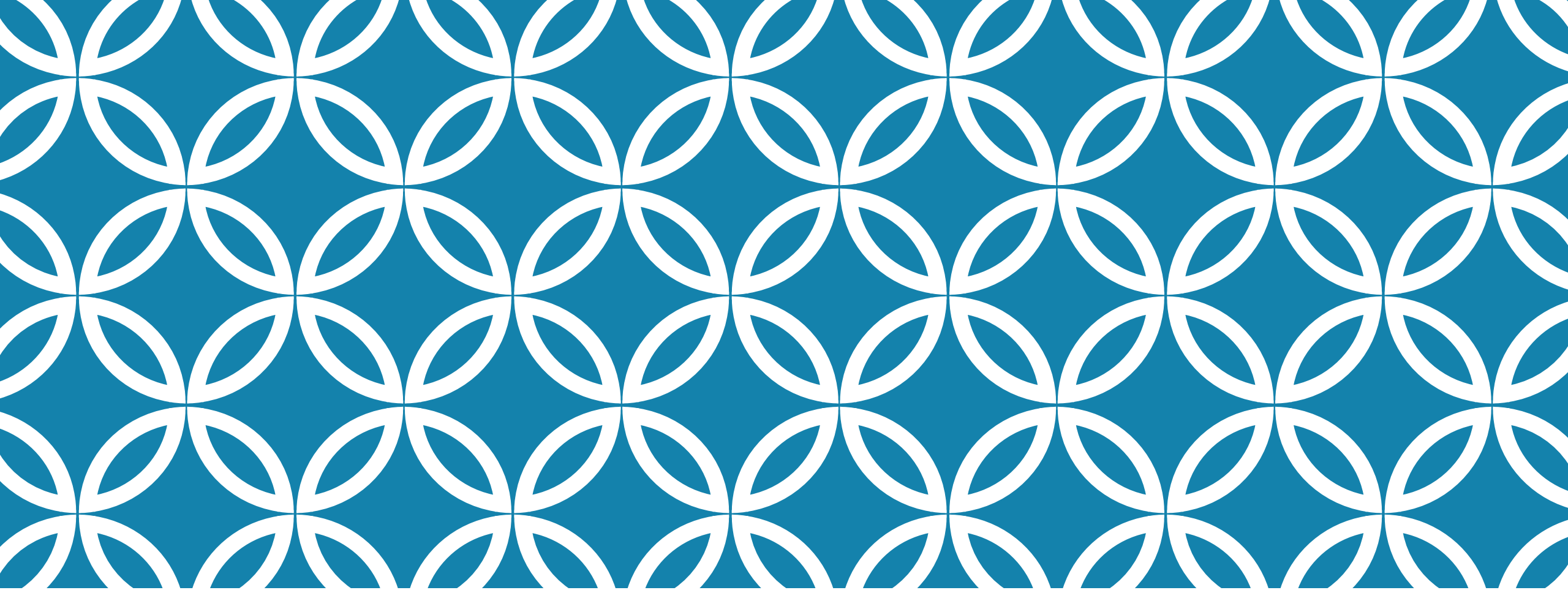


LENDING CLUB CASE STUDY

Financial Risks in Lending
and
Probability of Delinquency

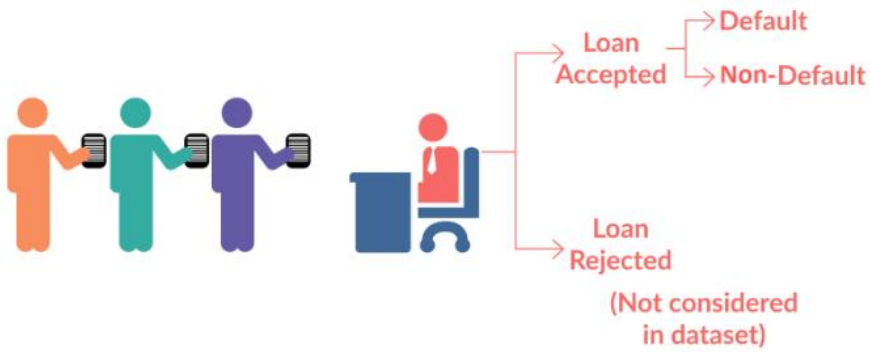
PROBLEM STATEMENT

In this Case Study of the various loans facilitated by Lending Club, we will try to assess the risks associated with the loans. We will use EDA on a dataset of historical and current loans to understand how various attributes could influence a tendency to default.

Two associated risks are:

- **Opportunity Risk:** Loans for deserving applicants are not funded
- **Default Risk:** Loans funded are not repaid by the applicant

LOAN DATASET



Some points to note for the loan dataset:

- A Loan is accepted if it is funded by investors. Status of the loan signifies the below categories:
  - **Charged Off:** Applicant was unable to pay/defaulted on the loan
  - **Current:** Applicant is in the process of paying the loan
  - **Fully Paid:** Applicant has fully paid off the loan

DATA CLEANSING APPROACH

NOTE:

- No records were filtered.
- Nulls were replaced based on the column definitions.
- Date columns in string format were converted to Dates
- Unwanted columns were dropped

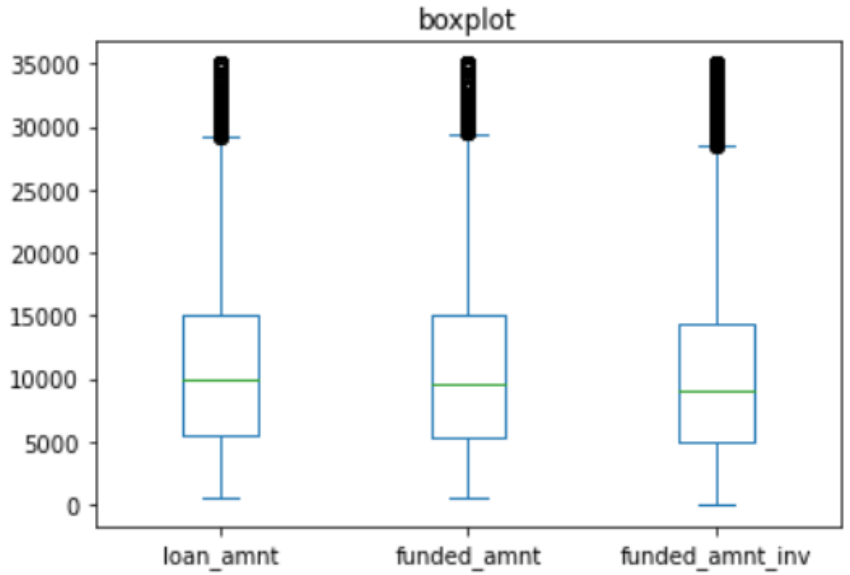
Below columns were handled

- **term** - Converted to int
- **int_rate** - Converted to float
- **int_rate_bucket** - Added new column to create buckets for int_rate
- **annual_inc** - Deleted annual_inc and added annual_inc_k with income in thousands (k)
- **emp_length** - Converted to float and replaced '< 1 year' with .5 and '10+ years' with 10. NOTE: Using an upper limit of 10 here as there is no exact data post 10 years
- **url** - Dropped 'url' column as the data seems redundant
- **revol_util** - Converted to float. Since all loans with Null revol_util are either closed or charged off, null values were replaced with 0
- **pub_rec_bankruptcies** - Converted to int and converted 'NA' to 0 as no records is equivalent to 0
- **mths_since_last_delinq** - Converted to int and converted 'NA' to 0 as no records is equivalent to 0
- **mths_since_last_record** - Converted to int and converted 'NA' to 0 as no records is equivalent to 0
- **issue_d** - Converted to date
- **earliest_cr_line** - Converted to date
- **earliest_cr_line_years** - Added with number of years difference between issue_d and earliest_cr_line

INITIAL UNIVARATE ANALYSYS
[All Loans]

INFERENCE:

- Approved/Investor Funded amounts have a 50th quartile of 9600/8975 and a mean of 10947.71/10397.44. The outliers are not causing a big skew between the 50th quartile and mean.
- Outlier counts of 1265/1293 seem insignificant compared to overall sample count of 38577. Any additional filtering if required on outliers will be addressed base on impact of skewness determined in further analysis
- **funded_amnt** ranges from 500 to 35000
- Initial look into the data shows that [minimum **funded_amnt_inv** == 0] which looks like a discrepancy/data issue since an unfunded loan should not have a status of Charged Off/Paid Off/Current



	loan_amnt	funded_amnt	funded_amnt_inv
count	39717.000000	39717.000000	39717.000000
mean	11219.443815	10947.713196	10397.448868
std	7456.670694	7187.238670	7128.450439
min	500.000000	500.000000	0.000000
25%	5500.000000	5400.000000	5000.000000
50%	10000.000000	9600.000000	8975.000000
75%	15000.000000	15000.000000	14400.000000
max	35000.000000	35000.000000	35000.000000

INITIAL UNIVARATE ANALYSYS
[All Loans]

INFERENCE:

- 116 rows had [**funded_amnt < total_rec_prncp**] but the difference was less than a dollar. Hence no data cleansing was done on these rows
- Number of rows with [funded_amnt_inv < total_rec_prncp] was found to be 16696. Since 16696 is sizeable sample of records, and since funded_amnt is more in line with total_rec_prncp, no corrections will be done for the funded_amnt_inv.
- funded_amnt will be used for future analysis to check for the probability of delinquency

```
Sample Size: 39717
funded_amnt outlier count and percentage: 1265[3.19]%
funded_amnt_inv outlier count and percentage: 1193[3.0]%
```

```
Number of rows with funded_amnt < total_rec_prncp : 116
Number of rows with funded_amnt < total_rec_prncp but difference is less than a dollar : 116

# NOTE: Since difference was less than a dollar, hence no data cleansing was done on these 116 rows
```

```
Count of rows where funded_amnt_inv == 0: 129
Number of rows with funded_amnt_inv < total_rec_prncp : 16696

# NOTE:
Since 16696 is sizeable sample of records, and since funded_amnt is more in line with total_rec_prncp,
no corrections will be done for the funded_amnt_inv
funded_amnt will be used for future analysis to check for the probability of delinquency
```

INITIAL UNIVARATE ANALYSYS

[All Loans]

categorical columns

INFERENCE:

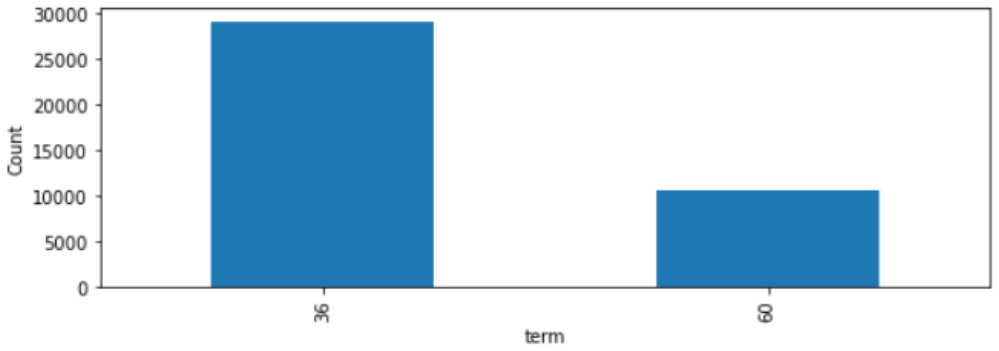
- **term** - Majority of customers prefer a shorter term loan
- **int_rate** - Majority of the loans have an interest rate between 6% to 18%. Highest number of loans were issued between 9% to 12% and 2nd highest is between 12% to 15%. This indicates most investors are willing to take on high risk while lending. Fewer investors prefer rates between 0 to 6 % due to lower yield and rates above 18% due to significantly high risk
- **grade** - Higher grades are preferred by investors as the probability of default is lower. Most preferred grade is B
- **sub_grade** - Higher sub grades are preferred by investors as the probability of default is lower. Most preferred grade is B3 and A4
- **emp_length** - Most Investors prefer loan applicants with more than 10 years of employment history
- **home_ownership** - Most loan applicants have a mortgage or are staying on rent which is a possible indicator that most applicants are already in debt before applying for the loan
- **verification_status** - Most investors are ok to fund applicants whose income has not been verified. This might be an indicator that investors are preferring other categories (say rates, loan_status history etc) over this category.
- **loan_status** - Most loans are paid off which is good indicator for both Lending Club as well as investors.
- purpose - Majority of applicants are applying for loans for service existing debt. Top two categories are debt_consolidation and credit_card. This indicates most applicants and investors are taking on high risk
- **addr_state** - Top two states where most loans are funded are CA and NY. This could because of high cost of living or high per capita debt in these states. Another possibility is Lending Club's market penetration can be improved in states with lower applicants.

