DATA VISUALIZATION REPORT

LENDING CLUB: A VISUAL ANALYSIS

12/18/2017

Akshay Deshpande
Chani Thakkar
Shun Yao
Xinbo Zhou

Table of Contents

Abstract

As institutionalized loans become harder to avail following the 2008 Subprime mortgage crisis, certain individuals remain deprived of the opportunity to take a personal loan in spite of being amply able to repay it, simply because they do not fulfill the criteria set by the banking institutions for personal loans. This has given birth to the concept of Lending Club - (www.lendingclub.com), a peer to peer online marketplace that allows borrowers to apply online for personal loans of up to 40,000$. Lending Club allows Investors to invest their capital on such applicants for a high return on investment.
Lending Club has been operational since 2007 and ever since has been keeping a meticulous record of all the loan applications that have been accepted or rejected by through their mediation. As this data has been made public by them on their website, it begs to be analyzed in depth to find hitherto unseen trends in non-institutional loans. By joining these datasets based on the geo-socio-economic information along with an applicant's loan repayment history, we can provide important indicators to investors regarding Lending Club loan applications currently under their consideration.

Introduction

Given the non-institutional nature of these loans, there is risk of loss of investment involved and it would not be a wise decision for small and medium investors to invest without conducting an in-depth analysis of this investment opportunity. To this end, Lending club has released datasets for loans and rejections form 2007 onwards. The lack of the knowledge of analytical tools and the knowhow of data analysis has given rise to the need for an analysis tool that can be used by a layman. This analysis is our first step towards the development of such a tool. This undertaking being an Exploratory Data analysis, could reveal relationships within the data that could be leveraged by banking institutions or institutional investors to have a surplus return on investment. With a more short-term goal, we as small investors could also profit from investing on the right individuals based on the results of our analysis.

Design Process

Lending Club has made two datasets (for every year from 2007) publicly available on their website for investor to perform research. These are:
LoanStats.csv – Information about all loan applications.
RejectStats.csv - Information about rejected loan applications.
These datasets, available here, (www.lendingclub.com/info/download-data.action) furnish data from 2007 onwards and although they are readily available, they have not been adequately exploited to maximize the return on investment(RoI) by investors due to

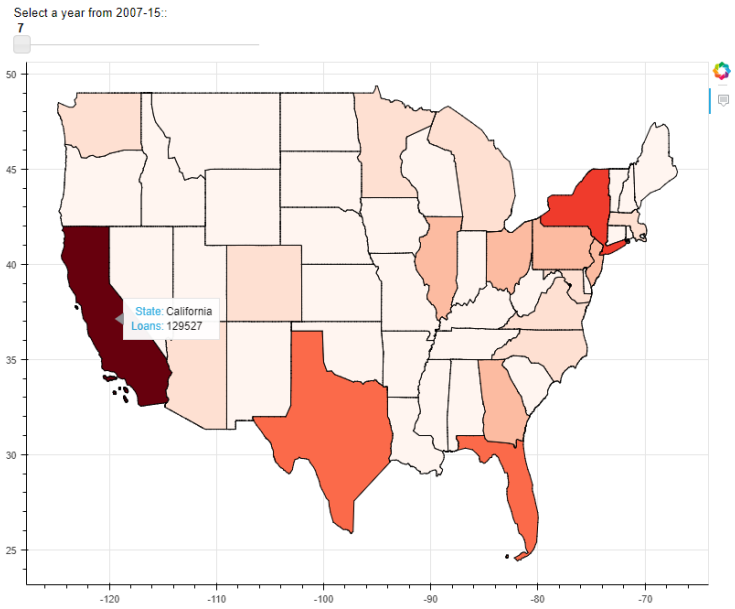
the lack of means or the knowledge to perform a comprehensive analysis of these multivariate datasets. As we set out exploring the dataset we noticed that:

- The files for years were in asymmetrical chunks. Some files were over multiple years while the recent years 2016 and 2017 are in the form of quarters.
- The Loans data contains well over 100 columns. Most of which contained null values. Also, we deemed that many of these columns were not useful for our initial Exploratory Data Analysis.
- The Data Dictionary provided contained many deprecated columns and has not been maintained by the Lending club authorities.
- All the datasets together form 2007- 2017 are a total of 1.06 GB of data. This could make our program incredibly slow if loaded all together.
- The Member Ids have been removed to conserve the anonymity of the members of lending club.
- In order to counter these issues, we first started with:
- A thorough exploration of the dataset.
- Formulation of questions based on our understanding of the dataset.
- Identifying the variables to answer these questions while conduction this EDA.
- Deciding the strategy of loading only the columns of interest to us into the data frames.
- Dropping null values to make to program faster by limiting the memory requirements.

These steps paved the way for the commencement of our Exploratory Data Analysis of the lending club dataset.

Results of the user study

As we started working with the dataset using python with bokeh, we quickly realized that the dataset is quite large, and the amount of time required for the initial loading of the data frames was quite long. Concatenating the dataset over the decade only made things worse. The whole system drew to a screeching halt and it became abundantly clear that the hardware configuration of the machine we were using was not up to the mark for such an undertaking. Nevertheless, we pushed on, at times including only one data file to make sure that our premise was working.



