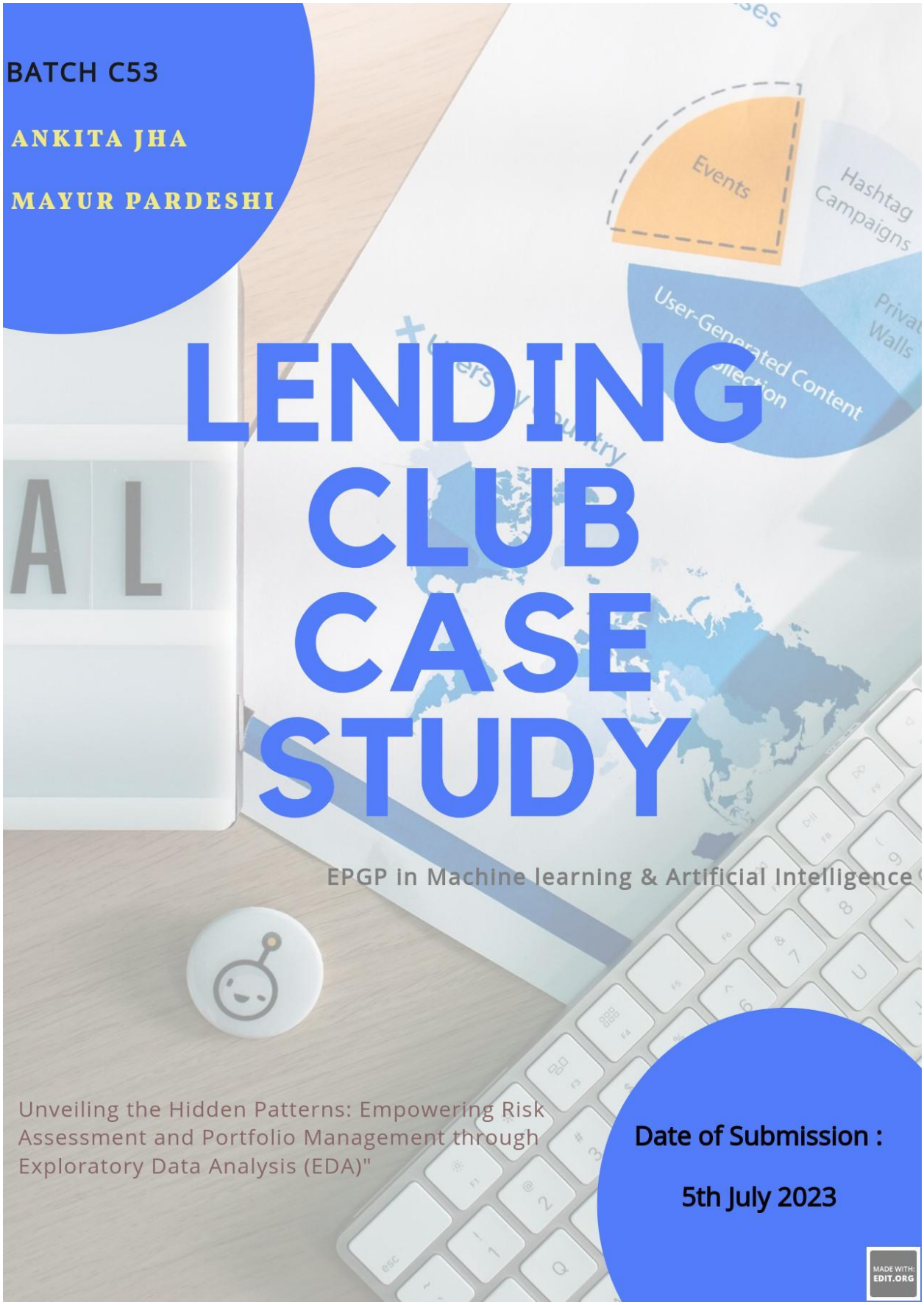


BATCH C53

ANKITA JHA

MAYUR PARDESHI

LENDING CLUB CASE STUDY

EPGP in Machine learning & Artificial Intelligence

Unveiling the Hidden Patterns: Empowering Risk Assessment and Portfolio Management through Exploratory Data Analysis (EDA)"

