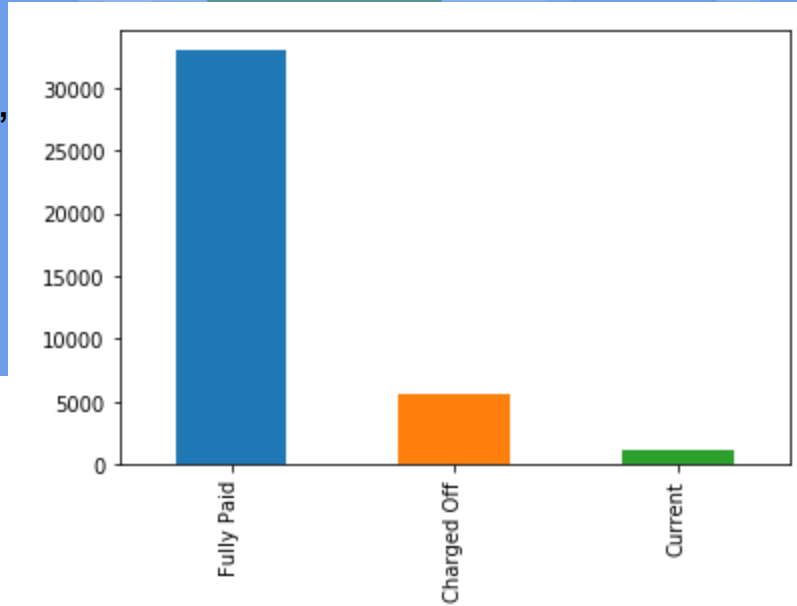
6. Some attributed have improper data type. Hence, dtype of some attributes have been changed.
Now the data is ready for Analysis.

Data Analysis

Unordered Categorical Variables - Univariate Analysis

- Calculating the number of application on the basis of loan status
- Conclusion- Though the number of applications with status as “Fully Paid” is high, the number of applications with status as “Charged Off” is also significant.
- Our further analysis will be centered around “Charged Off” loans as they are more related to loan defaults.

```
Fully Paid     32950
Charged Off     5627
Current         1140
Name: loan_status, dtype: int64
```

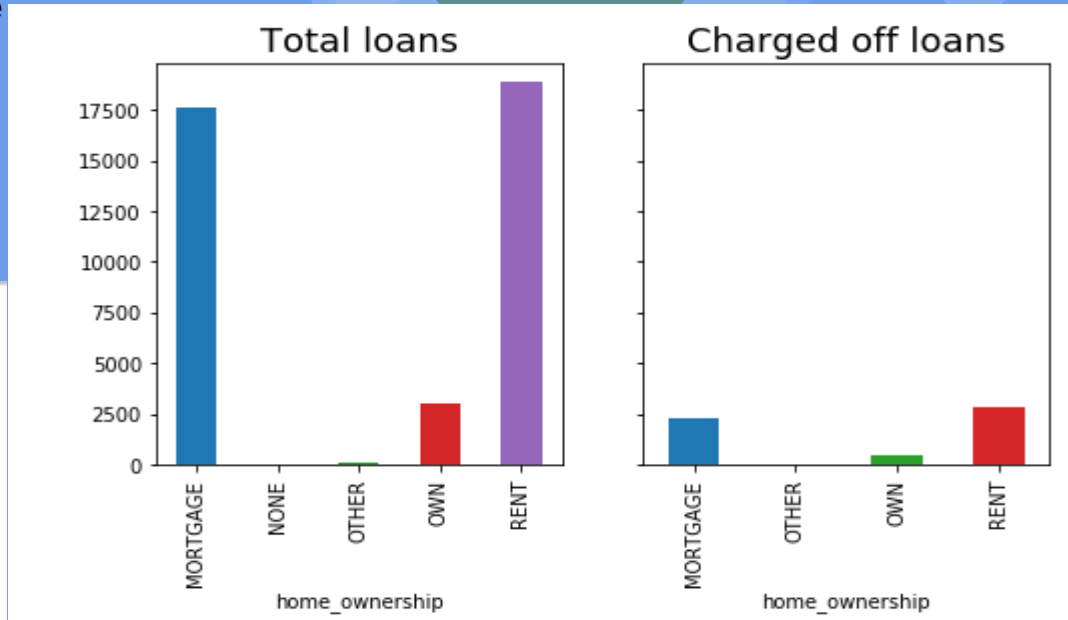


Data Analysis

Unordered Categorical Variables - Univariate Analysis

- **Analysing Charged Off loans- On the basis of home_ownership**
  - **Most of the Charged Off loans were with home_ownership as 'MORTGAGE' and 'RENT'**
  - **Proportion of Charged Off loans to the total loans in each segment of 'home_ownership' shows no significant variation and is close to the overall proportion of charged off loans.**

	Total	Charged Off	percentage
MORTGAGE	17659	2327.0	13.177417
NONE	3	NaN	NaN
OTHER	98	18.0	18.367347
OWN	3058	443.0	14.486593
RENT	18899	2839.0	15.021959

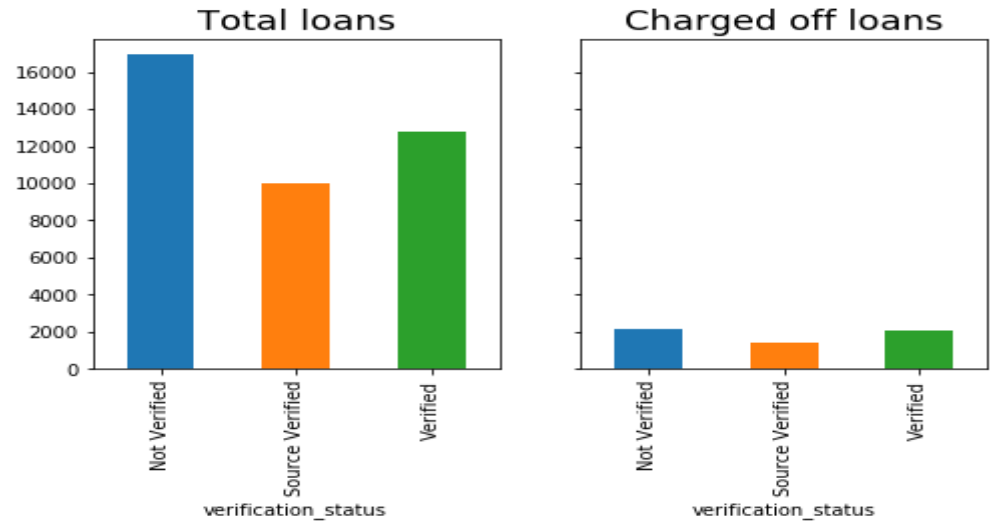


Data Analysis

Unordered Categorical Variables - Univariate Analysis

- **Analysing Charged Off loans- On the basis of verification_status**
  - **Proportion of Charged Off loans to the total loans in each segment of 'verification_status' shows no significant variation and is close to the overall proportion of charged off loans.**
  - **However there are significant number of charged off loan applications which are not verified i.e., 2142; This could be one of the reasons for loan defaults.**

verification_status	Total	Charged Off	percentage
Not Verified	16921	2142	12.658826
Source Verified	9987	1434	14.358666
Verified	12809	2051	16.012179

