Lending Club Case Study

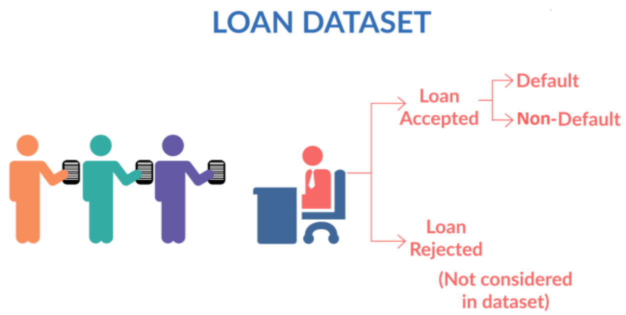
- Nidhi Mantri

Problem Statement

- You work for a **consumer finance company** which specialises 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
- The loan dataset contains information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

About Dataset

- Here, since, we don’t have data for “Loan Rejected” category. So let’s understand categories in “Loan Accepted” part.
- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
  - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
  - **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'.
  - **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.



Business Objective

The company wants 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.

1

Exploratory Data Analysis

Data Understanding

- There are total of 111 columns and 39717 rows
- float variables - 74, integer variables - 13, and string variables – 24
- No Duplicates in dataset

“

2. Handling Null Values

- Let’s drop all the columns, having null values > 60% of total rows
- Total number of such columns is 57

3. Drop below types of columns

- Columns having only one category throughout the dataset, such columns doesn’t add value in analysis.
- Customer Behavior Columns – Such columns are not available at the time of loan application, and thus they cannot be used as predictors for credit approval. Columns in dataset are - delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d.
- Not Important Variables – Columns for which have better alternates in dataset. Ex. Purpose is much better than title and desc (description). So we can drop title and desc.

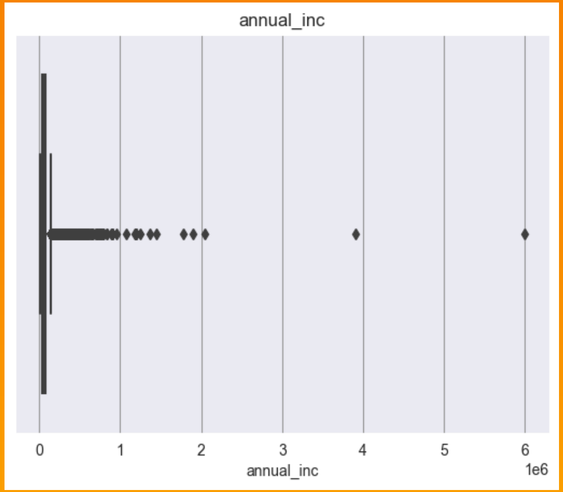
4. Data Preprocessing

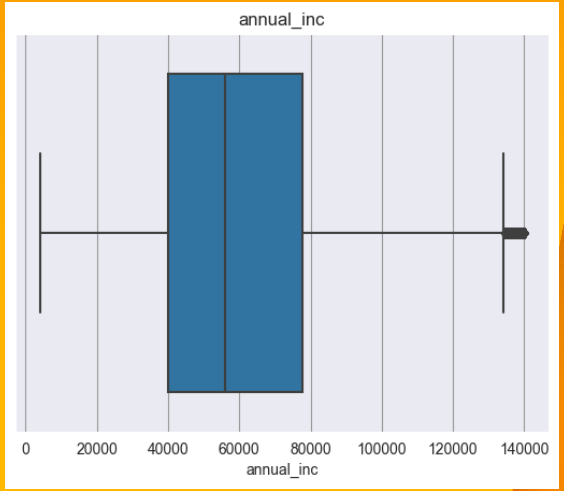
Data Cleaning and Manipulation

- Imputation of emp_length (Employment tenure) with it’s mode and of pub_rec_bankruptcies (public record bankruptcies) with “Not Known” (because bankruptcy is a very serious parameter, we shouldn’t randomly guess it).
- Combining NONE, ANY, OTHER into “OTHER” category.
- Fixing data e.g., by removing extra % sign from interest rate column.

Filtering

As we want to predict between defaulters and non defaulters, so data of borrowers whose loan_status is "Current" is of no use to us right now.. because we can't predict beforehand whether those borrowers would be defaulters or non-defaulters.





5. Outlier detection and Removal

On left side (1st plot), this is a box plot of annual_inc (annual income) column, we can see clearly that there are some outliers. So, in analysis let's drop all the values after 95th percentile. After dropping, see 2nd plot, now it's a continuous distribution

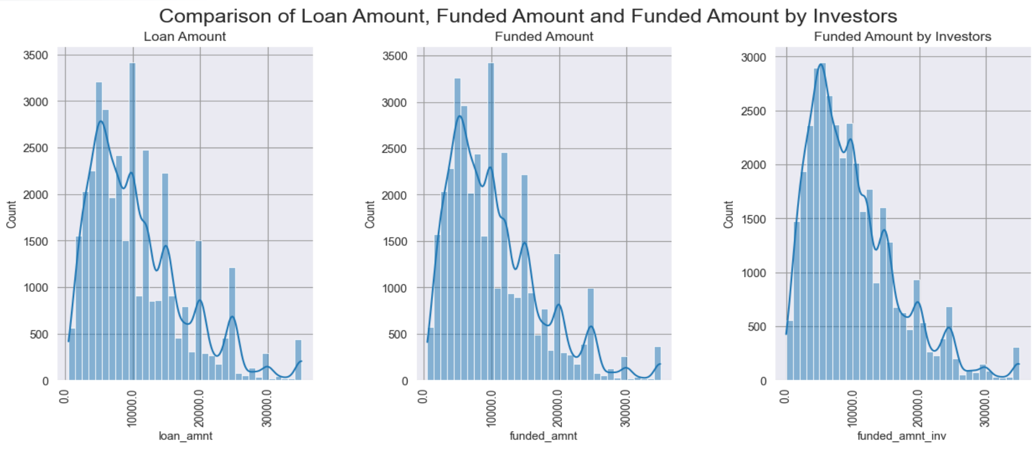
On the other hand, if distribution is continuous, then we don't drop any row.

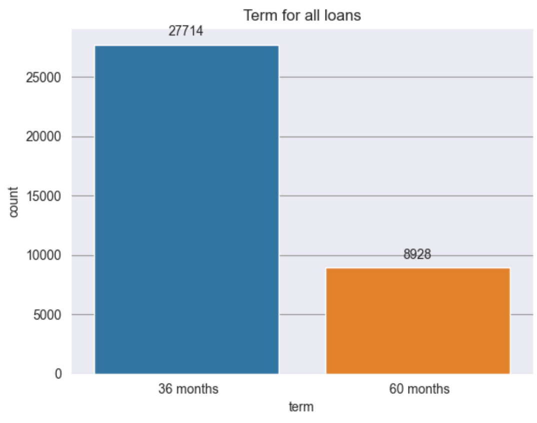
After these 5 steps, we are ready for Univariate and Bivariate analysis

Before moving forward, let’s look how are dataset looks like -
Shape of dataset now - (36642, 19)
Shape of dataset (Charged Off status) - (5416, 19)

6. Univariate Analysis

1. All three amounts are following same pattern, with slight changes in values in funded_amnt_inv
2. Maximum number of loans are in range 5k to 10k

