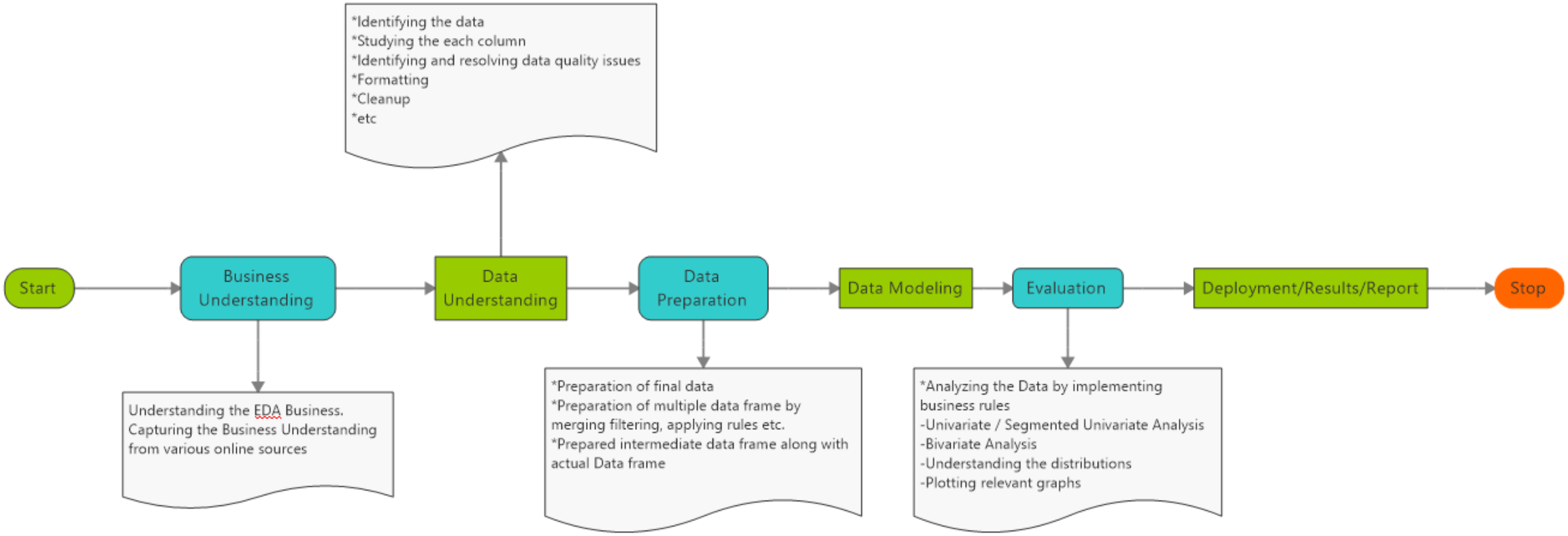
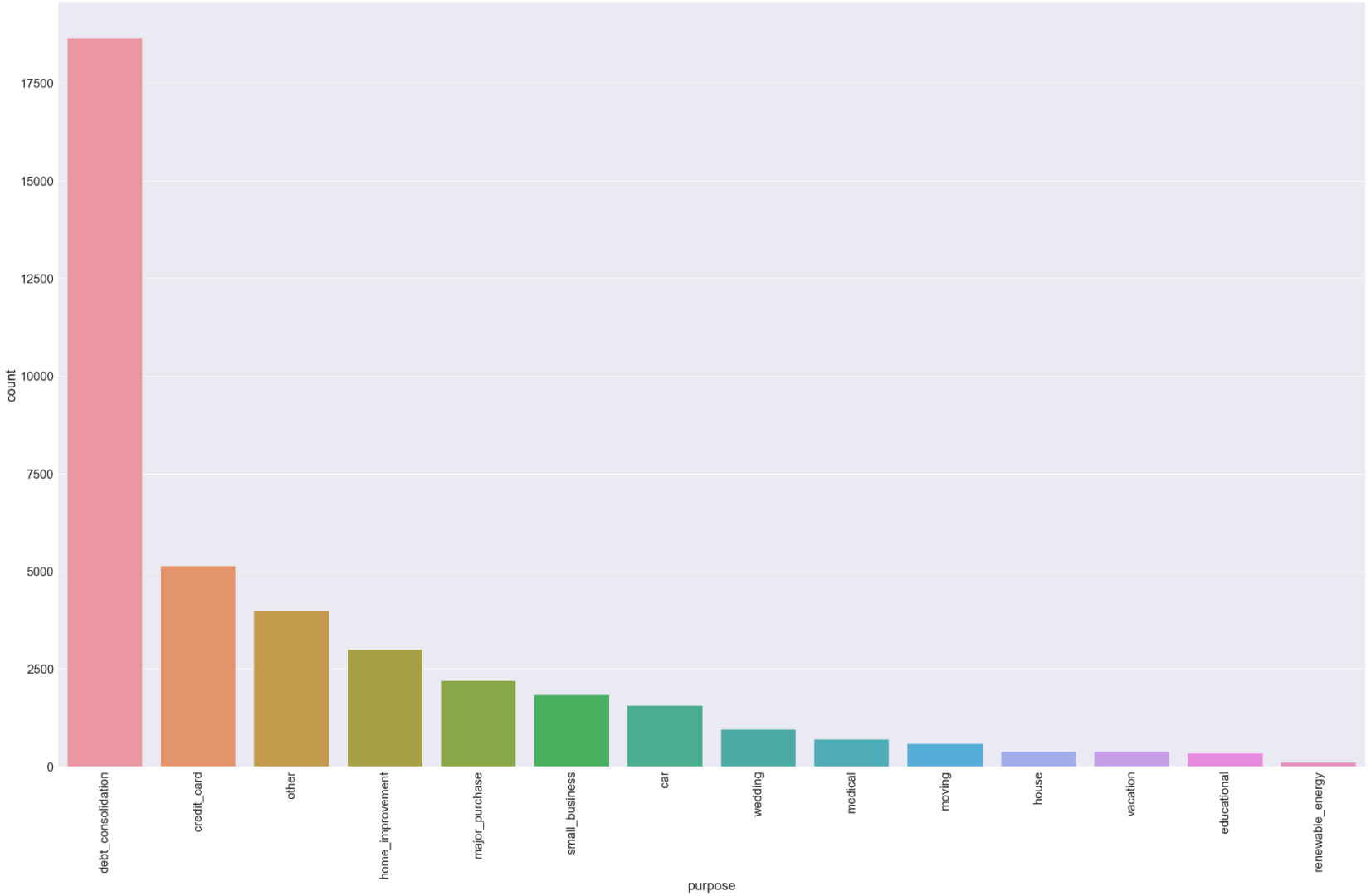
EDA CASE STUDY-GRAMENER

SUBMISSION

Problem Solving Methodology



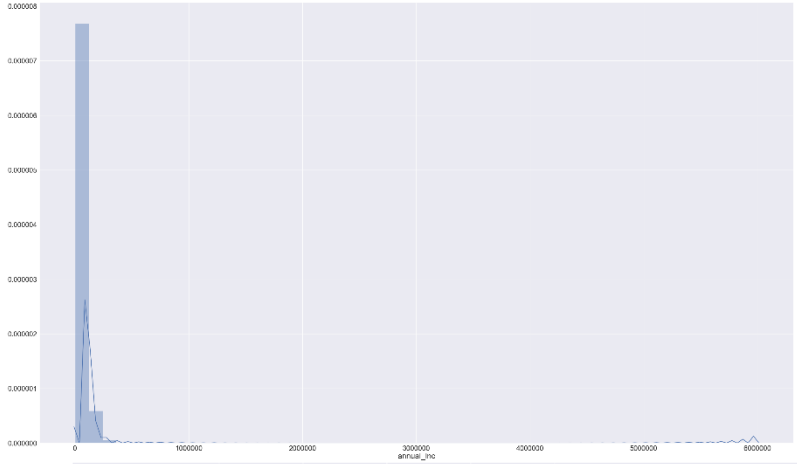
Univariate Analysis on Purpose

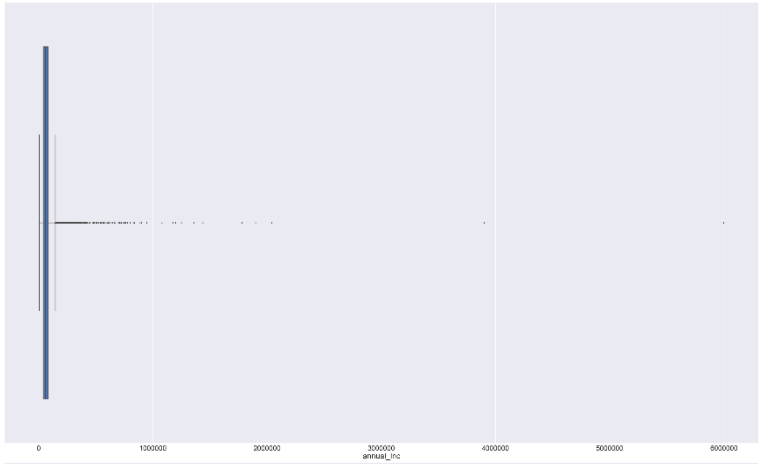


- The Histogram for “purpose” shows that most of the customer bought the loan for debt consolidation and credit card means maximum people has given unsecured loan to clear their old debit which is risky in terms of defaulters