For example, the first casualty of the user testing was the year slider for the choropleth maps. It made map update incredibly slow, so we preferred dropping it and using four static maps with hover tooltips instead of with different coloring methods i.e. Equal intervals and using Natural breaks.

For Loan Data we concluded :

The first thing is to not read the first row, so I use skiprows=[0] and the data read properly. Also I tried the drop function, it didn't work. Fortunately, skiprows works.

Examined is it worth to count 'emp_title', used d1['emp_title'].nunique() and find 3550, used d1['emp_title'] and find 96781, not worth.

Examined d1[d1['loan_amnt'] != d1['funded_amnt']], only 2 rows different. They are basically the same.

Examined d1['home_ownership'].nunique() and d1.groupby('home_ownership')['home_ownership'].count(), only five categories: ANY, MORTGAGE, NON, OWN RENT

Examined d1['revol_bal'].max() and d1['revol_bal'].min()

Examined d1['purpose'].nunique() and d1.groupby('purpose')['purpose'].count(). I did the same for 'title'. The finding is they are basically the same.

For Rejection Data we concluded:

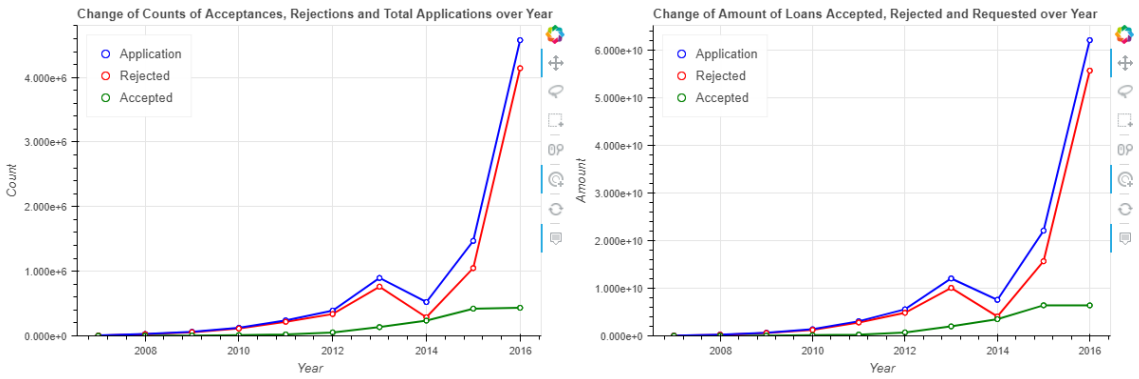
When it's time to analyze the Declined loan data (rejected data), my first observation is it's similar to Loan data with fewer columns, though the amount of useful and necessary data to create visualizations remains the same.

As with Loan data, the first thing is to not read the first row, so I use skiprows=[0] and the data read properly. Also I tried the drop function, it didn't work. Fortunately, skiprows works.

To union the data, I can use either append or concat, I choose concat since it is designed for union and the code looks cleaner and nicer.

Results of our EDA

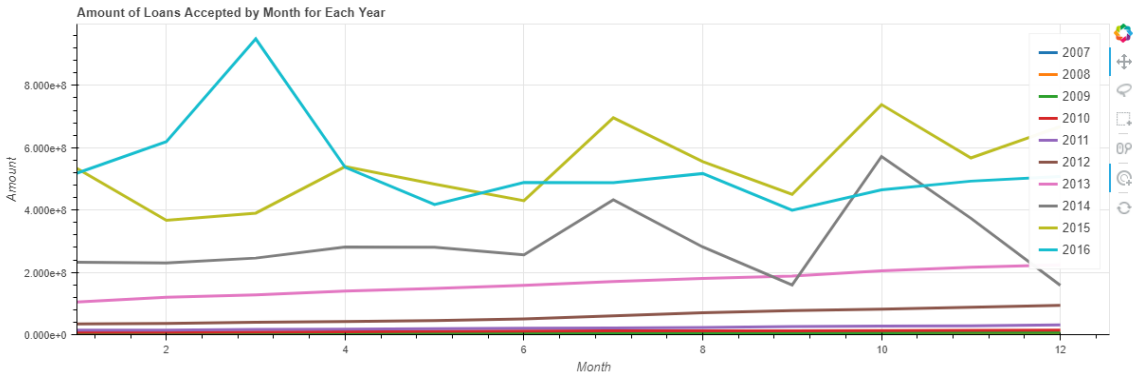
Tab 1:



The first two charts show the overall view of change of counts and amount of loans accepted, rejected and requested over year, which is the beginning of our exploratory data analysis. It includes data from both loan and reject datasets. Although a stacked bar chart can also be used to show these data, multi-line chart is more straightforward for users to observe the change and trend.

As you can see, the lines in these charts are basically in the same shape, which indicates that the amount might be highly related to the counts. The gap between the amount of loans requested and accepted, which is shown by the red rejected line, is very large. This reflects a very common phenomenon in the lending market for both banks and other P2P lending marketplaces that the demand for capital is always far more than the supply. And the situation might be more server in the near future. Therefore, institutions are using very strict evaluation criteria and procedures to select the most reliable and profitable applicants in the large amount of applications. This led to the further analysis in the factors that impact the result of the loans applications.