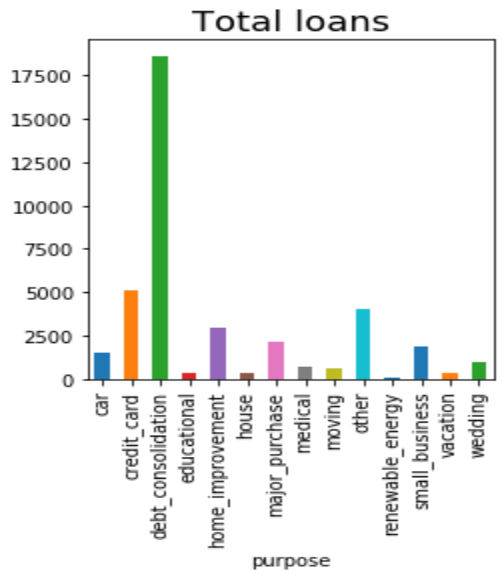


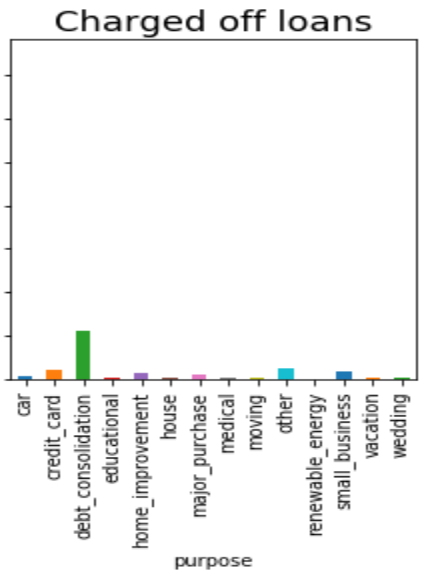
Data Analysis

Unordered Categorical Variables - Univariate Analysis

- Analysing Charged Off loans- On the basis of purpose
  - Here we observe the high number of loans with the purpose of “debt_consolidation”
  - Also the percentage of Charged Off loans to total loans for purpose as “small_business” is the highest at 25.98%

purpose	Total	Charged Off	percentage
car	1549	160	10.329245
credit_card	5130	542	10.565302
debt_consolidation	18641	2767	14.843624
educational	325	56	17.230769
home_improvement	2976	347	11.659946
house	381	59	15.485564
major_purchase	2187	222	10.150892
medical	693	106	15.295815
moving	583	92	15.780446
other	3993	633	15.852742
renewable_energy	103	19	18.446602
small_business	1828	475	25.984683
vacation	381	53	13.910761
wedding	947	96	10.137276





Data Analysis

Unordered Categorical Variables - Univariate Analysis

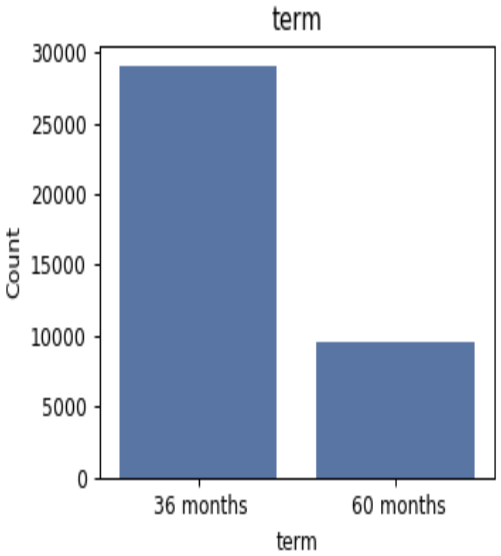
- **Analysing Charged Off loans- On the basis of term**
  - **It is observed that though the overall number of loan applications w 36 months is greater than wit 60 months, for charged off status, loans with 60 months term are more defaulted.**

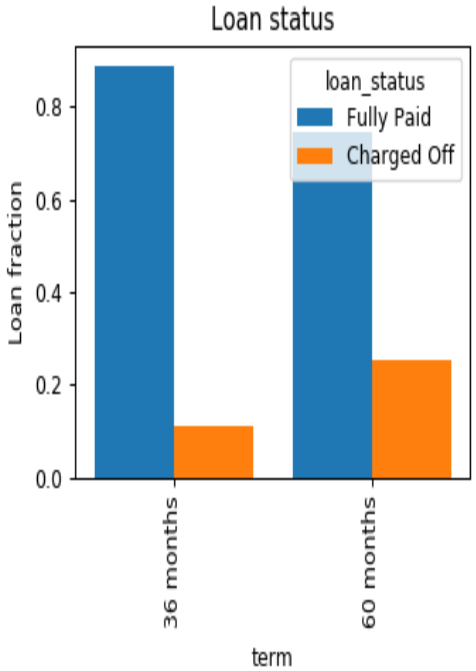
Out of total loan applications

```
 36 months    29096
 60 months     9481
Name: term, dtype: int64
 36 months    75.42318
 60 months    24.57682
Name: term, dtype: float64
```

For Charged Off loans

```
term
 36 months    3227
 60 months    2400
dtype: int64
```

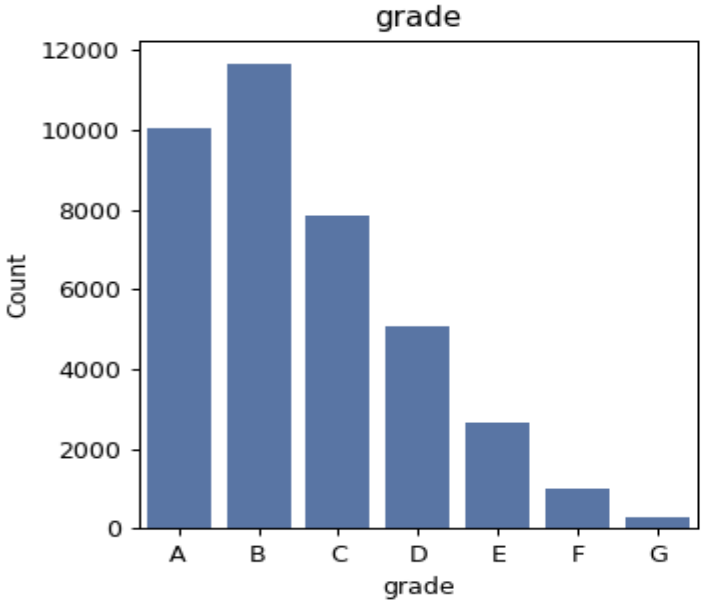


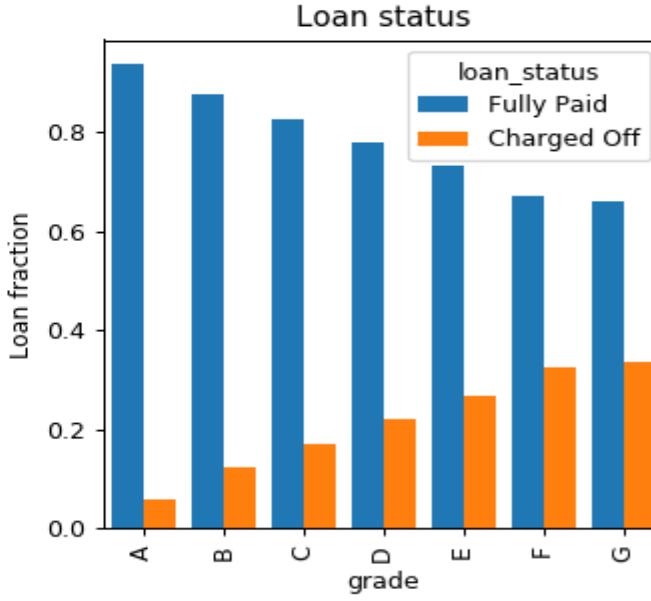


Data Analysis

Unordered Categorical Variables - Univariate Analysis

- Analysing Charged Off loans- On the basis of grade, sub_grade
  - The Charged off loan status is high and comparable for grades E, F and G