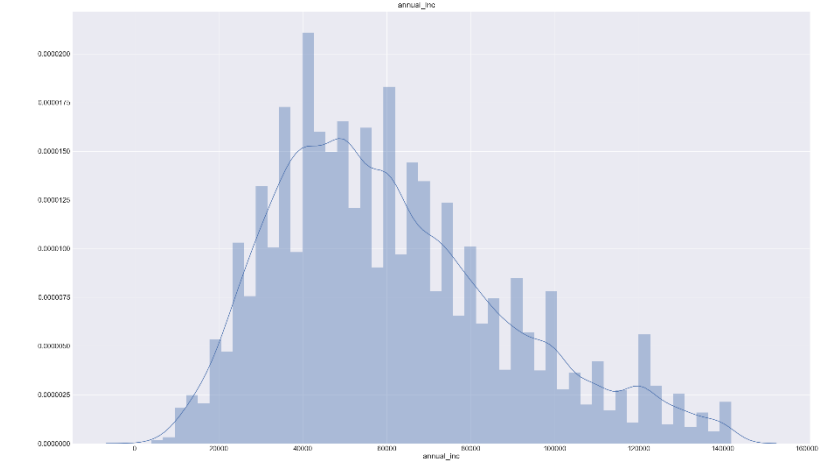
```
debt_consolidation    18641
credit_card            5130
other                  3993
home_improvement       2976
major_purchase         2187
small_business         1828
car                    1549
wedding                 947
medical                 693
moving                  583
house                   381
vacation                381
educational             325
renewable_energy        103
Name: purpose, dtype: int64
```

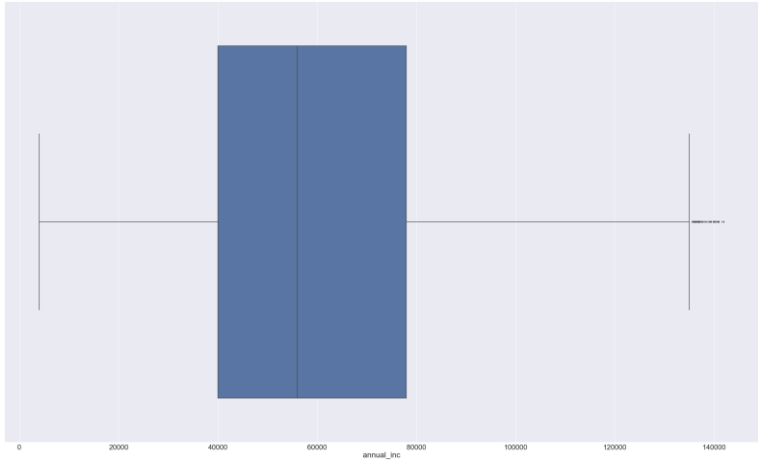
Univariate Analysis on Annual Income






```
count     3.971700e+04
mean      6.896893e+04
std       6.379377e+04
min       4.000000e+03
25%       4.040400e+04
50%       5.900000e+04
75%       8.230000e+04
max       6.000000e+06
Name: annual_inc, dtype: float64
```



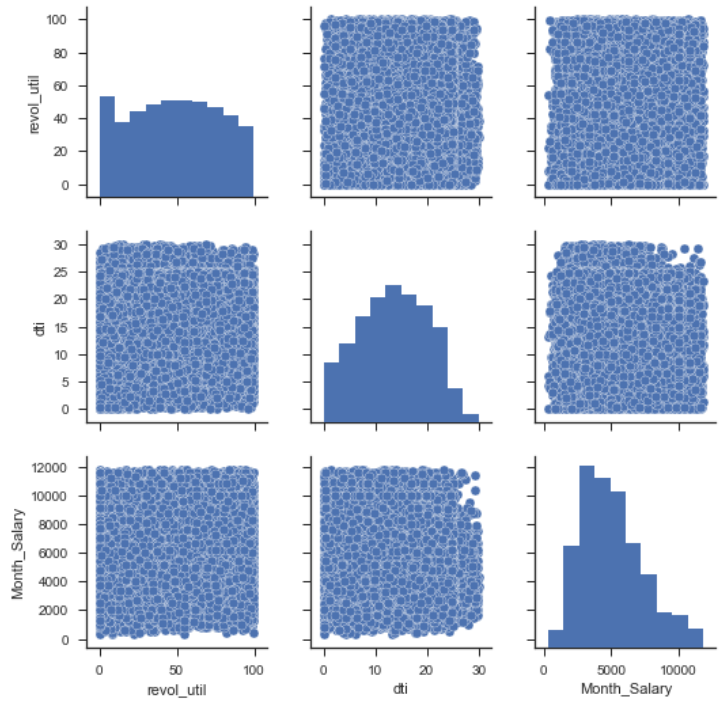



```
count     37743.000000
mean      61071.860572
std       27830.995882
min        4000.000000
25%       40000.000000
50%       56000.000000
75%       78000.000000
max      142000.000000
Name: annual_inc, dtype: float64
```

- The Histogram for annual income shows that there are very few customers with annual income is more than 1 million and less than 5000 . Income with income more than 1million can be treat as outlier which can possibly effect our normality
- Plot shown after removing values greater than 95 percentile and plotting it , this was done to remove certain outliers
- We also observed there were frequent spikes were loan amount was whole numbers like 5K,10K,20K,30K … etc

Segmented Univariate Analysis

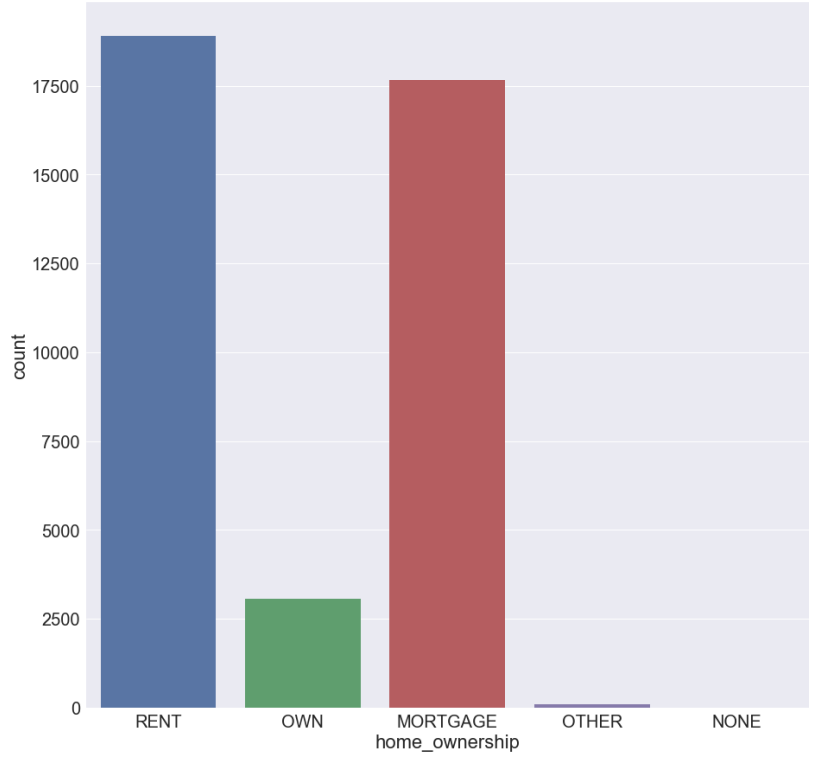
	dti		annual_inc		revol_util	
	mean	median	mean	median	mean	median
loan_status						
Charged Off	14.000624	14.29	62427.298034	53000.0	55.414095	58.20
Current	14.750009	15.05	75430.665105	65000.0	53.204482	54.95
Fully Paid	13.148421	13.20	69862.503328	60000.0	47.482755	47.50



With this analysis for loan status vs Average of different variable shows that

1) With Average loan status vs Annual Income, we can say that there is high chances of defaulters if annual income is low
2) With Average loan status vs revol_util , means we can say that the customer who are using relatively higher credit borrowing against available credit revolving credit are having higher chance of default
3) With loan status vs Average dti , we can say that the customer with higher debts are having higher chance of default

