Lending Club Case Study

GROUP MEMBERS:
ANSHUL AWASTHI
AMOOLYA BOORA

Business Objectives

- Lending Club 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.

The goal is to minimize financial loss for a consumer finance company by identifying risky loan applicants.

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

1. Loss of business if a loan is not approved for a reliable applicant.

2. Financial loss if a loan is approved for a risky applicant who defaults.

- Objective: Identify patterns indicating likelihood of default to aid in loan approval decisions.

Exploratory Data Analysis on Lending Club dataset

- **Data Cleaning**

1. Dataset having (39717, 111) values
2. Dropping columns having more than 30% null values,
3. This include dropping column 'desc' with 33% null values. As this column didn't have any important values that are going to impact our analysis.
4. Dropped rows where applicants are currently paying their debts i.e. dropping 1140 rows present of loan_status=‘current’
5. Dropping columns ‘url’ and ‘member_id’ as all values were unique in nature.
6. Reviewed for duplicate values (none present).
7. Treated textual data columns ‘title’ and ‘emp_title’ not having contribution to our analysis
8. Dropped sub_group column which further categorises the group column, not contributing to the analysis
9. Dropping following columns with only one unique value, not impacting our analysis: pymnt_plan, initial_list_status, out_prncp, out_prncp_inv, collections_12_mths_ex_med, policy_code, application_type, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt, tax_liens
10. Further, analysing the remaining columns from the data dictionary, identified and dropped the columns having behavioural data of the customers usually captured post the loan approval. Therefore treating 18 such columns.
11. Post the cleaning process, left with data with (38577, 19) [rows, column]

Data Manipulation: Data Conversion, Imputing, Derived Columns

- Convert the following columns:

loan_amnt and funded_amnt as flot64

term column into an integer from a string

int_rate to float by removing the "%" character

column issue_d from string object to DateTime

issue_d has been converted to date type.

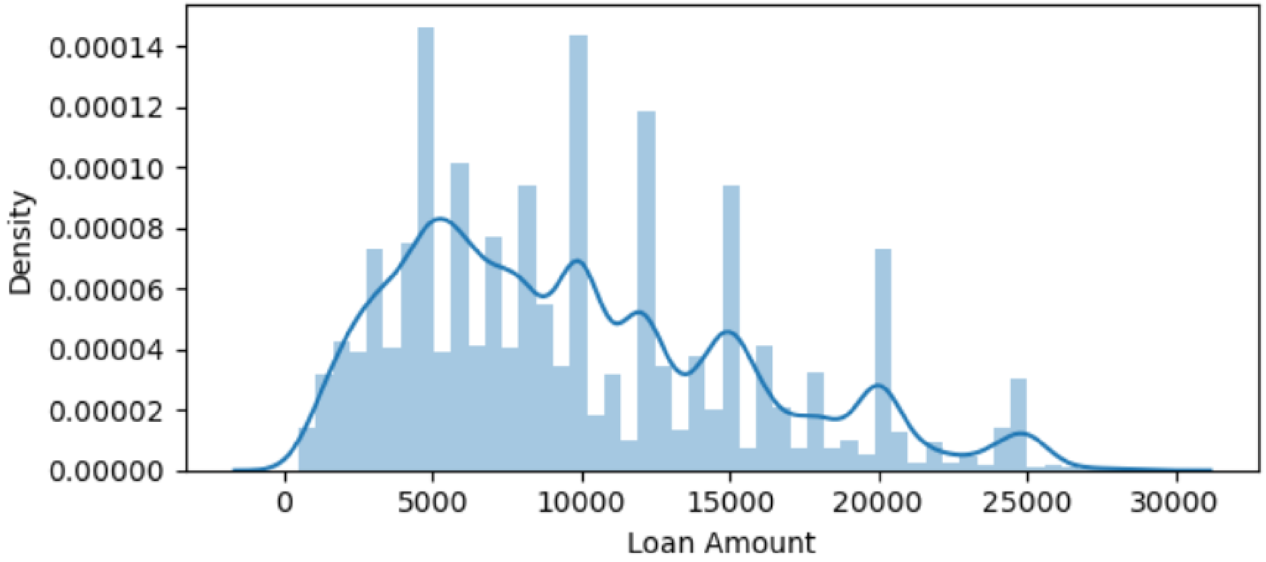
- Creating a derived columns for ‘issue_year’, ‘issue_month’, ‘issue_quarter’ from ‘issue_d’ which will be using for further analysis.
- ‘loan_amnt_cat’, ‘annual_inc_cat’, ‘int_rate_cat’, and ‘dti_cat’ derived columns(multiple buckets from respective data columns). These are created for better analysis
- There exists Outliers for numeric data 'loan_amnt', 'funded_amnt', ‘funded_amnt_inv','int_rate', 'installment’ and ‘annual_inc'. Outliers treatment has been done for above columns using quantile mechanism.

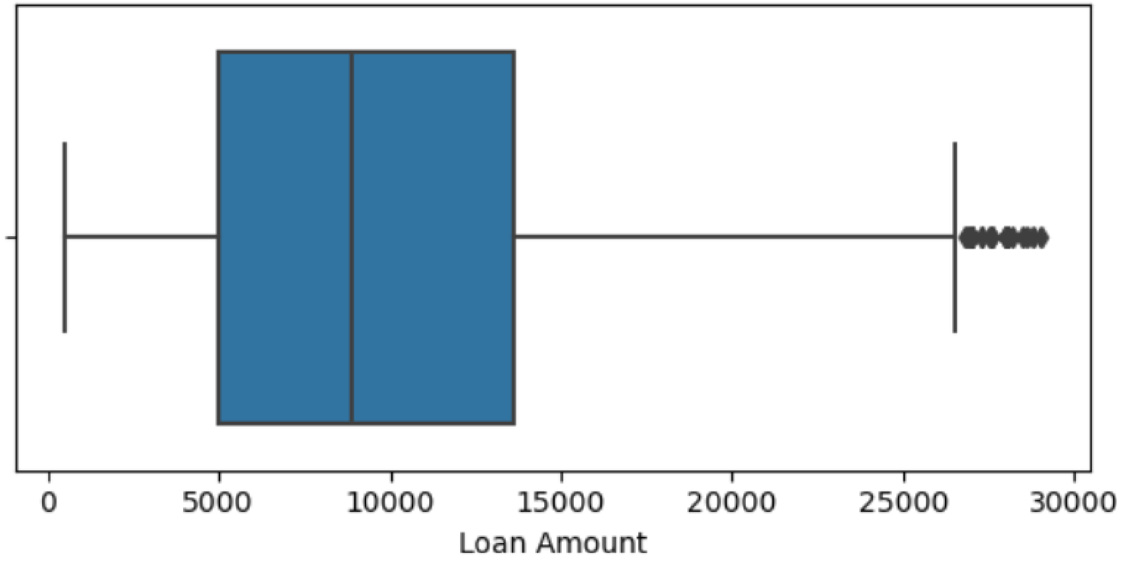
Univariate Analysis : with Data Visualization

- LOAN AMOUNT
- INTEREST RATE
- ANNUAL INCOME
- HOME OWNERSHIP
- PURPOSE
- EMP LENGTH
- STATE

Quantitative Univariate Analysis : Loan Amount

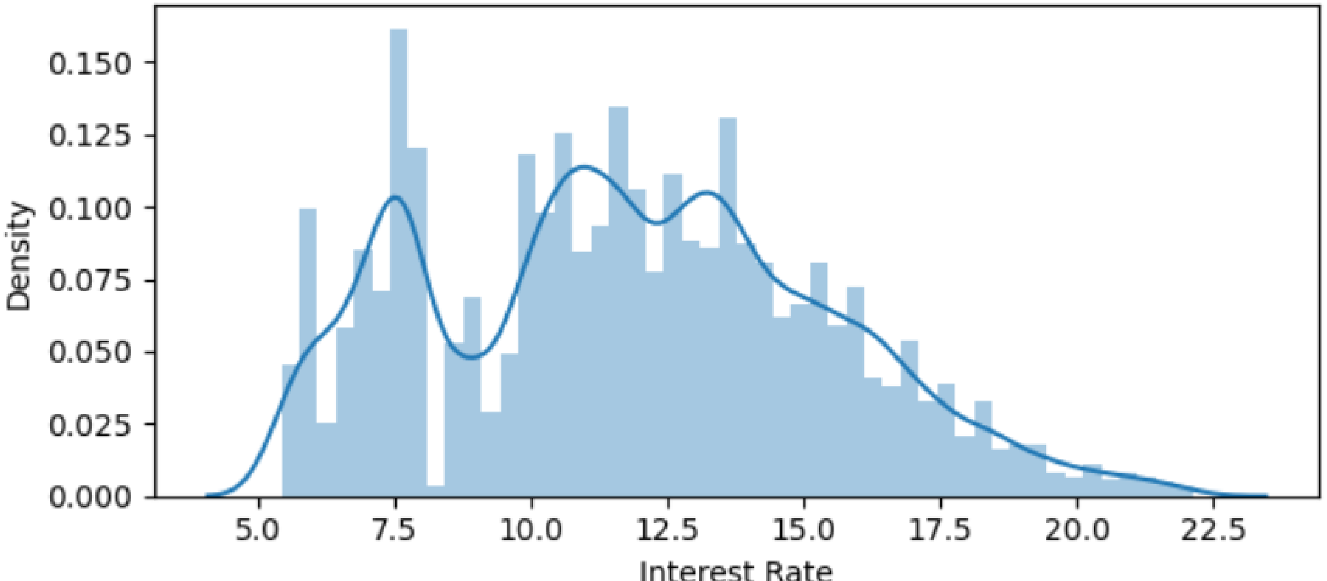
- Majority of the loan_amount is in the range of 5K to 14K
- Maximum amount applied for is 29K

