

LendingClub

EDA Case Study

Objectives

- Build risk profile to enable lending business
 - Develop recommendations on ability to repay loan, which will lead to loan approval, which in turn improve business to company.
- Identify key risks leading to loan default
 - Identify strong drivers (variables) for loan default.

Available Dataset

- Dataset contains information about past loan applicants.
- Columns with demographic and customer specific details have been removed from analysis as they do not help in predicting the business.
- Columns with more than 90% missing values
 - There are about 56 columns with more than 90% values missing.
- Rows that has more than 25% data missing has been removed.
- Fields with mixed data type has been converted to usable format (ex. Term field is a string, moved to integer).
- Columns that do not drive any impact has been removed.
- 16 columns have been identified and taken for analysis

RangeIndex: 39717 entries, 0 to 39716

Data columns (total 16 columns):

#	Column	Non-Null Count	Dtype
0	loan_amnt	39717 non-null	int64
1	funded_amnt	39717 non-null	int64
2	funded_amnt_inv	39717 non-null	float64
3	term	39717 non-null	object
4	int_rate	39717 non-null	object
5	installment	39717 non-null	float64
6	grade	39717 non-null	object
7	sub_grade	39717 non-null	object
8	emp_length	38642 non-null	object
9	home_ownership	39717 non-null	object
10	annual_inc	39717 non-null	float64
11	verification_status	39717 non-null	object
12	loan_status	39717 non-null	object
13	purpose	39717 non-null	object
14	addr_state	39717 non-null	object
15	dti	39717 non-null	float64

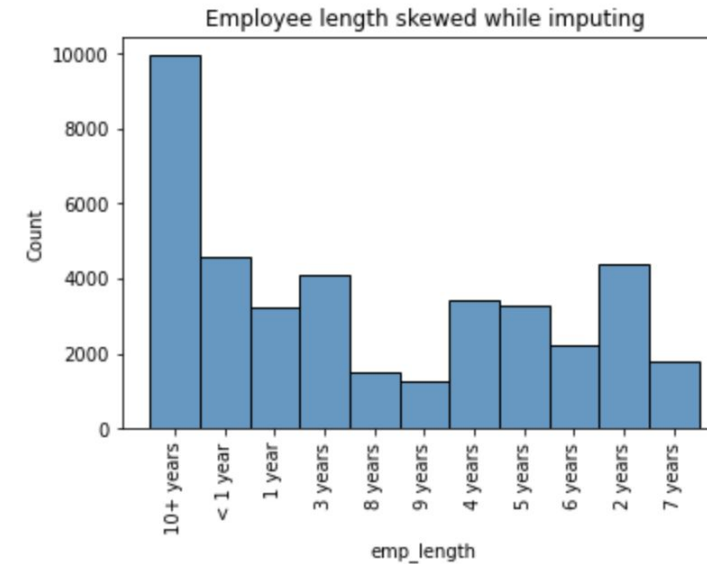
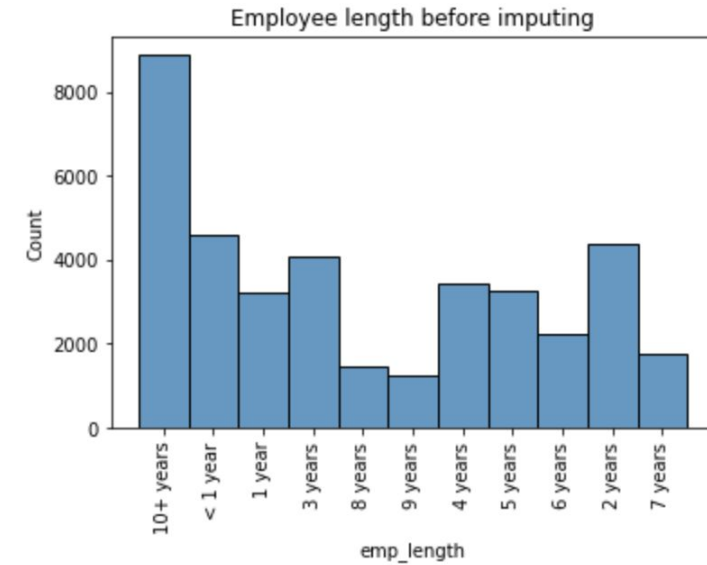
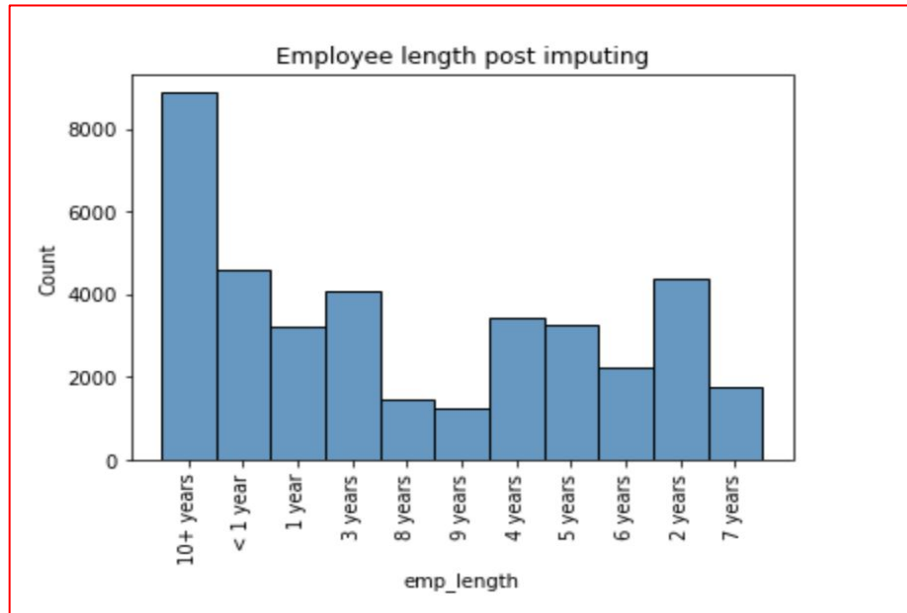
dtypes: float64(4), int64(2), object(10)

