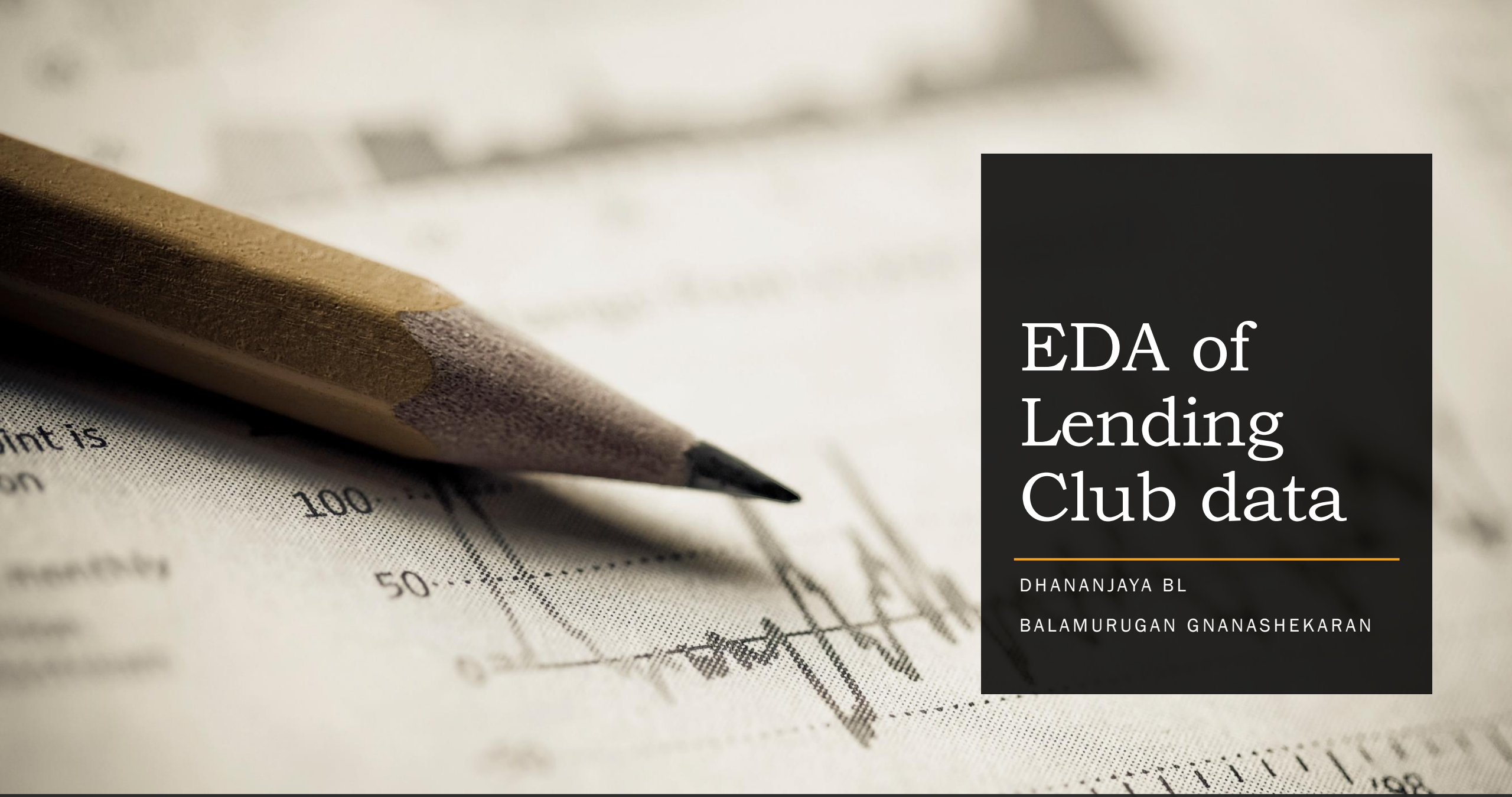


EDA of Lending Club data

DHANANJAYA BL

BALAMURUGAN GNANASHEKARAN

Content

- Overview
- Problem Statement
- Analytical Approach
- Observations and Findings
- Conclusions

Overview

Lending Club specializes in lending various types of loans to urban customers. LC is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters’.

Two types of risks are associated with the bank’s decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

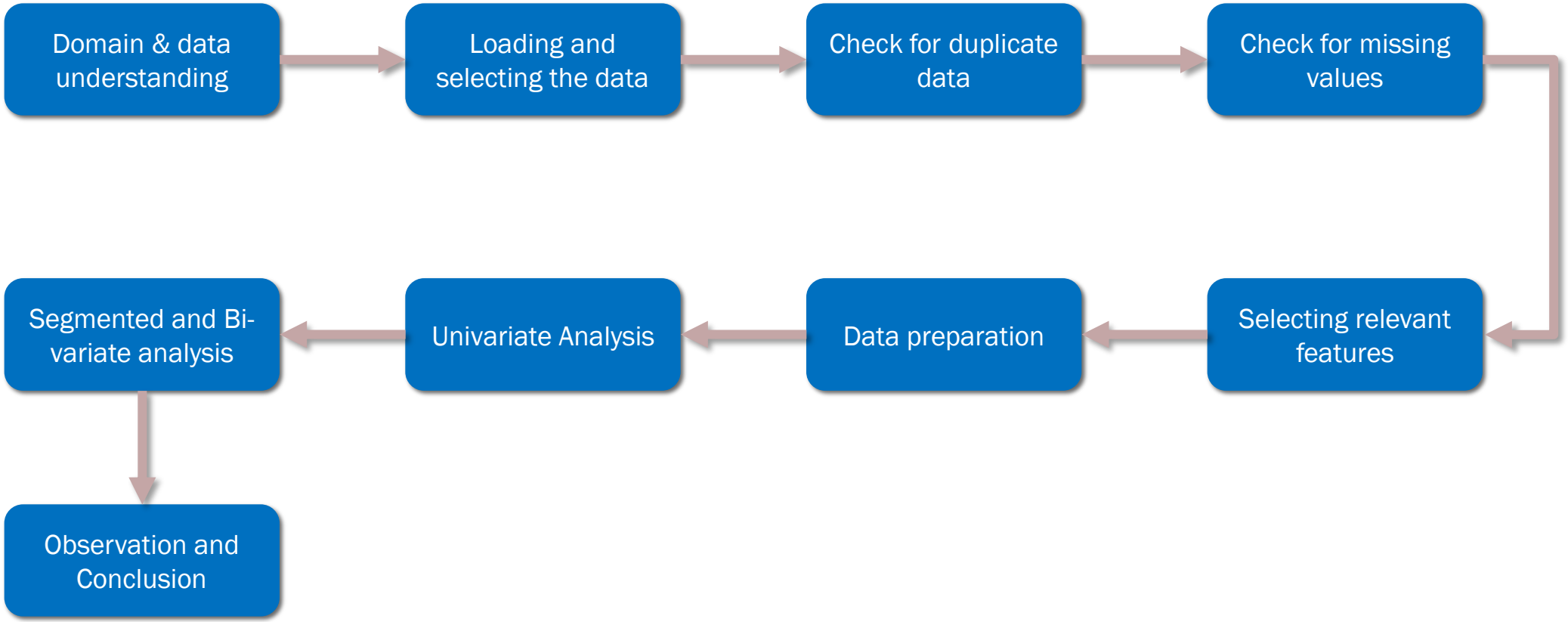
Problem Statement

Identifying the risky loan applicant, based on the data available from the loan application and other account and credit history available from other sources like public records.

- Business Problem:
  - Accurately identify the risky loan applicants so that appropriate decision can be taken, either to reject the loan application, accept the loan application with higher interest rates or customised repayment schedules
  - Accurately identify the non risky loan applications so that the lending opportunity is not lost

- Analytical Problem:
  - From the given data set identify the driving variables that are indicators of loan default, that helps in classifying the loan application as risky or non-risky

Analytical Approach



Analytical Approach Cont..

- Domain & Data understanding
  - Web research made to understand terminologies and parameters considered in risk analytics and data dictionary used to understand the data
- Loading and selecting the relevant data
  - Data is loaded using pandas
  - Relevant data is selected – ignored the records containing loan_status – 'Current'
- Duplicate data
  - Check for duplicate records using unique identifiers like 'id' and 'member_id'
- Missing value analysis
  - Identify rows and columns with missing values
  - Check for extent of missing values, if missing values are more than 20%, ignore the column or row
  - For continuous data based on distribution and presence of outliers either mean or median values are used for imputation
  - For categorical data mode value is used for imputation
  - For datatime data relevant columns are used to impute with making some assumptions
- Feature selection
  - Some of the features in the data are result of approving the loan, such as 'funded_amnt', 'url', 'total_pymnt', ect. Which are not useful in classifying a loan application as risk or not. Hence such columns are dropped
  - Some features have either high cardinality or single values, e.g, 'application_type' which has only one class – "INDIVIDUAL" or 'emp_title' which has 28820 unique values, such columns are dropped from analysis
- Data preparation
  - Remove unwanted characters in the data like – "%" symbols
  - Check for appropriate data type, caste the right data type, like 'issue_d' to datetime, 'open_acc' to category

Analytical Approach Cont..

- Univariate analysis
  - For continuous data;
    - Outlier are identified using box plot and used custom function to identify records with outliers and count number of outliers in a column
    - Summary statistics are computed and Plotted the distribution(hist plots) to understand how the data is spread
    - observations are made on outlier treatment
  - For categorical data;
    - Class-frequencies are obtained, frequency plots are made to understand the distribution of the classes
    - Observations are made to combine the minority classes
- Segmented and Bi-variate Analysis
  - Inspected the correlation among continuous variable using person correlation coefficient and pairplots
  - `'loan_status'` vs continuous variables;
    - used groupby method to check for mean and median values for each segment of the 'loan_status'
    - Used boxplots for each segment of loan_status on numerical variable to check for outliers, Q1, Median, Q3 values
    - If outliers presents, outliers are removed for the further analysis
    - Used KDE plots to understand the distribution of continuous data for each segment of loan_status
  - `'loan_status'` vs categorical variables;
    - Cross-tabulated the data to obtain the proportion of split between segments of the loan_status across various classes of a categorical variable
    - Pivoted the data for a given categorical variable and loan_status using numerical values to compute various metrics like mean, median and percentile for each segment of loan_status, same is visualized using line plots
- Observations and conclusions
  - Based on the observations made in segmented univariate and bi-variate analysis, driving variables are identified which are indicators of loan default

Observations and Findings – Data Cleaning

- Data:
  - Data has 39717 rows and 111 columns
  - 'loan_status' is the target column, has 3 categories;
    - Fully Paid - 32950
    - Charged Off - 5627
    - Current - 1140
  - 'Current' category is dropped from analysis as this represents the running loans

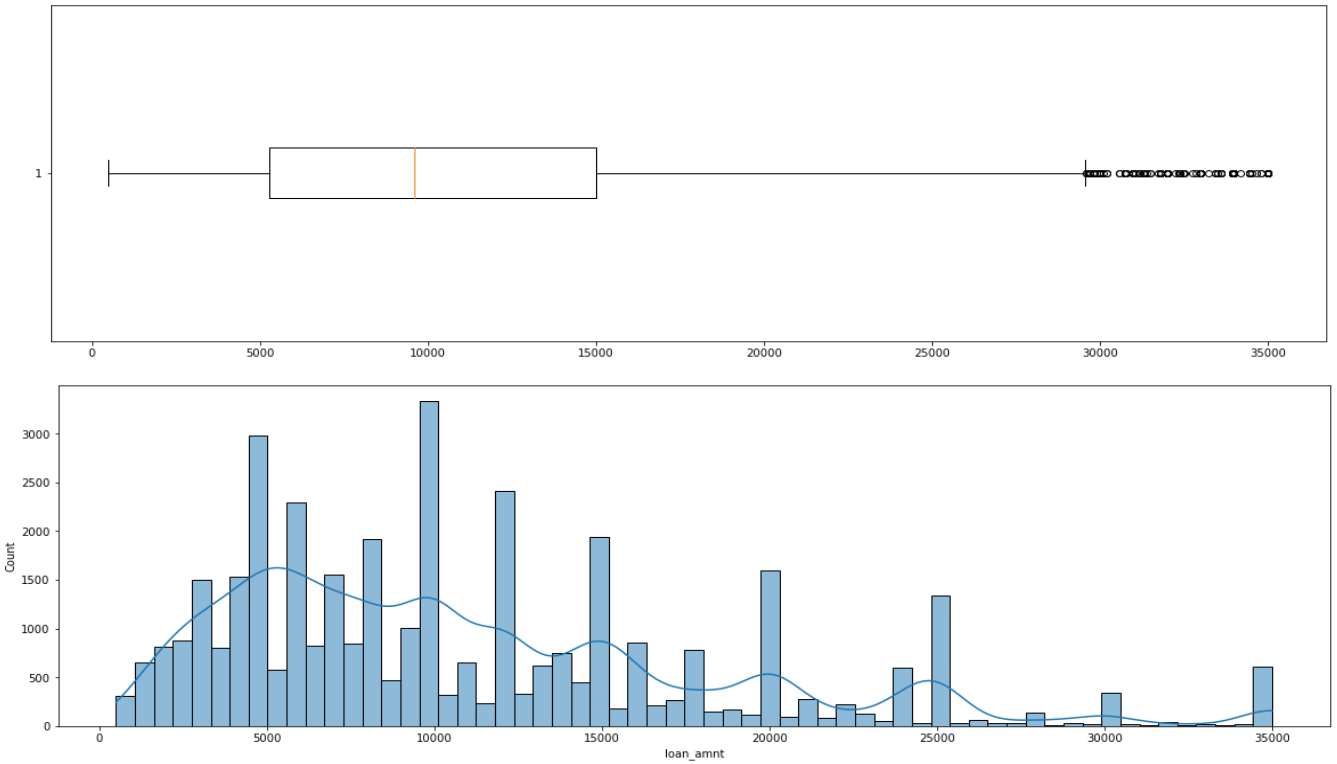
- Missing values:
  - There are 58 columns with more than 20% of missing values, these columns are dropped from the analysis
  - There are 10 columns with less than 20% missing values, categorical columns are imputed with mode values, continuous columns are imputed with median as there are outliers in the column

- Feature Selection:
  - Columns such as *['emp_title', 'title', 'collections_12_mths_ex_med', 'chargeoff_within_12_mths', 'tax_liens',]* have either high cardinality or single value are dropped from analysis
  - Columns such as *['funded_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'pymnt_plan', 'url', 'last_pymnt_d', 'initial_list_status', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', collection_recovery_fee', 'last_pymnt_amnt', 'policy_code']* are considered irrelevant for the analysis hance dropped.

- Data considered for analysis:
  - Rows – 38577 and Columns – 23
  - *'loan_amnt', 'grade', 'sub_grade', 'emp_length', 'home_ownership', 'annual_inc', 'verification_status', 'issue_d', 'loan_status', 'purpose', 'zip_code', 'addr_state', 'dti', 'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'last_credit_pull_d', 'pub_rec_bankruptcies'*

Observations and Findings – Univariate Analysis

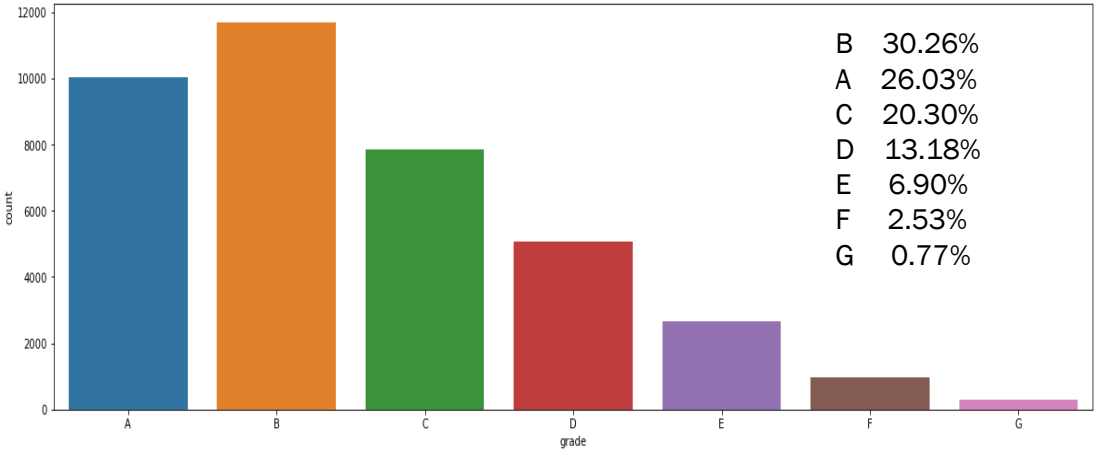
Variable: ‘loan_amnt’



Observations

- loan_amt - This is the amount of the loan applied by the borrower, can be considered for the analysis
- There are 1088 outliers, these outliers are actual loan amount applied for, hence will be analyzed against the loan status
- We can also bucket them into bins for analysing with loan status
- 50% of loan applicants have applied for loan amount less than or equal to 9600
- 95% of loan applicants have applied for loan amount less than or equal to 25000

Variable: ‘grade’

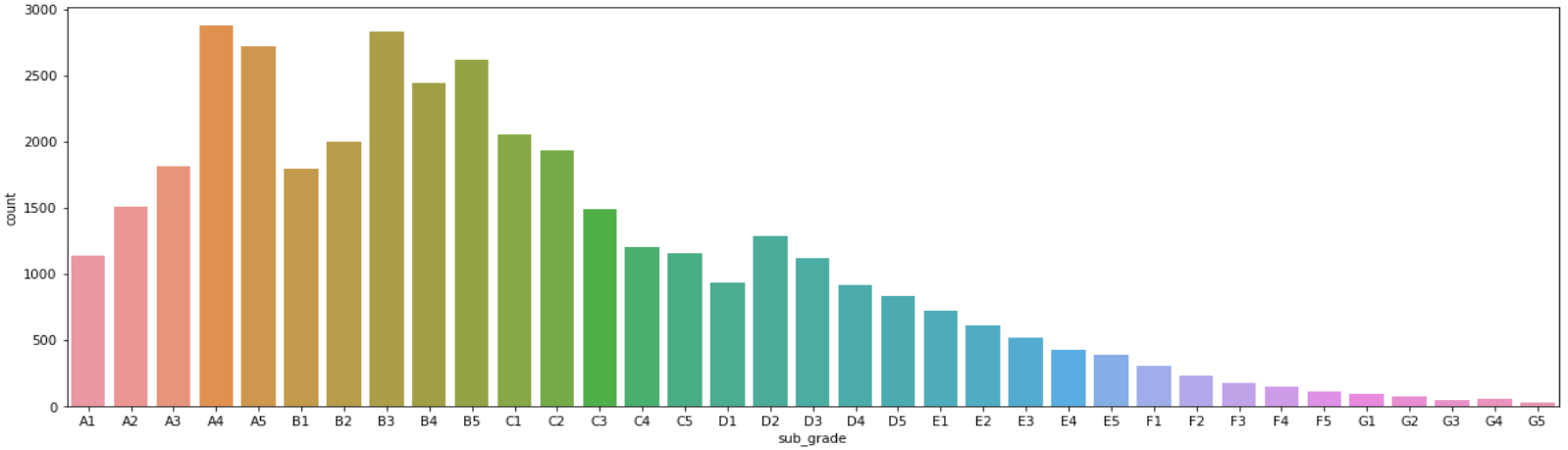


Observations

- 'grade' - is LC assigned loan grade
- This column has 7 categories
- 'grade' categories 'A', 'B', 'C' and 'D' constitute the majority, categories 'E', 'F' and 'G' combined represents < 11% data
- We can consider combining 'E', 'F', and 'G' categories into one category

