Data Source:

File Name	loan.csv
# Rows	39717
# Columns	111

Data Preparation:

1. Reading file into Python environment

2. Basic EDA
   a. Shape of imported data
   b. Column names and it's type
   c. Checking interquartile range for numeric variables

3. Creating derived variables

Existing	Derived	Purpose
Term	**term_int**	**to bring month in number (e.g.: From 36 month to 36**
int_rate	***int_rate_%***	***to remove "%" sign (e.g.: 10.65% to 10.65)***
revol_util	***revol_util_%***	***to remove "%" sign (e.g.: 10.65% to 10.65)***
Loan_amnt - funded_amnt_inv	***Diff_Amount***	to find any pattern for Diff_Amount
loan_status	loan_status_Target	***Converting Categorical Dependent Variable to Nominal data (to check correlation with independent variable)***

4. Removing few irrelevant variables as it would not be helpful for analysis

id', 'member_id', 'emp_title', 'pymnt_plan', 'url', 'desc', 'title', 'zip_code', 'initial_list_status', 'policy_code', 'application_type', 'acc_now_delinq', 'delinq_amnt', 'chargeoff_within_12_mths', 'tax_liens', 'last_pymnt_d', 'last_credit_pull_d', 'collections_12_mths_ex_med', 'delinq_amnt'

5. Checking for Missing Data and Missing value Imputation

(a) Those columns were removed where missing%>90%

mths_since_last_record , open_rv_24m , max_bal_bc , all_util , total_rev_hi_lim , inq_fi , total_cu_tl , inq_last_12m , acc_open_past_24mths , avg_cur_bal , bc_open_to_buy , bc_util , mo_sin_old_il_acct , mo_sin_old_rev_tl_op , mo_sin_rcnt_rev_tl_op , mo_sin_rcnt_tl , mort_acc , mths_since_recent_bc , mths_since_recent_bc_dlq , mths_since_recent_inq , mths_since_recent_revol_delinq , num_accts_ever_120_pd , num_actv_bc_tl , num_actv_rev_tl , num_bc_sats , num_bc_tl , num_il_tl , num_op_rev_tl , num_rev_accts , num_rev_tl_bal_gt_0 , num_sats , num_tl_120dpd_2m , num_tl_30dpd , num_tl_90g_dpd_24m , num_tl_op_past_12m , pct_tl_nvr_dlq , percent_bc_gt_75 , tot_hi_cred_lim , total_bal_ex_mort , total_bc_limit , total_il_high_credit_limit

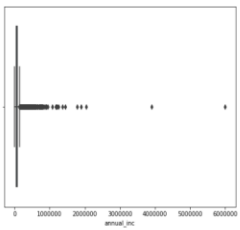
(b) Missing value Imputation

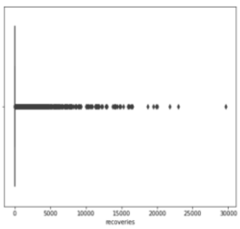
Var	# Total Missing	Imputation Logic
`emp_length`	1075	By looking at above analysis output, we can somehow conclude that the missing emp_length should fall into either '<1 year' or may be even more smaller. Eg.: It could be in months (say <6 months) or could be intern As we don't have any clear anaswer, so **Imputing mising values as '*<u>Not Mentioned</u>*'**
`mths_since_last_delinq`	25682	Imputing mean value for each loan_status against missing values of mth_since_last_deling for each loan_status
`pub_rec_bankruptcies`	697	imputuing missing value with **mode** of entire data
`revol_util_%`	50	imputuing missing value with **median** of entire data

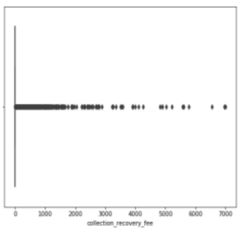
6. Checking for Outlier and Outlier Treatment

(a) Checking for Outlier using skew function

(b) Also used boxplot to check the nature of outlier (Right/Left Skewed)

