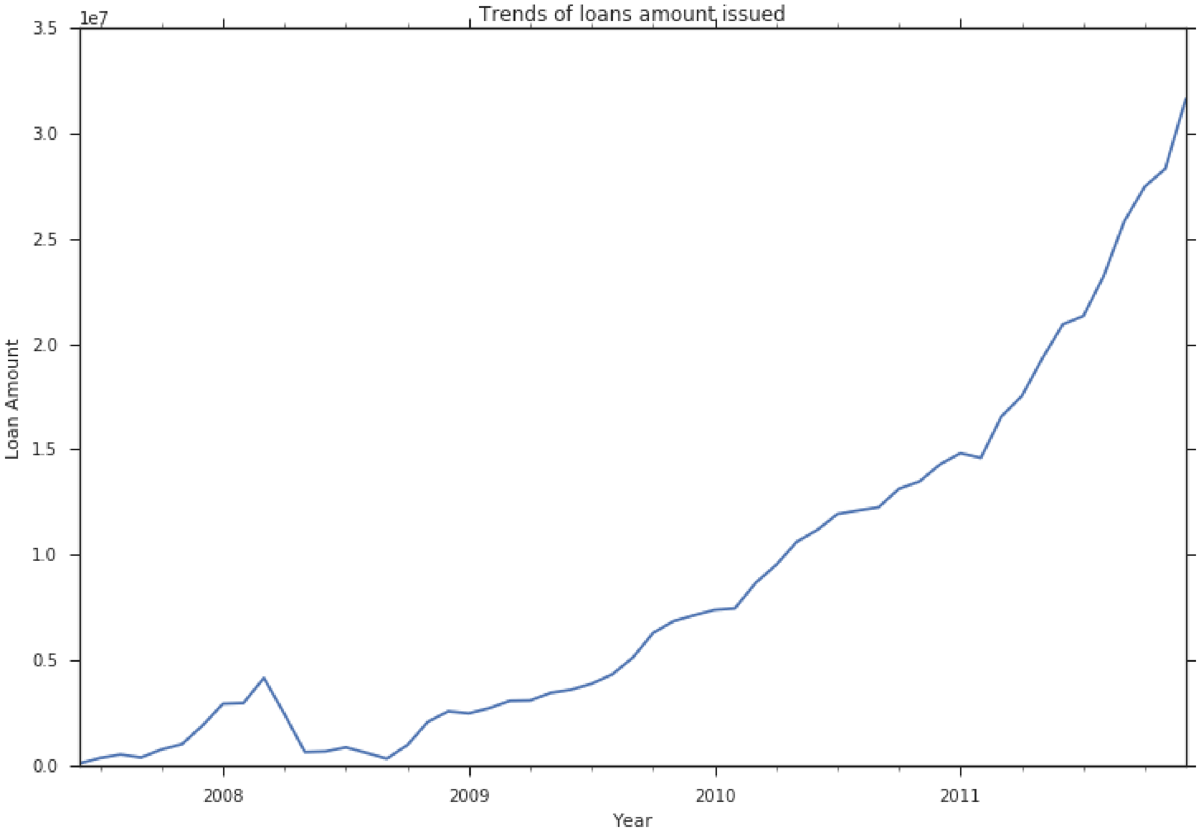
File Name for reference:
Data-Analysis.ipynb