memory usage: 4.8+ MB

Variables

- Null values imputation

- emp_length variable need to be imputed as it had 1000+ null data.
- Using mode function, data was left skewed.
- Hence removed the null values.

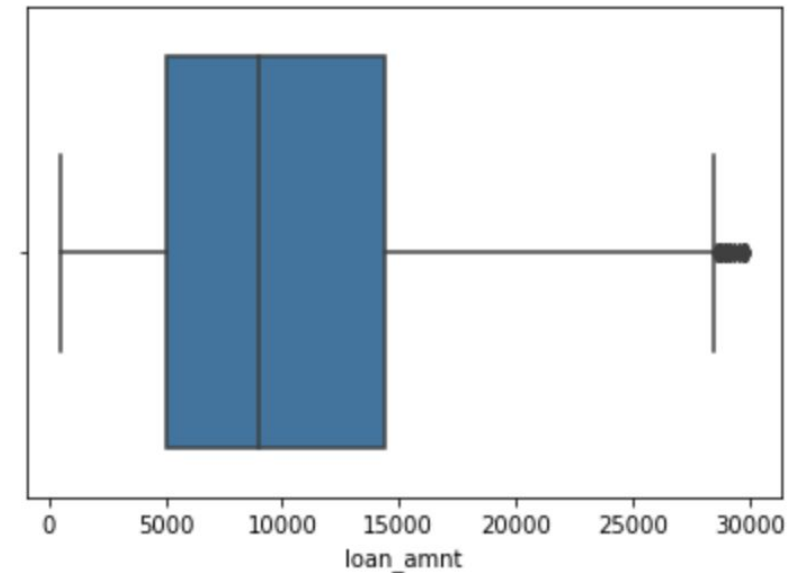
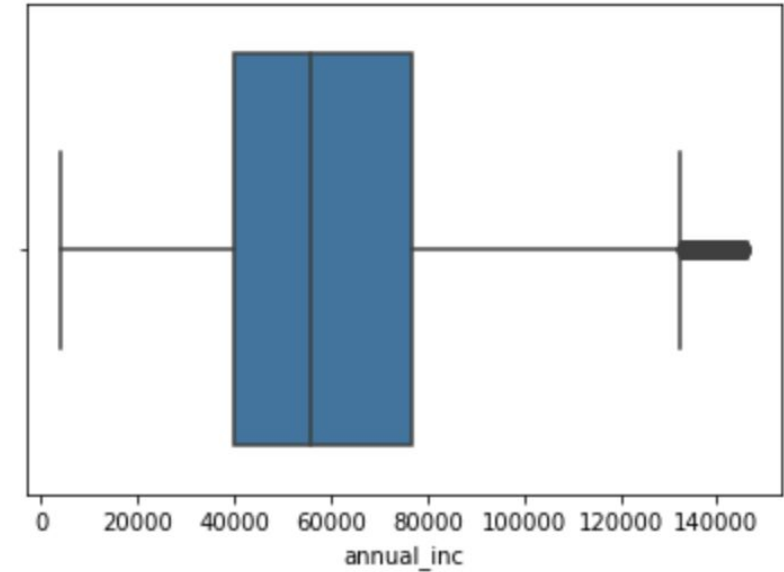