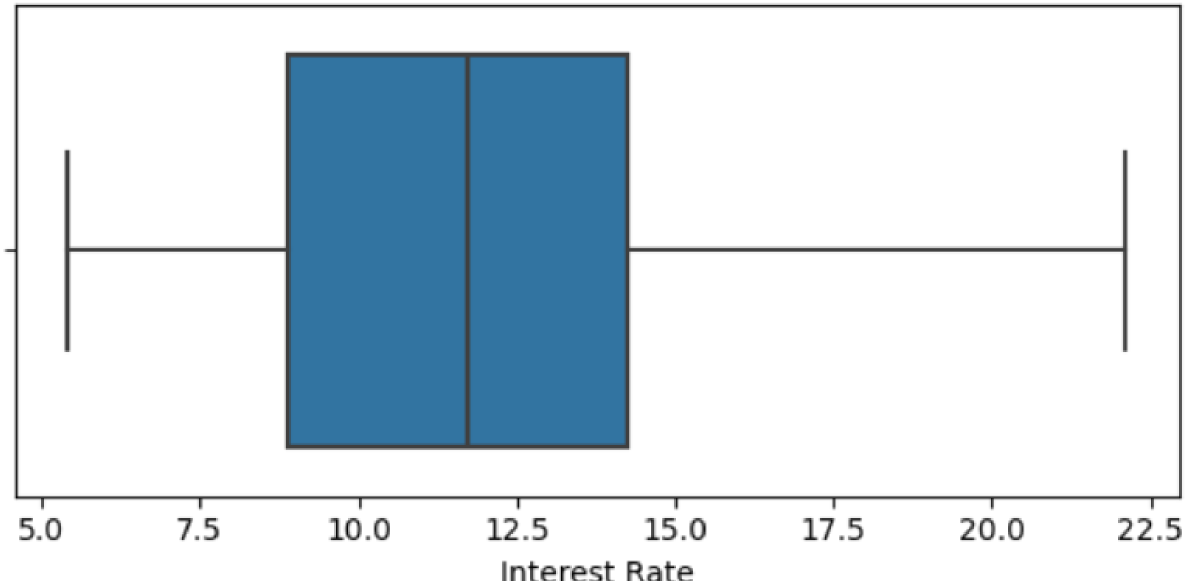




Quantitative Univariate Analysis : Interest Rate

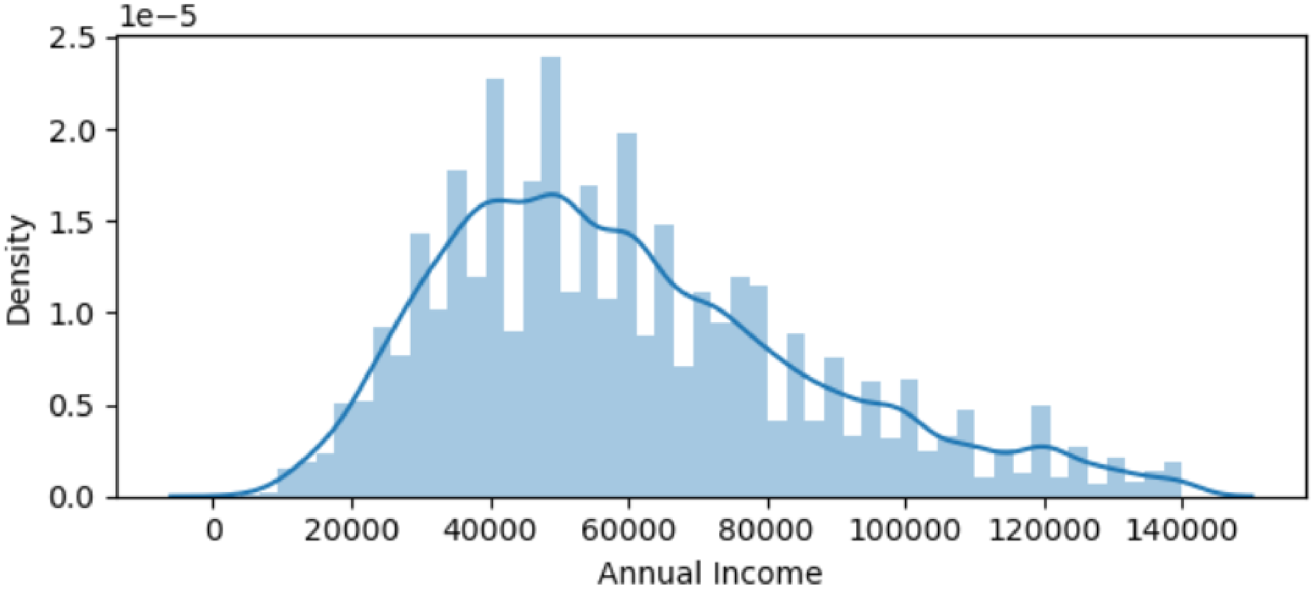
- Majority of the interest_rate is in the range of 8% to 14%
- The average rate of interest is 11.7%

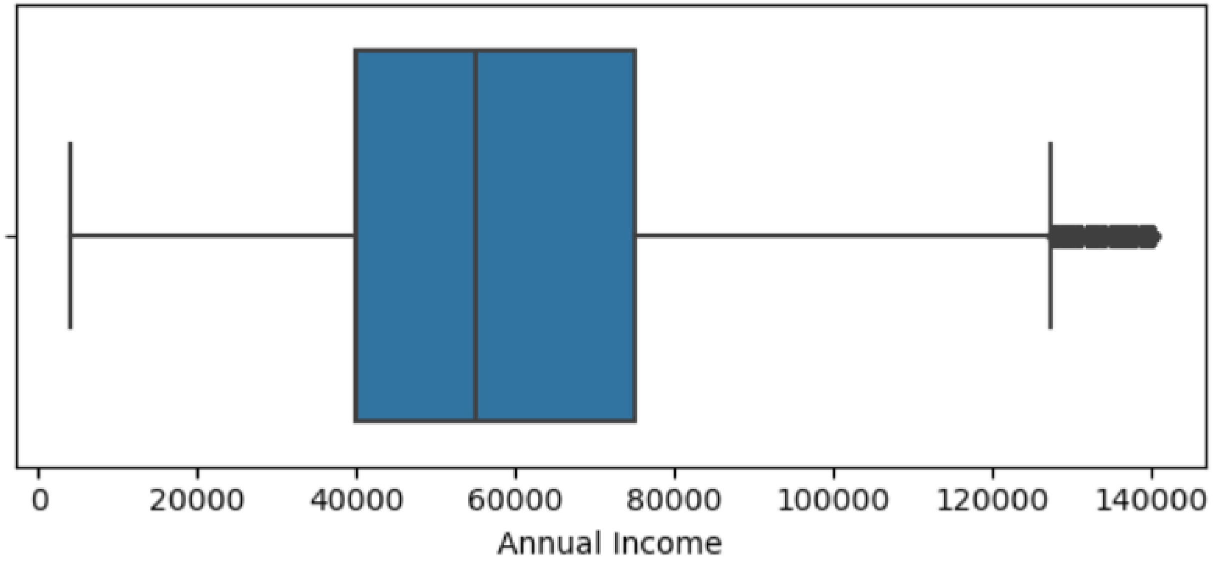




Quantitative Univariate Analysis : Annual Income

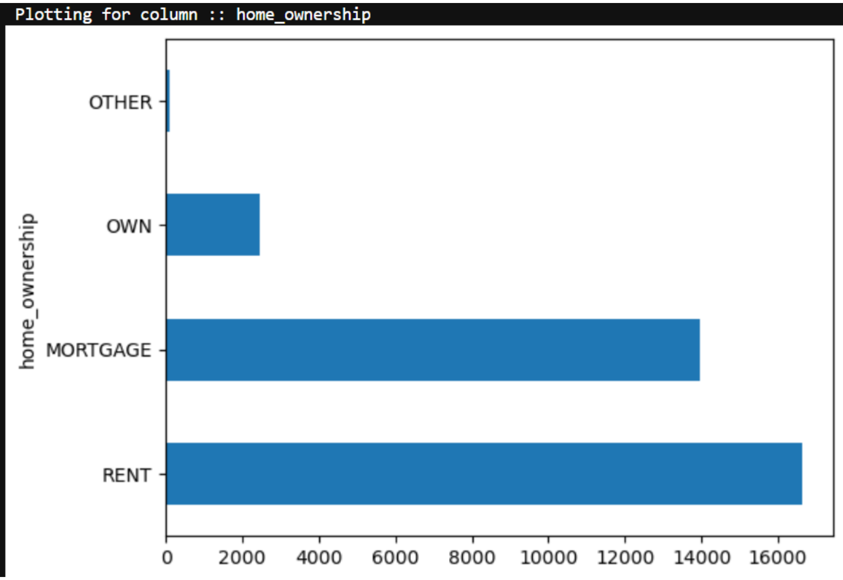
- The average annual income of Applicants fall between 40K to 75K which falls at 59K
- The lowest earning applicants are standing at a paycheque of 4k and the elite ones at 140K



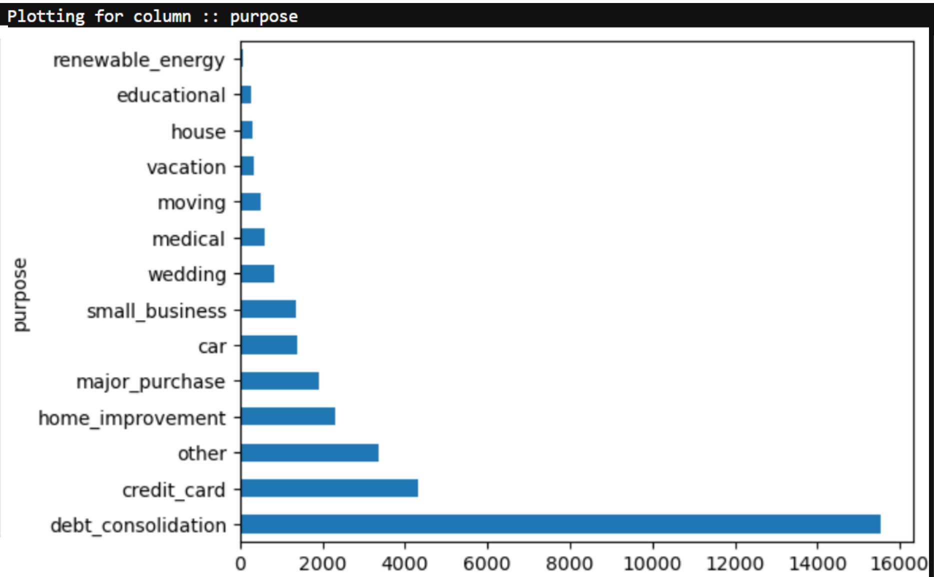


Univariant Analysis : Unordered & Ordered Categorical Variable Analysis : Visuals

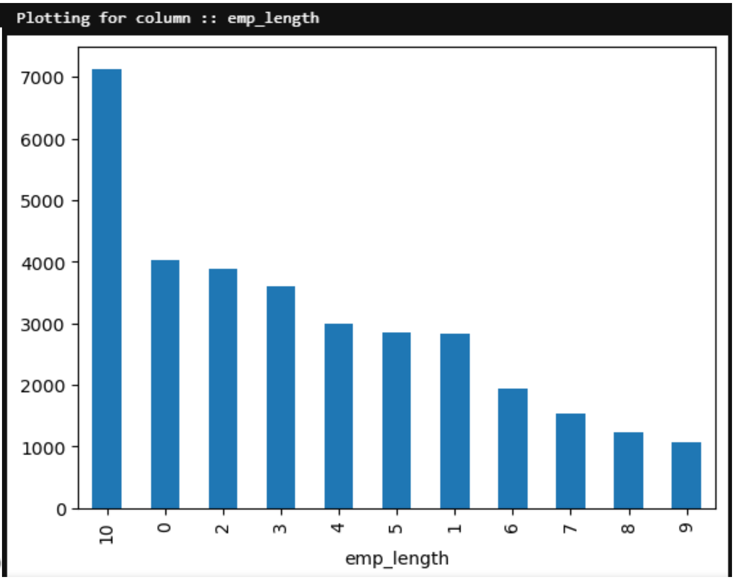
- **Observations:-**
- • Large share of loan applicants are either living on Rent or on Mortgage
- • The purpose majority applicants provided the purpose of debt_consolidations for their loan request.
- • Most of the applications are having 10+ yrs of Job experience.