There is a clear growing trend for the company's business. However, compared with the rapid growth of the counts and amount of loans requested, the counts and amount of loans accepted increased much more slowly. Thus, if the company can increase the return of the lending and enhance their propagation of the returns to attract more investors and lenders, it can better bridge the gap between the demand and supply.



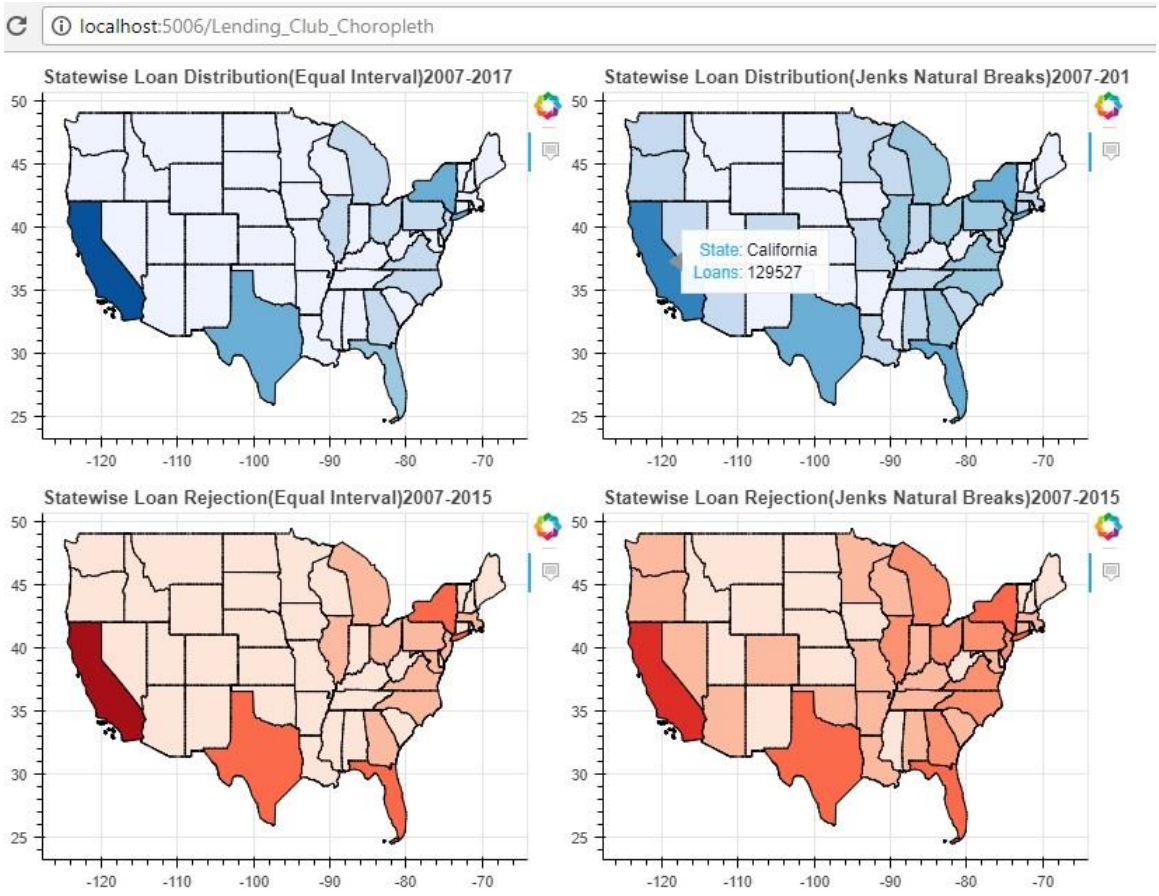
The third chart is designed to find out whether there is any pattern of the amount of loans accepted within a year. Since the amount of loans accepted highly depends on how much investors would like to invest, this chart could also be used to explore lenders' investing cycles and to assess the company's performance in attracting lenders. I thought about using acceptance rates (amount accepted/amount requested) to do this analysis, but then I figured out that the amount accepted does not rely on the amount requested but mainly depends on the investors' willingness to make investments, and that since the amount requested is continuously increasing, the line might be a slope down for each year, which doesn't make a lot of sense in the pattern analysis. So, I continued with using the amount of loans accepted.

The amount is grouped by months and each color represents a year. Generally speaking, there is a growing trend on amount accepted over months. But there are unexpected ups and downs in the lines as well. Lending Club could learn from this chart and dive deeper to analyze to see if there's any triggers for those specific drops or increases. For example, there were three drops in Septembers during the recent three years, is that caused by the reason that most of investors(families) have a certain amount of cash outflow during that period, so they don't have spare money to make investments. Questions like this might be aroused by the chart, so that the company can conduct more researches to find out the causes and to make better use of the features and patterns of the market.