- Target Variables

- loan_default_status variable has been identified to classify target as fully paid or defaulted
- Current employees are not added for analysis as we do not have information regarding they have defaulted any payment.

- Outlier Treatment

- annual inc & loan_amnt variable are treated for outliers using IRQ. Hence data used for analysis represents the entire customer database.



- Driver Variables

- grade
- purpose
- emp_length
- home_ownership
- addr_state
- annual_inc
- funded_amnt

are identified to have business impact on determining if a borrower tends to pay fully or default.

Conclusions

Risk profile recommendations based on loan data set research

Risk Profile based on Home Ownership

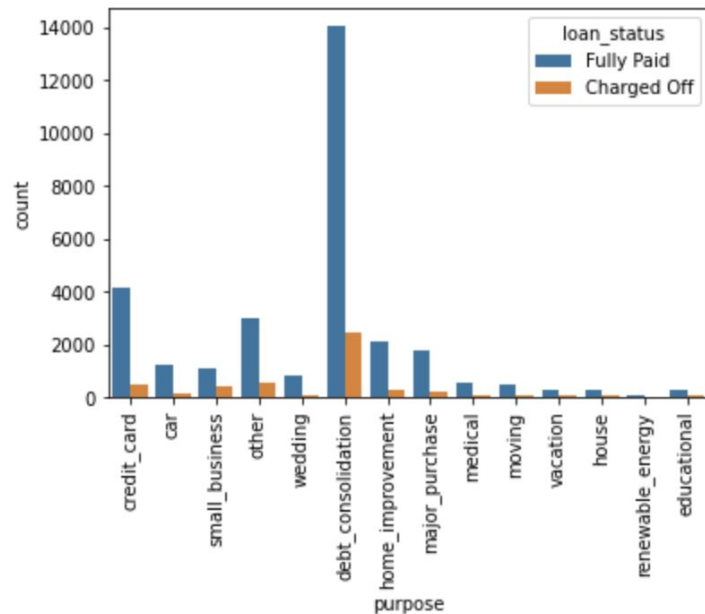
```
home_ownership  loan_status
MORTGAGE        Charged Off    1993
                Fully Paid    12860
NONE            Fully Paid         3
OTHER           Charged Off      17
                Fully Paid      76
OWN             Charged Off     379
                Fully Paid    2235
RENT            Charged Off    2609
                Fully Paid   14833
Name: loan_status, dtype: int64
```



An applicant is more likely to default loan re-payment, when type of residence is Rented or Mortgaged when compared to Owned.