Plotting for column :: home_ownership



Plotting for column :: purpose

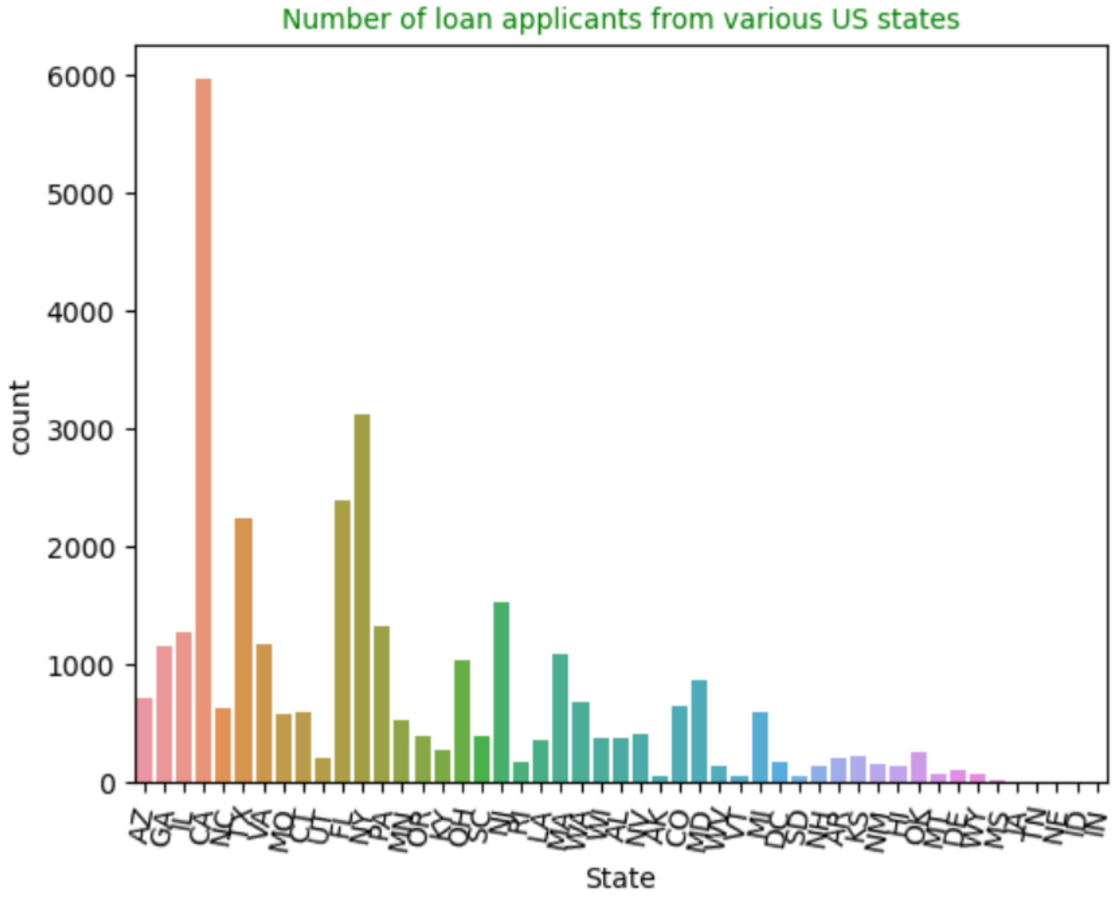


Plotting for column :: emp_length

Univariant Analysis : Unordered & Ordered Categorical Variable Analysis

- Most of the Loan applicants are from CA(State).

- Loan Demographics:-
  > Highest loan amount applications fall in the range of 5k to 10k
  > Majority of the interest rate is in the range of 5% to 16% going at the max to 22%.
  > Majority of the installment amount is in the range of 20.
  > Majority of the loan applications counts are in the term of 36 months.
  > Majority of loan application counts fall under the catogory of Grade B



Univariate Summary

- **Time Based Analysis**

1. The loan application count increases every year
2. The highest number of loan applications are in Quarter 4 of every year.
3. Lowest number of loan applications are in Q1 might be because :

   - By year ends people face the financial challenges
   - Holiday/Festive season
   - Possibly because they are consolidating debt by year end

- **Inferences**

1. The dataset helps understand which segment of customers, the Lending Club needs to target for highest volume of loan.
2. Highlights that more introspection is needed as why some categories are not as high as other few.
3. Signifies that the Lending Club has high volume in Q4 and it should target customers in other quarters to increase sales.

Bivariate Analysis : with Data Visualization

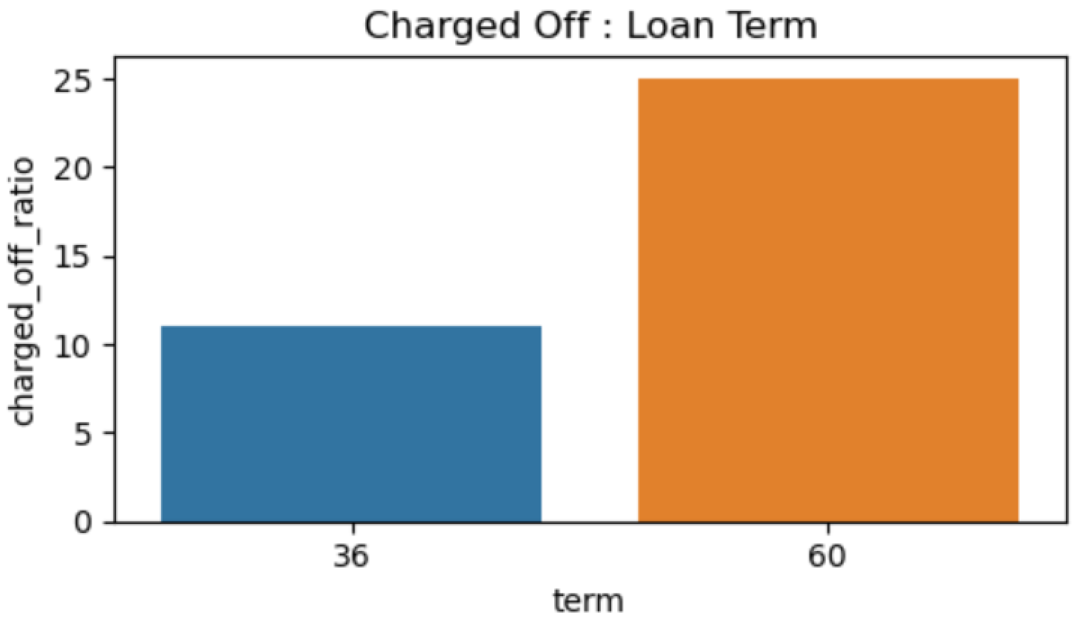
V/S CHARGED OFF STATUS:

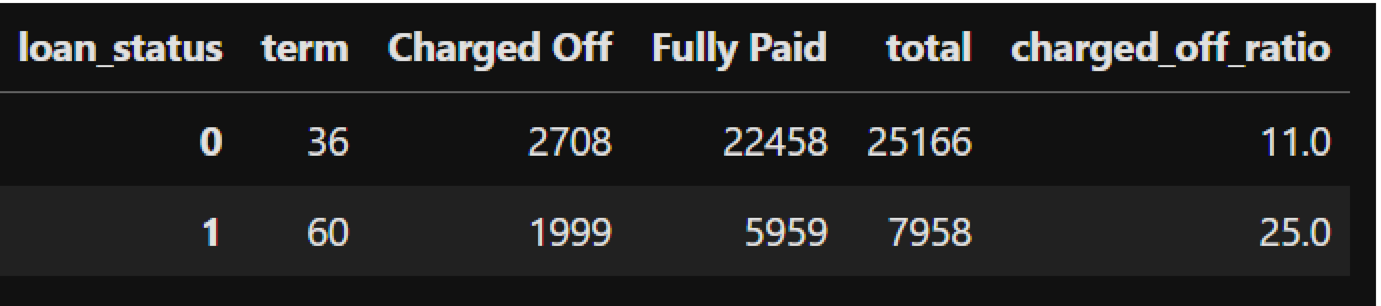
- TERM
- EMP LENGTH
- HOME OWNERSHIP
- BANKRUPTCIES RECORD
- ISSUE QUARTER
- ANNUAL INCOME CATEGORY
- LOAN AMOUNT CATEGORY
- INTEREST RATE CATEGORY
- STATE

IDENTIFYING CAUSES THAT CONTRIBUTE TO MORE CHARGE OFF'S

term

- The overall count of Charged Off's is slightly higher in term 36 as compared to term 60
- If we calculate the ratio of Charge Off's within a category
- Charge Offs ratio is for the term=60 is 25% which is much higher than term=36 (10%)
- term=60 is the loan applications which require more scrutiny
- Inferences
- Most of the applicants with term=60 potentially will have high Charge Offs

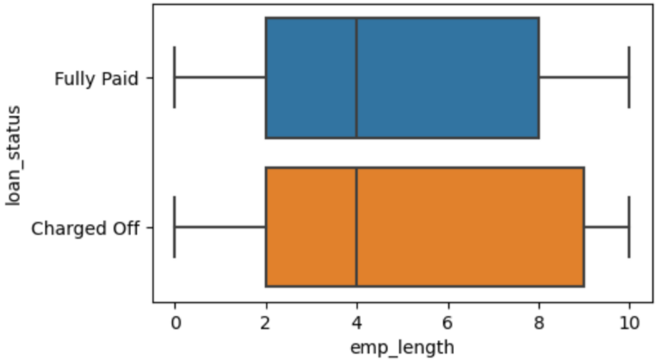


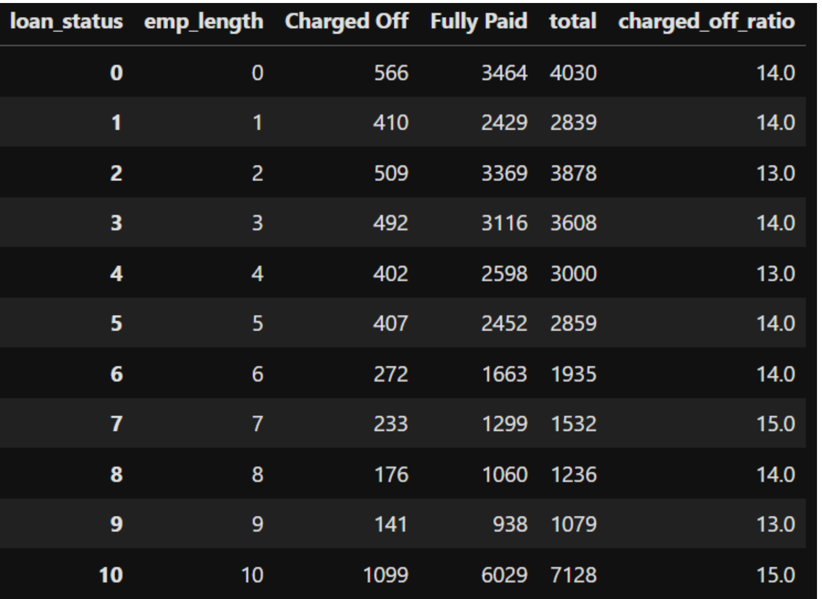


loan_status	term	Charged Off	Fully Paid	total	charged_off_ratio
0	36	2708	22458	25166	11.0
1	60	1999	5959	7958	25.0