Date of Submission :

5th July 2023

Table of Contents

1. Problem Statement

The project revolves around a consumer finance company specializing in lending various types of loans to urban customers. The company receives loan applications and makes decisions based on the applicant's profile. The project aims to use exploratory data analysis (EDA) to understand the influence of consumer attributes and loan attributes on the likelihood of loan default.

2. Analysis approach

The analysis approach includes below steps

2.1. Analyzing given data

Below points are observed while analyzing data

a. Shape and size of matrix
   Row count - 39717
   Column count - 111

b. Columns with null value
c. Columns with incorrect data type
d. Understanding different attributes with the help of Dictionary provided along with the data

2.2. Data cleaning

Below steps were taken for data cleaning

2.2.1. Removing columns with all null values

Below 54 columns were removed as part of it

S. No	Column names
1	mths_since_last_major_derog
2	annual_inc_joint
3	dti_joint
4	verification_status_joint
5	tot_coll_amt
6	tot_cur_bal
7	open_acc_6m
8	open_il_6m

9	open_il_12m
10	open_il_24m
11	mths_since_rcnt_il
12	total_bal_il
13	il_util
14	open_rv_12m
15	open_rv_24m
16	max_bal_bc
17	all_util
18	total_rev_hi_lim
19	inq_fi
20	total_cu_tl
21	inq_last_12m
22	acc_open_past_24mths
23	avg_cur_bal
24	bc_open_to_buy
25	bc_util
26	mo_sin_old_il_acct
27	mo_sin_old_rev_tl_op
28	mo_sin_rcnt_rev_tl_op
29	mo_sin_rcnt_tl
30	mort_acc
31	mths_since_recent_bc
32	mths_since_recent_bc_dlq
33	mths_since_recent_inq
34	mths_since_recent_revol_delinq
35	num_accts_ever_120_pd
36	num_actv_bc_tl
37	num_actv_rev_tl
38	num_bc_sats
39	um_bc_tl
40	num_il_tl
41	num_op_rev_tl
42	num_rev_accts
43	num_rev_tl_bal_gt_0
44	num_sats
45	num_tl_120dpd_2m
46	num_tl_30dpd
47	num_tl_90g_dpd_24m
48	num_tl_op_past_12m
49	pct_tl_nvr_dlq
50	percent_bc_gt_75
51	tot_hi_cred_lim
52	total_bal_ex_mort
53	total_bc_limit

54	total_il_high_credit_limit

2.2.2. Remove columns with same values

S No	Column Name	Value
1	application_type	INDIVIDUAL
2	collections_12_mths_ex_med	0.0
3	policy_code	1
4	acc_now_delinq	0
5	chargeoff_within_12_mths	0.0
6	delinq_amnt	0
7	tax_liens	0
8	pymnt_plan	n
9	initial_list_status	f

2.2.3. Remove columns with same values

S. No.	Column Name	Reason
1	desc	It had various different values, which could be categorized into categorical value
2	url	It is unique and could not be used for group by
3	Id	It is unique and could not be used for group by
4	Member_id	It is unique and could not be used for group by
5	title	Purpose column has details and is better for aggregation

6	Zip_code	It has only first three digits of zip-code.

2.2.4. Below columns were removed as they had more than 50% of data missing

S. No.	Column Name
1	last_credit_pull_d
2	pub_rec_bankruptcies
3	emp_length
4	last_pymnt_amnt
5	revol_util
6	emp_title
7	last_pymnt_d

2.2.5. Remove data not required

S. No.	Loan_status	Reason
1	Current	The loan is in progress, hence not useful to find defaulter traits And it contributes to only 3% of the total data