Observations and Findings – Univariate Analysis

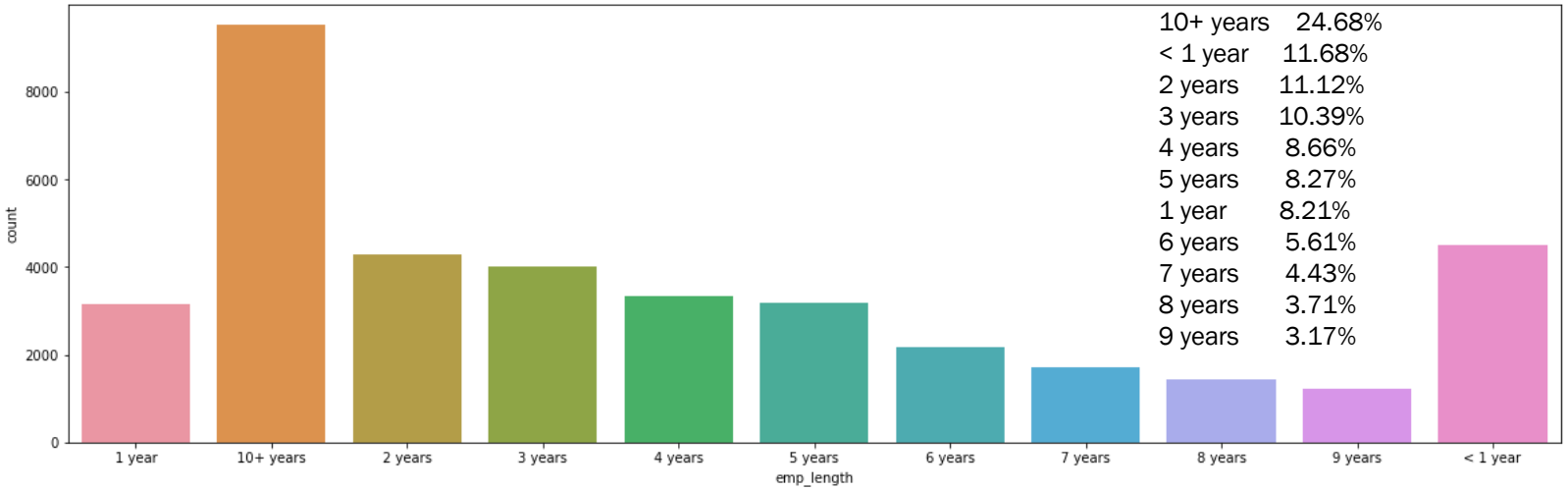
Variable - 'sub_grade'



Observations

- 'emp_length' - is Employment length in years, at the time of application
- This column has 11 categories
- Most of the loan applicant about 25% are with 10+ years of employment history
- Applicant with 7, 8, 9 years of employment length have less than 5% of data in each category

Variable - 'emp_length'

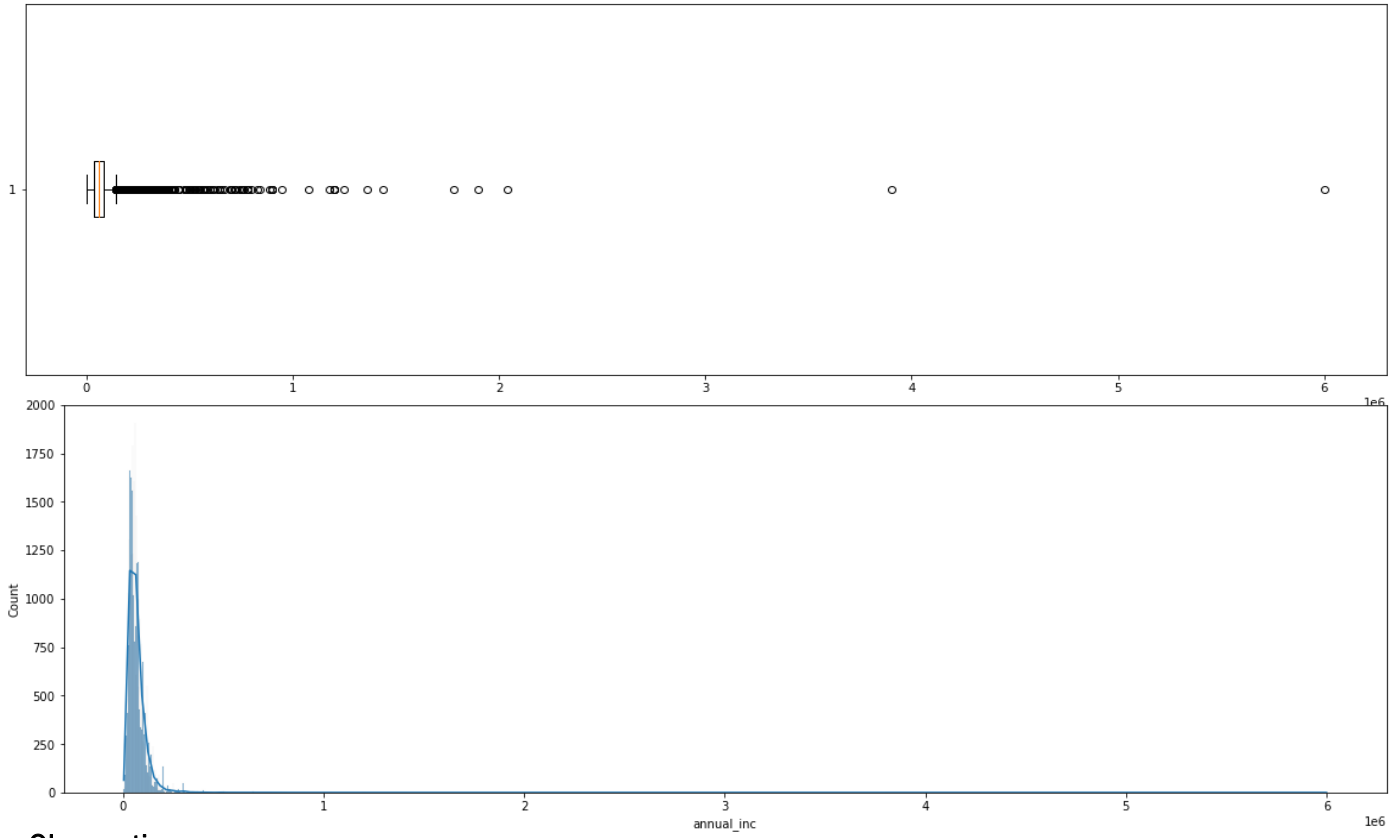


Observations

- 'emp_length' - is Employment length in years, at the time of application
- This column has 11 categories
- Most of the loan applicant about 25% are with 10+ years of employment history
- Applicant with 7, 8, 9 years of employment length have less than 5% of data in each category
- We can further regroup this column into fewer categories to get more insights wrt. loan status

Observations and Findings – Univariate Analysis

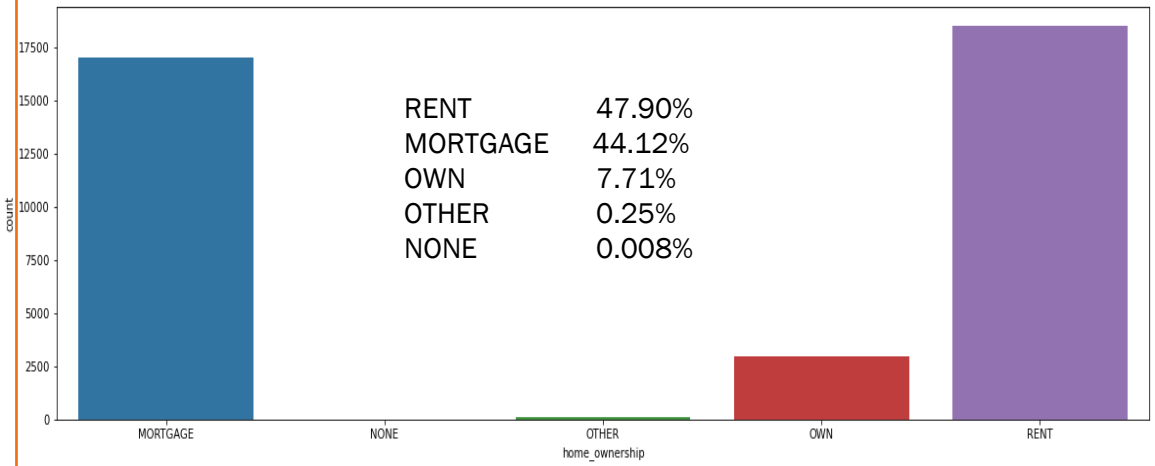
Variable: 'annual_inc'



Observations

- There are extreme outliers and in total 1762 outliers
- However, these outliers are actual income disclosed by the applicants and will be considered for analysis
- We can bucket them into bins for analysing with loan status
- 50% of loan applicants have annual income less than or equal to 58868, where as mean annual income is 68777 which is due to presence of outliers
- From the distribution plot, data looks extremely right skewed.

Variable - '**home_ownership**'

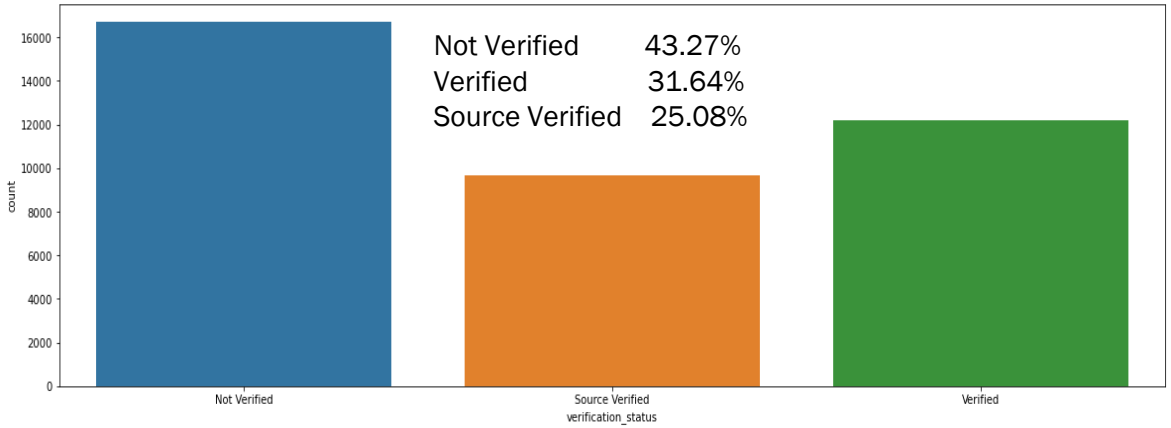


Observations

- 'home_ownership' - is The home ownership status provided by the borrower during registration. LC values are: RENT, OWN, MORTGAGE, OTHER.
- This column has 5 categories - RENT, OWN, MORTGAGE, OTHER and NONE
- None category can be considered as missing values and it is also the minor category, We can merge NONE with RENT, as RENT is the majority class (~48%)
- customers living in RENTED House and customers who have mortgaged their homes are the majority loan applicants
- 'OWN' and 'OTHER' class represent minority class
- NONE class will be added to RENT class

Observations and Findings – Univariate Analysis

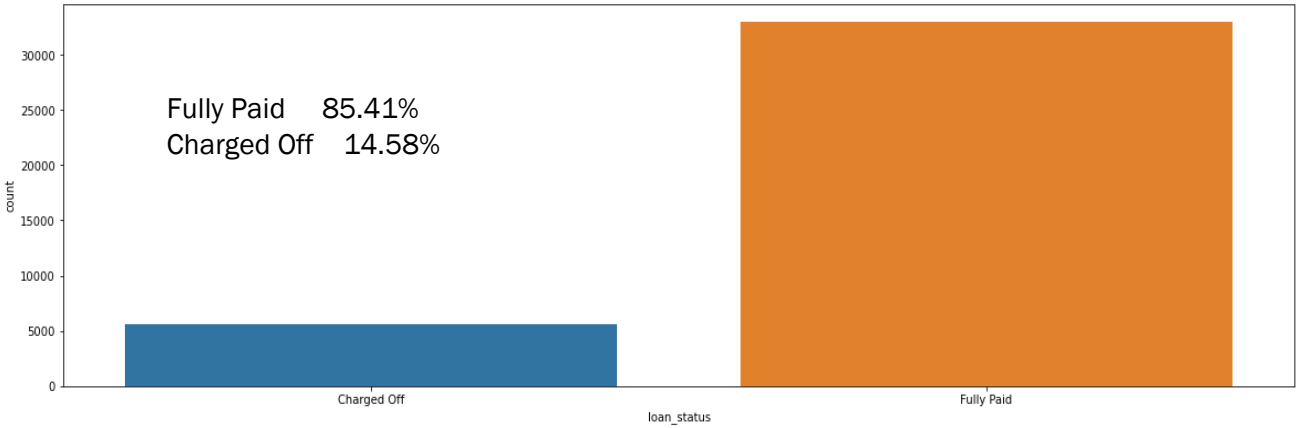
Variable: ‘verification_status’



Observations

- verification_status - Indicates if income was verified by LC, not verified, or if the income source was verified
- There are 3 categories
- For about 57% of applicants, there has been somre sort of varification done
- For rest 43% of applicants, there is no varification done
- can be regrouped into 2 categories - Verified and Not Verified by combining Verified with Source Verified

Variable: ‘loan_status’

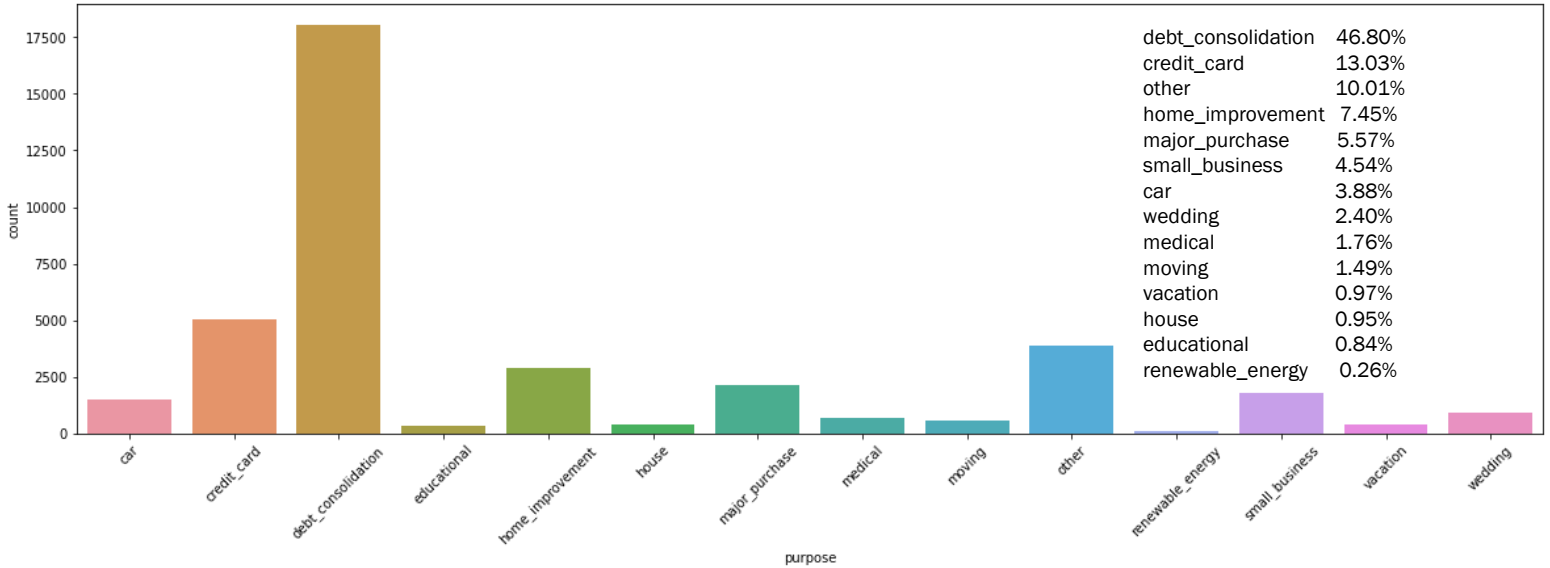


Observations

- 'loan_status' is the target column
- It has 2 categories 'Fully Pain' and 'Charged Off'
- Data is skewed as it has 85% of Fully Paid category compared to 15% of Charged off category

Observations and Findings – Univariate Analysis

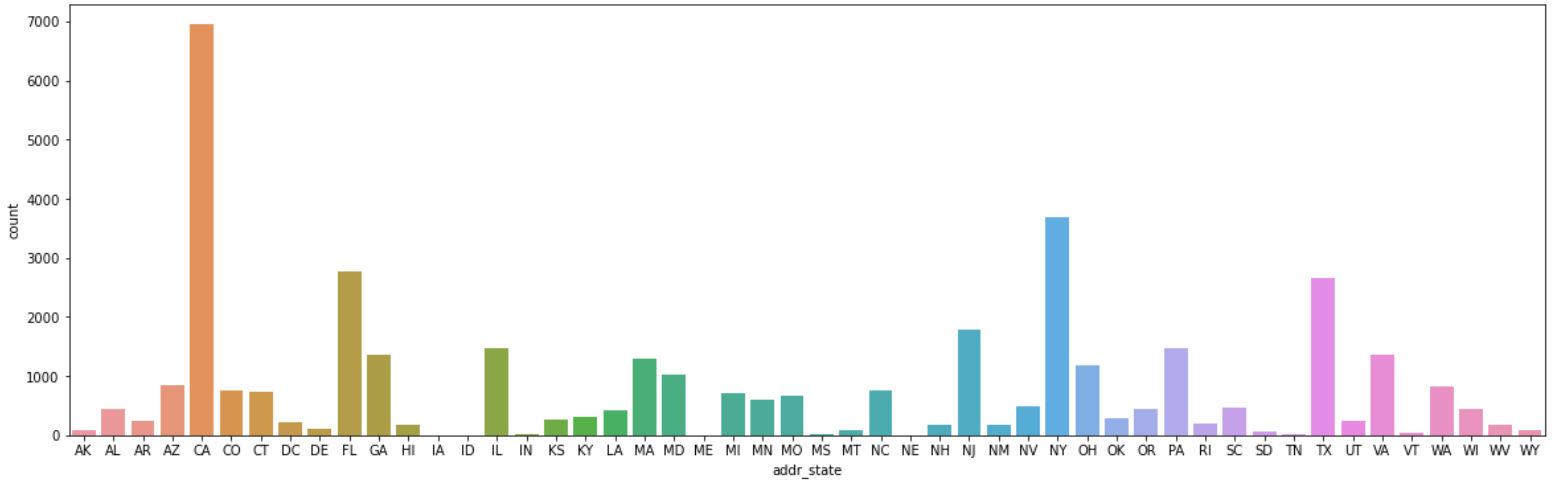
Variable - ‘purpose’