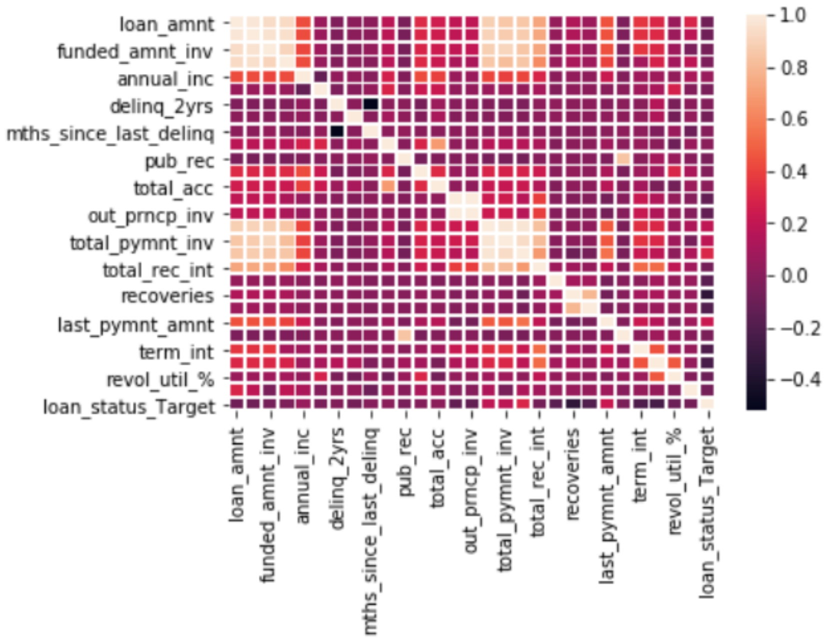






Var	Skew value and Type	Outlier Treatment
annual_inc	30 - Right skewed	IQR and Upper limit is calculated and all value above upper limit (`14,5144`) is capped.
recoveries	16 - Right skewed	Since there are many outliers in each Grades so we will not be going to remove these outliers. Also most of the values are '0' due to this IQR and upper limit is coming as '0'(Zero)
collection_recovery_fee	25 - Right skewed	

7. Correlation

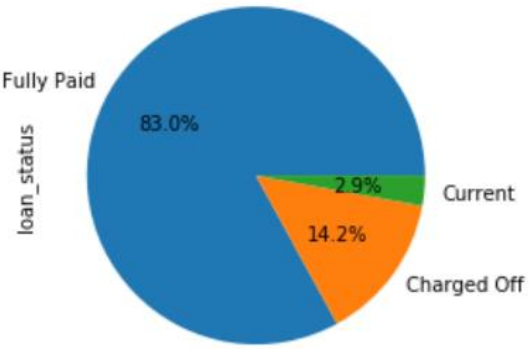


List of correlated columns

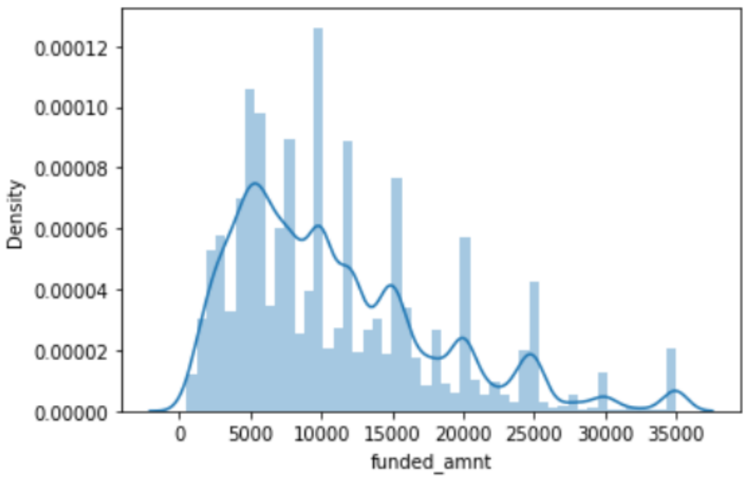
Var Name	Outlier Treatment
loan_amount	Thease variables are very correlated with 'funded_amnt'
funded_amnt_inv	
collection_recovery_fee	Highly correlated with 'recoveries'
out_prncp_inv	Highly correlated with 'out_prncp'
total_pymnt_inv	Highly correlated with 'total_pymnt'

Univariate Analysis:

1. **Loan Status**
   82.9% of the total population has Loan status as '**Fully Paid**'.

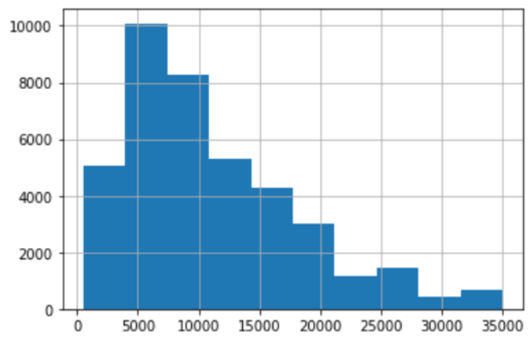


2. **Funded Amount** ('*funded_amnt*)'
   (a) Data Distribution: Data is 'Right Skewed'



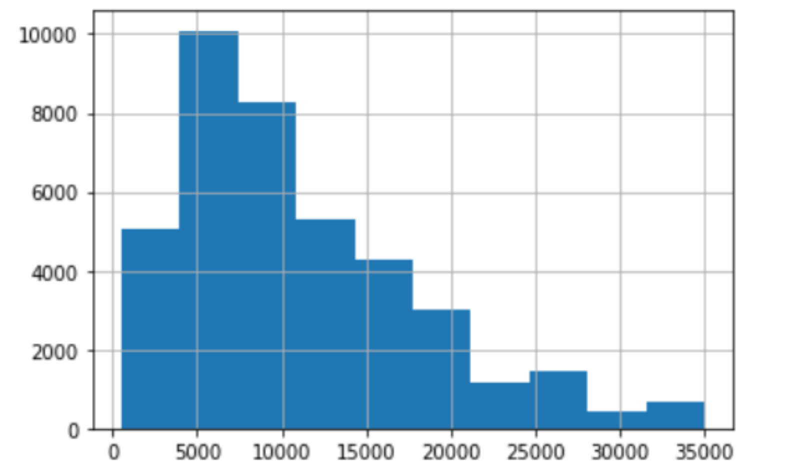
Blue line shows Kernal Density Estimation (KDE) which shows probability of certain value.

