EDA REPORT

Lending Club analysis

28th July 2018

Amirisetty Vijayaraghavan	**Hrudaya Ranjan Sahoo**
Aditya Mehta	**Divya R K**

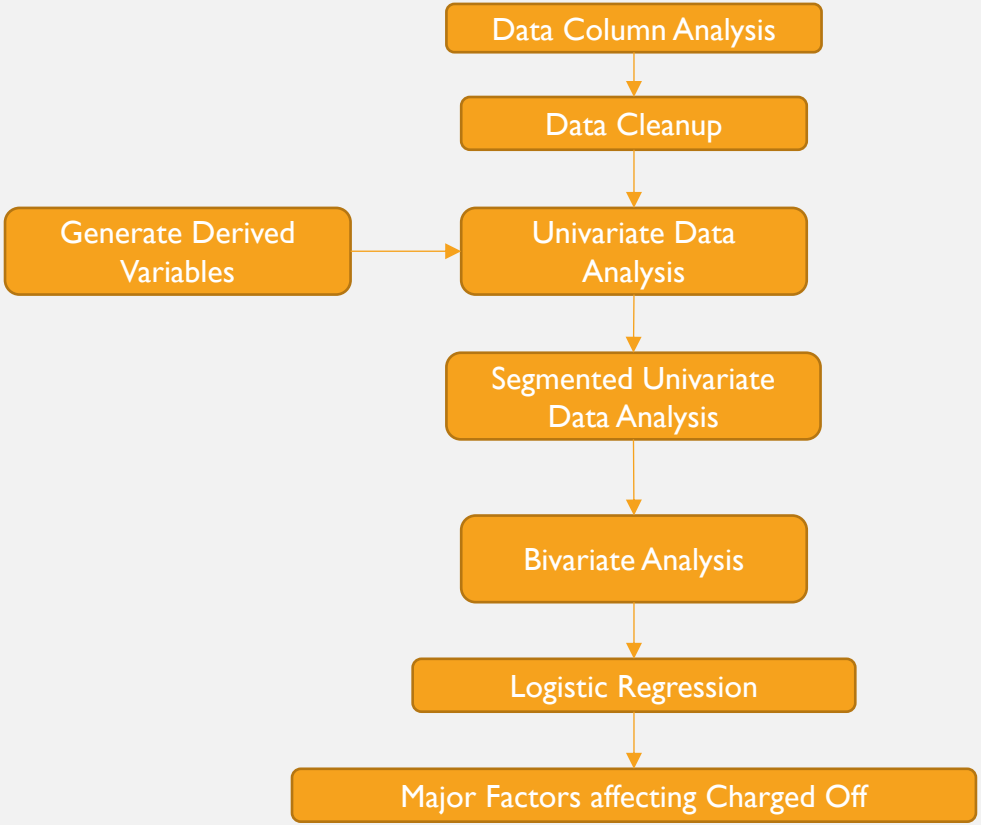
BUSINESS PROBLEM

- Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.
- Lending loans to ‘risky’ applicants is the largest source of financial loss.
- Identify these risky loan applicants and reduce such loans to minimize financial loss.
- Recommend guidelines to minimize this loss.

DATA AVAILABLE

- 39, 717 rows and 111 columns along with a column loan_status.
- Loan_status indicates of a loan is charged off (default and hence financial loss), fully paid or current.
- 83% Fully Paid Applications, 14% Charged Off and 3% Current Applications.
- Data for the 5 years from 2007 to 2011
- Data for Applications across all 50 US states.

SUMMARY OF THE APPROACH



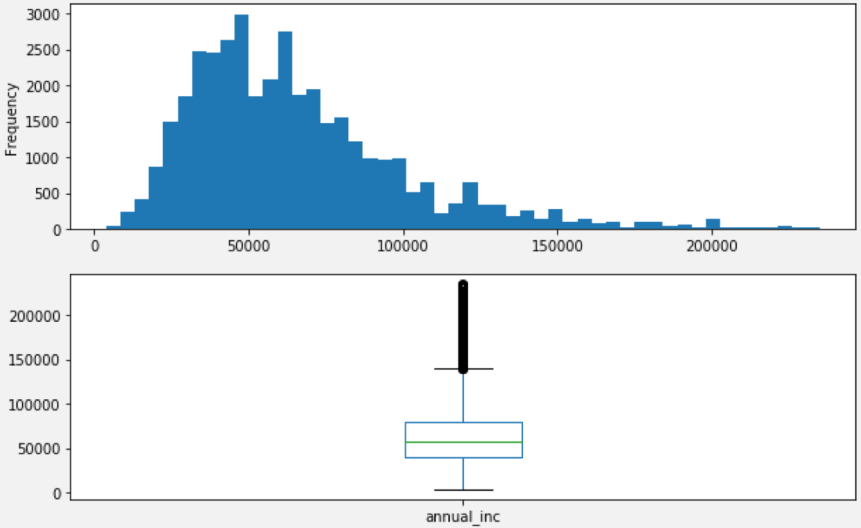
DATA CLEANUP

Column	Data Cleanup Done
54 columns have all null values	Dropped these columns
8 columns have a single unique value	Dropped these columns
Columns like int_rate, installment are strings	Convert these columns to numeric
Column like emp_length have characters like < and >	Round the column to the nearest number and convert to numeric
Column zipcode is of variable length and type object	Take the first 3 characters and convert to string.
Date Columns like next_pymnt_d, last_pymnt_d, last_credit_pull_d are strings in MON-YY format.	Convert these into sequence of integers and categorical variables with year and month.

