Lending Club Case Study

1. Reading & Understanding the data

- **Importing the input files**
- **Inspect data**

2. Data Cleaning and Manipulation

- **Remove columns where NA values are more than or equal to 30%**
- **Remove irrelevant columns.**

  Now let's look at each column from business perspective if that is required or not for our analysis such as Unique ID's, URL. As last 2 digits of zip code is masked 'xx', we can remove that as well.

- **Remove irrelevant records**

  Purpose of loan: Drop records where values are less than 0.75% We will analyse only those categories which contain more than 0.75% of records. Also, we are not aware what comes under 'Other' we will remove this category as well.

3. Derived Metrics

We will now derive some new columns based on our business understanding that will be helpful in our analysis.

1. **Loan amount to Annual Income ratio**
2. **Extract Year & Month from Issue date**
3. **Remove '%' from int_rate column**

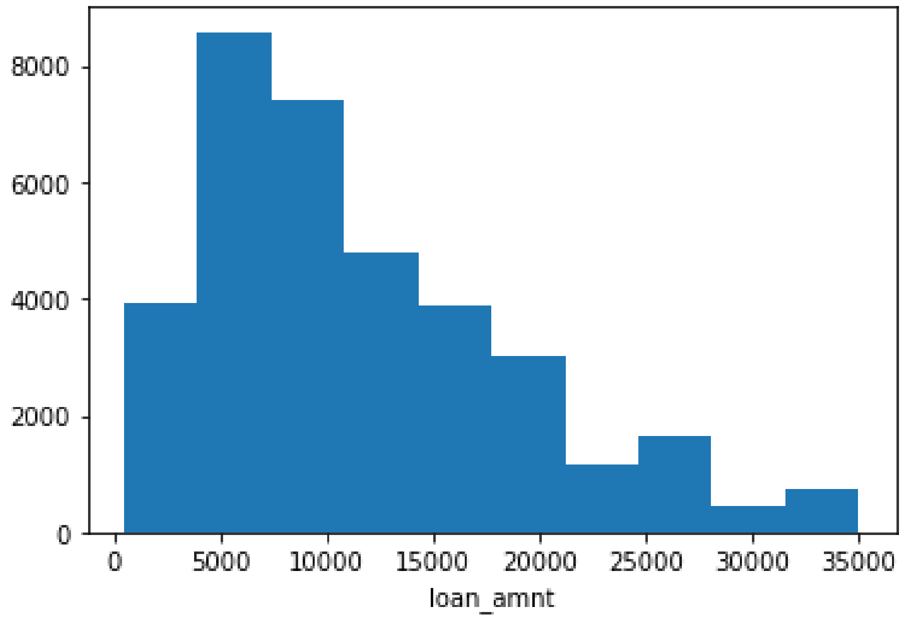
4. Univariate Analysis

- Continuous Variables In case of continuous variables, we need to understand the central tendency and spread of the variable.These are measured using various statistical metrics visualization methods such as Boxplot,Histogram/Distribution Plot etc.

- Categorical Variables For categorical variables, we'll use frequency table to understand distribution of each category. It can be be measured using two metrics, Count and Count% against each category. Countplot or Bar chart can be used as visualization.

