Lending Club Case Study:

Exploratory Data Analysis(EDA)

Index

- ✓ Problem Statement
- ✓ Data Understanding
- ✓ Data Cleaning
- ✓ Analysis and Observations

Problem Statement

consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile. Two **types of risks** are associated with the bank’s decision:

•If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
•If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

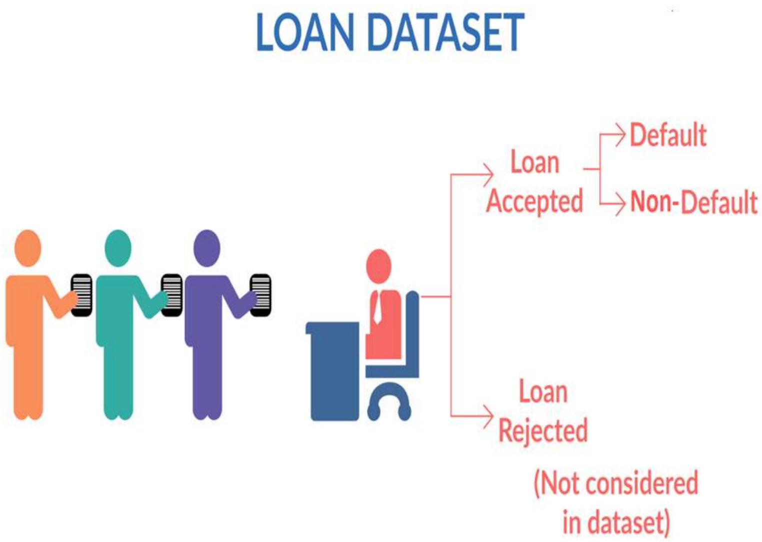
This case study is limited to data exploratory analysis. So no further action required to impute nulls or build and train model for prediction

In this case study, We will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

Data Understanding

Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
2. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
3. **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan



Data cleaning

We have imported required Python libraries for our case study and read the csv file into dataframe.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

The dataset contains 39717 rows with 111 columns we have taken appropriate action based on the column type.

1. Identified and deleted columns which has all null values (100%) because which the total columns reduces to 57.

```
loan_cust.dropna(axis=1,how='all',inplace=True)
```

2. dropped the columns which has more than 50% null values

```
loan_cust.drop(labels=['mths_since_last_record','next_pymnt_d’, ‘mths_since_last_delinq ’], axis=1,inplace=True)
```

3. removed columns based on Nan and zero values and also columns which are not required

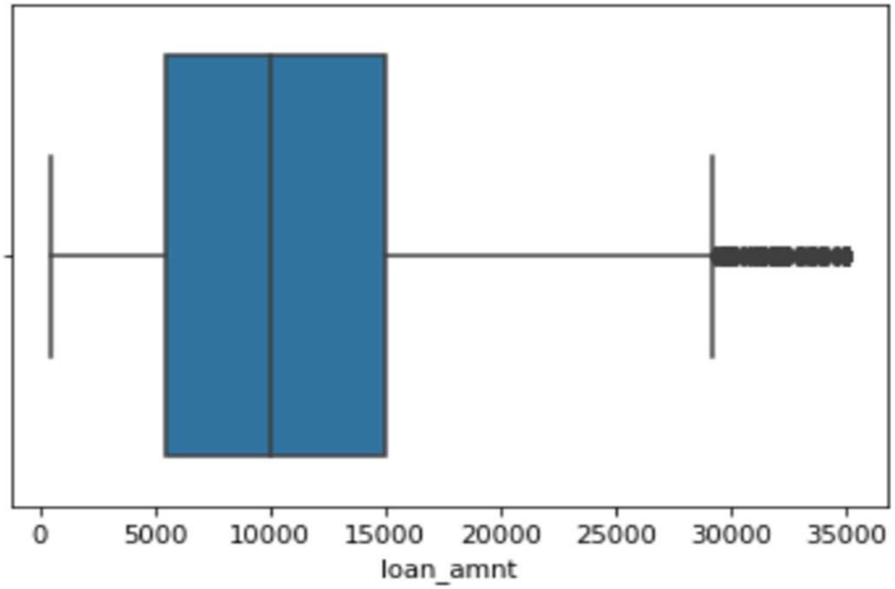
4. Removed string "months" - from term column
5. Column int_rate and revol_util - remove % symbol from the values
6. Treated the emp_length column to have only integers, so removed additional characters like +,<, years
7. Derived columns(Year and month) from issue_d variable
8. int_rate was still in “object” data type, We have converted it to float.