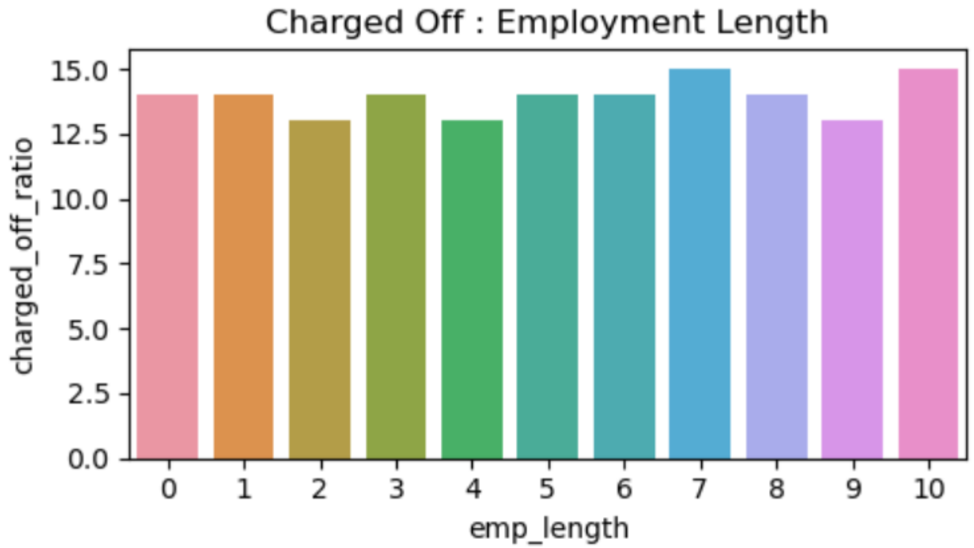
emp_length

- Highest Charge Offs are in the employee length of 10 Years and above
- High probablity of Charge Off's are the ones, having income range in less than 1 years
- The charge off ratio within the ranges are pretty much same (in conclusive)



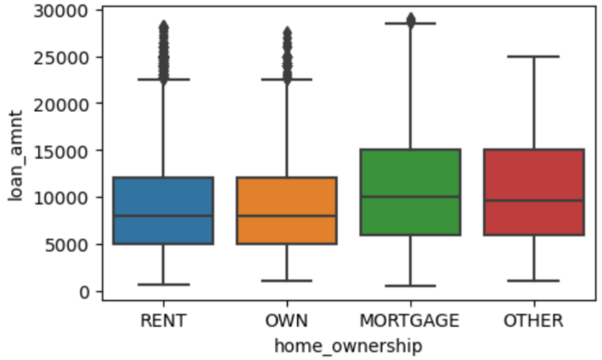


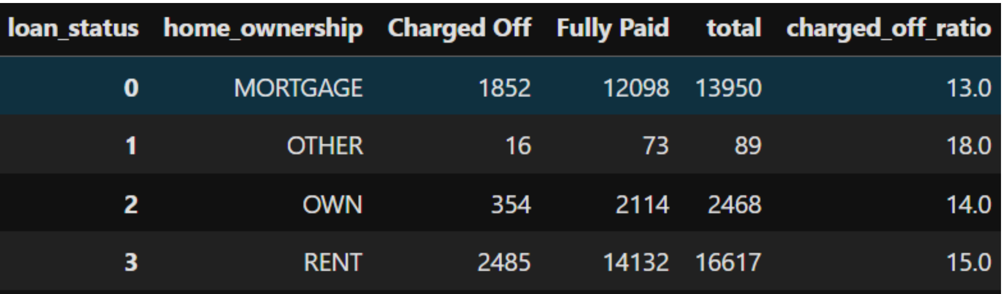
loan_status	emp_length	Charged Off	Fully Paid	total	charged_off_ratio
0	0	566	3464	4030	14.0
1	1	410	2429	2839	14.0
2	2	509	3369	3878	13.0
3	3	492	3116	3608	14.0
4	4	402	2598	3000	13.0
5	5	407	2452	2859	14.0
6	6	272	1663	1935	14.0
7	7	233	1299	1532	15.0
8	8	176	1060	1236	14.0
9	9	141	938	1079	13.0
10	10	1099	6029	7128	15.0



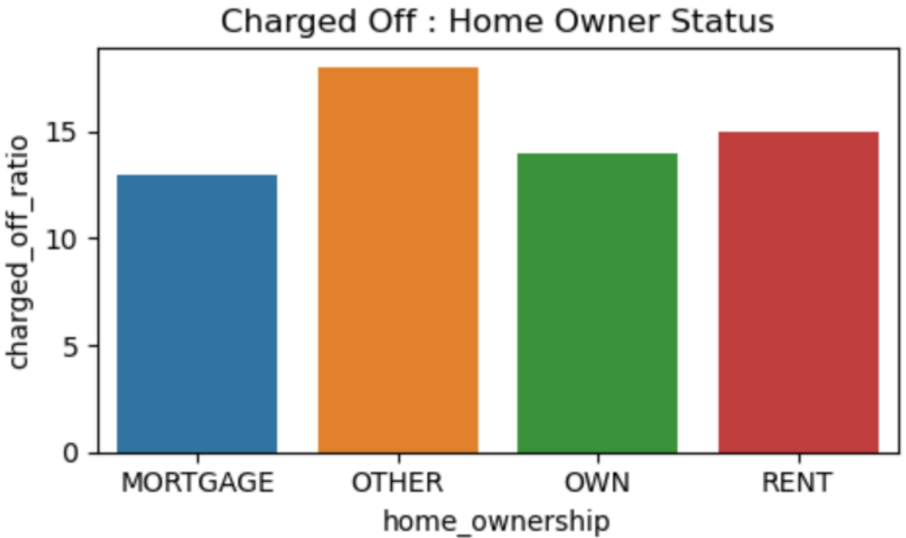
home_ownership

- Overall the highest Charge Off numbers are in the category of RENT and MORTGAGE
- Within each home_ownership category the Charge Off ratio of for Other is higher
- The MORTGAGE category of applicants are at the highest risk of Charge Offs. They also have the highest range of loan amounts increasing the risk



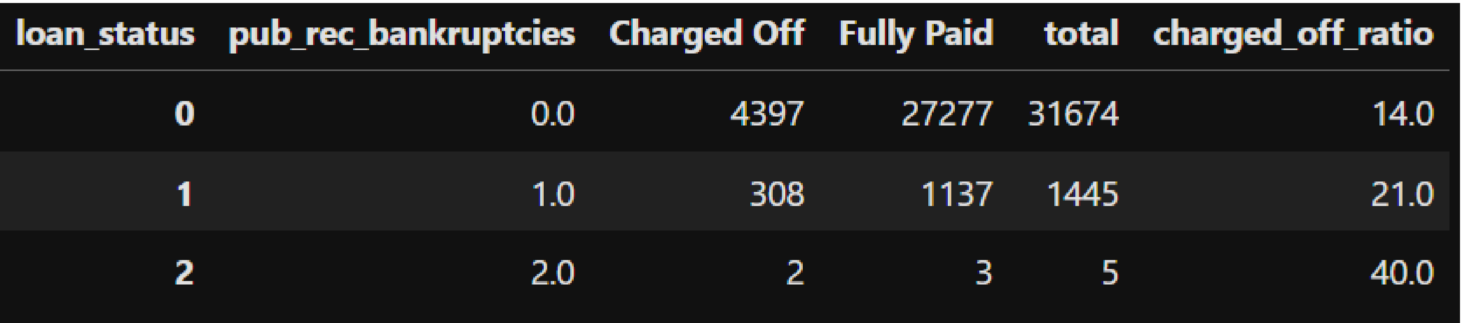


loan_status	home_ownership	Charged Off	Fully Paid	total	charged_off_ratio
0	MORTGAGE	1852	12098	13950	13.0
1	OTHER	16	73	89	18.0
2	OWN	354	2114	2468	14.0
3	RENT	2485	14132	16617	15.0

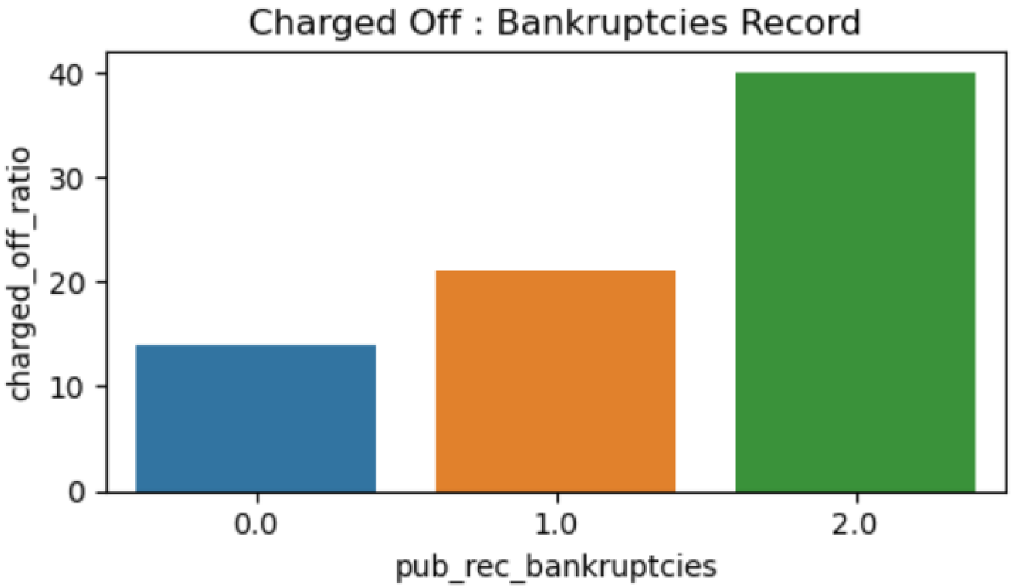


pub_rec_bankruptcies

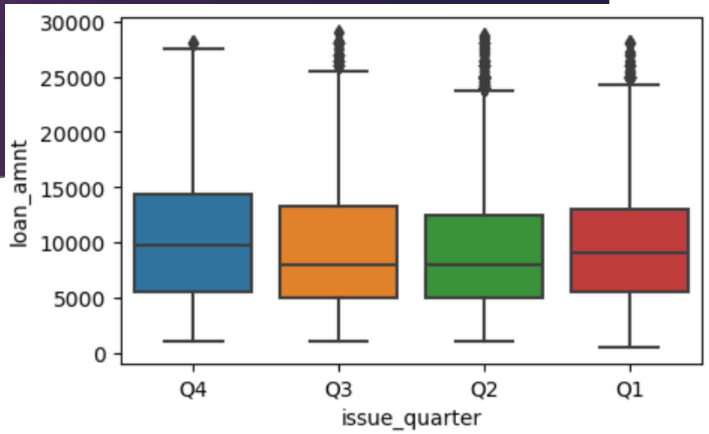
- The large number of charge_off falls under 0 category (i.e. no bankruptcy record)
- Reviewing the ratio within each category, customers having bankruptcy record have a high charge_off ratio
- Customers having bankruptcy record are at high risk of Charge Offs
- pub_rec_bankruptcies record 2, has the highest Charge Off ratio having only a fewer values for analysis



loan_status	pub_rec_bankruptcies	Charged Off	Fully Paid	total	charged_off_ratio
0	0.0	4397	27277	31674	14.0
1	1.0	308	1137	1445	21.0
2	2.0	2	3	5	40.0

