**LENDING CLUB DATA ANALYSIS**

**Data download:**

Lending club dataset has pre and post login option to download data. Post login, you can get the entire data which is tagged secure.

Mechanical soup was used to automatically login and download data. Mechanical soup has the combined features of request and BeautifulSoup.

The links for the secure files were found in a hidden div, pipe separated.

Once the data is downloaded, it is timestamped after reading and creating dataframes.

These dataframes are merged into a single csv for feature engineering.

**Missing data Analysis:**

The dataset consists of around 118 features out of which there are a few features which consists NaN values. By setting a threshold of 75%, if more than 75% or records are null, then they have been dropped.

Each feature is analyzed based on its distribution and an appropriate strategy is chosen to handle missing values and derive additional information from it. The data has been cleaned to ensure uniformity.

The following are a few techniques for handling missing values:

· Forward Fill

· Backward Fill

· Replace by Mean value

· Replace by Maximum occurring value

· Drop the record

Eg: Missing ‘Application Date’ is fetched using forward fill method. Missing ‘Term’ is replaced by the maximum occurring value. If ‘Interest Rate’ is missing the record is dropped.

*Note: Refer to the ‘*[DataCleaning&FeatureEngineering](https://github.com/jainpranj/lending-club-data-analysis/blob/master/DataCleaning_FeatureEngineering.ipynb)*’ notebook*

**Feature engineering:**

A few additional columns were derived from the data which influences the calculation of the interest rate.

1. As per an article written by [Credit Karma](https://www.creditkarma.com/article/age-of-credit-history), the age of the credit history is a fundamental factor in influencing a user’s credit score. Since credit score is an important aspect for calculating the interest rate, calculating the credit age would be useful for determining the borrower’s credit quality.

To compute this parameter we created a derived column ‘**CREDIT AGE**’, we utilized the columns ‘earliest\_cr\_line’ and ‘last\_credit\_pull\_d’ which gives the Month and Year when the borrower’s credit line began and latest credit pull.

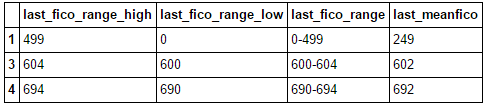
**CREDIT AGE= year of (earliest\_cr\_line) - year of (last\_credit\_pull\_d)**

2. As per an article written by [myFICO](http://www.myfico.com/credit-education/questions/how-do-inquiries-impact-credit-scores/), the number of inquiries done on the borrower has a key impact in their interest rate. Thus we bucketed our borrower’s on the basis of the number of inquiries in last 6 months. This is derived using the column ‘number\_of\_inquiries\_in\_last\_6\_months’ (inq\_last\_6mths).The following were the buckets:

|  |  |
| --- | --- |
| **Number of inquiries** | **Grade** |
| 0 | A |
| 1-2 | B |
| 3-6 | C |
| 7-10 | D |
| 10+ | E |

Note: These thresholds are set by [Credit Karma](https://www.creditkarma.com/question/hard-inquiries-how-many-is-to-many)

3. Since FICO score is an essential component in determining the interest rate. Using the Minimum and Maximum FICO score we are creating two derived columns, **MEANFICO\_score** and **FICO\_RANGE.**



*Note: Refer to the ‘*[DataCleaning&FeatureEngineering](https://github.com/jainpranj/lending-club-data-analysis/blob/master/DataCleaning_FeatureEngineering.ipynb)*’ notebook*

Thus the following are the list of derived columns:

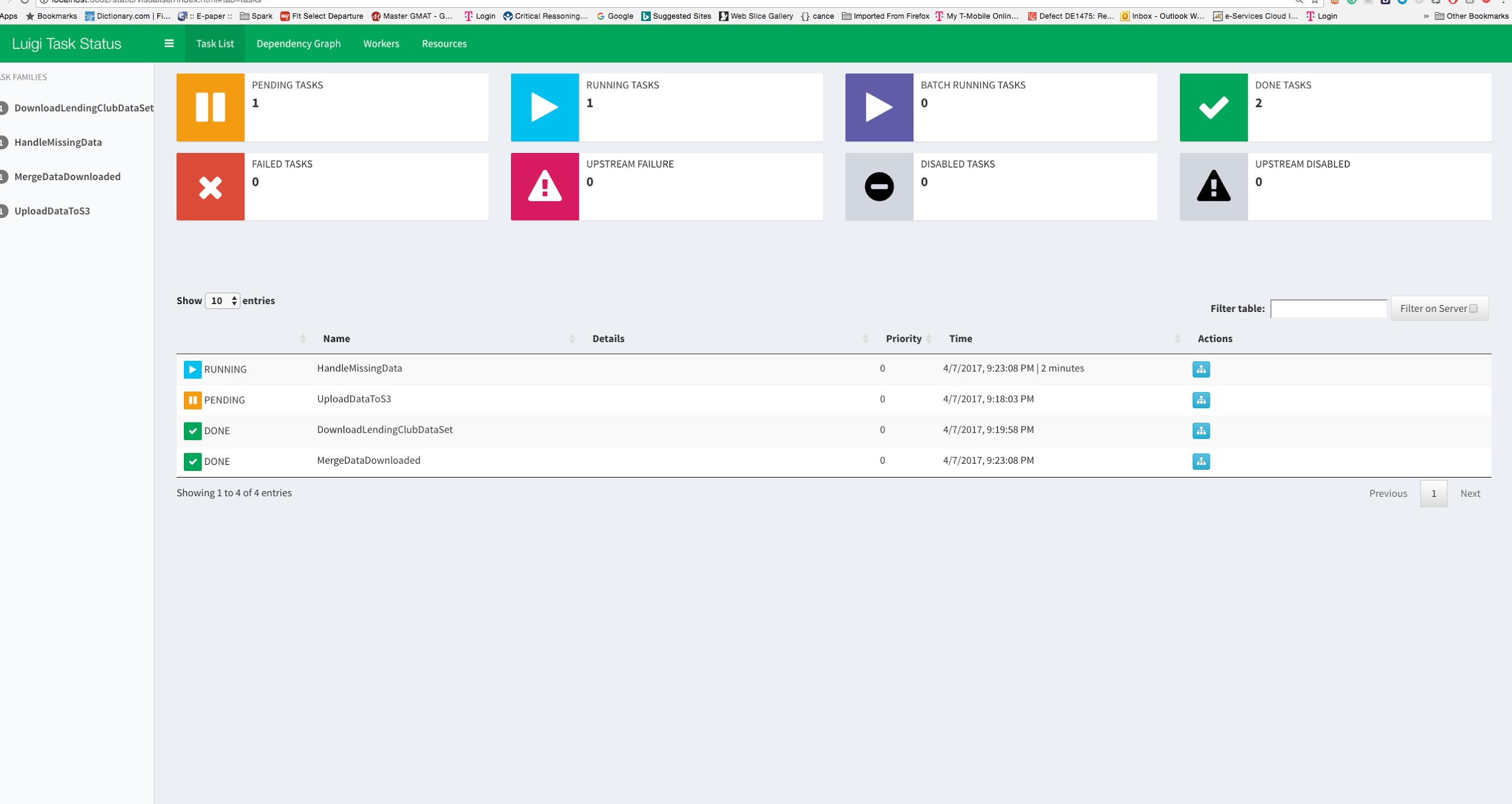
1. Credit\_age
2. grade\_based\_on\_number\_of\_inquiries
3. ~~fico\_range~~
4. ~~meanfico~~
5. ~~last\_fico\_range~~
6. ~~last\_meanfico~~
7. fico\_range\_high, fico\_range\_low, last\_fico\_range\_high, last\_fico\_range\_low

**Pipeline:**

Screenshots of luigi UI

Showing running and done tasks





The same steps were followed for Rejected/Declined Loan data except that it did not require any login for complete data

The credentials of the Lending Club login and the access keys for s3 upload are being taken as parameters during docker run image command. After this, no human intervention is required. The pipeline ends with the processed csv uploaded to s3.

