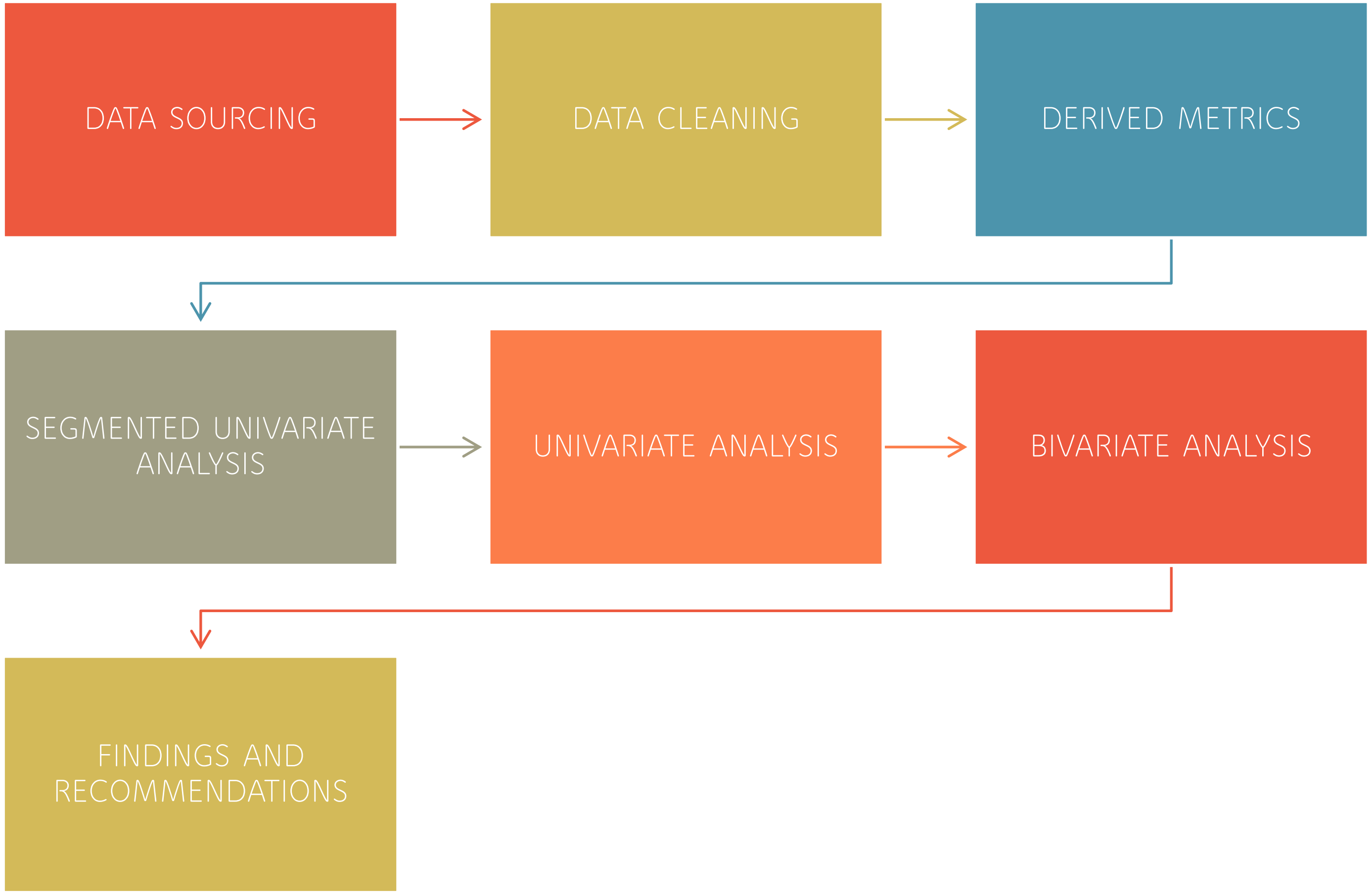
Lending Club Case Study

EDA Analysis

SIVA MUPPALLA

NAMRATA SHIVTARKAR

Road Map for Methodology



DATA SOURCING

DATA CLEANING

DERIVED METRICS

SEGMENTED UNIVARIATE ANALYSIS

UNIVARIATE ANALYSIS

BIVARIATE ANALYSIS

FINDINGS AND RECOMMENDATIONS

DATA SOURCING

- Loan.csv file provided by upgrad is used for data analysis
- Each variables mentioned in Data dictionary excel is used for better understanding provided by upgrad
- Internet is used for further more understanding of a variable

DATA CLEANING

- Checking for duplication in ‘id’ column. Not found any duplicates
- Finding % of missing values and null
- Check columns having more than 70% missing
- Standardizing the data
- Deleting columns if missing and Null values grater than 50%
- Dropping columns where only one value presented because It don't give any explanation