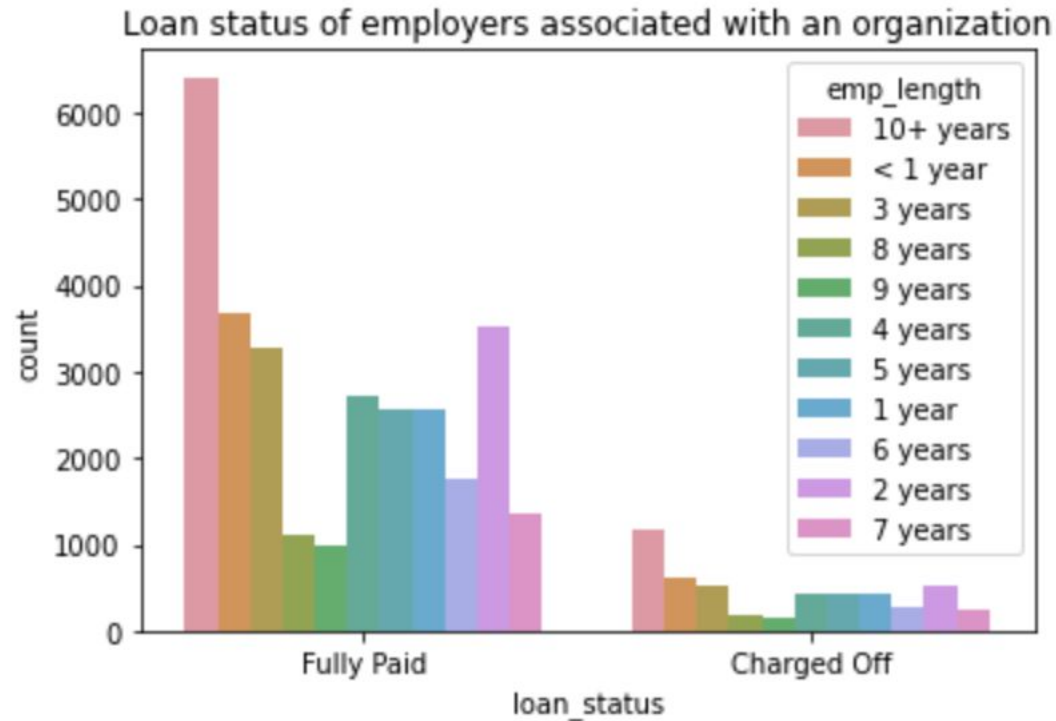
Risk Profile based on Loan purpose

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13]),  
[Text(0, 0, 'credit_card'),  
 Text(1, 0, 'car'),  
 Text(2, 0, 'small_business'),  
 Text(3, 0, 'other'),  
 Text(4, 0, 'wedding'),  
 Text(5, 0, 'debt_consolidation'),  
 Text(6, 0, 'home_improvement'),  
 Text(7, 0, 'major_purchase'),  
 Text(8, 0, 'medical'),  
 Text(9, 0, 'moving'),  