1. Initial Step
   Luigi Script Downloads data, handles missing values and add extra features.

2. Analysis is done on processed file generated

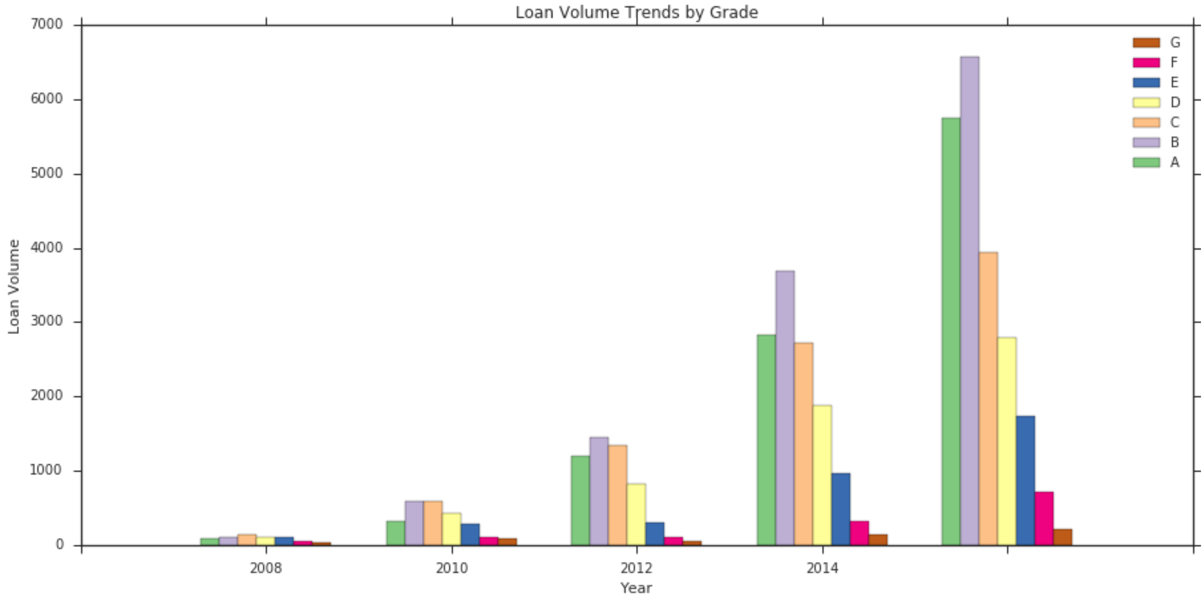
First Analysis on years 2007-2011

1. **Loan Amount vs year- Increase in number of loans issued rapidly. There is a downfall during 2008-2009 due to economic depression.**



2. **Loans Volume trends by Loan Grade and year. Loan Grade tells us how riskier the loan is.**
Loan Grade A is least risky and Loan G is highest.

There is rapid increase in loan B and E grade. Grade will be important parameter to determine if person is going to default loan. Special analysis should be done on parameters of loan grade E, to minimize its chance of defaulting.



The Data Dictionary provided is joined with Data frame columns to understand each column independently.

Analysis of first 19 columns

columns to be removed

id - Randomly genereated Unique Identification Number

member_id - randomly generated field by identification purposes

funded_amnt - Gives future information

funded_amnt_inv - Gives future information

sub_grade - Already defined in grade

int_rate - Already included in grade

emp_title - not much useful unless mapped wiht other information

issued_d - Gives future information

Analysis of next 19 columns

columns to be removed

zip_code - Only first 3 digits available

out_prncp - Gives future information

out_prncp_inv - Gives future information

total_pymnt_inv - Gives future information

Analysis of next 19 columns

columns to be removed

recoveries - Gives future information

collection_recovery_fee - Gives future information

last_pymnt_d - Gives future information

total_rec_late_fee - Gives future information

total_rec_prncp - Gives future information

total_rec_int - Gives future information

last_pymnt_amnt - Gives future information

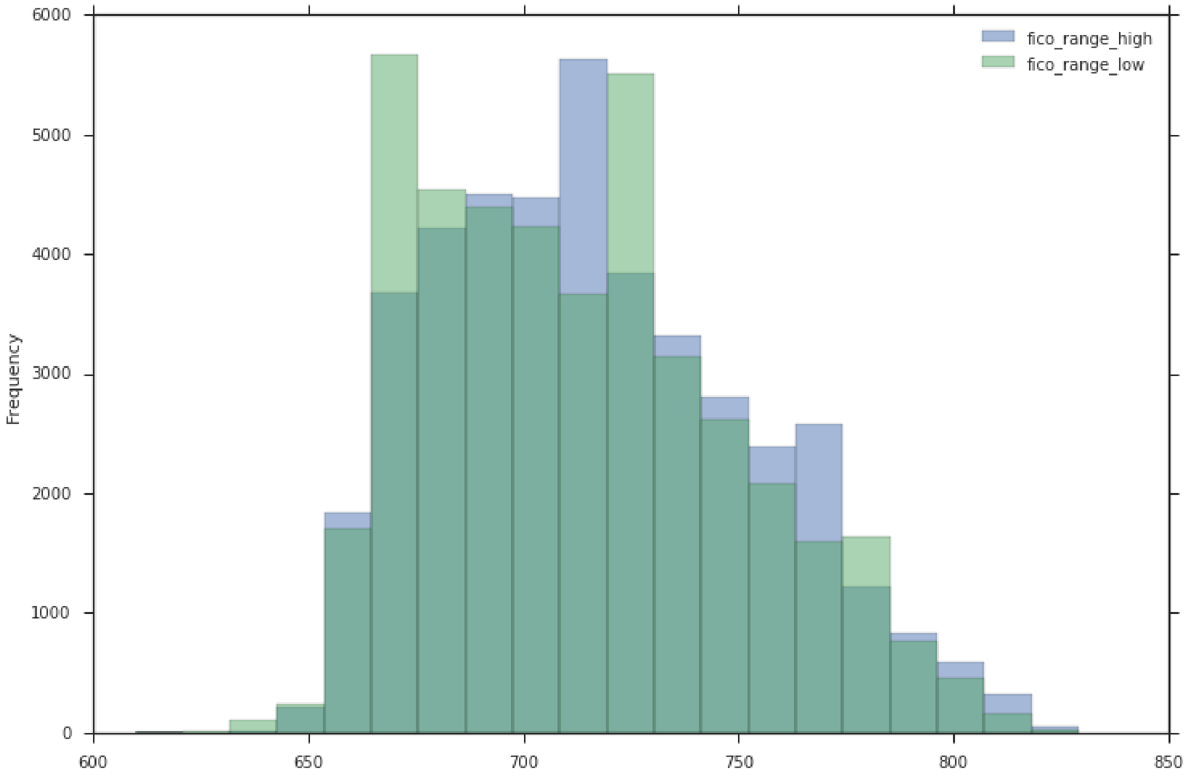
FICO

FICO scores are a credit score, or a number used by banks and credit cards to represent how credit-worthy a person is.