DATA CLEANING

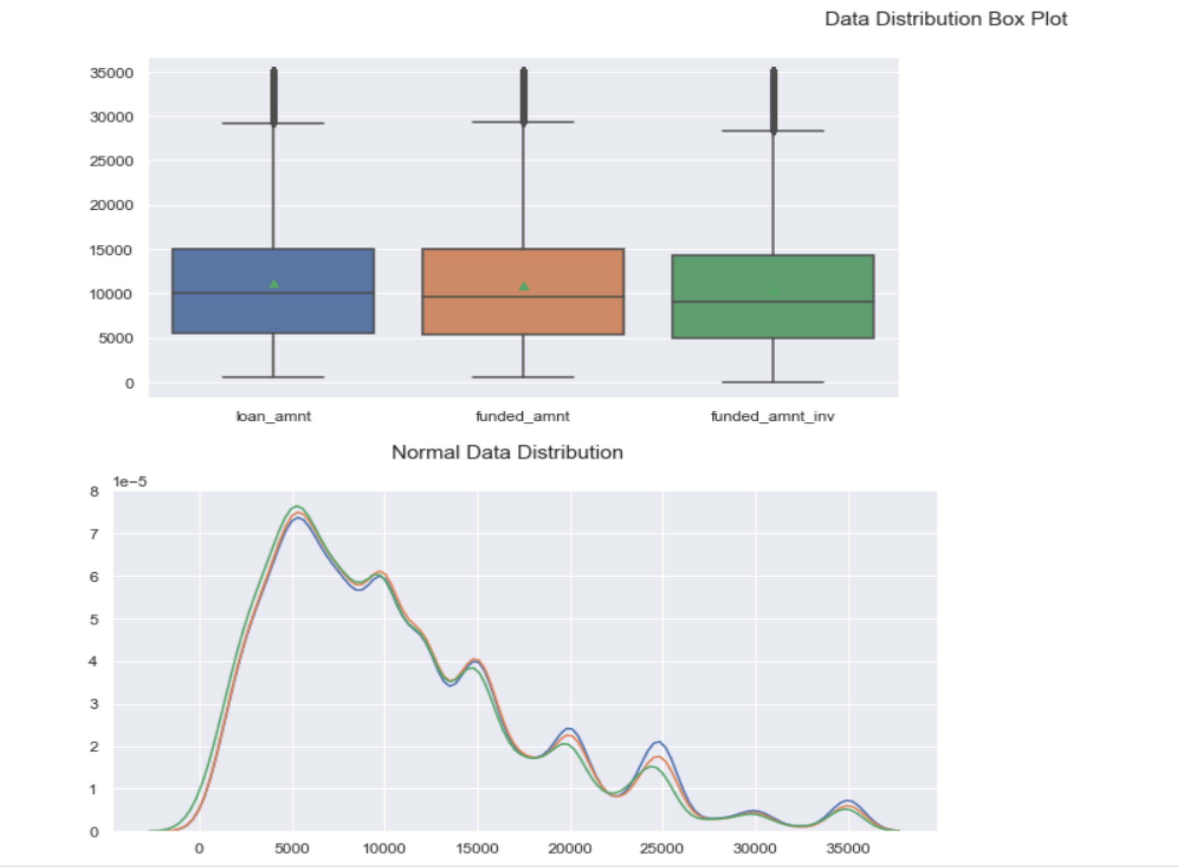
- Removing ‘Current’ loan status rows, not required for default/non default analysis
- Deleting columns like title, zip code, address, url which are not needed for loan default analysis
- Standardizing the data
  - Removing % symbol, removing + and < symbol from employee years format
  - Removing string ‘months’ from term_months column
  - Converting columns into integer
  - Converting DataTime type of column ‘issue_date’ and ‘last_pymnt_date’
- Renaming Column names for better understanding
  - Term -> term_months, interest_rate -> interest_rate_percentage, emp_length -> emp_length_year

DERIVED METRICS

- Derived new ‘total_mnths_paid’ column by Subtracting issue_d from last_pymnt_d
- Derived new ‘approved_loan_amnt_ratio’ column by taking ratio of ‘funded_amnt_inv’ and ‘loan_amnt’

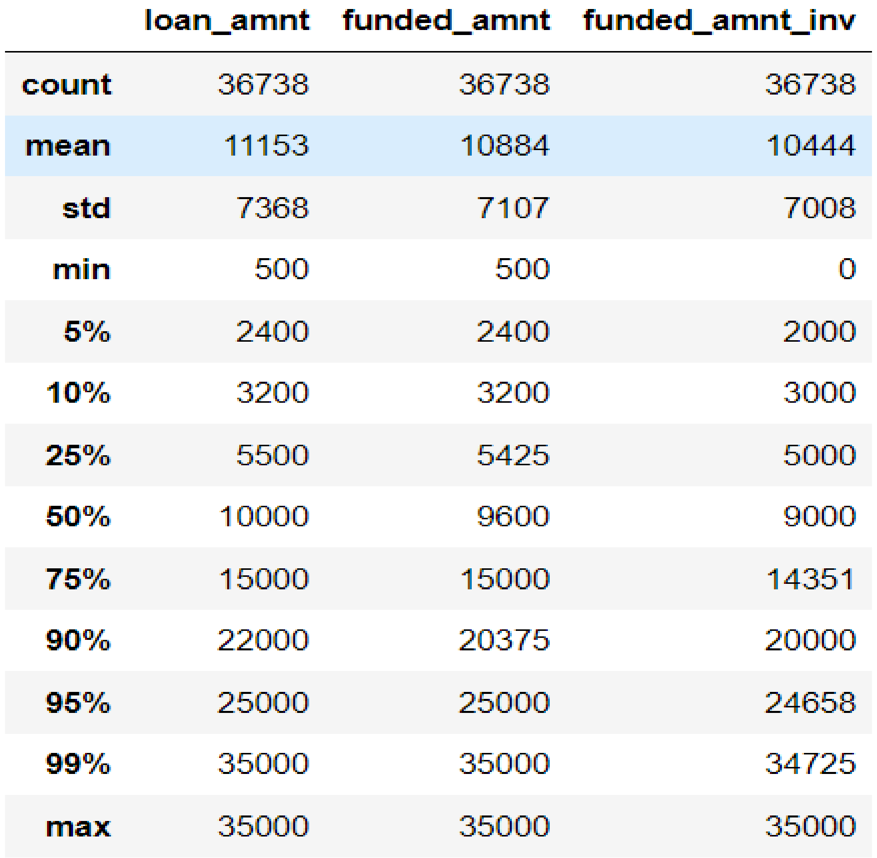
OUT LIER CHECKING

Box plot shows that funded_amnt_inv has outlier presented in the data, we need make the normalization data removing outliers so that mean and median will come inline with all loan amount.