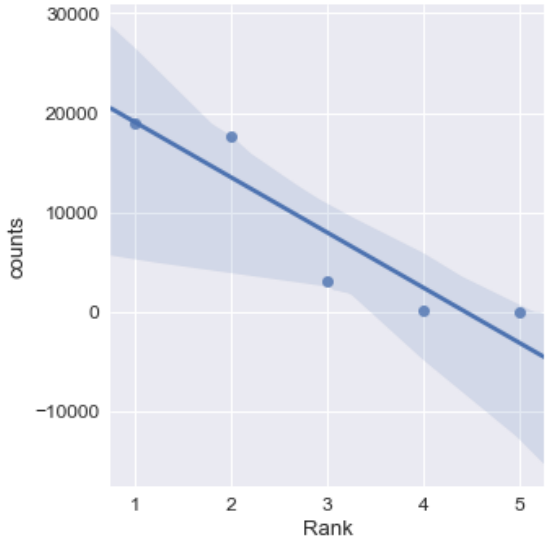
Univariate Analysis on Home Ownership



	home_ownership	counts	Rank
0	MORTGAGE	17659	2.0
1	NONE	3	5.0
2	OTHER	98	4.0
3	OWN	3058	3.0
4	RENT	18899	1.0

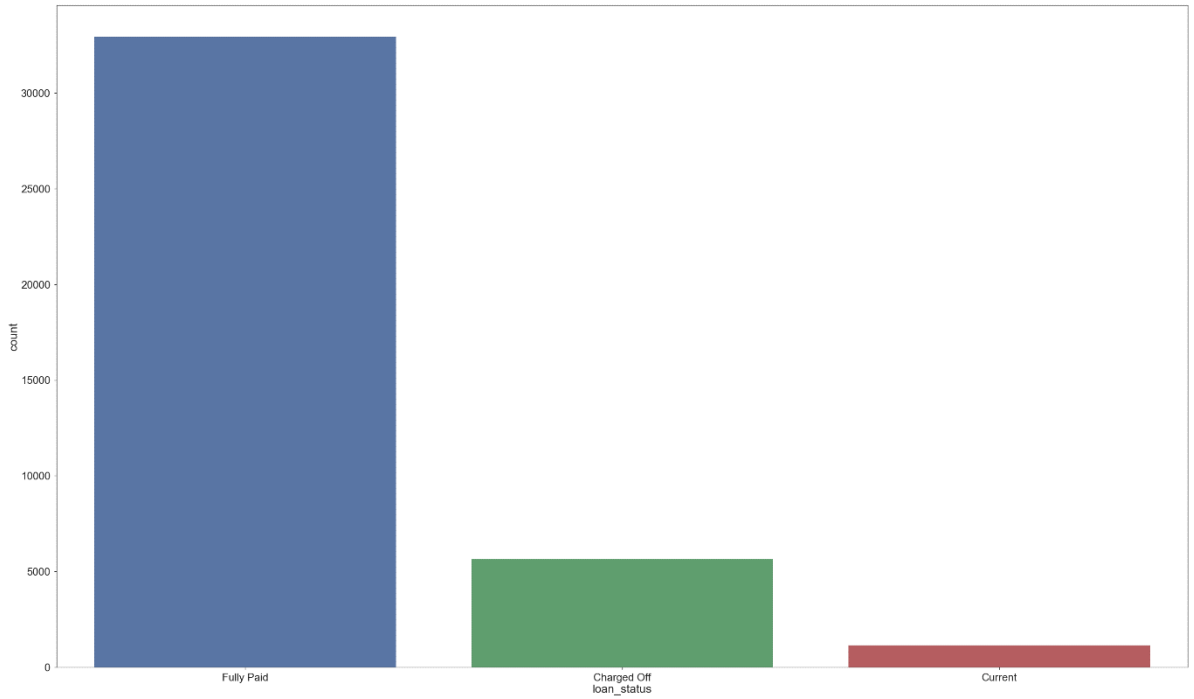
Out[84]: <seaborn.axisgrid.FacetGrid at 0x1e296cf55f8>

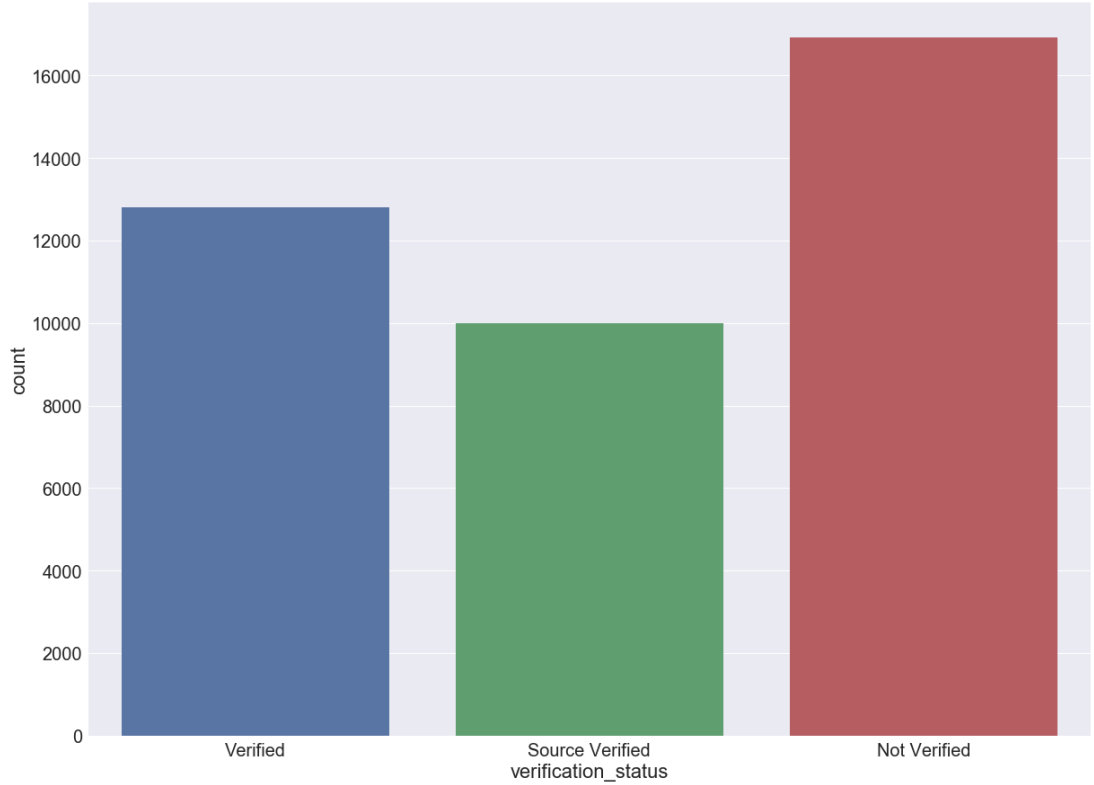
<matplotlib.figure.Figure at 0x1e296119390>



- Power law or rank frequency plot for home_ownership shows that irrespective of defaulters Rented customers are having more requirement of loan than the customer with mortgage and respectively own house customer