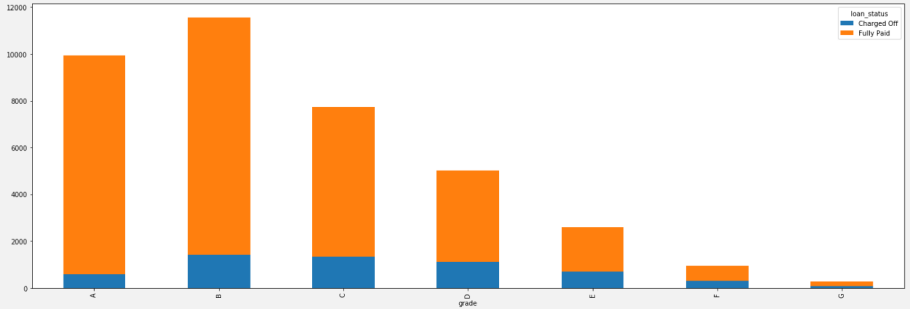
SAMPLE DERIVED COLUMNS

Column Name	Column Formula	Description
Annual Income Category	Segment the customers annual income into 4 categories.	Income Segment of the customer
grade_subgrade	Concatenate grade and subgrade	A categorical variable that shows the interactions of grade and subgrade
Revol_Bal_Category	Split revolving balance into 4 categories.	A categorical variable derived from the revolving balance.
Year and Month for dates	Extract the year and date from the given date format	Extracting the year and month from the date.
Requested Debt to Available Credit (rdac)	Ratio of loan amount to available revolving balance	Business metric indicating the requested loan amount to available credit of the customer
Loss percentage	Total received late fee divided by total payment	Out of the total amount paid, what percentage is the late fee.
Credit Limit	Revolving balance/ Revolving utilization	Calculate the business variable to compute the credit limit
Available Credit	Credit limit – revolving balance	Calculate the business variable available credit

DATA ANALYSIS - UNIVARIATE

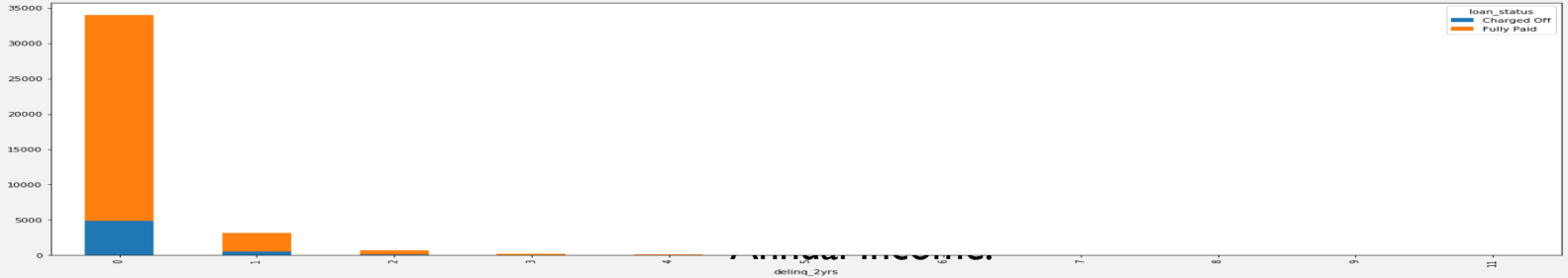


Annual Income:

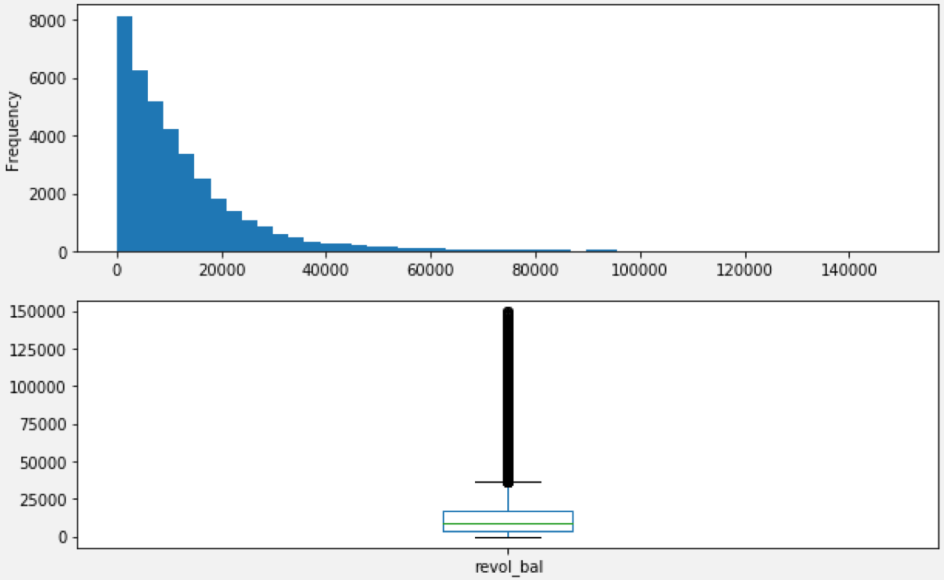


grade histogram

delinq_2_yrs histogram

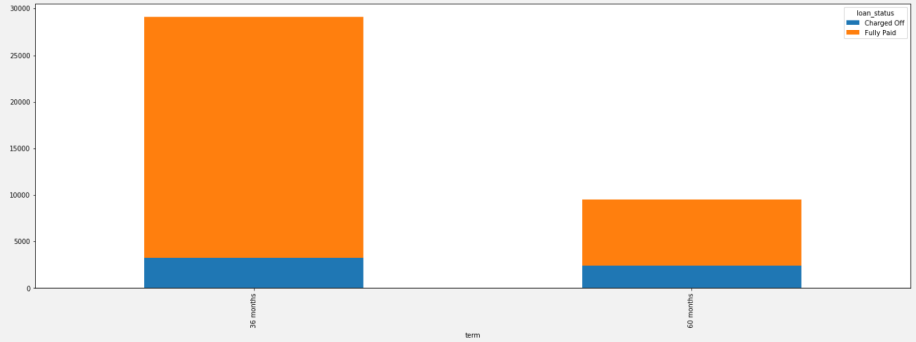


DATA ANALYSIS - UNIVARIATE



Revolving Balance Histogram and BoxPlot

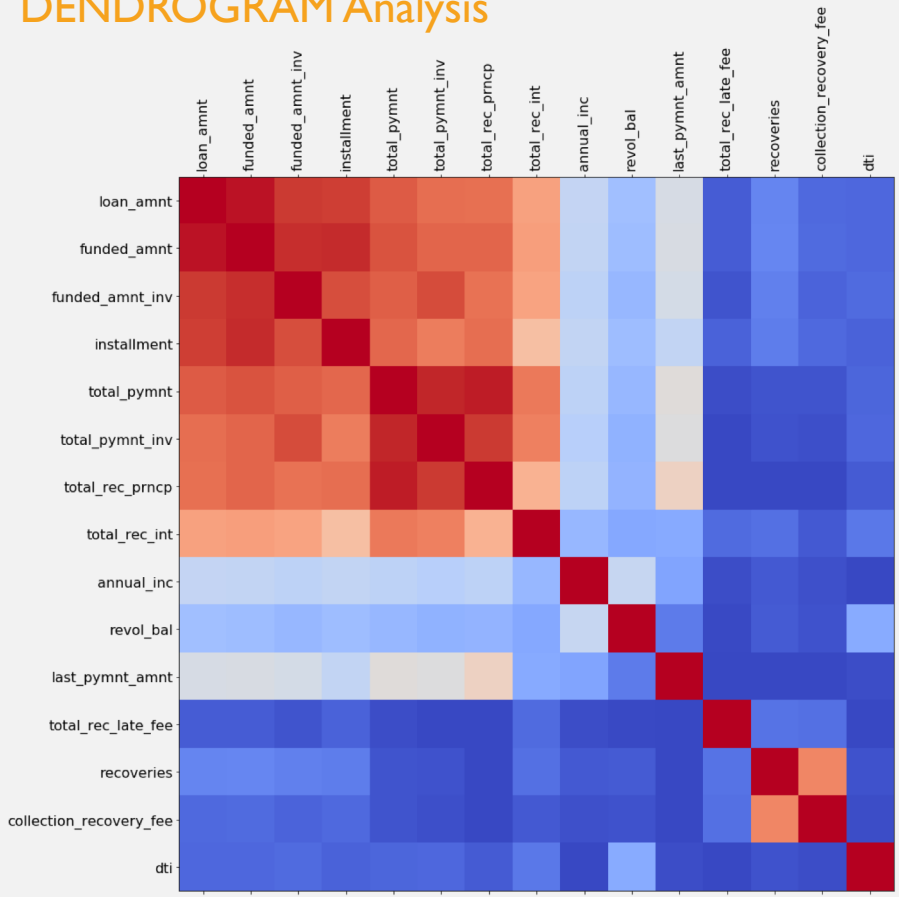
public_rec_bankruptcies histogram



term histogram

DATA ANALYSIS - BIVARIATE

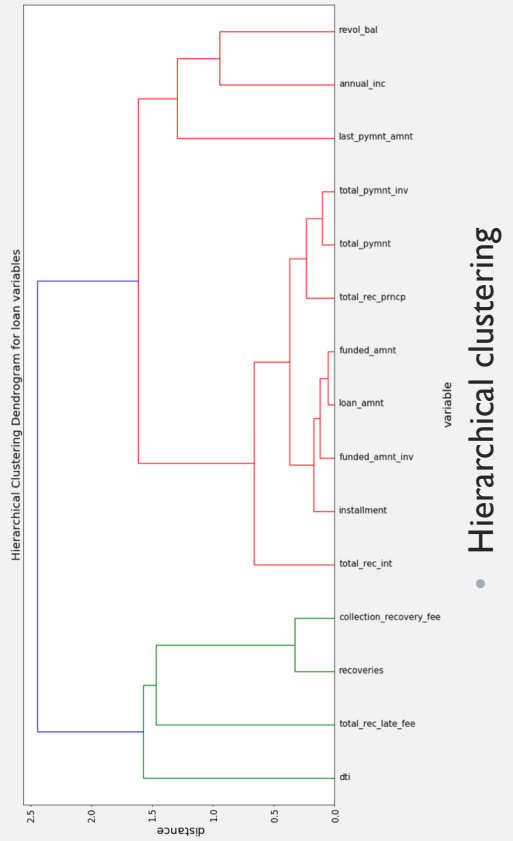
Correlation analysis for numeric columns related to the loan.
DENDROGRAM Analysis