OUTLIER CHECKING

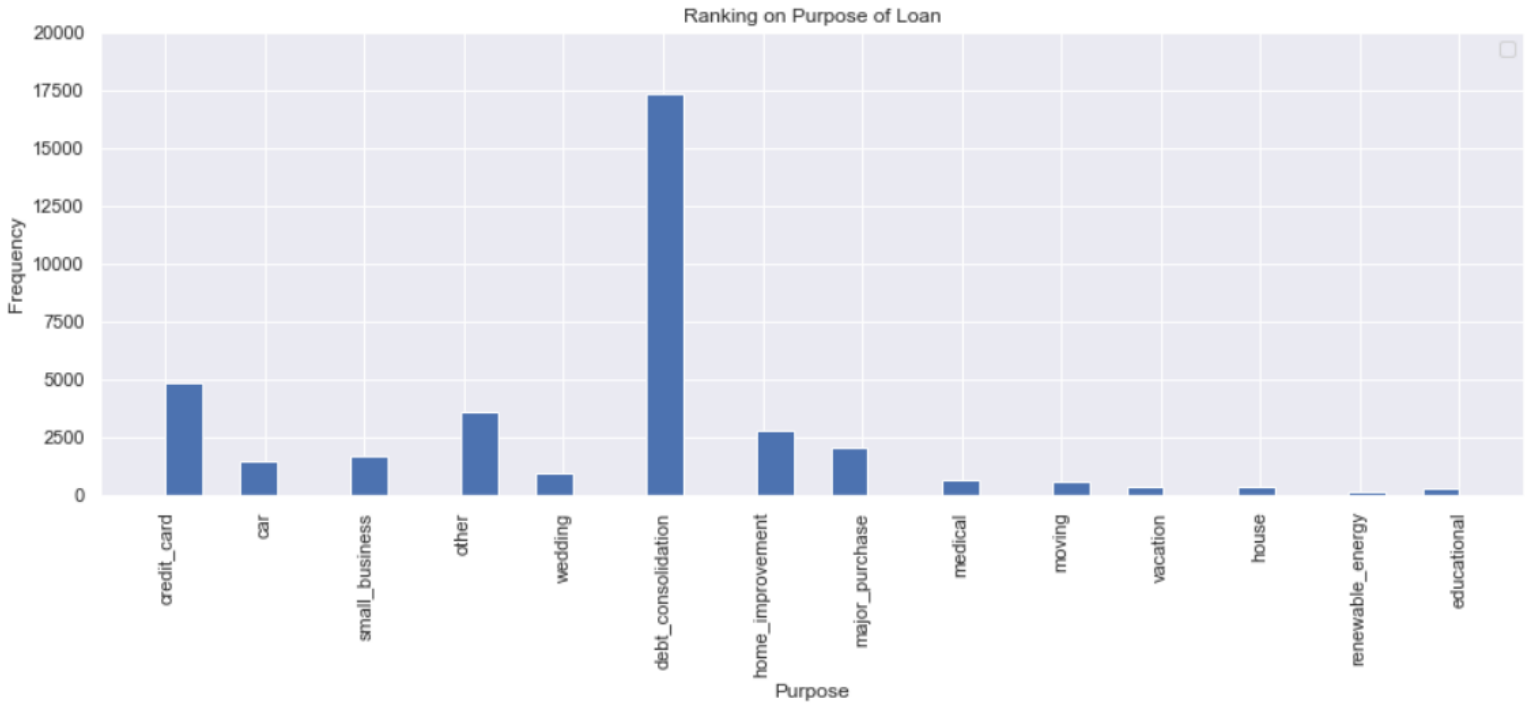
Descriptive Statistics Table shows that funded amount investment has zero values presented in data. Due this it will cause outlier presented in Funded amount investment.



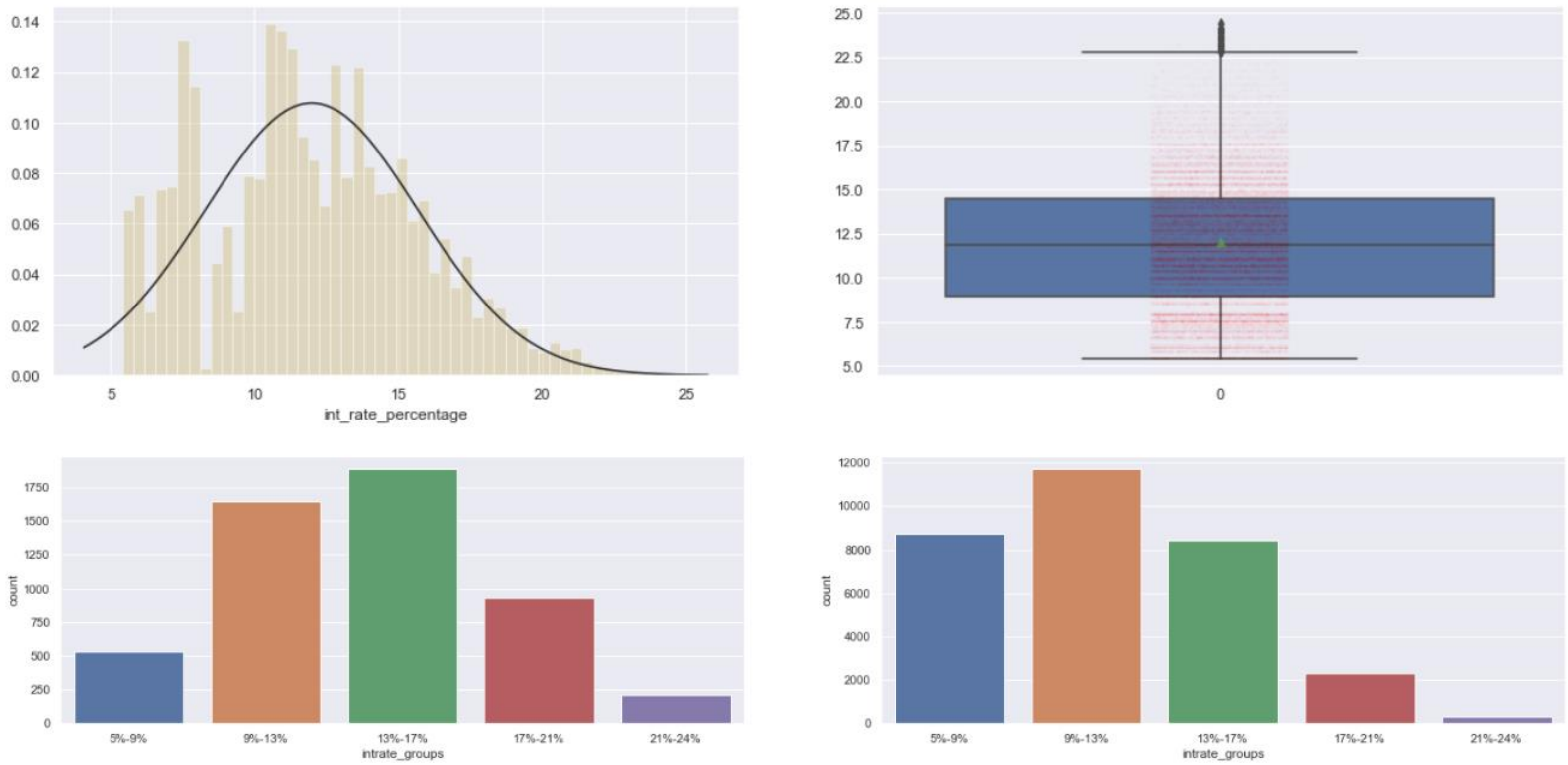
	loan_amnt	funded_amnt	funded_amnt_inv
count	36738	36738	36738
mean	11153	10884	10444
std	7368	7107	7008
min	500	500	0
5%	2400	2400	2000
10%	3200	3200	3000
25%	5500	5425	5000
50%	10000	9600	9000
75%	15000	15000	14351
90%	22000	20375	20000
95%	25000	25000	24658
99%	35000	35000	34725
max	35000	35000	35000