When a borrower applies for a loan, Lending Club gets the borrowers credit score from FICO - they are given a lower and upper limit of the range that the borrower's score belongs to, and they store those values as fico_range_low, fico_range_high.

After that, any updates to the borrowers score are recorded as last_fico_range_low, and last_fico_range_high.
Reference-http://cs229.stanford.edu/proj2014/Kevin%20Tsai,Sivagami%20Ramiah,Sudhanshu%20Singh,Peer%20Lending%20Risk%20Predictor.pdf

FICO Score Distribution



Average of high and low fico score range

fico_range_low	fico_range_high	fico_average	
0	735.0	739.0	737.0
1	740.0	744.0	742.0
2	735.0	739.0	737.0
3	690.0	694.0	692.0

4	695.0	699.0	

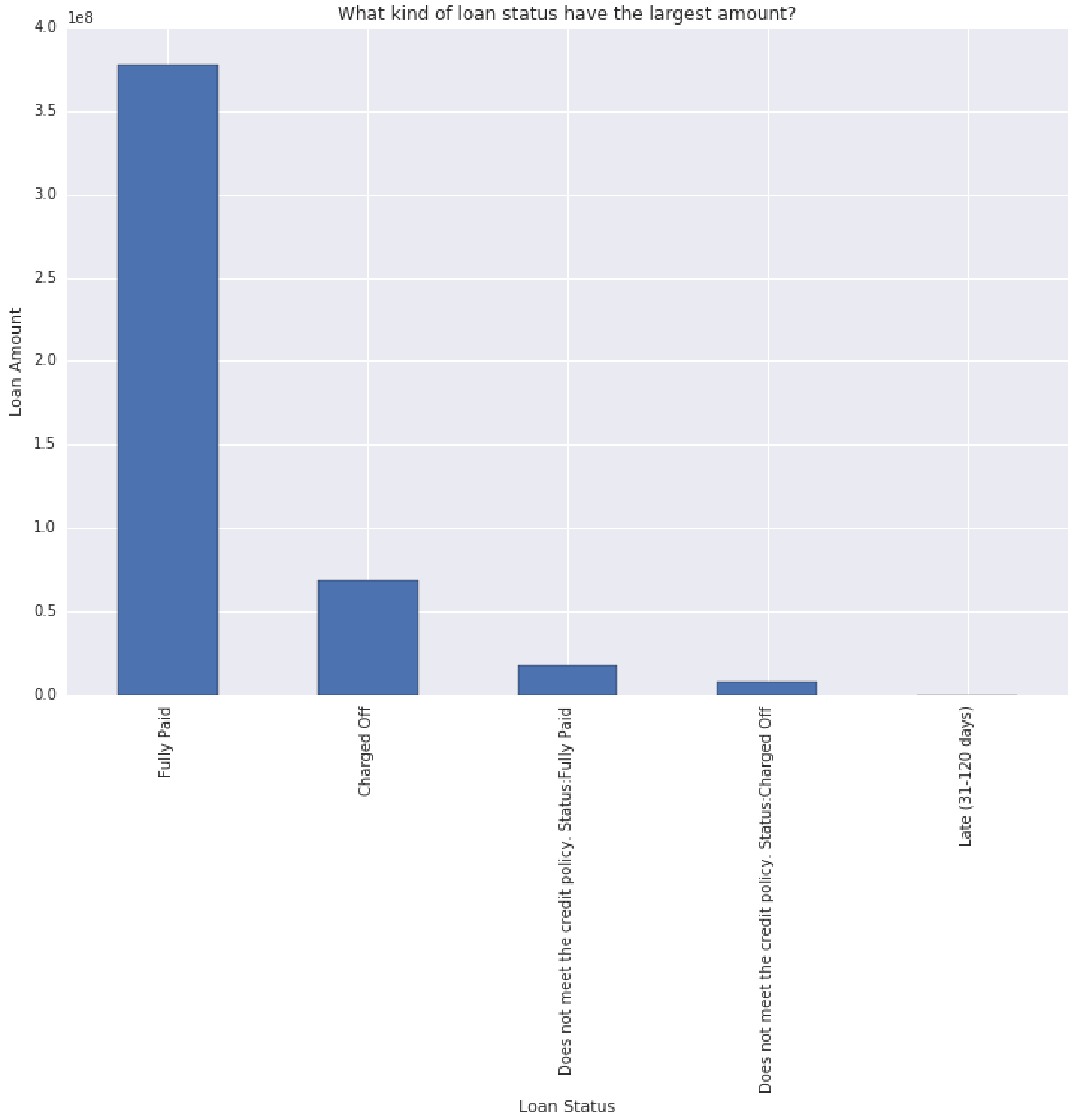
Important to determine who will pay loan and who will default