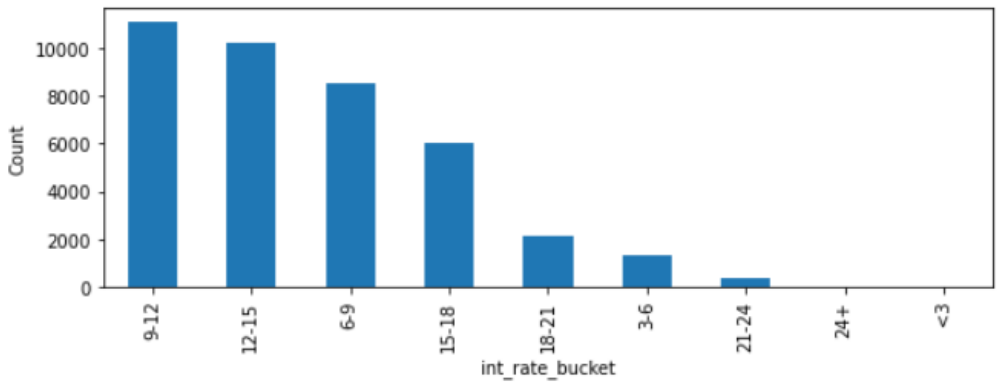
INITIAL UNIVARATE ANALYSYS

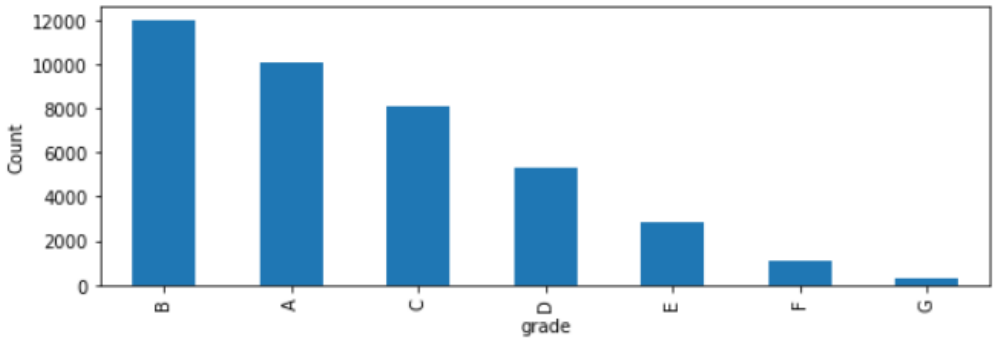
[All Loans]

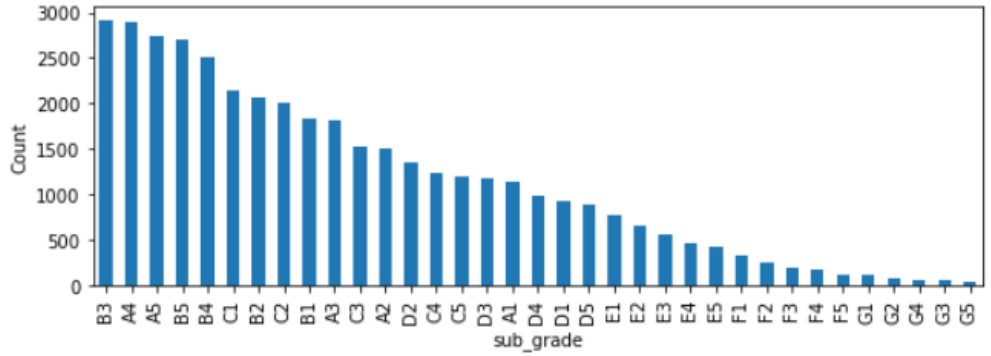
categorical columns

PLOTS:







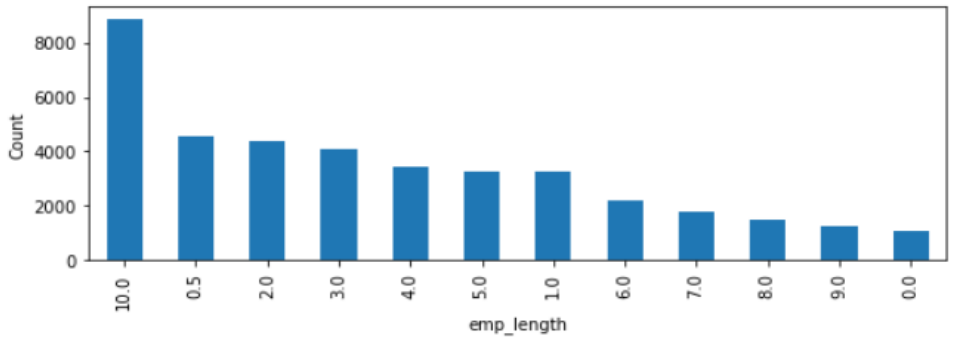


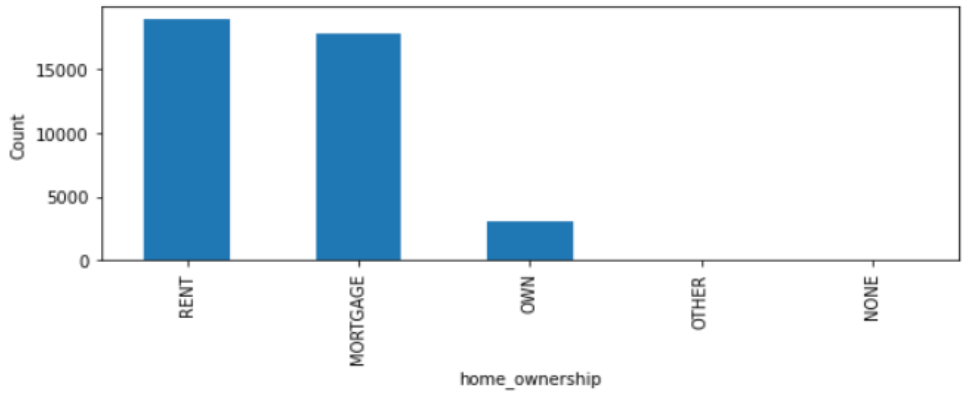
INITIAL UNIVARATE ANALYSYS

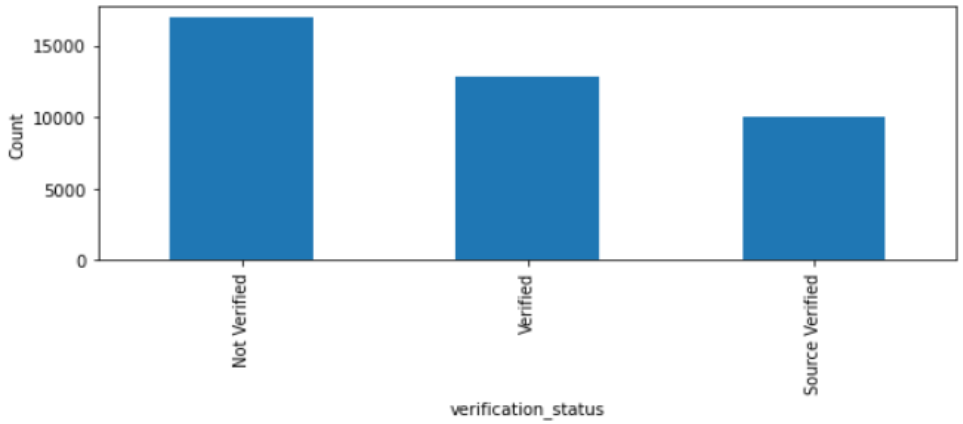
[All Loans]

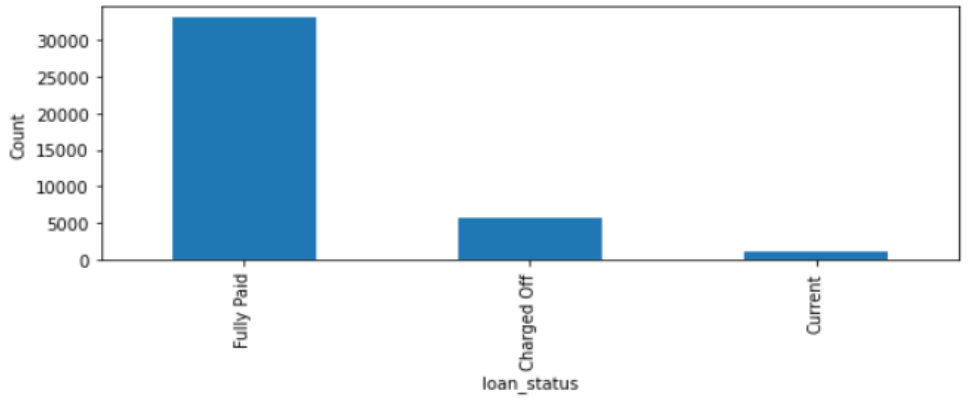
categorical columns

PLOTS:







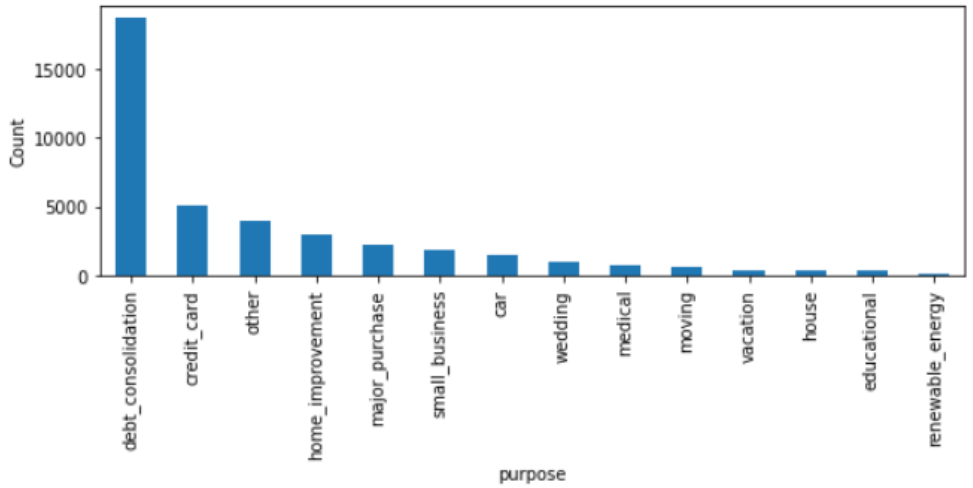


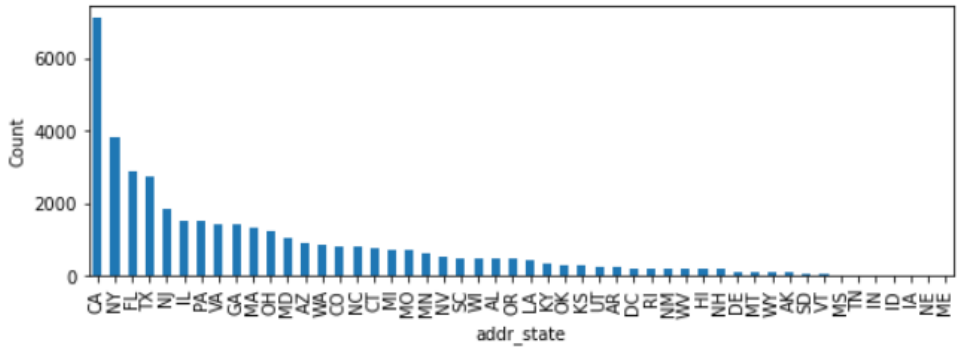
INITIAL UNIVARATE ANALYSYS

[All Loans]

categorical columns

PLOTS:

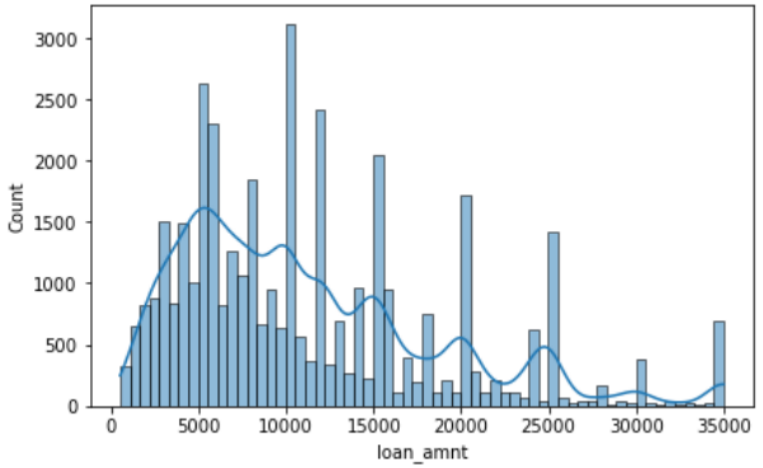


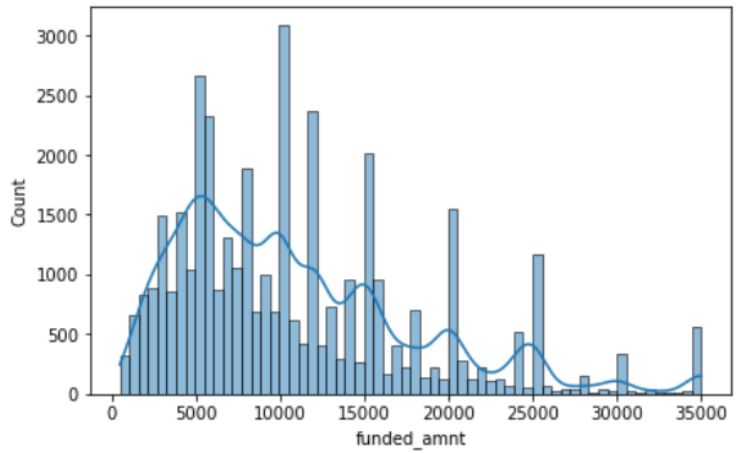


INITIAL UNIVARATE ANALYSYS
[All Loans]
Continuous Variables

INFERENCE:

- **loan_amnt** and **funded_amnt** - Majority of requested and approved loan amounts are between 400 to 15000
- **int_rate** - Majority of the loans have an interest rate between 6% to 18% as determined earlier.
- **dti** - dti has a gradual incline up to 15 and a gradual decline till 25 post which it is flat.
- **annual_inc_k** - Has big outliers. Majority of borrowers have incomes ranging from 40K to 82K
- **revol_bal** - Declines exponentially
- **total_acc** - Majority of total_acc are between 13 to 29
- **total_pymnt** and **total_pymnt_inv** - Are correlated as expected
- **total_rec_prncp** and **total_rec_int** - No significant conclusions yet