 Text(10, 0, 'vacation'),  
 Text(11, 0, 'house'),  
 Text(12, 0, 'renewable_energy'),  
 Text(13, 0, 'educational')])
```



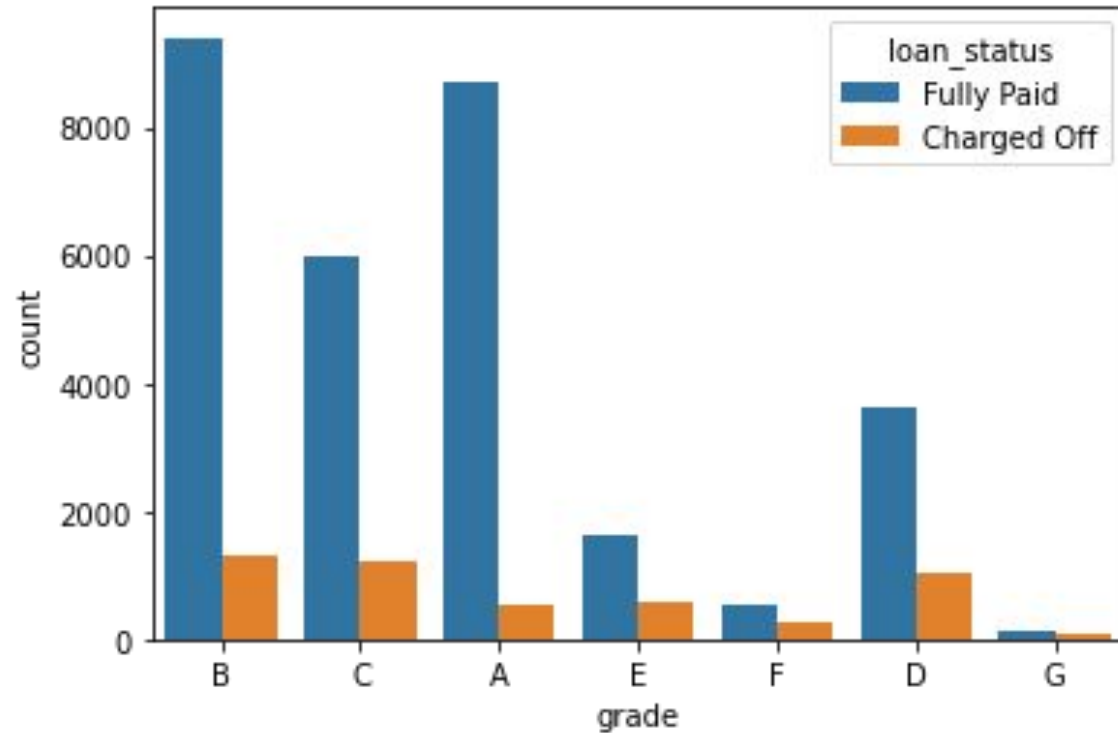
An applicant is more likely to default loan re-payment, when purpose of loan is for debt consolidation as opposed to any other purpose.

Risk Profile based on Employment Length



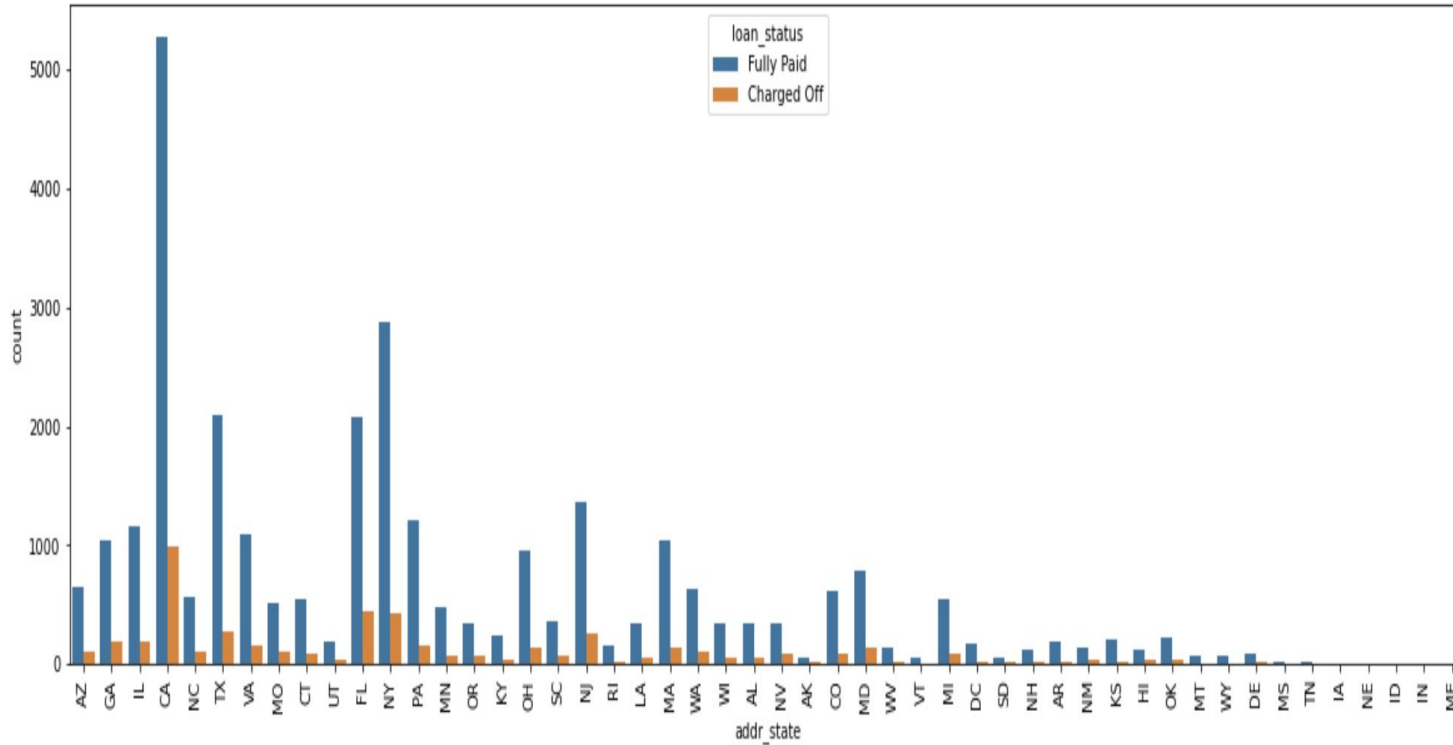
Applicant's that has worked at an employer for longer term, i.e., 10+ years tend to re-pay better.