Data Analysis

Post the data cleaning we have used seaborn and matplotlib libraries to analyze the data through plots

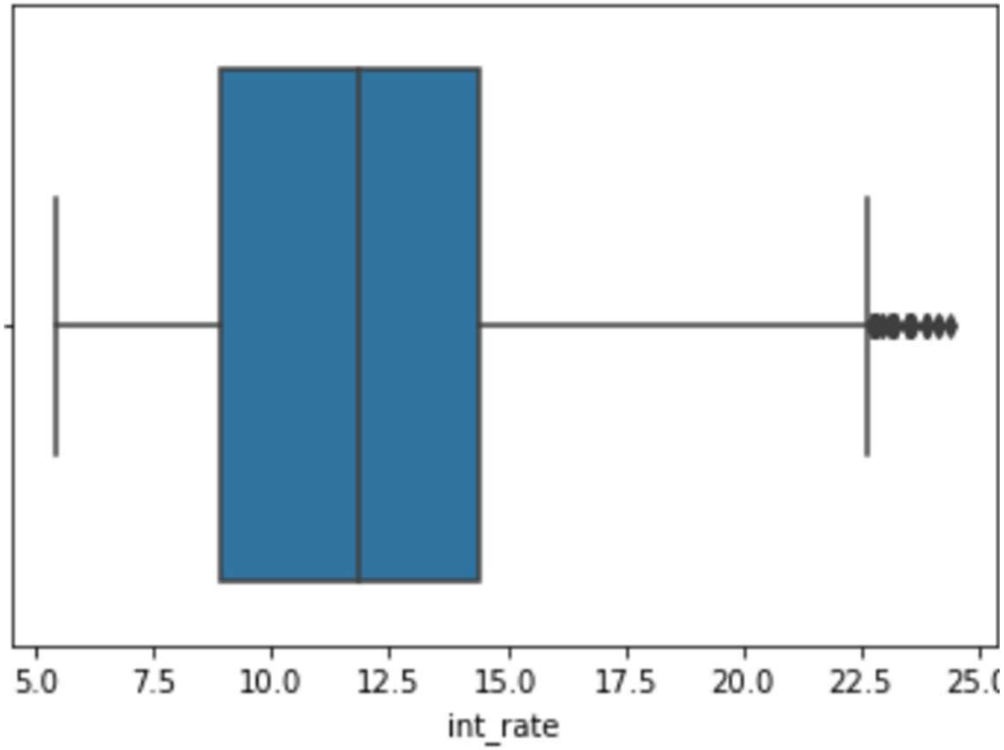
1. Univariate Analysis - variable- loan_amnt

From the boxplot of variable/column, loan_amnt 75% of customers has taken loan amount of 15000 and below.



2. Univariate Analysis - variable- int_rate

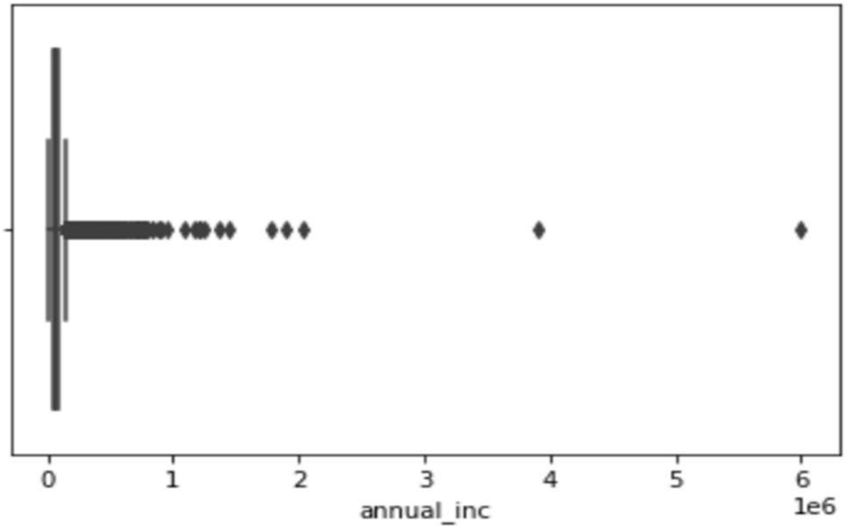
From the boxplot of variable/column, int_rate
It is clear that 75% of customers are paying less interest than 15%.