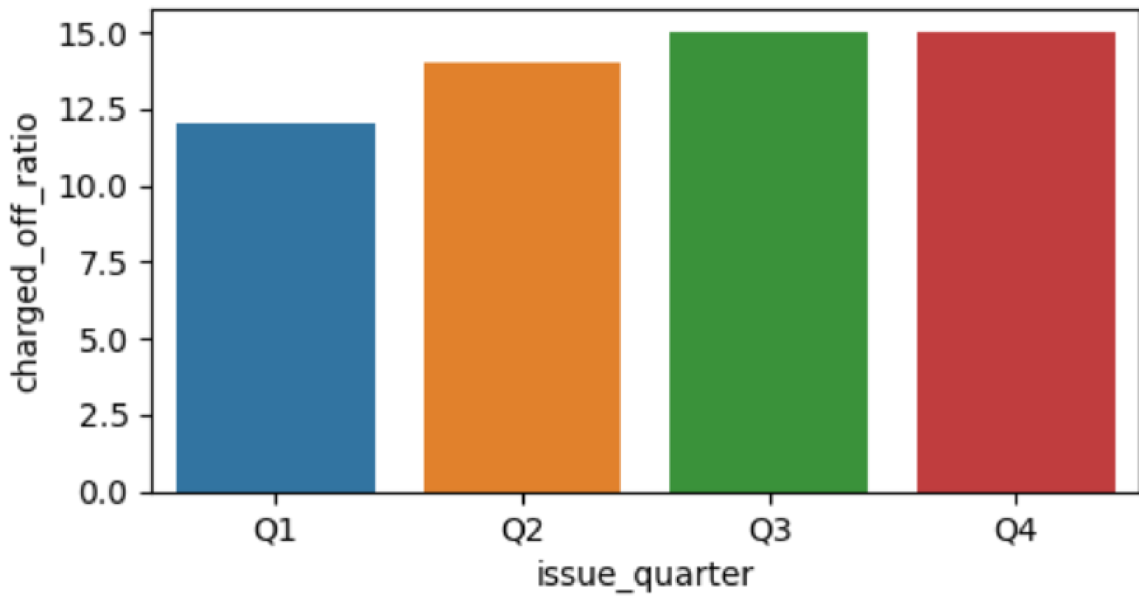


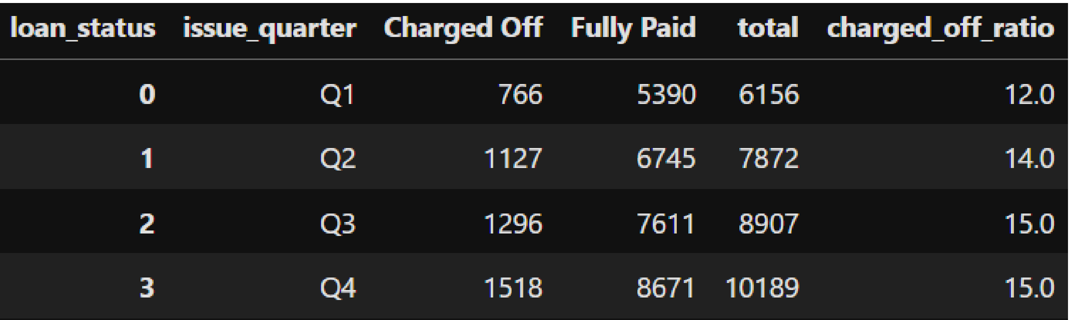
issue_quarter



- Q 4 of every year has the highest ratio of Charge Offs
- Years has no significant impact a apart from increasing the volume year over year, which is impacting the charge offs.
- The year 2007 has the maximum Charge Offs which means current running loan that started in 2007 may have risk of defaulting.

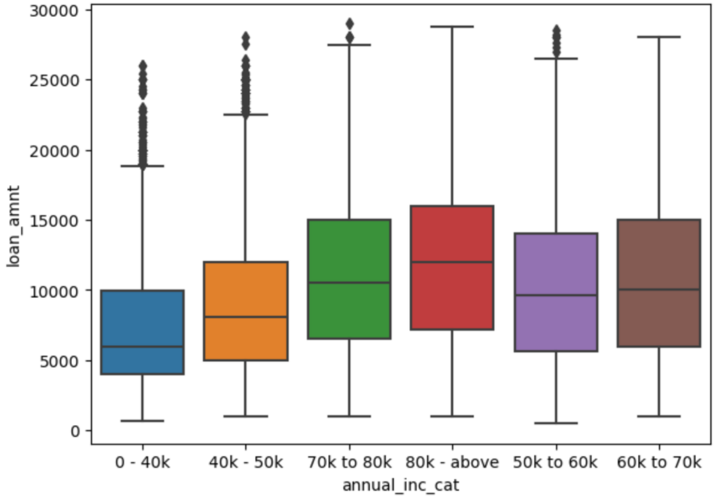
Charged Off : Issue Quarter



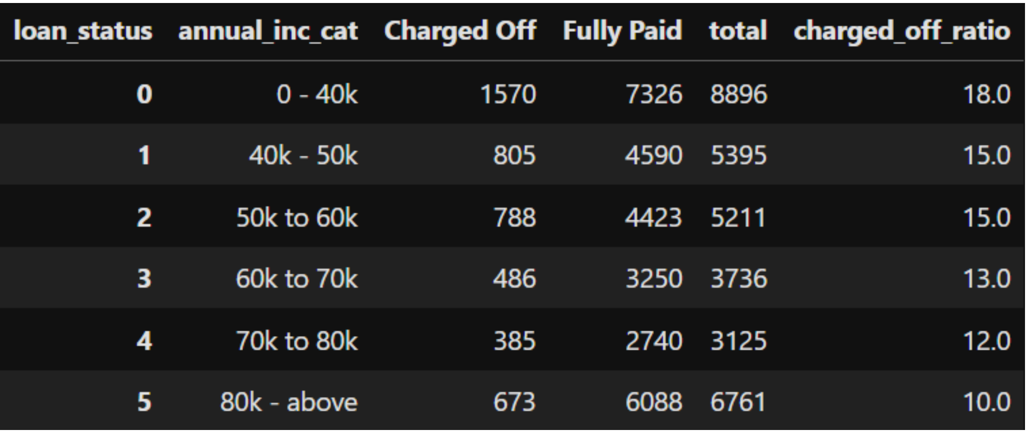


loan_status	issue_quarter	Charged Off	Fully Paid	total	charged_off_ratio
0	Q1	766	5390	6156	12.0
1	Q2	1127	6745	7872	14.0
2	Q3	1296	7611	8907	15.0
3	Q4	1518	8671	10189	15.0

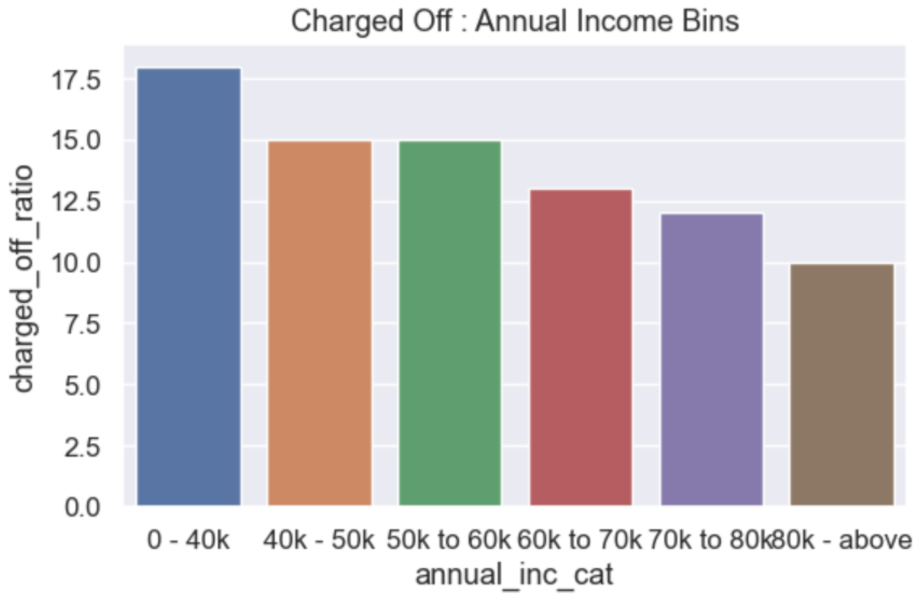
annual_inc_cat



- The Annual income range of 0-40K has the highest charge offs
- The Charge off ratio within the bucket of 0-40K have highest Charge Offs
- The income range of 0-40K have the highest risk
- Income range 80000+ has less chances of charged off.
- Increase in annual income charged off proportion decreases.