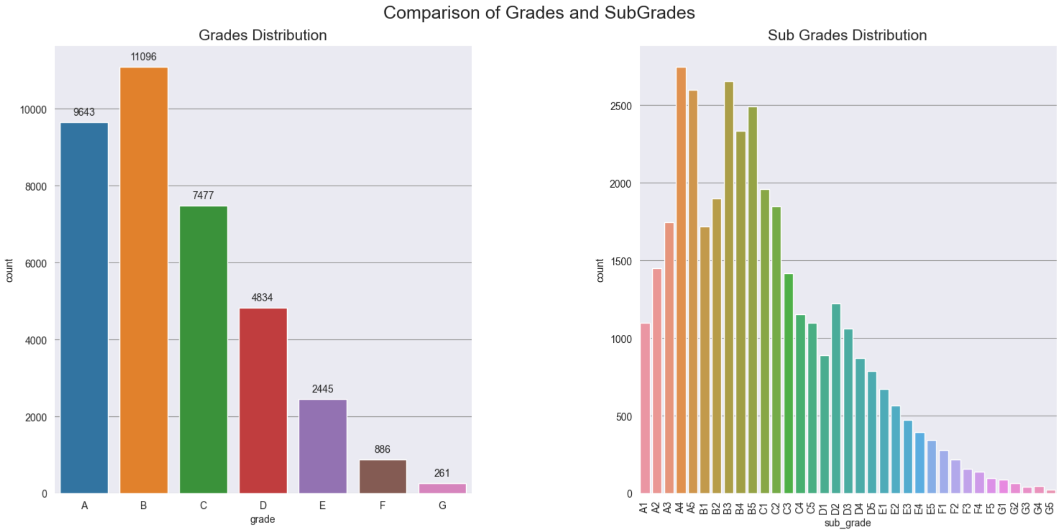


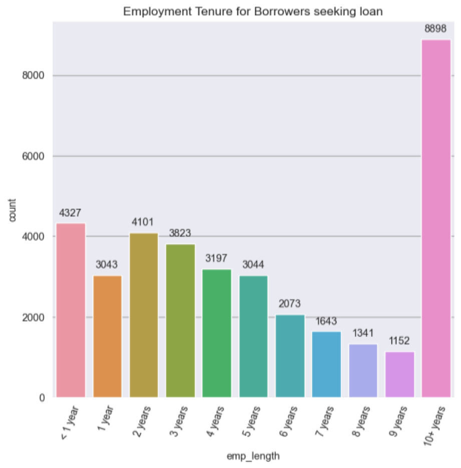


Maximum loans are for period of 36 months

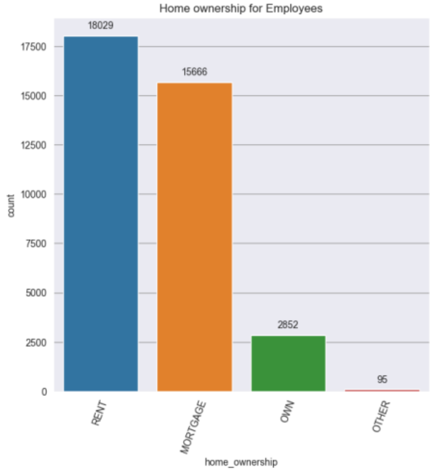
Continued

1. Maximum count is for grade B (subgrade B1 to B5) and
2. lowest is for grade G (subgrade G1 to G5)


Comparison of Grades and SubGrades


Employment Tenure for Borrowers seeking loan

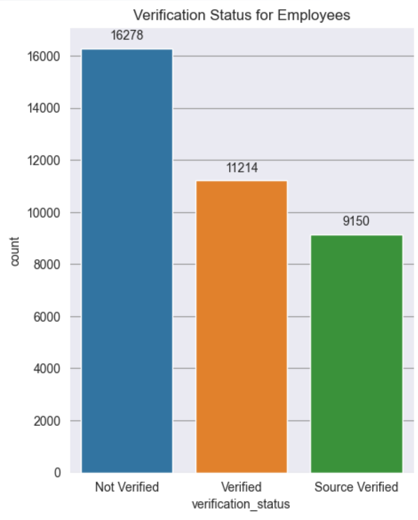
1. Maximum employees have 10 or 10+ years of experience
2. Least having 9 years of experience


Home ownership for Employees

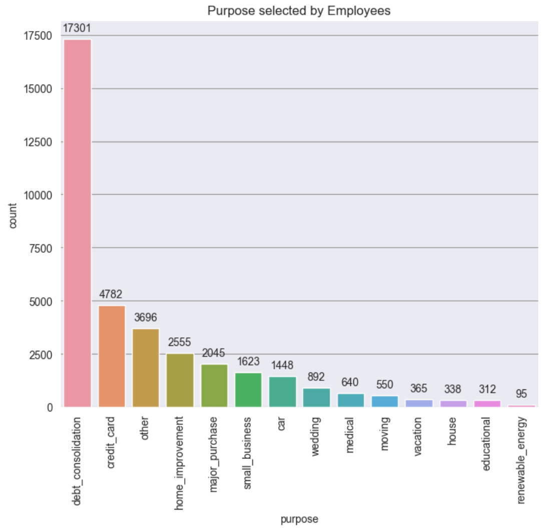
Maximum borrowers have either living in rented home or mortgaged their property

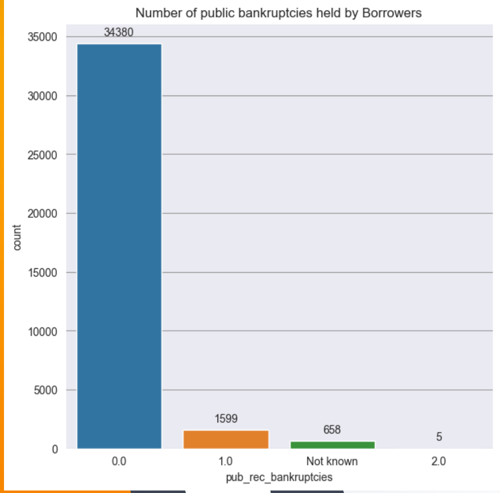
Continued

1. income sources of employees are least verified and
2. employees with incomes not verified are maximum