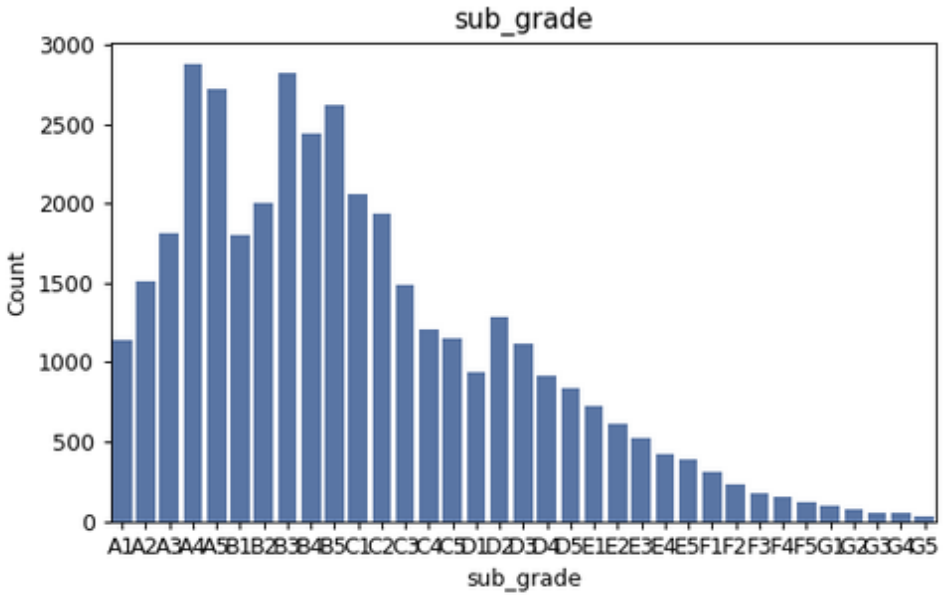


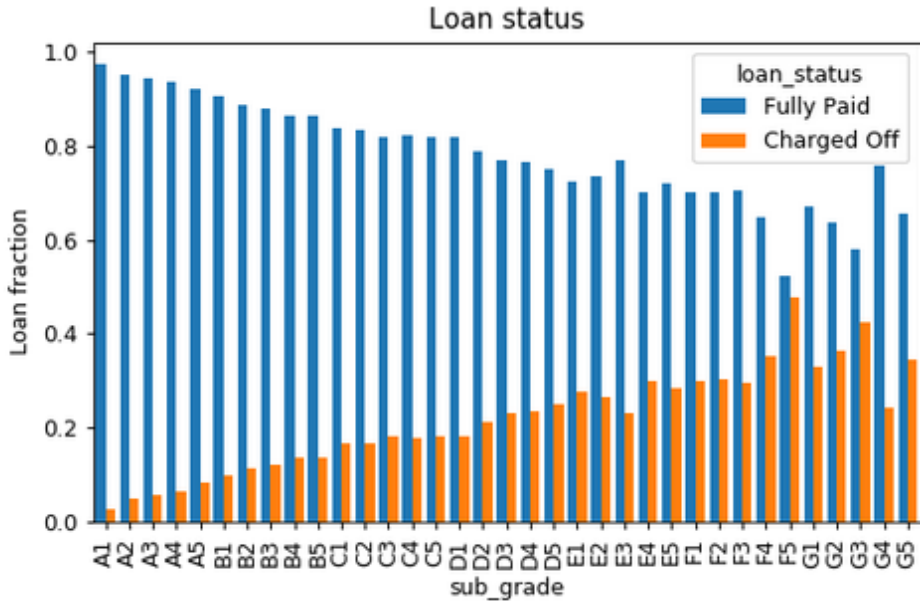


Data Analysis

Unordered Categorical Variables - Univariate Analysis

- Analysing Charged Off loans- On the basis of grade, sub_grade
  - Sub_grade has granular information; The insights from both grade and sub_grade columns is similar

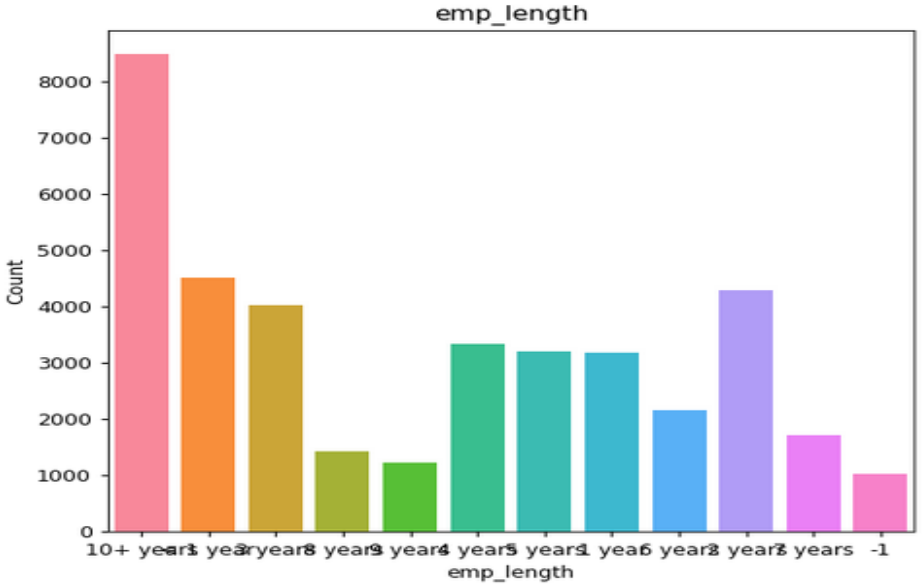


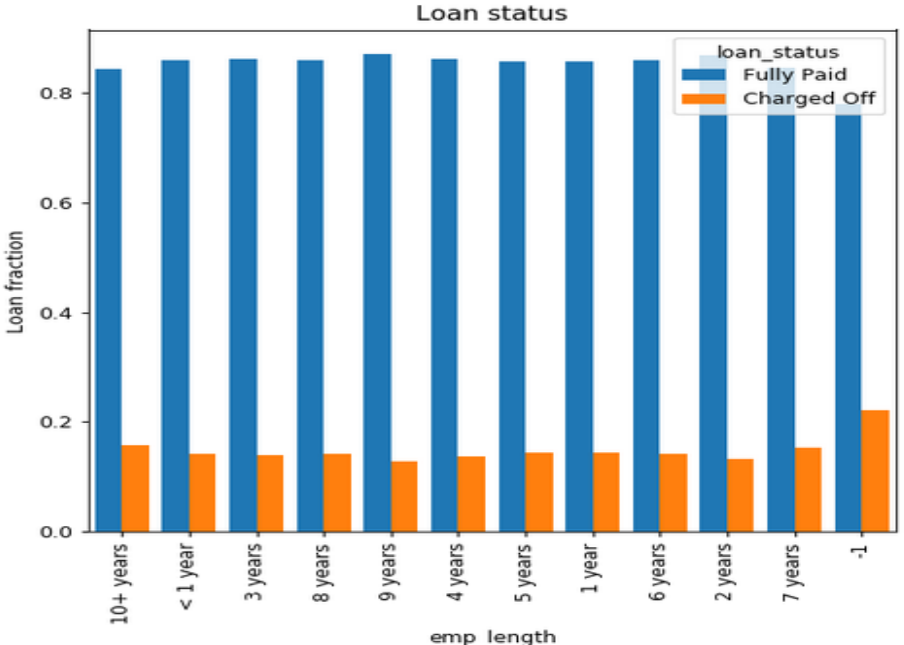


Data Analysis

Unordered Categorical Variables - Univariate Analysis

- Analysis categorical variable- emp_length
  - Loan applications is the highest for 10+ years of emp_length