Univariate Analysis on Verification and Loan Status





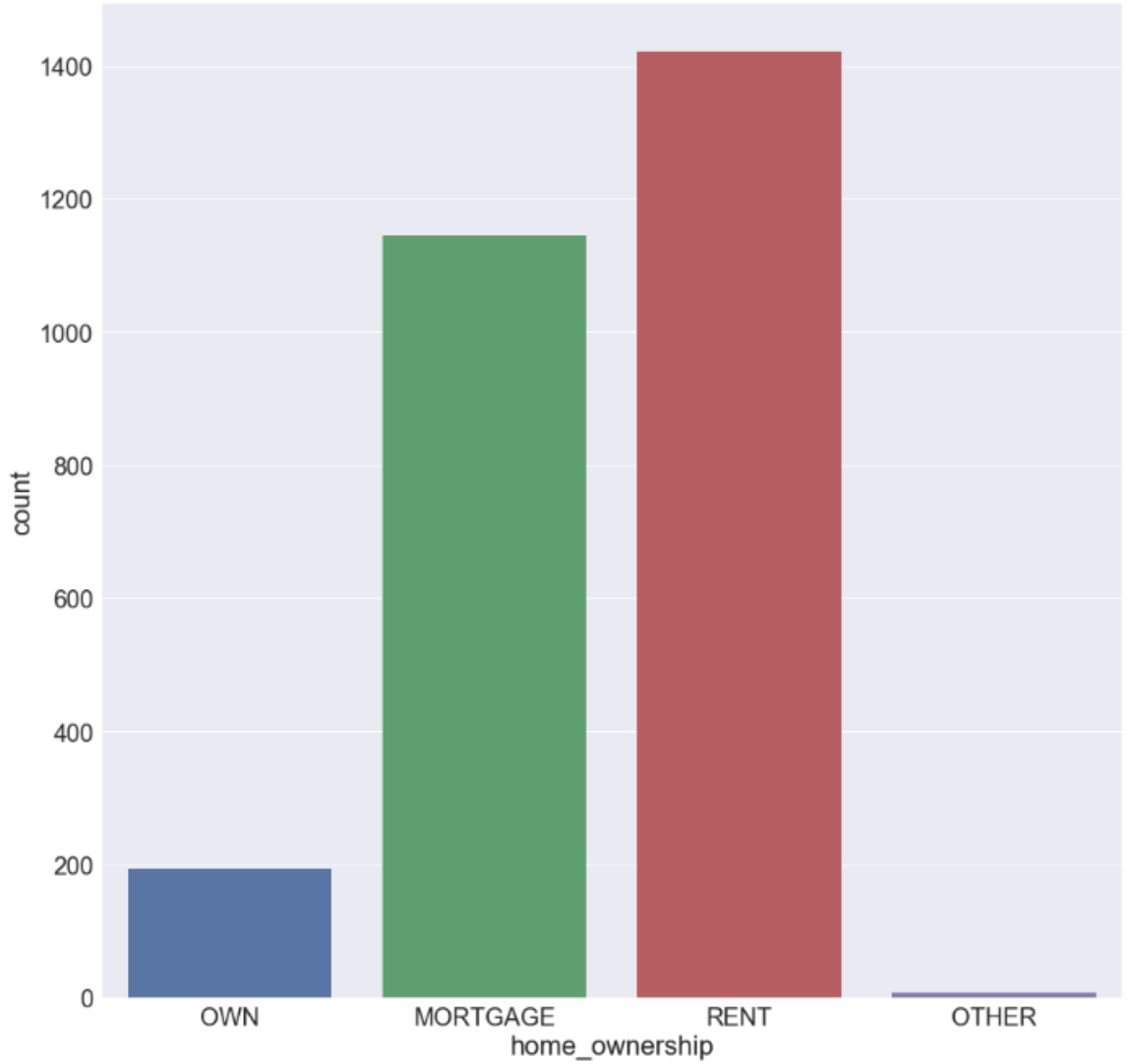
- Loan Status count plot shows the number of account count is highest for Fully Paid followed by Charged off and current
- Count Plot for for verification status shows that there are more customer with Not Verified status than Verified

Home Ownership Analysis

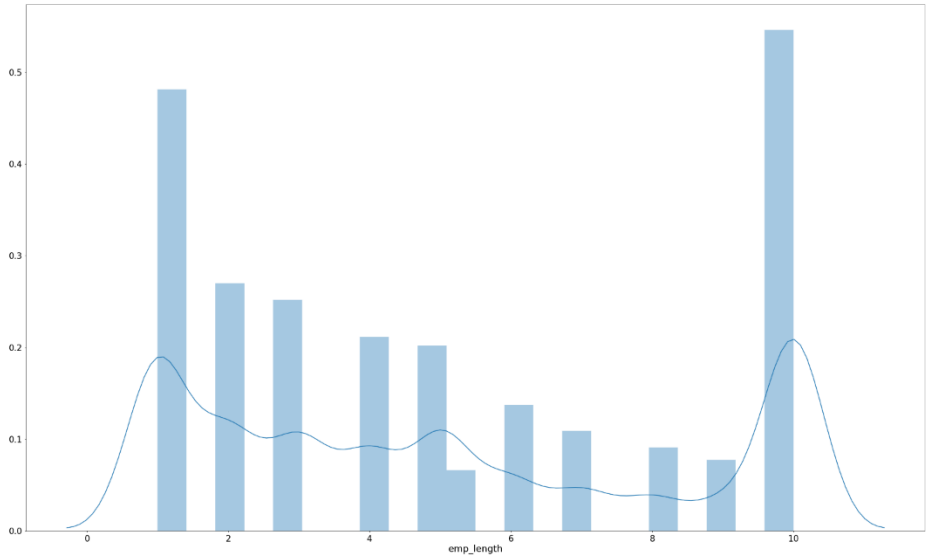
Home_ownership analysis:

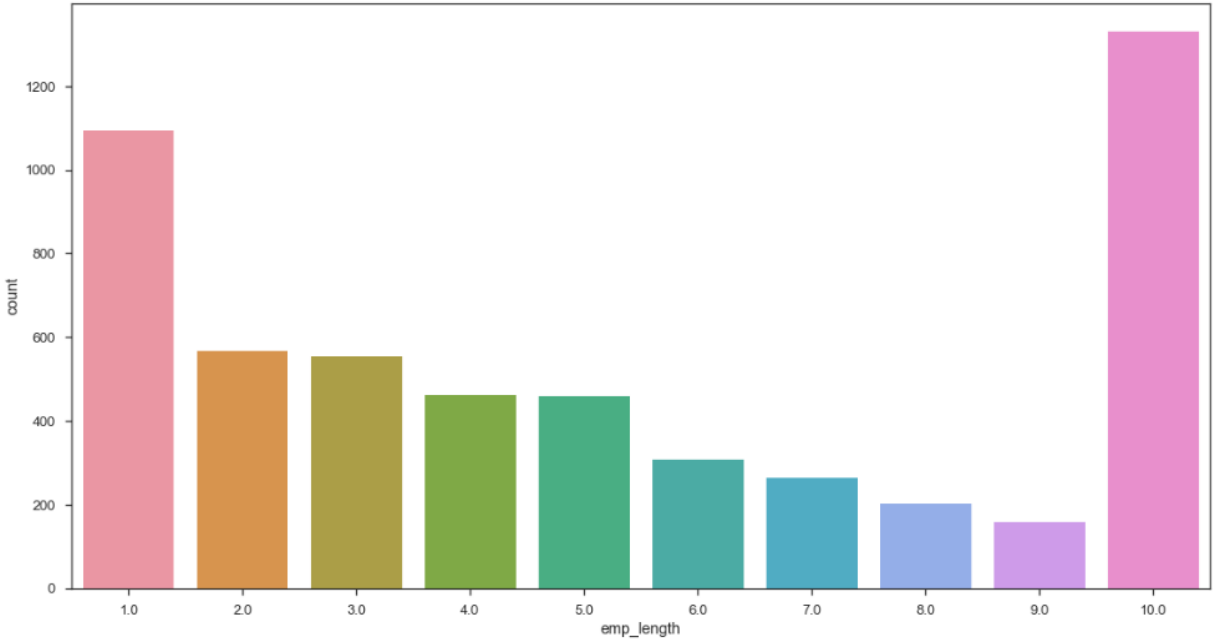
Interestingly the people leaving in Rent have defaulted more than people having Mortage