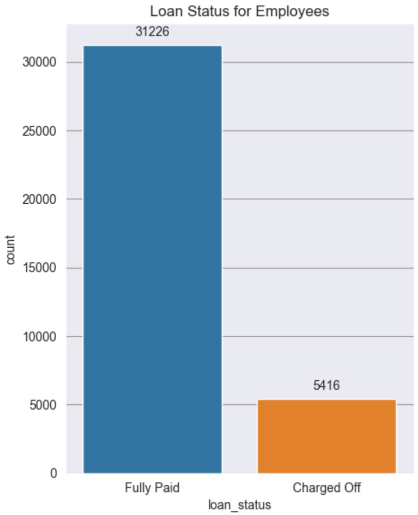


1. Maximum people have taken loan for debt consolidation and credit card.,
2. least for renewable energy



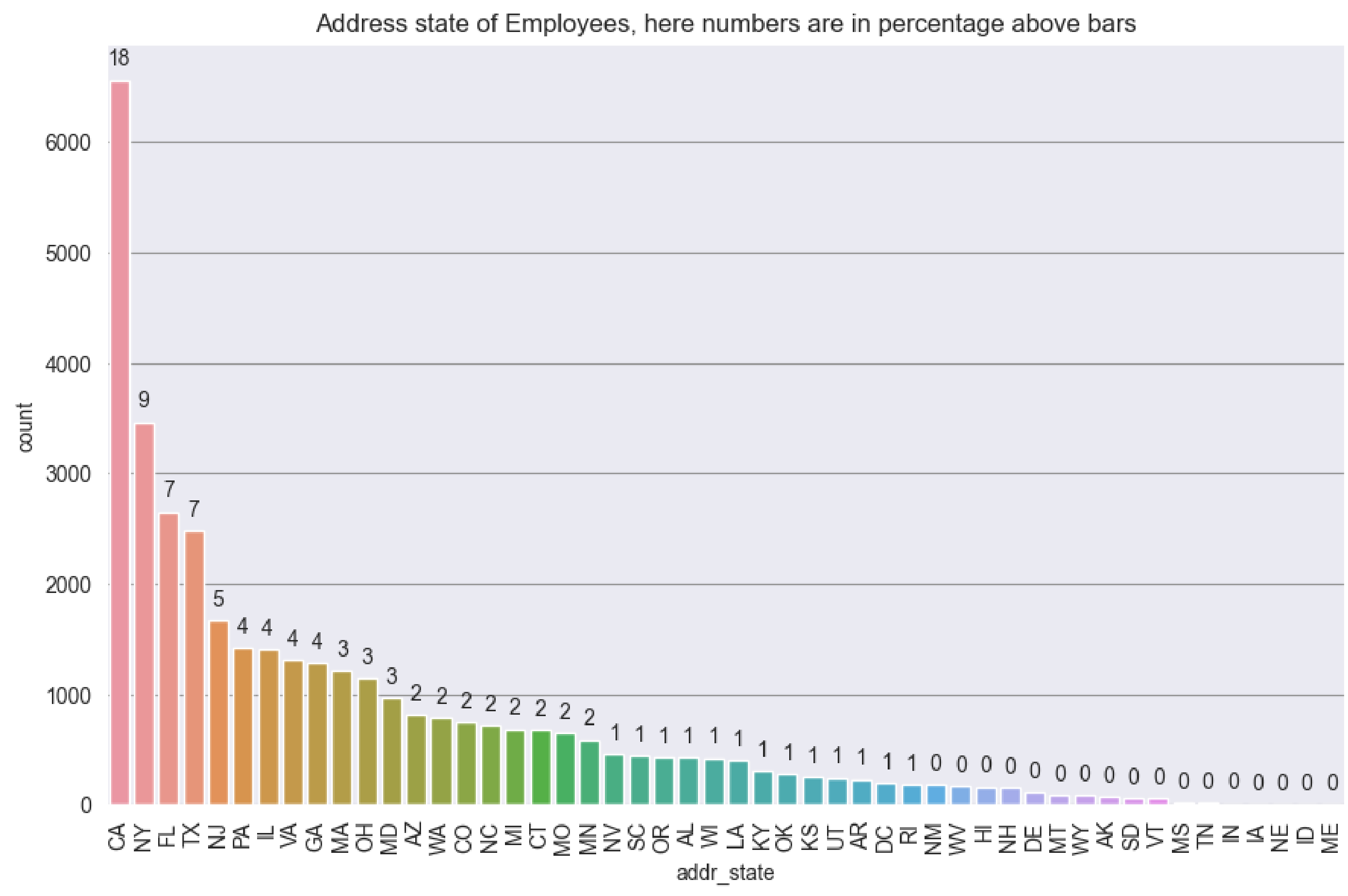


Maximum borrowers have 0 record of bankruptcy



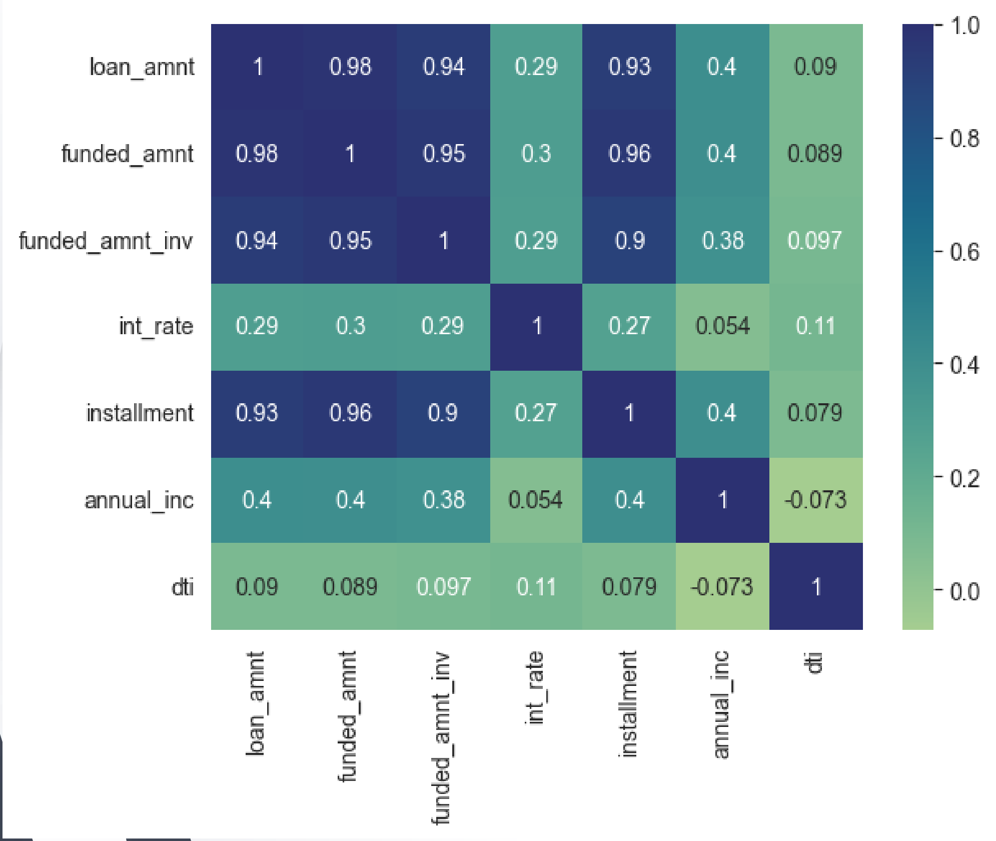
Maximum employees have paid their loans fully

Continued


Address state of Employees, here numbers are in percentage above bars

Maximum people (18% of total) are from CA

7. Bivariate Analysis



	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti
loan_amnt	1	0.98	0.94	0.29	0.93	0.4	0.09
funded_amnt	0.98	1	0.95	0.3	0.96	0.4	0.089
funded_amnt_inv	0.94	0.95	1	0.29	0.9	0.38	0.097
int_rate	0.29	0.3	0.29	1	0.27	0.054	0.11
installment	0.93	0.96	0.9	0.27	1	0.4	0.079
annual_inc	0.4	0.4	0.38	0.054	0.4	1	-0.073
dti	0.09	0.089	0.097	0.11	0.079	-0.073	1

1. loan_amnt, funded_amnt, funded_amnt_inv (inv stands for investor), installment are highly correlated with each other in positive direction
2. dti (debt to income ratio) column is not correlated to any other column, same holds for annual_inc also.