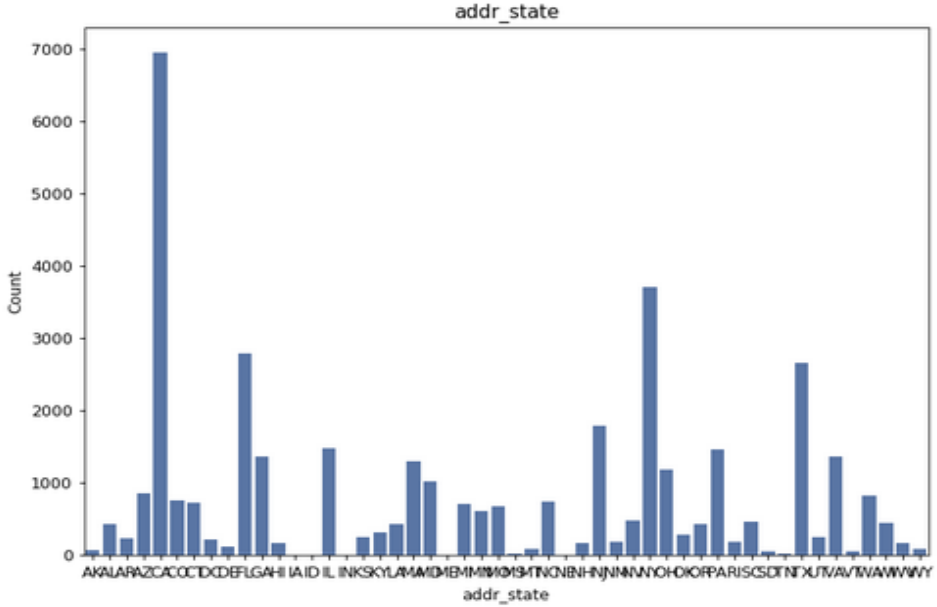


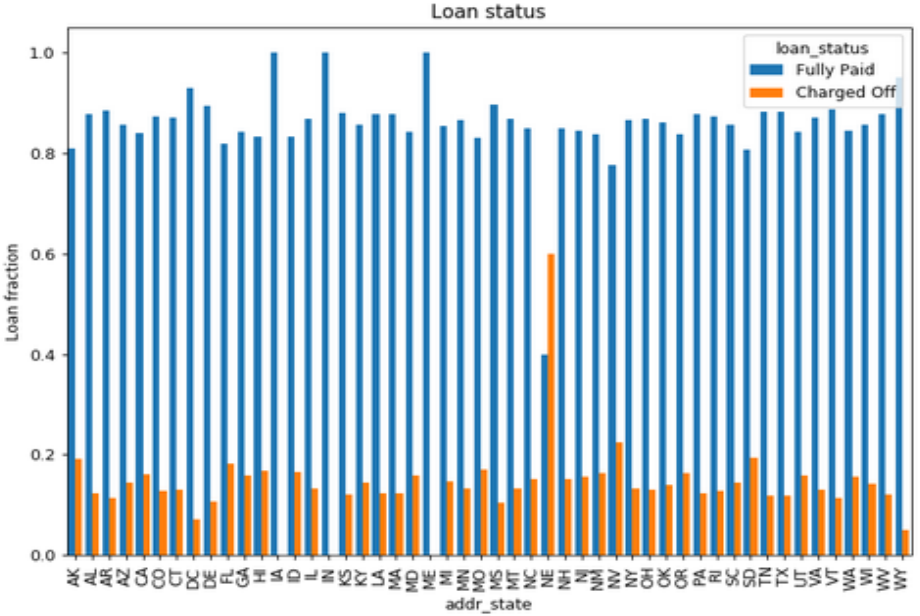


Data Analysis

Unordered Categorical Variables - Univariate Analysis

- Analysis of categorical variable- addr_state
  - Observe the largest percentage segments - 'CA', 'NY', 'FL'
  - Fraction of 'charged_off' loans - shows a lot of variation for each segment-- 'NE', 'NV' and 'AK' are the 3 highest

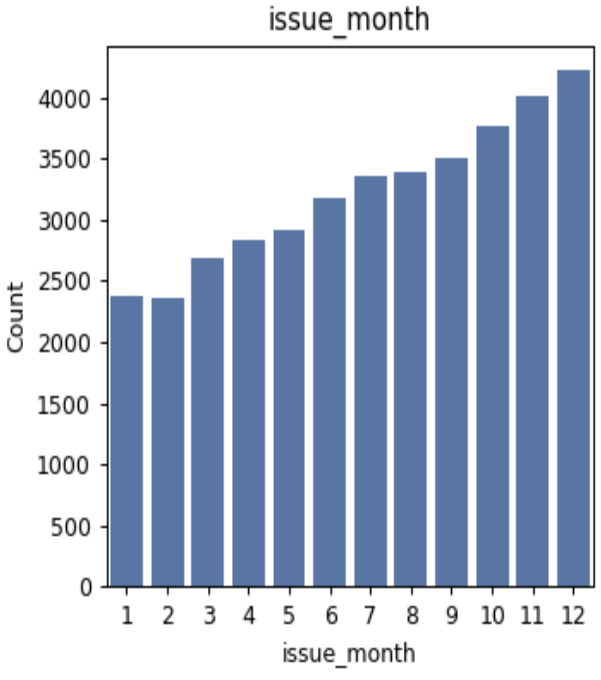


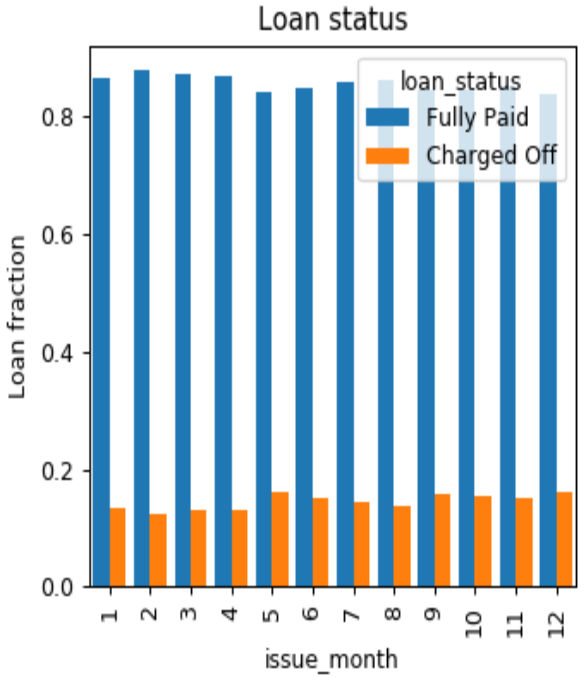


Data Analysis

Unordered Categorical Variables - Univariate Analysis

- Analysis categorical variable- issue_d
  - Consider the months-- fraction of charged_off loans is similar in all months;
  - Does not help in analysis and can be removed