Observations

- 'purpose' - A category provided by the borrower for the loan request.
- There are 14 categories
- 'debt_consolidation' category represents the majority class < 46% of data

Variable - ‘addr_state’

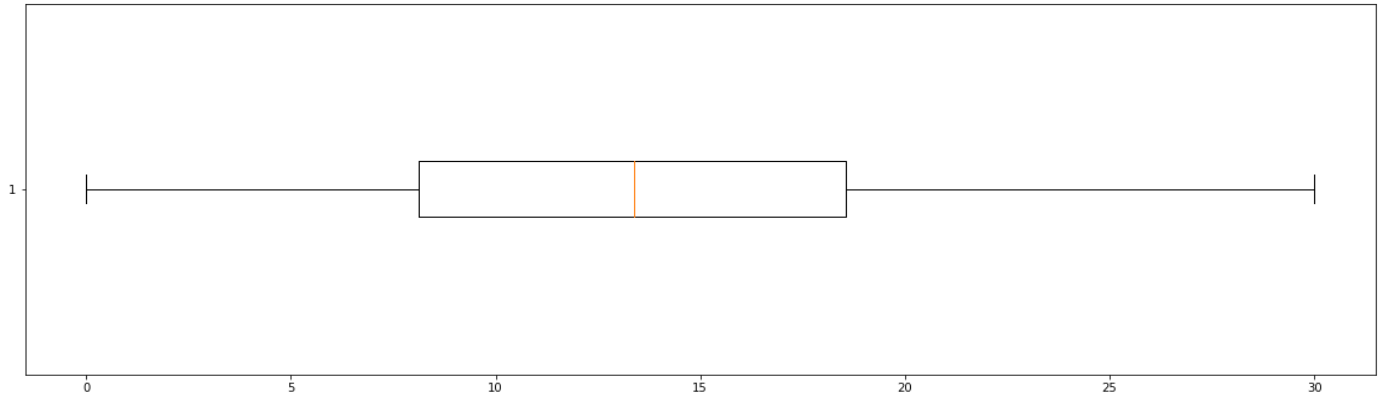


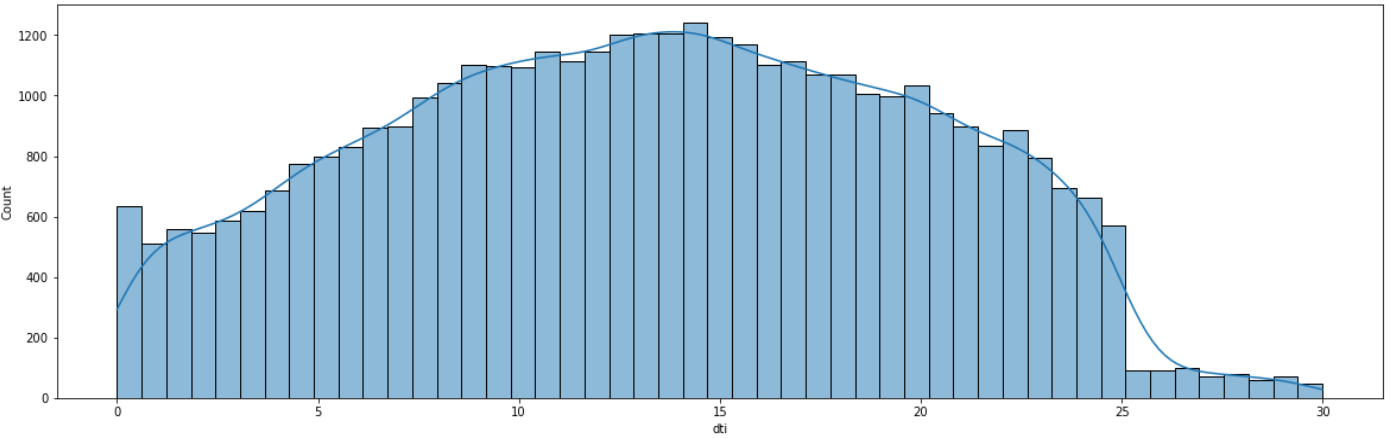
Observations

- 'addr_state' - The state provided by the borrower in the loan application
- The data contains applicants from all 50 states of USA
- The state 'CA' has the highest loan issuance and state 'ME' the lowest
- Minority states can be combined

Observations and Findings – Univariate Analysis

Variable: ‘dti’

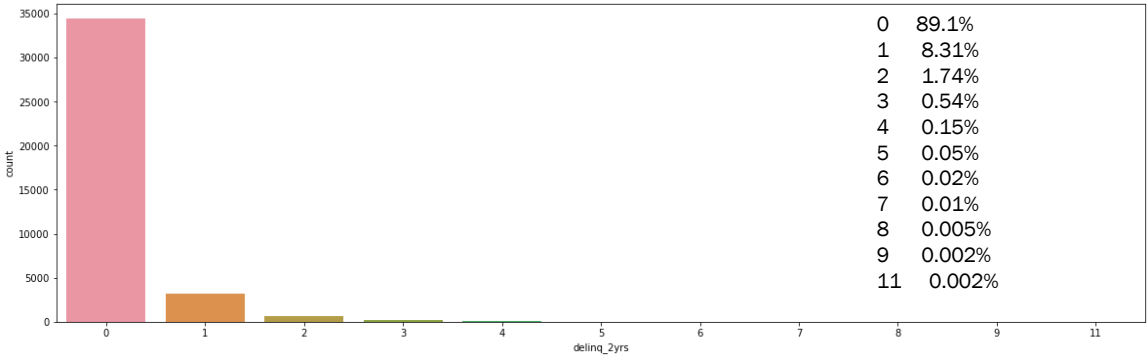




Observations

- 'dti': "debt-to-income" expressed as percentage - A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
- This is a good indicator of borrowers ability to repay the loan, higher the ratio indicates lower the ability to pay the newer debts
- There are no outliers in the 'dti' data
- The mean and median dti are 13.27 and 13.37, which are close to each other
- Distribution of dti looks Normal excpet for right tail,
- Max dti is close to 30%

Variable - ‘delinq_2yrs’



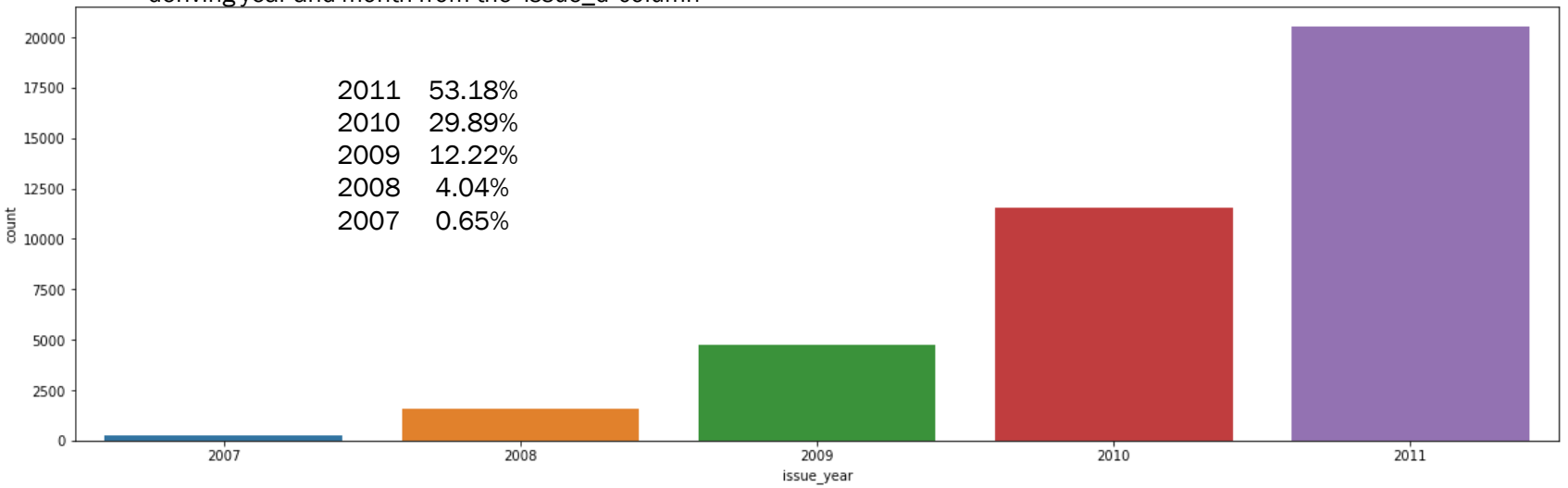
Observations

- 'delinq_2yrs': The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2
- Indicates the troubles the borrower has in making timely repayments
- Close to 90% of borrowers have zero delinquency in past 2 years
- Around 10% of borrowers have at least 1 delinquent payment history
- We can create binary categories - 0 and 1+ and check with loan status

Observations and Findings – Univariate Analysis

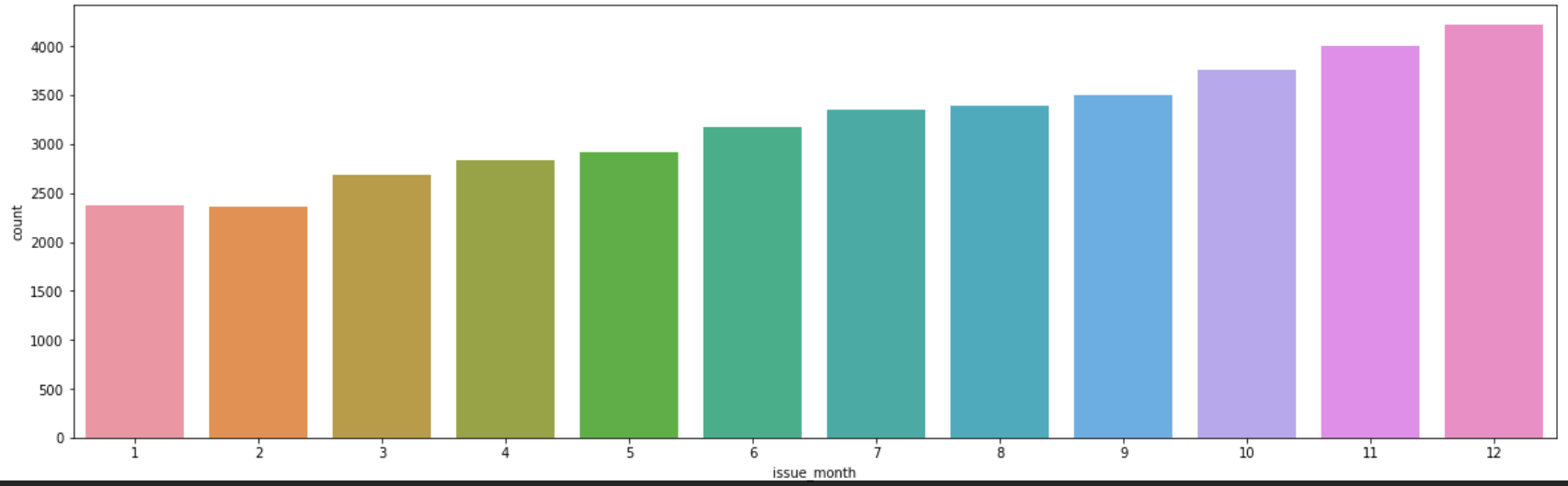
Variable - ‘earliest_cr_line’

- issue_d - data issuance of the loan
- Converted to datatime type created new features – year and month from it
- May not directly help in the analysis, but can be used to derive new feature when used with earliest_cr_line
- deriving year and month from the 'issue_d' column



Observations

- data represents loan issued between 2007 and 2011
- There is an explonential increase in loan issuance year by year
- data contains less than 10% of loan issued in 2007 and 2008
- data contains more than 53% of loan been issued in 2011



Observations

- There is a gradual increase in the loan issuance from January to December
- number of loans issued at the end of the year is more compared to beginning of the year
- November and December are the months having highest loan issuance
- May be attributed to Christmas and New year shopping

Observations and Findings – Univariate Analysis

Variable - ‘earliest_cr_line’

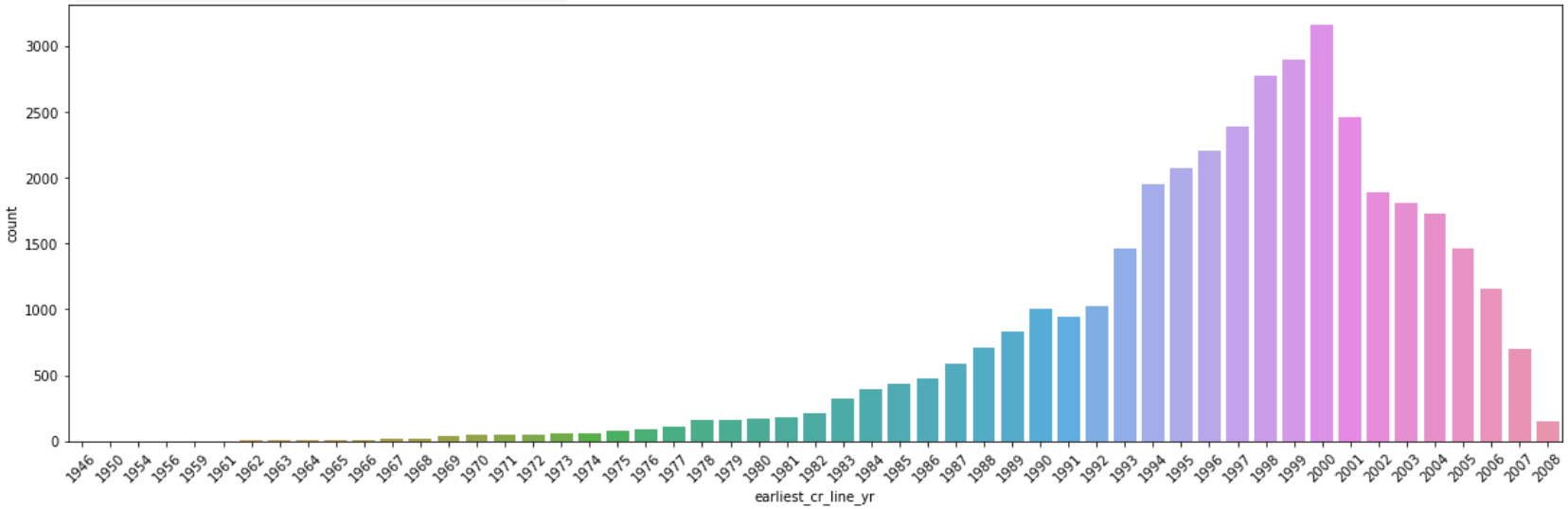
	earliest_cr_line	last_credit_pull_d	loan_status
1576	2062-09-01	2013-09-01	Fully Paid
1764	2068-09-01	2015-09-01	Fully Paid
3274	2067-09-01	2015-05-01	Fully Paid
3349	2065-02-01	2014-11-01	Fully Paid
3403	2067-06-01	2013-04-01	Charged Off
3595	2067-08-01	2014-04-01	Fully Paid
3976	2063-12-01	2014-11-01	Fully Paid
4426	2068-09-01	2016-05-01	Fully Paid
4435	2068-09-01	2015-10-01	Fully Paid
4478	2063-03-01	2016-05-01	Fully Paid
5092	2068-08-01	2013-02-01	Charged Off
5399	2065-11-01	2014-11-01	Fully Paid
5673	2065-05-01	2013-07-01	Fully Paid
6118	2054-10-01	2014-10-01	Fully Paid
6416	2068-06-01	2016-05-01	Fully Paid

	earliest_cr_line	adj_earliest_cr_line
39617	2063-05-01	1963-05-02
39618	1998-10-01	1998-10-01
39619	1993-05-01	1993-05-01
39620	2001-10-01	2001-10-01
39621	1989-11-01	1989-11-01
39622	1982-07-01	1982-07-01
39623	1997-08-01	1997-08-01
39624	2000-11-01	2000-11-01
39625	1997-06-01	1997-06-01
39626	1999-04-01	1999-04-01
39627	1998-07-01	1998-07-01
39628	2000-01-01	2000-01-01
39629	1999-04-01	1999-04-01

Observations

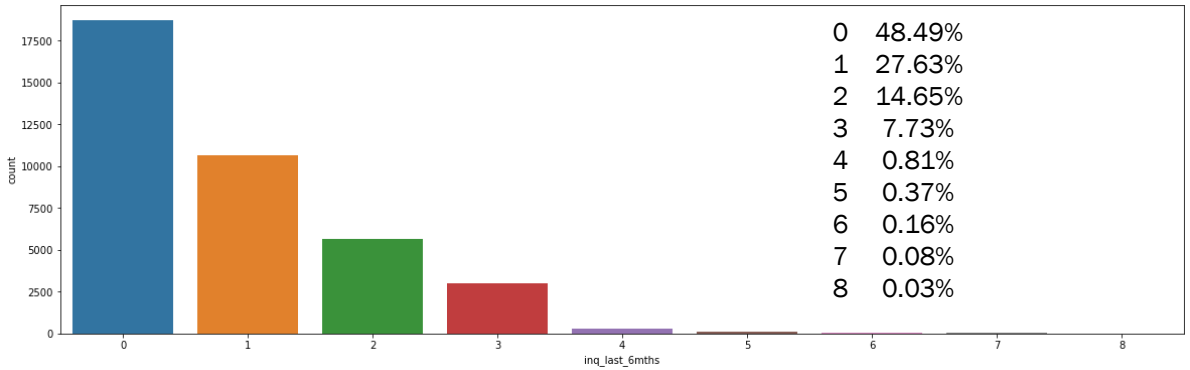
- 'earliest_cr_line_yr' The month the borrower's earliest reported credit line was opened
- It has been observed that data contain future dates eg, 2068, 2054, 2046..etc.
- After 2008, the dates are jumping to 2048 till 2068
- This may be due to system miss interpreting 1948 till 1968 as 2048 and 2068. Because the only two digit of the year is captured in the original data
- This can be verified with ‘last_credit_pull_d’ date

- we can see that after adjusting for the year, most of the borrowers have their earliest credit line availed between 1994 and 2006
- We can use this data to derive new feature