Loan Status

Meaning for each loan status-https://help.lendingclub.com/hc/en-us/articles/215488038

Loan Status	Count	Meaning	
0	Fully Paid	34115	Loan has been fully repaid, either at the expiration of the 3- or 5-year year term or as a result of a prepayment.
1	Charged Off	5670	Loan for which there is no longer a reasonable expectation of further payments.Charge Off occurs no later than 30 days after the Default status is reached.
2	Does not meet the credit policy. Status:Fully Paid	1988	While the loan was fully paid off, the loan application today would no longer meet the credit policy and wouldn't be approved on to the marketplace.
3	Does not meet the credit policy. Status:Charged Off	761	While the loan was charged off, the loan application today would no longer meet the credit policy and wouldn't be approved on to the marketplace.

4	Late (31-120 days)	1	Loan has not been current for 31 to 120 days.(late on the current payment).



machine learning model goal

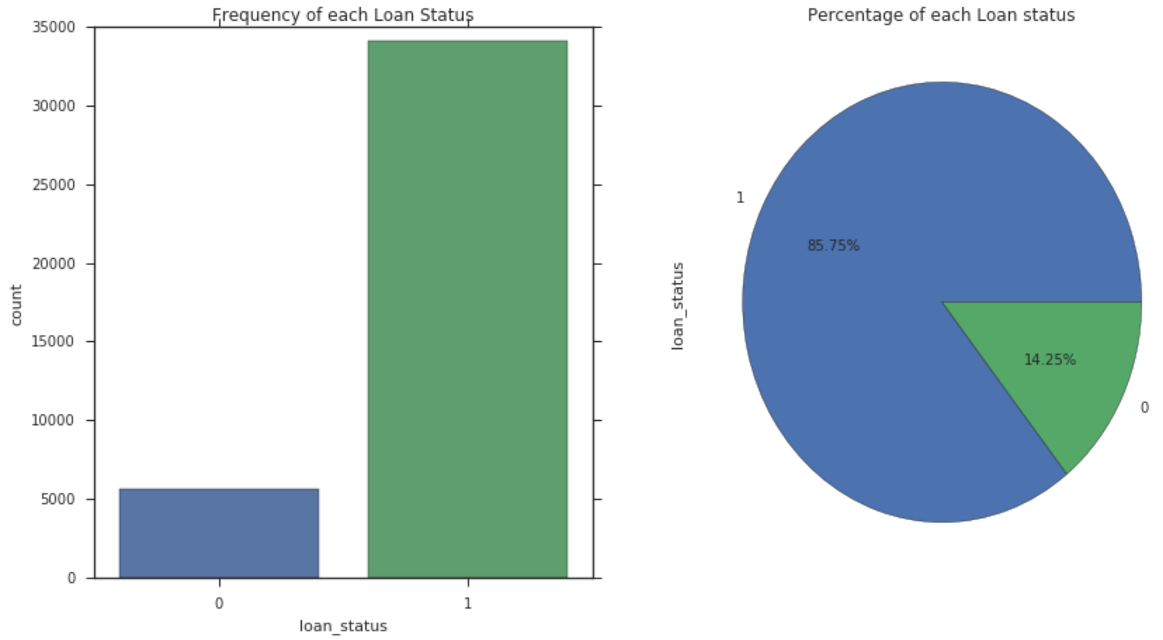
Predict Defaulting loans. From the above table, only the Fully Paid and Charged Off values describe the final outcome of a loan. The other values describe loans that are still on going, and even though some loans are late on payments, we can’t jump the gun and classify them as Charged Off.

Also, while the Default status resembles the Charged Off status, in Lending Club’s eyes, loans that are charged off have essentially no chance of being repaid while default ones have a small chance. Therefore, we should use only samples where the loan_status column is 'Fully Paid' or 'Charged Off'.

We’re not interested in any statuses that indicate that the loan is ongoing or in progress, because predicting that something is in progress doesn’t tell us anything.

Since we’re interested in being able to predict which of these 2 values a loan will fall under, we can treat the problem as binary classification.