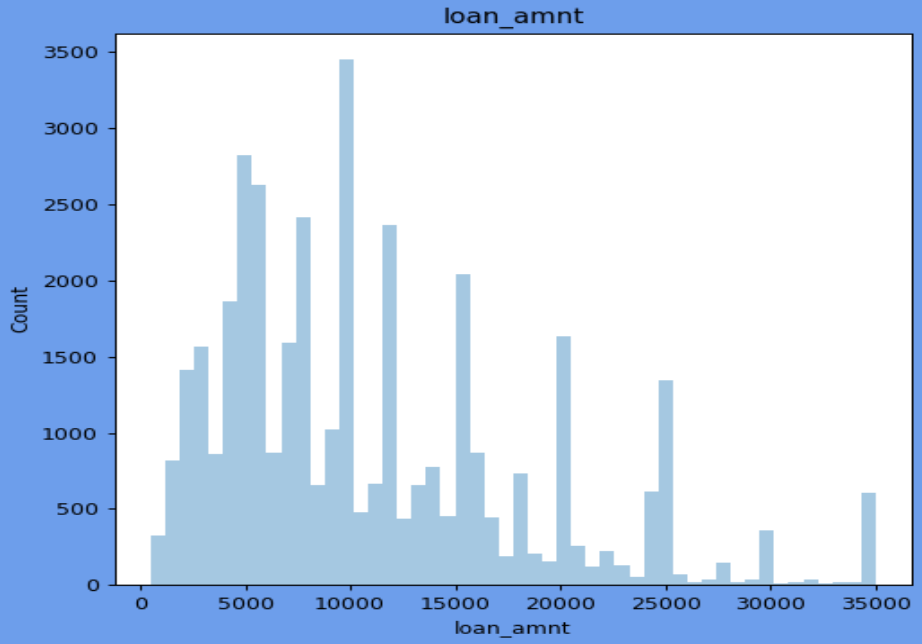


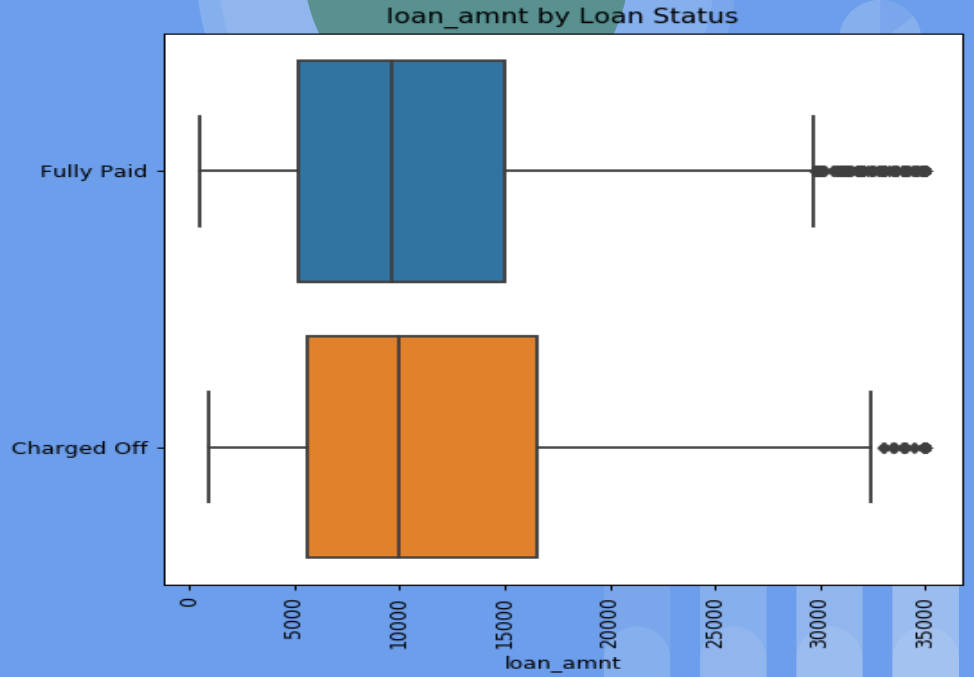


Data Analysis

Quantitative Variables - Univariate Analysis

- **Analysis of quantitative variable- loan_amt**
  - **The median loan_amt is approximately Rs. 10000 for fully paid and charged off loan applications**

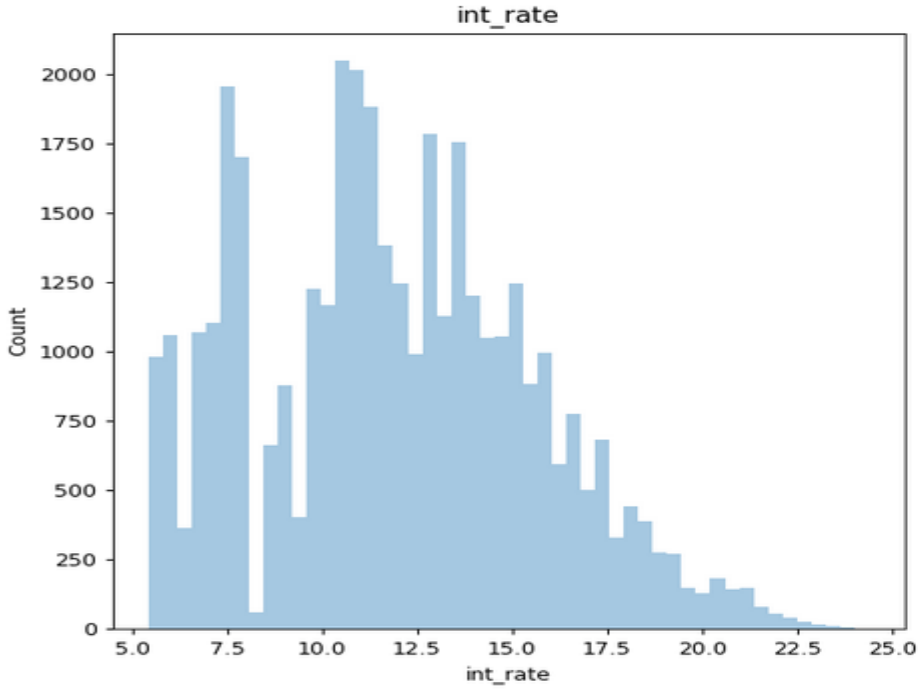


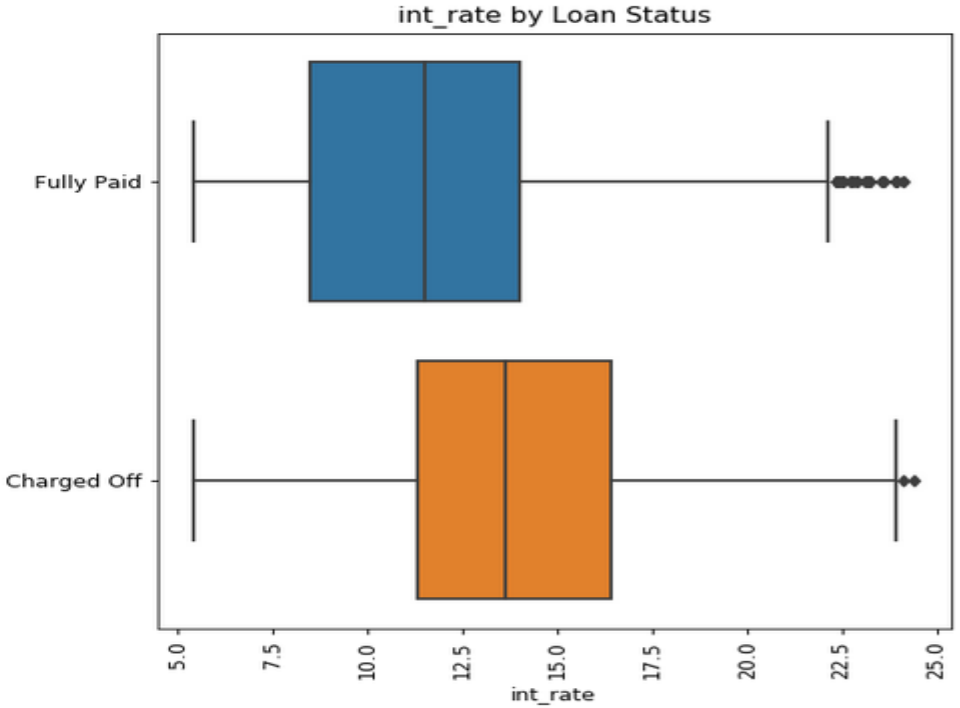


Data Analysis

Quantitative Variables - Univariate Analysis

- Analysis of quantitative variable- int_rate
  - The median interest rate for charged off loan applications is greater than fully paid loan applications