UNIVARIATE ANALYSIS

Debt_Consolidation has more applicants, which shows high risk in many other variables like late fees/recoveries/collections/delinquency , we can reduce the risk by checking past records of the applicant



DISTRIBUTED GRAPH FOR INTEREST RATES FOR FULLY PAID AND CHARGED OFF CUSTOMER



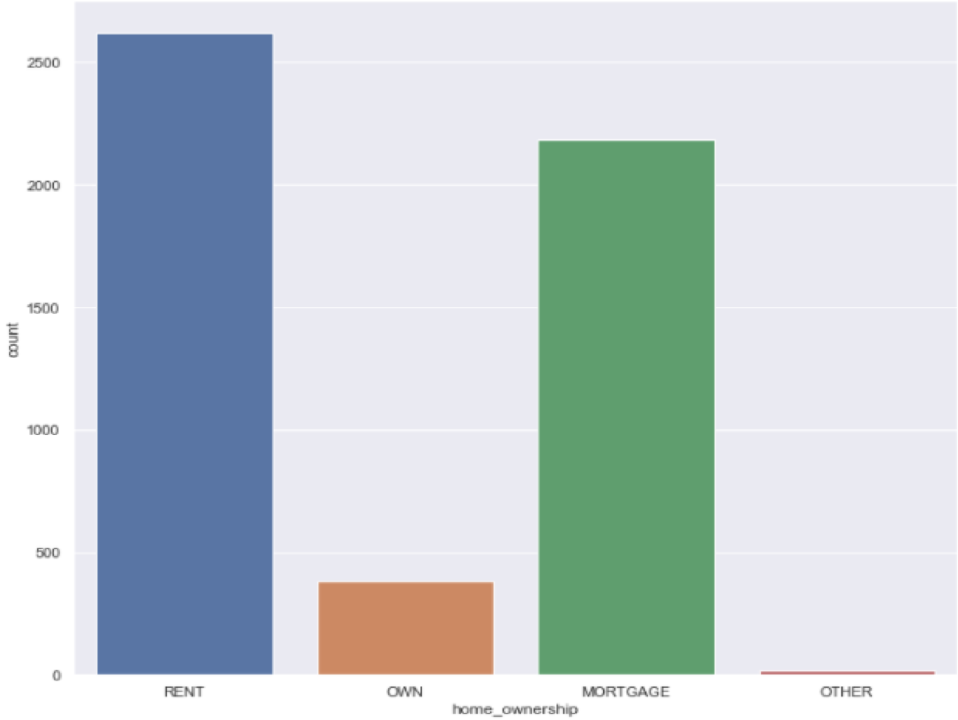
DISTRIBUTED GRAPH FOR INTEREST RATES FOR FULLY PAID AND CHARGED OFF CUSTOMER

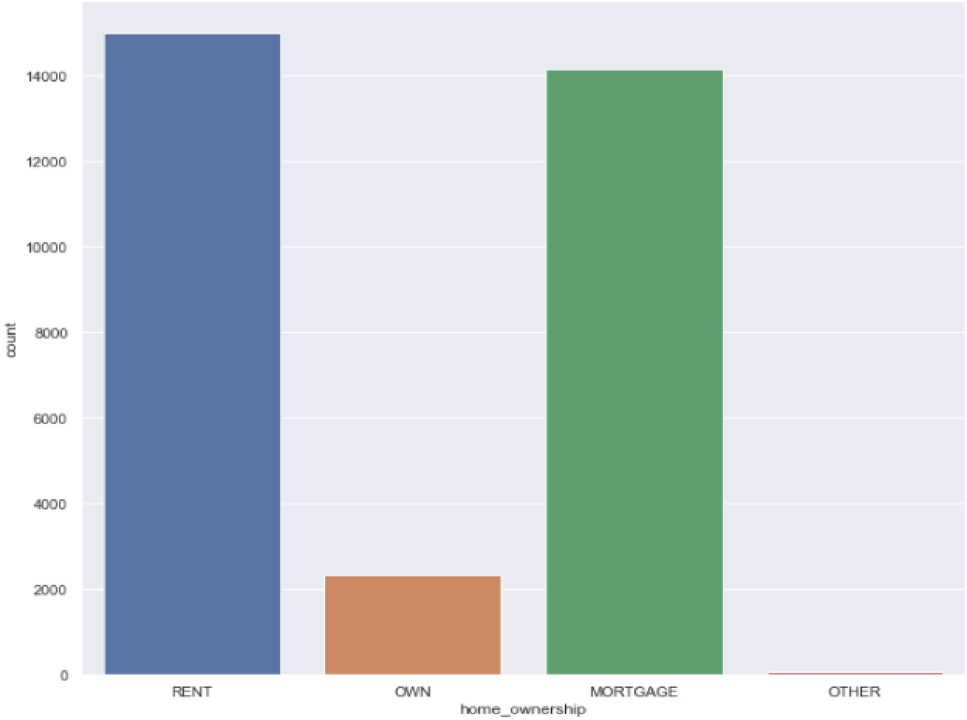
Observation :