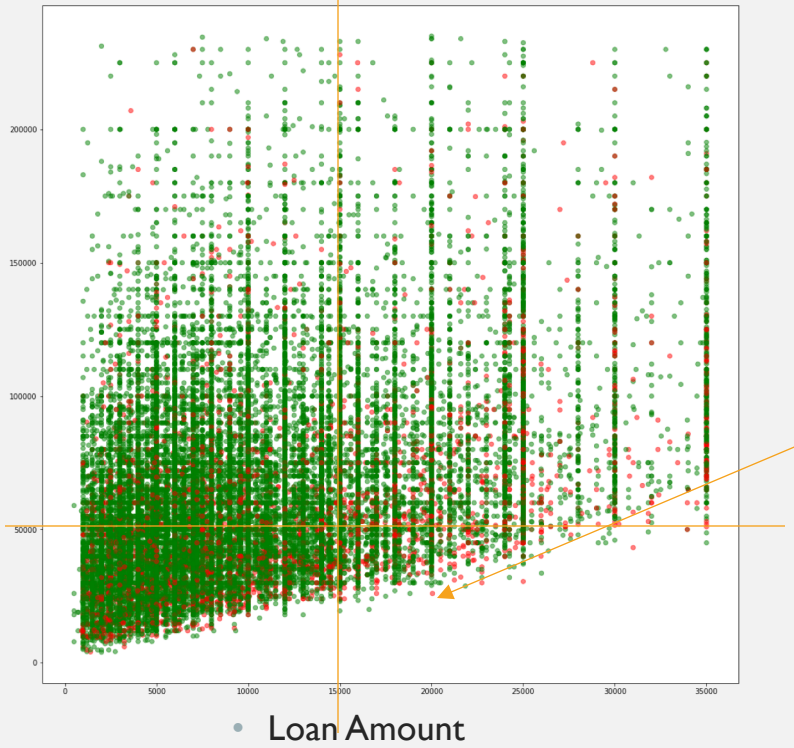
Insights:

- Loan amount related columns like installment, funded amount are highly corelated to loan amount. (This is seen in the top left cluster)
- Annual Income is not corelated with loan amount.
- Debt to income ratio is not corelated with loan amount.

BIVARIATE – CLUSTER ANALYSIS AND CLUSTER PLOT



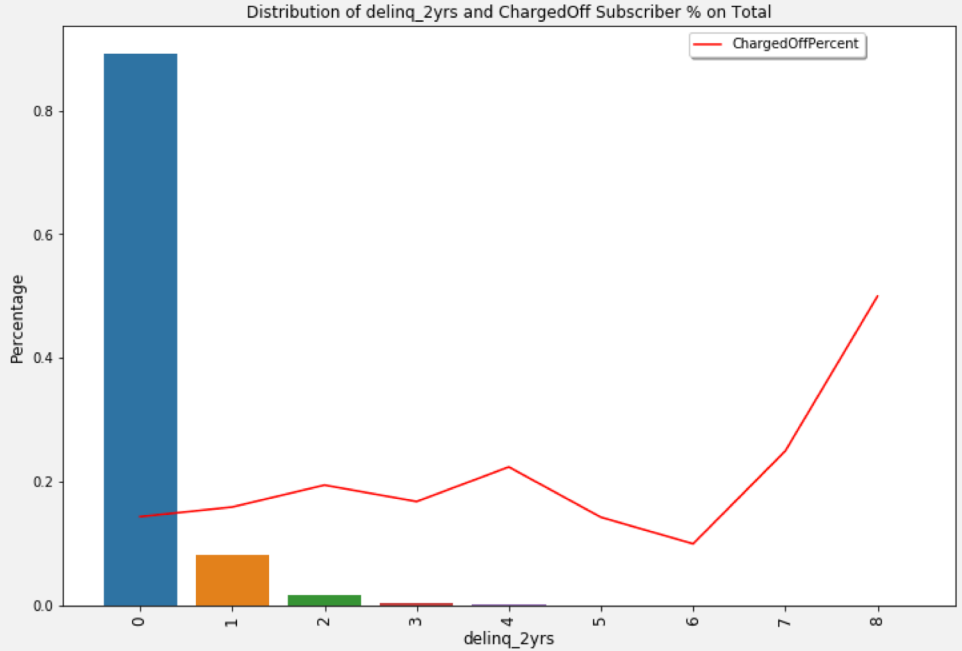
- Hierarchical clustering



- At lower income levels, below 50, 000 USD and higher loan amounts – above 15, 000 USD there are higher defaults.

BIVARIATE ANALYSIS – DELINQ_2YRS

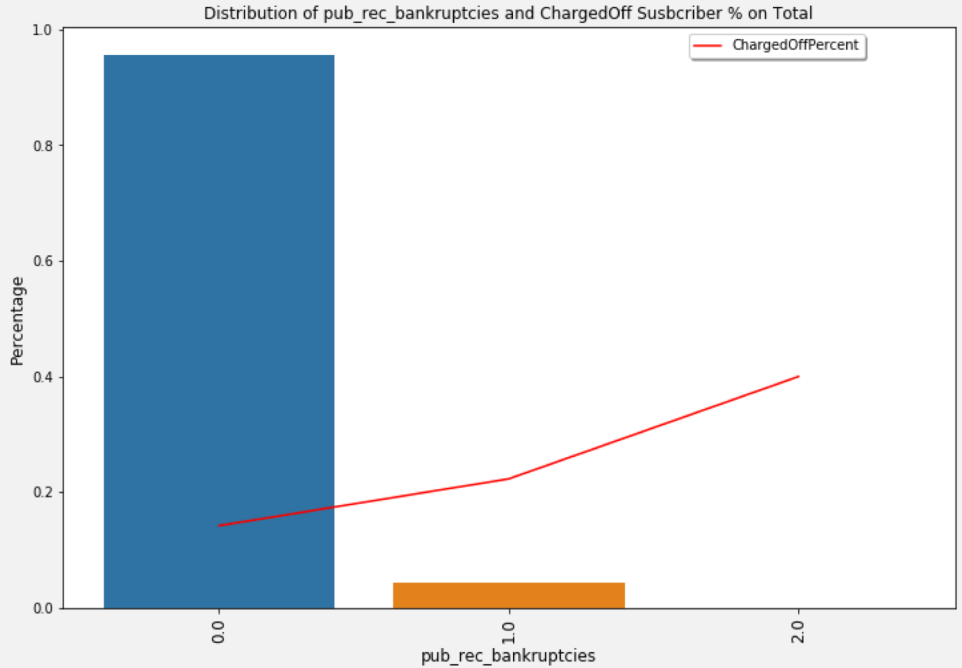
The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years