Continued

Let’s create bucket for continuous variables to analyse in more better way.

Loan amount - ['0-5000', '5000-10000', '10000-15000', '15000-20000', '20000-25000', '25000-30000', '30000+']

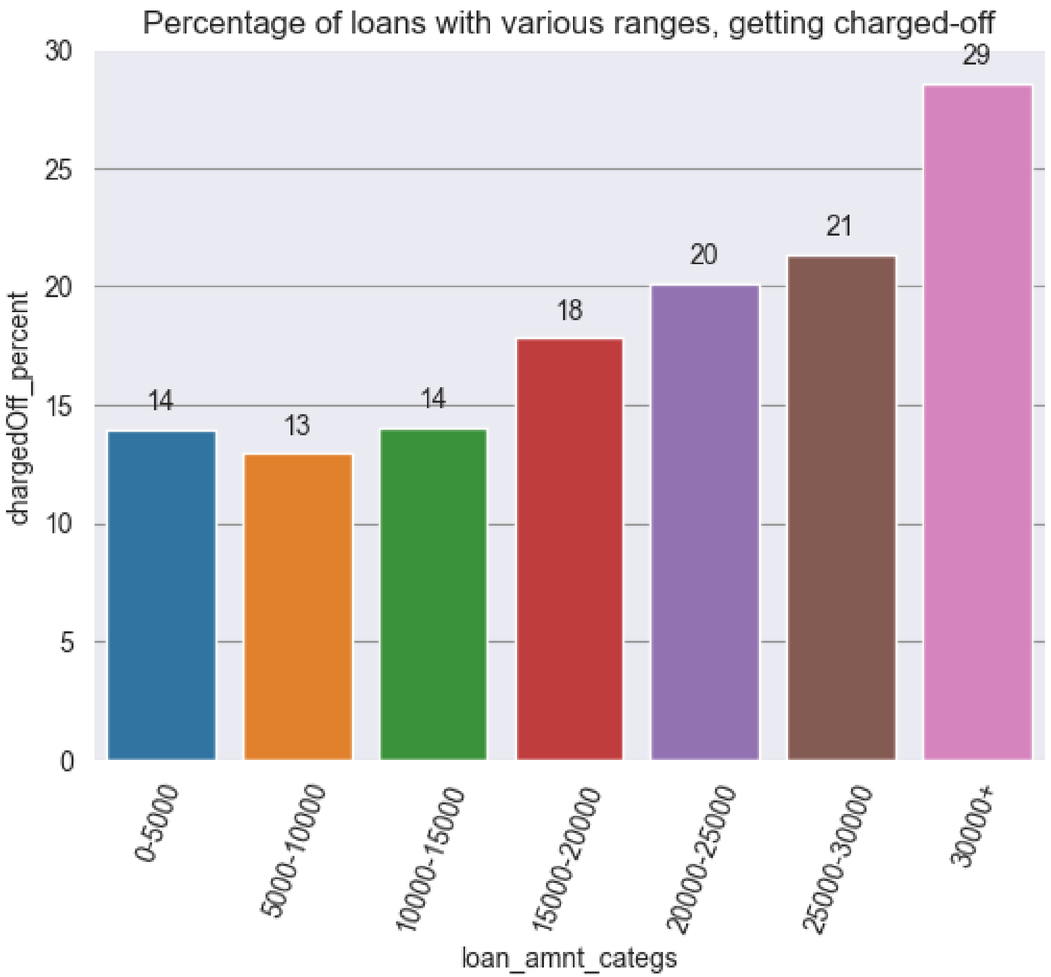
Interest rate - ['0-10','10-12.5','12.5-16','16+']
Instalments - ['0-100', '100-200', '200-300', '300-400', '400-500', '500+']

Annual income - ['0-30000', '30000-50000', '50000-70000', '70000+']

Debt to income ratio - ['0-5', '5-10', '10-15', '15-20', '20-25', '25+']

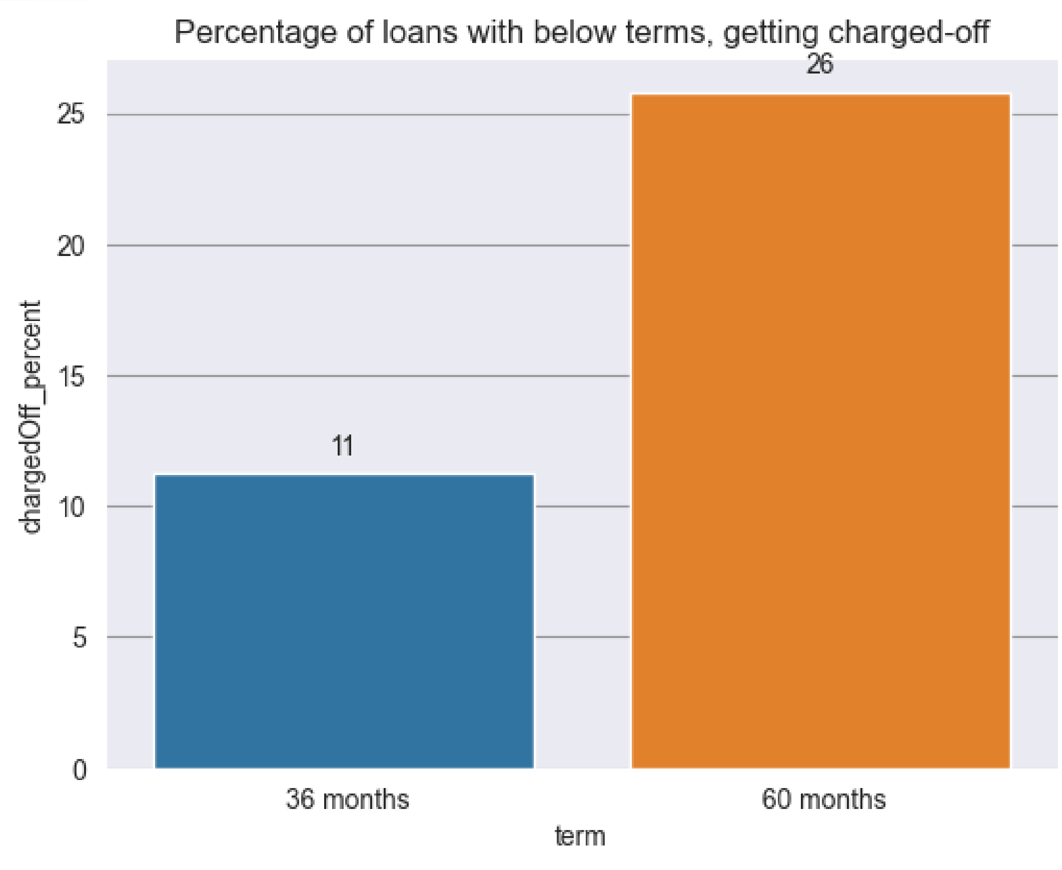
Continued

1. 30000+ amounts have 29% chances of getting charged off
2. 25000-30000 range has 21% chances of getting charged off
3. 20000-25000 range has 20% chances of getting charged off
4. 15000-20000 range has 18% chances of getting charged off