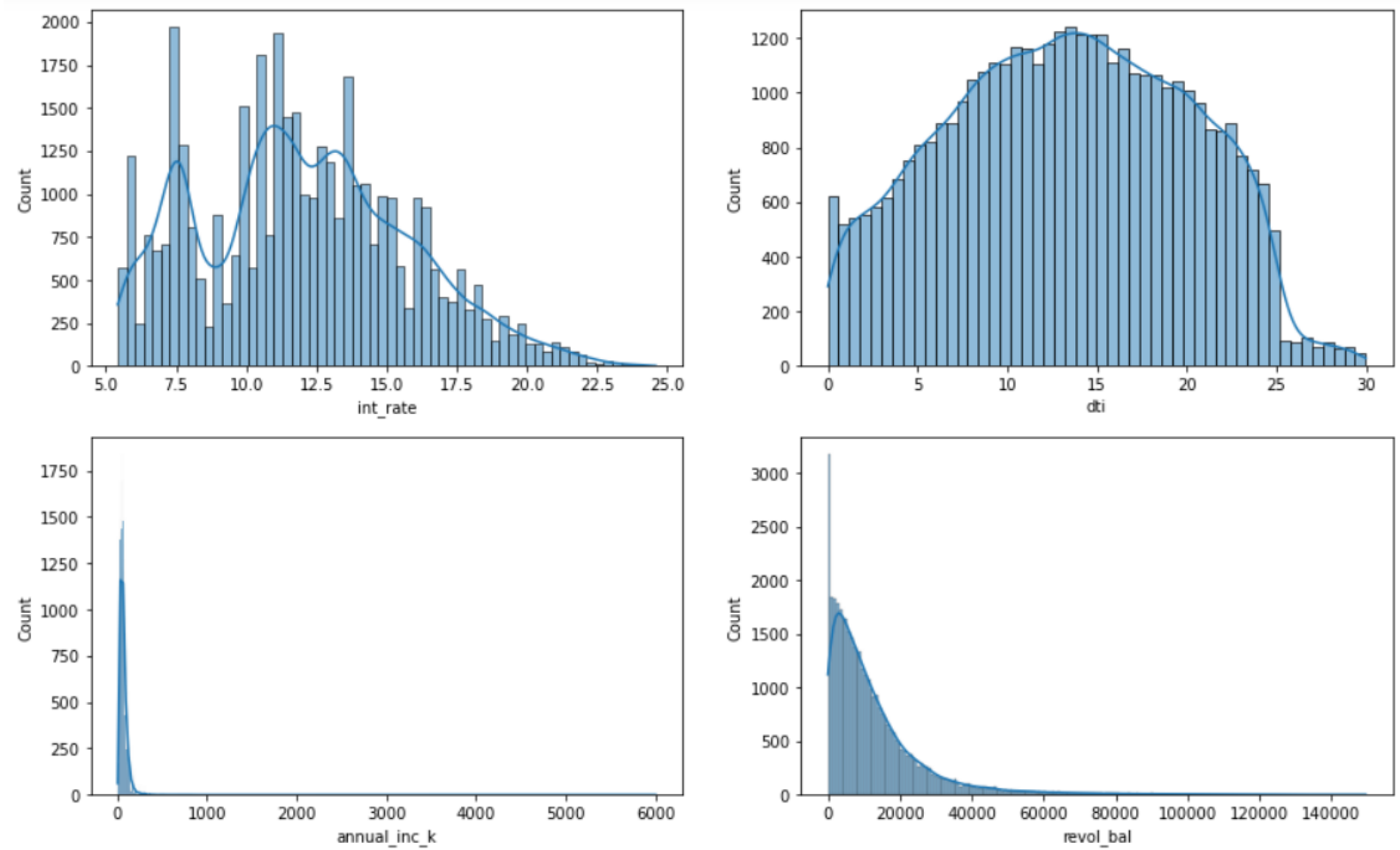


INITIAL UNIVARATE ANALYSYS

[All Loans]

Continuous Variables

PLOTS:

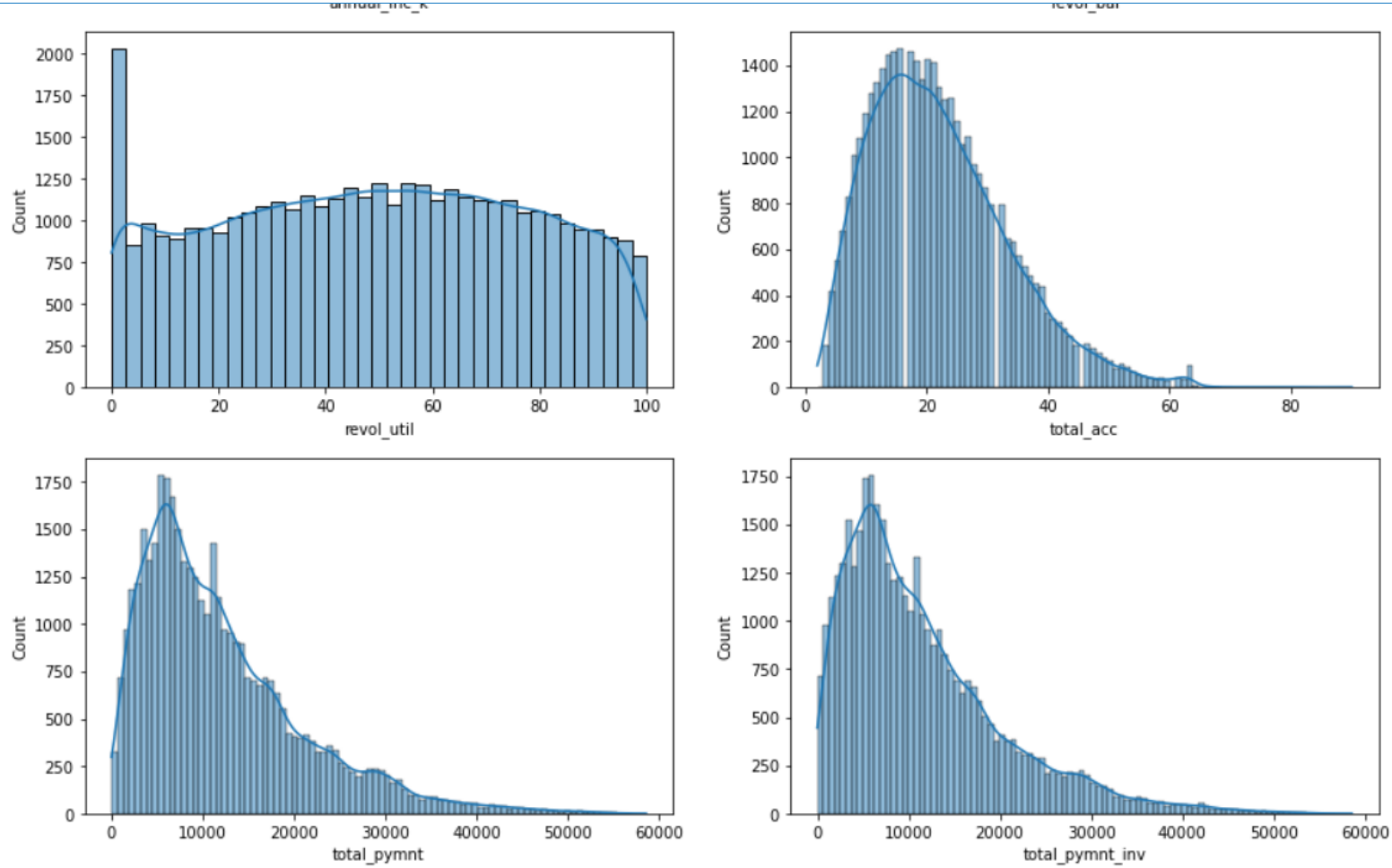


INITIAL UNIVARATE ANALYSYS

[All Loans]

Continuous Variables

PLOTS:

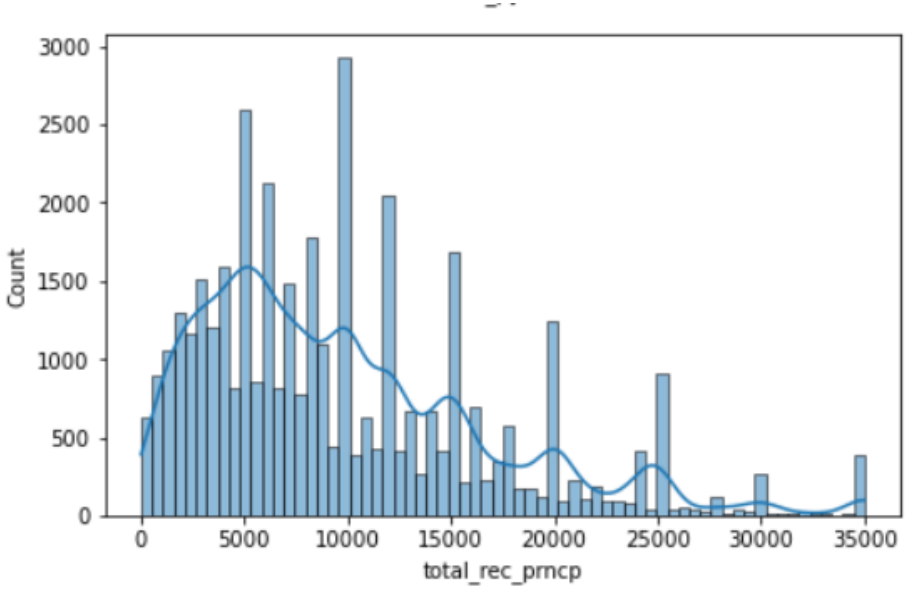


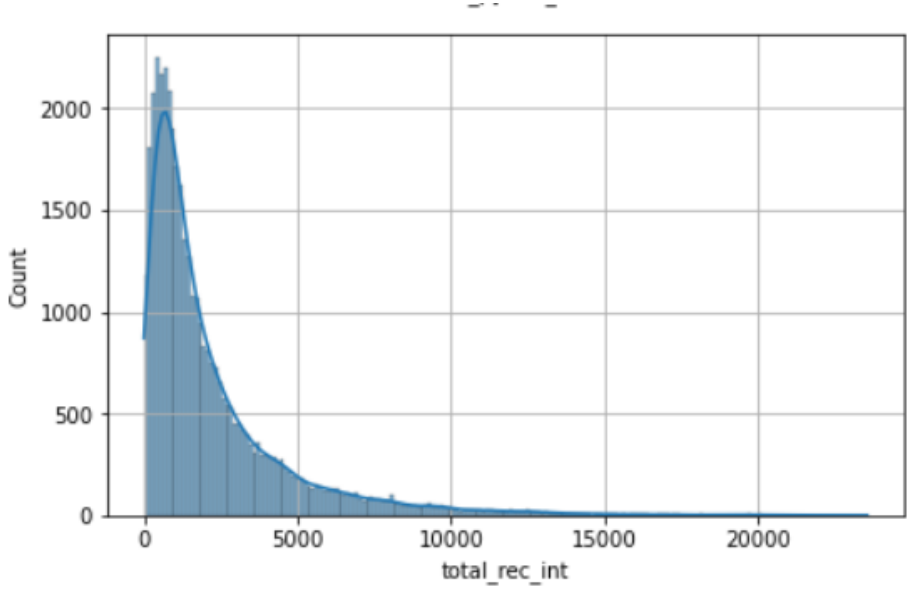
INITIAL UNIVARATE ANALYSYS

[All Loans]

Continuous Variables

PLOTS:





SEGMENTED UNIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

INFERENCE:

- **term** - Probablity of paying off the loan is more if the term is less. But, Delinquency may not be accurately predicted by this attribute alone.
- **int_rate_bucket** [**Good Criteria**] - Most pay-off and delinquencies are between 6% to 15%. However, ratio of payoff vs delinqunecy increases after 12%
- **grade** [**Good Criteria**] - Highest payoffs are for A and B as one would expect. Delinquency ratio to loans funded rises as we move from A to F as expected. Grade seems like a robust standalone criteria for predicting outcome.
- **emp_length** - Probability of both delinquency and pay-off is high for applicants with more than 10 years of employment. However, this may not be an accurate guess as data for 10+ years could have been classified into further bins but this data is unavailable. Volume of payoffs/delinquency is more for < 5 years and > 10 years
- **sub_grade** [Good Criteria] - Same as grades since this is a subset and outcomes are as one would expect.
- **home_ownership** [Good Criteria] - Probability of default is less of applicant owns a house. But loan volume is also comparatively low
- **verification_status** - Not sufficient to predict outcome. There is a high loan volume for 'Not Verified' which indicates that Lending Club should look into verification for these applicants
- **loan_status** - Most loans are paid off which is good indicator for both Lending Club as well as investors.
- purpose [**Good Criteria**] - Majority of applicants are applying for loans for service existing debt (debt_consolidation and credit_card). Likelyhood of predicting a default is high
- **addr_state** - Top three states where most defaults happen CA, NY and FL. This could because of high cost of living or high per capita debt in these states.

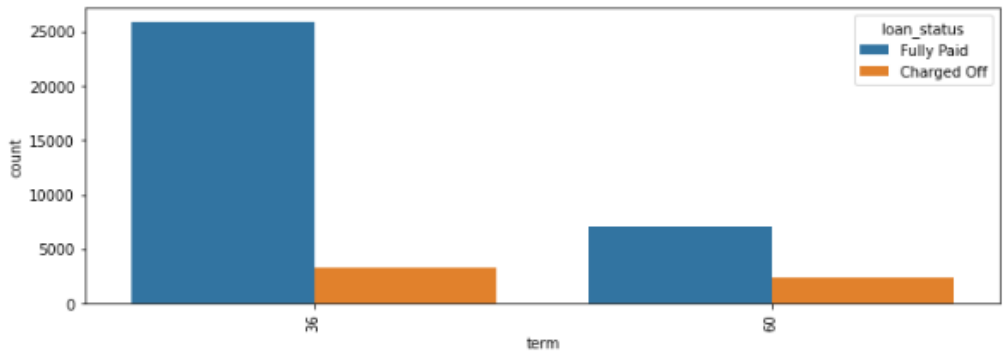
NOTE: Attributes not classified as good predictors of delinquency/payoff may still be useful if clubbed with other attributes

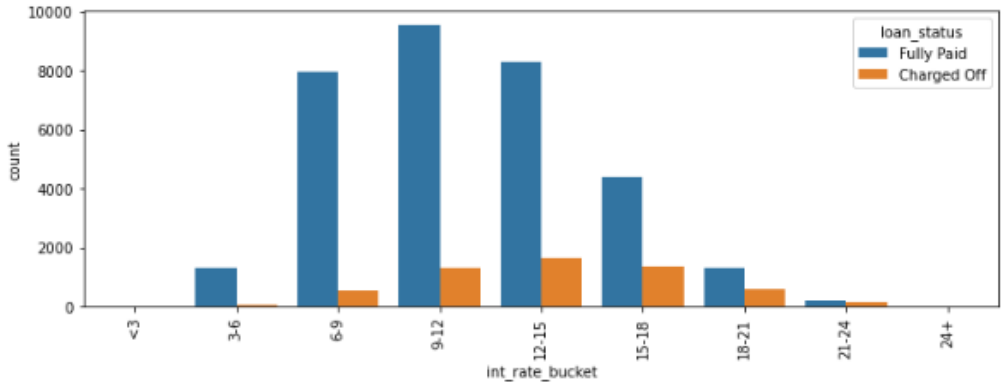
SEGMENTED UNIVARIATE ANALYSIS

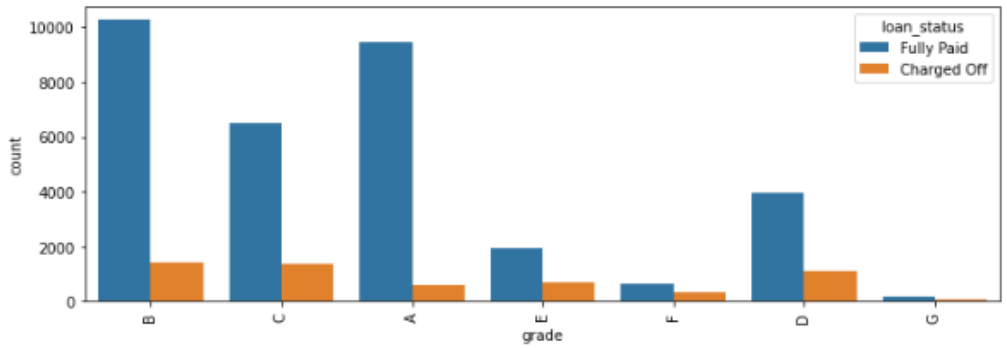
[loan status != Current]

