Lending Loan Case Study

Jeevitha Ravi
Anand Kumar Keshri

Problem Statement

Problem & Objective

Analysis of data using EDA (Exploratory Data Analysis) to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Benefits

Identification of risky loan applicants resulting in reducing the amount of credit loss.

Learning

Solving this assignment will give us an idea about how real business problems are solved using Exploratory Data Analysis (EDA). In this case study, apart from applying the techniques we have learnt in EDA, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Data Understanding

Data file - Contains loan data for the organization with following details
loan.csv

1. Rows/Columns - (39717, 111)

Metadata file - There is a separate file to provide the description of the columns for the data file.
Data_Dictionary.xlsx

Data type of the data files
{dtype('int64'): 13, dtype('float64'): 74, dtype('O'): 24}

Columns having null values more than 60%

- 57 columns

Columns having single value across the table

- 60 columns

Data cleaning and Manipulation

Data file - Contains loan data for the organization with following details
loan.csv

Cleaning - We need removed the following columns for various reasons listed below.

63 - Columns which have null values for 60% data or single values in the entire table

After manual look up following action has to be taken:

Unwanted column removal

13 other columns has to be removed

chargeoff_within_12_mths, collections_12_mths_ex_med, desc, emp_length, emp_title, id, il_util,open_rv_12m, initial_list_status, member_id, pub_rec_bankruptcies, tax_liens, url, zip_code

Unwanted rows removal from column

For the loan_status column we have removed the rows which contains the values ‘Current’

Data Analysis - 1 - loan_status

Percentage of data which is distributed in this two segment "['Fully Paid' 'Charged Off']

	counts	per	per100
Fully Paid	32950	0.854136	85.4%
Charged Off	5627	0.145864	14.6%

Current defaulter is ~15%

Data Analysis - 2 - annual_inc

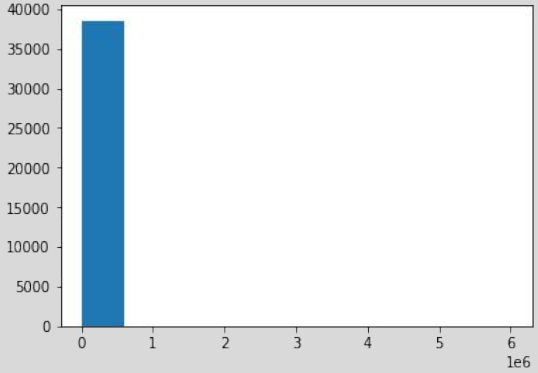
Understanding of the annual income

- median = 58868.0 max = 6000000.0
- It has been identified that the outlier in annual_inc is more than 200000

Conclusion

Annual income has very less impact on the charged off cases. 0.75 quantile for both 'Charged Off' and 'Fully Paid' are

```
loan_status
Charged Off  0.75    74000.0
Fully Paid   0.75    81085.0
Name: annual_inc, dtype: float64
```



Data Analysis - 3 - home_ownership

Understanding of the ownership of the home variable

home_ownership	loan_status	
MORTGAGE	Charged Off	13.671347
	Fully Paid	86.328653
NONE	Fully Paid	100.000000
OTHER	Charged Off	18.367347
	Fully Paid	81.632653
OWN	Charged Off	14.890756
	Fully Paid	85.109244
RENT	Charged Off	15.362554
	Fully Paid	84.637446

Conclusion
The above data clarifies that there is a very less change in the percentage contribution to the loan status by home_ownership variable

Data Analysis - 4 - verification_status

Understanding of the ownership of the home variable

		loan_status	percentage
verification_status	loan_status		
Not Verified	Charged Off	2142	12.830957
	Fully Paid	14552	87.169043
Source Verified	Charged Off	1434	14.818642
	Fully Paid	8243	85.181358
Verified	Charged Off	2051	16.803212
	Fully Paid	10155	83.196788