Observations and Findings – Univariate Analysis

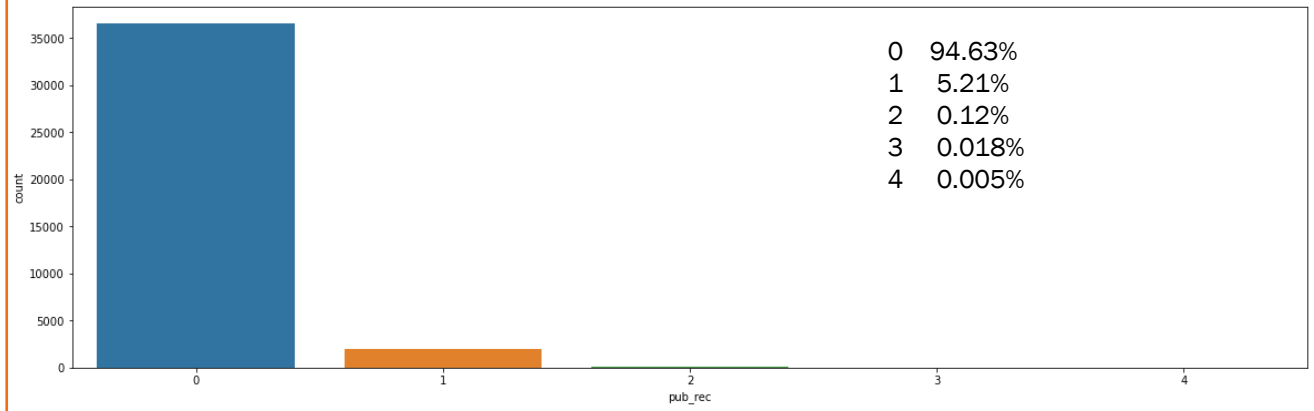
Variable: ‘inq_last_6mths’



Observations

- 'inq_last_6mths' - The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
- There are 9 categories
- ‘0' category represents the majority class < 48% of data

Variable: ‘'pub_rec’

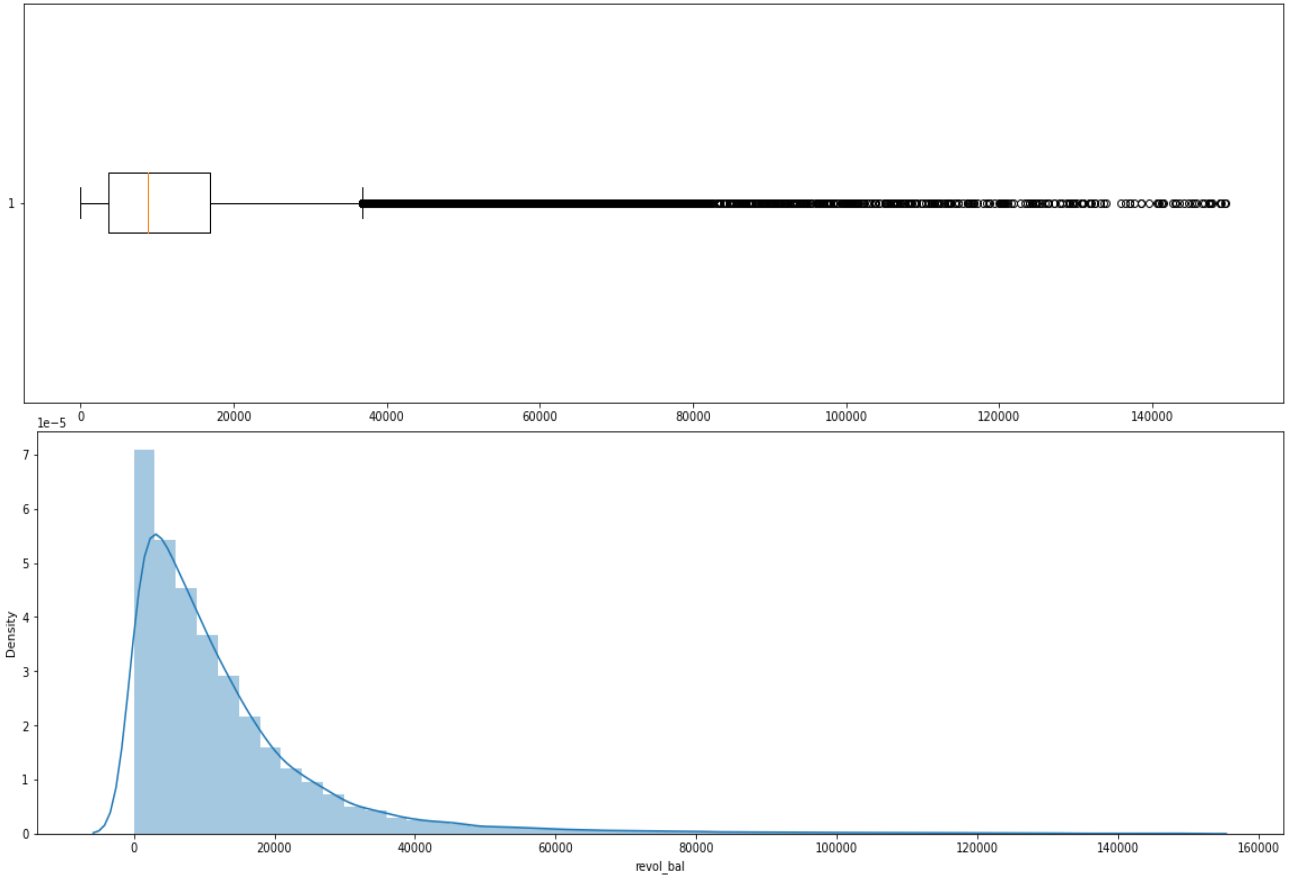


Observations

- 'pub_rec' - Number of derogatory public records.
- it is assumed that more the pub_rec counts, bad on the profile of the borrower
- There are 4 categories, representing number of records in public
- Majority of the data is concentrated in categories - 0
- Data can be regrouped as binary 0 and 1, obtained by combining all non 0 into one category

Observations and Findings – Univariate Analysis

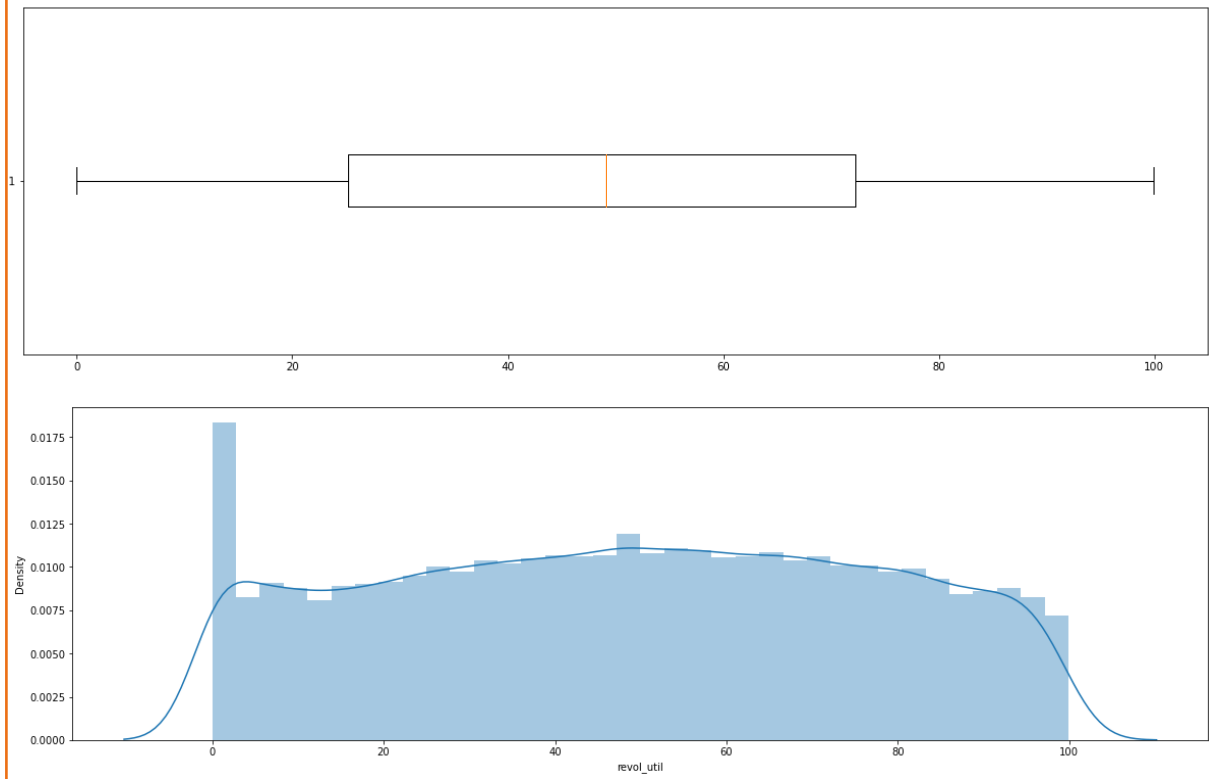
Variable: ''revol_bal'



Observations

- 'revol_bal' - Total credit revolving balance.
- This can be a risk indicator, higher the revolving balance riskier is the applicant
- Data has some large outliers, there are total 2423 outliers
- Data is right skewed
- Median is much smaller compared to Mean reflecting the affect of the outlier

Variable - 'revol_util'

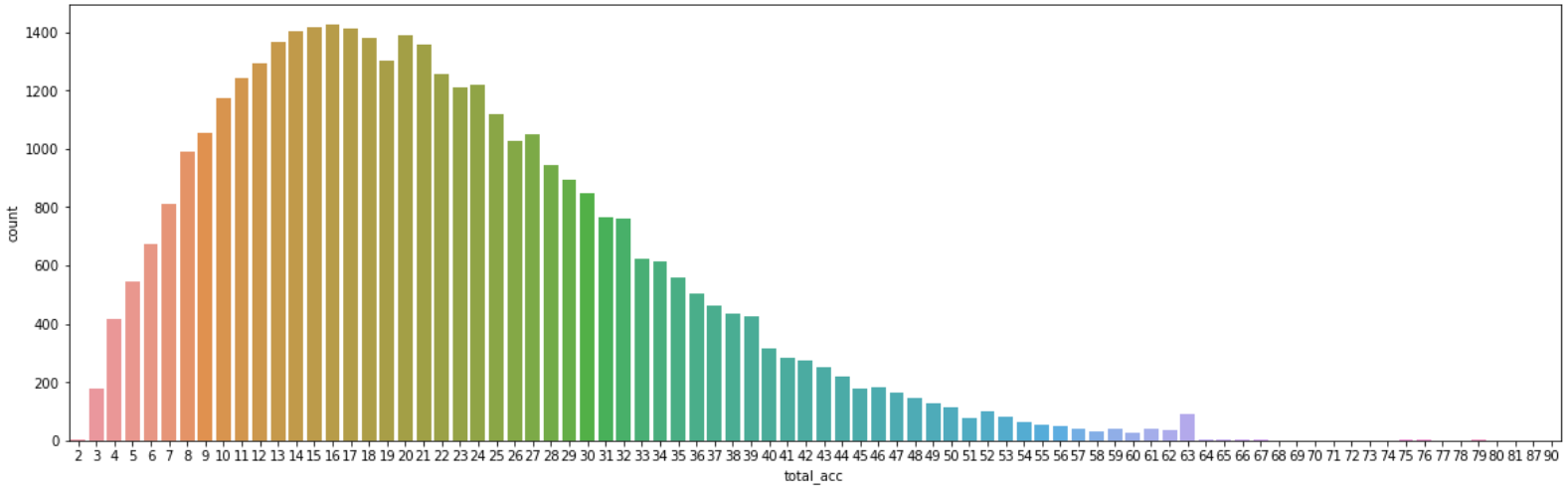


Observations

- 'revol_util' - Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
- This can be a risk indicator, higher the revolving utilization riskier is the applicant
- Data has no outliers
- Data looks uniformly distributed except for the 0th value
- Median and Mean are closer to each other

Observations and Findings – Univariate Analysis

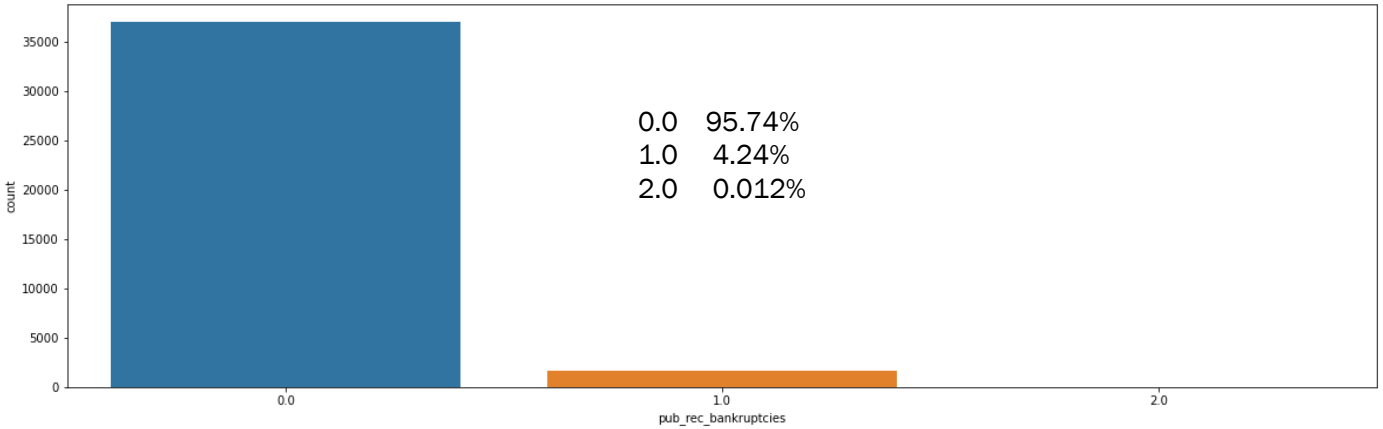
Variable - ‘total_acc’



Observations

- 'total_acc' - The total number of credit lines currently in the borrower's credit file
- it is assumed that more number of credit lines, bad on the profile of the borrower
- Number of credit lines are treated as categorical values
- The mode is 16, majority of data is concentrated between 5 and 39
- There are some extreme values like - 87, 90

Variable - ‘pub_rec_bankruptcies’

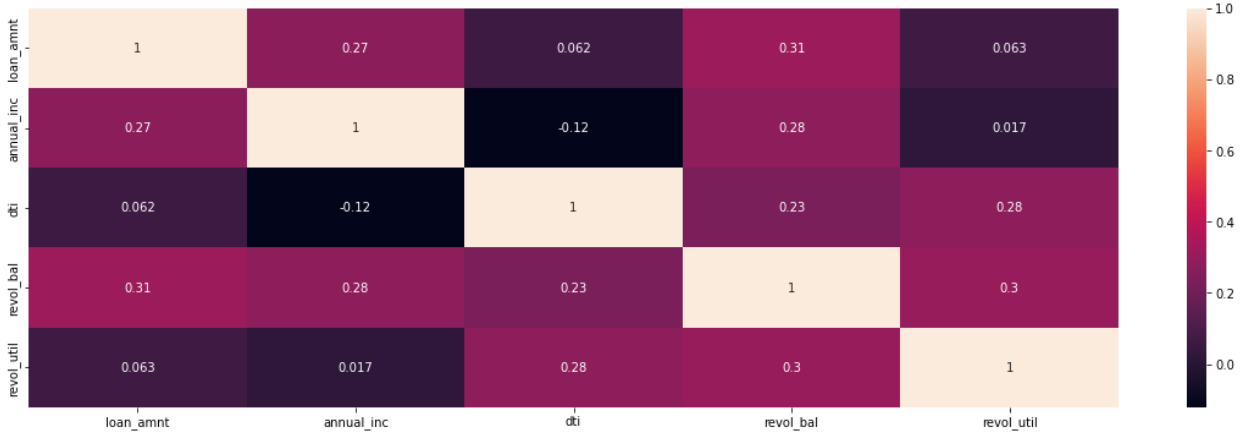


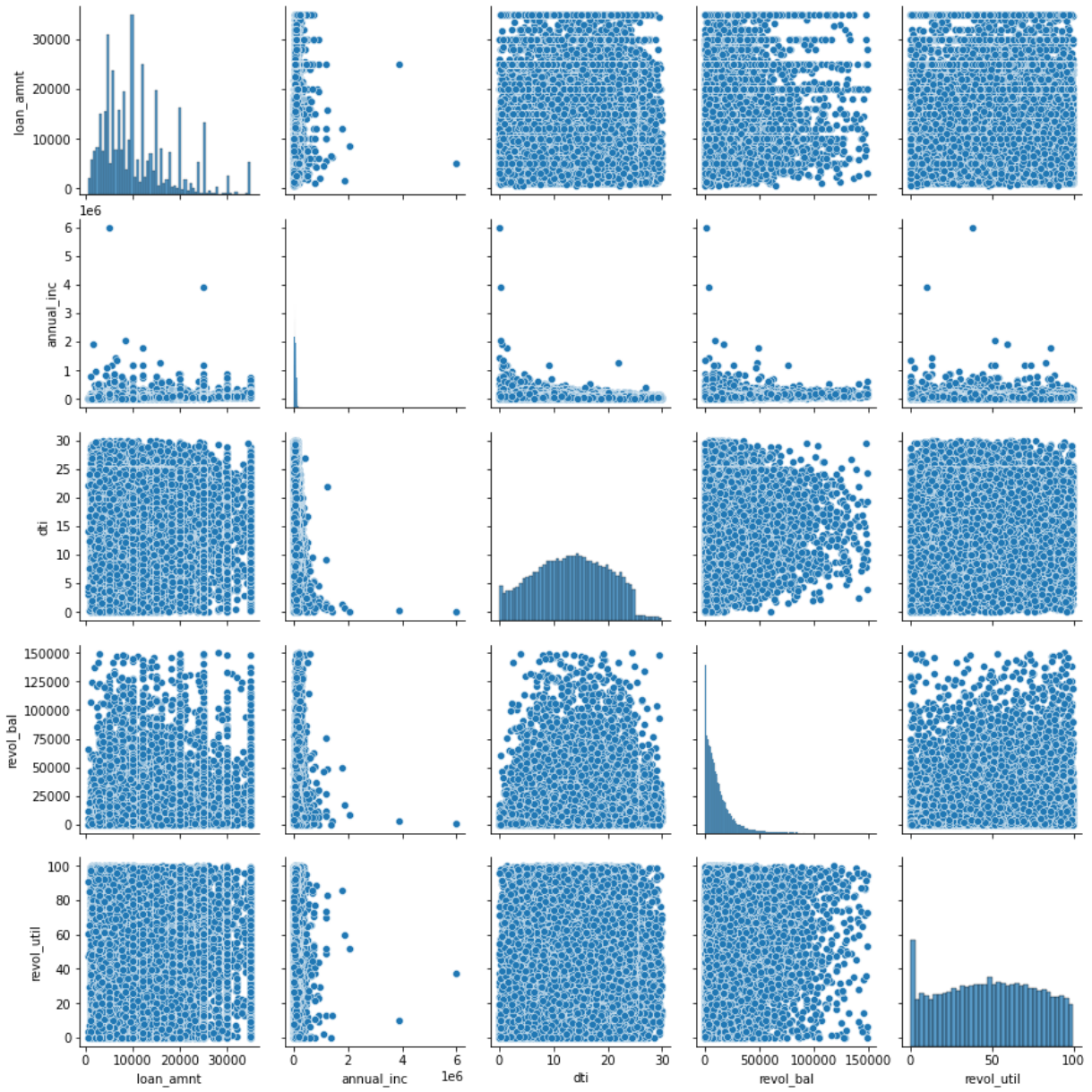
Observations

- 'pub_rec_bankruptcies' - Number of public record bankruptcies
- it is assumed that more number of bankruptcies, is bad on the profile of the borrower
- There are 3 categories 0, 1, 2 representing number of bankruptcies
- more than 95% of borrowers has 0 bankruptcies
- can be converted into binary class - 0 and more than 0