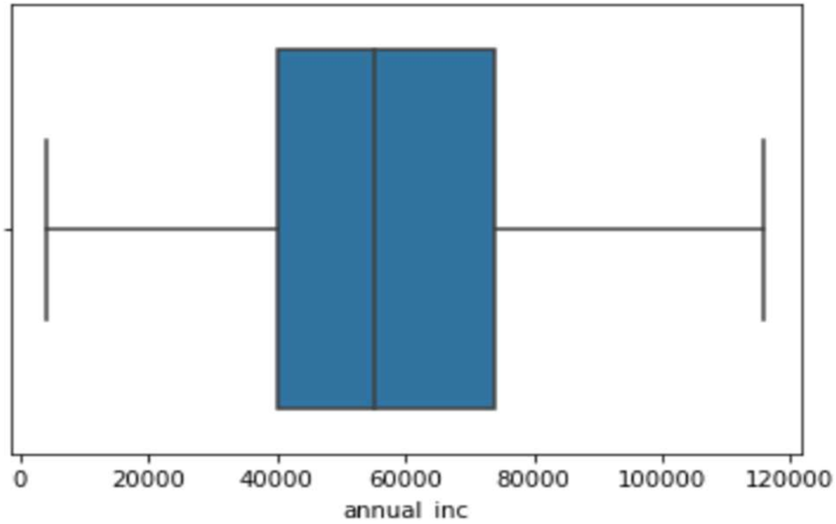
3. Univariate Analysis - variable- annual_inc

From the boxplot of variable/column, annul_inc

Initially we have identified that there were outliers. We have removed all the outliers above 90%



After removing the outliers , We can observe that customers are distributed across the income range and is equally distributed across all quartiles.

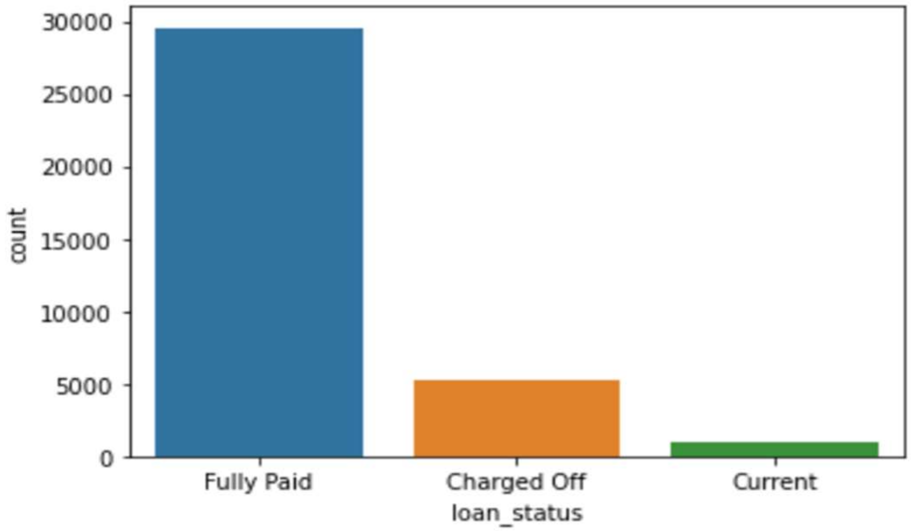


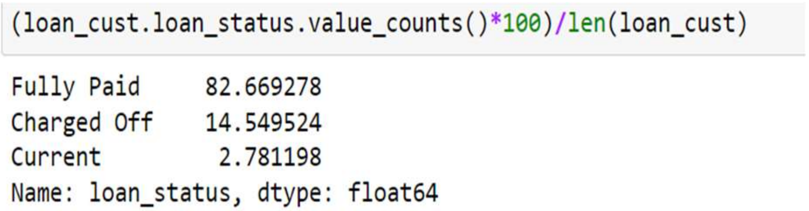
4. Univariate Analysis - variable- loan_status

From the boxplot of variable/column,   loan_status

We can observe that many of customer are fully paid their loan amount. We can ignore the customers whose loan_status is current.