Conclusion

Not Verified source has lesser charged off percentage

Data Analysis - 5 - purpose, addr_state

purpose

Conclusion
purpose has highest number charged off percentage for 'renewable_energy', 'small_business' more than 18%

addr_state

Conclusion
No big impact

Data Analysis - 6 - grade/subgrade

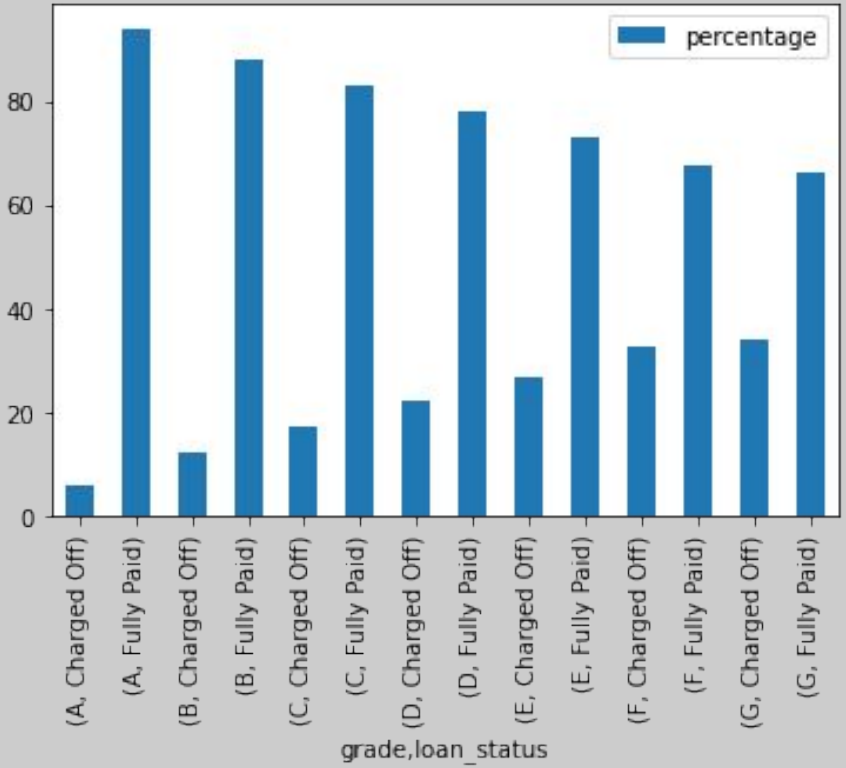
grade impact on charge off cases

grade	loan_status	loan_status_count	percentage
A	Charged Off	602	5.993031
	Fully Paid	9443	94.006969
B	Charged Off	1425	12.205567
	Fully Paid	10250	87.794433
C	Charged Off	1347	17.194281
	Fully Paid	6487	82.805719
D	Charged Off	1118	21.986234
	Fully Paid	3967	78.013766
E	Charged Off	715	26.849418
	Fully Paid	1948	73.150582
F	Charged Off	319	32.684426
	Fully Paid	657	67.315574
G	Charged Off	101	33.779264
	Fully Paid	198	66.220736

Conclusion

grade - Grade D E G and F has highest number of charged offs more than 21%
sub_grade - lower the sub_grade, more the charged off percentage