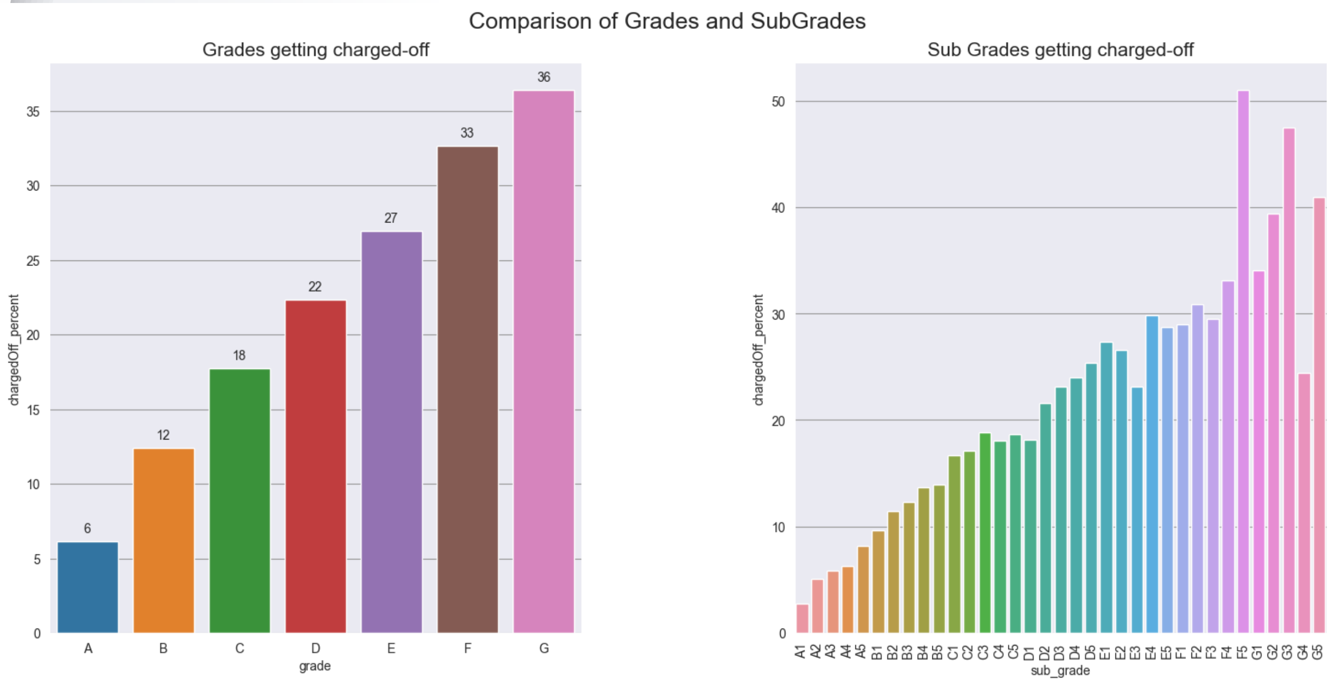


Continued

- 60 months term has 26% chances of getting charged off
- 36 months term has 11% chances of getting charged off



Continued

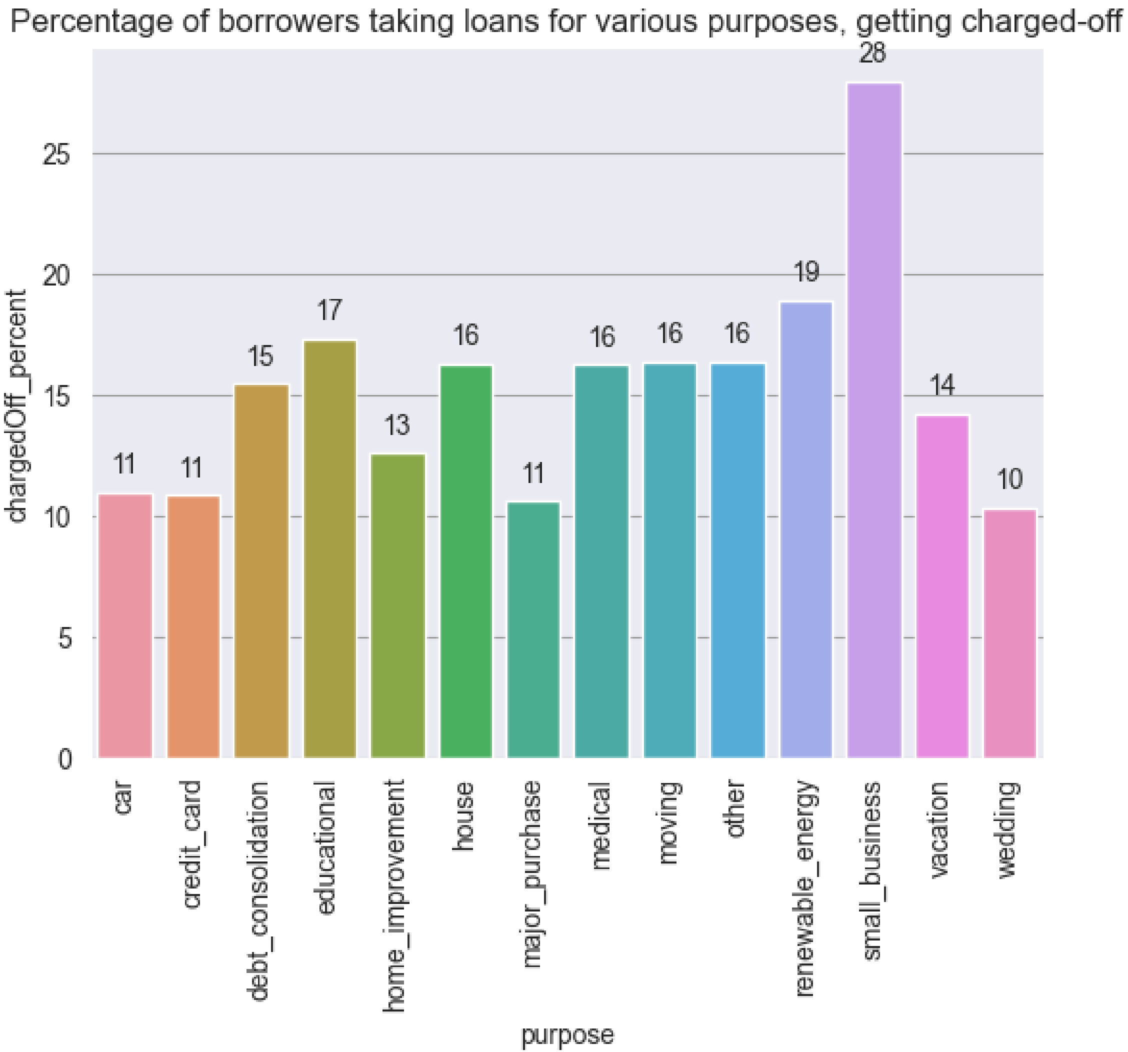


1. loans with G grade have 36% chances of getting charged off
2. loans with F grade have 33% chances of getting charged off