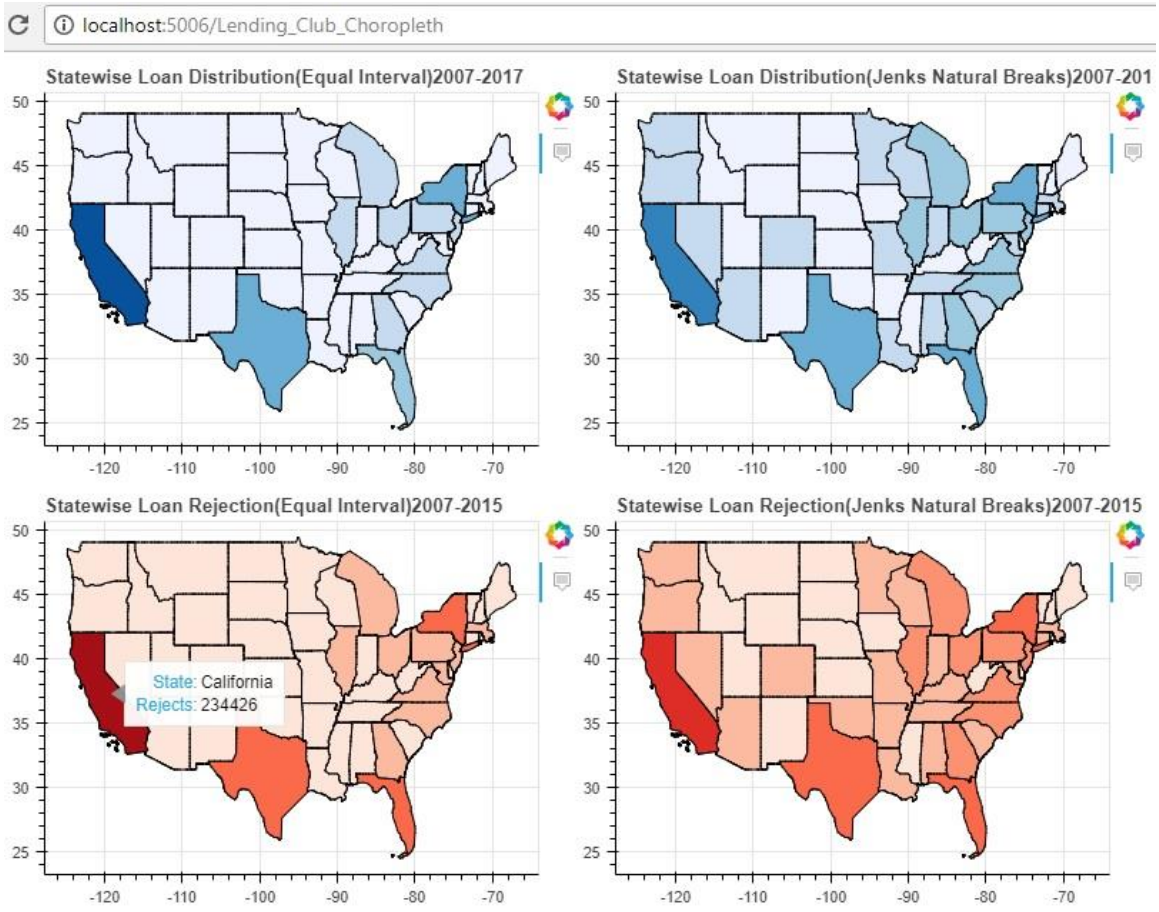
Since the y axis ranges for all of the charts are large, I added the wheel zoom-in tool for them so that by clicking on the legend and zooming in the y axis only, users can have a close and clear observation of the pattern for each year.

The problem I met when creating these charts was that the code ran very slowly because of the large amount of the data. Especially when I was trying to convert all date values from object to datetime type, it took me almost half an hour to finish the conversion. Realizing this, I attempted to lower the amount of data by first grouping the data by dates so that I could get distinct dates. This effectively reduced the time I need for converting them into datetime.

Tab 2:



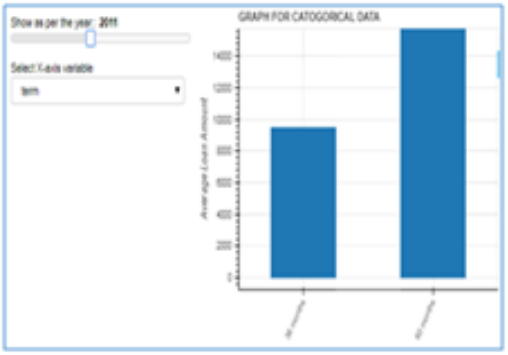
The Blue maps represent the state wise loan distribution from 2007 to 2015. The blue map on the left represents the equal intervals method implemented using the Linear color mapper. As we can see the whole of the Midwest region seems to have the same shade in spite of difference in the number of loans awarded in those states. While California is a dark blue because of the highest number of loans awarded to applicants form that state. This discrepancy has been fixed in the blue map on the right-hand side. It uses the Fisher-Jenks method i.e. Natural breaks using the pysal library. This shows a more equitable distribution of the shades which is closer to the reality. Even the states in the Midwest display different shaded based on the bins in which they were place. These maps show that California is clearly the state that is more accepting of the concept of lending club and investors are more open to the idea of investing in an online peer to peer lending network.