2.2.6. Data type correction

S. No.	Column Name	Reason

1	term	Remove the string **month** from column values, and change it to INT data type
2	int_rate	Remove **%** from column values, and change it to float data type
3	revol_util	Remove **%** from column values, and change it to float data type
4	emp_length	Remove **experience**, **>**, **<** from column values, and change it to float data type
5	funded_amnt_inv	Standardize precision. Four decimal point converted to two

2.2.7. Removing Outliers

S. No.	Column Name	Reason
1	annual_inc	Value more than Q3 + 1.5*IQR And Value less than Q1 - 1.5*IQR Is removed

2.3. Analyze data after cleaning

Once the data is cleaned by following the above steps, it is analyzed using below steps:-

1. Univariate analysis
2. Bivariate analysis
3. Derived matrix

The result of this analysis is elaborated in next section – Analyze results

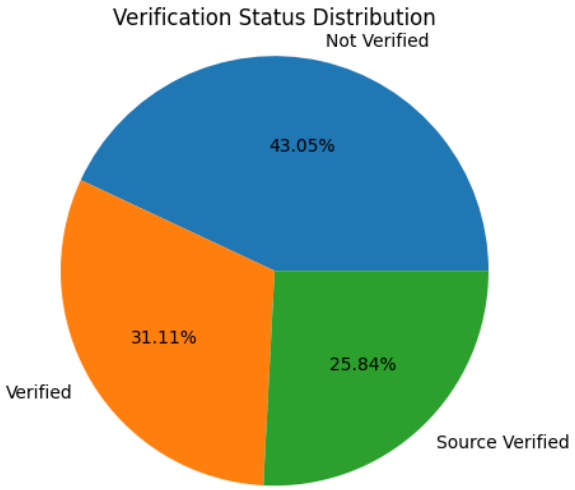