(b) Frequency Distribution: Maximum frequency of Loan funded amount is ~ 8k

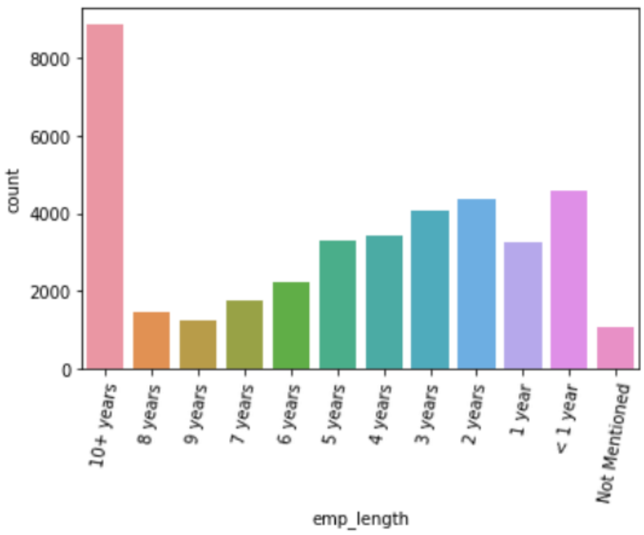


3. **Interest Rate**('int_rate_%')
    - Min Interest Rate is 5% and max is 25% .
    - 75% of population are paying >= ~14.5% Rate of interest
    - There are few people who pay high rate of interest (>22.55, these are considered to be outlier)

```
count    39717.000000
mean        12.021177
std          3.724825
min          5.420000
25%          9.250000
50%         11.860000
75%         14.590000
max         24.590000
Name: int_rate_%, dtype: float64
```



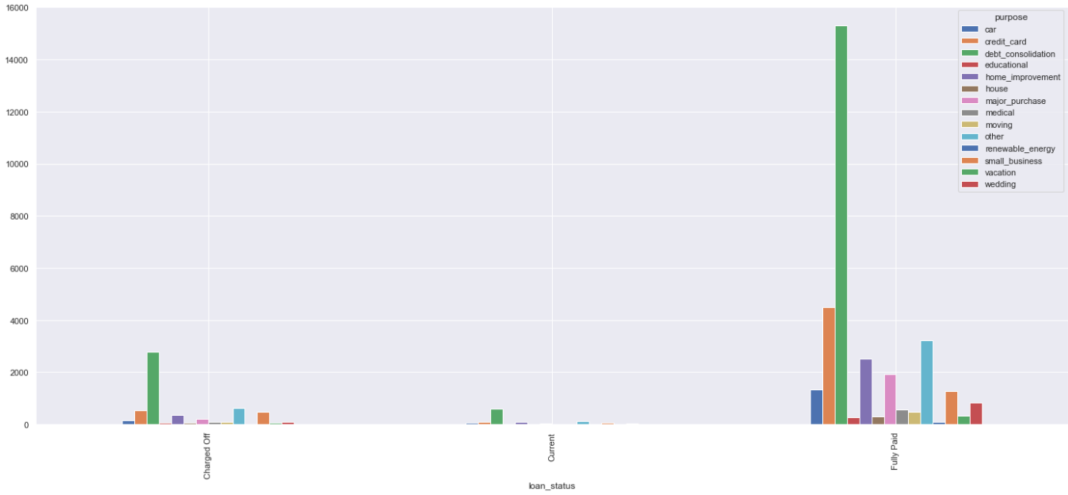
4. **Emp Length**
    Maximum applicant in given sample has Experience greater than 10 years.



Bivariate and Multivariate Analysis:

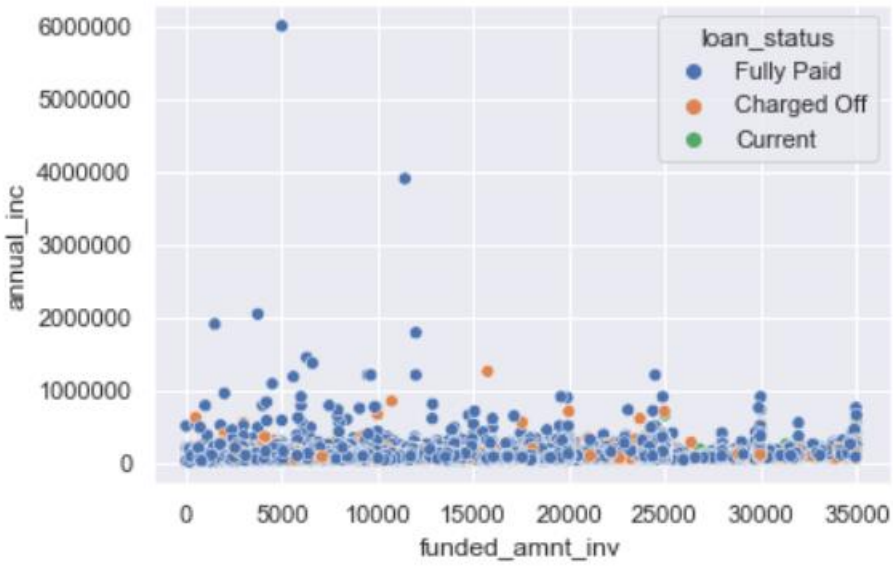
5. **Loan Status Vs Purpose**

   For fully paid customers maximum number od loan is for debt_consolidation

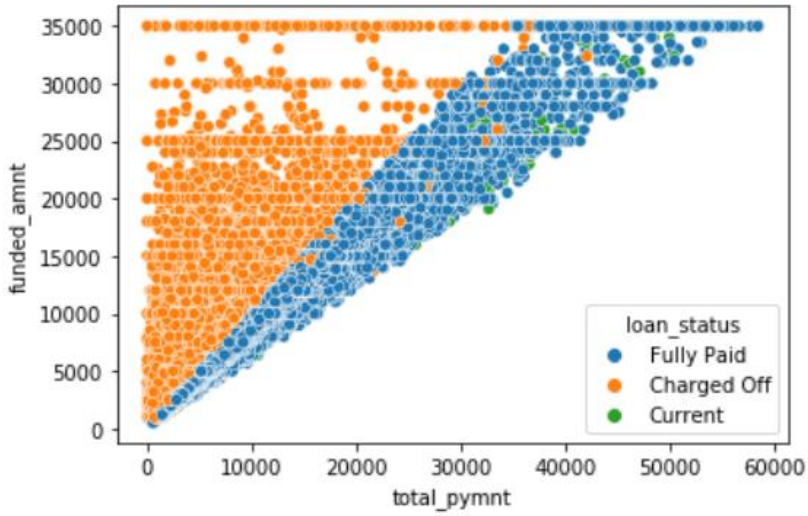


6. **Annual Income Vs Loan Funded Amount Vs Loan Status**

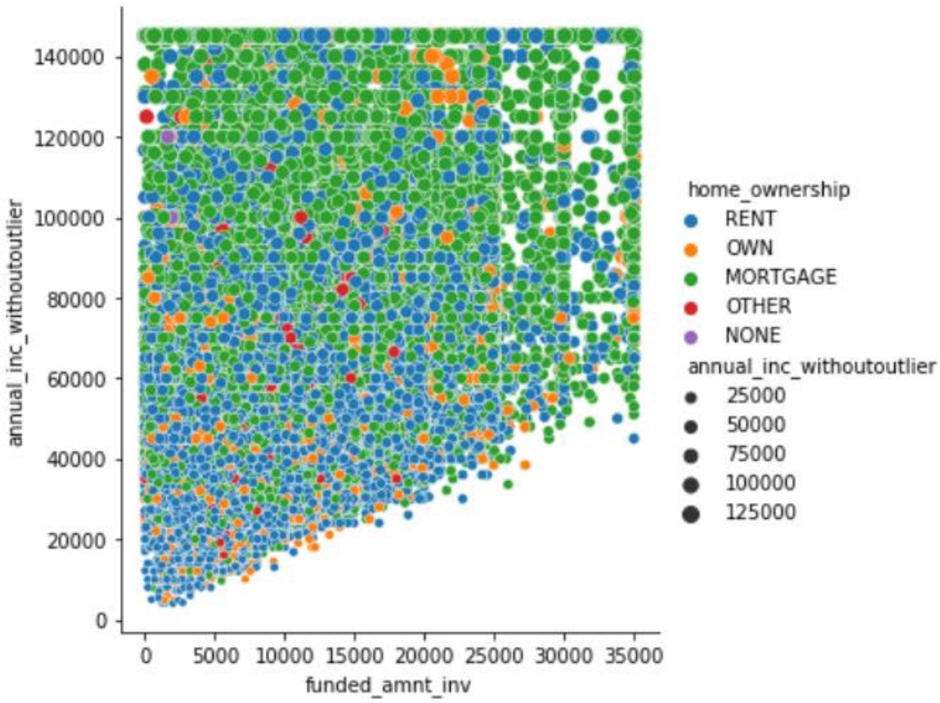
   Mostly people income < 1 million are more tend to take loan. There are only few applicants whose yearly income > 1 million



7. **Annual Income Vs Loan Funded Amount Vs Loan Status**
   There is a clear separation in Loan Status (Defaulter Vs Non_Defaulter). for variable