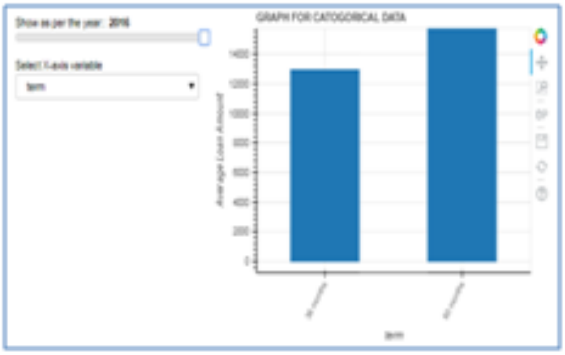


The red maps represent the state wise loan rejection from 2007 to 2015. The red map on the left represents the equal intervals method implemented using the Linear color mapper. As we can see the whole of the Midwest region seems to have the same shade in spite of difference in the number of loans rejected in those states. While California is a dark red because of the highest number of loans rejected to applicants form that state. This discrepancy has been fixed in the blue map on the right-hand side. It uses the Fisher-Jenks method i.e. Natural breaks using the pysal library. This shows a more equitable distribution of the shades which is closer to the reality. Even the states in the Midwest display different shaded based on the bins in which they were place. These maps again show that California is clearly the state that sees more value in borrowing from lending club.

Tab 3:

Using bokeh server for Loan Statistics dataset involved identifying the important variable out of the 140+ variables that were provided to us. We considered 8 main variables for our analysis- and created two visualization based on those 8 parameters, 1 visualization involves a 4 categorical variables that are 'term', 'employee length', 'grade' and 'employee tenure/length'. These parameters play a very important role while accepting any loan application.
In the first bokeh server for Loan statistics data visualization we have shown variation of





all the 4 categorical variables with respect to the average loan amount.

Chart 1: Term versus Average Loan Amount (2011 and 2016)

The above graph shows the variation in the average loan amount from 2007 till 2016. For the first three years only 36 months term was provided but from 2010 onwards Lending club gave an option of provided two different term options "36 months and 60 months". If we see the variation in the average loan amount the different in the average loan amount for 36 months and 60 months is decreased with every year.

Chart 2:Grade versus Average Loan Amount from 2007 till 2016

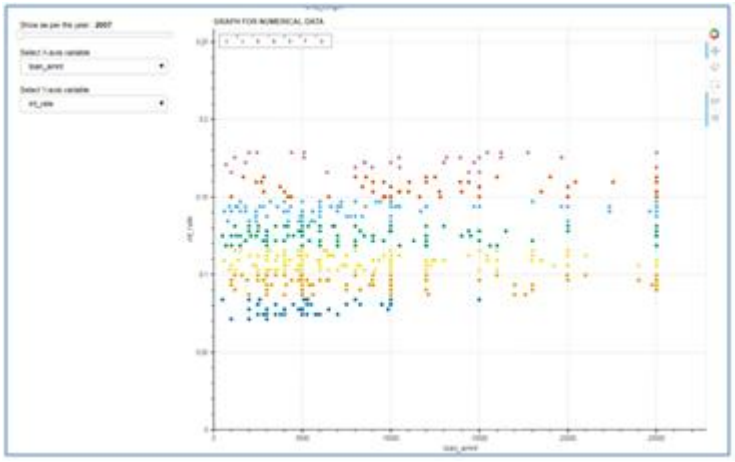
Figure 2 represents the variation in the average loan amount with respect to the grade assigned to the loan. Grade is assigned by Lending club based on the credit risk and other variables based on the application of the borrower. In the graph we can see that from 2007 till 2016 Grade F and G always had the highest average loan amount and the average loan amount increased from $5000 to $14500 for grade A loans. The variation in other grade types can also be seen.

Chart 3:Employee Tenure(Length) versus Average Loan Amount from 2007 till 2016

The above figure shows how the employee length for the applicants vary with time. In 2007 the employee with tenure of 4 years and 8 years applied for the loan with the highest average loan amount. By 2016 the average loan amount requested were highest for 10+years employee tenure.

Numerical Data for Loan Statistics

For our analysis we have worked on 4 numerical variables that are loan amount requested by the borrower, interest rate, annual income of the employee and the dti ratio.
The below chart the loan amount and the interest rates for all the loans from 2007 to 2016.



This chart shows that in 2007 very few borrowers applied for loan by Lending club and the interest rate varied from 5% to 15%.
We can also see that all these grades were shown as per their grade type and grade type shows a direct correlation with the interest rate .