3. Analyze results

3.1. Univariate analysis

3.1.1. Categorical – Unordered

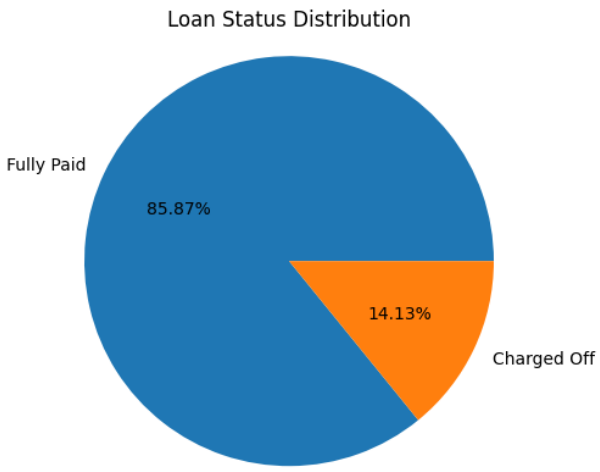
3.1.1.1. Verification status

Majority of the loans are Not Verified



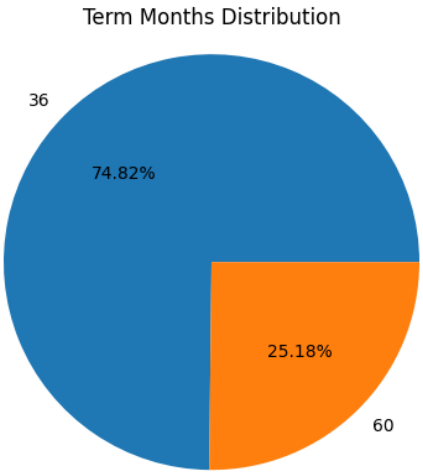
3.1.1.2. Loan status

Around 14.13% of loan are in Charged Off state



3.1.1.3. Term period

Around 75% of people take loan for a period of 36 months



3.1.1.4. Purpose of Loan

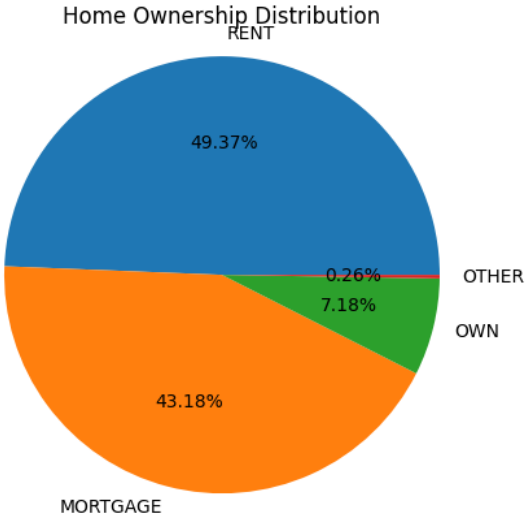
Around 50% of the people take loan for Debt consolidation

Purpose	Percentage
debt_consolidation	48.18
credit_card	13.07
other	9.84

home_improvement	6.98
major_purchase	5.59
car	4.02
small_business	3.74
wedding	2.48
medical	1.74
moving	1.48
vacation	0.95
house	0.91
educational	0.79
renewable_energy	0.24

3.1.1.5. Home Ownership

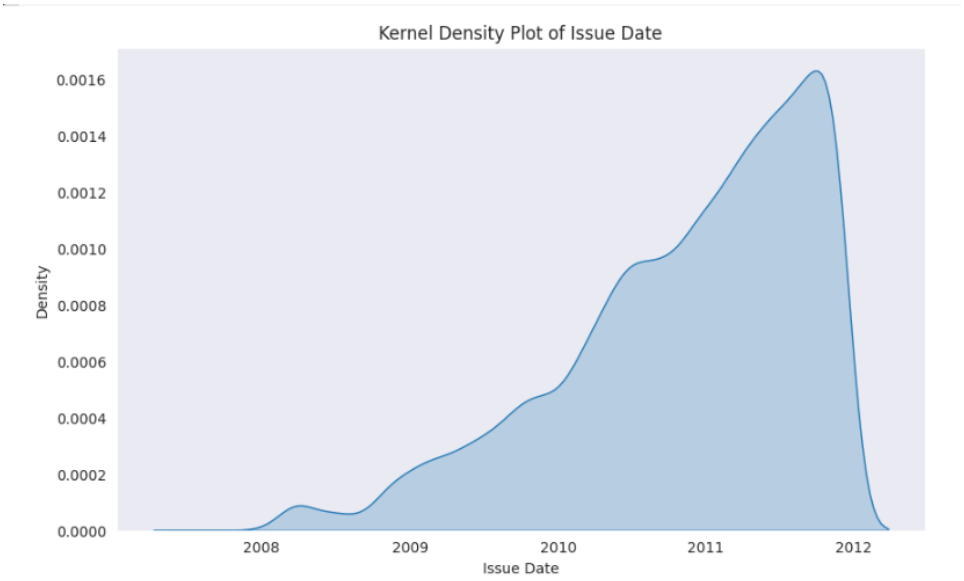
Majority of the people who take loan do not own a house