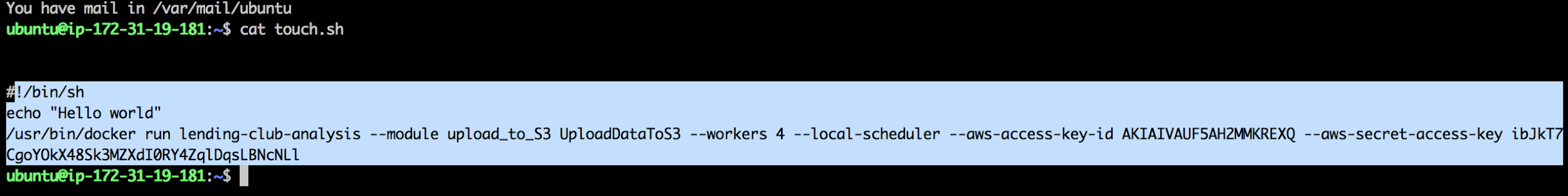
**Dockerizing:**

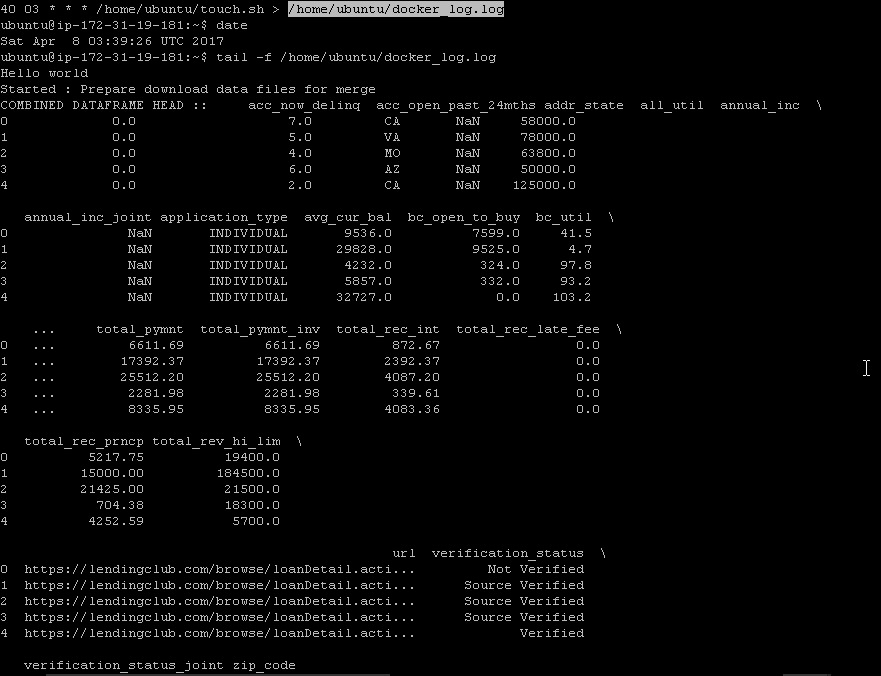
The docker image is simple using Dockerfile. A standard project structure for luigi project is followed.

**Scheduling the pipeline:**

It is scheduled using CRON job which runs every 30 mins using crontab.

A shell script runs the docker image with luigi pipeline.The output of cron job is in /home/ubuntu/docker\_log.log





**Exploratory Data Analysis :**

File Name for reference:

Data-Analysis.ipynb

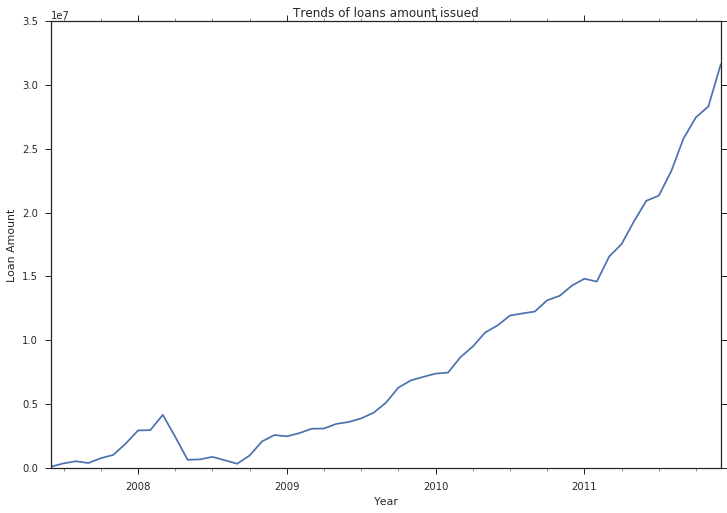
1. Initial Step

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

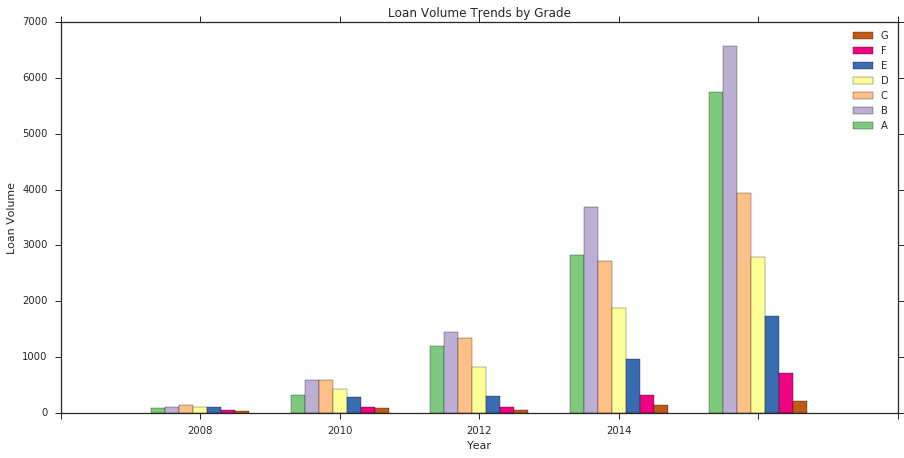
1. **Loan Amount vs year- Increase in number of loans issued rapidly. There is a downfall during 2008-2009 due to economic depression.**



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

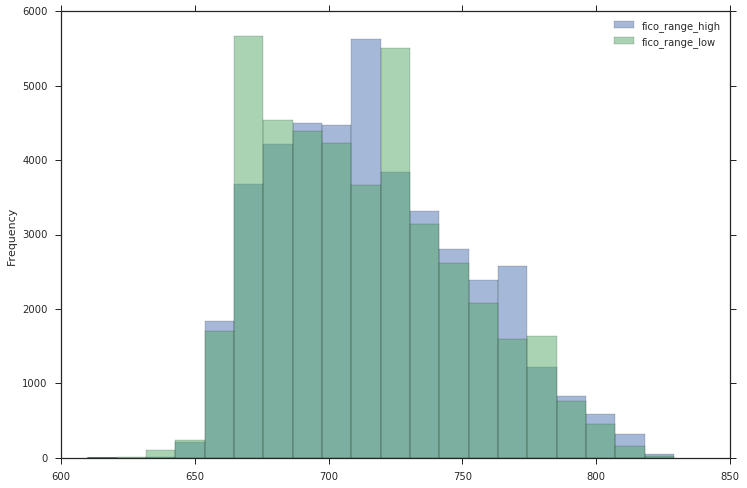
last\_pymnt\_amnt - Gives future information