People having home are less likely to default on loan



Analysis on employee length





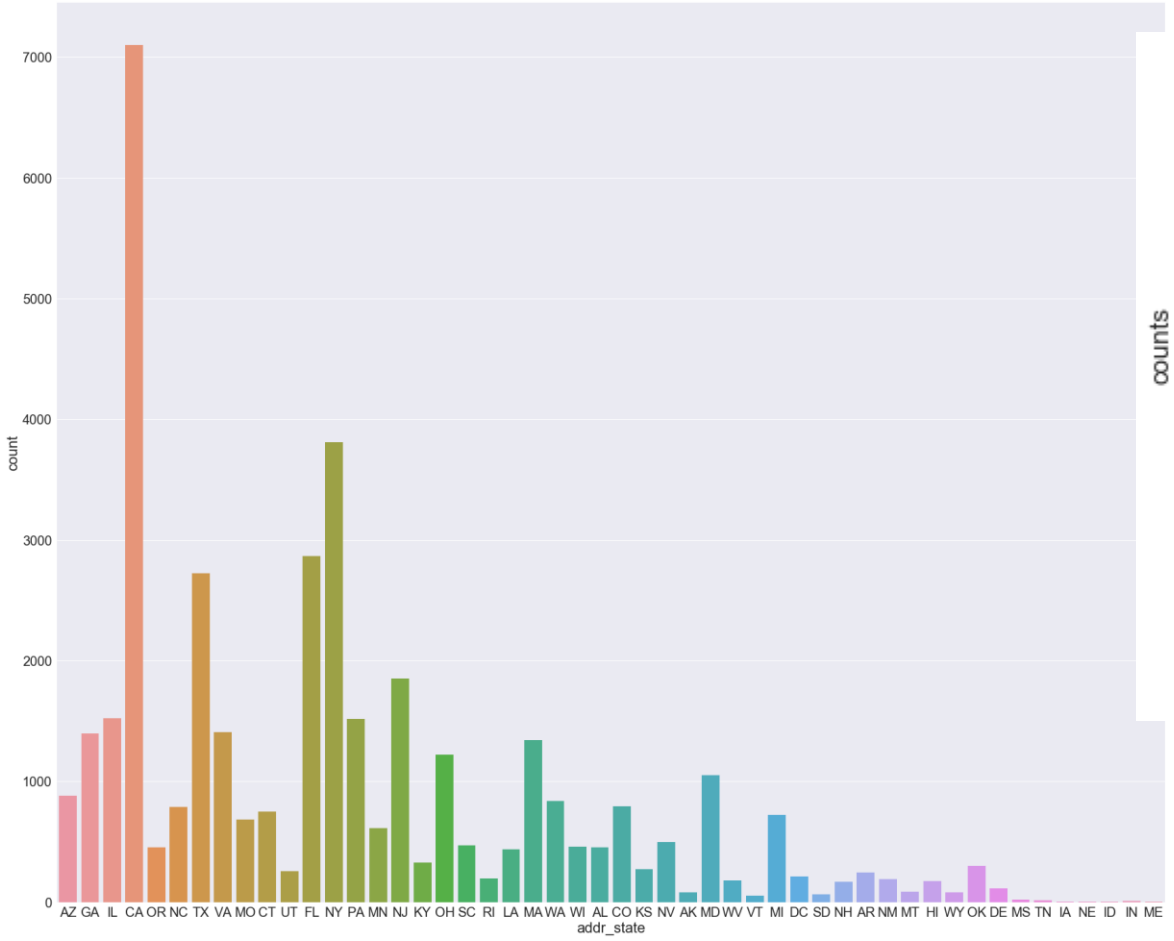
- Employee with experience 1 Year Likely to default experience in work
- 1 or less than 1 year of experience
- 10 or more than 10 year of experience
- - They are the most contributer in loan application
- *Images shown before and after applying Charged Off status Filter

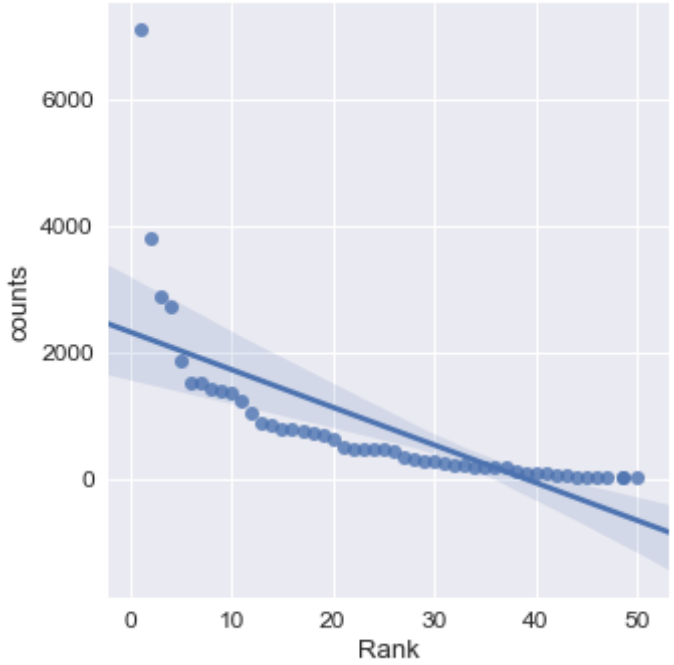
Analysis on address

Out[82]:

	loan_status	addr_state	percentage
4	Charged Off	CA	19.992891
9	Charged Off	FL	8.956815
30	Charged Off	NY	8.796872
39	Charged Off	TX	5.615781
27	Charged Off	NJ	4.940466
10	Charged Off	GA	3.820864
13	Charged Off	IL	3.500977
34	Charged Off	PA	3.198863
41	Charged Off	VA	3.145548

*addr_state code converted to state from source: https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations





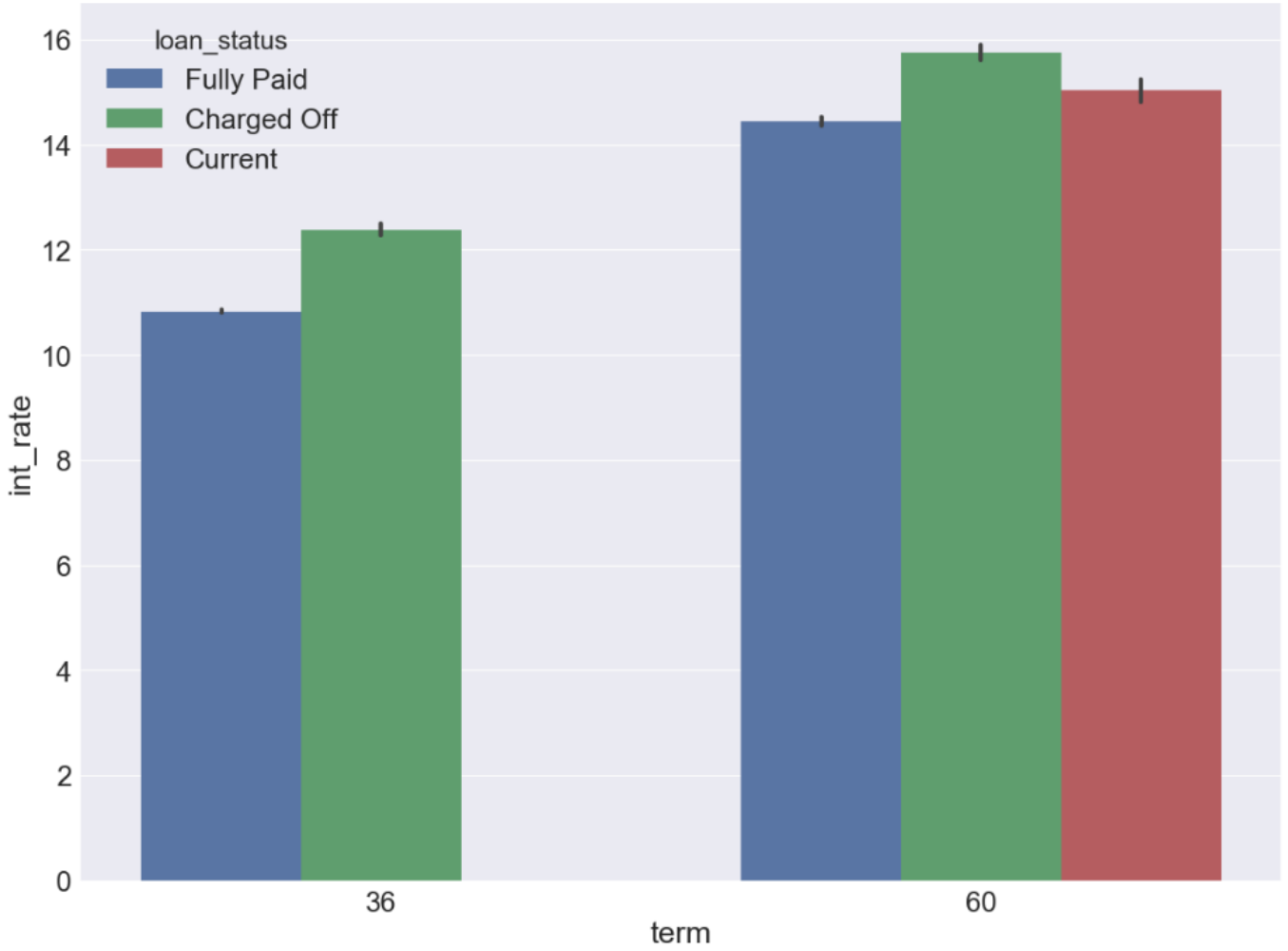
- Power law or rank frequency plot for addr_state shows that irrespective of defaulters there is more customer base in California >(than) New York>Florida >Texas>New Jersey >……and so on