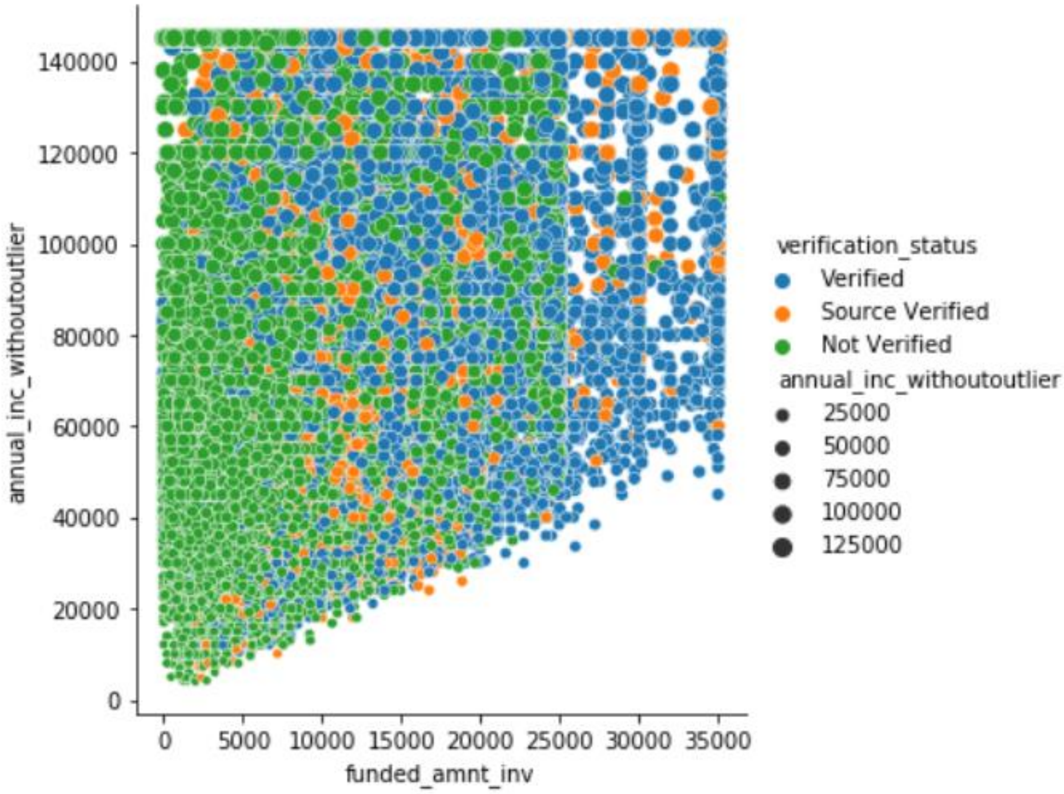


8. **Annual Income Vs Funded Amount Vs Home Ownership**
   Applicant who having high annual income are mostly have home ownership as '*Mortgage*'



9. Annual Income Vs Funded Amount Vs Verification_Status

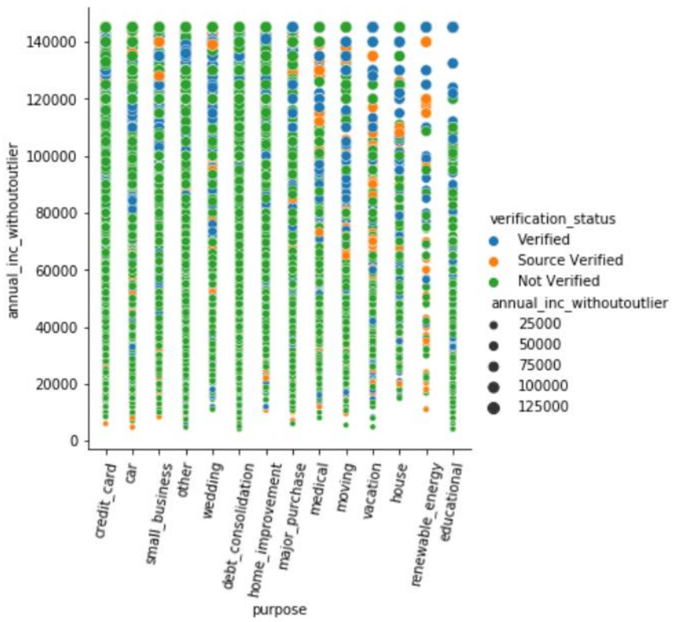
Non verified applicant mostly gets lower funded amount and vice versa for Verified applicant