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**

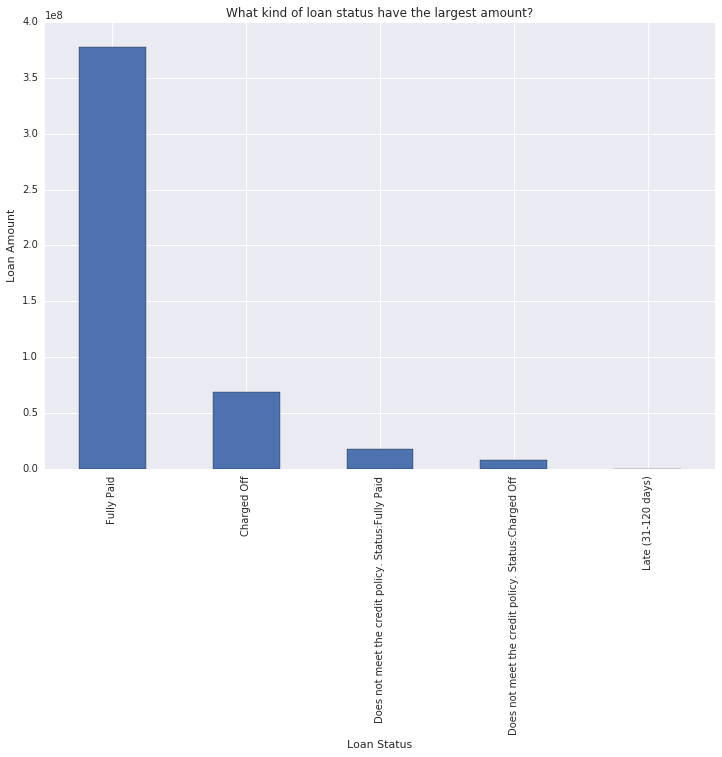
|  |  |  |  |
| --- | --- | --- | --- |
| **fico\_range\_low** | **fico\_range\_high** | **fico\_average** |  |
| **0** | 735.0 | 739.0 | 737.0 |
| **1** | 740.0 | 744.0 | 742.0 |
| **2** | 735.0 | 739.0 | 737.0 |
| **3** | 690.0 | 694.0 | 692.0 |
| **4** | 695.0 | 699.0 |  |

# **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**](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**

# 

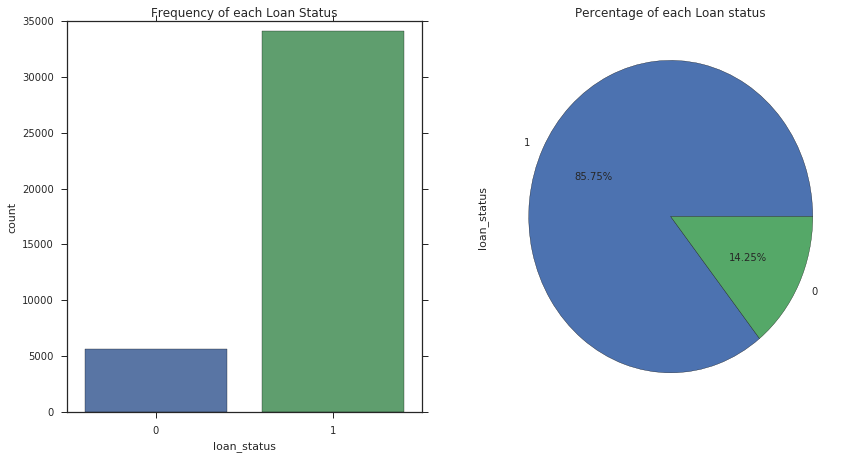
Predict Defaulting loans. From the above table, only the Fully Paid and Charged Off values describe the final outcome of a loan. The other values describe loans that are still on going, and even though some loans are late on payments, we can’t jump the gun and classify them as Charged Off.

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.

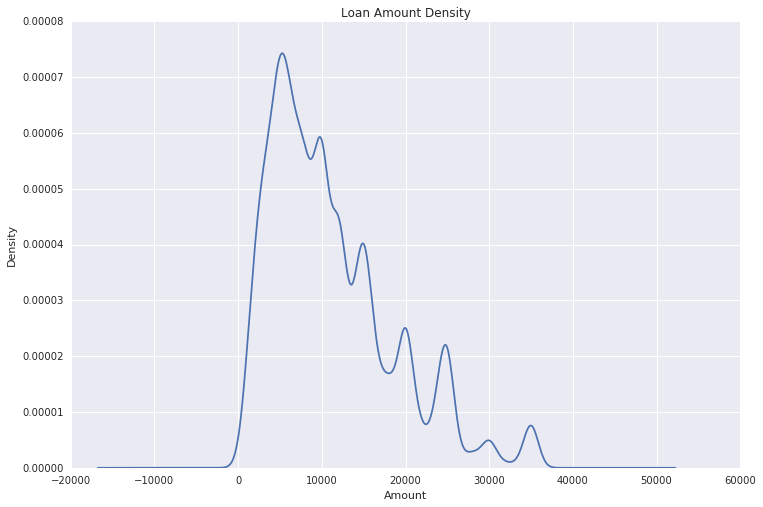
Let’s remove all the loans that don’t contain either 'Fully Paid' or 'Charged Off' as the loan’s status and then transform the 'Fully Paid' values to 1 for the positive case and the 'Charged Off' values to 0 for the negative case.