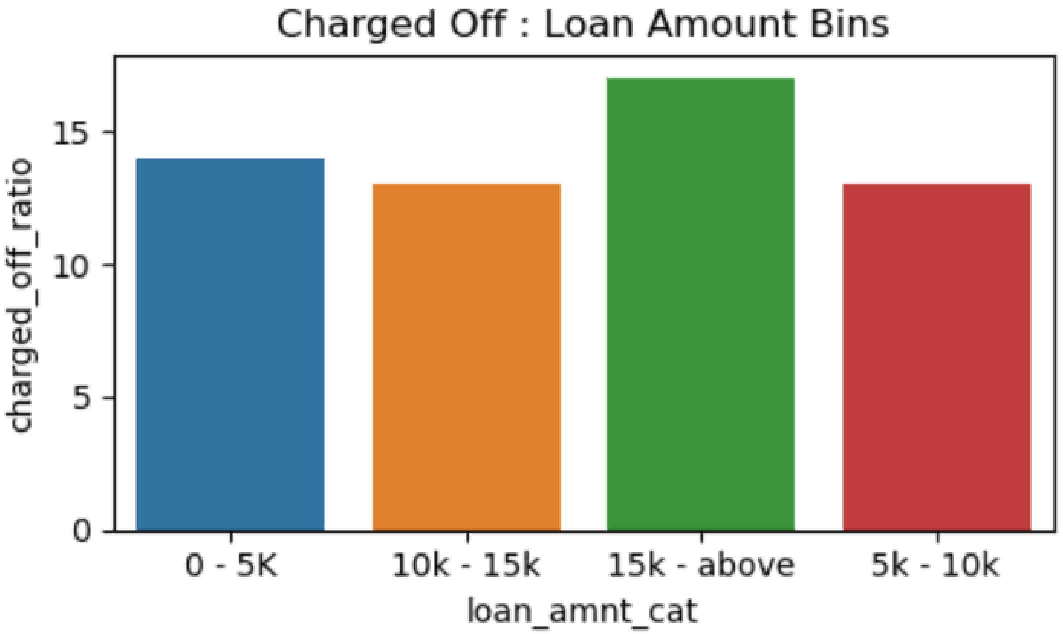


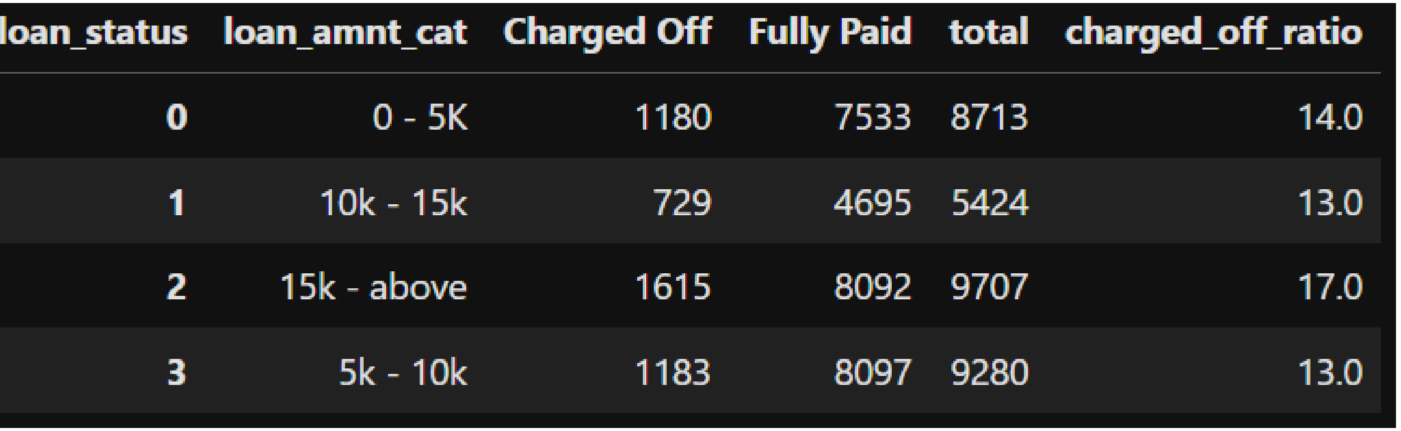
loan_status	annual_inc_cat	Charged Off	Fully Paid	total	charged_off_ratio
0	0 - 40k	1570	7326	8896	18.0
1	40k - 50k	805	4590	5395	15.0
2	50k to 60k	788	4423	5211	15.0
3	60k to 70k	486	3250	3736	13.0
4	70k to 80k	385	2740	3125	12.0
5	80k - above	673	6088	6761	10.0



loan_amnt_cat

- The highest percentage of Charge Offs are in the category 5K to 10k of the loan_amount
- The Charge Off ratio of all the customers within the loan_amount of 15K and above is at the highest Charge Off risk

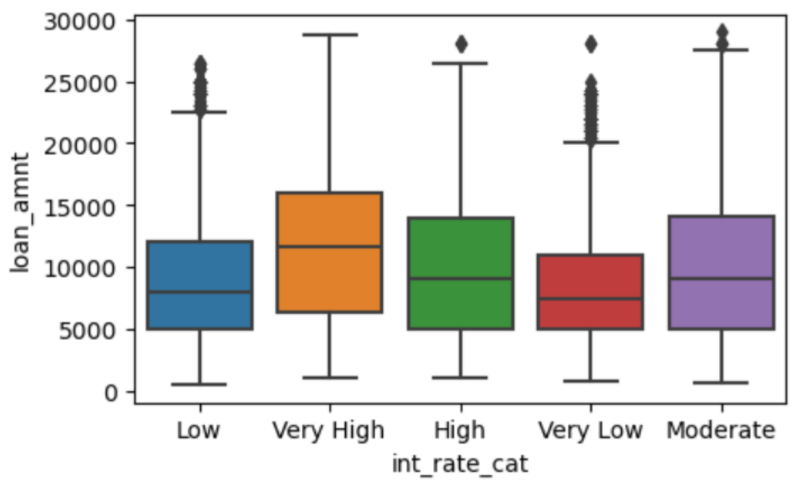


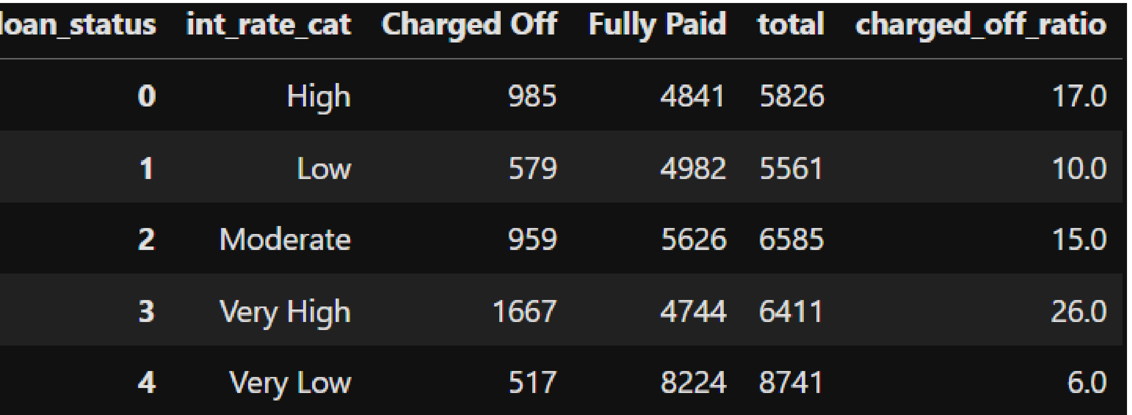


loan_status	loan_amnt_cat	Charged Off	Fully Paid	total	charged_off_ratio
0	0 - 5K	1180	7533	8713	14.0
1	10k - 15k	729	4695	5424	13.0
2	15k - above	1615	8092	9707	17.0
3	5k - 10k	1183	8097	9280	13.0

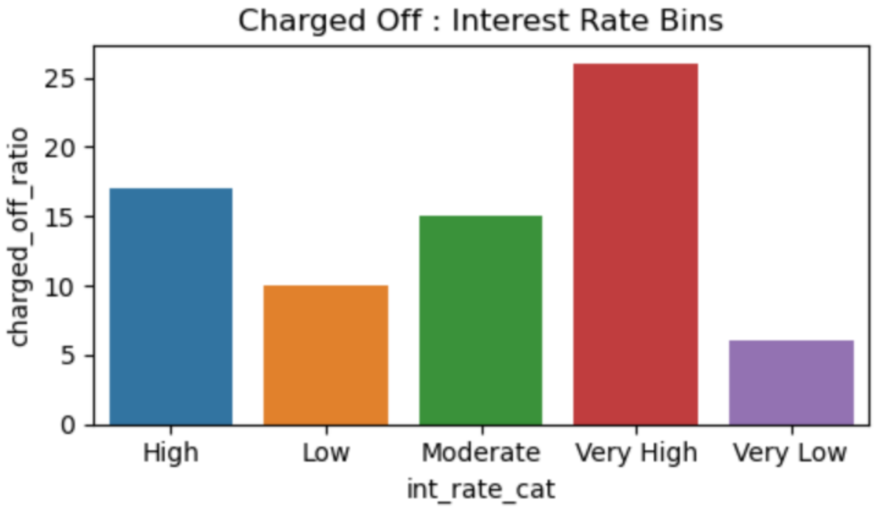
int_rate_cat

- The Charge Off ratio within the category ‘Very High’ interest rates are at a risk of Charge Off
- The category of Very High interest rate is 15% and above



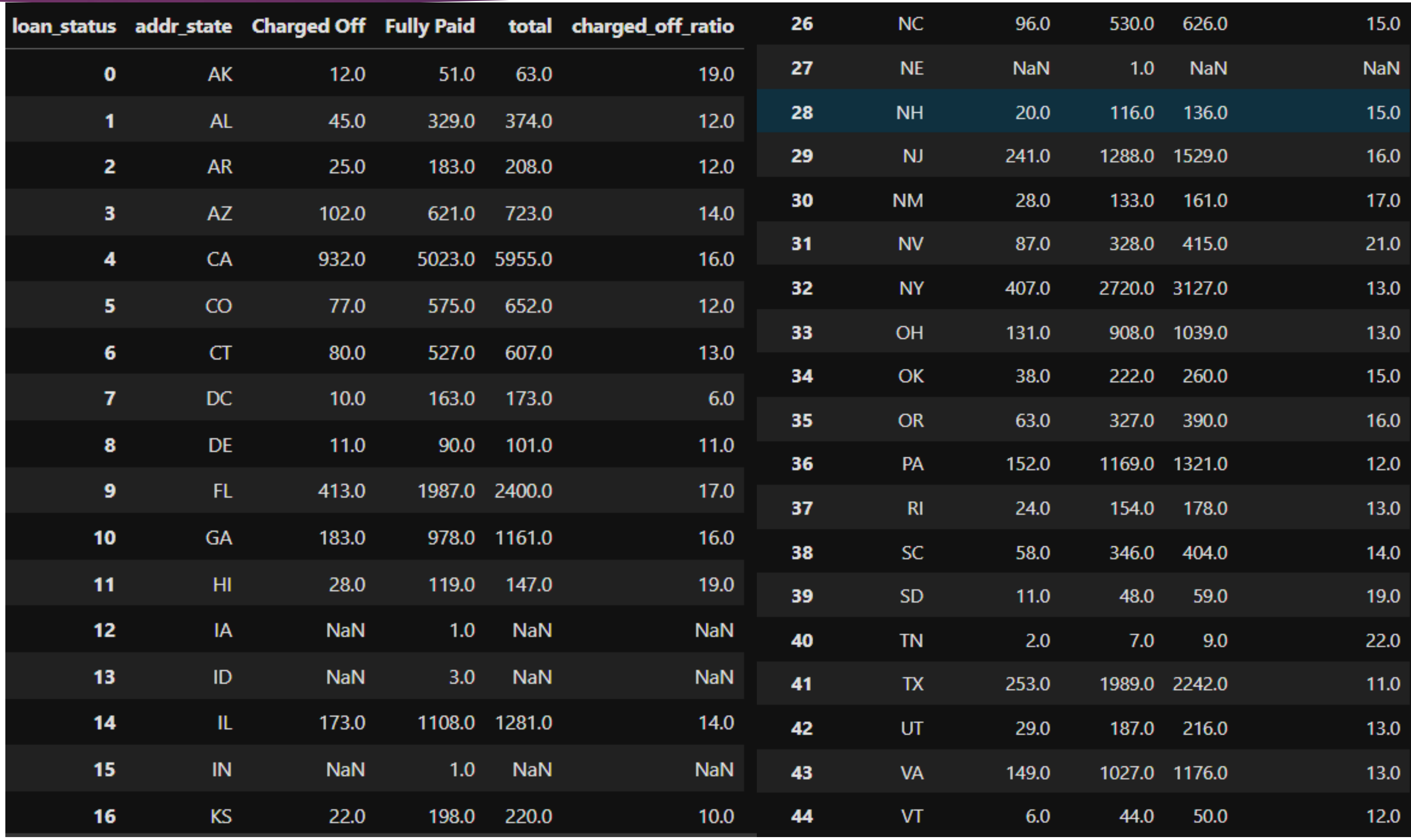


loan_status	int_rate_cat	Charged Off	Fully Paid	total	charged_off_ratio
0	High	985	4841	5826	17.0
1	Low	579	4982	5561	10.0
2	Moderate	959	5626	6585	15.0
3	Very High	1667	4744	6411	26.0
4	Very Low	517	8224	8741	6.0