10. Annual Income Vs Purpose Vs Verification_Status

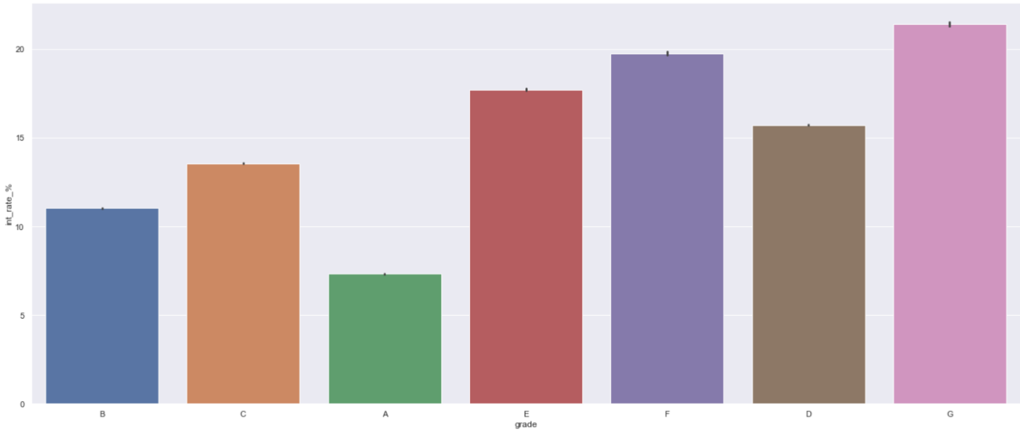
Most Non-Verified applicants are those customers who took loan for making Credit Card payments.

Where as most Verified customer are those whose purpose of loan is “moving”



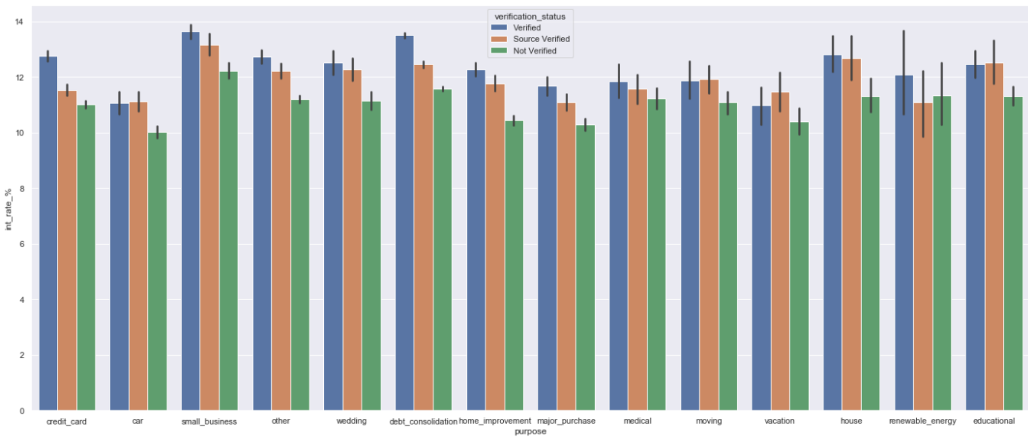
11. Grade Vs Interest Rate

Grade 'G' pays highest Average Rate of Interest where as the lowest Rate of interest is for Grade 'A'



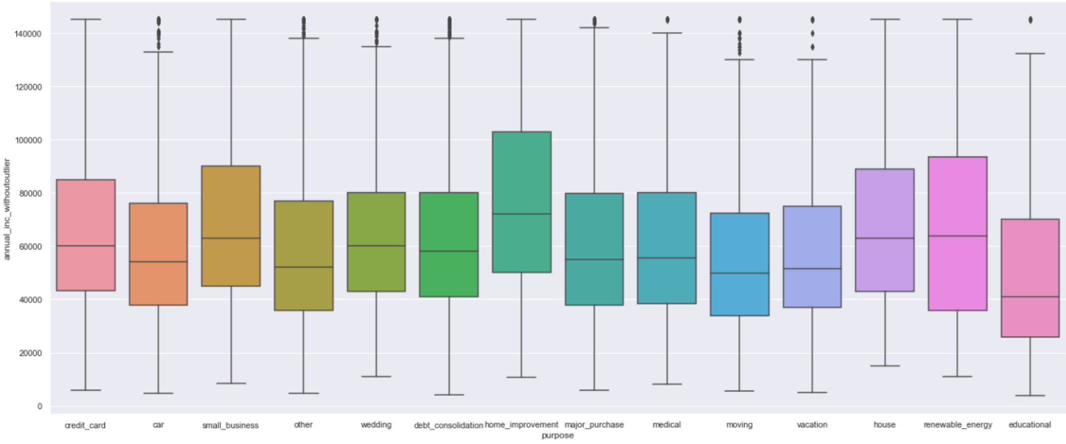
12. Purpose Vs Interest Rate Vs Verification status

Verified customer pay higher Rate of Interest compare to non-verified customers.
And Rate of Interest is independent of Purpose of Loan