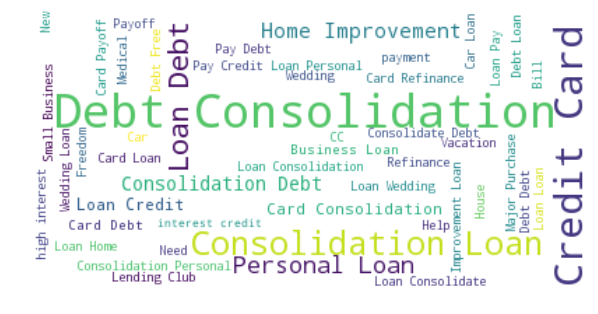
### **These plots indicate that a significant number of borrowers in our dataset paid off their loan - 85.75% of loan borrowers paid off amount borrowed, while 14.25% unfortunately defaulted. So now our interest is in these defaulters**

#### **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**

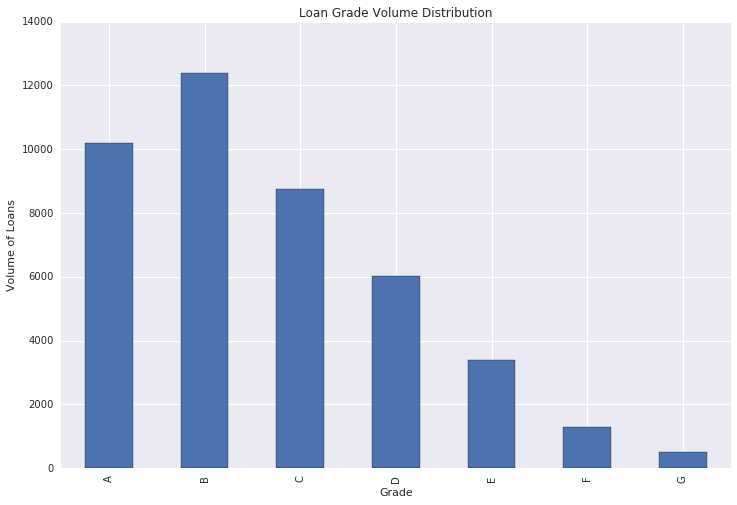
Loan Amount Density



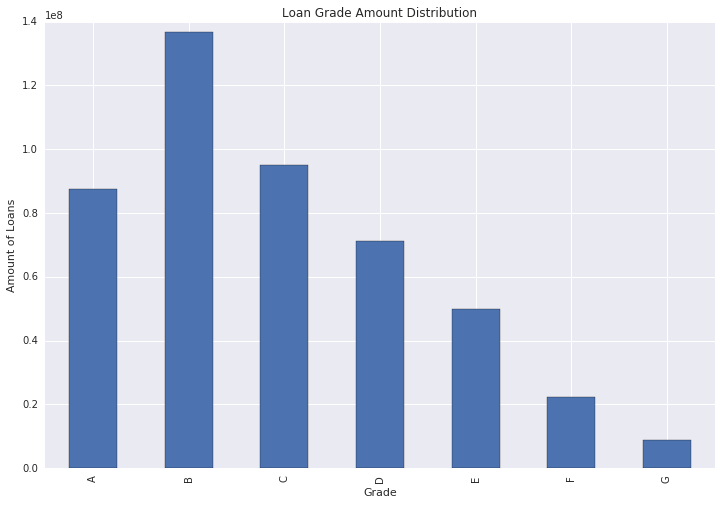
Loan Title



# **Loan Grade distribution**

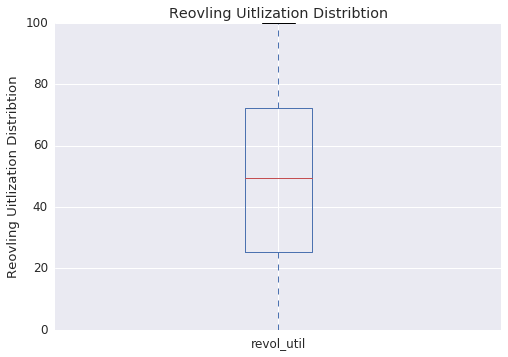


# **Loan Grade distribution with loan amount**



# **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 credit<http://blog.credit.com/2013/04/what-is-revolving-utilization-65530/>



# **Rejected loans Analysis**

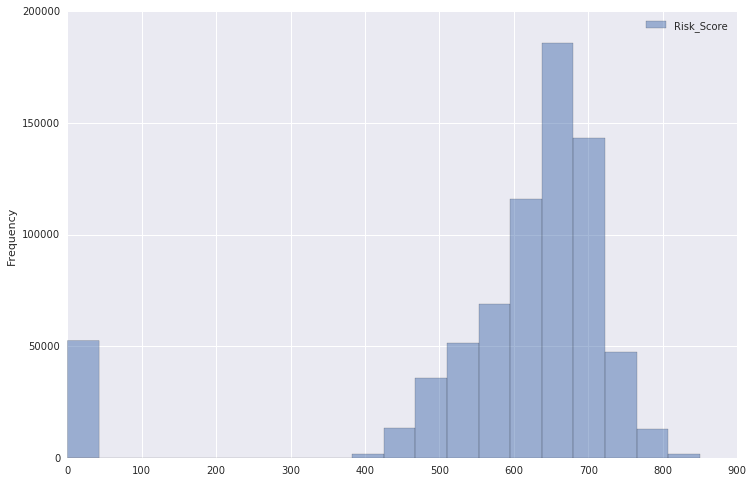