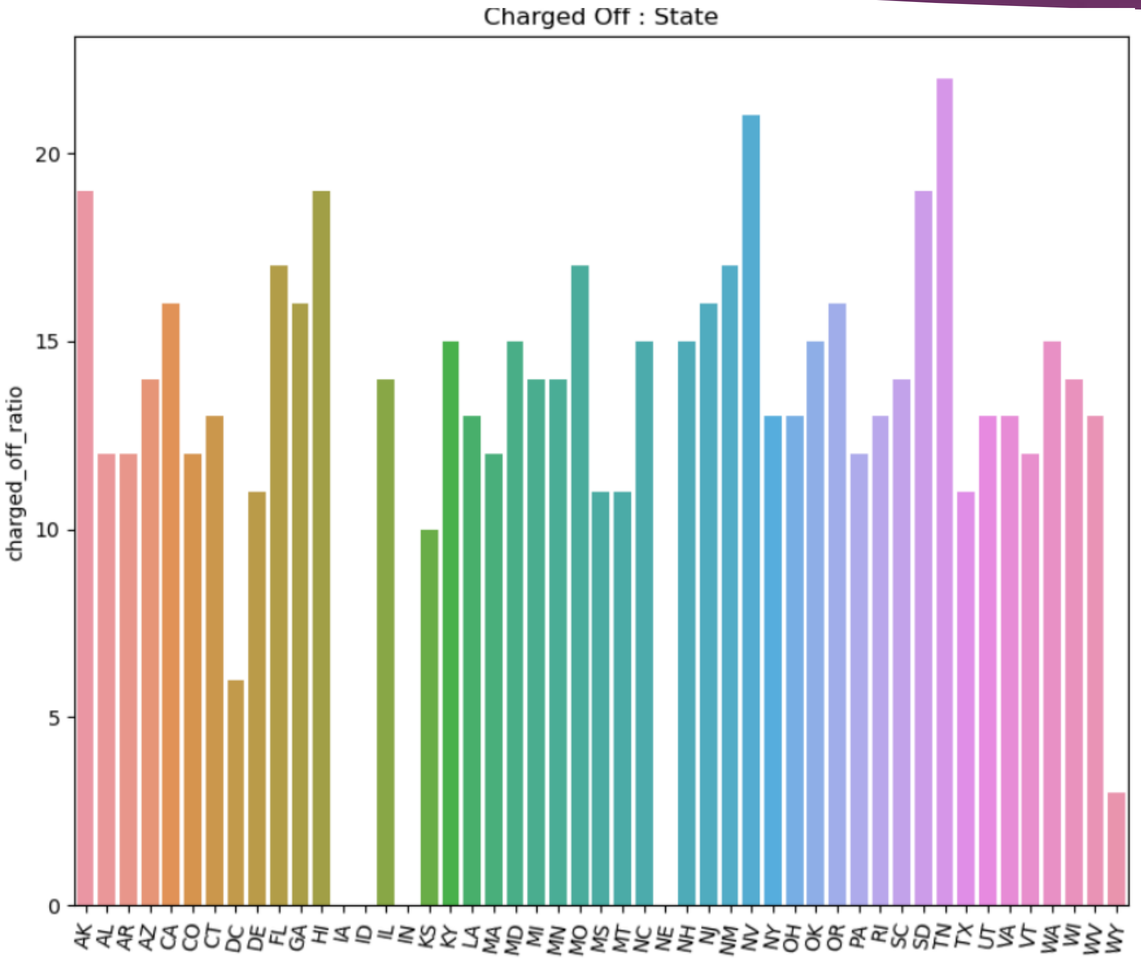
addr_state : State column

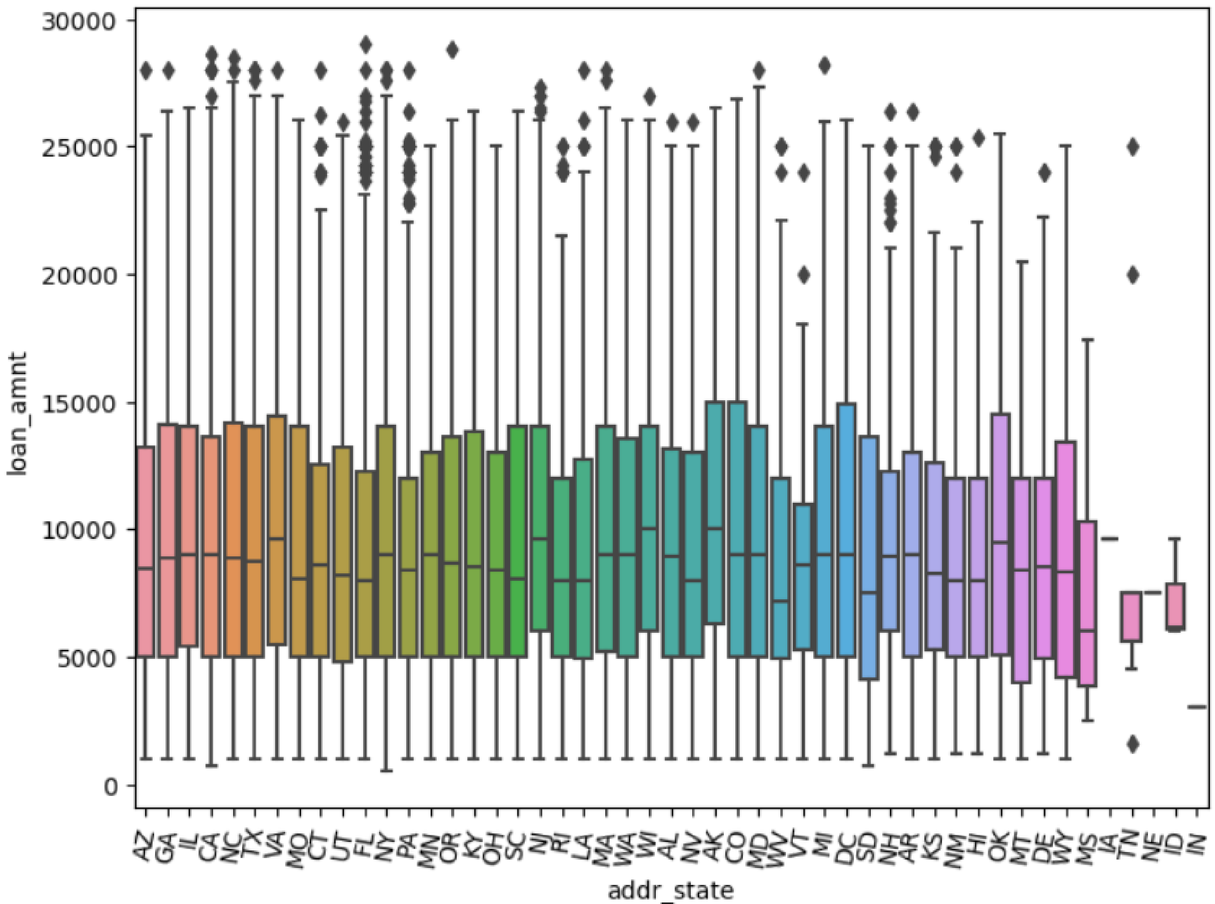
- Highest volume of loans is from CA and the highest Charge Off's are from CA
- Within each state NE and NV has the highest Charge Offs
- NE has very low volume which is not very clear for our analysis on the state.
- Loan applications from NV have high risk of Charge Offs
- TN, NV, CA and FL have high percentage of Charge Off's



loan_status	addr_state	Charged Off	Fully Paid	total	charged_off_ratio
0	AK	12.0	51.0	63.0	19.0
1	AL	45.0	329.0	374.0	12.0
2	AR	25.0	183.0	208.0	12.0
3	AZ	102.0	621.0	723.0	14.0
4	CA	932.0	5023.0	5955.0	16.0
5	CO	77.0	575.0	652.0	12.0
6	CT	80.0	527.0	607.0	13.0
7	DC	10.0	163.0	173.0	6.0
8	DE	11.0	90.0	101.0	11.0
9	FL	413.0	1987.0	2400.0	17.0
10	GA	183.0	978.0	1161.0	16.0
11	HI	28.0	119.0	147.0	19.0
12	IA	NaN	1.0	NaN	NaN
13	ID	NaN	3.0	NaN	NaN
14	IL	173.0	1108.0	1281.0	14.0
15	IN	NaN	1.0	NaN	NaN
16	KS	22.0	198.0	220.0	10.0
26	NC	96.0	530.0	626.0	15.0
27	NE	NaN	1.0	NaN	NaN
28	NH	20.0	116.0	136.0	15.0
29	NJ	241.0	1288.0	1529.0	16.0
30	NM	28.0	133.0	161.0	17.0
31	NV	87.0	328.0	415.0	21.0
32	NY	407.0	2720.0	3127.0	13.0
33	OH	131.0	908.0	1039.0	13.0
34	OK	38.0	222.0	260.0	15.0
35	OR	63.0	327.0	390.0	16.0
36	PA	152.0	1169.0	1321.0	12.0
37	RI	24.0	154.0	178.0	13.0
38	SC	58.0	346.0	404.0	14.0
39	SD	11.0	48.0	59.0	19.0
40	TN	2.0	7.0	9.0	22.0
41	TX	253.0	1989.0	2242.0	11.0
42	UT	29.0	187.0	216.0	13.0
43	VA	149.0	1027.0	1176.0	13.0
44	VT	6.0	44.0	50.0	12.0