1.In the above graph of interest rate distribution, you can see that the highest rate is between 10-15% for Fully Paid and Charged Off.

2.Also, the two graphs below show a comparison of fully paid and charged Off applicants, the 'charged off' category has a large number of people with interest rates ranging from 13 to 17 percent, while the 'Fully paid' category has an interest rate ranging from 9 to 13 percent

HOME OWNER SHIP FOR CUSTOMER

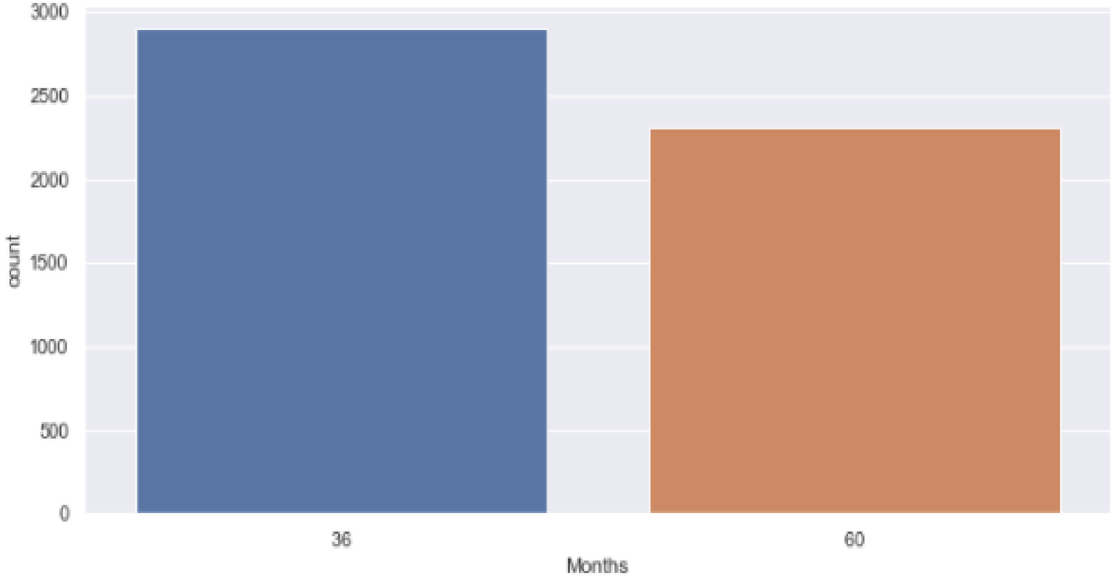


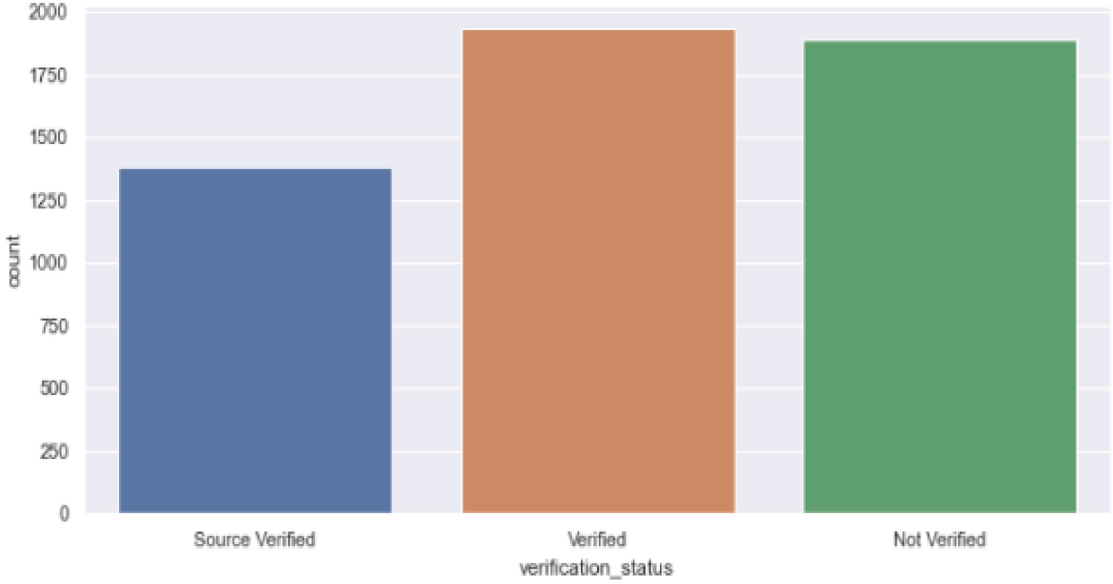


Observation:

- Ratio between Rent and Mortage of Charged off customers are higher than Fully Paid
- Rented Charged off customers are more than own house customers