# **As rejected loans so policy code zero. Not needed can be dropped**

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

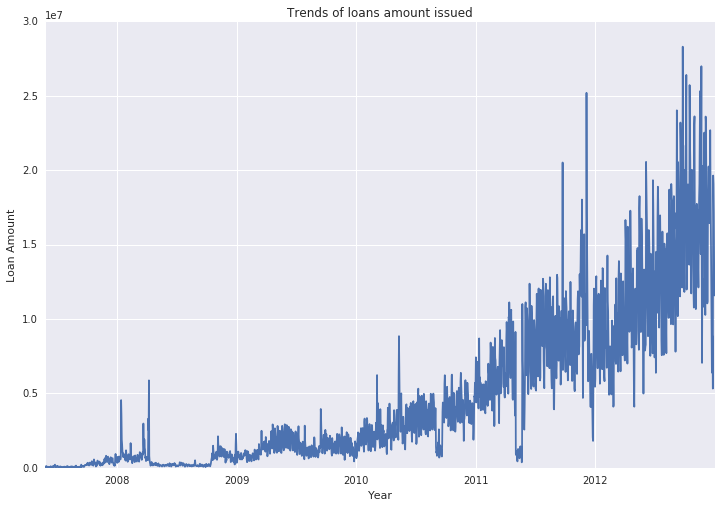
Loan Title visualization



**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.**



POWER BI:

<https://app.powerbi.com/view?r=eyJrIjoiNmMzZmY3ZTAtMTRiNy00NzAyLTgzZTgtM2NiNjMyZmQzZGVjIiwidCI6IjZhYmZjNzNmLWRhNjQtNDEzNy05ZjlmLTE1ZmFhZTU2ZjY4NSIsImMiOjN9>

References

[1] [hickford/MechanicalSoup]: A Python library for automating interaction with websites (https://github.com/hickford/MechanicalSoup)

[2]<https://www.creditkarma.com/article/age-of-credit-history>

[3]<https://www.creditkarma.com/question/hard-inquiries-how-many-is-to-many>