State column : Diagrams for analysis





Multivariate/Correlation Analysis

- Explored interactions between multiple variables.
- Negative Correlation

loan_amnt has negative correlation with pub_rec_bankrupticies

annual income has a negative correlation with dti field

- Strong Correlation

term has a strong correlation with loan amount

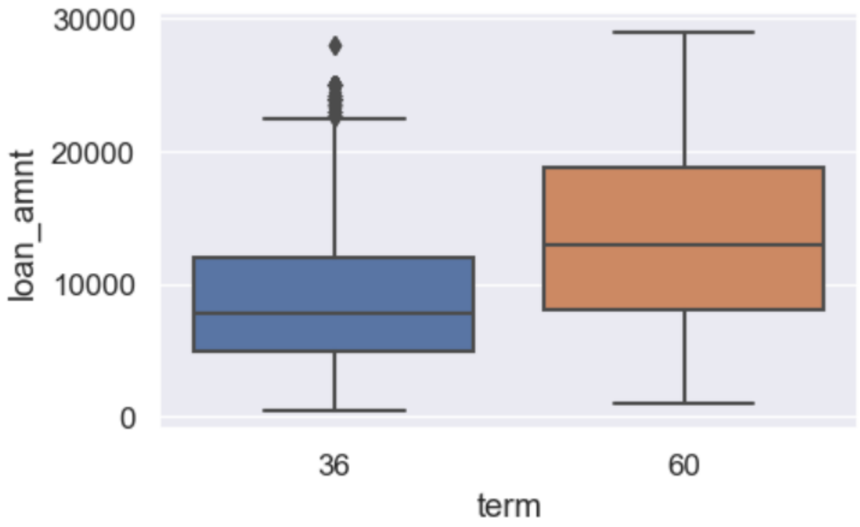
term has a strong correlation with interest rate

annual income has a strong correlation with loan_amount

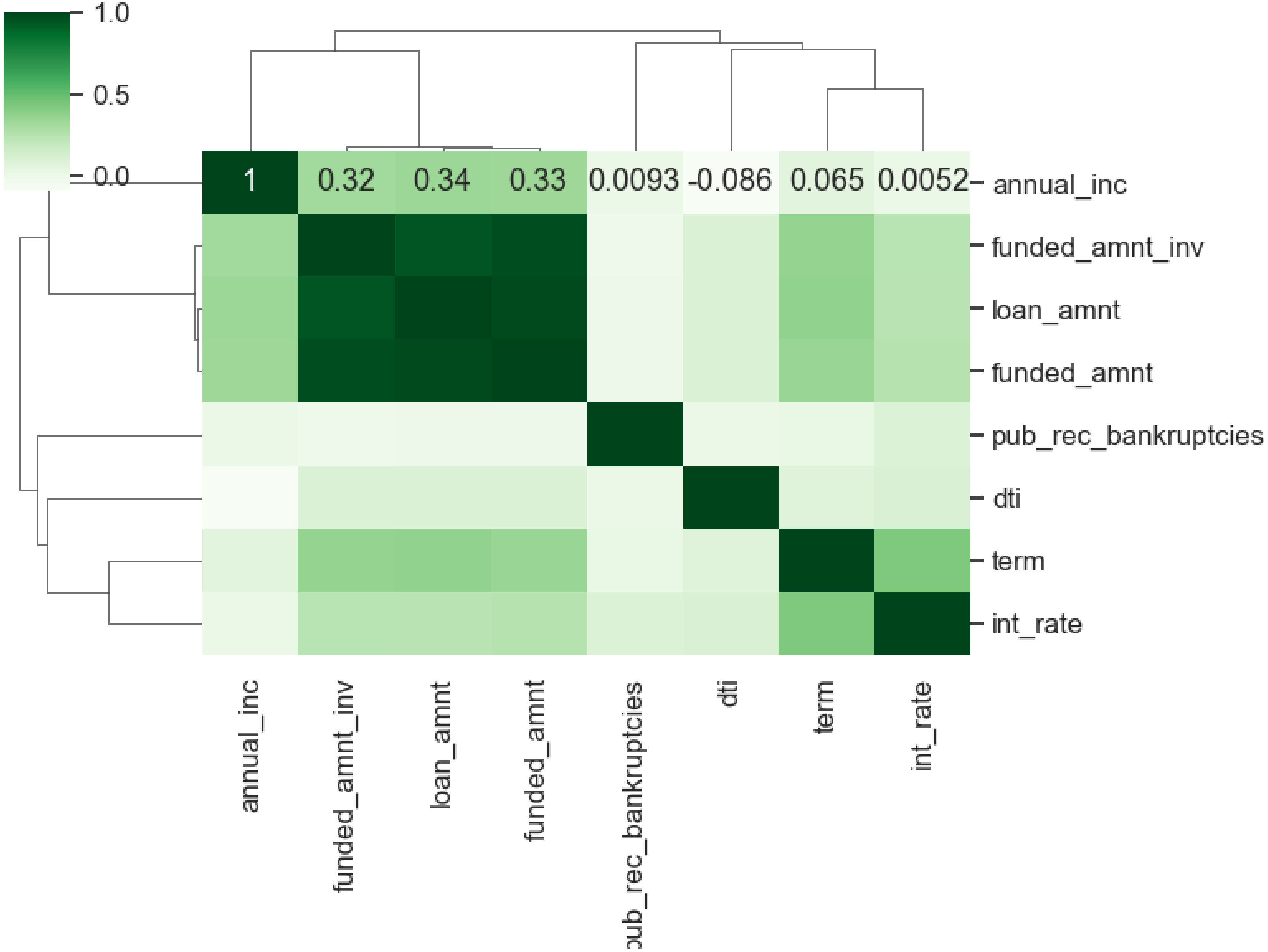
- Weak Correlation

debt-to-income ratio and annual income show correlations with loan status.

pub_rec_bankruptcies has weak correlation with majority of the attributes

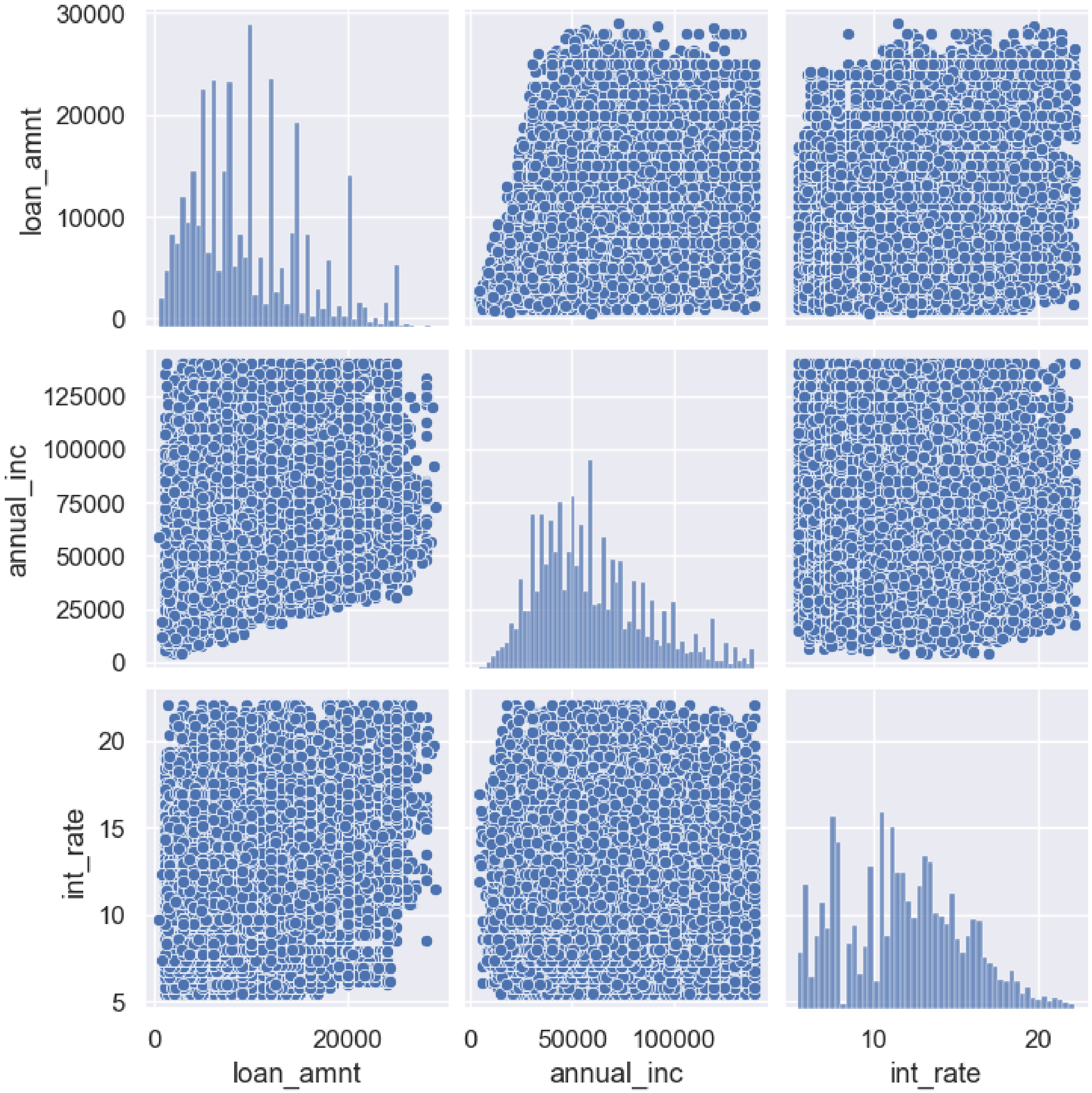


E.g. of Strong Correlation between term and loan amount


1.0
0.5
0.0
1
0.32
0.34
0.33
0.0093
-0.086
0.065
0.0052
annual_inc
funded_amnt_inv
loan_amnt
funded_amnt
pub_rec_bankruptcies
dti
term
int_rate
annual_inc
funded_amnt_inv
loan_amnt
funded_amnt
pub_rec_bankruptcies
dti
term
int_rate

Pair plot

- Plotting a pair plot among ‘loan_amnt’, 'annual_inc', 'int_rate'



Insights and Conclusions

- Loans with higher amounts have a higher default rate.
- Higher debt-to-income ratio is associated with higher risk.
- Emp length shows a trend where shorter employment duration correlates with higher default rates.

- Business Implications:
Adjust loan approval criteria to minimize defaults.
Implement higher interest rates for riskier applicants.

- Individuals with the income range between 0-20000 have a high chances of charged off.
- Interest rate of more than 16% has good chances of charged off as compared to other category.
- Individuals who are not owning the home is having high chances of loan defaulter.
- The high DTI value having high risk of defaults.
- Higher the Bankruptcy record higher the chance of loan defaults.
- The applicants with loan Grade G is having highest Loan Defaults

Thank You!