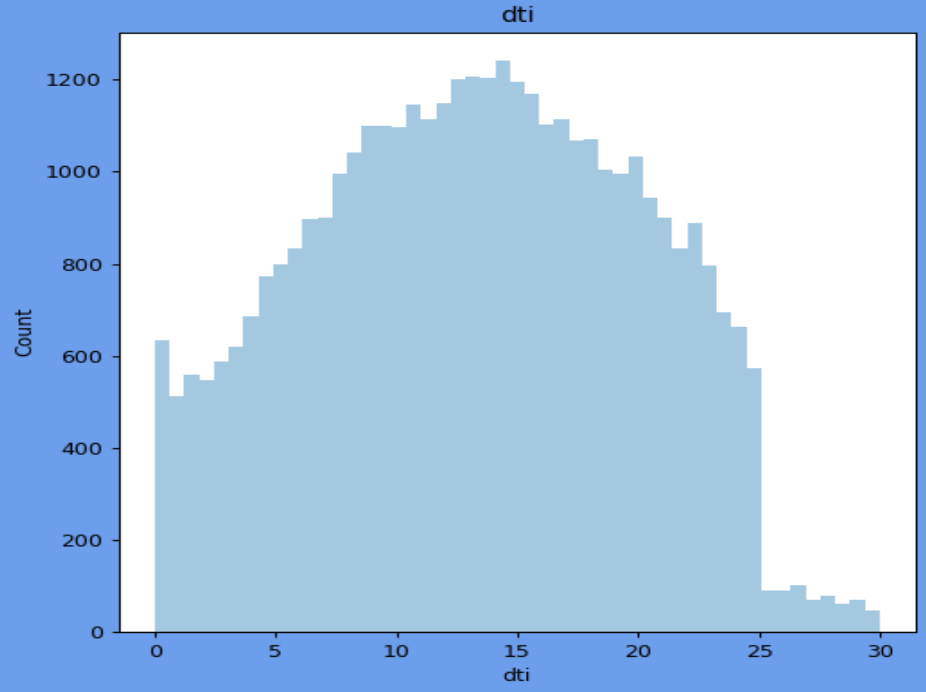


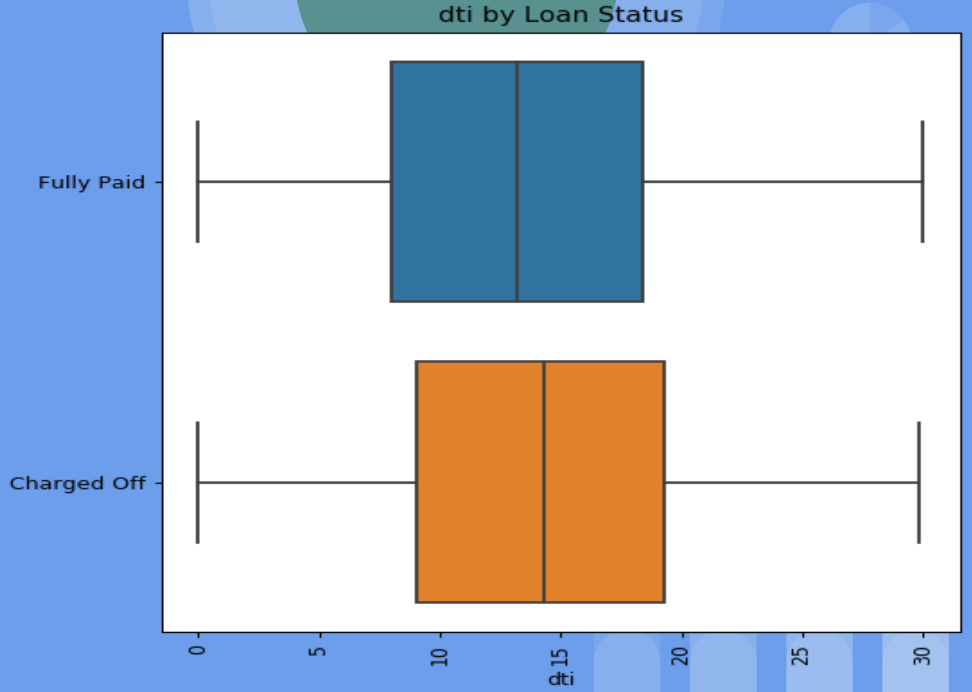


Data Analysis

Quantitative Variables - Univariate Analysis

- Analysis of quantitative variable- dti(debt to income ratio)
  - Median dti for both fully paid and charged off loan applications is approximately same