- 89% of the subscribers have the delinquency instances as 0.
- As delinquencies increase, the charged off percentage increases.
- Data beyond 3 delinquencies is sparse and cannot be considered.

BIVARIATE ANALYSIS – PUB_REC_BANKRUPTCIES

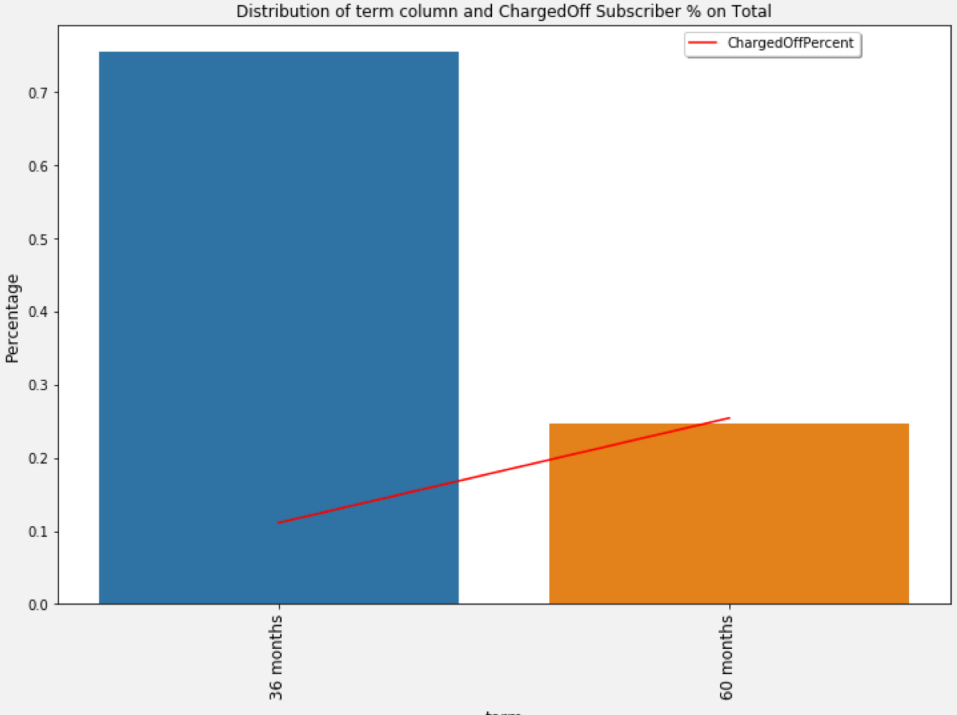
This column indicates the Number of public record bankruptcies



- As the bankruptcies increases from 0 to 1, the charged off percentage increases from 14% to 22%.
- For bankruptcies of 2, the charged of percentage is 40% however the data has only two samples.
- 95.6% of the subscribers do not have bankruptcy records.

BIVARIATE ANALYSIS - TERM

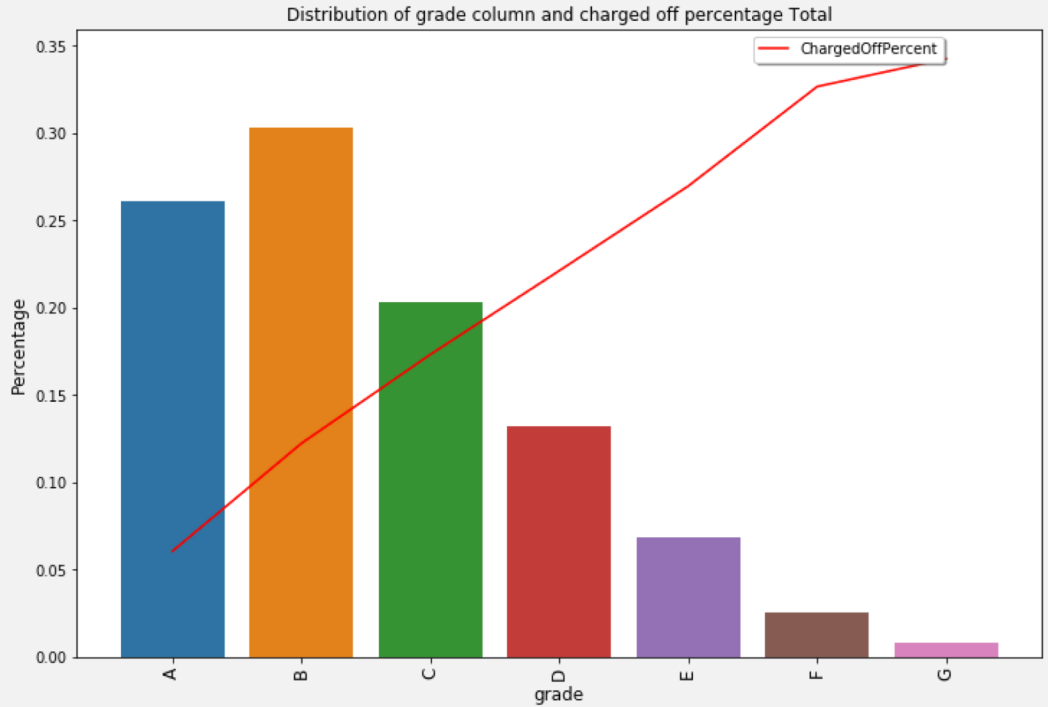
Term indicates The the number of payments months. This can be either 36 or 60.