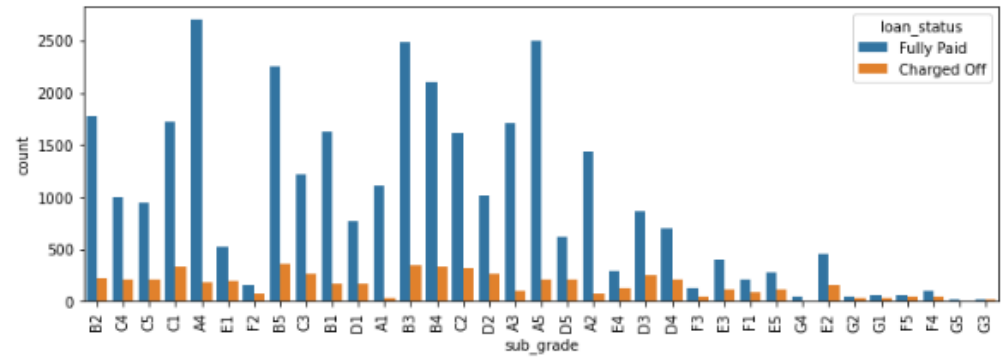
Categorical columns

PLOTS:







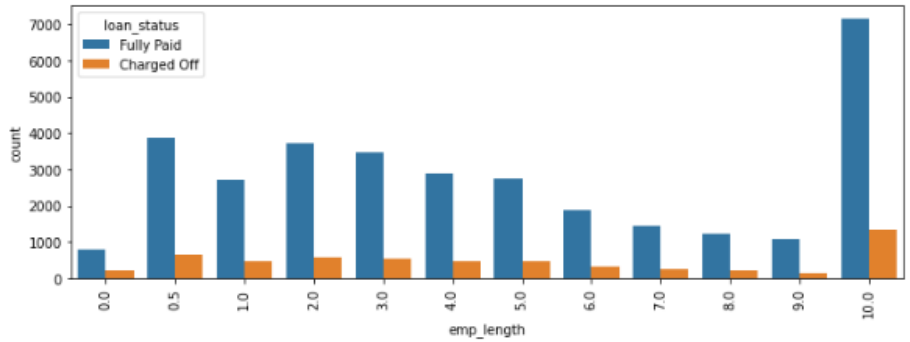


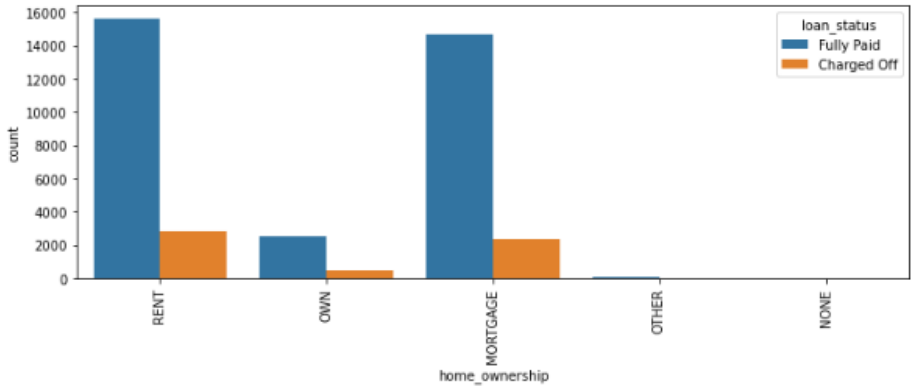
SEGMENTED UNIVARIATE ANALYSIS

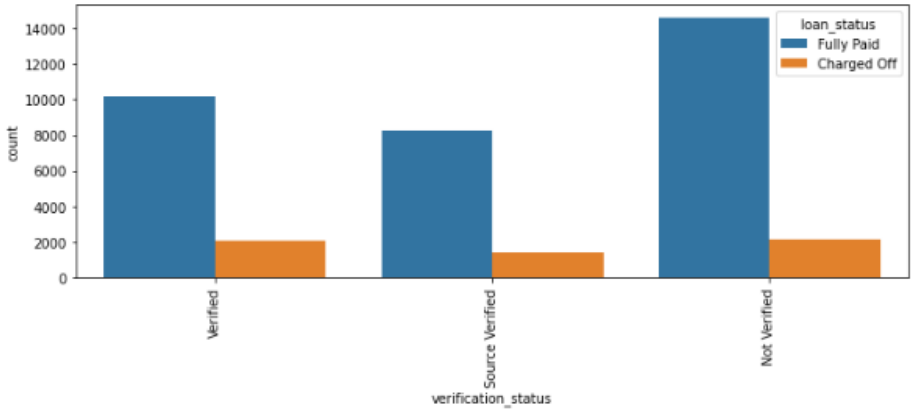
[loan status != Current]

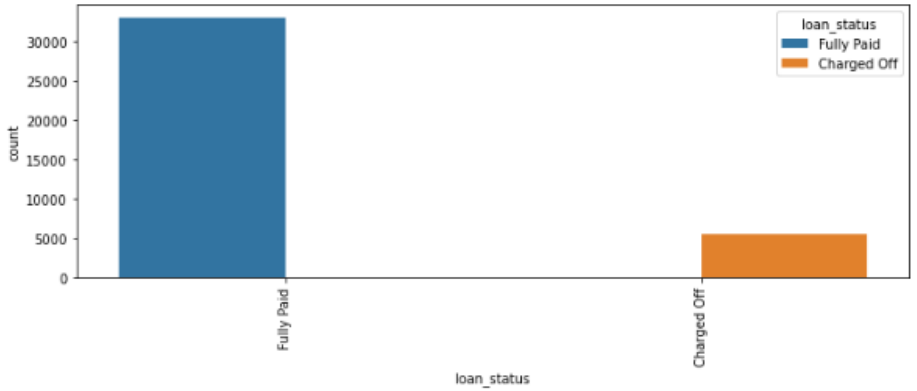
Categorical Variables

PLOTS:







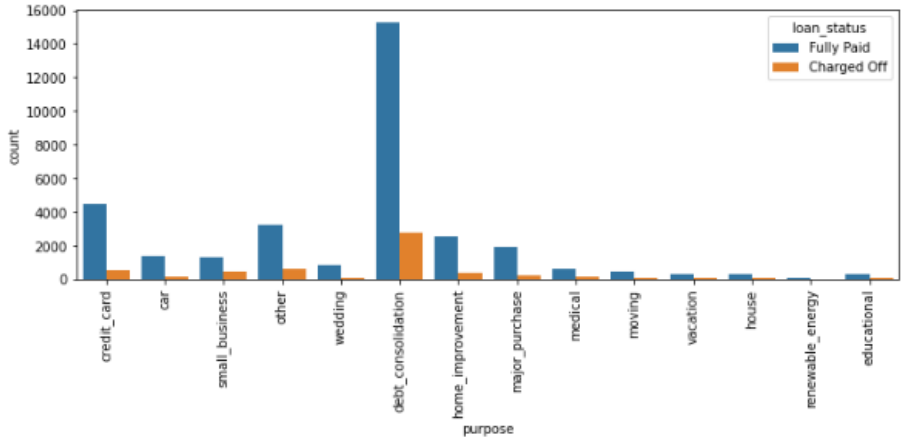


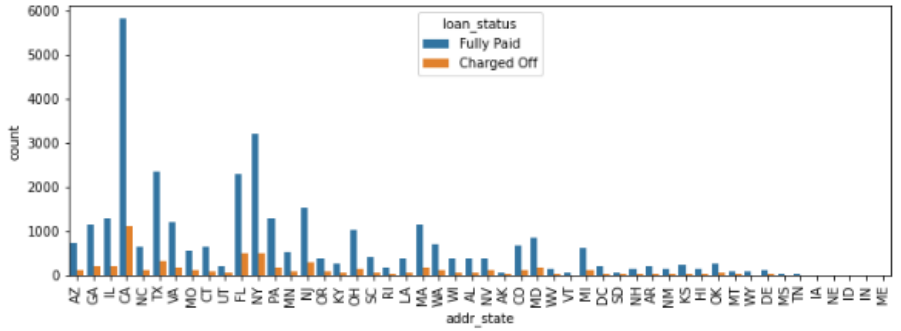
SEGMENTED UNIVARIATE ANALYSIS

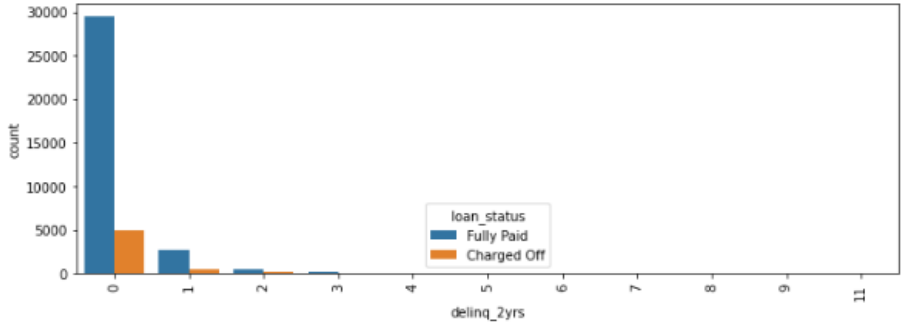
[loan status != Current]

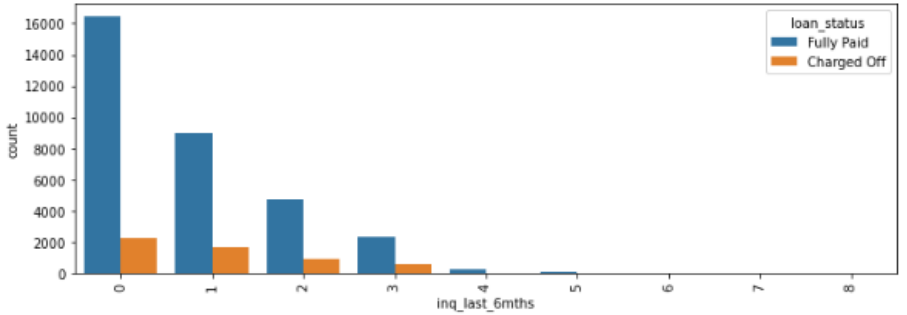
Categorical Variables

PLOTS:







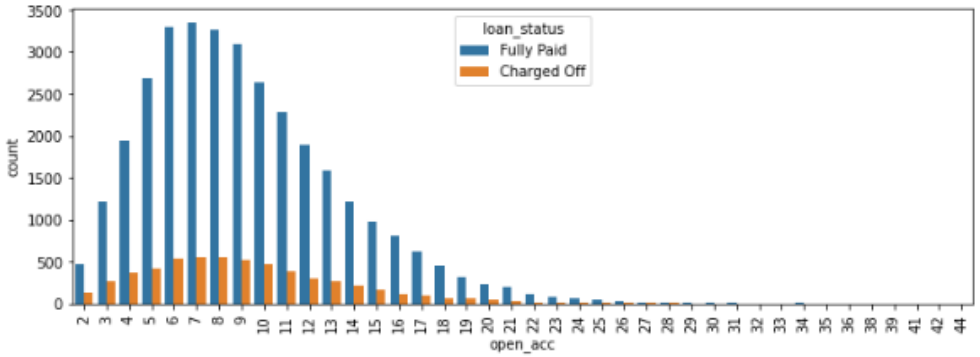


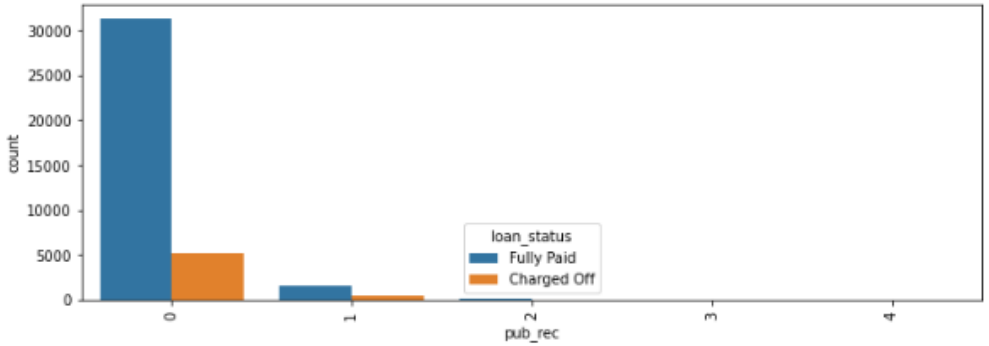
SEGMENTED UNIVARIATE ANALYSIS

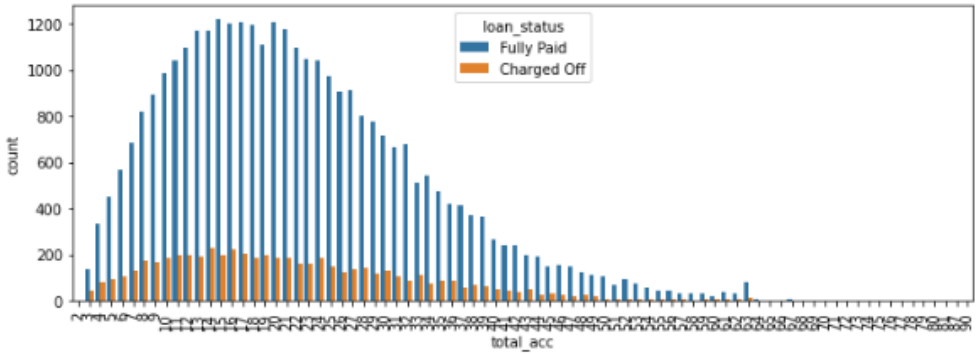
[loan status != Current]