Observations and Findings – Correlation

Inspecting the correlation among independent variables





Observations

- There is no indication of strong correlation among independent variables

Observations and Findings – Segmented Analysis

loan_status	loan_amnt	annual_inc	dti	revol_bal	revol_util
Charged Off	10000.0	53000.0	14.29	9211.0	58.2
Fully Paid	9600.0	60000.0	13.20	8682.5	47.6

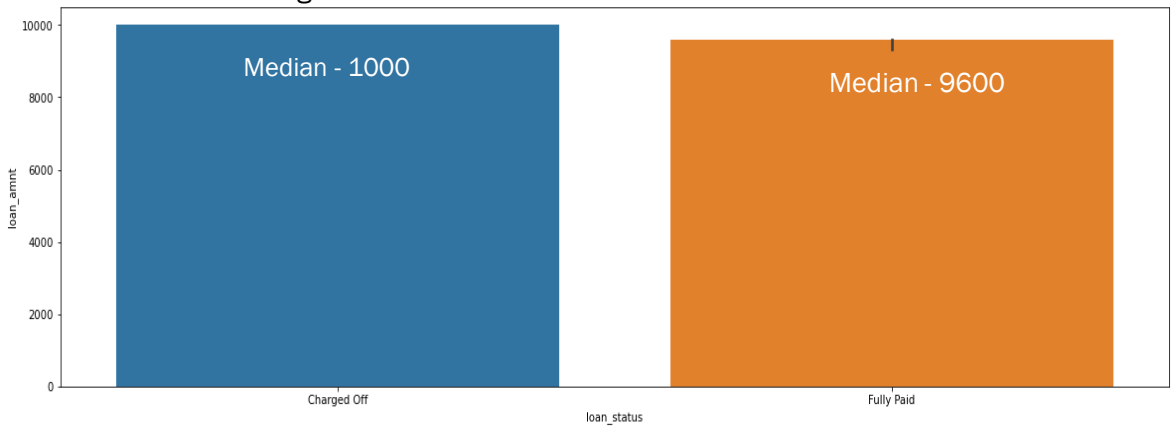
Observations

- Median is used as the metrics as some variables has outliers
- Median of – 'loan-amnt', 'dti', 'revol_bal' and 'revol_util' are higher for Charge Off compared to Fully Paid
- Median of 'annual_inc' is higher for fully paid compared to Charge off
- These numerical variable can distinguish between these two classes
- Will analyse each numerical variables vs Loan_amnt individually

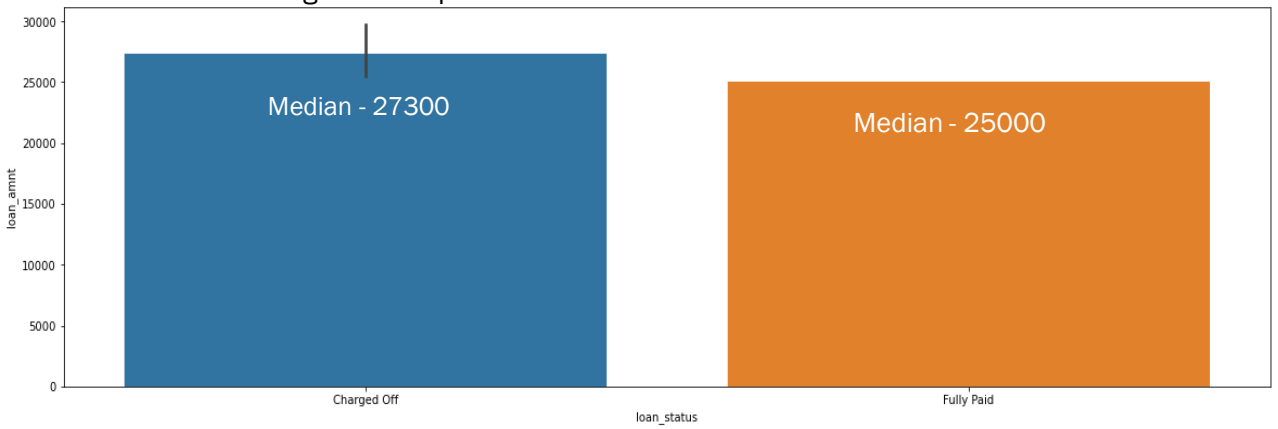
Observations and Findings – Bivariate Analysis

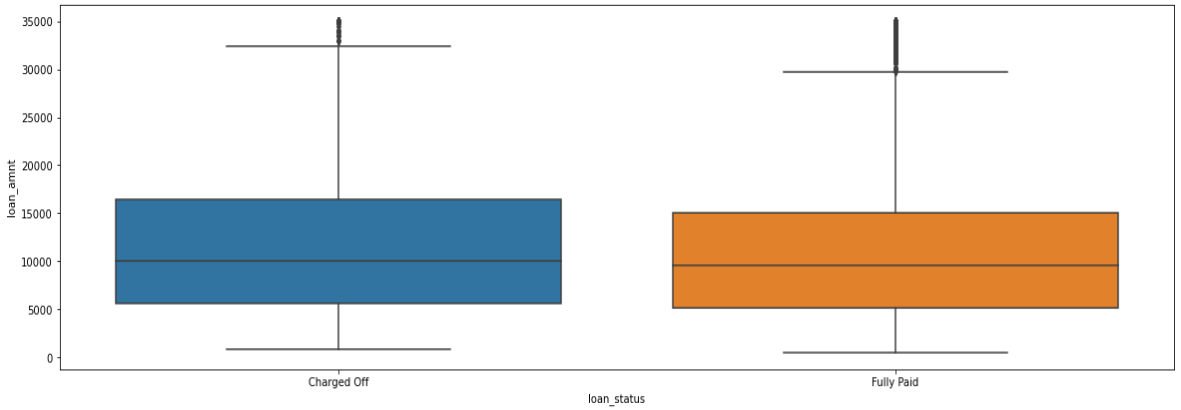
Loan_status vs Loan_amt

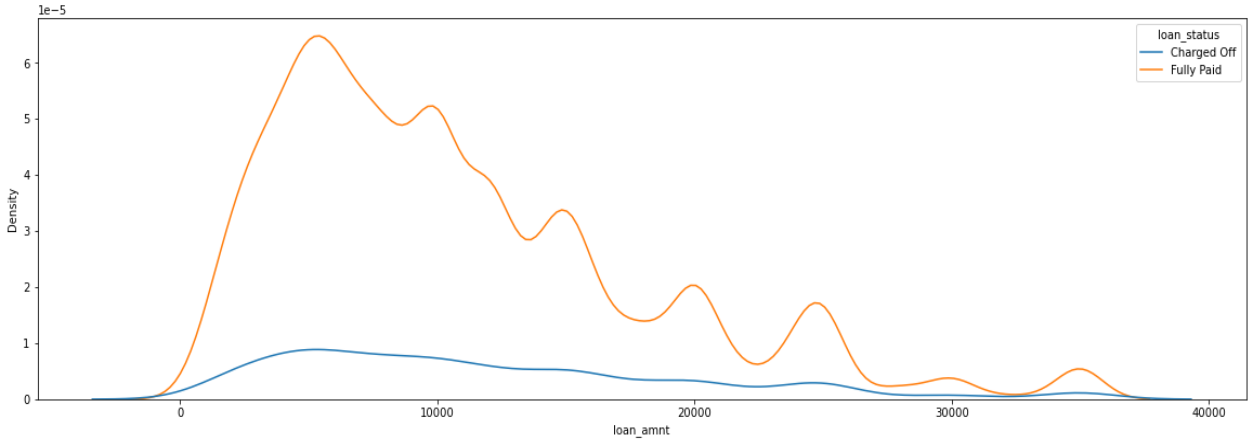
Plotting the median values for loan amount different loan status



Plotting the 95th percentile values for loan amount different loan status





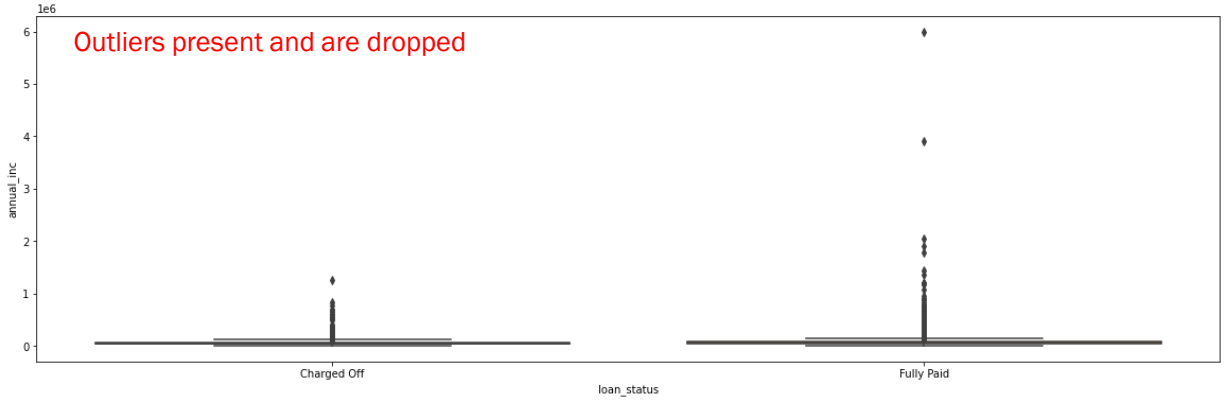


Observations

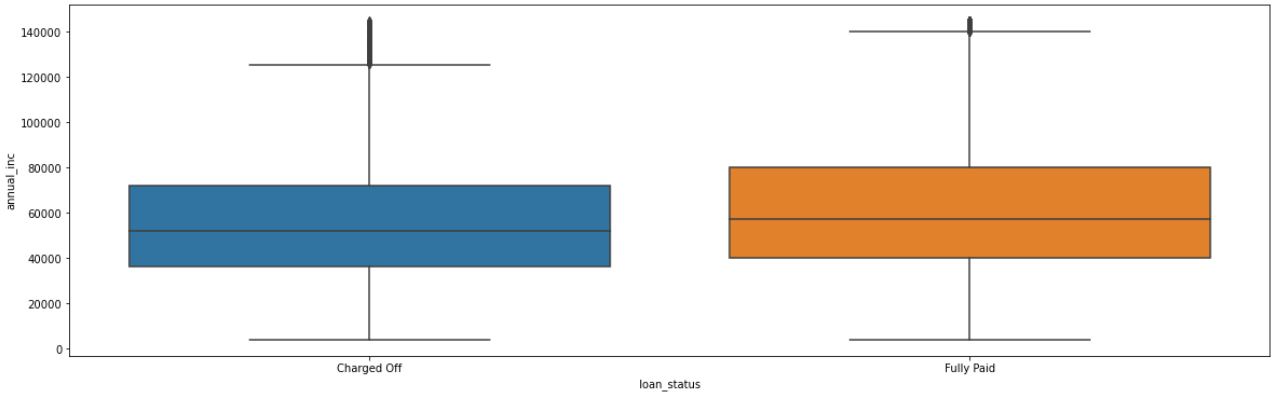
- We observe that median, 75% percentile, 95% percentile and upper whisker are higher for charged off loans
- This indicates high value loans have higher risk of being charged off, but Fully paid loans also has some outliers
- As we can see from the KDE plot both Fully paid and Charged off loans spread over the range, Fully paid loans are multi modal and peaking around 5000 mark, even the Charged off segment
- Hence, only considering loan amount independent to other borrowers attributes like bti, annual income etc is not a good measure

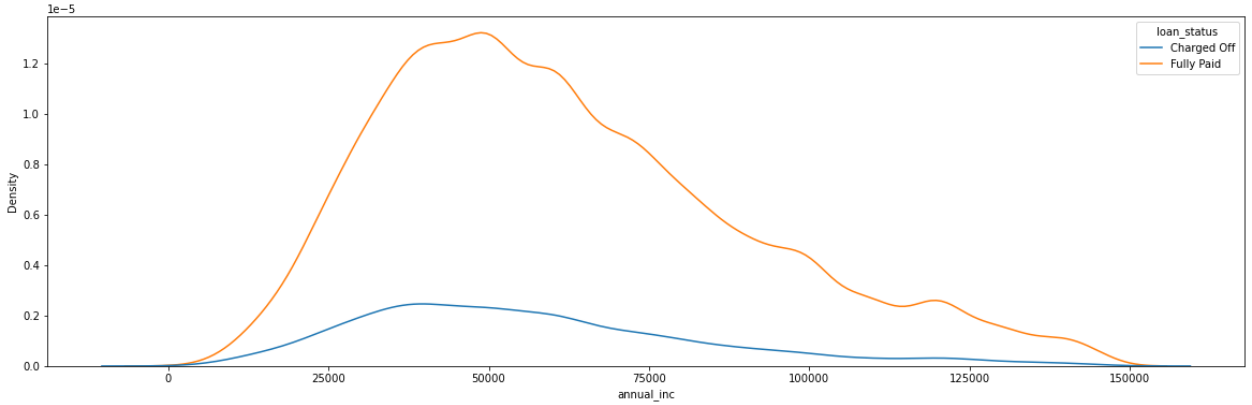
Observations and Findings – Bivariate Analysis

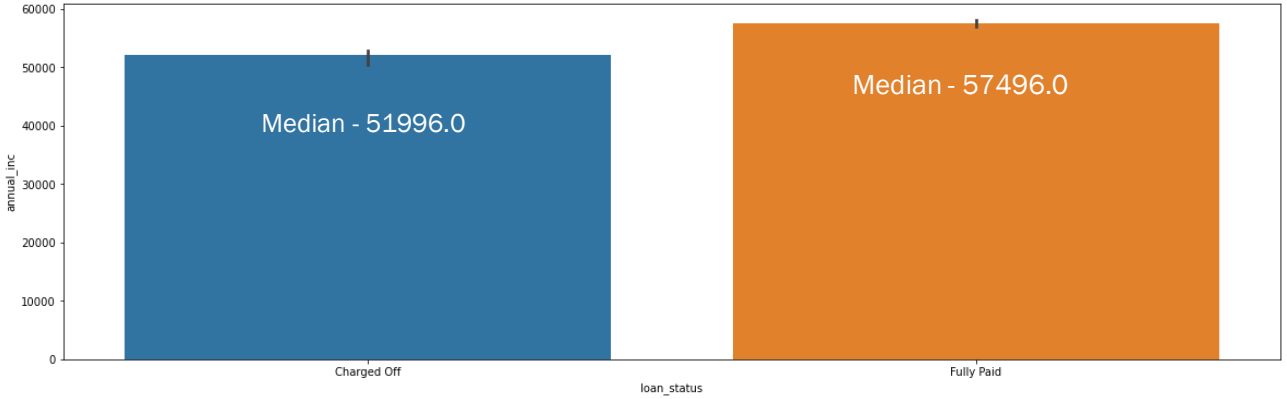
Loan_status vs annual_inc



After dropping the outliers





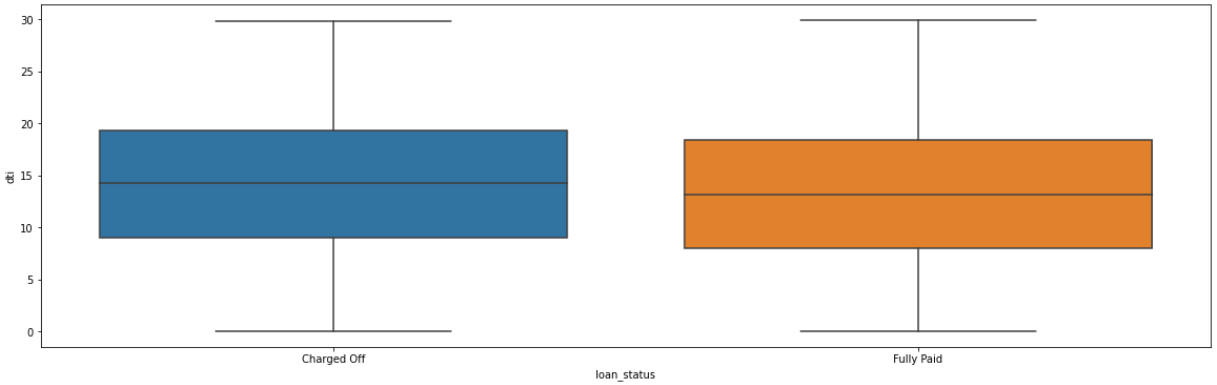


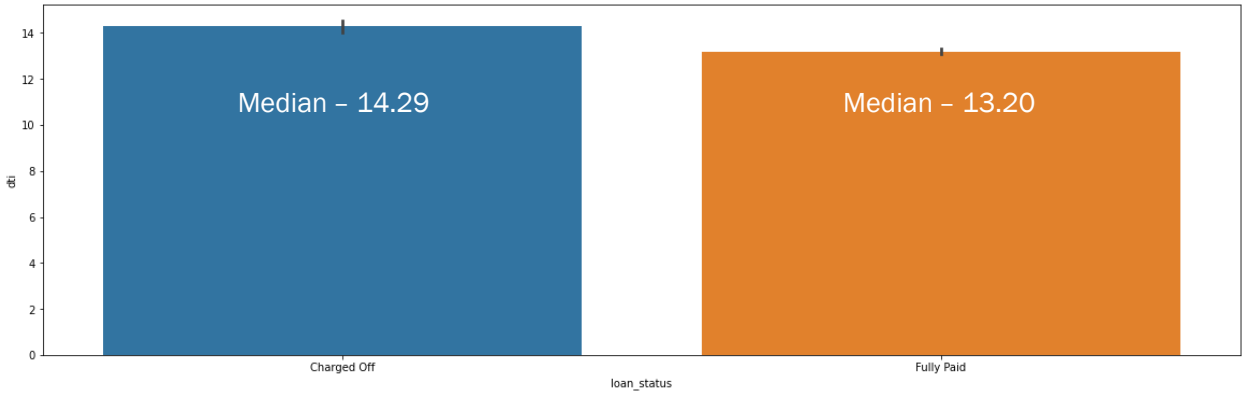
Observations

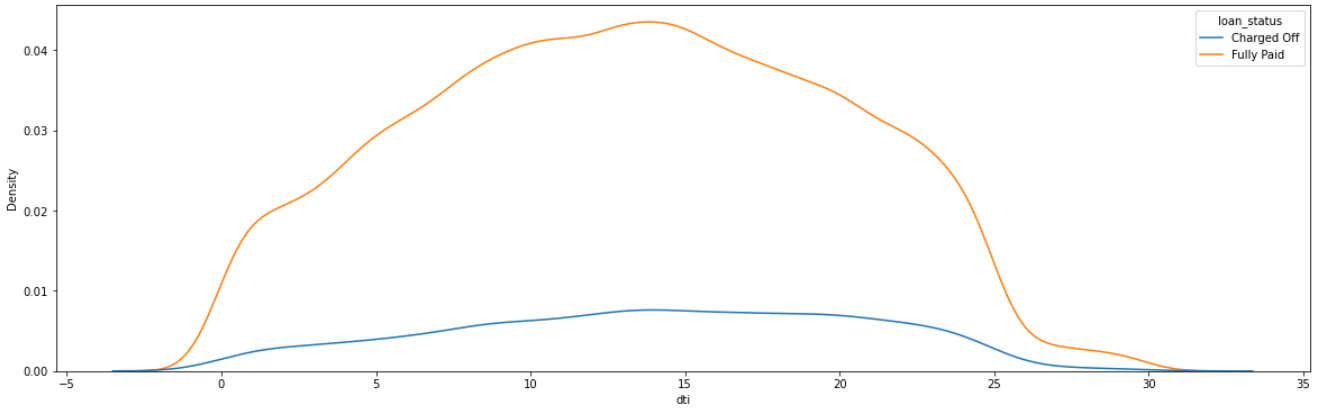
- after removing the outliers the median annual income (57496) of the borrowers who have fully paid the loan is significantly higher than those who defaulted the loan (51996)
- Same can also be observed in the box plot where the median, Q3 and top whisker is higher for fully paid borrowers
- ‘annual_inc’, when used along with other variables can be a good indicator of risk, we will use this variable with other categorical variables to ensure the usefulness