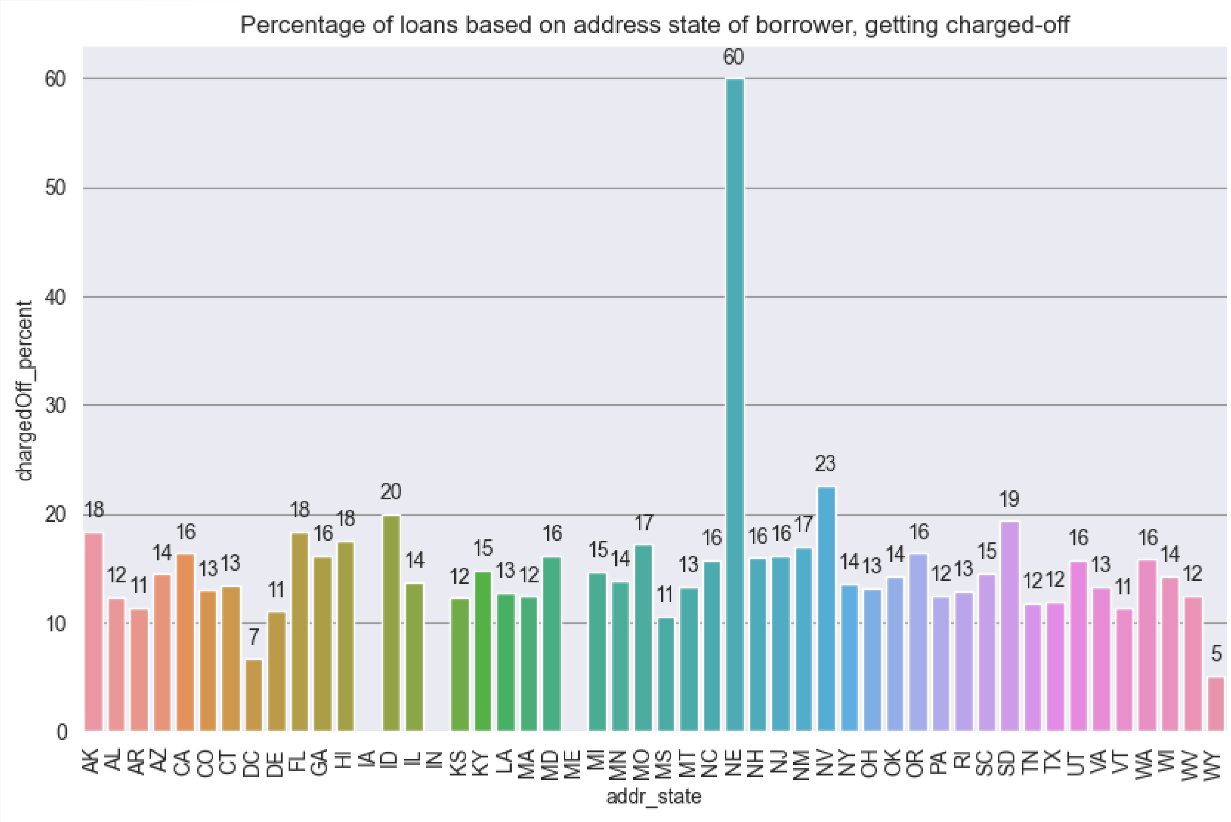
SubGrade is also follows similar trend as Grade, because it's a part of Grade only.. so sub grades from F1 to G5 have maximum chances of getting charged off among other sub grades.

Continued

- loan with small business purpose has 36% chances of getting charged off
- loan with renewable energy purpose has 33% chances of getting charged off