Analysis on Term Vs Loan Status

We also analysed from Bivariate Analysis of Loan_status,term and int_rate that

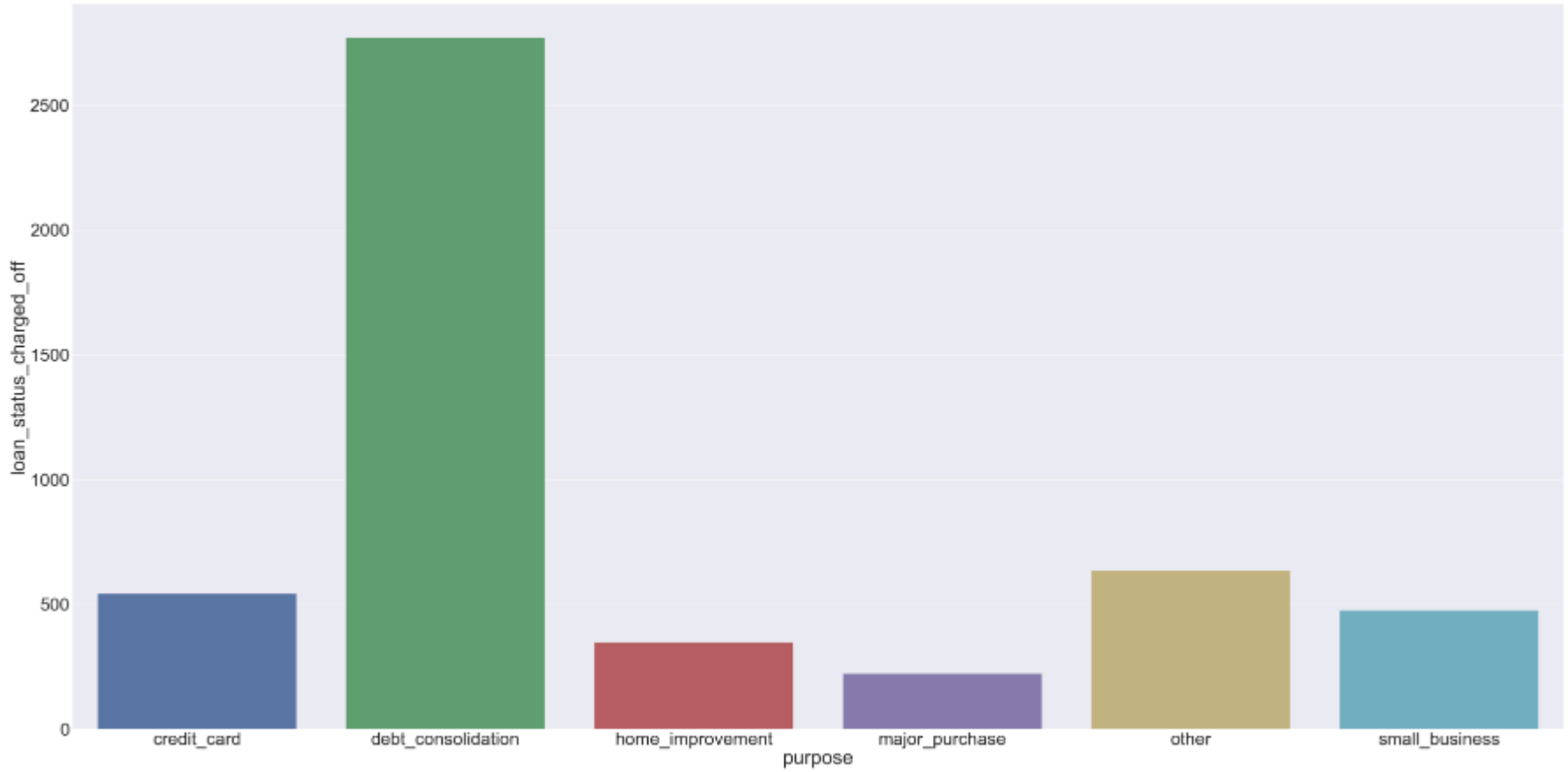
1.Higher the term higher the interest rate

2. High number of Charged Off account in highest term of 60



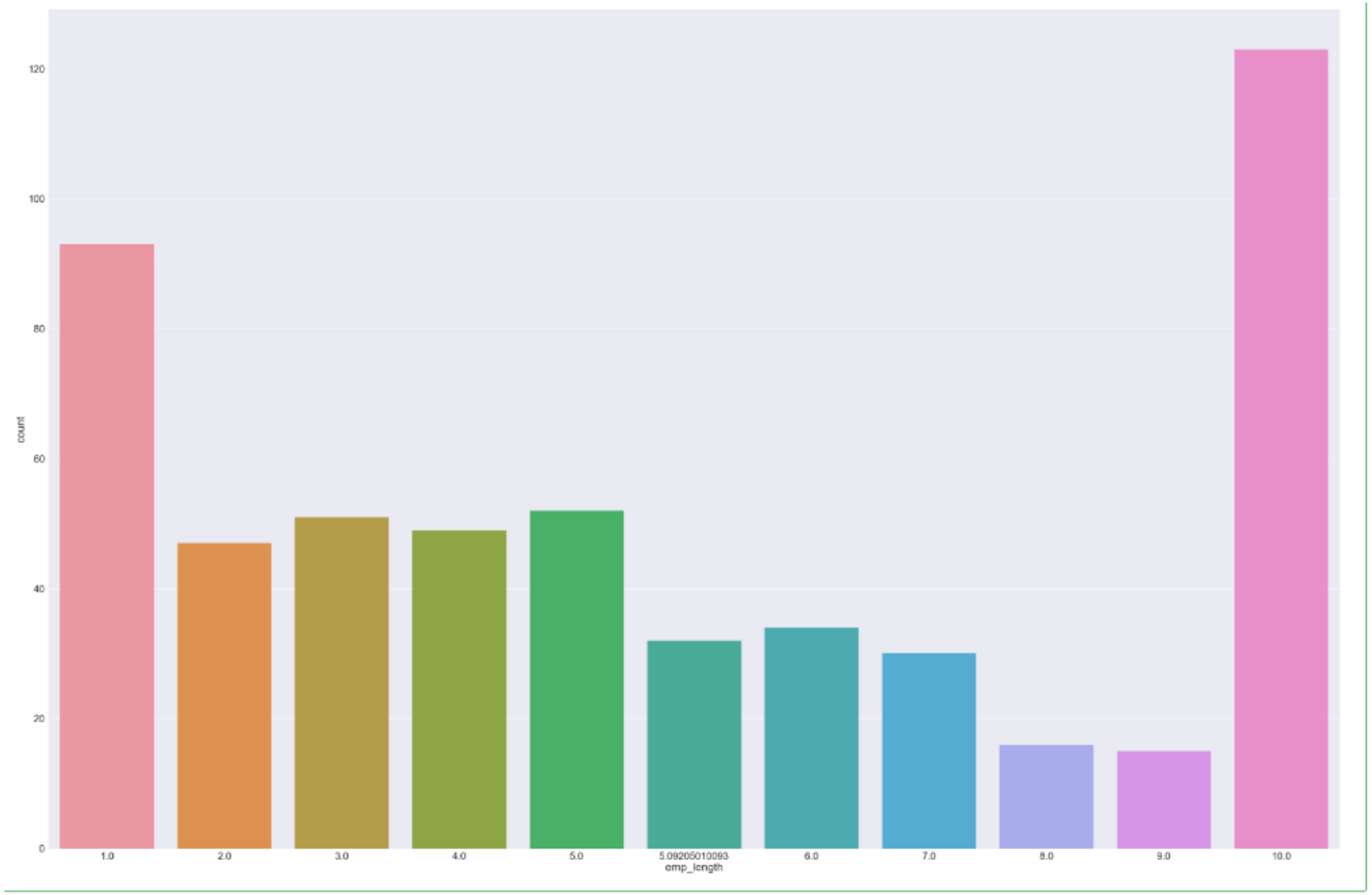
Analysis on Charged Off against purpose

- The Order of risk on loan default in area of debt consolidation>Credit Card>Small business or others