Observations and Findings – Bivariate Analysis

Loan_status vs dti





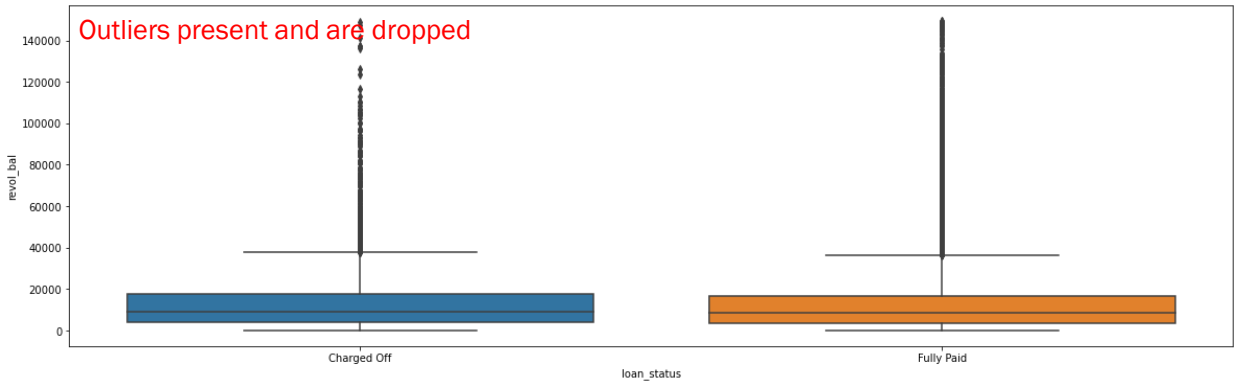


Observations

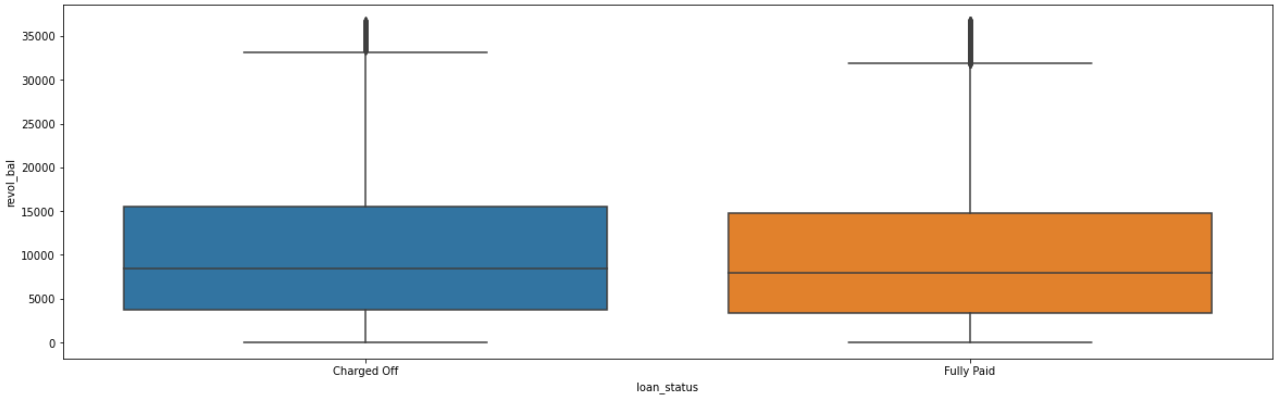
- Median for dti for Charge off segment is 14.29, which is higher compared to Fully Paid segment, which is 13.20
- From boxplot Median, Q3 are higher for Charged off segment
- It indicates higher 'dti' has higher risk of loan been Charged off
- ‘dit’ to be analysed along with other categorical variables

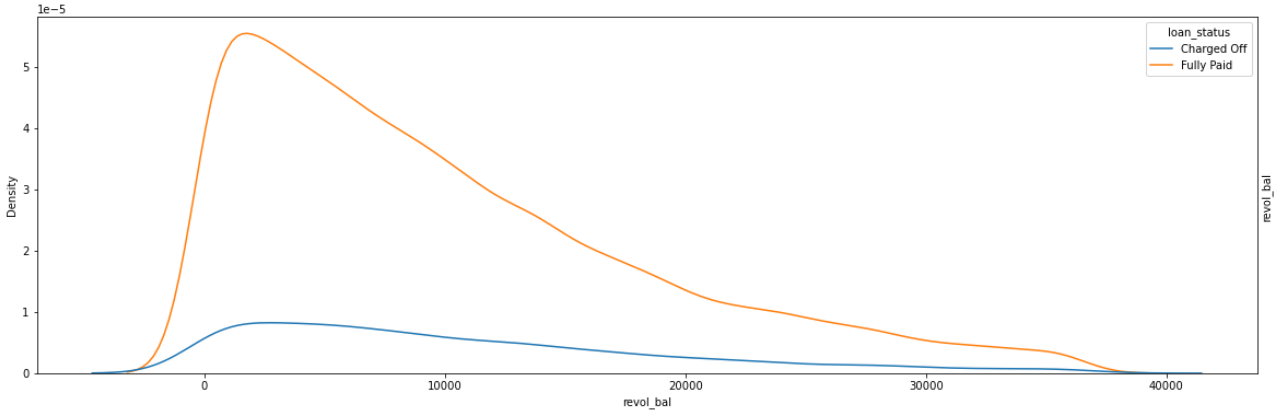
Observations and Findings – Bivariate Analysis

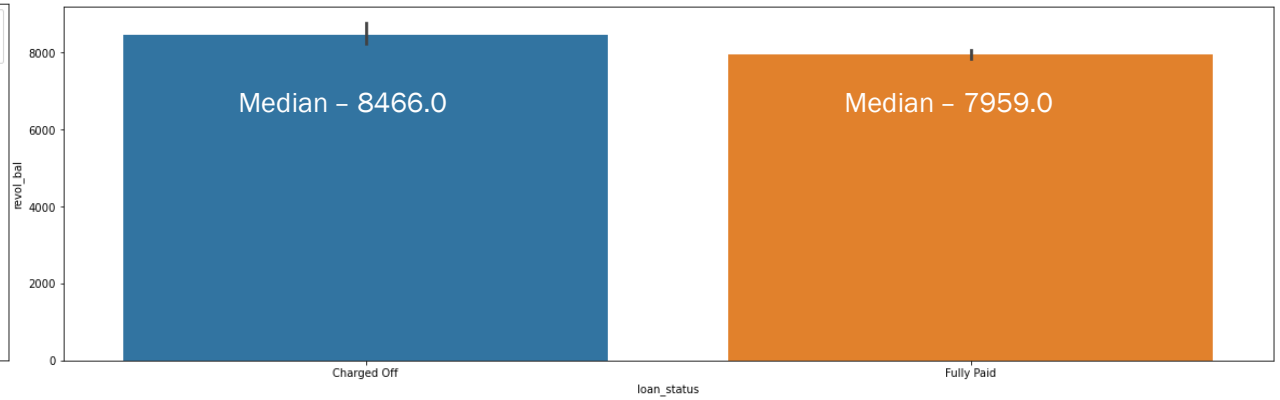
Loan_status vs revol_bal



After dropping the outliers





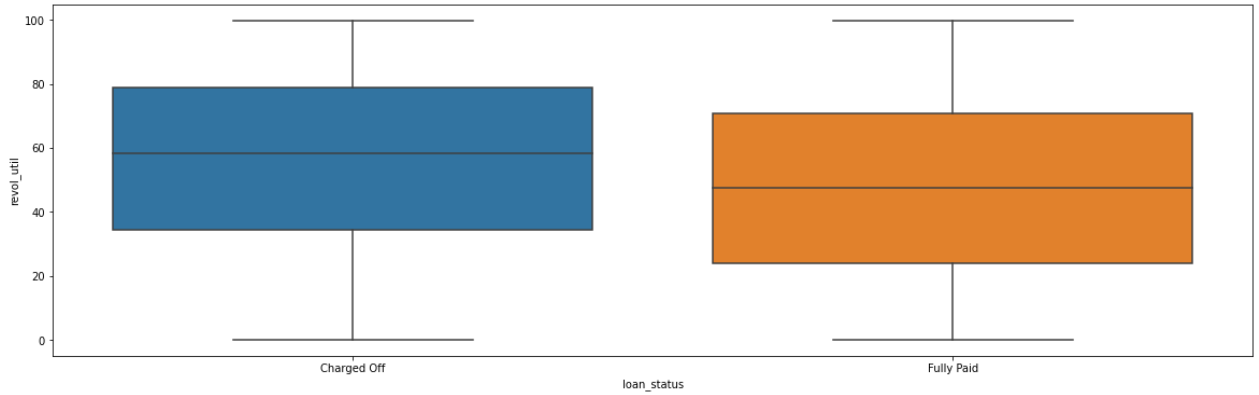


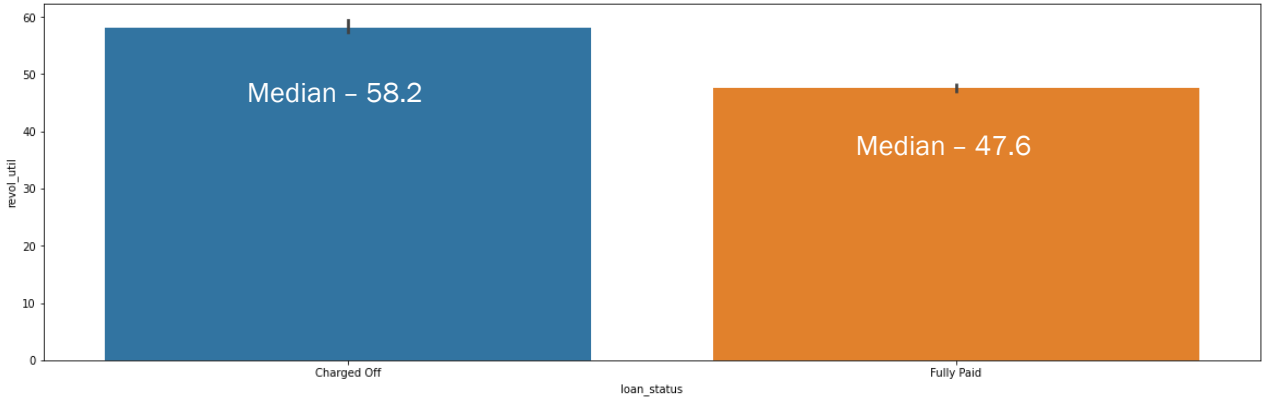
Observations

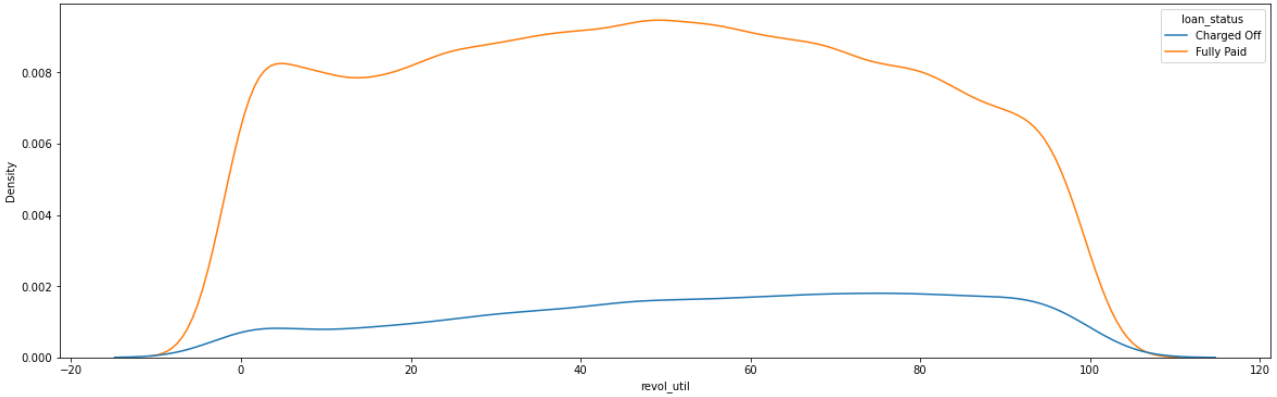
- Median for revol_bal for Charge off segment is 8466, which is quite high compared to Fully Paid segment, which is 7959
- From boxplot Median, Q3 and top whisker are higher for Charged off segment
- It indicates higher 'revol_bal' has higher risk of loan been Charged off
- ‘revol_bal’ to be analysed along with other categorical values

Observations and Findings – Bivariate Analysis

Loan_status vs revol_util







Observations

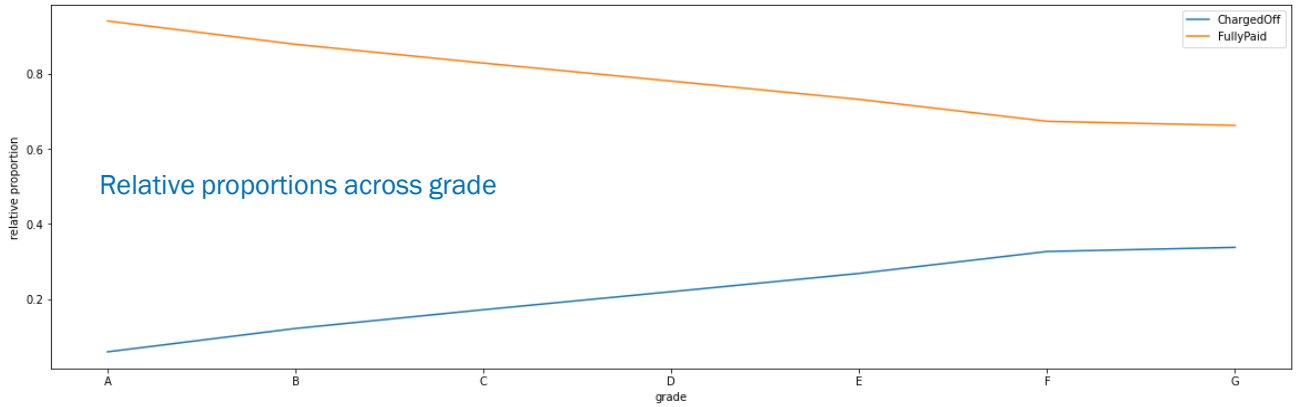
- Median for revol_util for Charge off segment is 58.2, which is quite high compared to Fully Paid segment, which is 47.6
- From boxplot Q1, Median, Q3 are higher for Charged off segment
- It indicates higher 'revol_util' has higher risk of loan been Charged off
- ‘revol_util’ to be analysed along with other categorical values

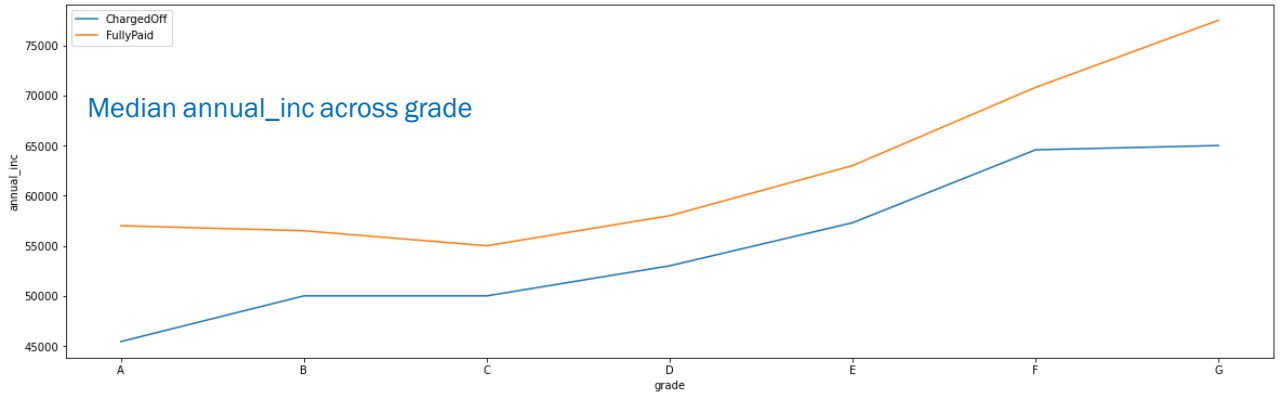
Observations and Findings – Bivariate Analysis

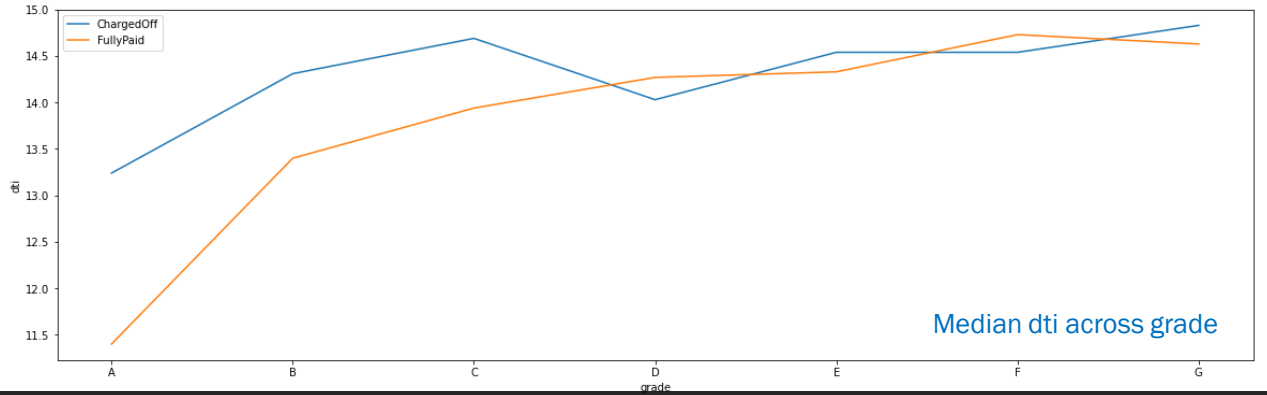
Loan_status vs grade

loan_status grade	Charged Off	Fully Paid
A	0.059930	0.940070
B	0.122056	0.877944
C	0.171943	0.828057
D	0.219862	0.780138
E	0.268494	0.731506
F	0.326844	0.673156
G	0.337793	0.662207