Risk Profile based on Applicant's grade



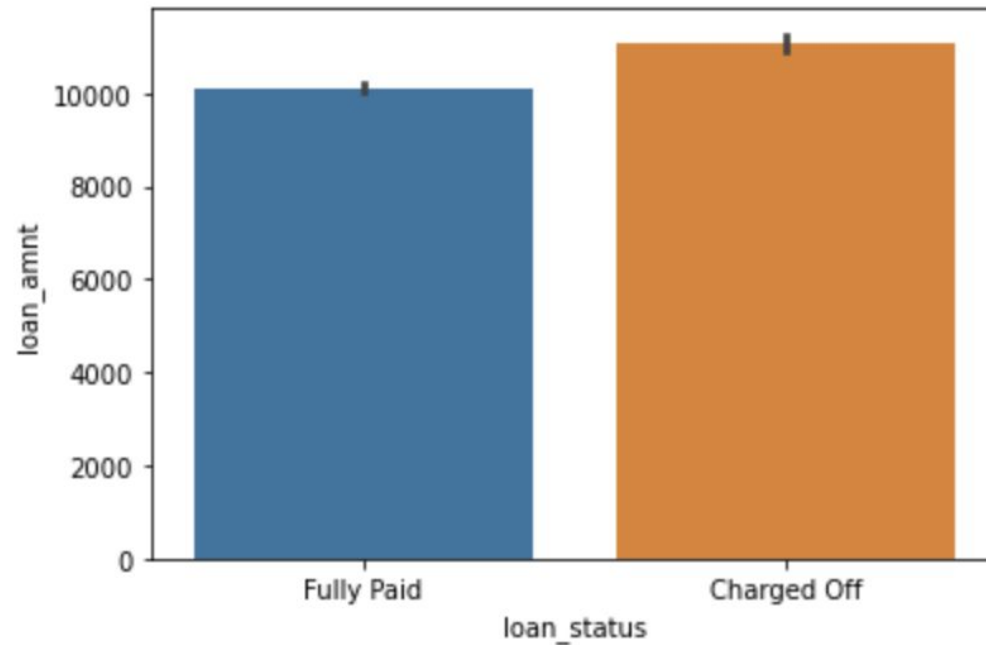
Applicants belonging to E,F,G grades are more likely to default on loan re-payment.

Risk Profile based on Home Ownership



Applicants from state of California has higher changes of a loan default than most other states.

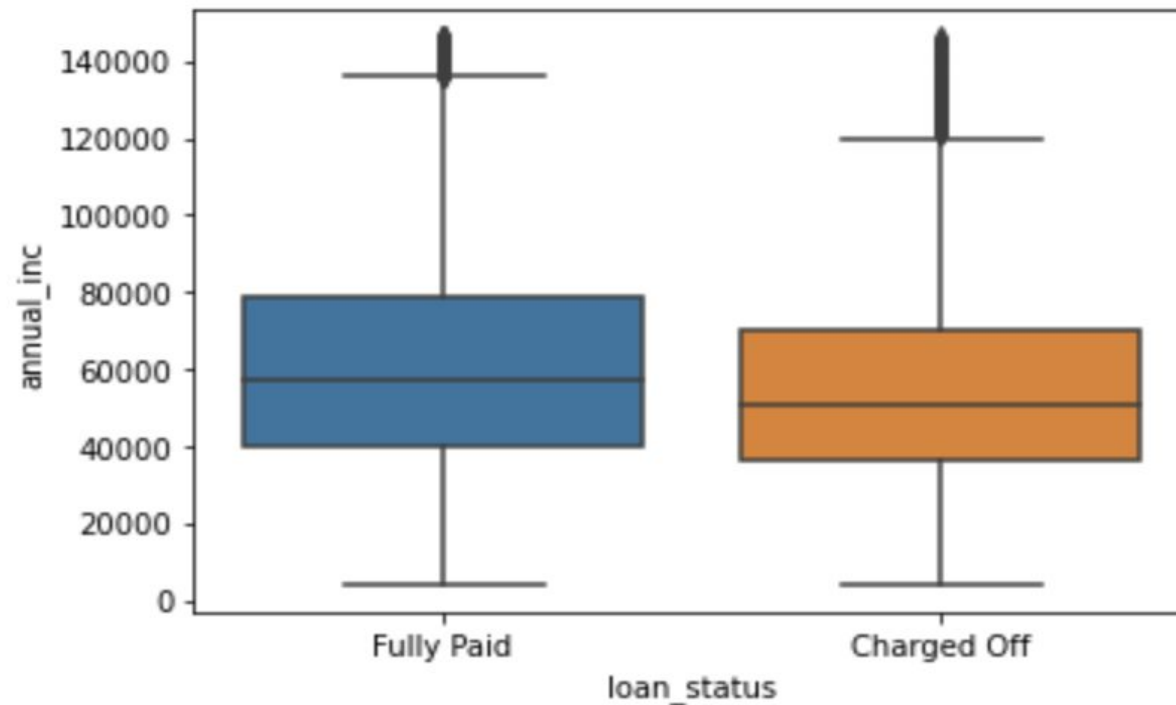
Risk Profile based on loan amount