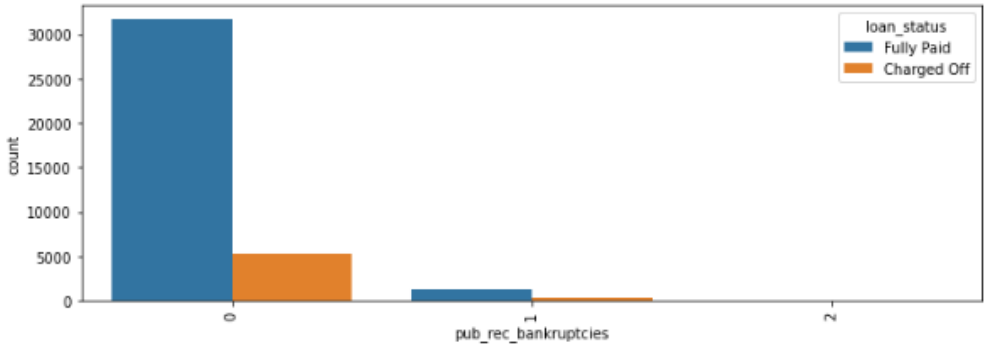
Categorical Variables

PLOTS:









SEGMENTED UNIVARIATE ANALYSIS

[loan status != Current]

Continuous Variables

INFERENCE:

- **funded_amnt** - Chances of default increases with funded_amount. But this not a good predictor alone. Perhaps, higher amounts should not be loaned if other factors determine possibility of default.
- **dti** [**Good Criteria**] - As expected, a higher dti ratio may lead to default and this is a good predictor
- **annual_inc_k** [**Good Criteria**] - As expected, higher incomes lead to less default. This is also a good criteria
- **revol_bal** - Moderate chance of default if revolving balance is more but not a good criteria
- **revol_util** [Moderate Criteria] - Good chance of default if the revol_util is high. Note that this could be a good predictor of default only once the loan is current and a few payments have been made. This information may not be available prior to loan is disbursed.
- **total_acc** - Moderate chance of default if total accounts is less but not a good criteria
- **total_pymnt, total_pymnt_inv, total_rec_prncp and total_rec_int** - Cannot be used as a predictor as a portion of the payments will not be made in case of default

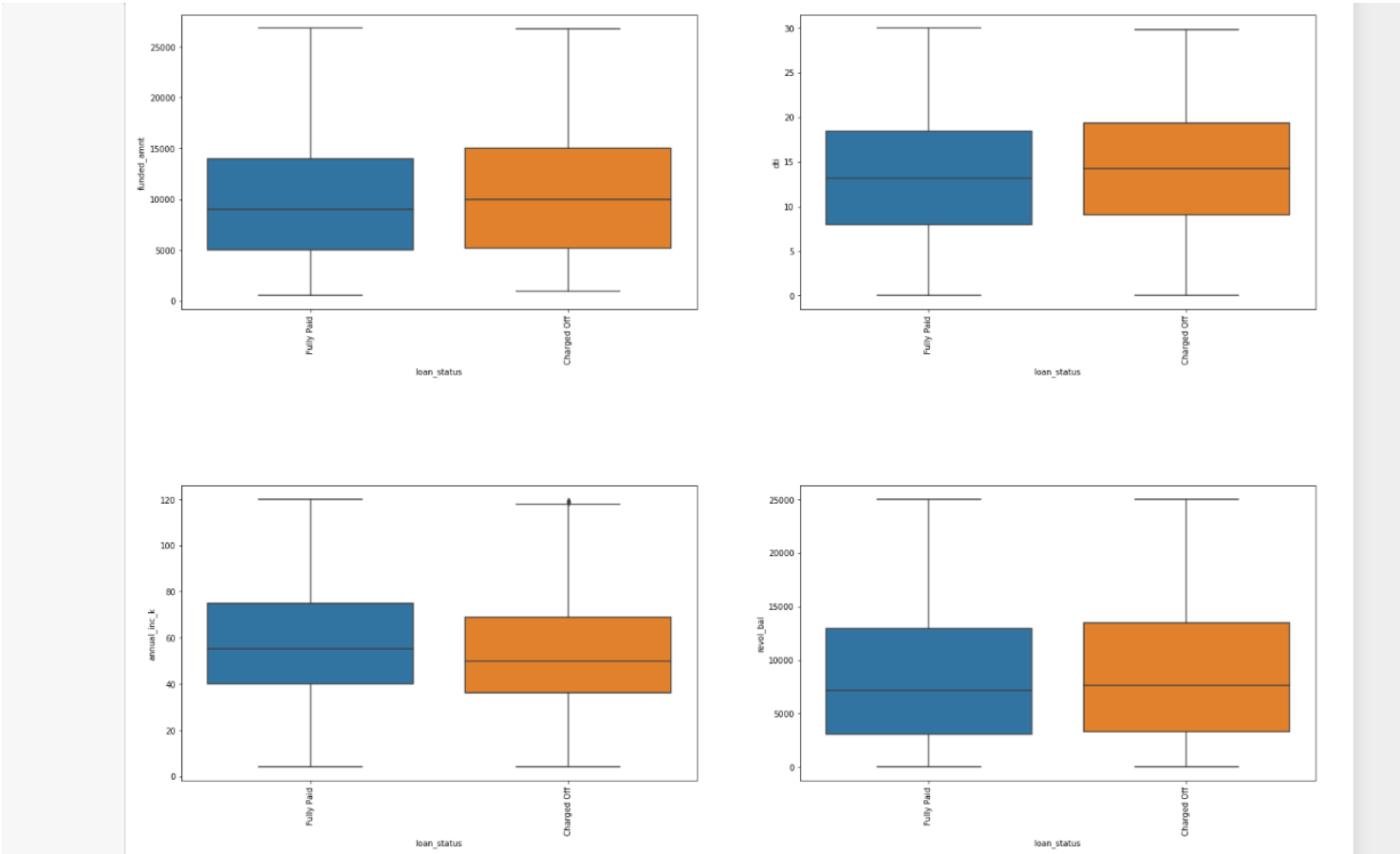
NOTE: Some applicant behaviours take effect after the loan is provided and hence may not be used accurately to predict outcomes of new loans

SEGMENTED UNIVARIATE ANALYSIS

[loan status != Current]

Continuous Variables

PLOTS:

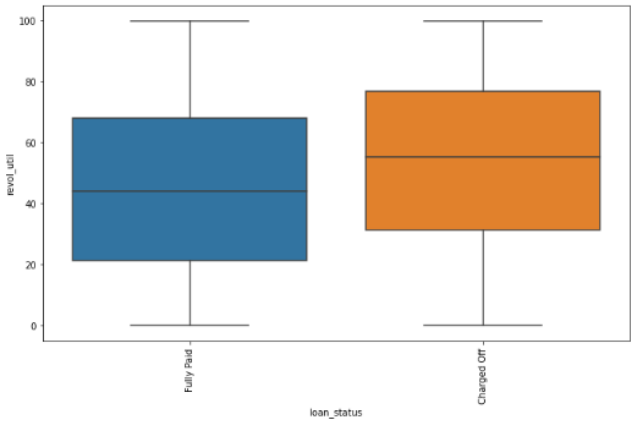


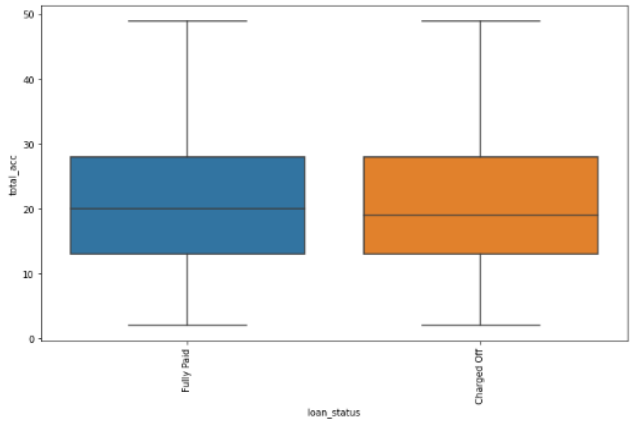
SEGMENTED UNIVARIATE ANALYSIS

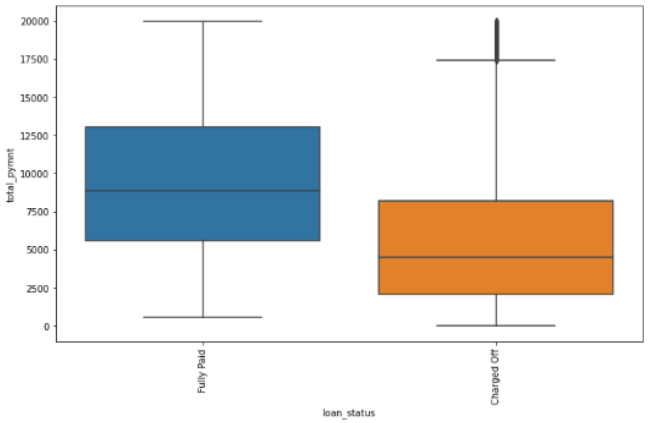
[loan status != Current]

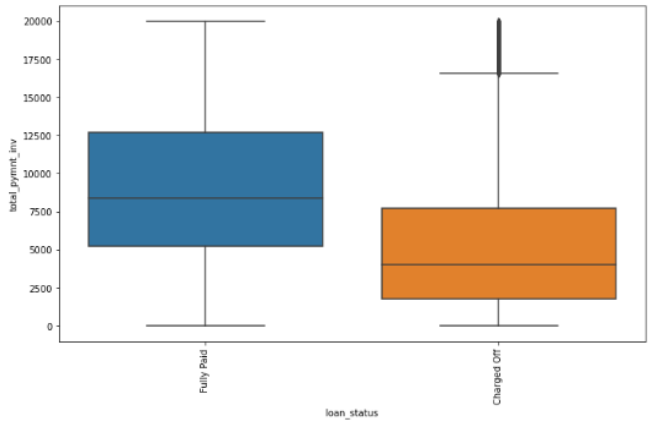
Continuous Variables

PLOTS:







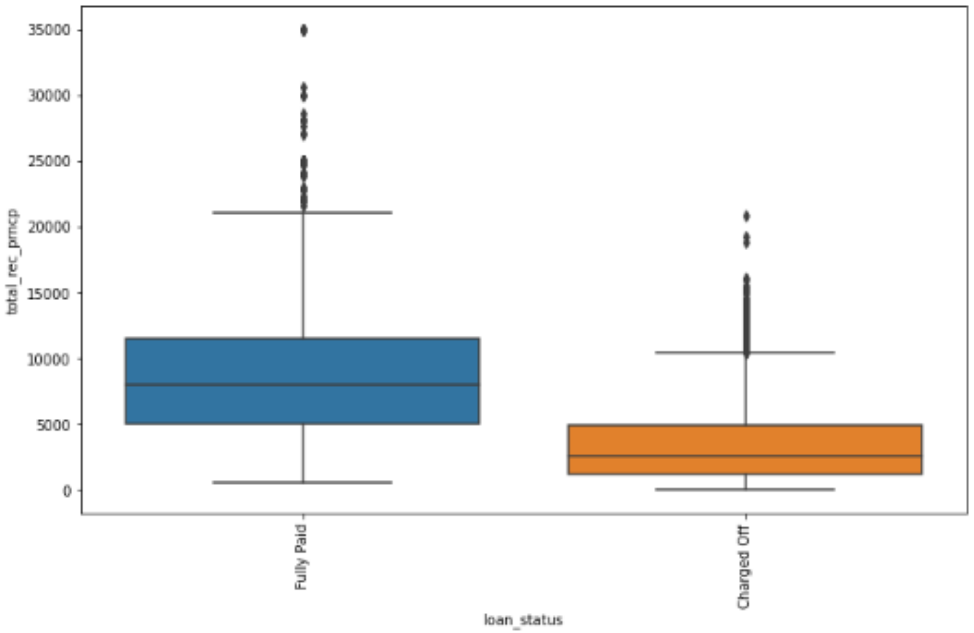


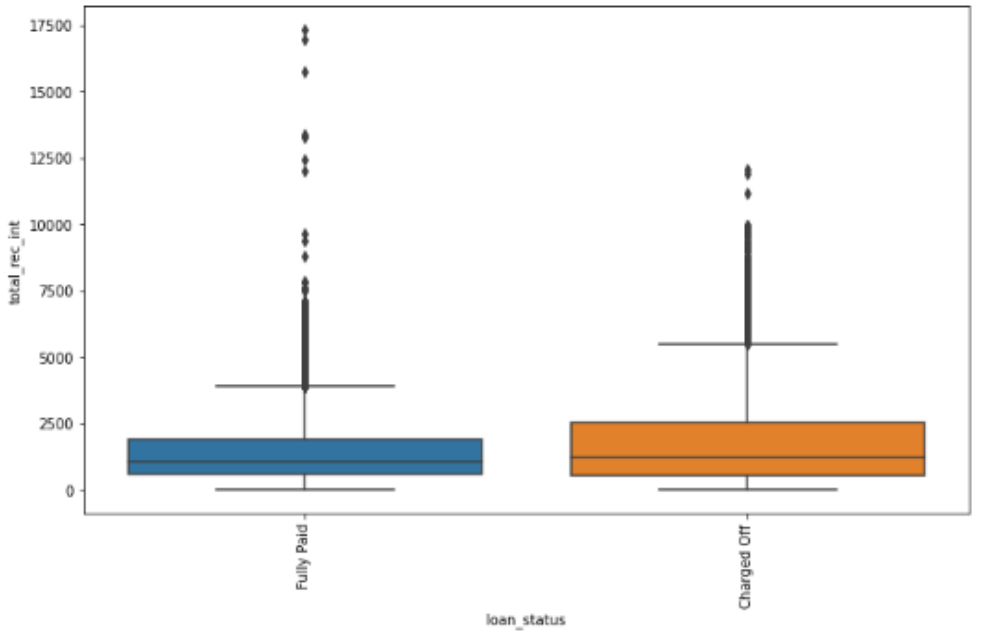
SEGMENTED UNIVARIATE ANALYSIS

[loan status != Current]

Continuous Variables

PLOTS:





BIVARIATE ANALYSIS
[loan status != Current]
Continuous Variables

INFERENCE:

- The [Good Criteria] Continious columns [annual_inc_k, dti, revol_util] seem to predict the credit worthiness of the applicant. This analysis will figure out the correlation of these columns compared to other continuous attributes
- **NOTE:** This analysis considers data

1. where 'loan_status' != 'Current'
2. where 'loan_status' == 'Charged Off'

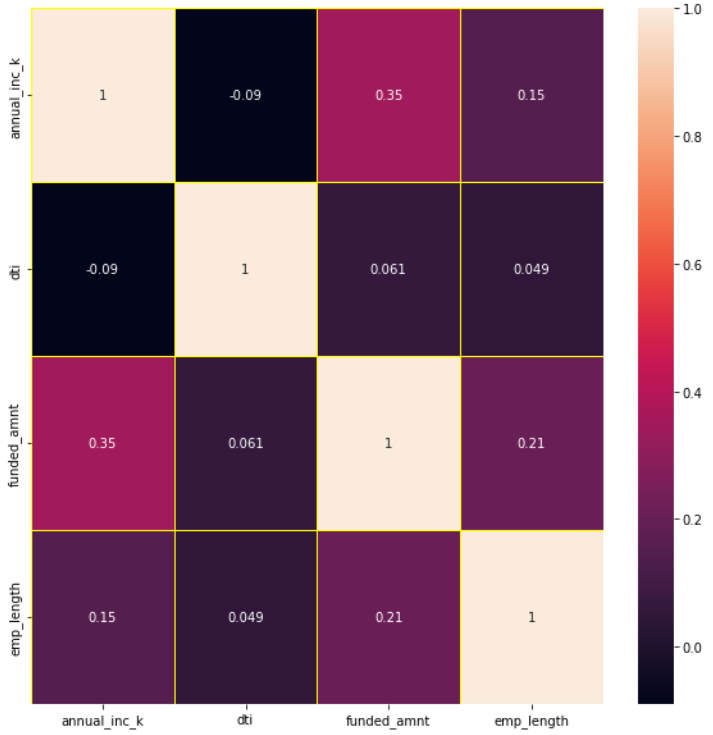
- funded_amnt vs annual_inc_k - are positively correlated which is good. Loan approval should factor in annual income. However, Correlation increases for Charged Off loans.