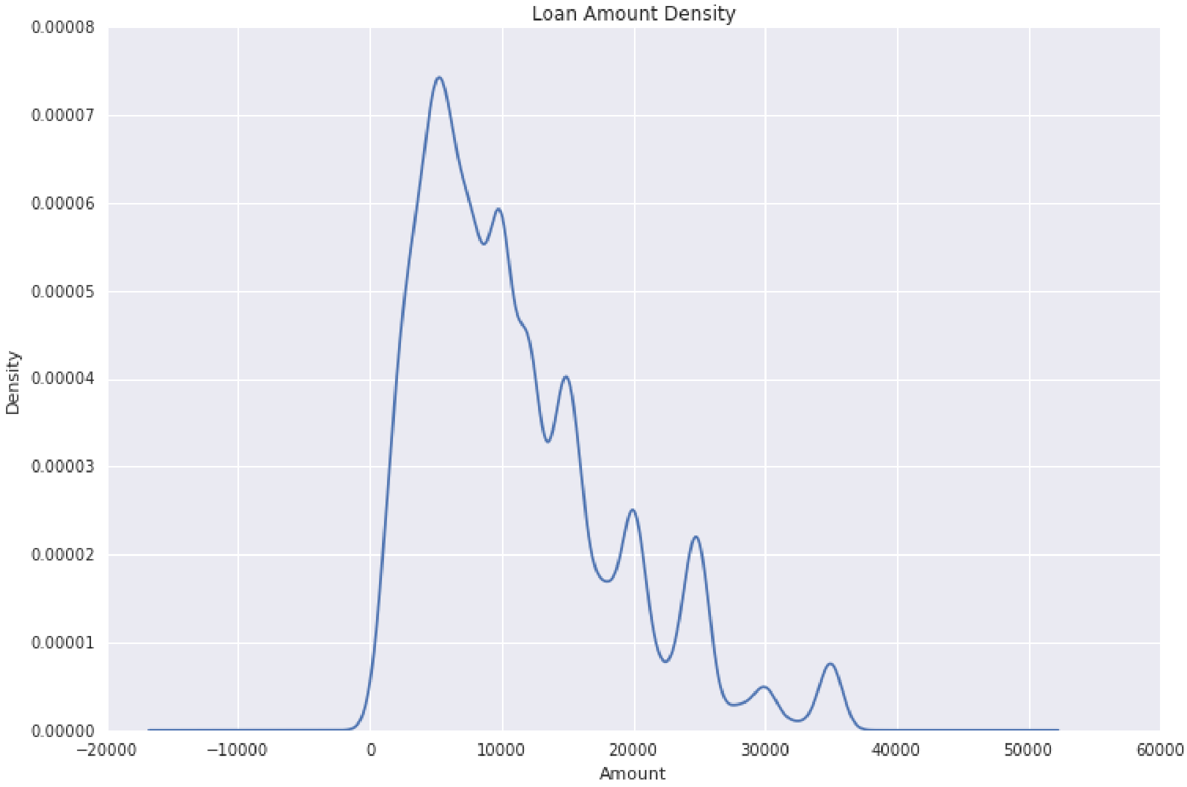
Let’s remove all the loans that don’t contain either 'Fully Paid' or 'Charged Off' as the loan’s status and then transform the 'Fully Paid' values to 1 for the positive case and the 'Charged Off' values to 0 for the negative case.



These plots indicate that a significant number of borrowers in our dataset paid off their loan - 85.75% of loan borrowers paid off amount borrowed, while 14.25% unfortunately defaulted. So now our interest is in these defaulters

any columns that contain only one unique value and remove them. These columns won’t be useful for the model since they don’t add any information to each loan application