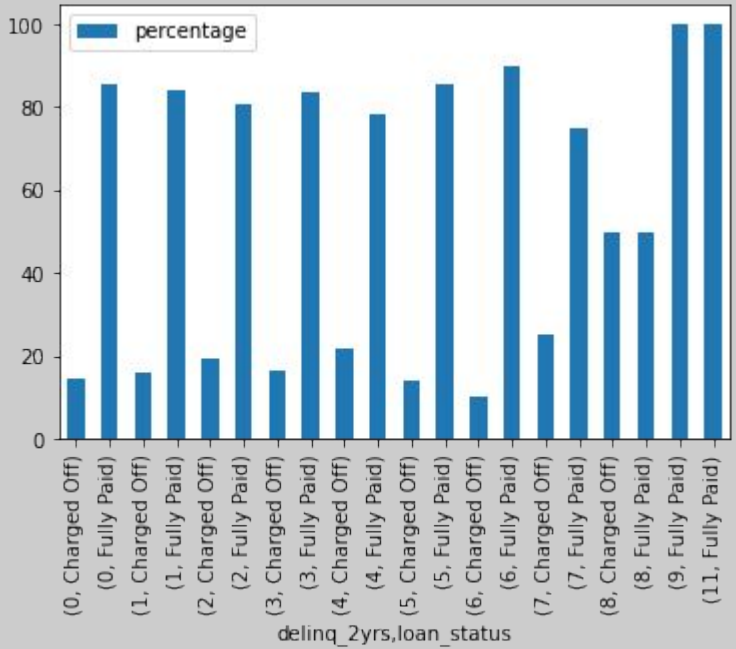
Data Analysis - 7 - grade using pivot table



Data Analysis - 8 - delinq_2yrs

delinq_2yrs

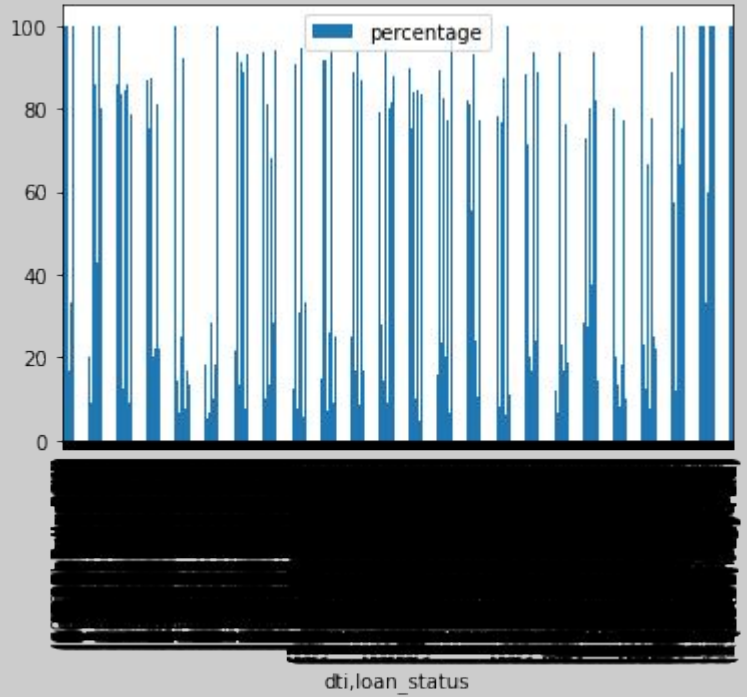
Conclusion
For more than 1 delinquency incidents, customer charged off percentage increases by 4%