Cross tabulation of grade vs loan_status showing the proporations





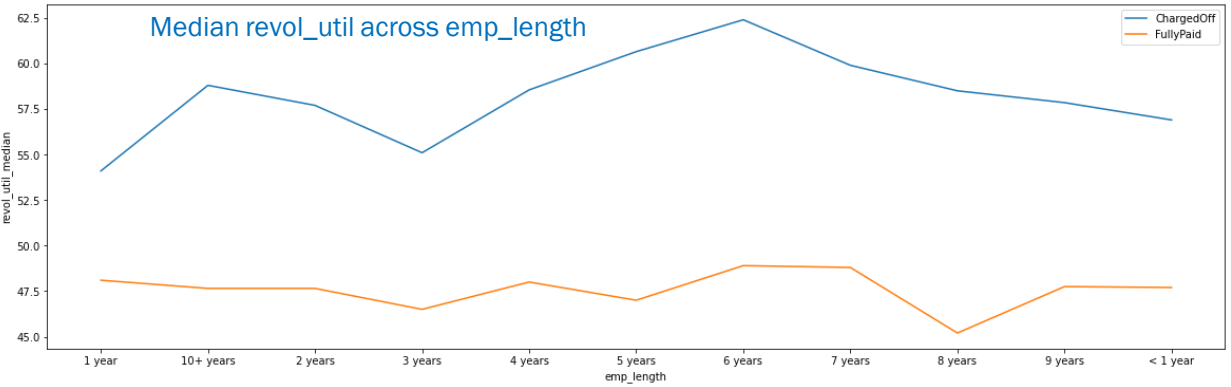


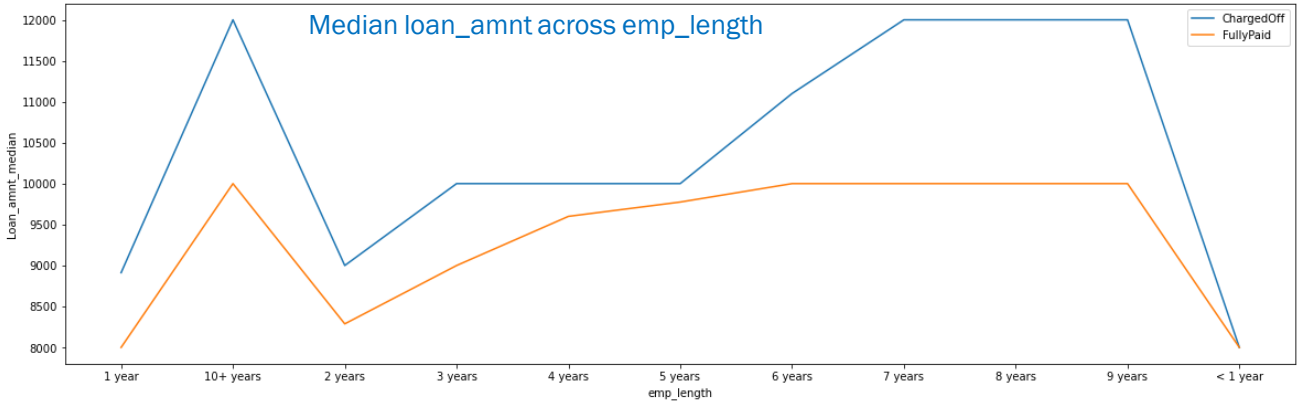
Observations

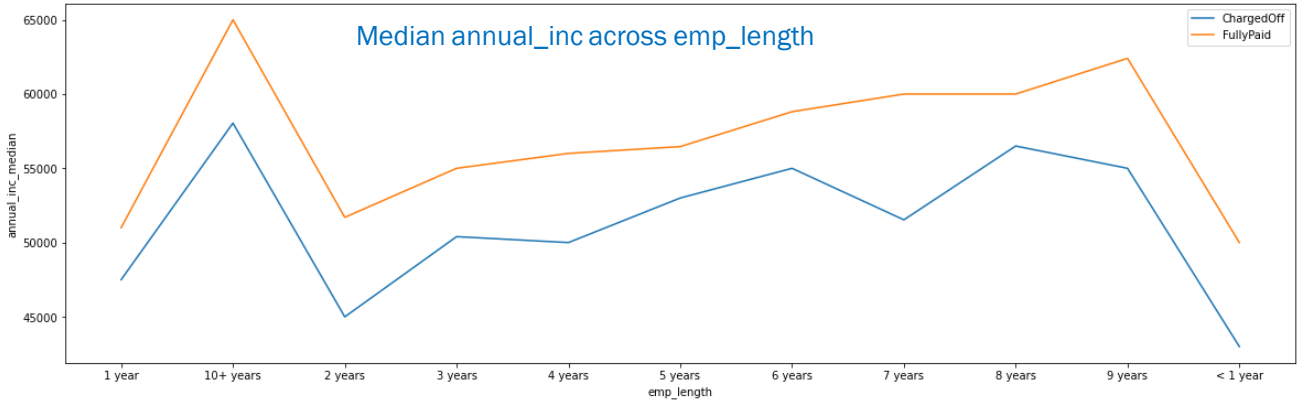
- The risk of Charge Off increase linearly for Grade segments from A to G, higher for grades D, E, F and G
- The median annual_inc across the grade is low for Charged off segment
- Higher median value for dti for grades A, B, C good indicators to detect risks for these grades
- ‘grade’ when used with ‘annual_inc’ and ‘dti’ can be a good indicator

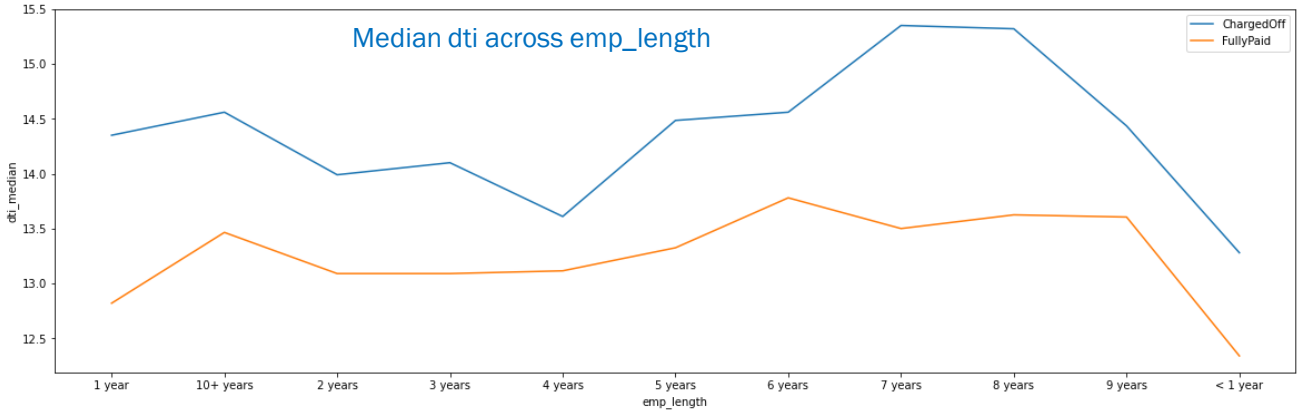
Observations and Findings – Bivariate Analysis

Loan_status vs emp_length







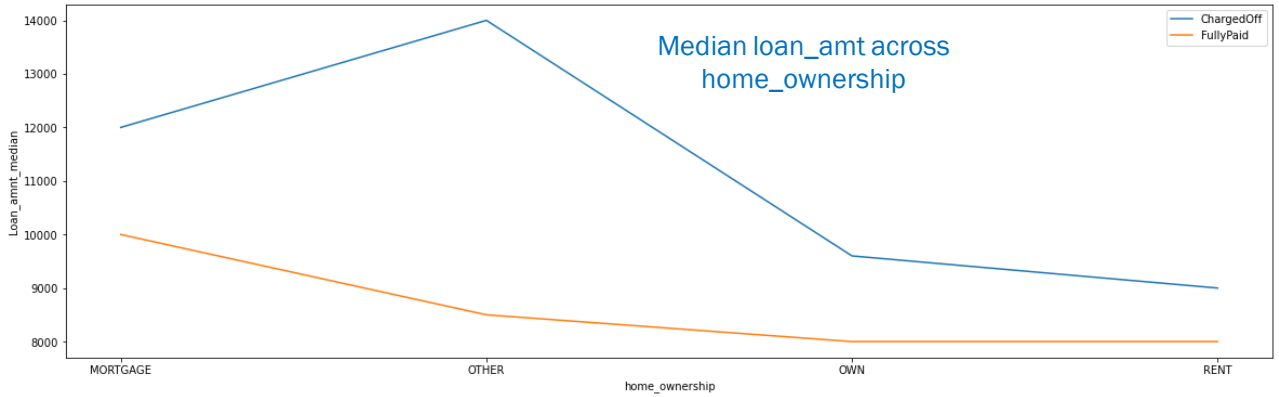


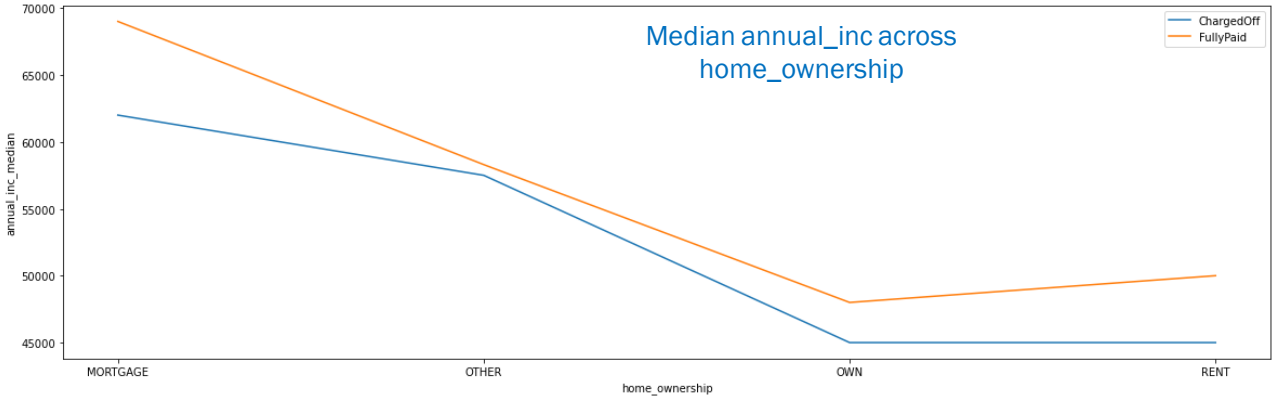
Observations

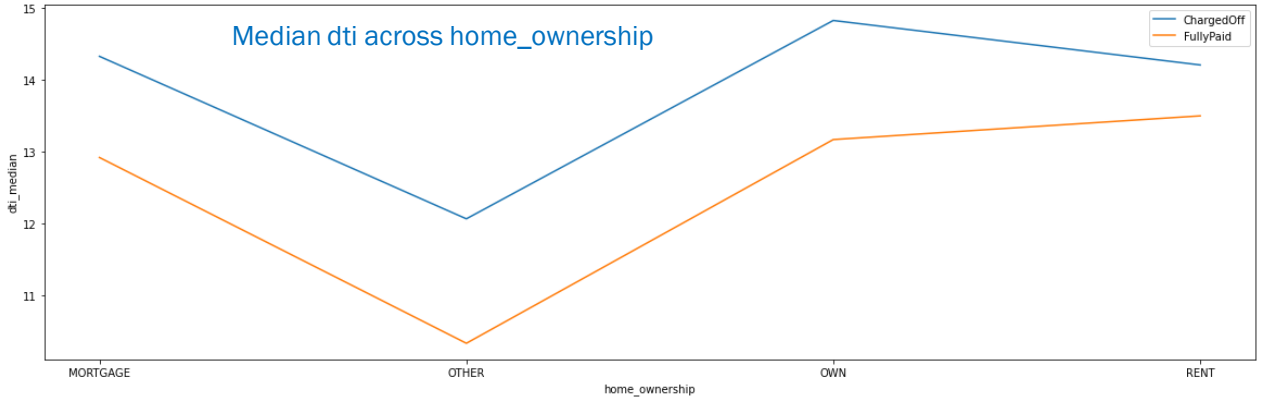
- The median values of 'loan_amnt', 'annual_inc, 'dti', and 'revol_util' differ greatly for both segments of loan_status across all the categories of 'emp_length', median value is higher for above variable for Charged Off Segment
- ‘emp_length’ when used with these variables can be a good indicator

Observations and Findings – Bivariate Analysis

Loan_status vs home_ownership





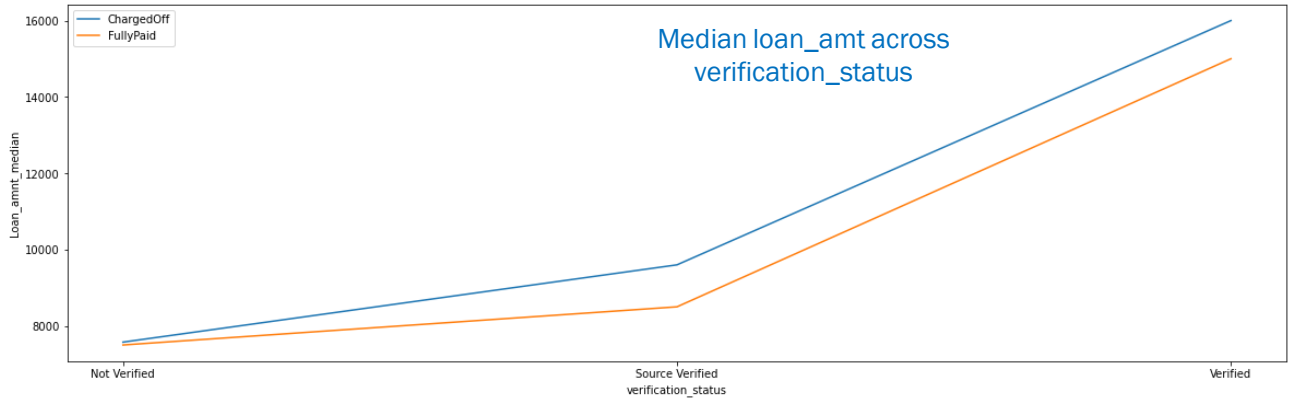


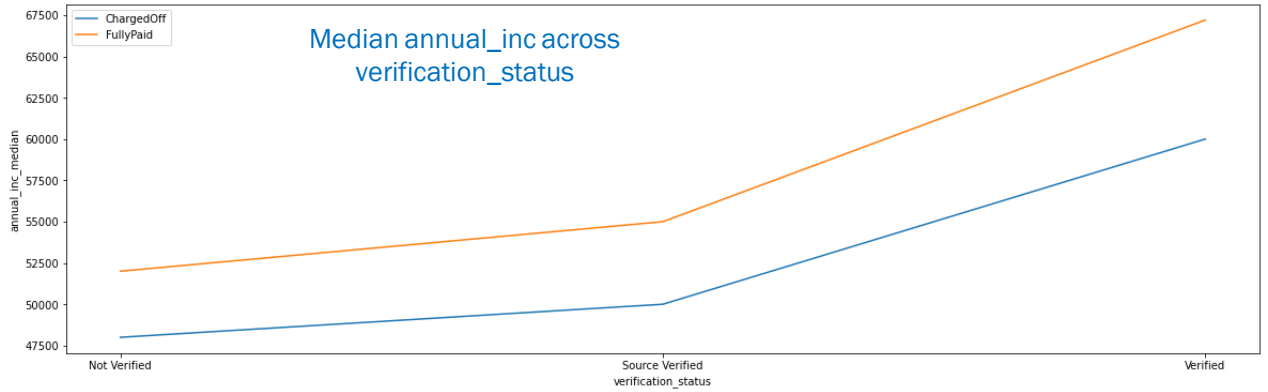
Observations

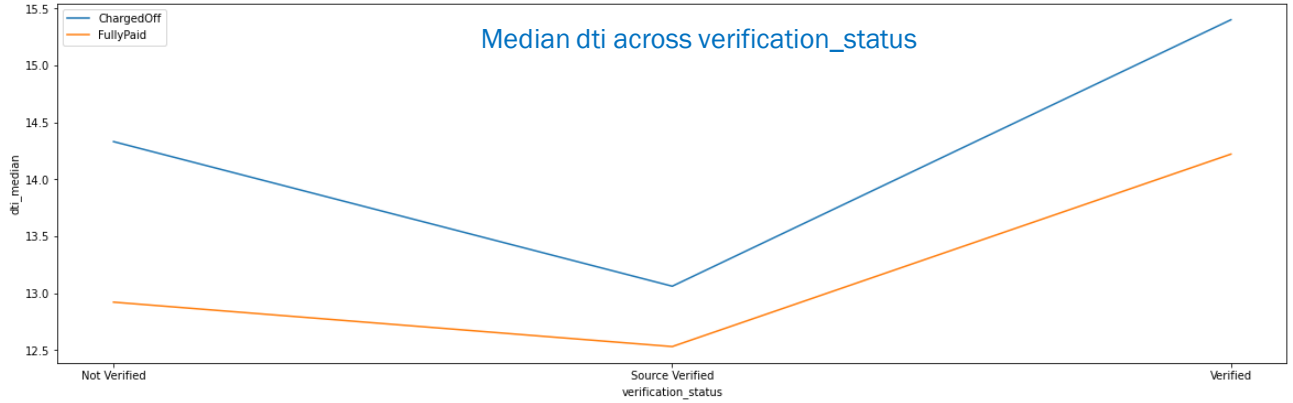
- The median values of 'loan_amnt', 'annual_inc, 'dti', differ greatly for both segments of loan_status across all the categories of 'home_ownership', median value is higher for Charged Off Segment
- 'home_ownership' can be a good indicator, when used with above mentioned variables