DISTRIBUTION OF GRAPH FOR VERIFICATION STATUS FOR CHARGED OFF CUSTOMER



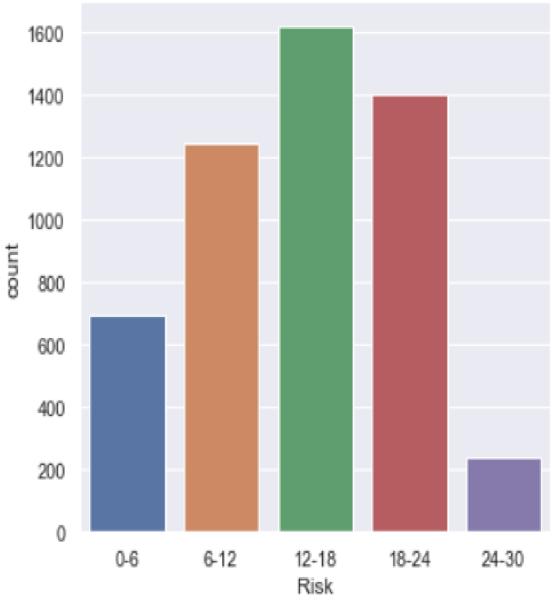


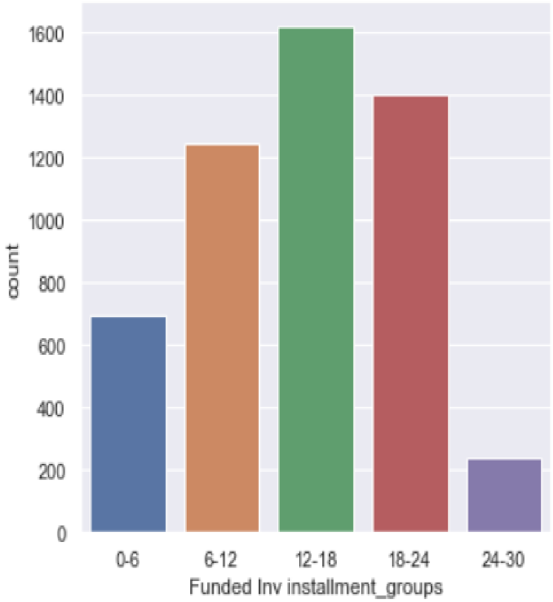
Observation:

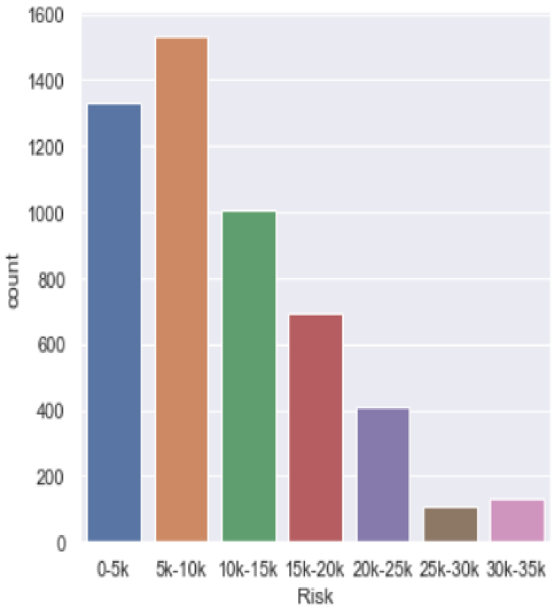
- 'Charged Off' customers with less term period are high as they want to finish off the loan soon.
- 'Charged Off' customers with non-verified income source is very high, Lender should verify source of income and should reduce the risk, it will reduce the loss

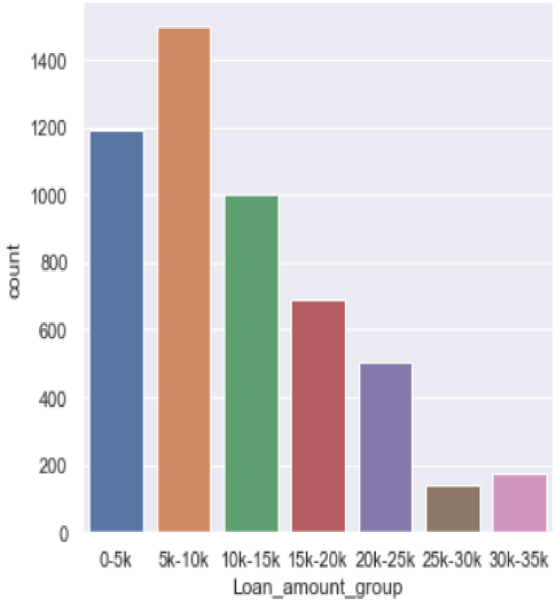
SEGMENTED UNIVARIATE ANALYSIS

DTI is one of the imp risk variable, higher the dti high chances to not pay back or late payments