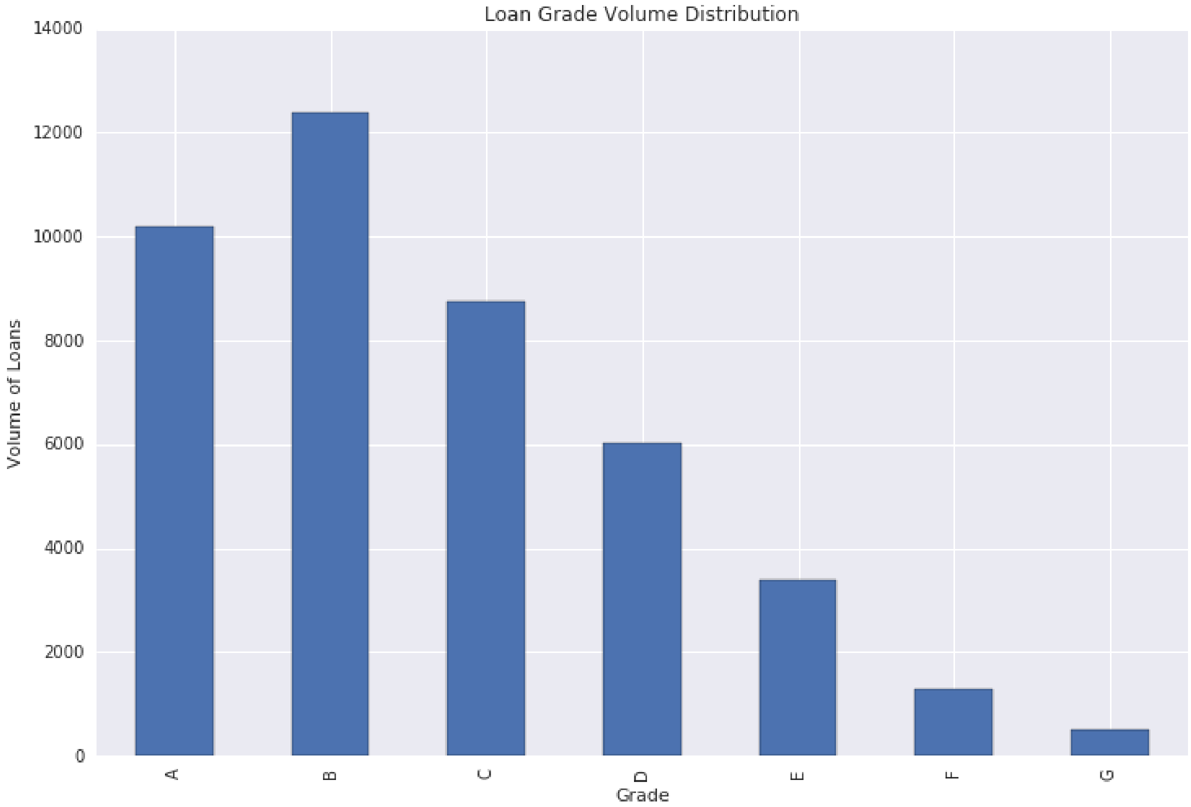
Loan Amount Density



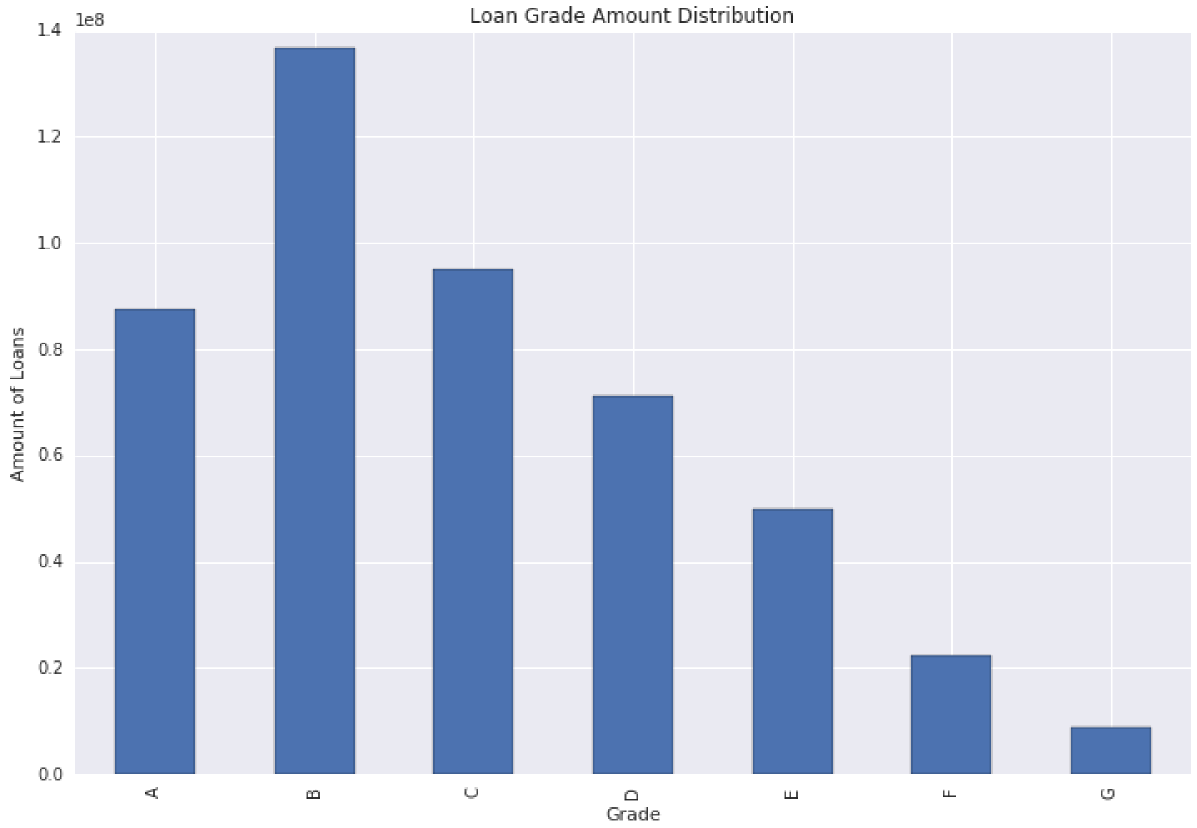
Loan Title

New
Payoff
Card Payoff
Medical
Debt Free
Loan Debt
Home Improvement
Pay Debt
Pay Credit
Loan Personal
Wedding
payment
Card Refinance
Car Loan
Loan Pay
Debt Loan
Bill
Credit Card
Small Business
Debt Consolidation
Wedding Loan
Freedom
Car
Card Loan
CC
Consolidate Debt
Vacation
Business Loan
Refinance
Loan Consolidation
high interest
Consolidation Debt
Loan Wedding
Loan Credit
Card Consolidation
Improvement Loan
House
Major Purchase
Debt Debt
Loan Loan
Card Debt
interest credit
Help
Loan Home
Need
Consolidation Loan
Consolidation Personal
Personal Loan
Lending Club
Loan Consolidate