3.1.2. Categorical – Ordered

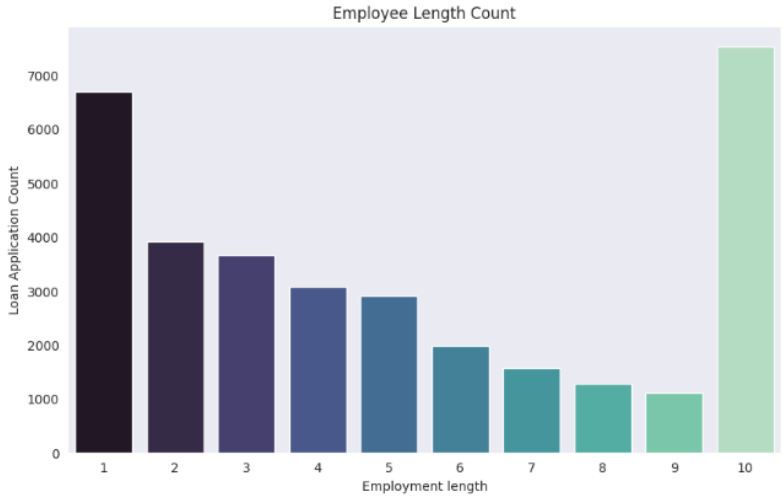
3.1.2.1. Issue date

The numbers of loans have increased with time



3.1.2.2. Employee length

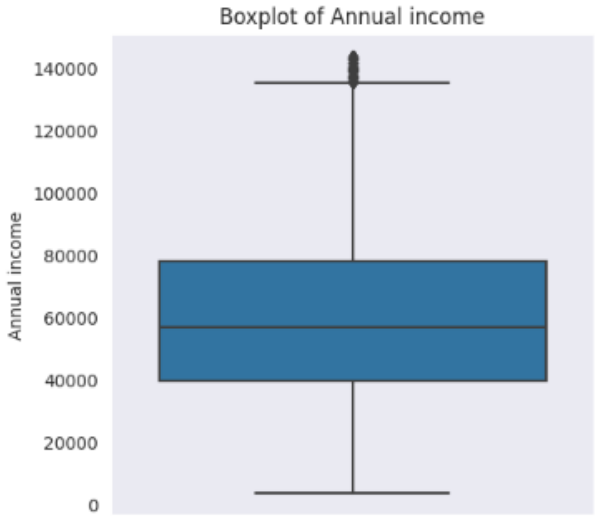
Majority of the loan is taken by people with less than 1 year of experience or more than 10 years of experience.



3.1.3. Quantitative

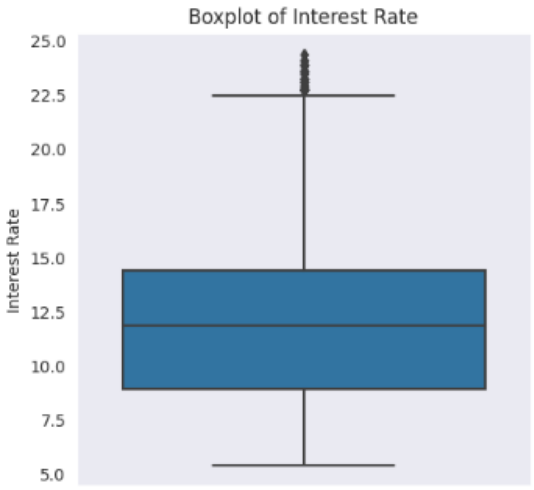
3.1.3.1. Annual income

Majority of the people's annual income lies between 40000 and 80000.



3.1.3.2. Interest Rate

Majority of people have interest rate in between 9% and 14.5%



3.2. Bivariate analysis

3.2.1. Loan status and Annual income

Salary is inversely proportional to Charge off. With decrease in salary, charge off possibility increases.

Annual Income	Charged Off	Fully Paid	Charged Off percentage
0-20000	162	716	18.45
20000-40000	1305	6308	17.14
40000-60000	1555	8902	14.87
60000-80000	918	6246	12.81
80000 +	831	6813	10.87