Data Analysis - 9 - dti

dti

Conclusion
No conclusion

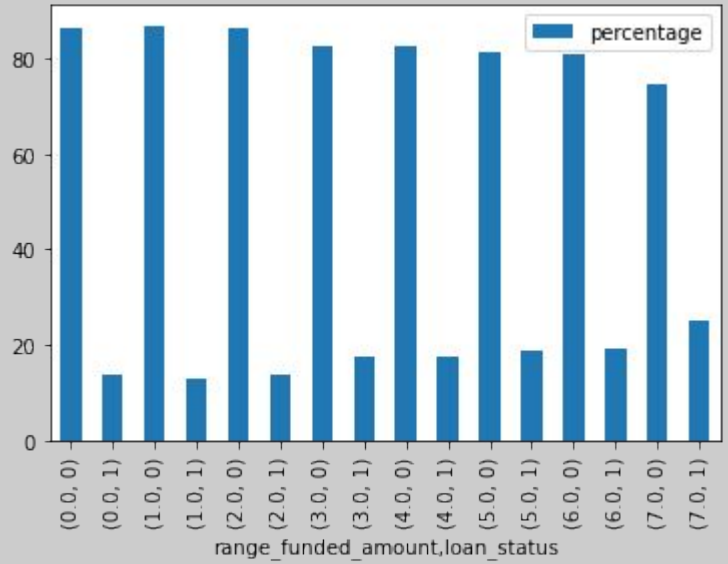


Data Analysis - 10 - funded_amount

funded_amount

Conclusion

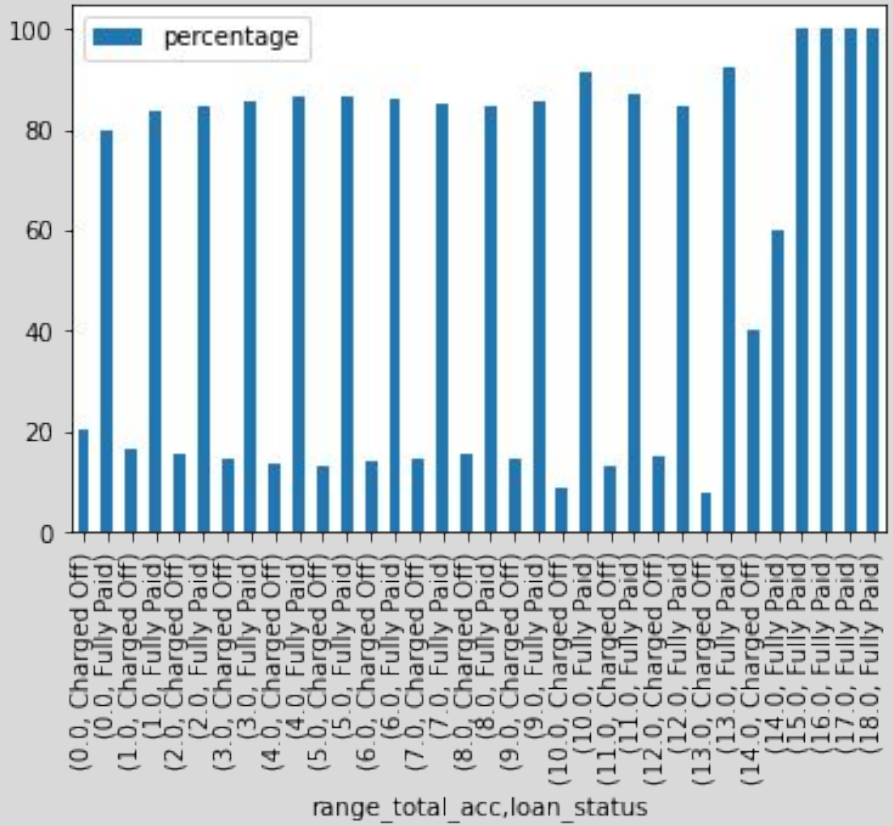
Charge off percentage increases with the increase in funded_amnt it is more than 18% if the funded amount goes beyond 25K