Chart 4 : Loan amount versus Interest rate for year 2007

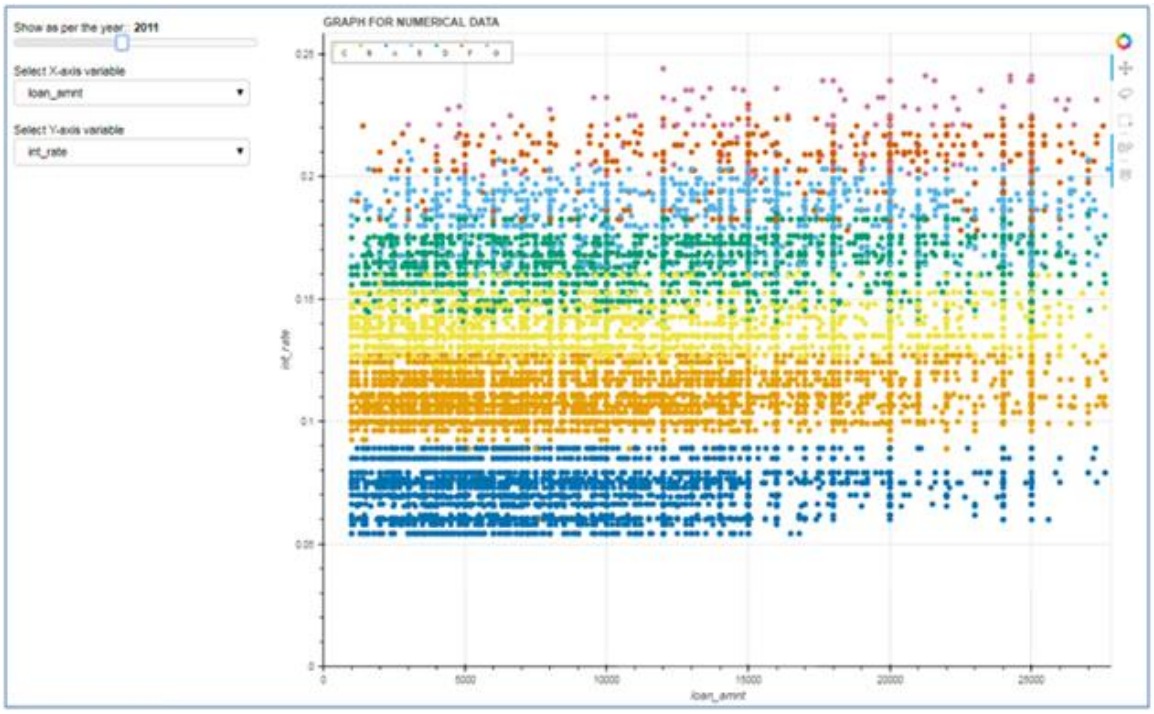


Chart 5 : Loan amount versus Interest rate for year 2011
The above graph shows the scatter plot highlighting the loan amount and interest rates. The requested loan amount has increased till $35000 and even the interest rate ranges from 5% to 25%.
We have plotted the graph till 2016 but due to 500,000 + data records for year 2016 the graph is not clear.
Strength: Provides details for all the records in one graph based on year

Weakness: Due to large number of data records, it takes a lot of time to connect to the server.

Tab 4:

Problem: the first file – RejectStatsA – contains values in 'Loan Title' that are very different than all other files. This file contains data from 2007 to 2012. As a common issue with early data, the contents in the 'Loan Title' column are very unorganized and uncategorized - they are like notes – whereas 'Loan Title' in the other files are categorized into short terminologies. What's worse, many of them are blank. Fortunately, since the year 2010 data, the contents become similar to the data in the later years.

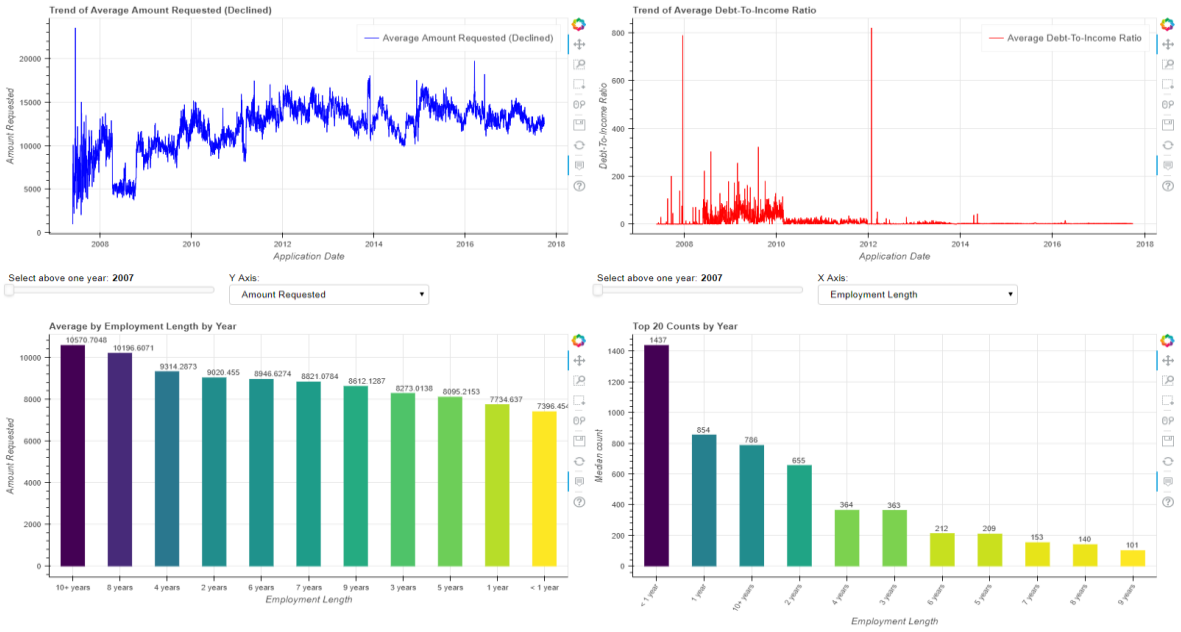
Problem: since the total data volume is huge for personal computers to process, sometimes I have to run it multiple times for the Bokeh server to work.

Problem: I tested putting the two time series charts into one (multi plot in one chart), but the chart looks very ugly as the y axis are very further apart: around 15000 vs from 1.5 – 200. So I have to use two charts.