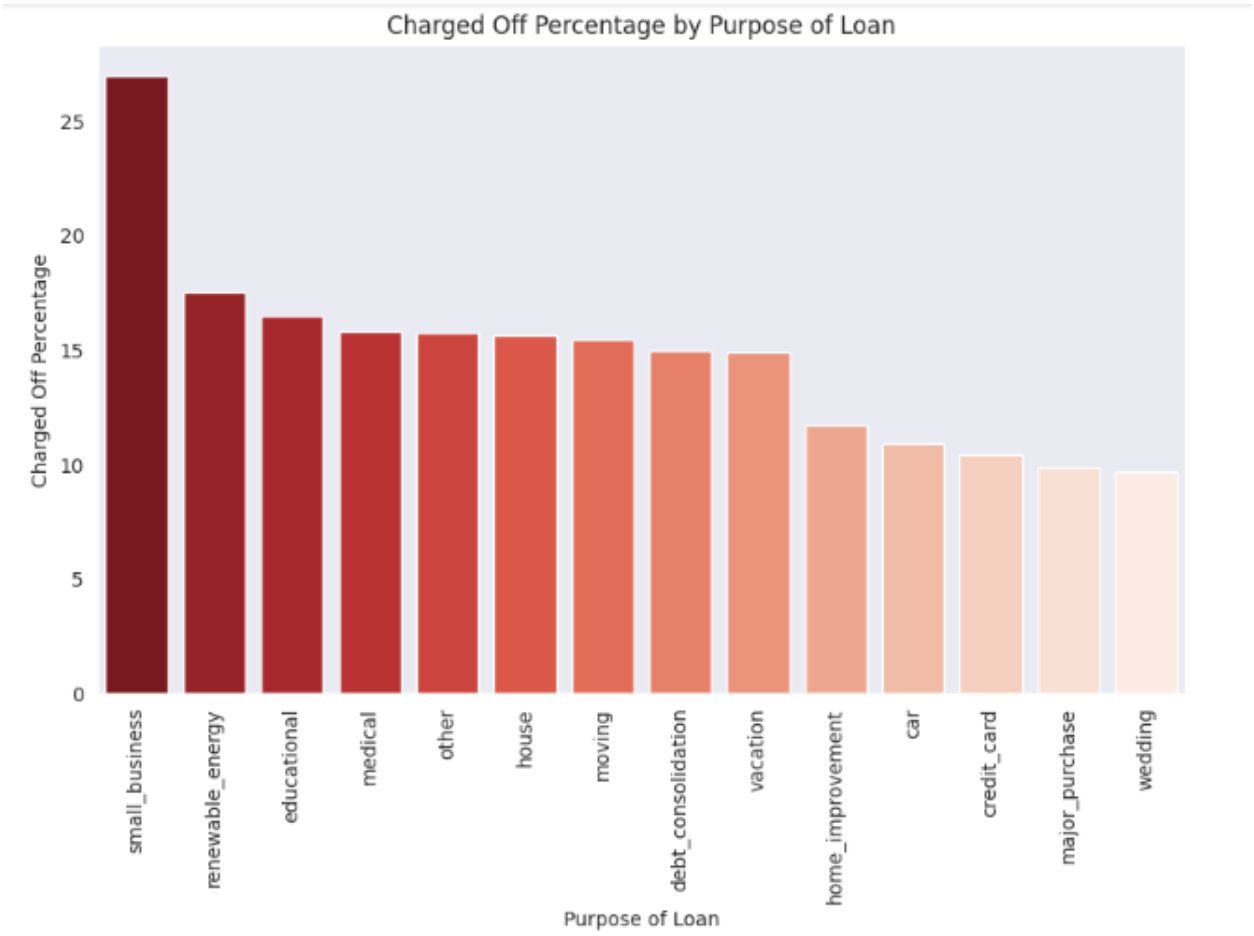
3.2.2. Loan status and Loan amount

Loan amount is directly proportional to Charge off possibility. Especially, for loan of more than 28000.