83% of customers are completely paid their loan where are 14.5% customers are defaulted or partially paid the loan and defaulted.





```
(loan_cust.loan_status.value_counts()*100)/len(loan_cust)
```

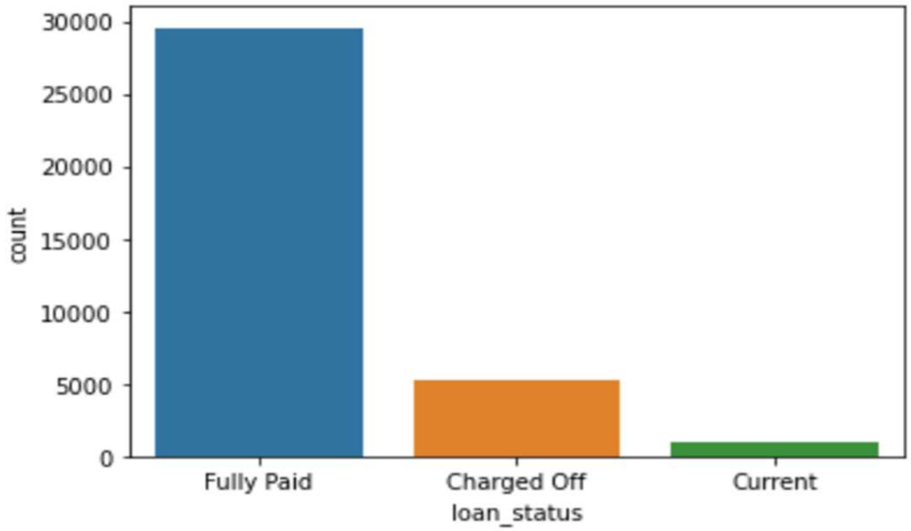
```
Fully Paid     82.669278
Charged Off    14.549524
Current         2.781198
Name: loan_status, dtype: float64
```

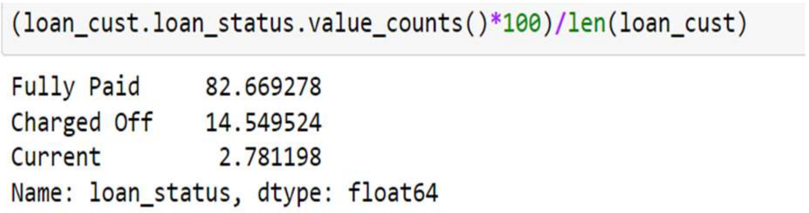
5. Univariate Analysis - variable- loan_status

From the boxplot of variable/column, loan_status

We can observe that many of customer are fully paid their loan amount. We can ignore the customers whose loan_status is current.

83% of customers are completely paid their loan where are 14.5% customers are defaulted or partially paid the loan and defaulted.



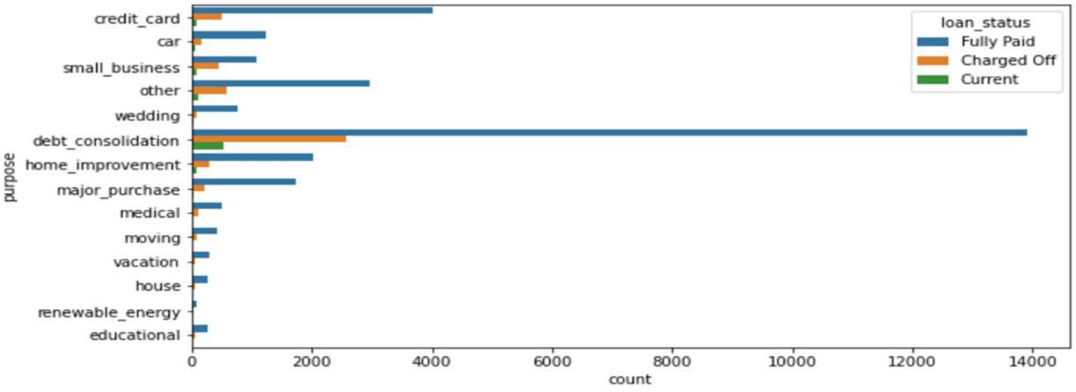


```
(loan_cust.loan_status.value_counts()*100)/len(loan_cust)
```

```
Fully Paid     82.669278
Charged Off    14.549524
Current         2.781198
Name: loan_status, dtype: float64
```