13. Purpose Vs Annual Income

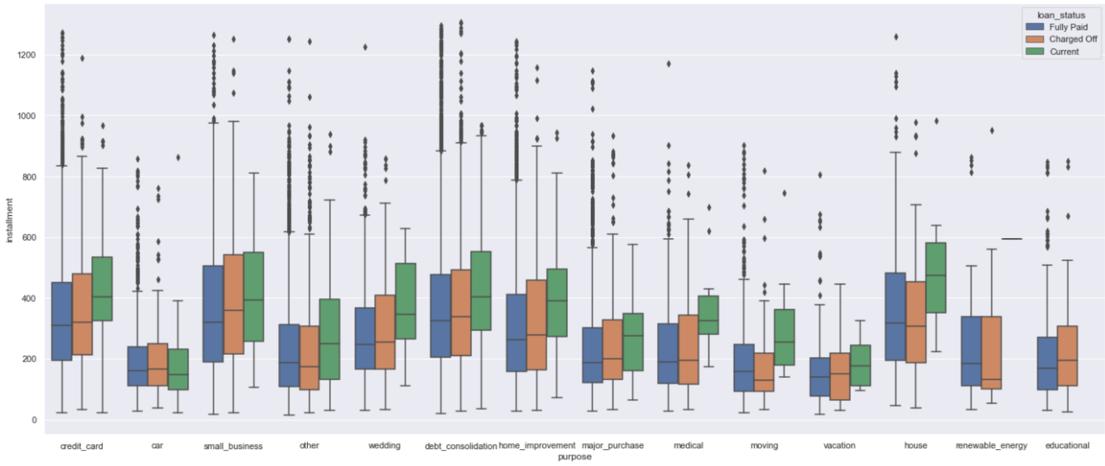
Customers having higher average income mostly take lone for ‘home improvement’ purpose



14. Purpose Vs Installment Vs Loan Status

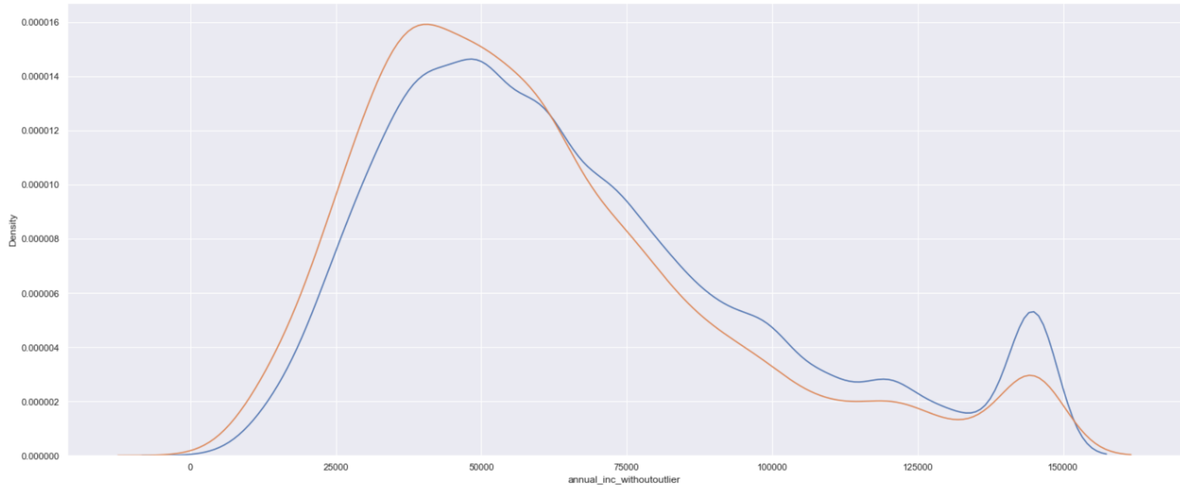
Average loan installments of each purpose are almost same for Defaulter and non-Defaulter except purpose = ‘*renewal energy*’

Average installments for Loan Status=’charged off’ is higher than ‘Fully Paid’ for all purpose of loan except ‘house’ and ‘moving’

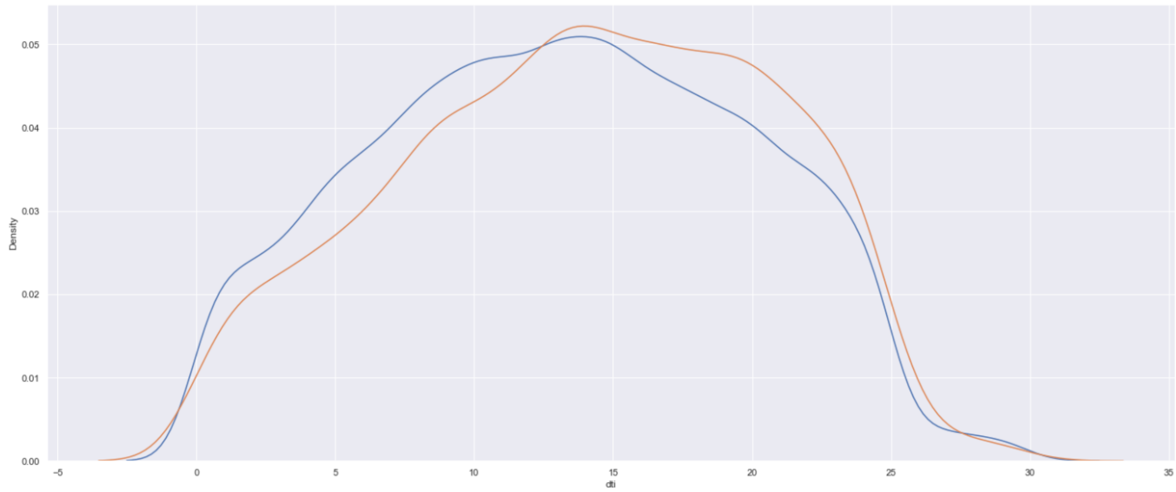


Recommendation:

1. If Annual Income is between **25k to 50k** then chance of '***Charged off***' is high whereas chance of '***Fully Paid***' is high at salary **~14 million** (1400000).