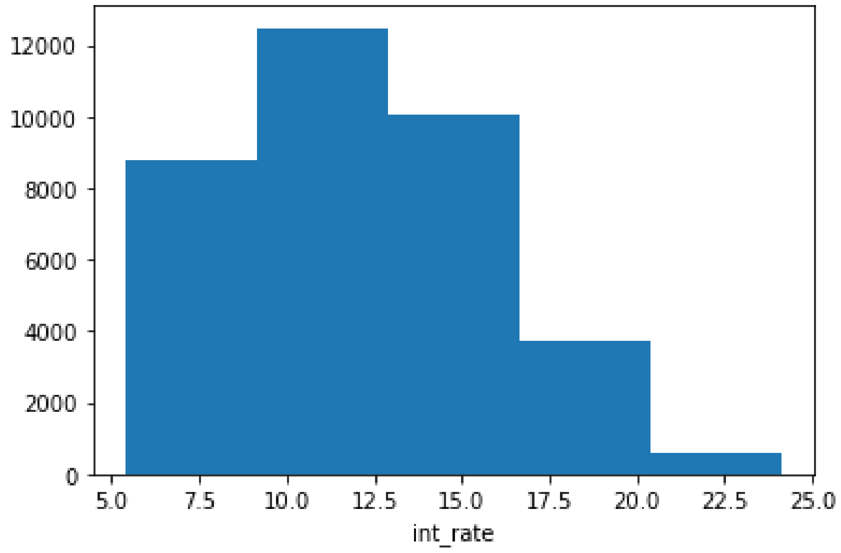
Continuous Variables

1. Loan Amount



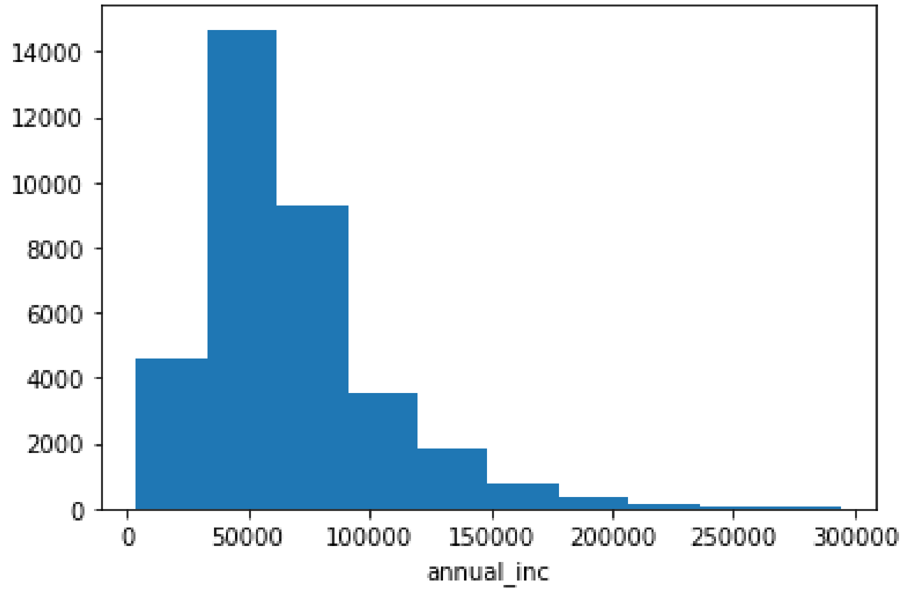
Insights: Most of the loan amounts are distributed between 5000 to 10000 USD.

2. Interest Rate



Insights: Most of the loans interest rates are distributed between 10% to 16%.

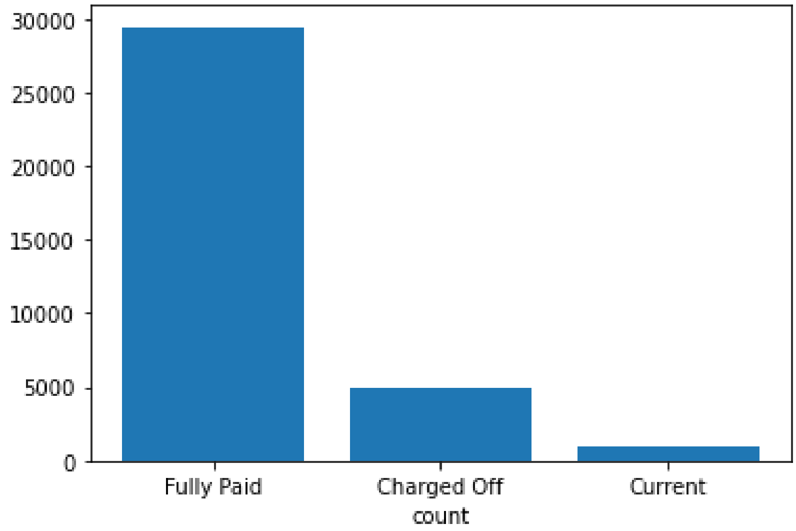
3. Annual Income



Insights: Most of the applicants earns between 40000 to 90000 USD annually.