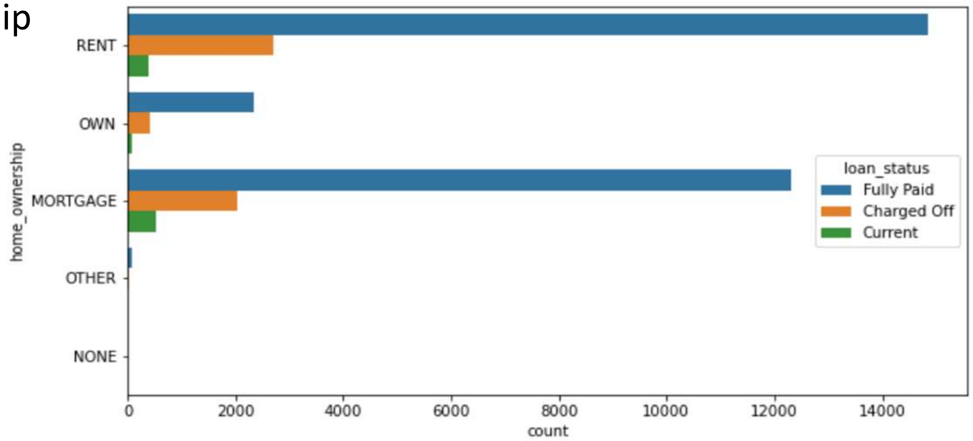
6. Bivariate Analysis - variable- loan_status and Purpose

More customers are opted loan for their debit consolidation, i.e, to repay other loans . the customers who paid their loan fully and most number of defaulted customers are also from debt_consolidation purpose only.



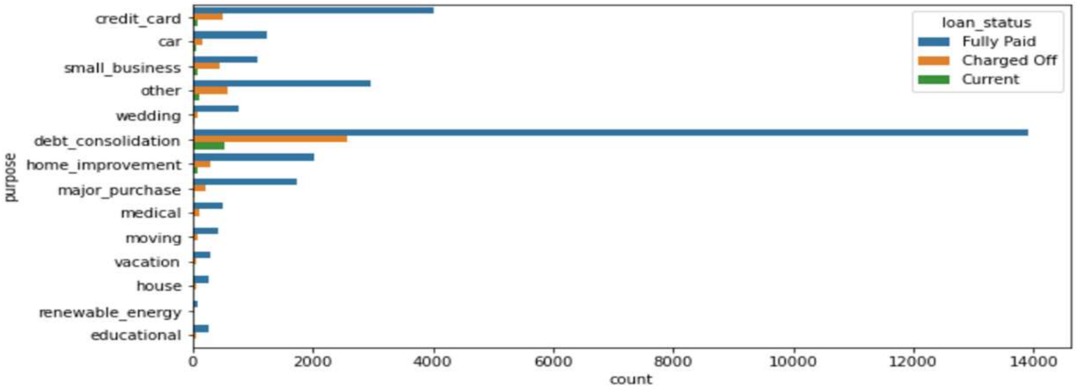
7. Bivariate Analysis - variable- loan_status Home_ownership

We can observer here that who ever are living in rented home, they have procured loan more than the other category in Home_ownership categories.