Loan Grade distribution

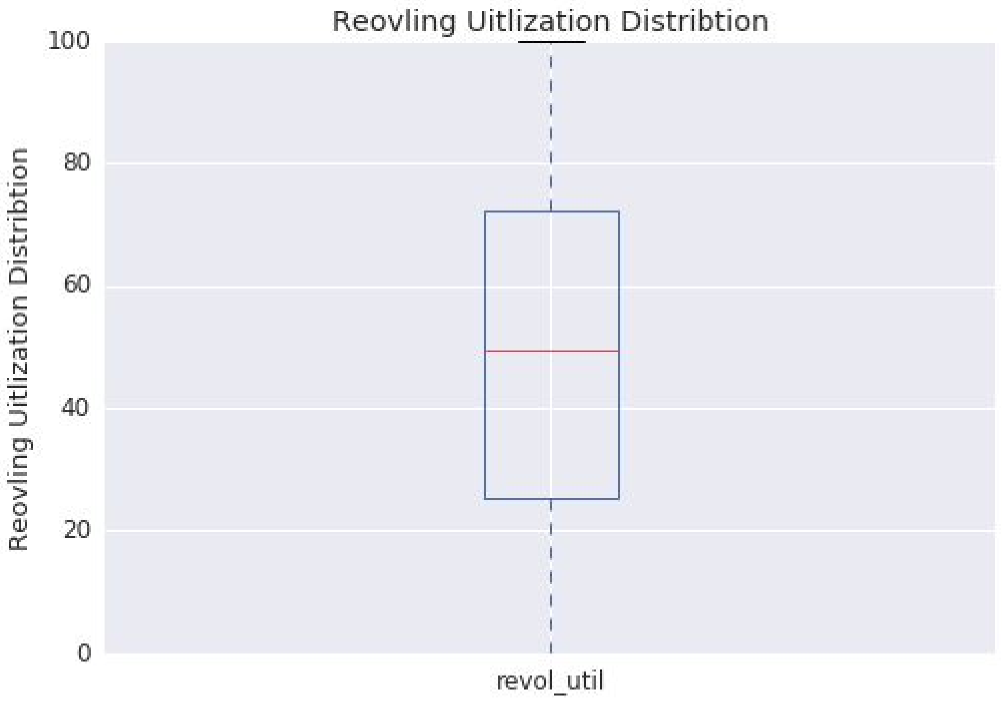


Loan Grade distribution with loan amount



Revolving line Utilization-Important feture

revol_util is a revolving line utilization rate or the amount of credit the borrower is using relative to all available credithttp://blog.credit.com/2013/04/what-is-revolving-utilization-65530/



Rejected loans Analysis

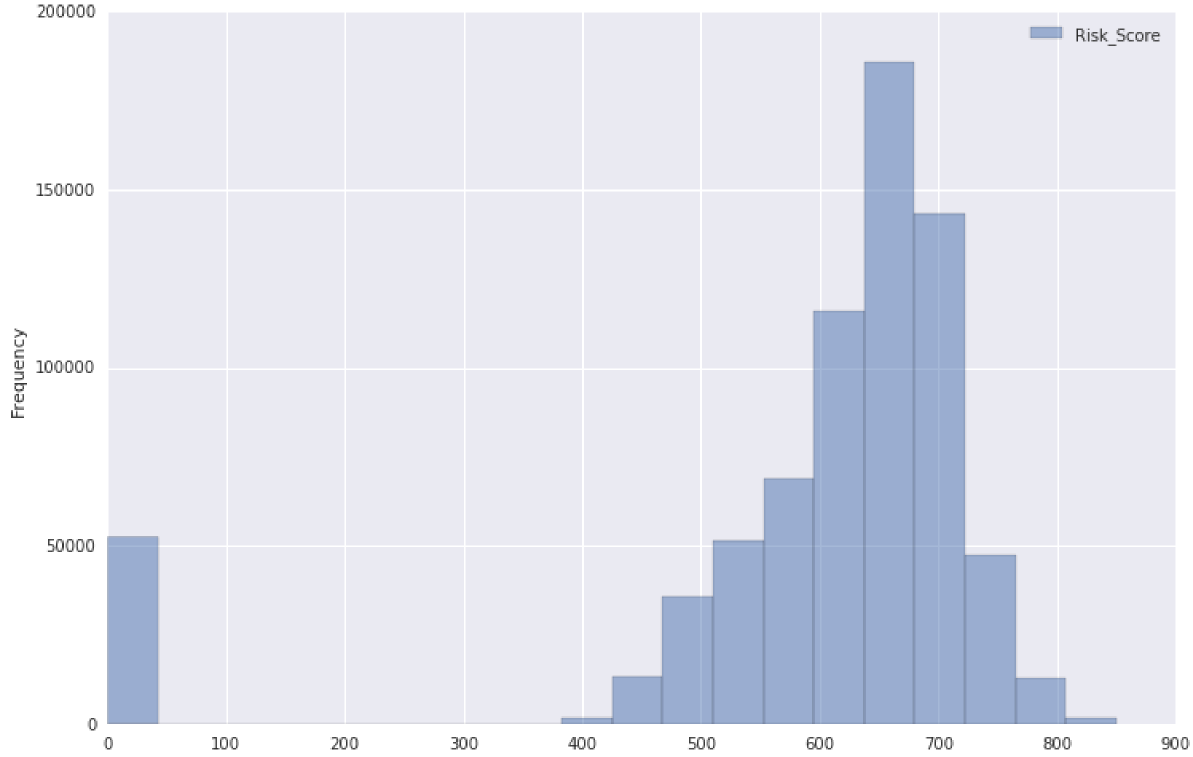
As rejected loans so policy code zero. Not needed can be dropped