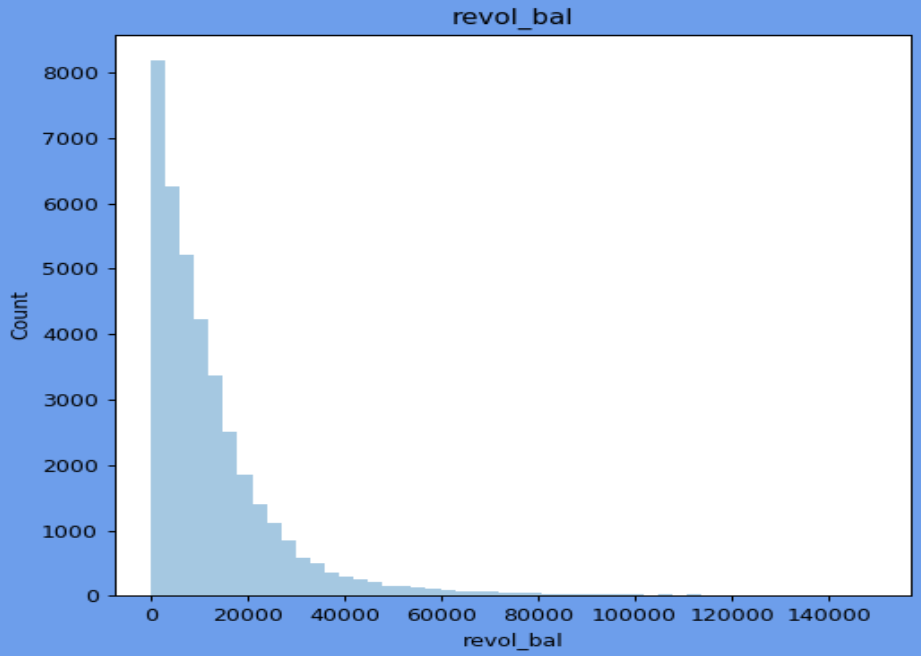


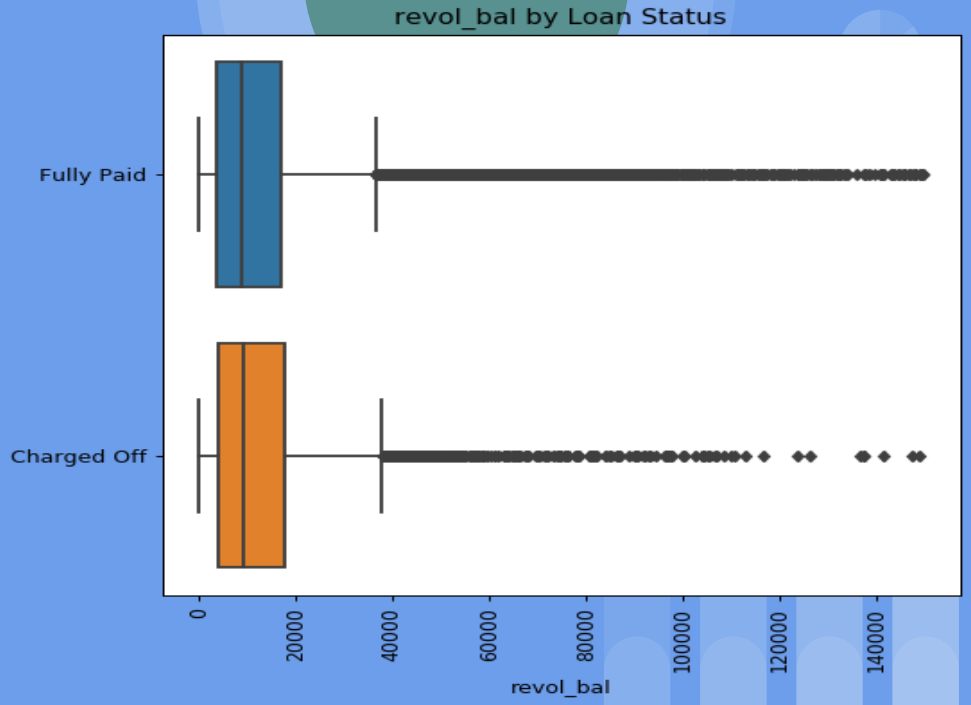


Data Analysis

Quantitative Variables - Univariate Analysis

- **Analysis of quantitative variable- revol_bal**
  - **The median revolving balance seems to be approximately Rs. 5000**

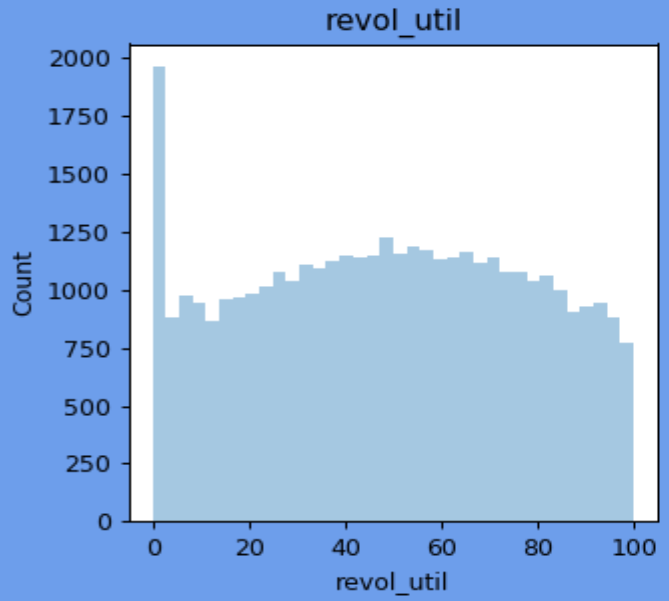


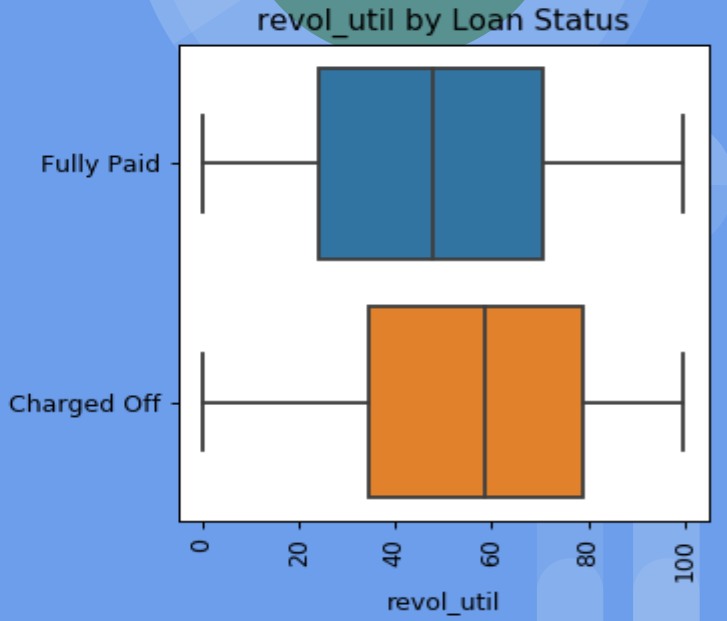


Data Analysis

Quantitative Variables - Univariate Analysis

- **Analysis of quantitative variable- revol_util**
  - **Revolving utilization, also known as your “debt-to-limit ratio” or “credit utilization,” measures the amount of your revolving credit limits that you are currently using.**
  - **Median revolving utilization for charged off loan applications is greater than for fully paid loan applications**



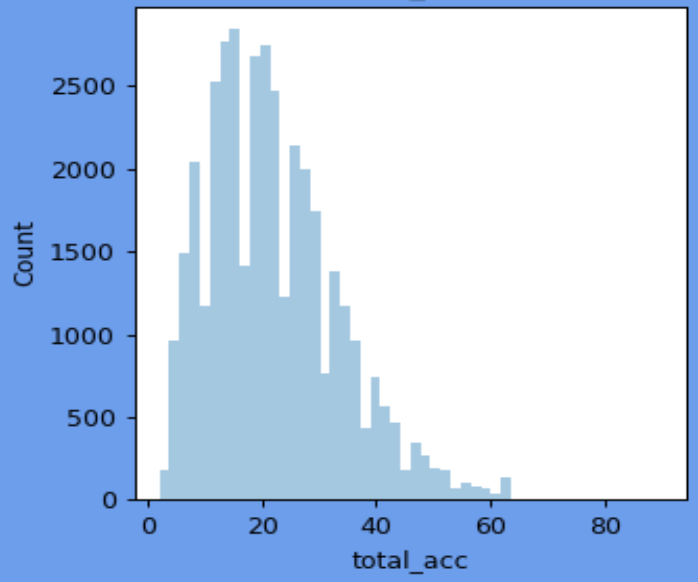


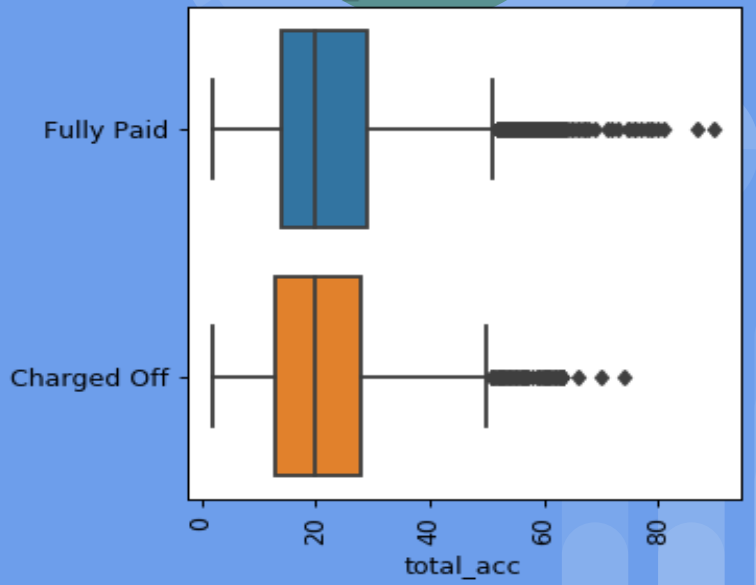
Data Analysis

Quantitative Variables - Univariate Analysis

- **Analysis of quantitative variable- total_acc**
  - **Total_acc represents the total number of credit lines currently in the borrower's credit file**
  - **The median value of total_acc for both fully paid and charged off loan applications is approximately the same**