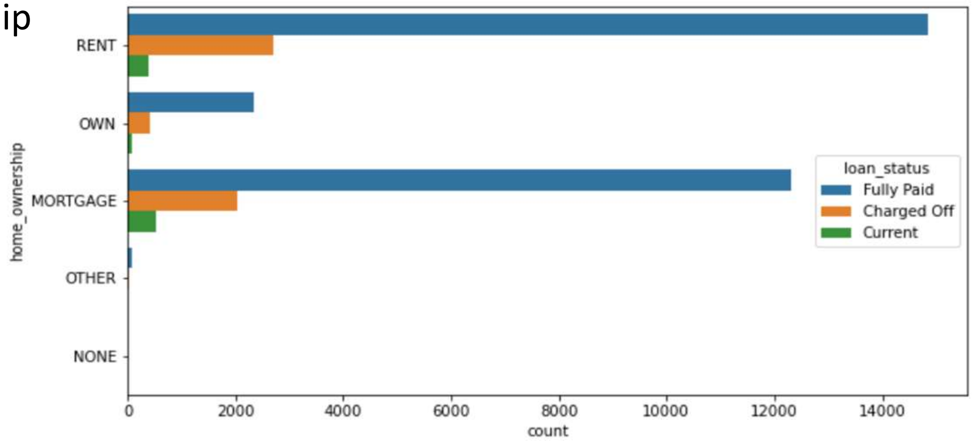
8. Bivariate Analysis - variable- loan_status and Purpose

More customers are opted loan for their debit consolidation, i.e, to repay other loans . the customers who paid their loan fully and most number of defaulted customers are also from debt_consolidation purpose only.



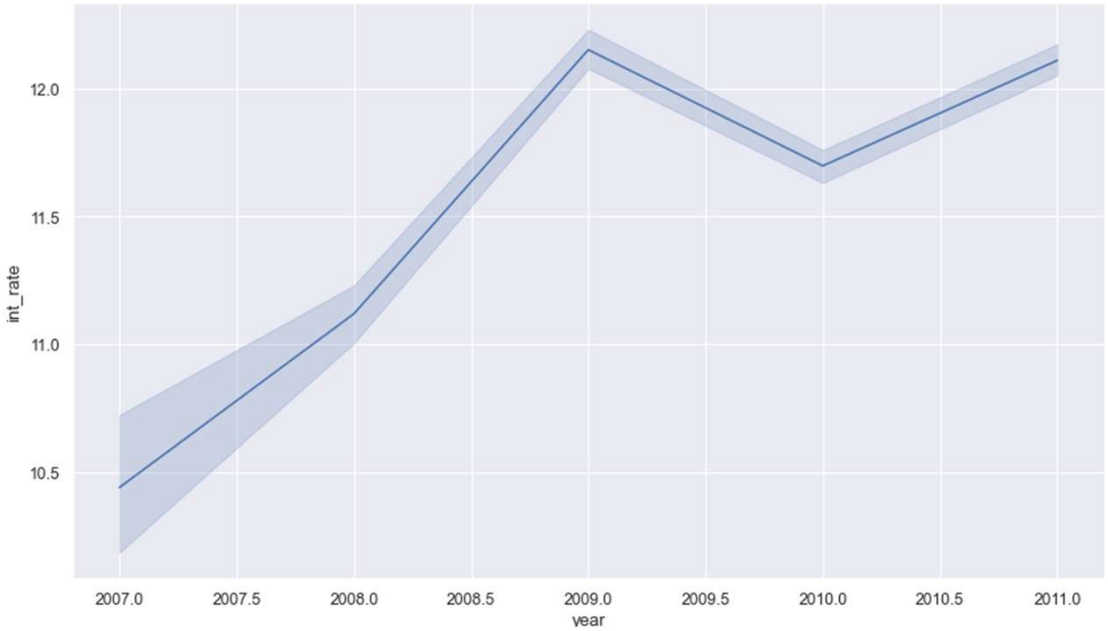
9. Bivariate Analysis - variable- loan_status Home_ownership

We can observer here that who ever are living in rented home, they have procured loan more than the other category in Home_ownership categories.