Zip code has only 3 digits. And we have state so can be removed

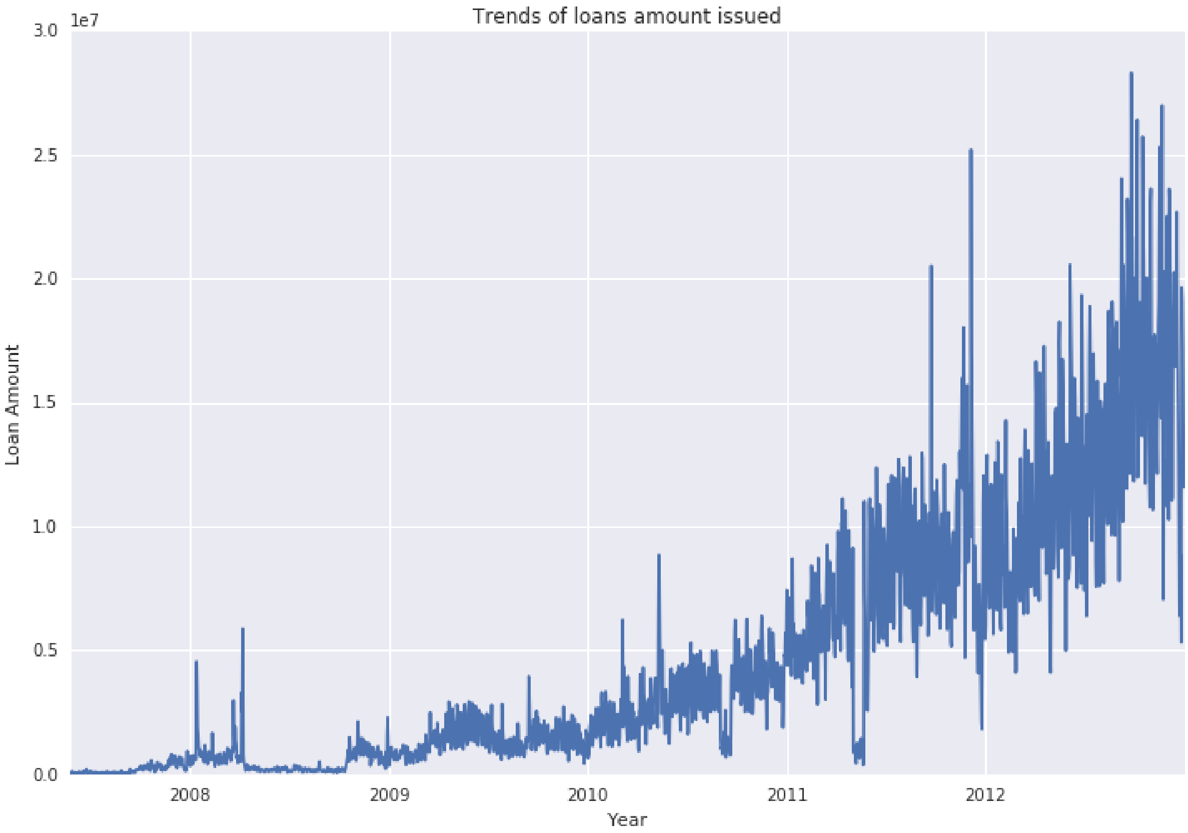
Loan Title visualization

small_business debt_consolidation
debt_consolidation home_improvement
debt_consolidation moving
debt_consolidation car
debt_consolidation major_purchase
medical debt_consolidation
debt_consolidation debt_consolidation
debt_consolidation medical
home_improvement debt_consolidation
debt_consolidation small_business
car debt_consolidation
credit_card debt_consolidation
debt_consolidation credit_card
major_purchase debt_consolidation
moving debt_consolidation
debt_consolidation
credit_card

Risk Score Distribution- One with Zero are ones who did not report their credit score.



Amount Requested Over the years- High rate of increase in loan rejection and amount requested also increased.


Trends of loans amount issued
1e7
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Loan Amount
2008
2009
2010
2011
2012
Year