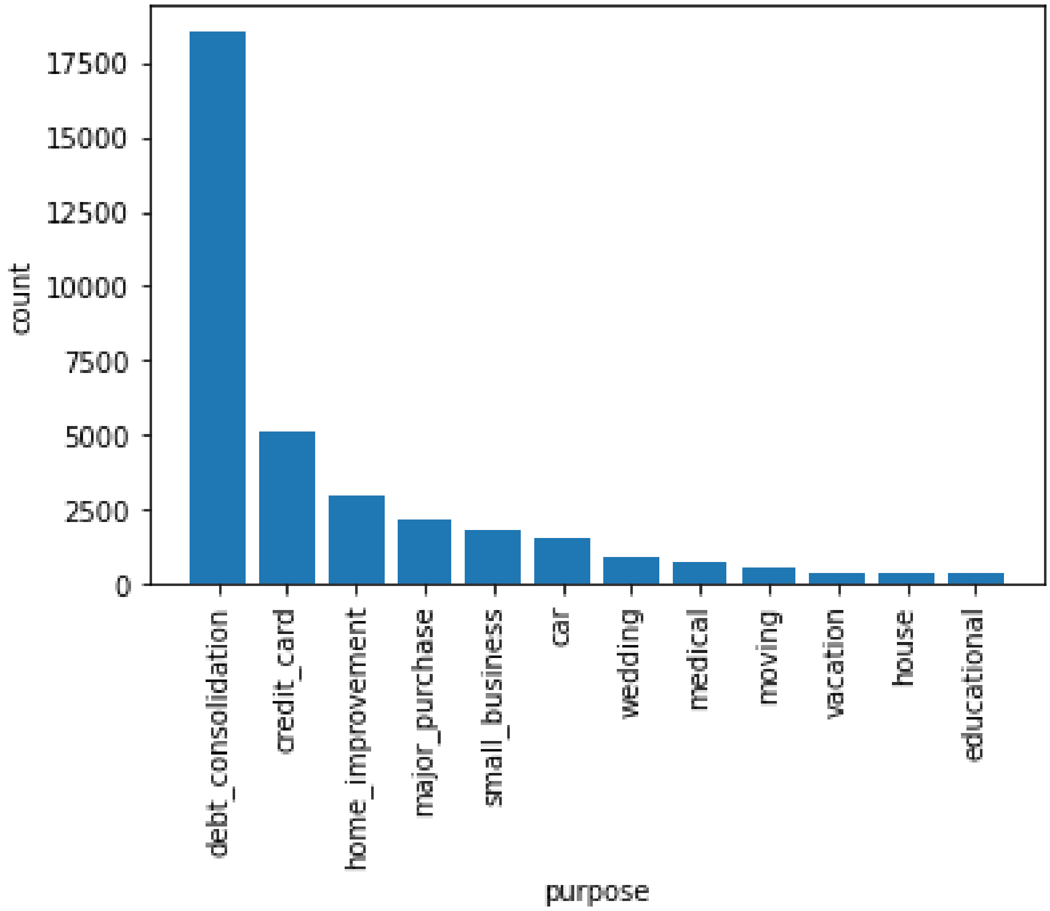
Categorical Variables

4. Loan Status



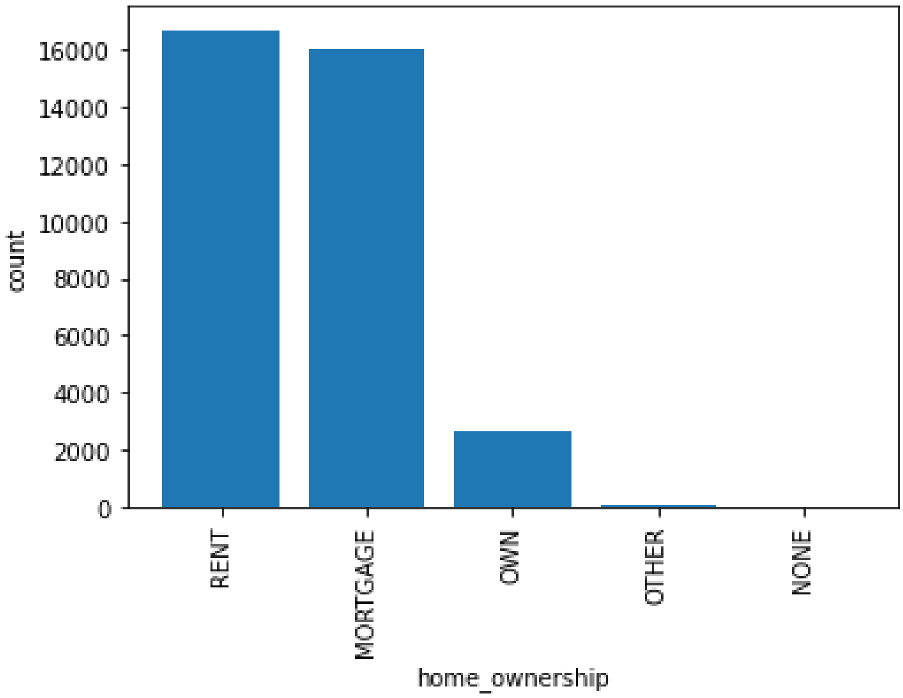
Insights: 5% of the applicants Charged off.