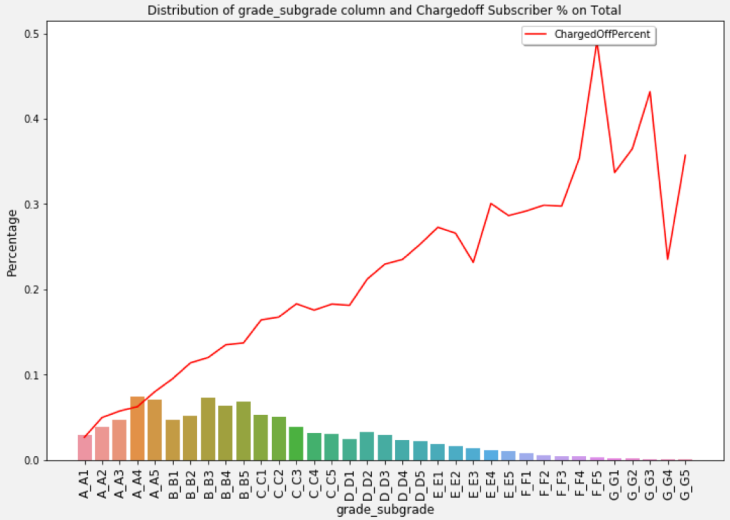
- Higher term leads to higher default rates.
  - 25.42 % of the customers with tenure of 60 months default
  - 11% of customers with tenure of 36 months default

BIVARIATE ANALYSIS – GRADE AND GRADE_SUBGRADE

Grade is a lending club assigned loan grade. This is ordinal categorical variable and interest rate increases grade increases to G.

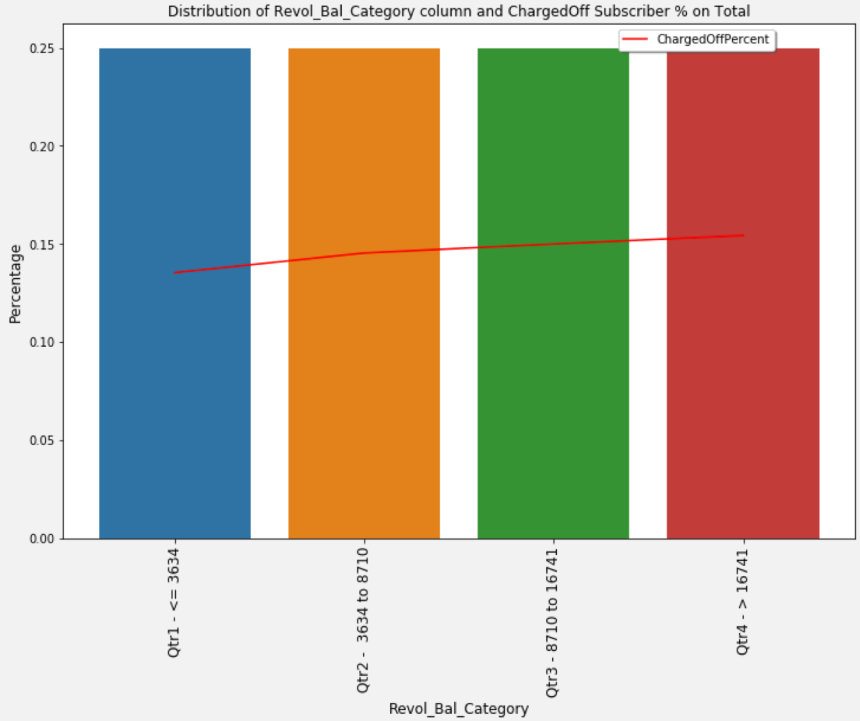


- As loan grade increases the charged off percentage also increases.
- As grade_subgrade increases, the charged off % increases till D_D5, however the trend is not consistent after that.