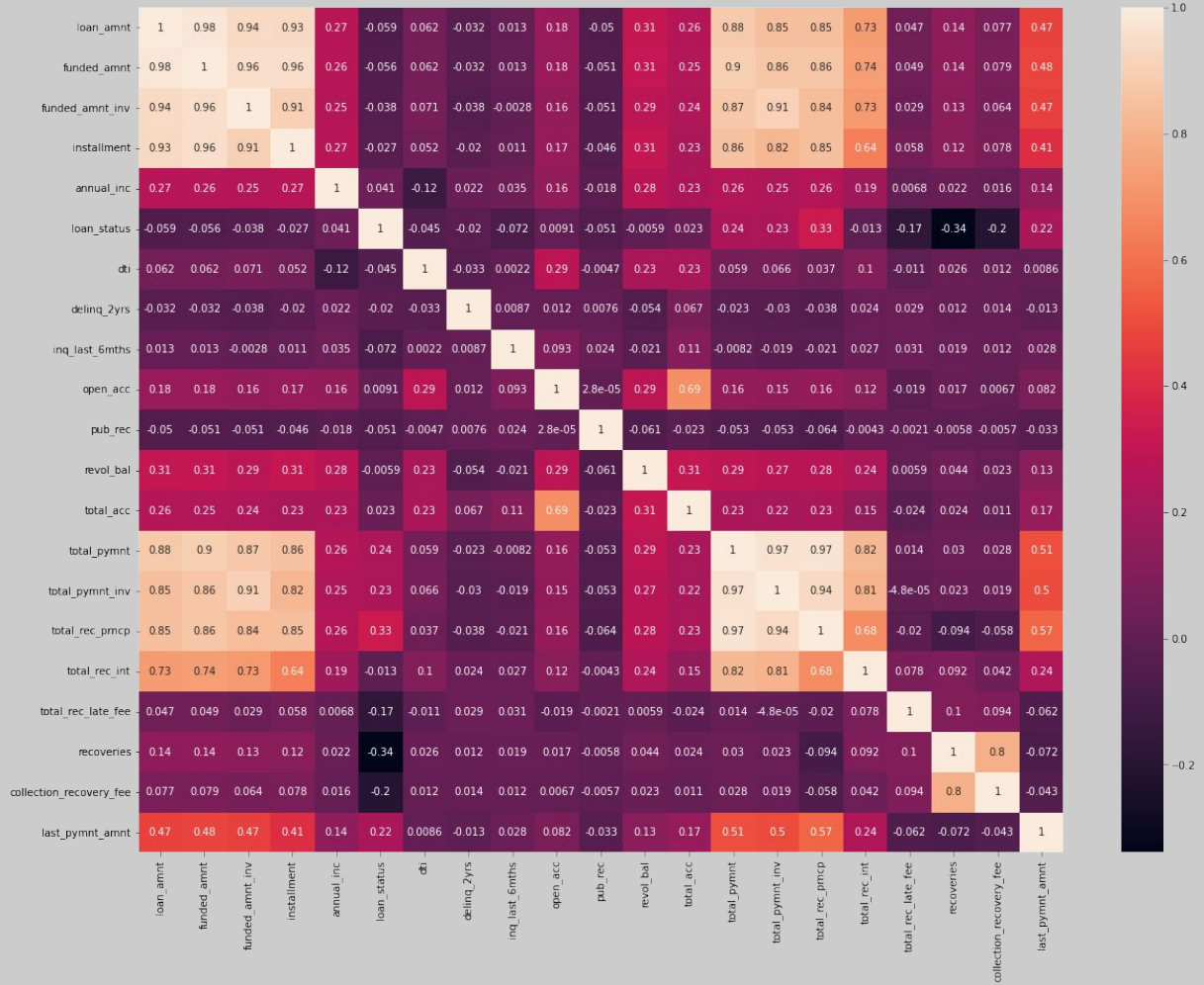
Data Analysis - 11 - total_acc

total_acc

Conclusion
Charge off percentage increases with the (0-5)
in total_acc it is more than 15% which is base average



Data Analysis - 12 - correlation matrix



Conclusion

- Annual income has very less impact on the charged off cases
- home_ownership has very less change in the percentage contribution to the loan status
- Not Verified source in verification_status has lesser charged off percentage
- purpose has highest number charged off percentage for 'renewable_energy', 'small_business' more than 18%
- Addr_state has no impact
- Grade D E G and F has highest number of charged offs more than 21%
- sub_grade - lower the sub_grade, more the charged off percentage
- For more than 1 delinquency incidents, customer charged off percentage increases by 4%
- Charge off percentage increases with the increase in funded_amnt it is more than 18% if the funded amount goes beyond 25K
- Charge off percentage increases with the (0-5) in total_acc it is more than 15% which is base average