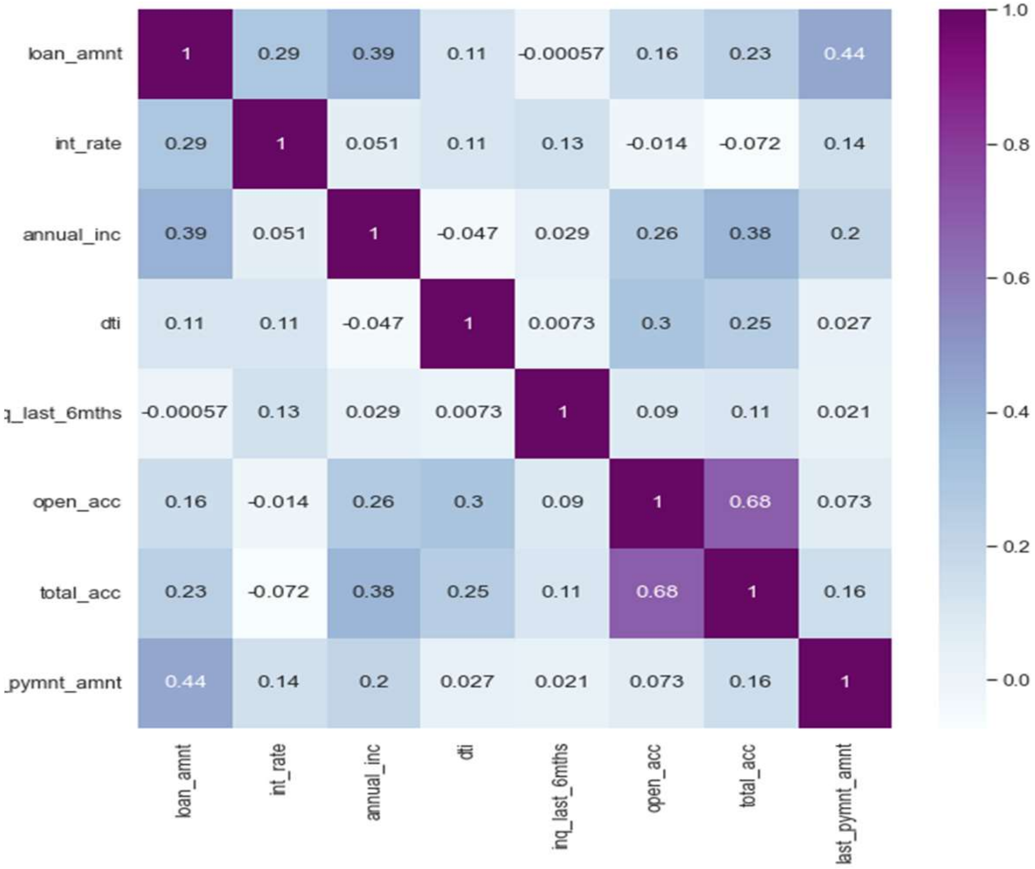
10. Bivariate Analysis - variable- interest rate and Year

We can notice that the interest rate is increased every year. But in 2010 the financial institute has reduced interest than previous year.

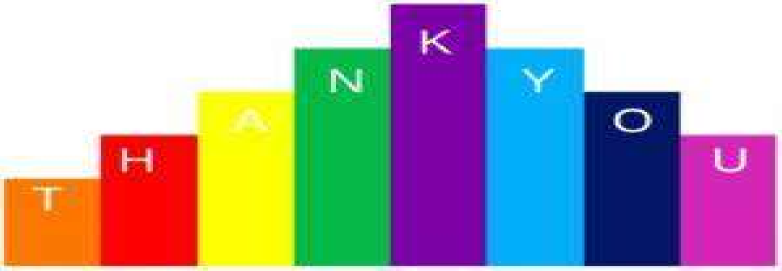


11. Multi variate Analysis - Heatmap

We have observed that the annual_income is positively correlated to loan_amount. If the customer has more annual income they are opting high amount of loan.
Total_acc is also positively correlated to Annual_income.



	loan_amnt	int_rate	annual_inc	dti	inq_last_6mths	open_acc	total_acc	last_pymnt_amnt
loan_amnt	1	0.29	0.39	0.11	-0.00057	0.16	0.23	0.44
int_rate	0.29	1	0.051	0.11	0.13	-0.014	-0.072	0.14
annual_inc	0.39	0.051	1	-0.047	0.029	0.26	0.38	0.2
dti	0.11	0.11	-0.047	1	0.0073	0.3	0.25	0.027
q_last_6mths	-0.00057	0.13	0.029	0.0073	1	0.09	0.11	0.021
open_acc	0.16	-0.014	0.26	0.3	0.09	1	0.68	0.073
total_acc	0.23	-0.072	0.38	0.25	0.11	0.68	1	0.16
_pymnt_amnt	0.44	0.14	0.2	0.027	0.021	0.073	0.16	1



THANK YOU