BIVARIATE ANALYSIS – REVOL BALANCE

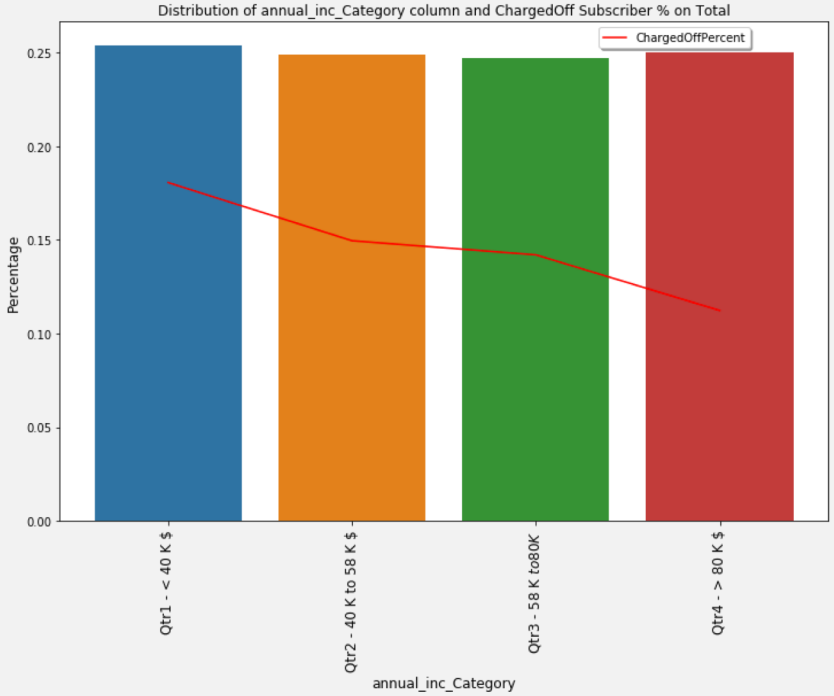
In **credit** card terms, a **revolving balance** is the portion of **credit** card spending that goes unpaid at the end of a billing cycle.



- As revolving balance increases, the charged off percentage increases from 13.5% to 15.4%.

BIVARIATE ANALYSIS ANNUAL INCOME

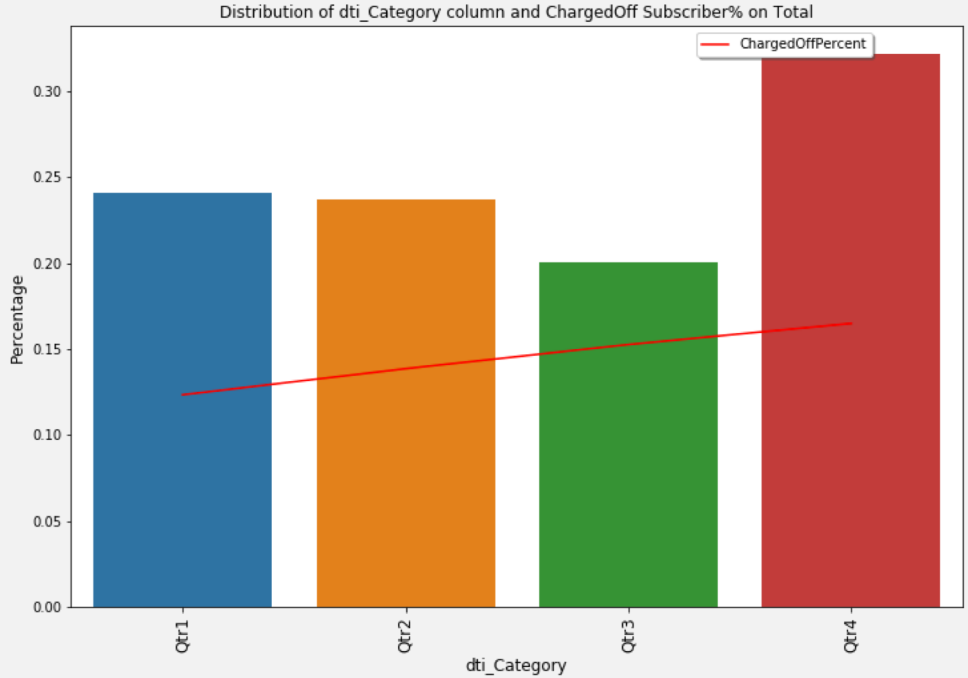
The self-reported annual income provided by the borrower during registration.



- As annual income increases, the charged off % decreases from 18% to 11%.

BIVARIATE ANALYSIS DTI

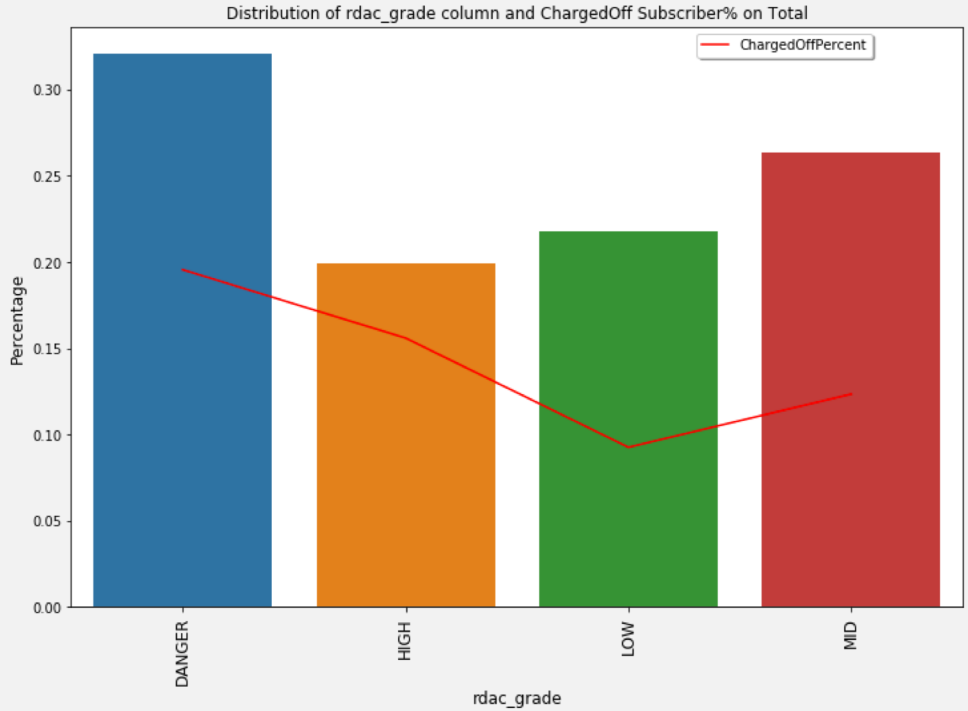
DTI is the ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.