Loan amount	Charged Off	Fully Paid	Charged Off percentage
0-5000	1084	7228	13.04
5000-10000	1414	10025	12.36
10000-20000	1588	9111	14.84
20000-30000	543	2254	19.41
30000-40000 +	142	367	27.9

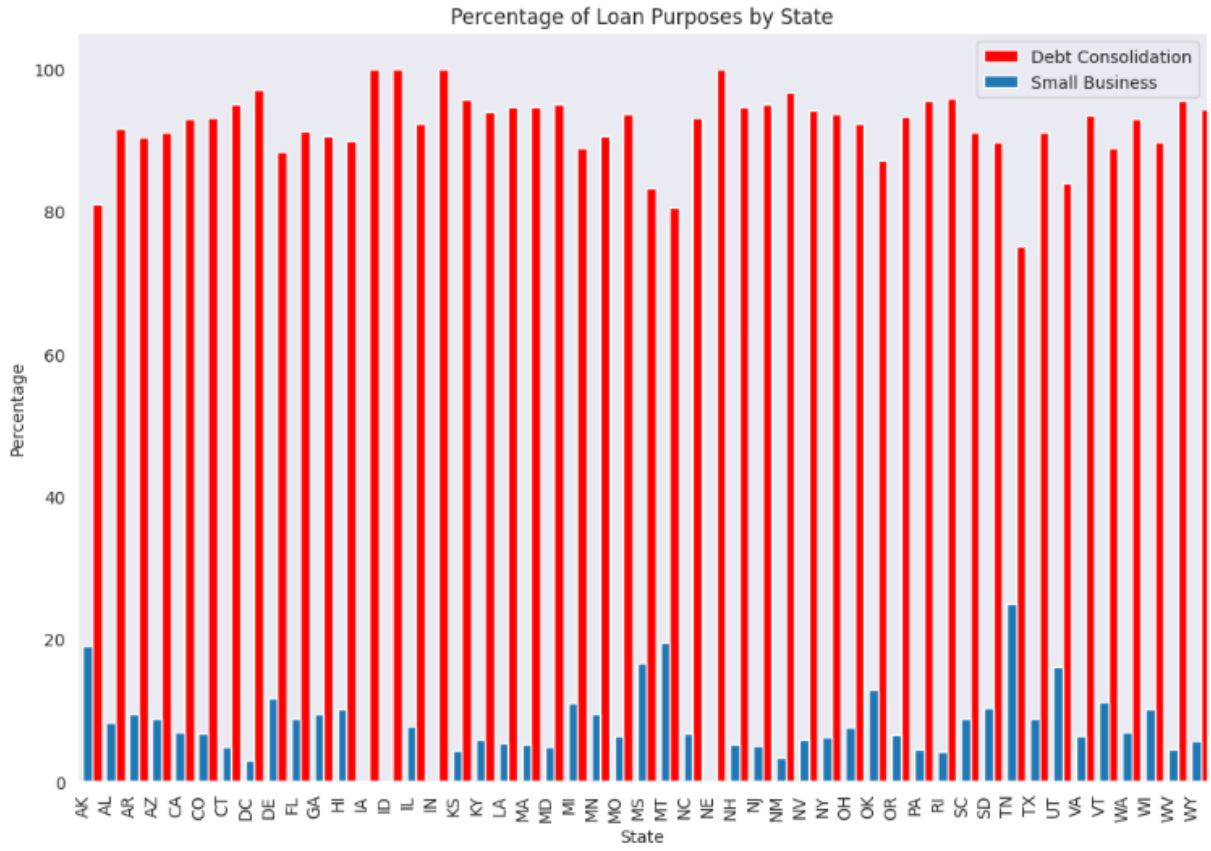
3.2.3. Loan status and Purpose

Loan for the purpose of small business has more chances of Charge Off.



3.2.4. Loan purpose and State

The percentage of 'Debt Consolidation' loans does not vary significantly across states. However, states such as 'TN,' 'AK,' 'MT,' and 'MS' have a higher percentage of 'Small Business' loans. This chart can help make informed decisions by identifying states with a higher proportion of charged-off loans.



3.3. Derived Metrics

3.3.1. Business Driven Metrics

3.3.1.1. Debt to income ratio

Debt to income ratio is calculated by dividing installment by month salary (month salary derived from annual salary divided by 12)

On calculating Debt to Income ratio it is observed that the chances of Charged Off increases with increase in ratio. However, it is not a clear increase, as there is decrease for 0.2+ debt to income ratio.