5. Purpose of loan



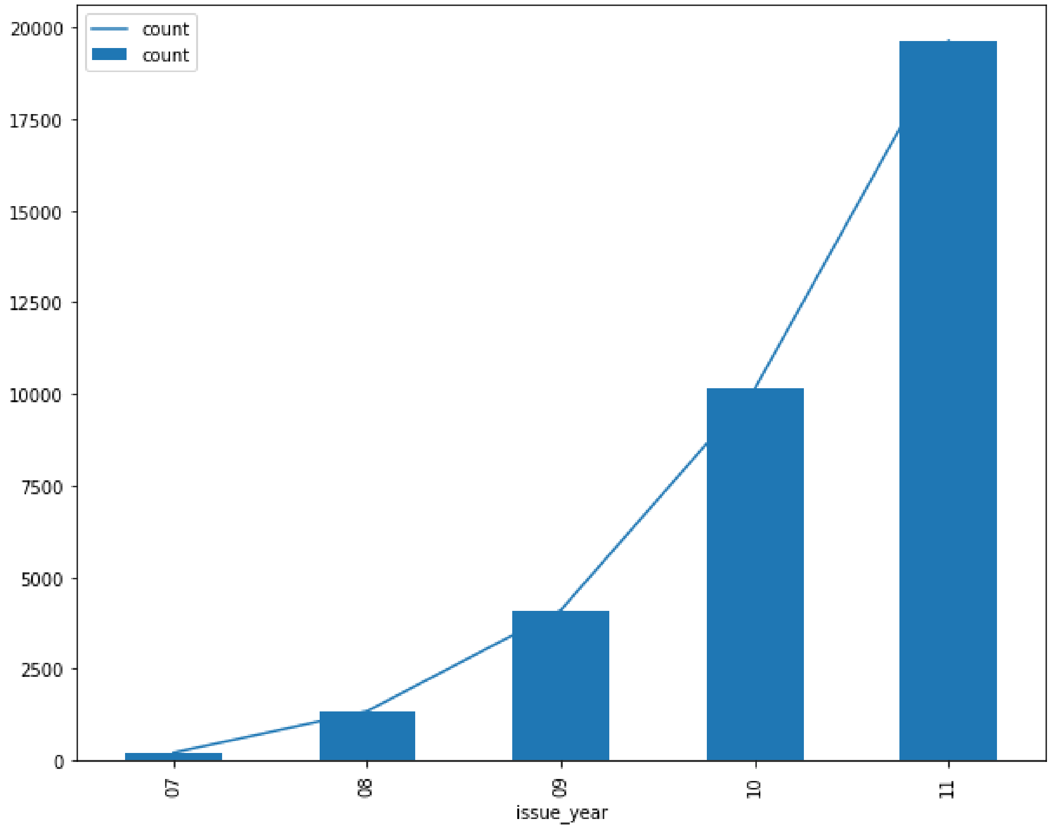
Insights: Approx 60% of the applicants applied loan for paying their other loans (Debt Consolidation)