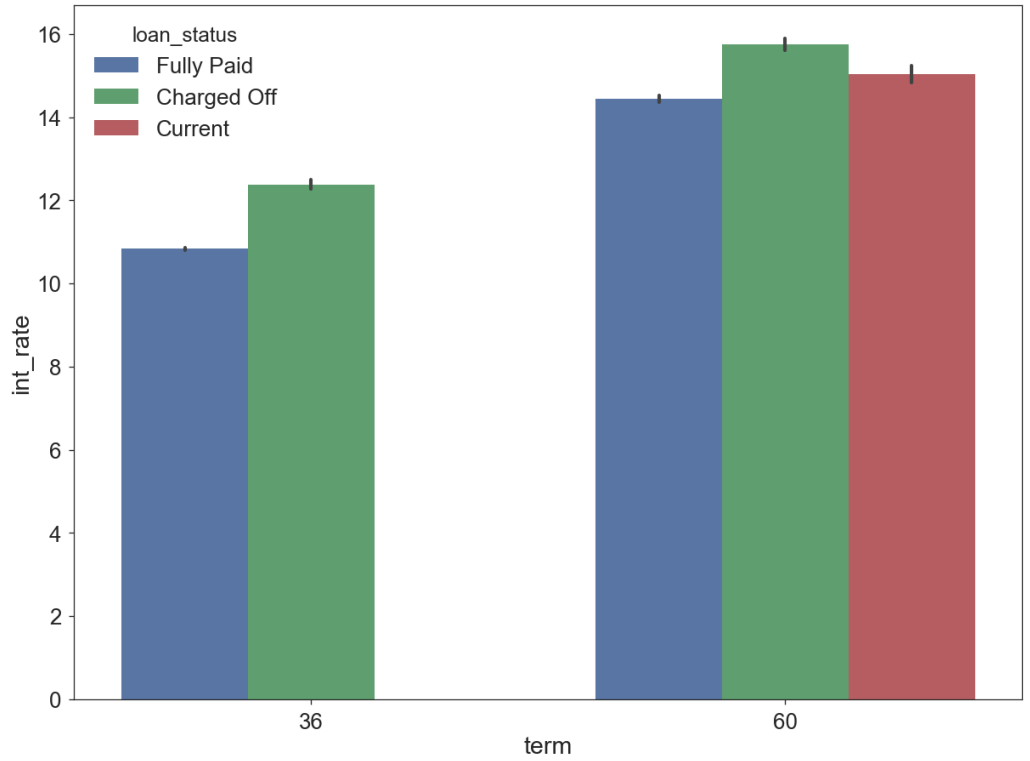


Analysis on Credit Card vs Emp_Length on Charged Off acc.

- Loan_status as Charged of for
- Loan_purpose Credit_card
- Based on the graph we can conclude
- 1.Years of experience in job for 1 or 10 years are more likely to default where as years of experience of 8 and 9 are less likely to default
- This was done with credit card with charged off status against emp_length

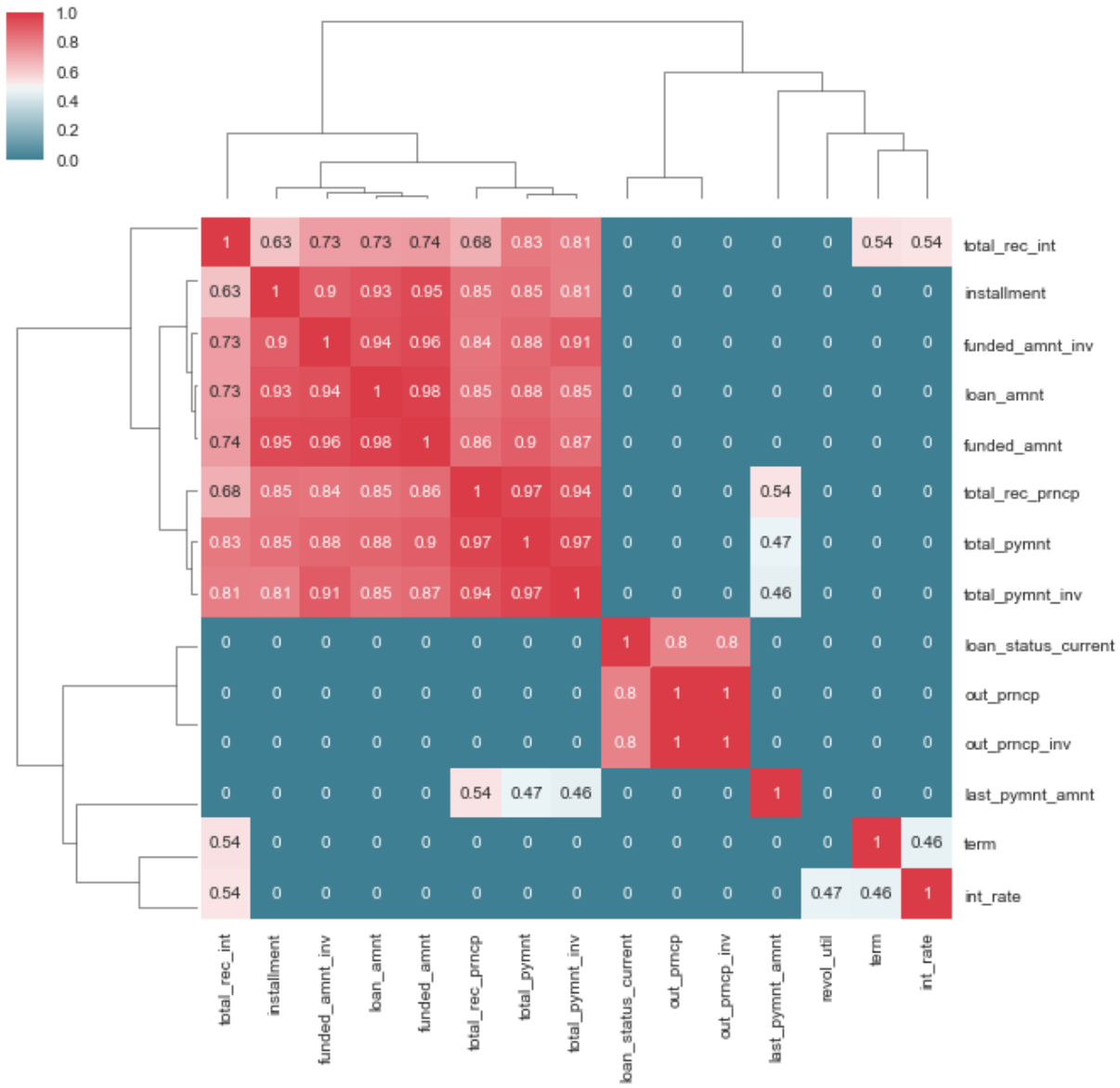


Analysis on Interest Rate vs Term



- Customers with maximum repayment term charging more interest rate from financial company

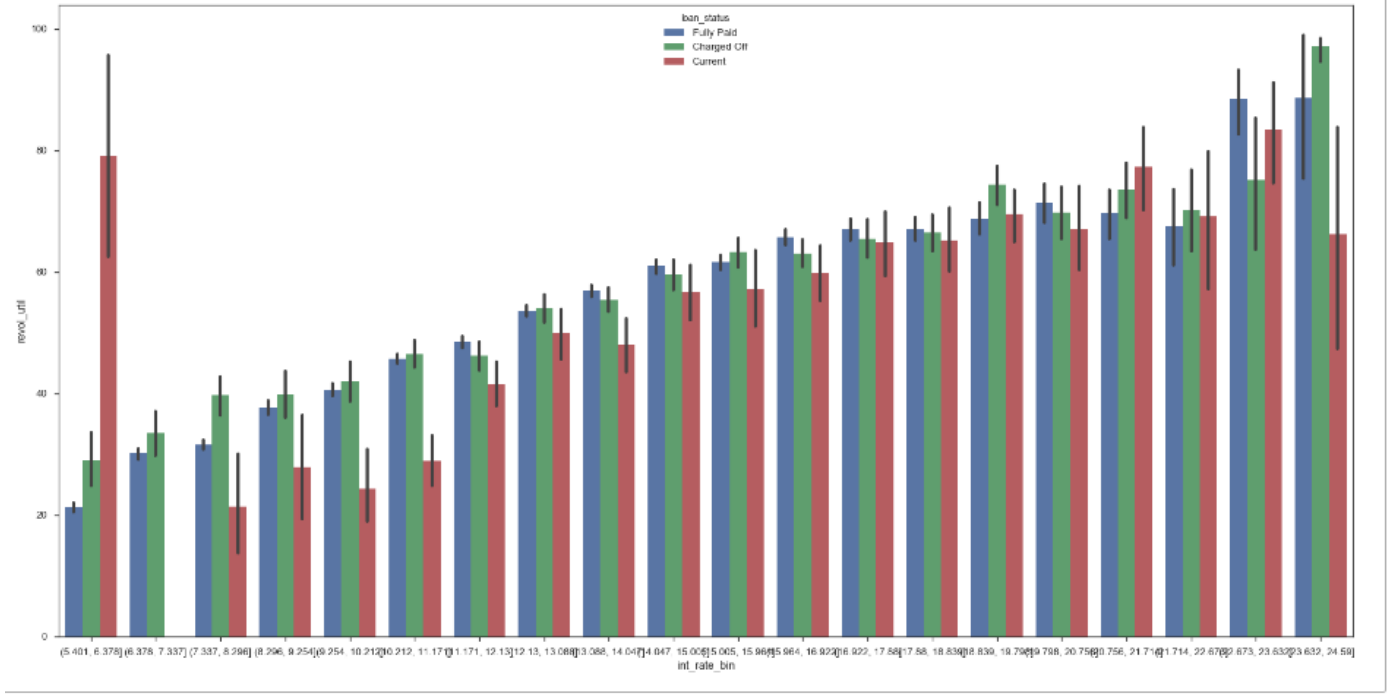
Analysis on correlation



- We have created correlation matrix to identify correlated variable
- Initially we got very large matrix of attribute
- It was reduced to few variables considering correl>0.45
- From this matrix we can see the correlation like
- Int_rate, revol_unit and term show good correlation
- Installments and total_rec_principal
- We can find many from here which we analysed in due course of investigation