- As dti increases from 8% to above 17%, charged off % increases from 12 % to 16.5%

BIVARIATE ANALYSIS – RDAC

Request Debt to Available Credit (RDAC) is a derived business variable which is a ratio of requested debt to the available credit of the customer.



- RDAC of the DANGER (greater than 2) category has a charged off percentage of 19.5% whereas the LOW category (less than 0.4) is only 9%.

KEY INSIGHTS SUMMARY

Delinquencies ^, Public Recorded Bankruptcies ^, Loan Term ^, Loan Grade ^, Revolving Balance ^ , dti ^ , requested debt to available credit ^ and Annual Income v are the key factors affecting loan being charged off. ^ are positively corelated and v is negatively corelated to loan default.

1. Longer term leads to higher default rates.
2. As delinquencies increase, the charged off percentage increases.
3. Customers with higher bankruptcies have a higher chance of default.
4. The charged off percentage for grade increases linearly as loan grade moves from A to G.
5. As the revolving balance increases, the charged off percentage increases slightly.
6. At lower annual incomes (below 40 K USD) the charged off percentage is 18%.
    1. As the annual income increases, the charged off percentage decreases.
7. Higher term leads to higher default rates.
    1. 25.42 % of the customers with tenure of 60 months default
8. If a borrowers requested loan amount to total available revolving credit is greater than 2, the borrower has a 20% chance of default.

ADDITIONAL INSIGHTS

- The loan amount has spikes at the round numbers like 5, 000,10,000,10,000, 15, 000,20,000,20,000 25, 000 $ etc. This can be an input to the UX team or the Business Development team to create loan packages at these round numbers.
- The Loan amount is corelated positively with the annual income.
- Most of the fully paid loans are with installments between 200 to 400
- Number of loans disbursed to employees with employment experience greater than 10 is the highest.
- Most defaulters are in the Rent and Mortgage category of home ownership.
- Number of loans issues are exponentially increasing over years from 2007 to 2011.
- Most of the customers take the loan for the purpose of debt consolidation and credit card.
- Top states are CA, NY and FL. CA, NY, and TX pay off the most loans, whereas CA, NY and FL default on the most.
- At lower income levels, below 10, 000 USD and higher loan amounts –above 15, 000 USD there are higher defaults.

RECOMMENDATIONS

1. Delinquencies, Public Recorded Bankruptcies, Loan Term, Loan Grade, Revolving Balance, Requested Debt to Available credit and Annual Income are the key factors affecting loan being charged off.
2. Customers with poor financial health as indicated by Delinquencies , Public Recorded Bankruptcies, Revolving Balance, dti and loan amount to credit ratio are the highest risk customers.
3. Avoid giving bigger loans to these high risk customers. Give loans below 15, 000 to these customers.
4. If these highest risk customers fall into the high loan grades of F and G, don't give them loans.
5. Keep the loan term short at 36 months and below.
6. Give loan that is lower than total available revolving credit of the customer (low RDAC).

APPENDIX – LOGISTIC REGRESSION RESULTS

Logistic regression is a **statistical method** for analyzing a dataset in which there are independent **variables** that determine a binary outcome.

- **Formula: 'loan_status_binary = f (annual_inc + loan_amnt + dti + revol_bal + C(income_level) + delinq_2yrs + pub_rec_bankruptcies + funded_amnt + C(term))**
  - **P values less than 0.05 are stastically significant variables affecting the charged off**

Results: Logit

Model:	Logit	No. Iterations:	6.0000
Dependent Variable:	loan_status_binary	Pseudo R-squared:	0.049
Date:	2018-07-29 16:02	AIC:	29633.0648
No. Observations:	37503	BIC:	29735.4509
Df Model:	11	Log-Likelihood:	-14805.
Df Residuals:	37491	LL-Null:	-15575.
Converged:	1.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-1.9604	0.1116	-17.5587	0.0000	-2.1792	-1.7416
C(income_level)[T.LOW]	0.2448	0.1005	2.4360	0.0148	0.0478	0.4418
C(income_level)[T.MID]	0.0148	0.0683	0.2172	0.8281	-0.1190	0.1487
C(term)[T. 60 months]	0.9769	0.0334	29.2578	0.0000	0.9114	1.0423
annual_inc	-2.0937	0.2185	-9.5824	0.0000	-2.5219	-1.6654
loan_amnt	0.1915	0.3394	0.5643	0.5725	-0.4737	0.8568
dti	0.1828	0.0750	2.4376	0.0148	0.0358	0.3298
revol_bal	0.9879	0.1733	5.6994	0.0000	0.6482	1.3276
delinq_2yrs	1.5692	0.3097	5.0661	0.0000	0.9621	2.1763
pub_rec_bankruptcies	1.1315	0.1256	9.0090	0.0000	0.8853	1.3776
funded_amnt	0.4161	0.3484	1.1944	0.2323	-0.2667	1.0990
open_acc	-0.2069	0.1617	-1.2798	0.2006	-0.5238	0.1100