6. Home Ownership wise Loan



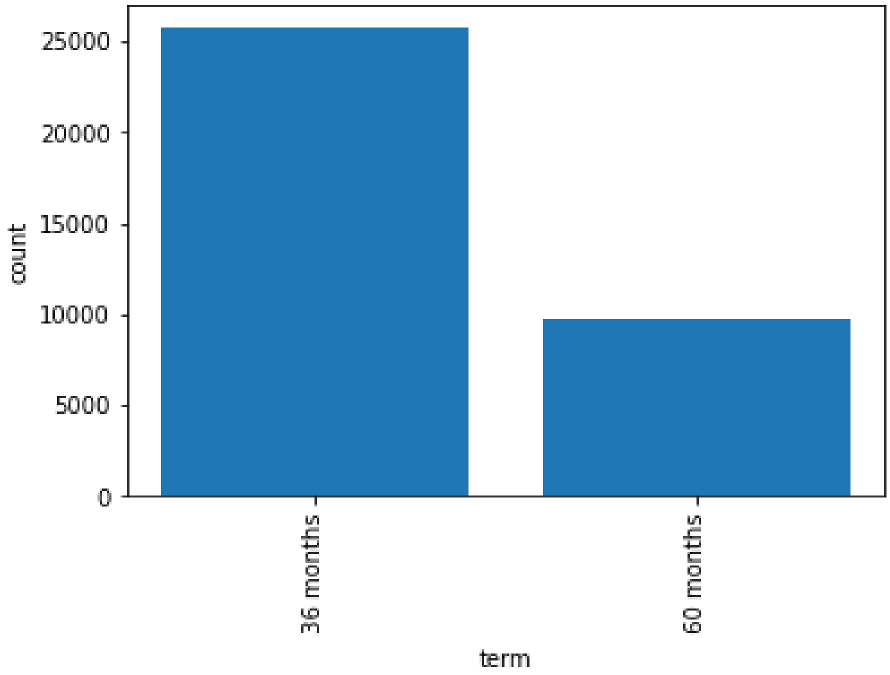
Insights: Most of applicants are living in rented home and mortagaged their home

7. Year wise Loan



Insights: loan applicants are increasing year on year, approx 47% of loan applicants received loans in 2011