Debt-income-ratio	Charged Off	Fully Paid	Charged Off percentage
0-0.04	1126	8640	11.53
0.04-0.08	1711	11570	12.88
0.08-0.12	1222	5880	17.21
0.12-0.16	529	2218	19.26
0.16-0.2	154	554	21.75

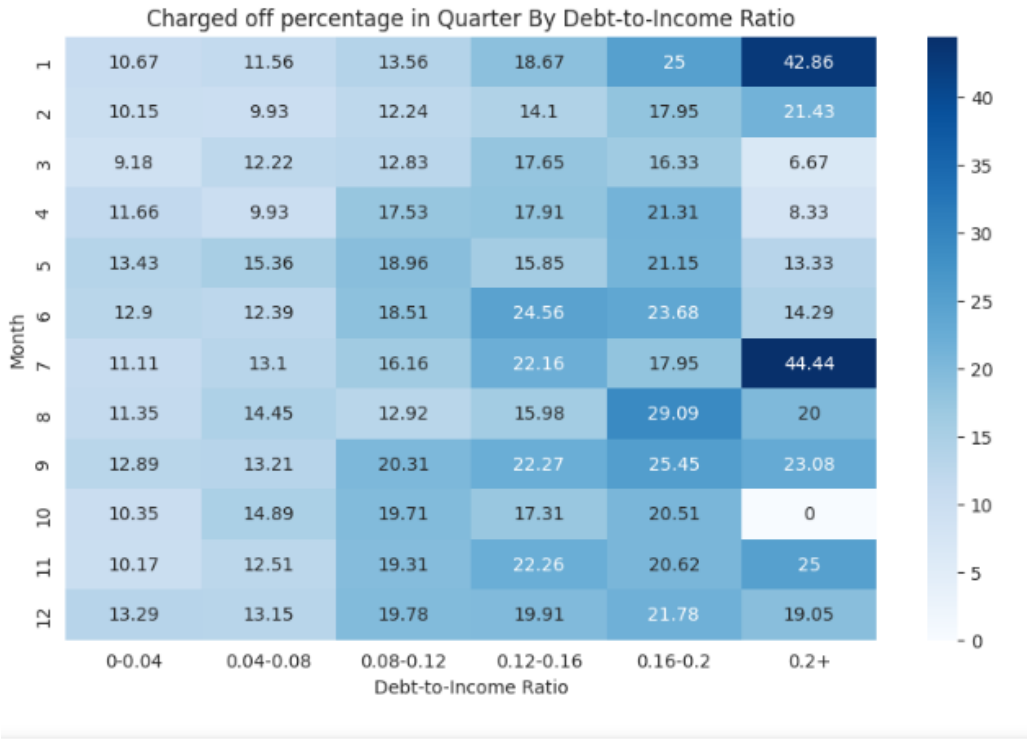
0.2+	29	123	19.08

3.3.2. Data Driven Metrics

3.3.2.1. Charged Off

Created a column to indicate charged off (value 0 for False, 1 for True) from column loan_status. It was used to create the graph below.

As per below graph, around close to 40% of Charged Off happen in month of January and July having maximum Debt to Income ratio (0.2+)



3.3.3. Type Driven Metrics

3.3.3.1. Month

Month is derived from issue date.

This data is used to derive conclusions in the previous section.

4. Conclusion

Below are the key conclusions to determine the Charged Off cases:-

1. Charged Off possibility increases with increase in loan amount. Around 25% for loans more than 28000.
2. Charged Off possibility increases with decrease in annual income.
3. Close to 40% of Charged Off happen in the months of January and July having maximum Debt to Income ratio (0.2+). It could be due to the Festive season in December and Summer vacation in May-June.
4. Majority of the loans are taken for the purpose of Debt Consolidation and it is 8th highest in terms of Charged of percentage.
5. Chances of Debt are observed to be more when loan taken for Small business, especially in states such as TN, AK, MT and MS.