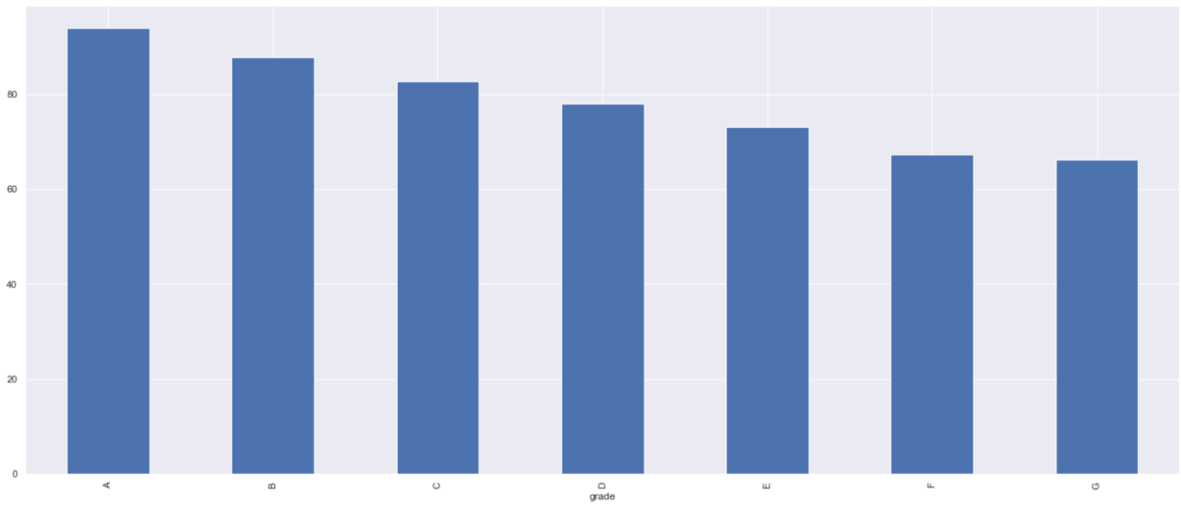


2. If DTI of the customer is between 0 to 12.5 then chance of Loan status= 'Fully Paid' is significantly high compare to if DTI is between 12.5 to 30



3. Loan status has strong relationship with grades of customer. ~94% of Grade A customer has paid the loan (*i.e.*, Loan Status='***Fully Paid***')

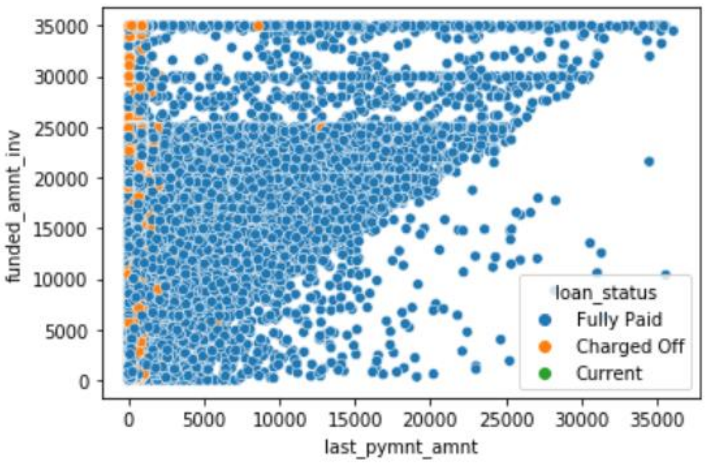
```
grade
A    94.006969
B    87.794433
C    82.805719
D    78.013766
E    73.150582
F    67.315574
G    66.220736
```



4. The lowest % for 'Fully Paid' customer are those whose emp_length is not mentioned.

```
emp_length
1 year            85.610603
10+ years         84.319039
2 years           86.786297
3 years           86.166500
4 years           86.175943
5 years           85.660614
6 years           85.839483
7 years           84.628872
8 years           85.853659
9 years           87.112561
< 1 year          85.825200
Not Mentioned     77.928364
Name: loan_status_New, dtype: float64
```

5. Applicant who paid less (~ <2000) 'Last Payment Amount' are more likely to be Defaulter.



6. Verified customers who still defaults to pay loan are those customers whose purpose of loan is ‘*education*’
   Maximum number of defaulter are those whose verification_status is ‘*Source Verified*’ and Purpose of loan are either ‘*small business*’, ‘*other*’ or ‘*moving*’