8. Loan Term

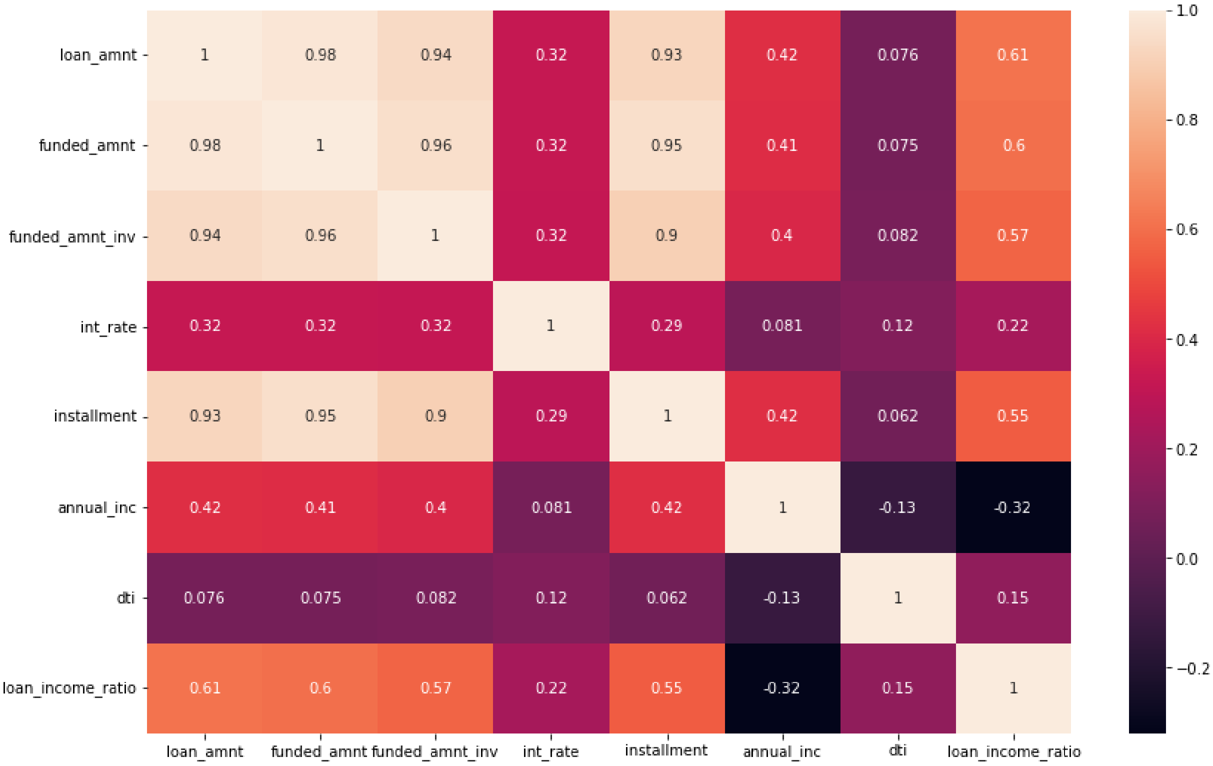


Insights: 70% of applicants applied loan for 36 months term period

5. Bivariate/Multivariate Analysis

Bivariate/Multivariate Analysis finds out the relationship between two/two or more variables. We can perform Bivariate/Multivariate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous.

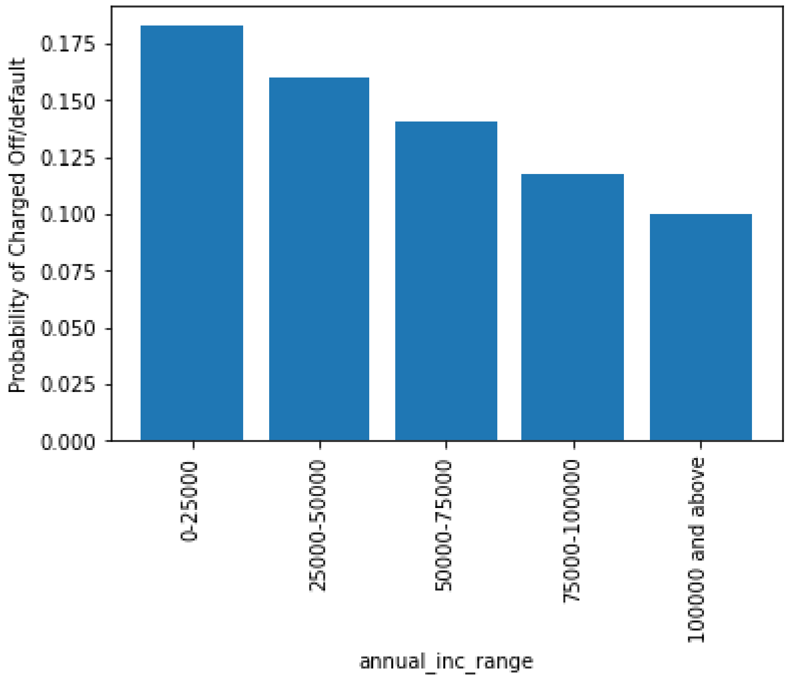
1. Correlation Matrix: All Continuous (Numeric) Variables