which

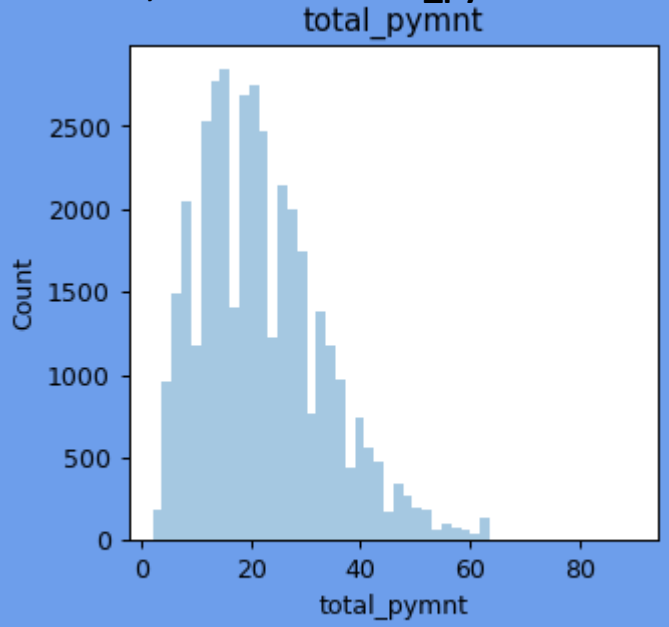


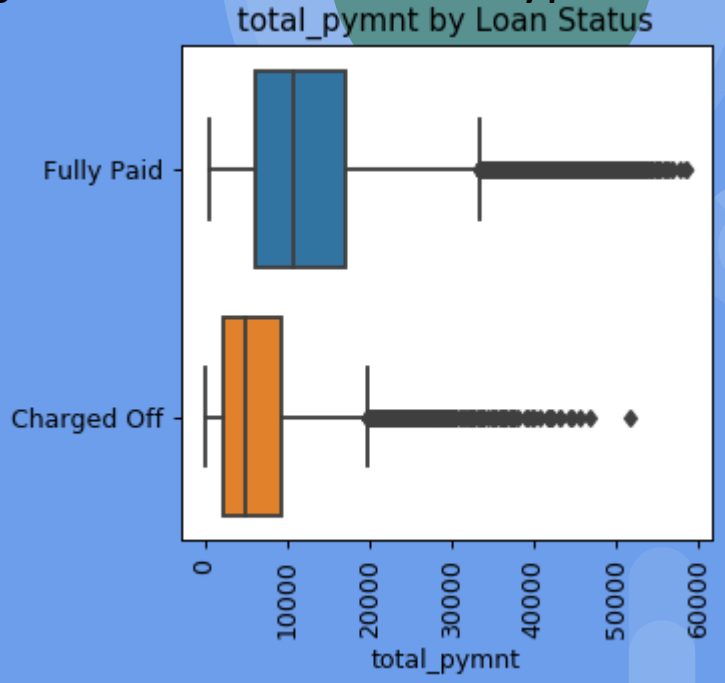


Data Analysis

Quantitative Variables - Univariate Analysis

- **Analysis of quantitative variable- total_pymnt**
  - **Column total_payment represents payments received to date for total amount funded**
  - **As obvious, the median total_pymnt for charged off loans is much less than fully paid loans**

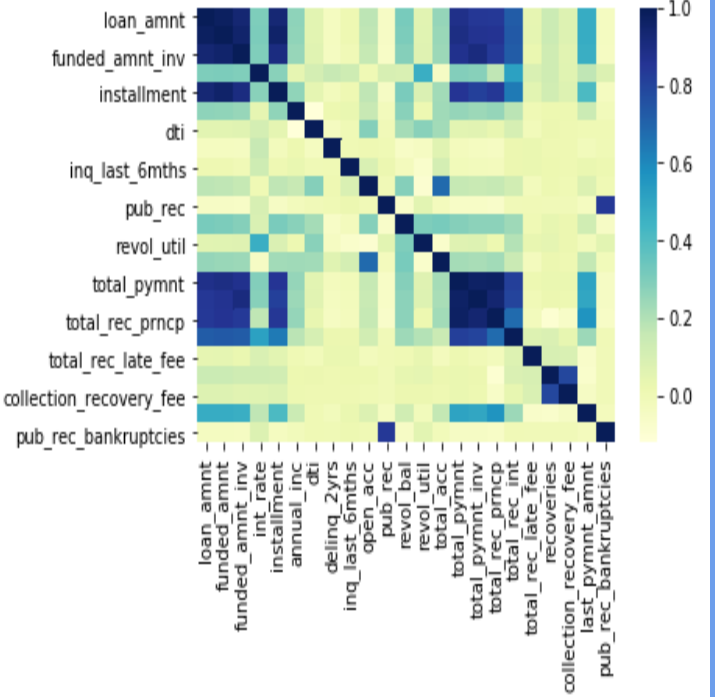




Data Analysis

Bivariate Analysis

➢ **To find correlation between the quantitative/numerical variables. We identify groups of closely related variables through heatmap**



From the above the following have a strong correlation and so pick one from the set

- loan_amnt, funded_amnt, funded_amnt_inv, installment :analyse 'loan_amnt'
- total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int:analyse 'total_pymnt'

Following are post charged_off actions, hence can't help identify risky loan applications

- recoveries, collection_recovery_fee :delete these columns

Following have strong correlation only with each other

- pub_rec, pub_rec_bankruptcies :delete
- -delinq_2yrs, inq_last_6mnths :delete have no significance to the analysis

Conclusions

Univariate and Bivariate Analysis

- The correlation is 0.9 which means that with increase in funded amount invested the installments increase.
- The correlation is 0.95 which means that with increase in loan amount there is increase in funded amount invested.
- Relationship between loan_status as charged off and home_ownership variable- Around 28% of the charged off loan applications were with home_ownership as mortgage and rent.
- Relationship between loan_status and verification status- Nearly 12% of the charged off loans were not verified.
- Relationship between loan_status as charged off and purpose-
  - By number, charged off loans were high for debt_consolidation
  - By percentage, charged off loans were high for small businesses
- The median loan_amt is approximately Rs. 10000 for fully paid and charged off loan applications
- The median total_pymnt for charged off loans is much less than fully paid loans
- Median revolving utilization for charged off loan applications is greater than for fully paid loan applications
- Median dti for both fully paid and charged off loan applications is approximately same
- The median interest rate for charged off loan applications is greater than fully paid loan applications