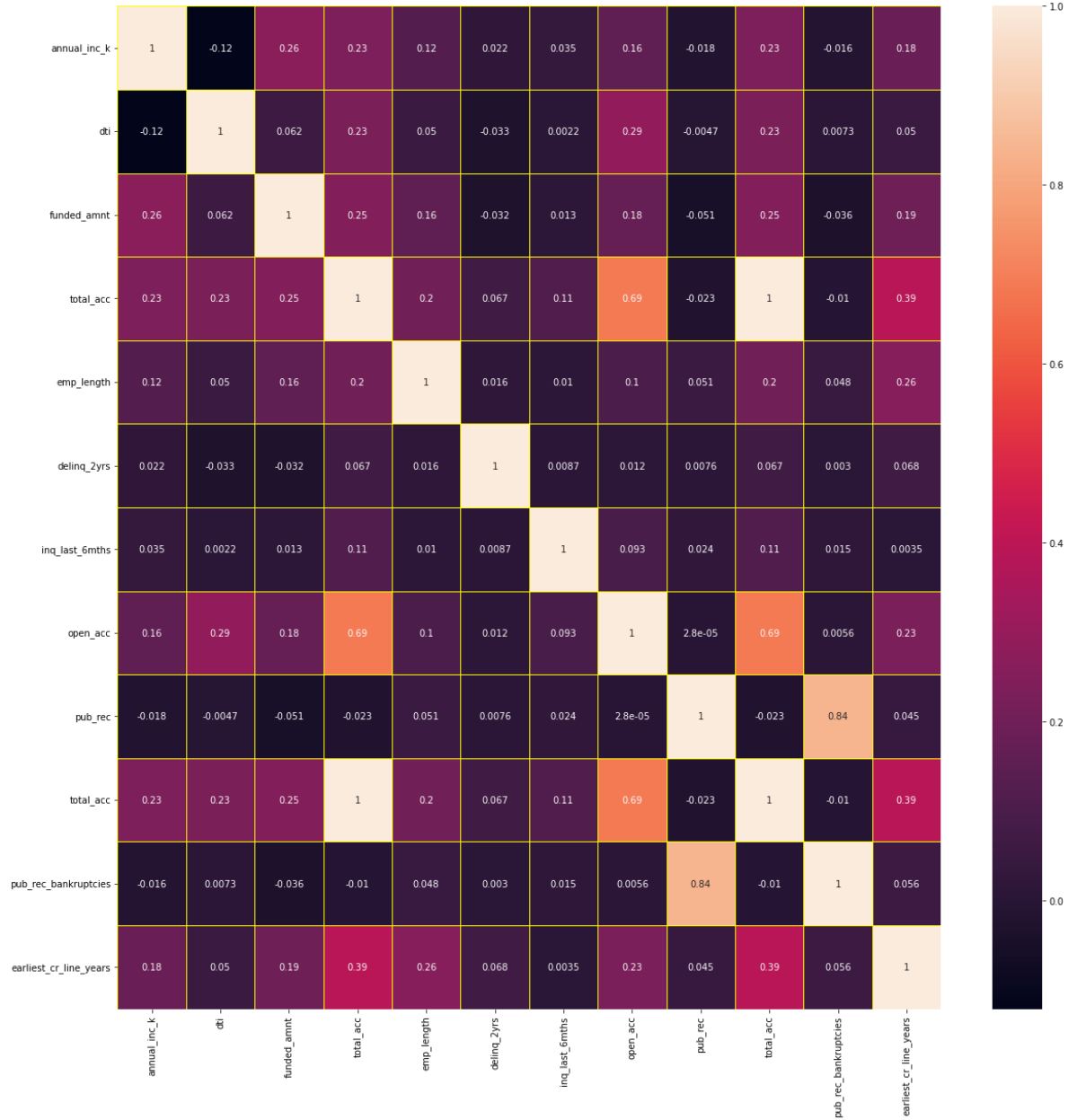
- funded_amnt vs dti - This ratio seems low (including defaults) indicating dti is not a major factor in approving higher loan amount. Perhaps this needs to factor in more while approving loans
- funded_amnt vs emp_length - Ratio is positive but is comparitively low indicating that employment length may not be a major criteria for loan approval
- pub_rec_bankruptcies and pub_rec - Correlation of these attributes are low which is preferred. This indicates less loans are approved with applicants having derogatory public records or history of bankruptcy

NOTE: Some high correlations are ignored as they are obvious

BIVARIATE ANALYSIS
[loan status != Current]
Continuous Variables

PLOTS:





BIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

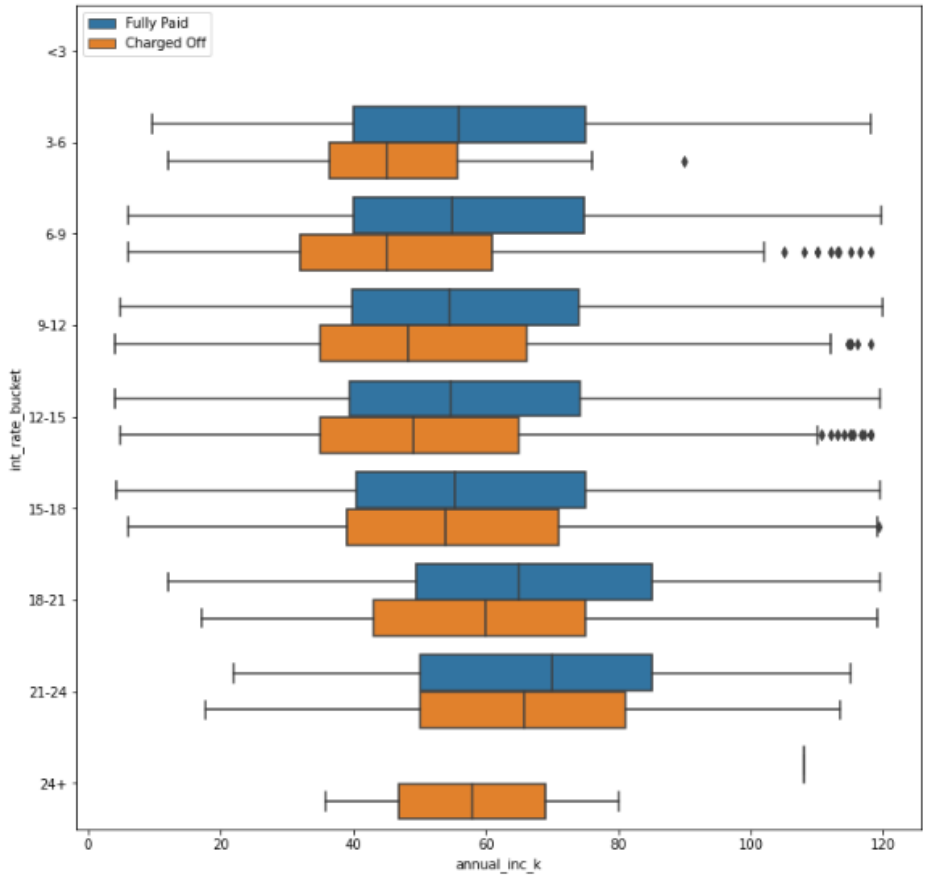
INFERENCE:

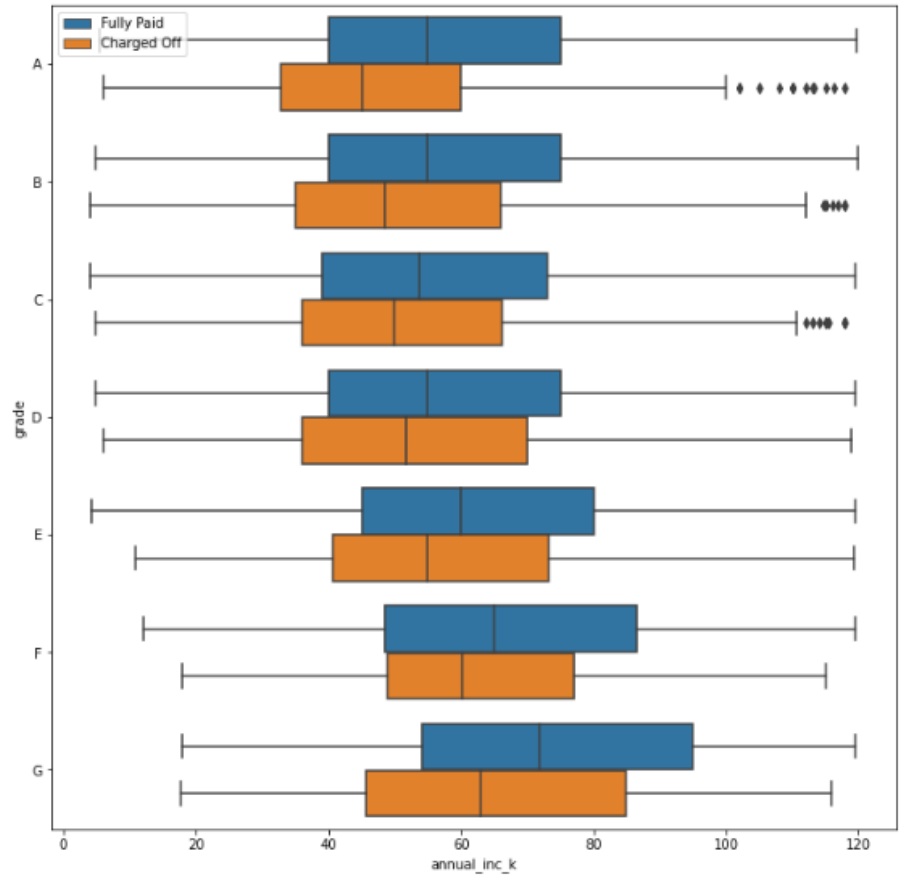
- **annual_inc_k vs int_rate_bucket** - Higher Interest rate is a stronger candidate for default than income.
- **annual_inc_k vs grade** - Higher grade is a stronger candidate for default than income as already guessed.
- **annual_inc_k vs home_ownership** Highest default still shows for Mortgages despite higher income
- **annual_inc_k vs purpose** We earlier saw most of defaults were happening for debt consolidation and credit cards. Within the sub-categories of home improvement and small businesses, defaults are high.
- **dti vs int_rate_bucket** Defaults increases with interest rates as dti increases
- **dti vs grade higher** dti may have higher defaults
- **dti vs home_ownership** Higher defaults are seen as dti increases for Loan, Rent and Mortgage

- **dti vs purpose** high for or debt consolidation and credit cards. Loans taken for education purpose happen even for lower dti.
- **funded_amnt vs int_rate_bucket** - Higher Interest rate increases with funded amount as guessed earlier. However, for interest rate greater than 18%, the funded amount seems to increase along with high defaults. Reduced funded amounts can be considered for such categories to reduce default
- **funded_amnt vs grade** - Same conclusion as int_rate_bucket
- **funded_amnt vs home_ownership** - Consistent with earlier findings but reduced funded amount can be considered for 'Other' category as funded amounts seem high along with defaults
- **funded_amnt vs purpose** - Funded amount seems high for debt consolidation, credit cards and small businesses. Reduced funded amounts can be considered for such categories to reduce default

BIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

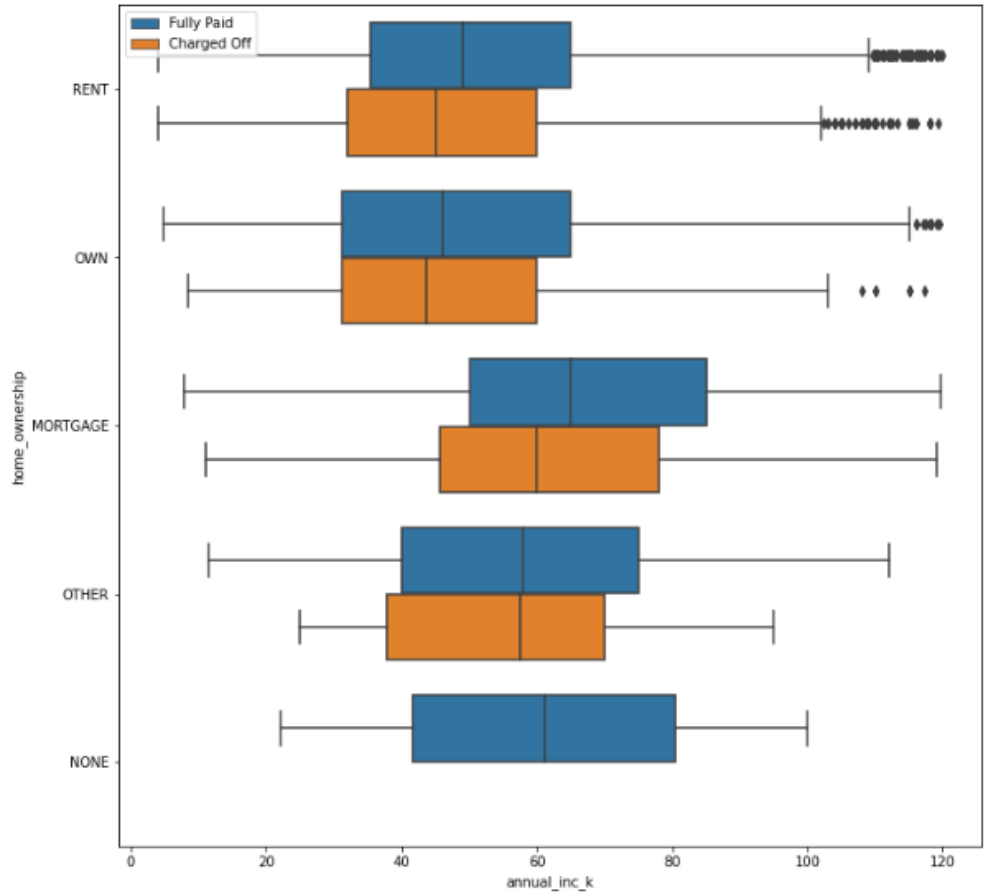
PLOTS:

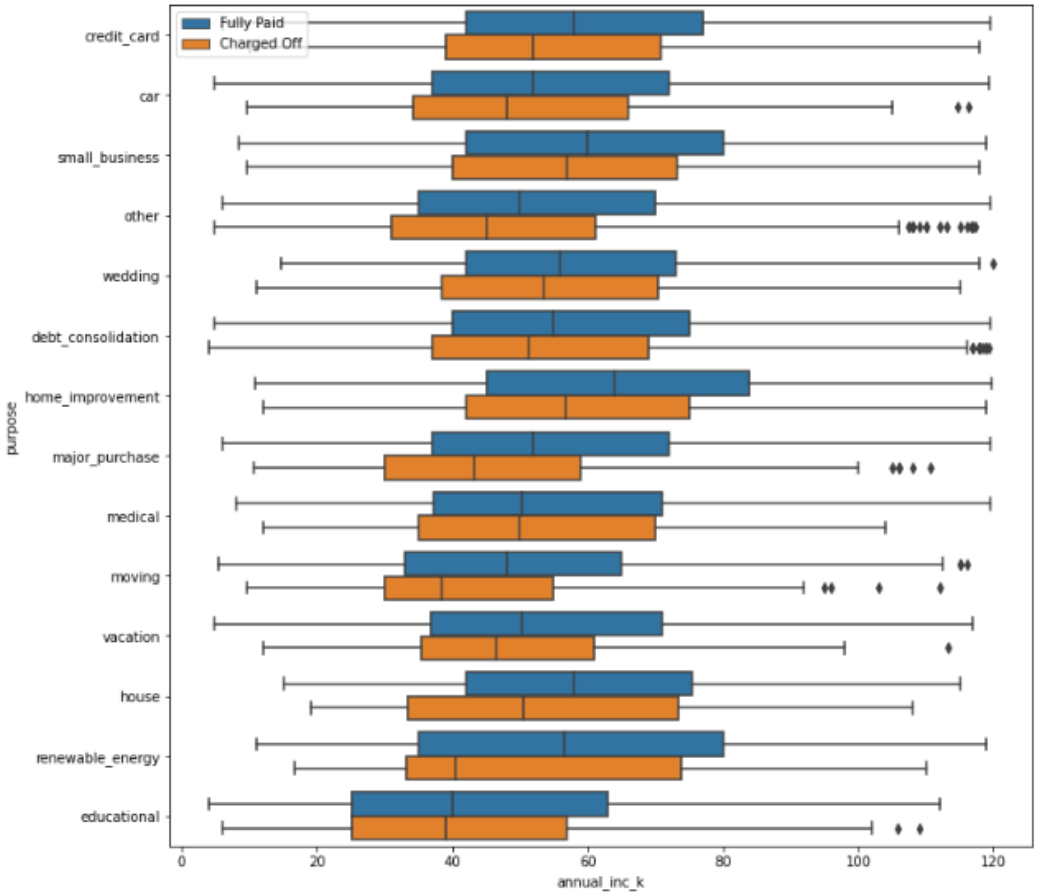




BIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

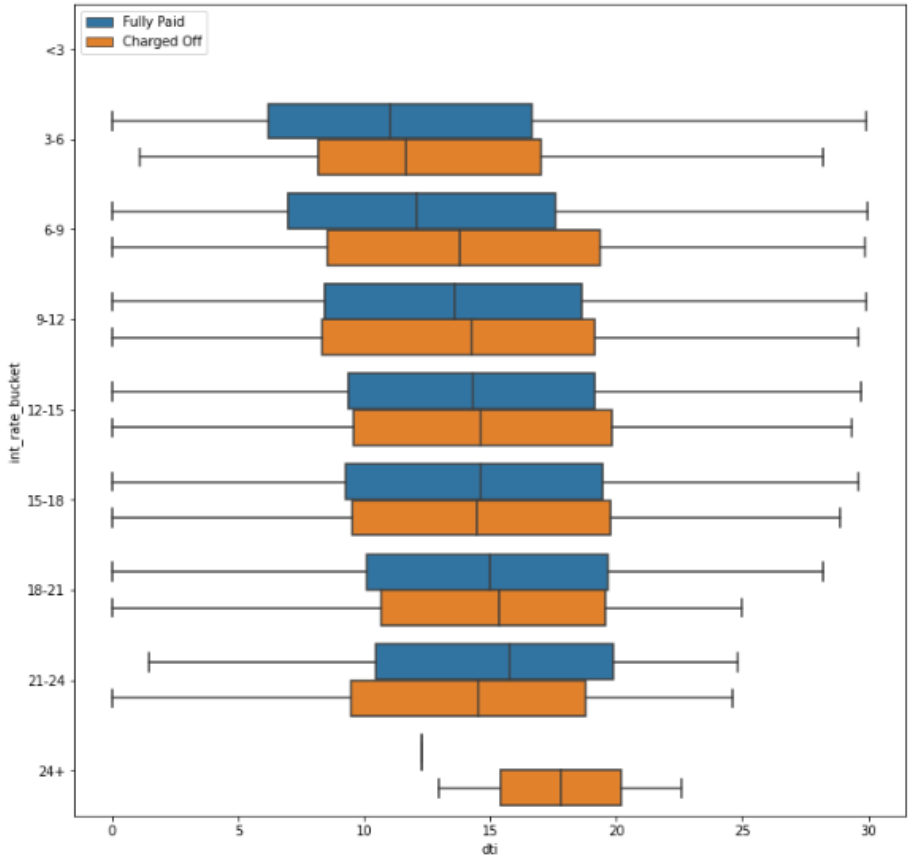
PLOTS

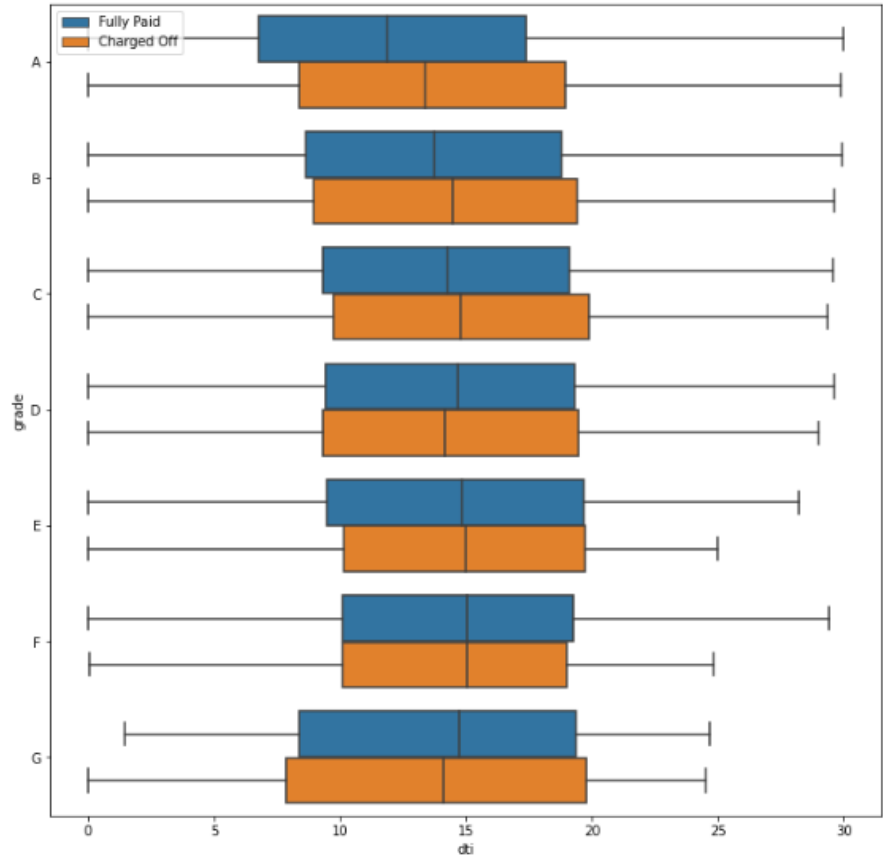




BIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

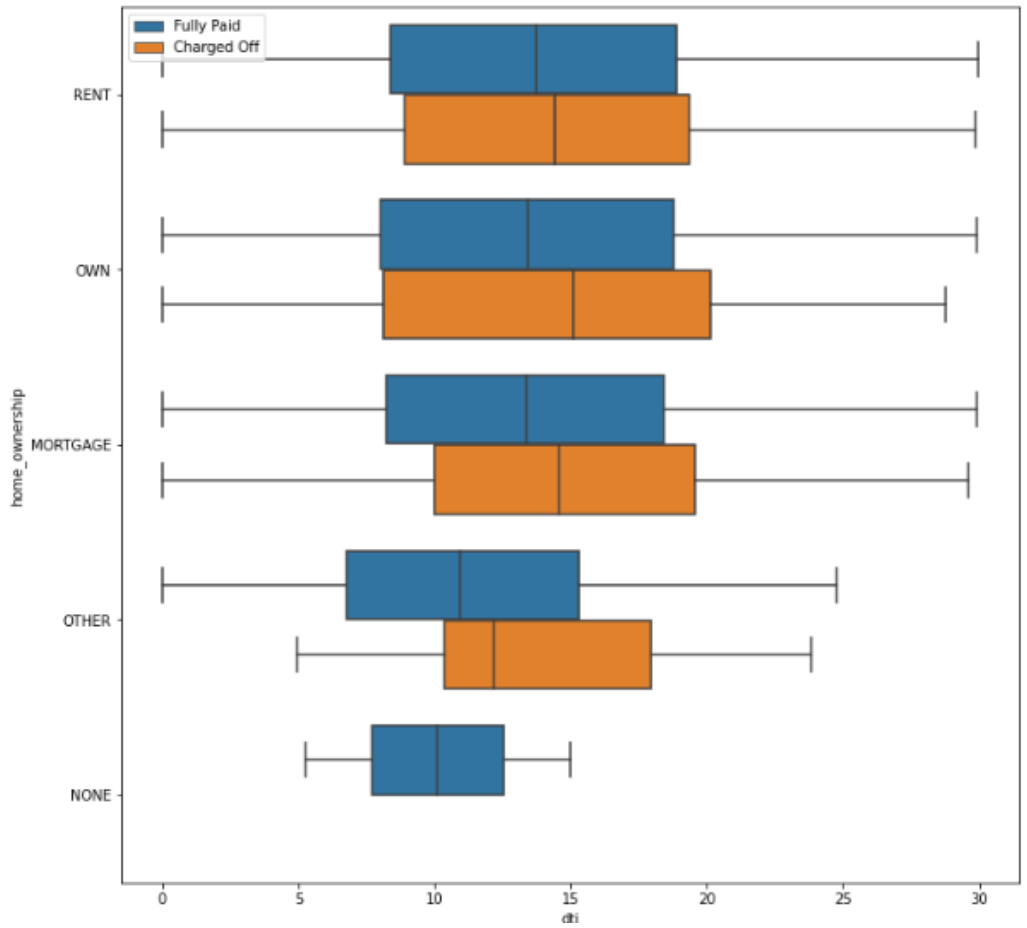
PLOTS:

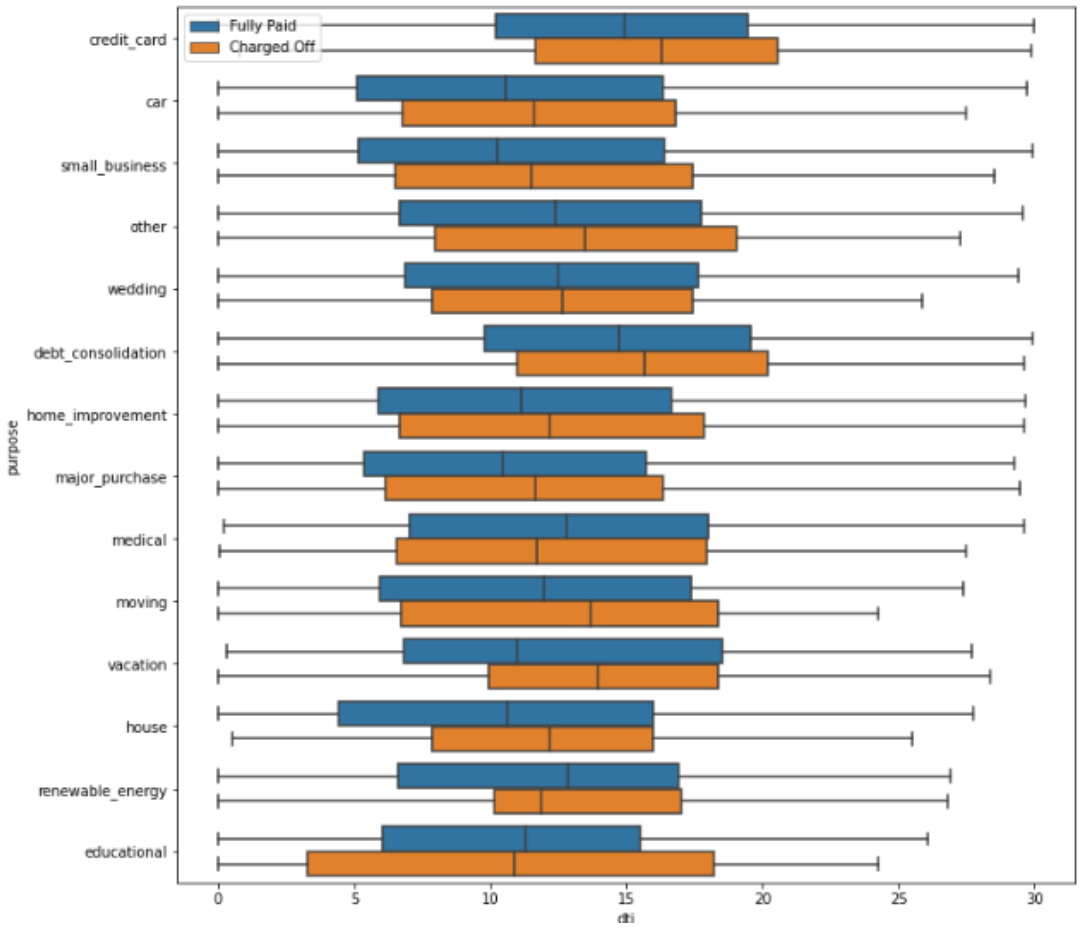




BIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

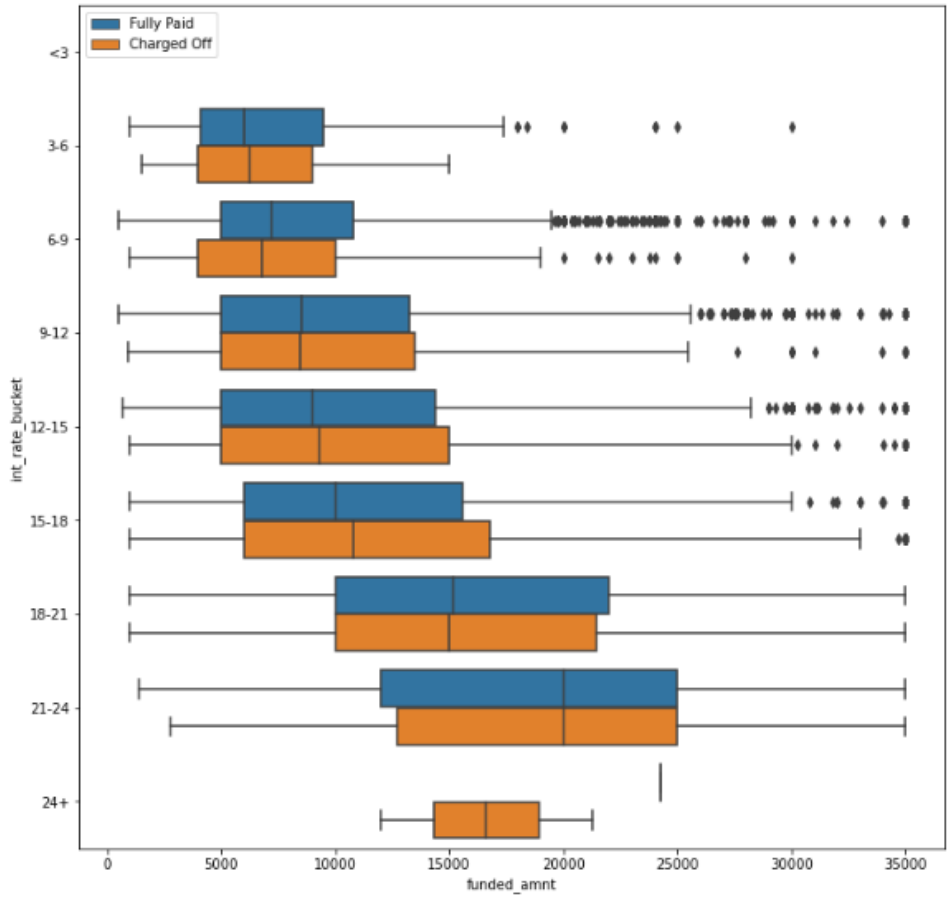
PLC

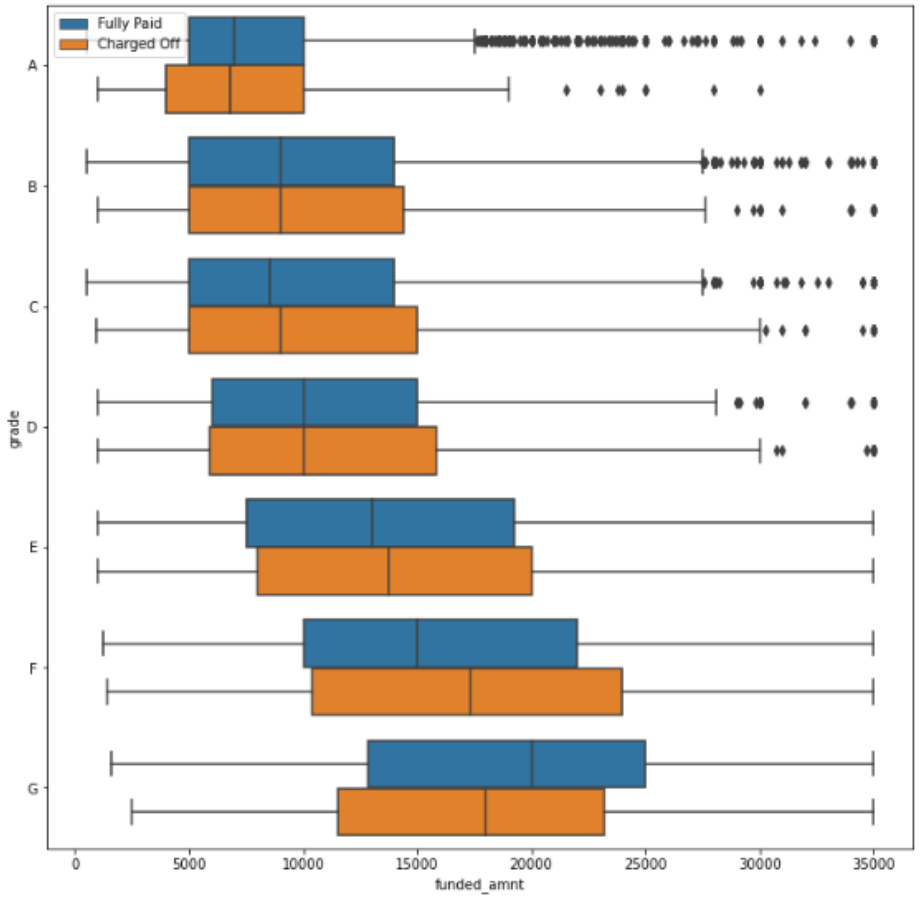




BIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

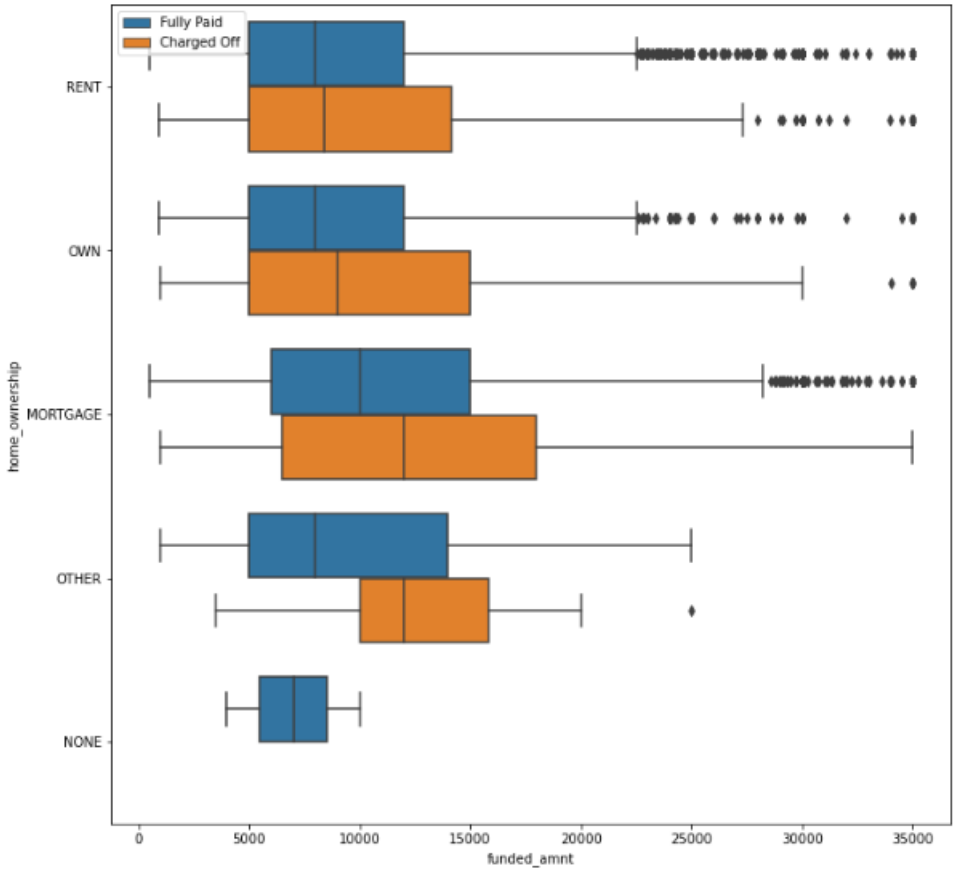
PLOTS:

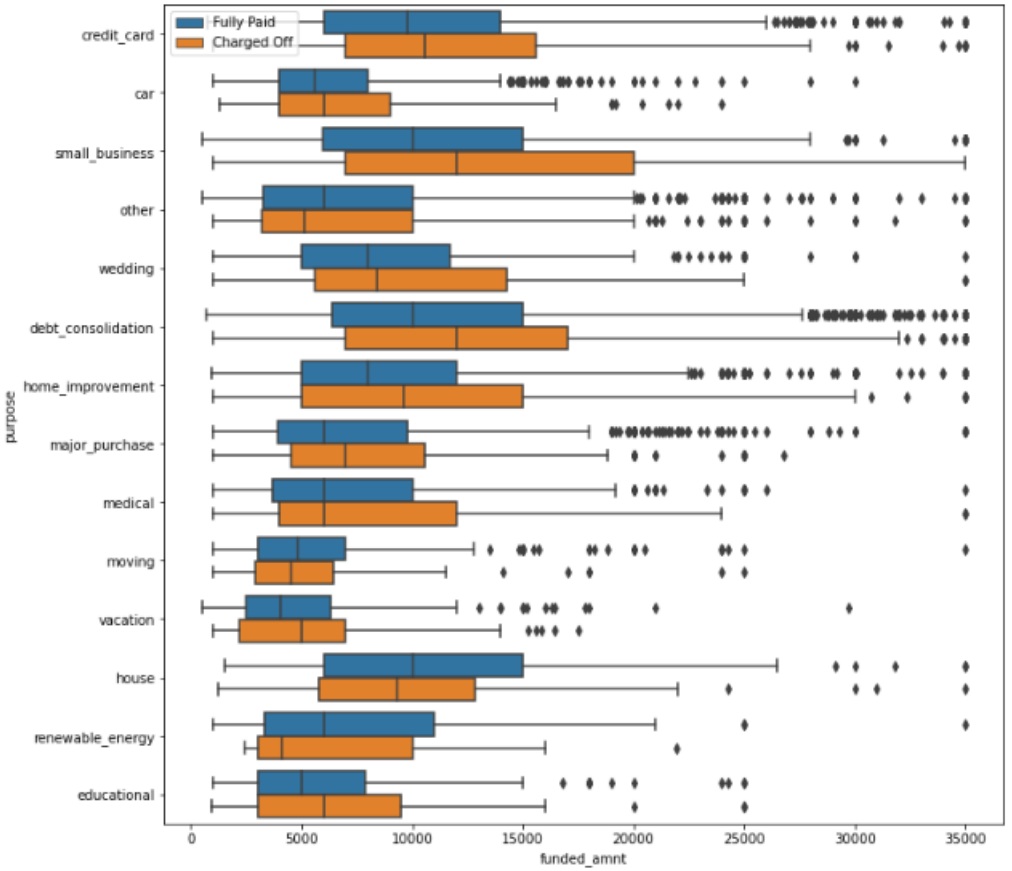




BIVARIATE ANALYSIS
[loan status != Current]
Categorical Variables

PLOTS:





FINAL CONCLUSION

The following points can be factored to determine risks of lending as well help reduce defaults.

- **int_rate_bucket** :
  - Delinquency increases significantly post 12% indicating higher risk.
  - For interest rate greater than 18%, the funded amount seems to increase along with high defaults. Reduced funded amounts can be considered for such categories to reduce default
- **grade** and **sub_grade** :
  - Lower grades is a robust criteria for predicting default
  - For interest rate greater than 18%, the funded amount seems to increase for lower grades. Reduced funded amounts can be considered for such categories to reduce default
- **home_ownership**:
  - Probability of default is less of applicant owns a house. But loan volume is also comparatively low
  - Highest default is for Mortgages and this criteria can be flagged as high risk
  - Higher defaults are seen as dti increases for Loan, Rent and Mortgage which is another criteria to look into
  - Reduced funded amount can be considered for 'Other' category as funded amounts seem high along with defaults
- **purpose**:
  - Most of defaults happen for debt refinincing i.e. debt consolidation and credit cards.
  - Within the sub-categories of home improvement and small businesses, defaults are high.
  - Loans taken for education purpose happen even for lower dti.

FINAL CONCLUSION

- **dti**:
  - A higher debt to income ratio may lead to default and this is a good predictor
  - dti may not be a major factor in approving higher loan amount. Perhaps this needs to factor in more while approving loans
  - Higher defaults are seen as dti increases for Loan, Rent and Mortgage categories of home ownership.
  - Reduced funded amount can be considered for 'Other' category of home ownership as funded amounts seem high along with defaults
- **annual_inc_k**:
  - Higher incomes lead to less default. This is also a good criteria to assess risk
- **addr_state**:
  - Top three states where most defaults happen CA, NY and FL. This could because of high cost of living or high per capita debt in these states.