	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	loan_income_ratio
loan_amnt	1	0.98	0.94	0.32	0.93	0.42	0.076	0.61
funded_amnt	0.98	1	0.96	0.32	0.95	0.41	0.075	0.6
funded_amnt_inv	0.94	0.96	1	0.32	0.9	0.4	0.082	0.57
int_rate	0.32	0.32	0.32	1	0.29	0.081	0.12	0.22
installment	0.93	0.95	0.9	0.29	1	0.42	0.062	0.55
annual_inc	0.42	0.41	0.4	0.081	0.42	1	-0.13	-0.32
dti	0.076	0.075	0.082	0.12	0.062	-0.13	1	0.15
loan_income_ratio	0.61	0.6	0.57	0.22	0.55	-0.32	0.15	1

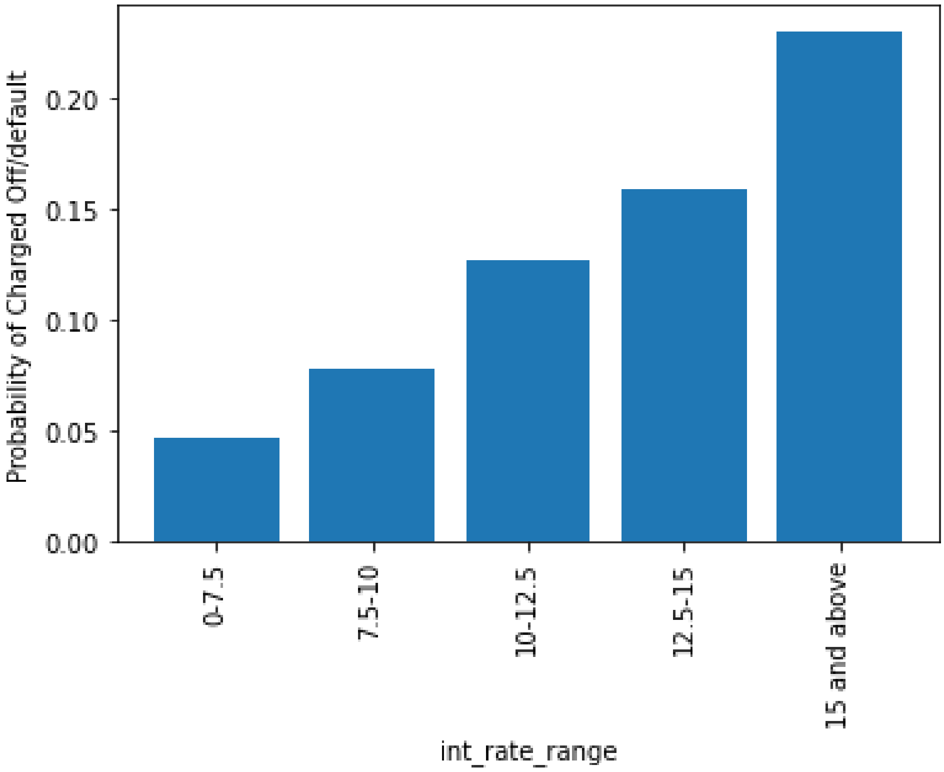
Insights: It is clear from the Heatmap that how 'loan_amnt','funded_amnt' & 'funded_amnt_inv' are closely interrelated.So we can take any one column out of them for our analysis

3. Annual Income Range vs Probability Charge Off



Insights: As the annual income is decreasing the probability that person will default is increasing with highest at (0 to 25000) salary bracket.

4. Interest rate Vs Probability Charge Off



Insights: As the interest rate is increasing the probability that person will default is increasing with highest of 9% at 15% & above bracket