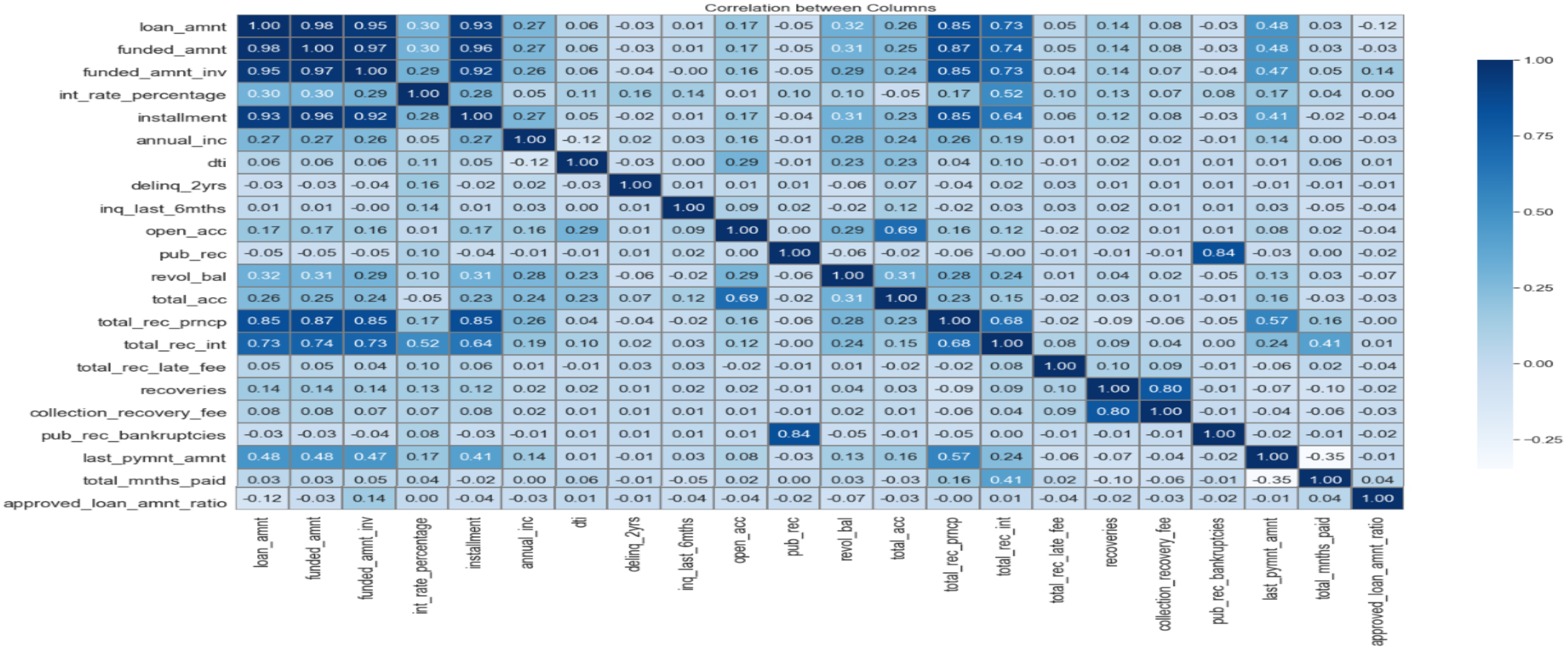








BIVARIATE ANALYSIS

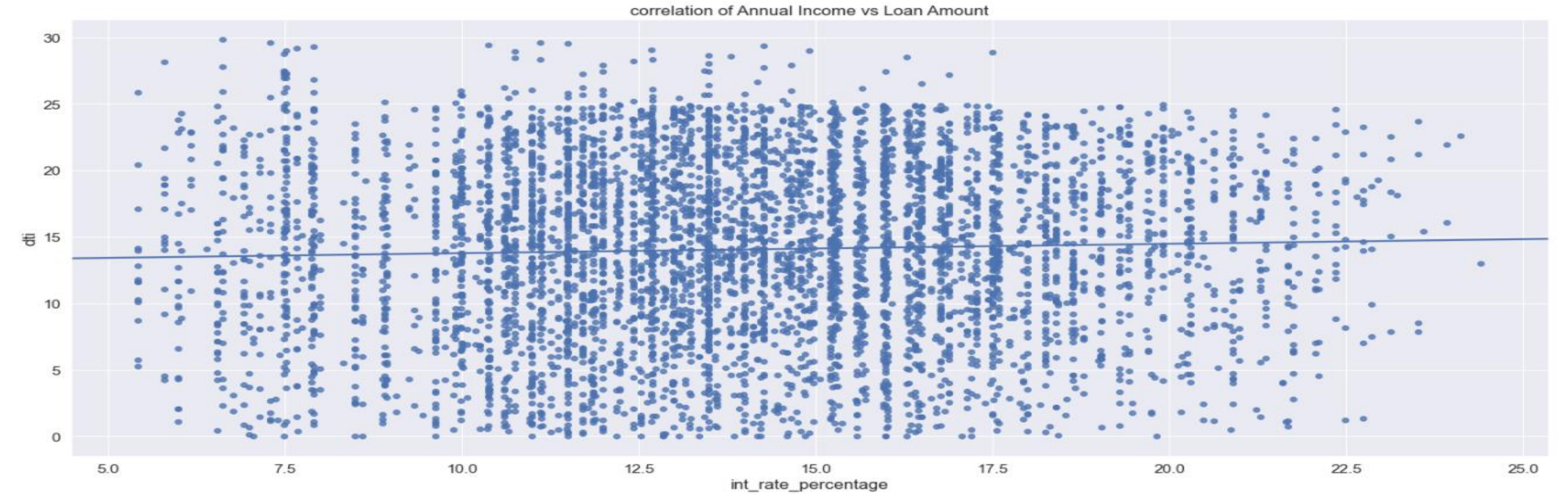


	loan_amnt	funded_amnt	funded_amnt_inv	int_rate_percentage	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	total_acc	total_rec_prncp	total_rec_int	total_rec_late_fee	recoveries	collection_recovery_fee	pub_rec_bankruptcies	last_pymnt_amnt	total_mnths_paid	approved_loan_amnt_ratio
loan_amnt	1.00	0.98	0.95	0.30	0.93	0.27	0.06	-0.03	0.01	0.17	-0.05	0.32	0.26	0.85	0.73	0.05	0.14	0.08	-0.03	0.48	0.03	-0.12
funded_amnt	0.98	1.00	0.97	0.30	0.96	0.27	0.06	-0.03	0.01	0.17	-0.05	0.31	0.25	0.87	0.74	0.05	0.14	0.08	-0.03	0.48	0.03	-0.03
funded_amnt_inv	0.95	0.97	1.00	0.29	0.92	0.26	0.06	-0.04	-0.00	0.16	-0.05	0.29	0.24	0.85	0.73	0.04	0.14	0.07	-0.04	0.47	0.05	0.14
int_rate_percentage	0.30	0.30	0.29	1.00	0.28	0.05	0.11	0.16	0.14	0.01	0.10	0.10	-0.05	0.17	0.52	0.10	0.13	0.07	0.08	0.17	0.04	0.00
installment	0.93	0.96	0.92	0.28	1.00	0.27	0.05	-0.02	0.01	0.17	-0.04	0.31	0.23	0.85	0.64	0.06	0.12	0.08	-0.03	0.41	-0.02	-0.04
annual_inc	0.27	0.27	0.26	0.05	0.27	1.00	-0.12	0.02	0.03	0.16	-0.01	0.28	0.24	0.26	0.19	0.01	0.02	0.02	-0.01	0.14	0.00	-0.03
dti	0.06	0.06	0.06	0.11	0.05	-0.12	1.00	-0.03	0.00	0.29	-0.01	0.23	0.23	0.04	0.10	-0.01	0.02	0.01	0.01	0.01	0.06	0.01
delinq_2yrs	-0.03	-0.03	-0.04	0.16	-0.02	0.02	-0.03	1.00	0.01	0.01	0.01	-0.06	0.07	-0.04	0.02	0.03	0.01	0.01	0.01	-0.01	-0.01	-0.01
inq_last_6mths	0.01	0.01	-0.00	0.14	0.01	0.03	0.00	0.01	1.00	0.09	0.02	-0.02	0.12	-0.02	0.03	0.03	0.02	0.01	0.01	0.03	-0.05	-0.04
open_acc	0.17	0.17	0.16	0.01	0.17	0.16	0.29	0.01	0.09	1.00	0.00	0.29	0.69	0.16	0.12	-0.02	0.02	0.01	0.01	0.08	0.02	-0.04
pub_rec	-0.05	-0.05	-0.05	0.10	-0.04	-0.01	-0.01	0.01	0.02	0.00	1.00	-0.06	-0.02	-0.06	-0.00	-0.01	-0.01	-0.01	0.84	-0.03	0.00	-0.02
revol_bal	0.32	0.31	0.29	0.10	0.31	0.28	0.23	-0.06	-0.02	0.29	-0.06	1.00	0.31	0.28	0.24	0.01	0.04	0.02	-0.05	0.13	0.03	-0.07
total_acc	0.26	0.25	0.24	-0.05	0.23	0.24	0.23	0.07	0.12	0.69	-0.02	0.31	1.00	0.23	0.15	-0.02	0.03	0.01	-0.01	0.16	-0.03	-0.03
total_rec_prncp	0.85	0.87	0.85	0.17	0.85	0.26	0.04	-0.04	-0.02	0.16	-0.06	0.28	0.23	1.00	0.68	-0.02	-0.09	-0.06	-0.05	0.57	0.16	-0.00
total_rec_int	0.73	0.74	0.73	0.52	0.64	0.19	0.10	0.02	0.03	0.12	-0.00	0.24	0.15	0.68	1.00	0.08	0.09	0.04	0.00	0.24	0.41	0.01