Applicants requesting for higher loan amount tends to default more.

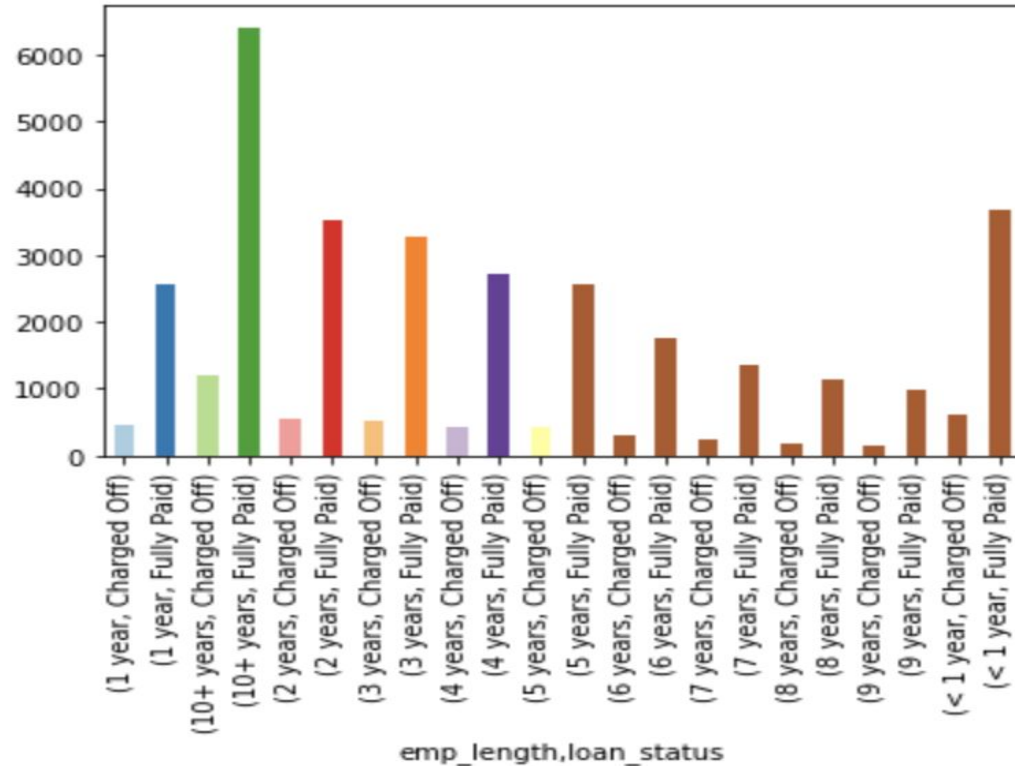
Lesser the loan amount, lesser chances of defaulting

Risk Profile based on annual income



Applicants having lesser annual income more likely to default.

Risk Profile based on funded amount vs emp_length



Applicants having lesser annual income more likely to default.

Thank you!

Univariate and Bivariate Analysis of variables in the
ipynb files