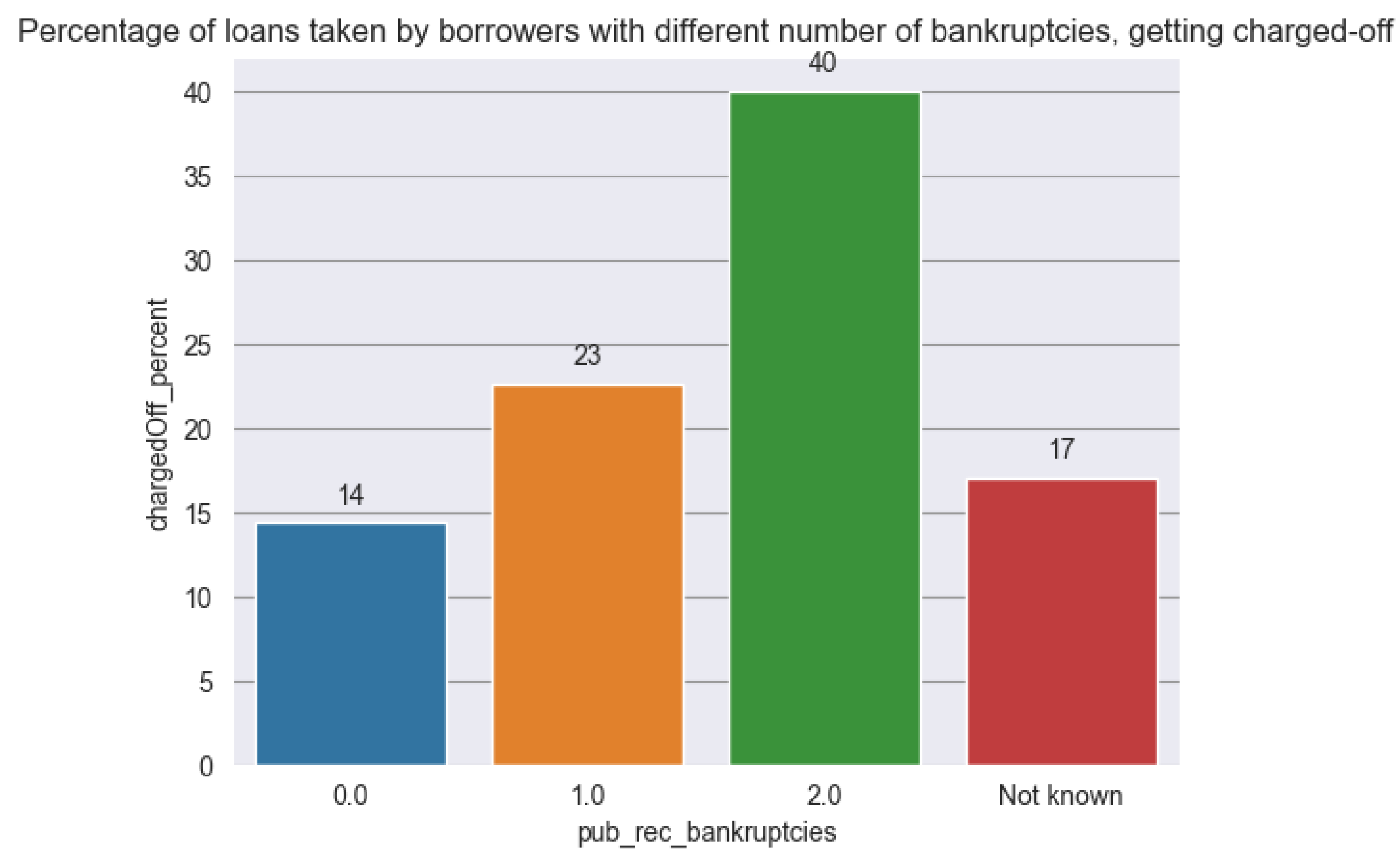
Continued



loans of borrowers from state "NE" have 60% chances of getting charged off

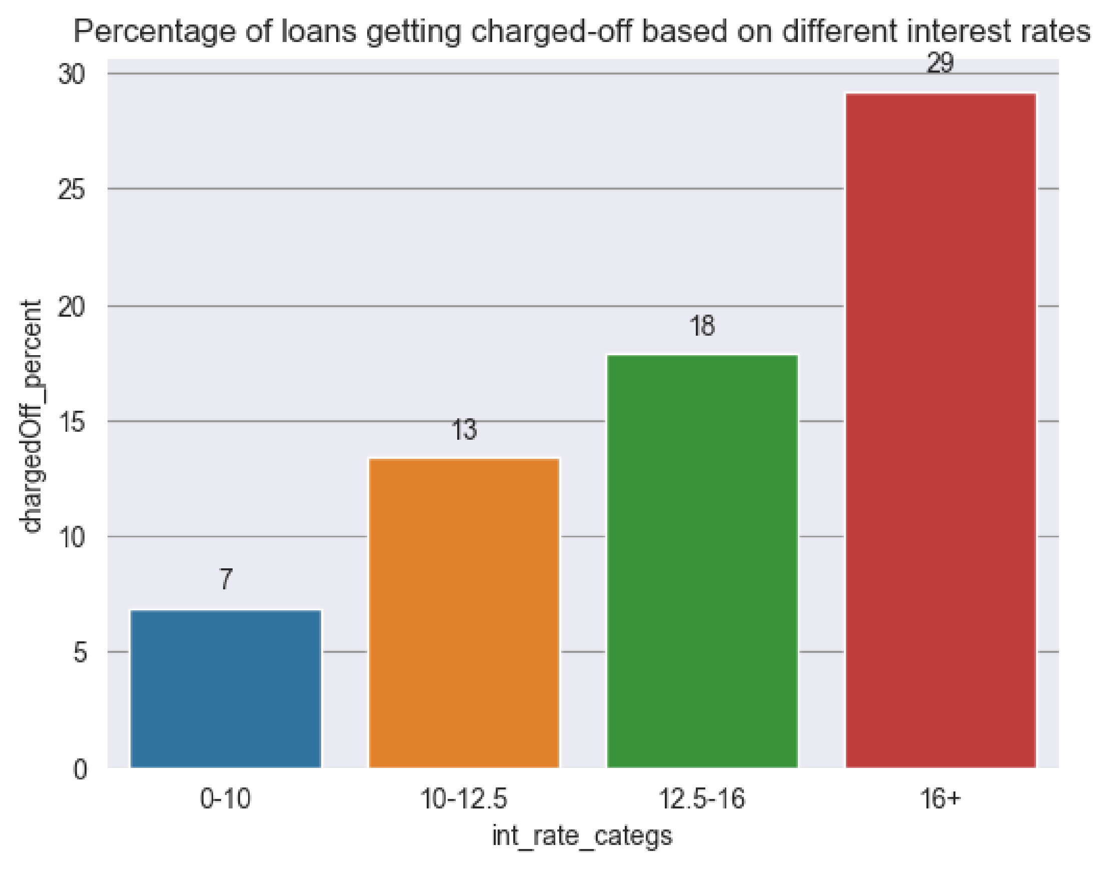
Continued

- borrowers with number of bankruptcies as 2 have 40% chances of loan getting charged off
- borrowers with number of bankruptcies as 1 have 23% chances of loan getting charged off



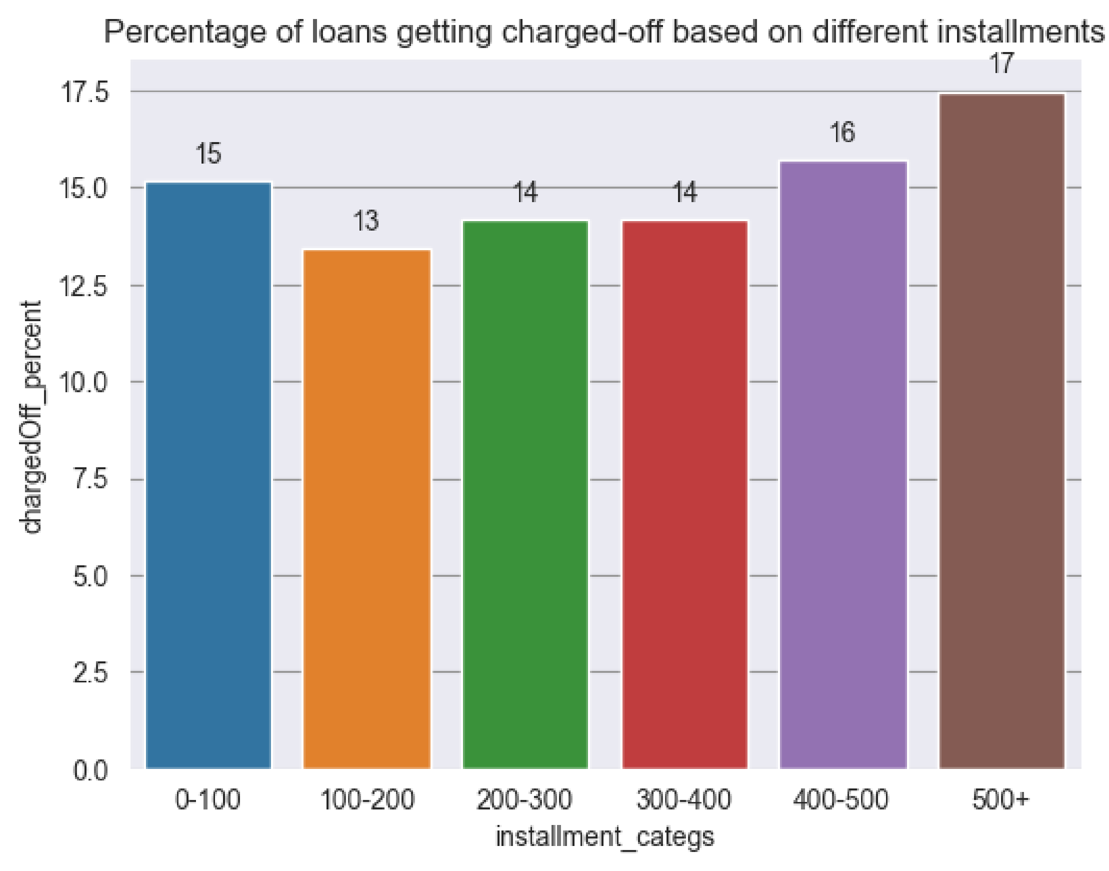
Continued

- interest rates greater than 16 have 29% chances of getting charged off
- interest rates between 12.5-16 have 18% chances of getting charged off