total_rec_late_fee	0.05	0.05	0.04	0.10	0.06	0.01	-0.01	0.03	0.03	-0.02	-0.01	0.01	-0.02	-0.02	0.08	1.00	0.10	0.09	-0.01	-0.06	0.02	-0.04
recoveries	0.14	0.14	0.14	0.13	0.12	0.02	0.02	0.01	0.02	0.02	-0.01	0.04	0.03	-0.09	0.09	0.10	1.00	0.80	-0.01	-0.07	-0.10	-0.02
collection_recovery_fee	0.08	0.08	0.07	0.07	0.08	0.02	0.01	0.01	0.01	0.01	-0.01	0.02	0.01	-0.06	0.04	0.09	0.80	1.00	-0.01	-0.04	-0.06	-0.03
pub_rec_bankruptcies	-0.03	-0.03	-0.04	0.08	-0.03	-0.01	0.01	0.01	0.01	0.01	0.84	-0.05	-0.01	-0.05	0.00	-0.01	-0.01	-0.01	1.00	-0.02	-0.01	-0.02
last_pymnt_amnt	0.48	0.48	0.47	0.17	0.41	0.14	0.01	-0.01	0.03	0.08	-0.03	0.13	0.16	0.57	0.24	-0.06	-0.07	-0.04	-0.02	1.00	-0.35	-0.01
total_mnths_paid	0.03	0.03	0.05	0.04	-0.02	0.00	0.06	-0.01	-0.05	0.02	0.00	0.03	-0.03	0.16	0.41	0.02	-0.10	-0.06	-0.01	-0.35	1.00	0.04
approved_loan_amnt_ratio	-0.12	-0.03	0.14	0.00	-0.04	-0.03	0.01	-0.01	-0.04	-0.04	-0.02	-0.07	-0.03	-0.00	0.01	-0.04	-0.02	-0.03	-0.02	-0.01	0.04	1.00

BIVARIATE ANALYSIS

Observations: Positive correlation between loan amount, funded amount and funded amount investment, say, most of funded amount approved by loan amount if a person funded amount, it increases the likelihood of loan amount. Now, dti(Risk) is negatively correlated with annual income, if dti(Risk) increase than annual income customer is decrease Hence, Low income customer easily charged off. There is -ve correlation between the dti and annual income, if annual income is less the risk will increase.

BIVARIATE ANALYSIS



correlation of Annual Income vs Loan Amount

Observation: 1) The above graph correlation has positive correlation. Hence if interest rates increase, risk(dti) is also increase. Hence, Charged off customers are more when interest rate increase while risk increases. 2) Interest rates are high for people who use their credit cards frequently.

FINDINGS AND RECOMMENDATIONS

Findings	Recommendations
Debt_Consolidation has more applicants	Reduce the risk by checking past records of late fees/recoveries/collections/delinquency of the applicant
'Charged Off' customers with non-verified income source is very high	Lender should verify source of income and should reduce the risk, it will reduce the loss
dti(Risk) is negatively correlated with annual income	Lenders can keep Applicants property as Mortage as applicants with low income and high DTI are risky and less likely to repay loan