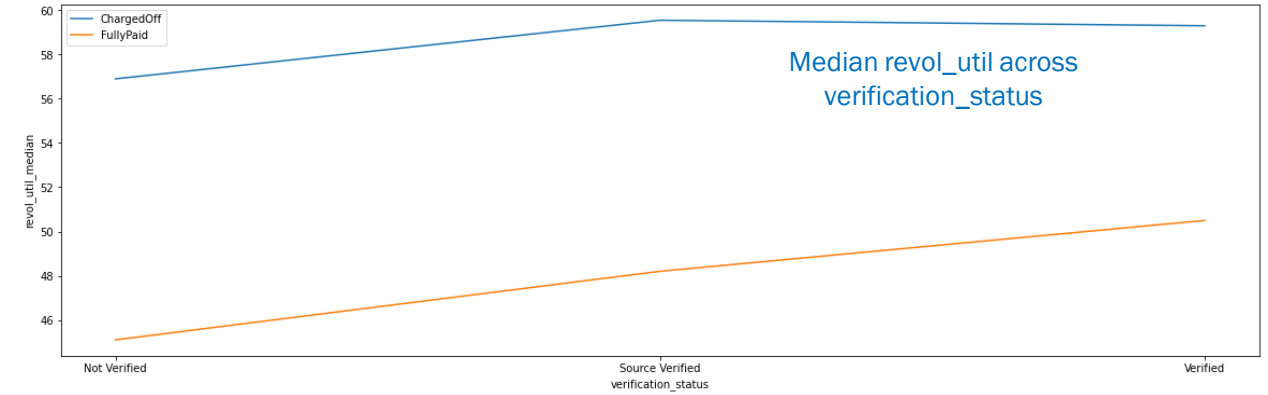
Observations and Findings – Bivariate Analysis

Loan_status vs verification_status









Observations

- The median values of 'loan_amnt', 'annual_inc, 'dti', 'revol_bal' and 'revol_util' differ greatly for both segments of loan_status across all the categories of 'verification_status'
- Median value is higher for Charged Off Segment
- 'verification_status' can be a good indicator, when used with above mentioned variables

Observations and Findings – Bivariate Analysis

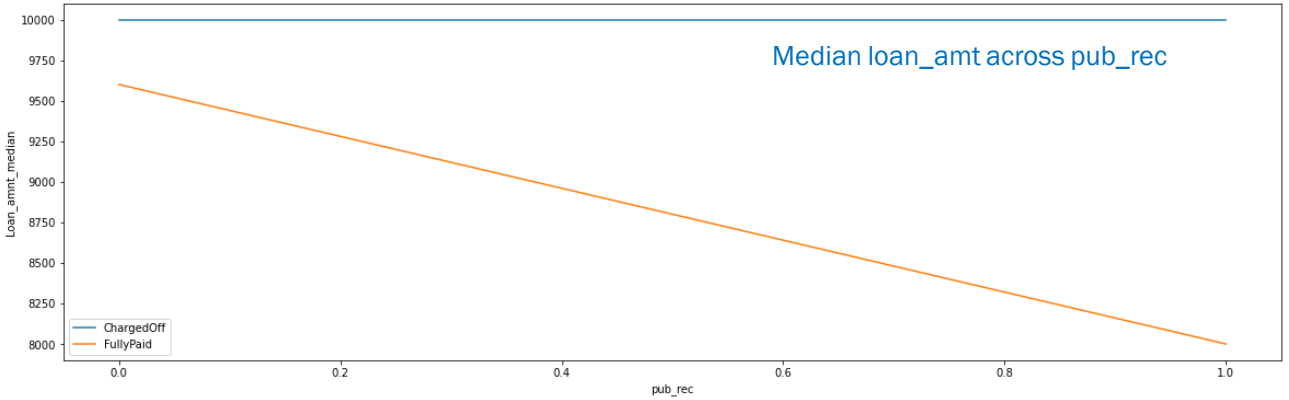
Loan_status vs pub_rec

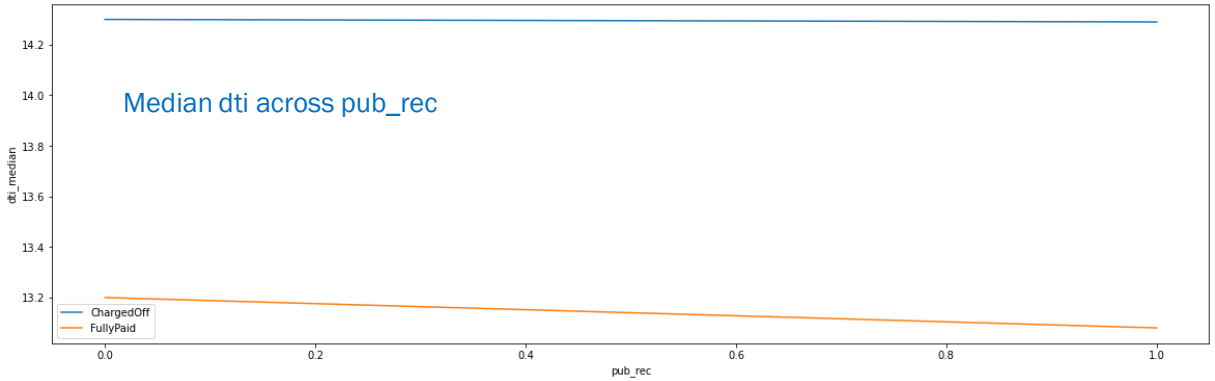
Original categories of pub_rec

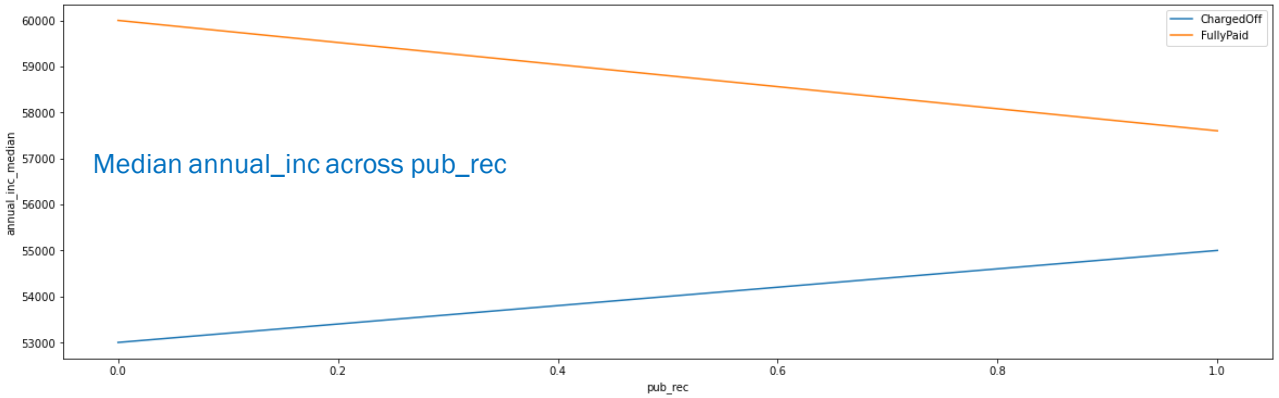
Category	Count
0	36507
1	2013
2	48
3	7
4	2

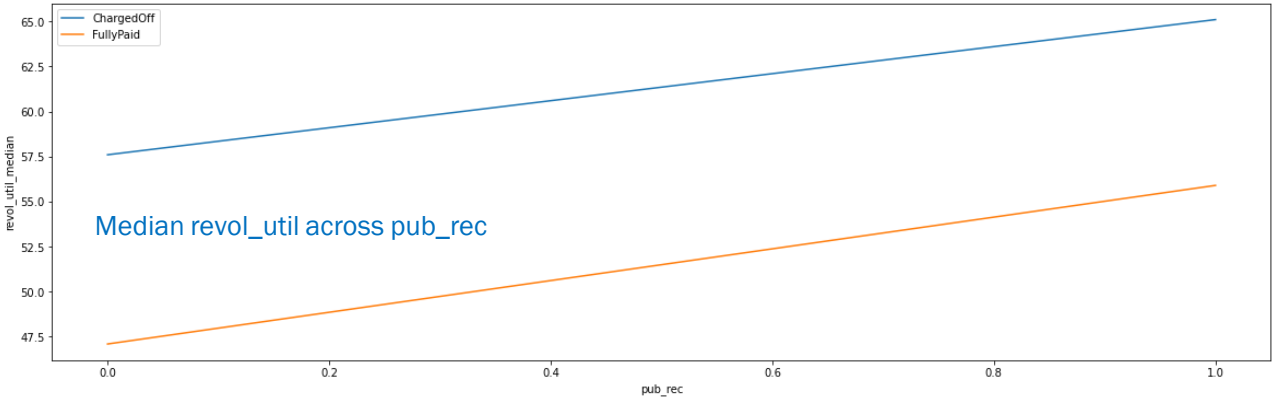
Regrouped categories of pub_rec

Category	Count
0	36507
1	2070









Observations

- 'pub_rec' had 5 different categories, we have created 2 categories out of it, 0 and 1, 1 is created by combining 1, 2, 3 & 4 categories, as they had very small representations
- After combining the category 1 has the high proportion of Charge Off
- The median values of 'loan_amt', 'annual_inc', 'dti', 'revol_bal' and 'revol_util' for Charge off segment is higher
- pub_rec after re-grouping can be a driving factor

Observations and Findings – Bivariate Analysis

Loan_status vs pub_rec_bankruptcies

Original categories of pub_rec_bankruptcies

Category	Count
0	36935
1	1637
2	5

Regrouped categories of pub_rec_bankruptcies

Category	Count
0	36935
1	1645

loan_status	Charged Off	Fully Paid
pub_rec_bankruptcies		
0	10000	9600
1	10000	8000

Pivot table for loan_status vs pub_rec_bankruptcies with loan_amount

loan_status	Charged Off	Fully Paid
pub_rec_bankruptcies		
0	53000.0	60000.0
1	56000.0	58000.0

Pivot table for loan_status vs pub_rec_bankruptcies with annual_inc

loan_status	Charged Off	Fully Paid
pub_rec_bankruptcies		
1	0.224117	0.775883
0	0.142385	0.857615

Cross tabulation of loan_status vs pub_rec_bankruptcies

loan_status	Charged Off	Fully Paid
pub_rec_bankruptcies		
0	14.28	13.19
1	14.55	13.41

Pivot table for loan_status vs pub_rec_bankruptcies with dti

loan_status	Charged Off	Fully Paid
pub_rec_bankruptcies		
0	9273.0	8774.0
1	8682.5	7255.5

Pivot table for loan_status vs pub_rec_bankruptcies with revol_bal

loan_status	Charged Off	Fully Paid
pub_rec_bankruptcies		
0	57.6	47.2
1	66.7	57.2

Pivot table for loan_status vs pub_rec_bankruptcies with revol_util

Observations

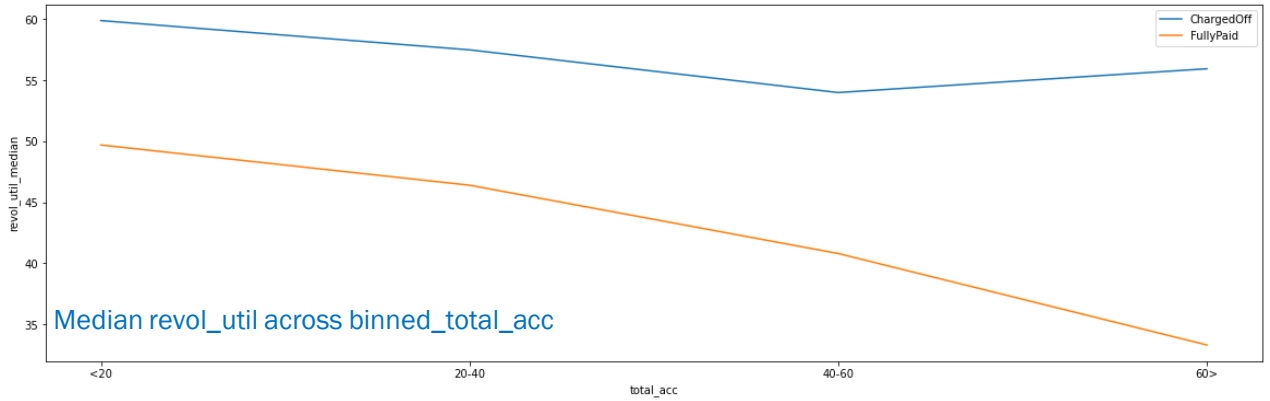
- 'pub_rec_bankruptcies' had 3 different categories, we have created 2 categories out of it, 0 and 1, 1 is created by combining 1, and 2 categories, as they had very small representations
- After combining the category 1 has the high proportion of Charge Off
- The median values of 'loan_amt', 'annual_inc', 'dti', 'revol_bal' and 'revol_util' for Charge off segment is higher
- pub_rec_bankruptcies after re-grouping can be a driving factor

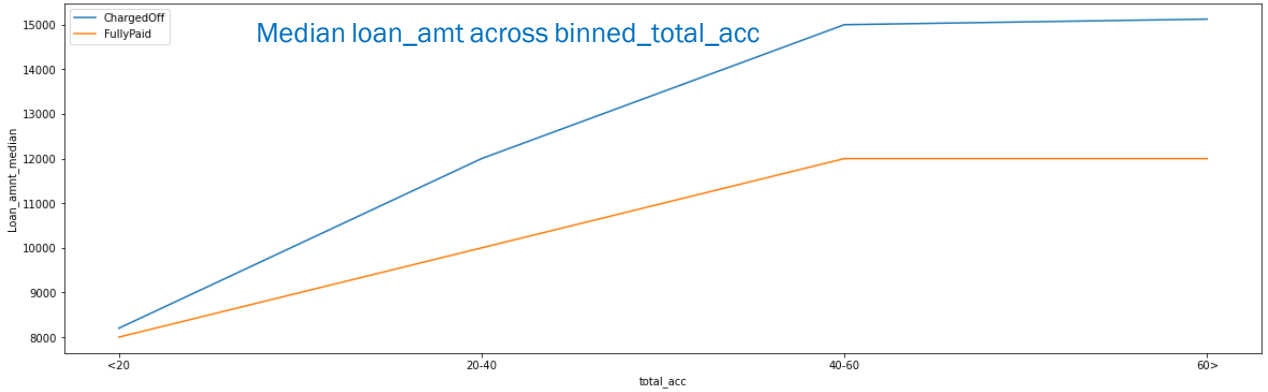
Observations and Findings – Bivariate Analysis

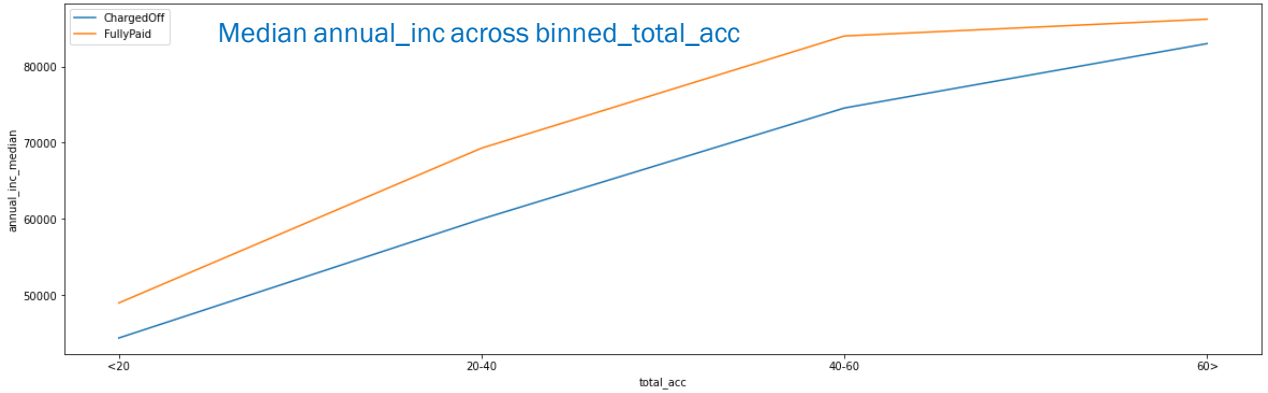
Loan_status vs total_acc

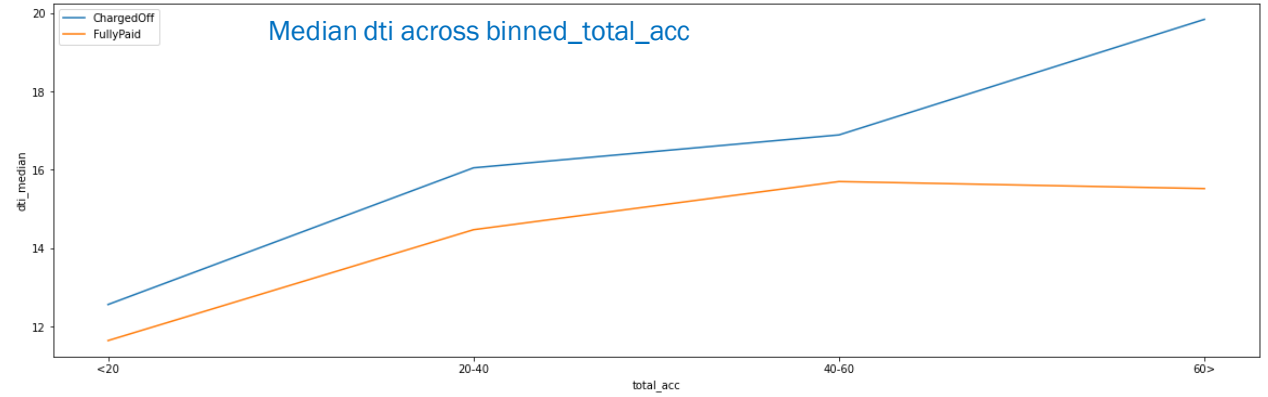
loan_status	Charged Off	Fully Paid
binned_total_acc		
<20	0.153716	0.846284
40-60	0.139099	0.860901
20-40	0.137820	0.862180
60>	0.126214	0.873786

Proportions of Charge off vs Fully Paid after binning the 'total_acc'









Observations

- 'total_acc' had 82 different categories, we have binned thouse into 4 groupd - 0-20, 20-40, 40-60 and 60+
- After binning, proportion of Charged off to Fully paid is constant across 4 groups
- The median values of 'loan_amnt', 'annual_inc', 'dti', 'revol_bal' and 'revol_util' for Charge off segment is higher
- total_acc after binning can be a driving factor

Conclusion

- Continuous values like – ‘loan_amnt’, ‘dti’, ‘revol_bal’ and ‘revol_util’, have clear higher median values for Charge Off and lower median values for Fully Paid
- ‘annual_inc’, has clear higher median value for Fully paid and lower median value for Charge Off
- However, continuous variables on their own has less power to distinguish between risky application vs non-risky application. But when used along with categorical variables like ‘emp_lenght’, ‘verification-status’, they are good indicator of risky application
- This holds same for categorical variables, hence propose the use of categorical variables along with continuous variable to classify risky application
- From our analysis we propose these variables are driving variables that are indicators of loan default;

Continuous variables	Categorical variables
loan_amnt	emp_length
annual_inc	Verification_status
dti	pub_rec (regrouped)
revol_bal	Pub_rec_bankruptcies (regrouped)
revol_util	total_acc (binned)
	grade
	open_acc