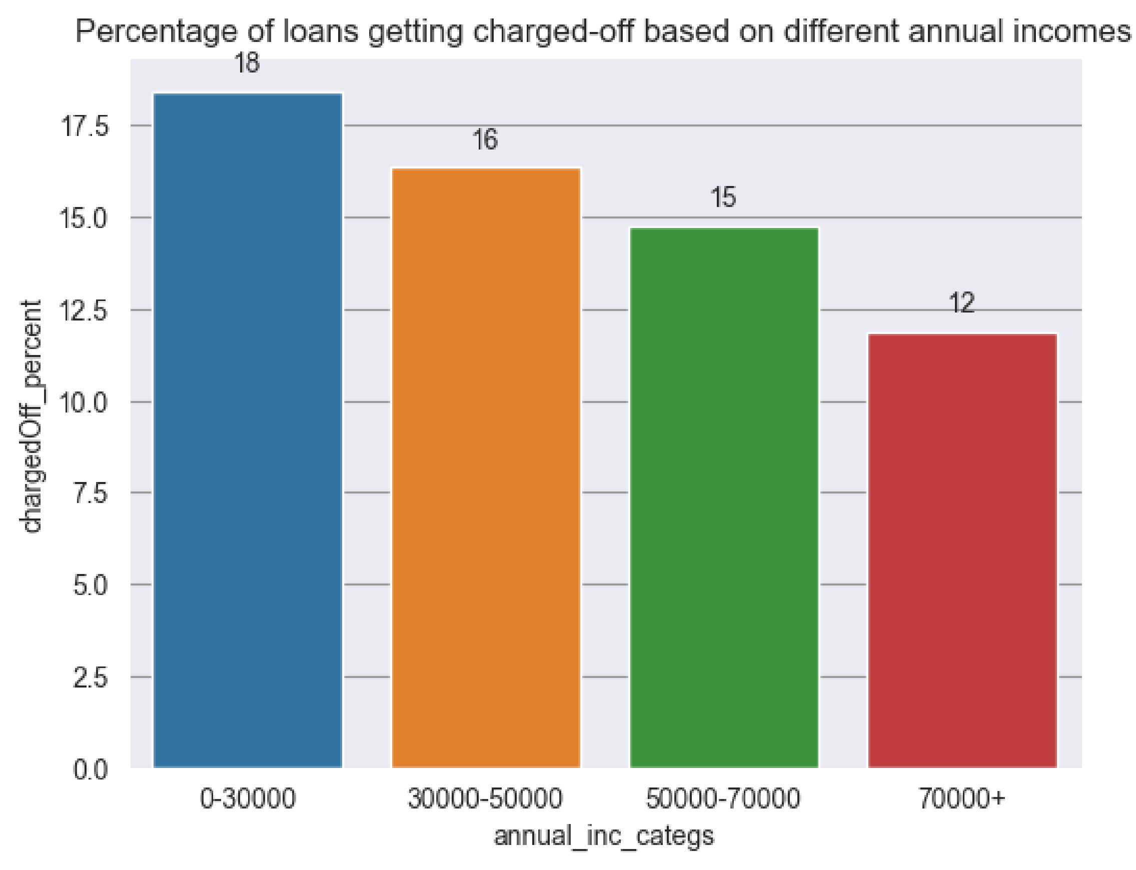
Continued

- installments greater than 500 have 17% chances of getting charged off
- installments between 400-500 have 16% chances of getting charged off



Continued

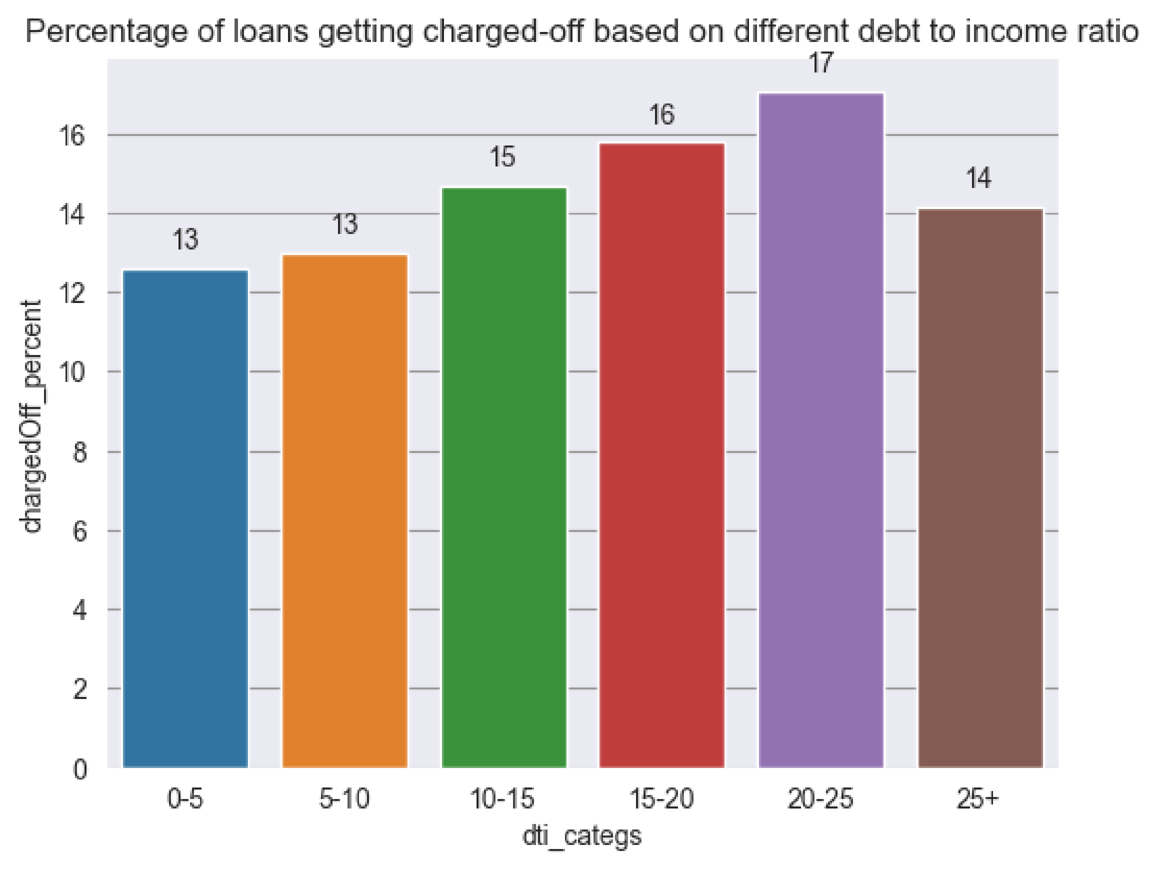
- annual incomes lesser than 30000 have 18% chances of getting charged off
- annual incomes between 30000-50000 have 16% chances of getting charged off



Continued

- dti between 20-25 has 17% chances of getting charged off
- dti between 15-20 has 16% chances of getting charged off

<u>Note</u> - dti greater than 25+ has lower value in our dataset because data is less for these dti ratios as compared to other



Driving Variables, which plays a key role in identifying chances of defaults

As value of below variables increases, chances of loan getting charged off also rises –

1. **Loan Amount**
2. **Interest Rate**
3. **Installment**
4. **Debt to income ratio**
5. **Term**
6. **Public record bankruptcies**
7. **Address State -** loans of borrowers from state "NE" have 60% chances of getting charged off
8. **Grade & SubGrade -** As grade moves from A to G, chances of loan getting charged off also rises
9. **Purpose -** if purpose for loan is small business or renewable energy, then chances of loan getting charged off are huge
10. **Annual Income -** As annual income decreases, chances of loan getting charged off also rises

Thanks!

- Nidhi Mantri