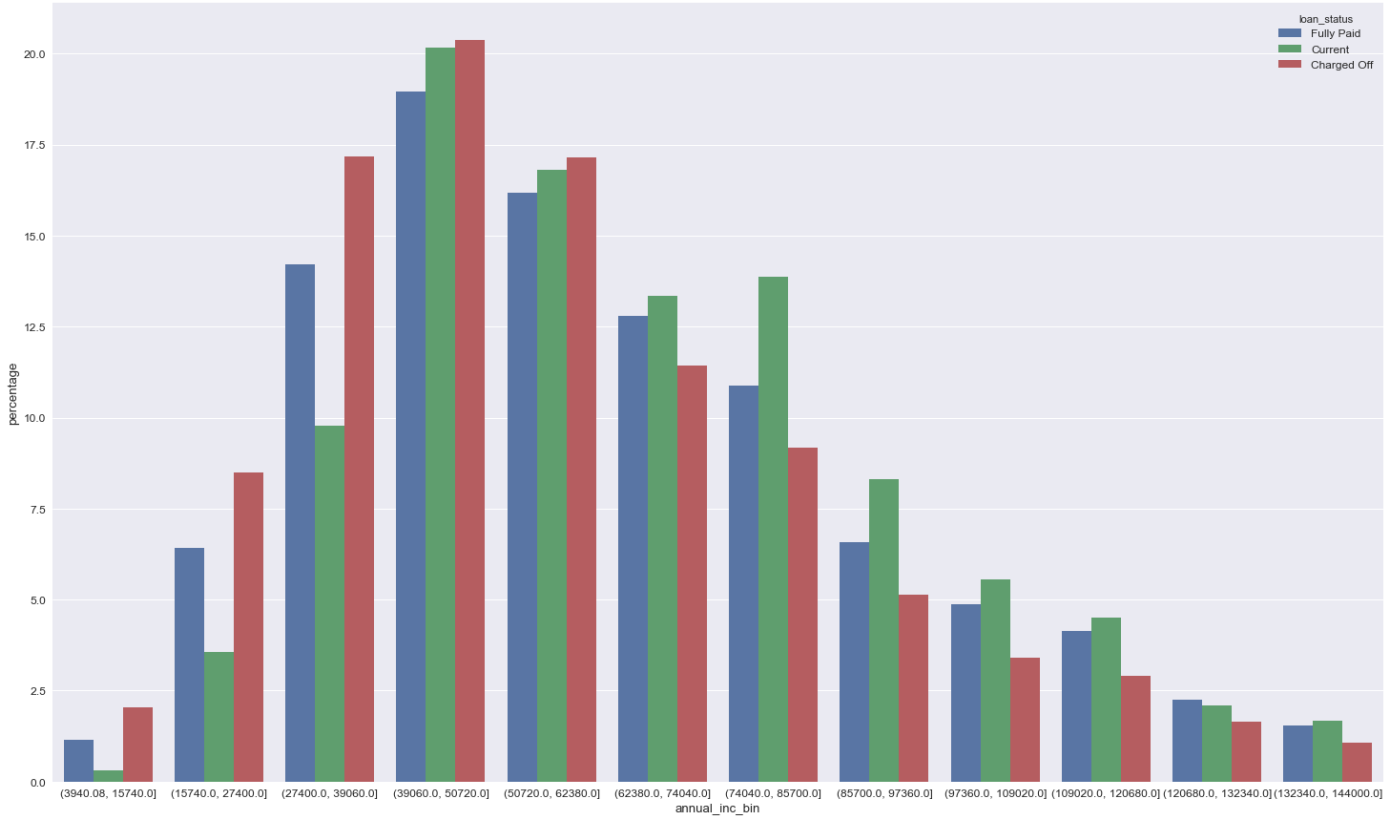
Analysis on Interest Rate vs revol_util

	revol_util	int_rate
count	39717.000000	39717.000000
mean	48.770677	12.021177
std	28.367689	3.724825
min	0.000000	5.420000
25%	25.300000	9.250000
50%	49.200000	11.860000
75%	72.300000	14.590000
max	99.900000	24.590000



- Customers with high Revolving utilization rate are charged with more interest rate from financial company
- Charged Off Accounts are maximum in trend in lower interest rate trends and decreases as interest rate increases
- The maximum current account was observed in the lowest interest rate category

multivariate on annual_inc and loan_status with grouped percentage

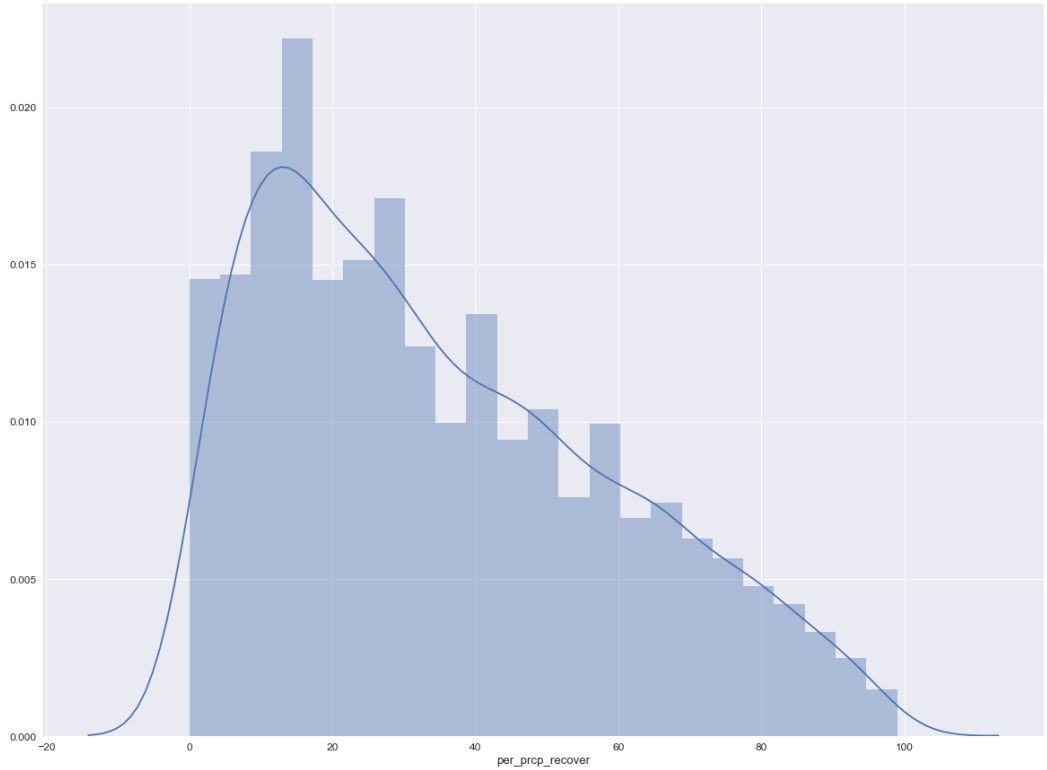


- From Above graph some of the highrisk Annual income group where we found maximum Defaults are (Approx Range rounded to near 1000s) 4000-70,000 where as in range 15-40,000 we have maximum default %

% Defaulter Vs Principal Recovered

We added one extra column per_prcp_recover which shows that how much principle amount finance company recover till date specially from defaulters

% Defaulters	% Part of Principal Recovered Already
19	60
11	70
6	80
2	90
1	95



Conclusions and Recommendations

Major Points

- Debt Consolidation is the Major Area followed by Credit Card where maximum number of loan application are there
- Higher the term higher the interest rate the higher probability of defaulting on loan
- Work experience group of 1 and 10 are found to be most defaulting group on loan in credit card as well as in debt_consolidation
- Some of driving variable for defaulters is **Address_state**, **Annual_income** , **Emp_length** means employee working experience and **revol_util** means higher credit borrowing against available credit revolving credit.

Recommendations

- We have Identified California or CA where maximum number of defaults were founds, we need to process loan applications stringently in these areas
- People with experience <1 and >=10 are tend to default more loans compare to other group, there application needed to be handled carefully
- Customers with high Revolving utilization rate needed to be charged with more interest rate from financial company considering default probability
- As we have seen people leaving on Rents are Defaulting more than those who leave in Mortage or own the home so we need to investigate more for this (rent) category before approving loans for them