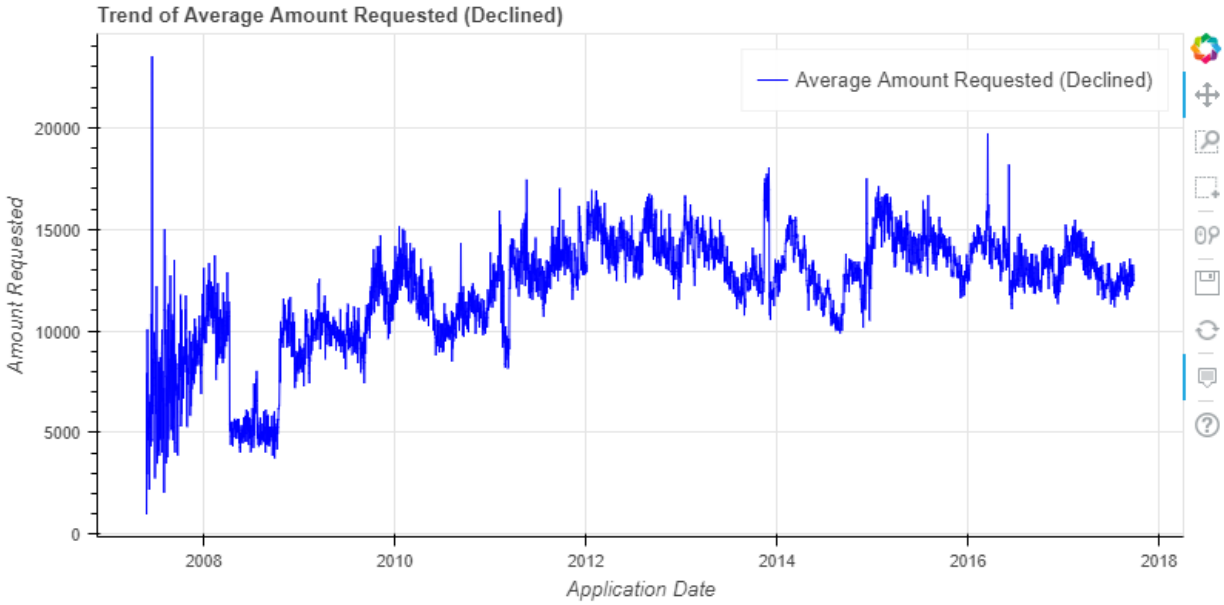
Problem: The color scheme for the bottom left chart on Debt-to-Income Ratio doesn't vary, as compared to Amount Requested. This is because of linear data mapper and the different range of data: thousands to tens of thousands for Amount Requested but a hundred and below for Debt-to-Income Ratio. But this is still a suitable color chose and doesn't affect readability.

Analysis:

This is a screenshot of my visualization for the data from 2007 to 2017 Q3 after my code is written:



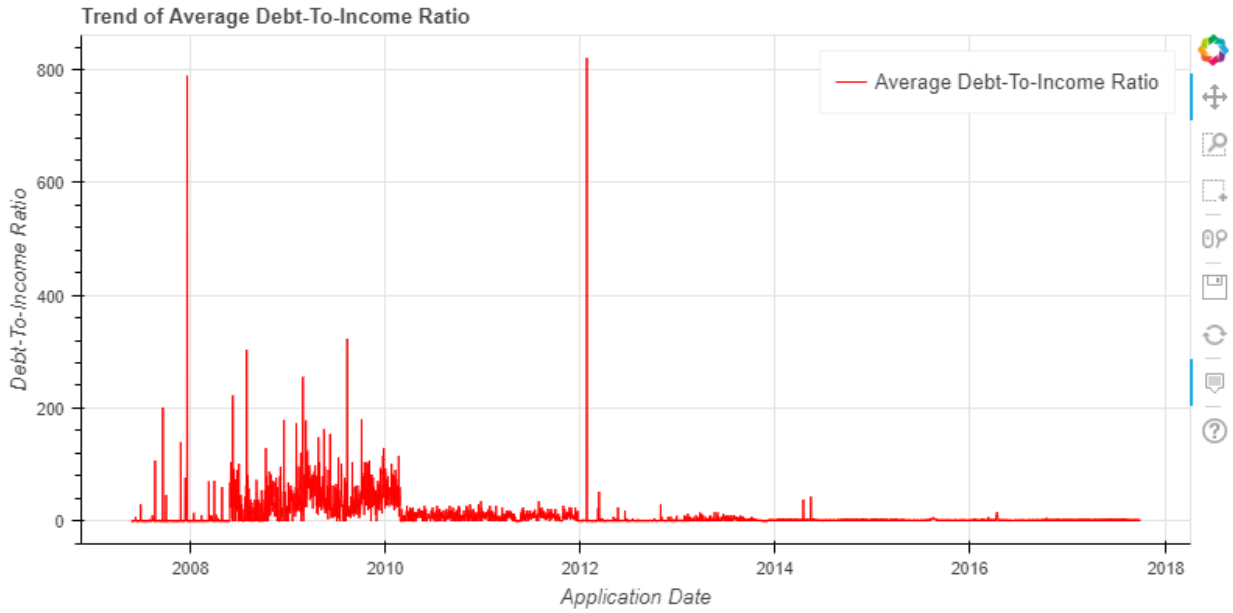
The first (top left) chart:



There are some spikes around the middle of 2007, how coincidental it is with the Financial crisis of 2007–2008? The crash of economy made people went in debt quickly, people got panic and suddenly requested a large amount of loan. Moreover, as people become deeper in debt, the average loan amount requested increases – it shows a trend on the graph. And the trend didn't stop until the middle of 2008, when there's a sudden drop. My theses are as the amount of loan declined increased, people gave up on getting a loan or they

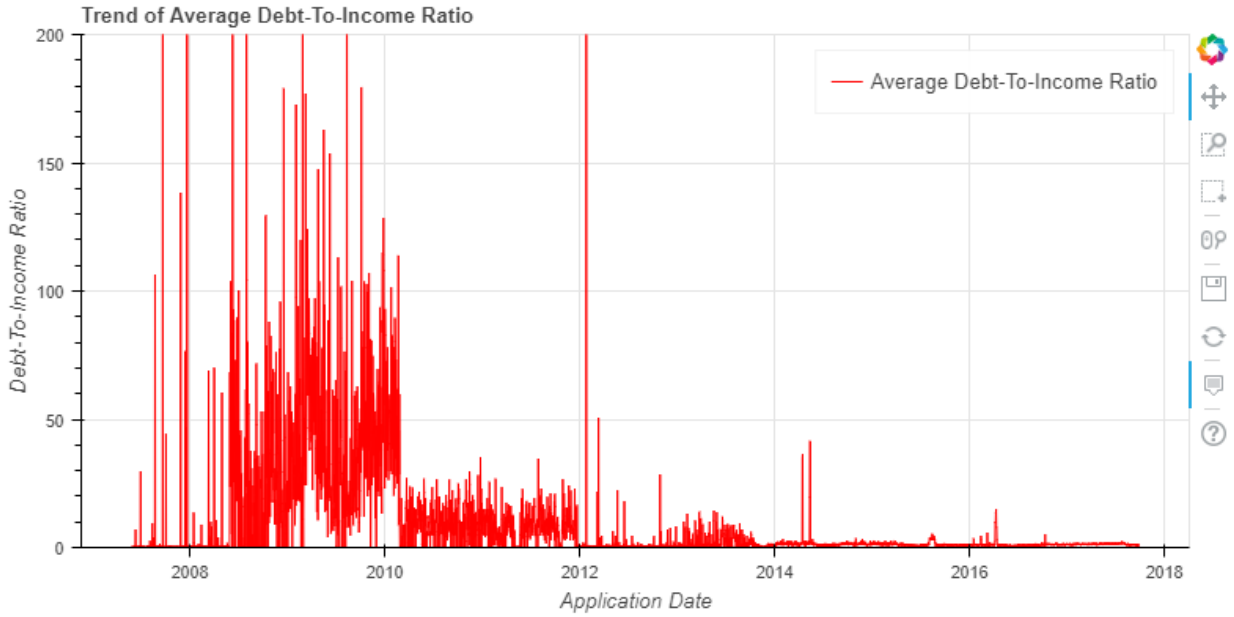
couldn't pay back the loan. And after a short while, the trend continued as the declined loan amount continued to increase until it topped at the middle of 2012. Again, how coincidental it is that with the start of our economy recovery? As our economy slowly recovers, the average declined loan amounted stabilizes.

The second (top right) chart:



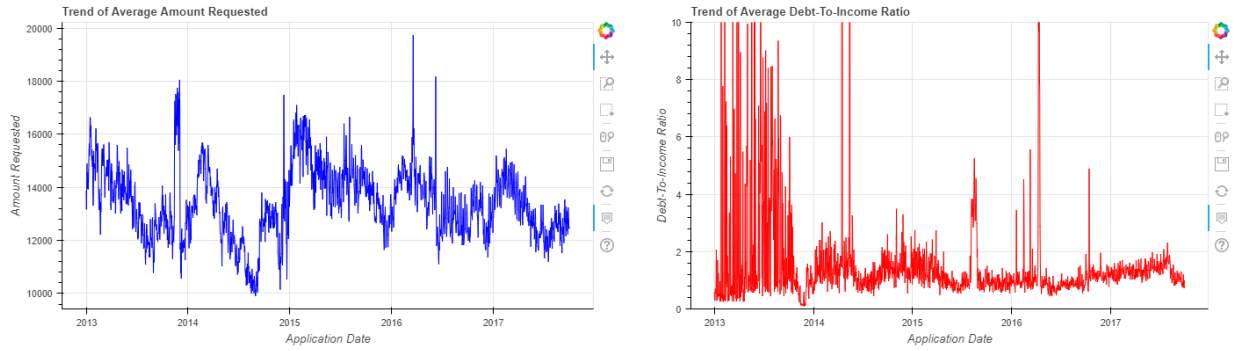
As we can see, there are quite some noise in the chart 'Trend of Average Debt-To-Income Ratio'. It can be an occasion thing, but also likely some typo. I need to set some filter to exclude the noise for a better visualization to analyze.

One simple solution is the add y_range=(0,200) to the figure, and below is the new graph:



As we can see, consistent with the first chart, the average debt-to-income ratio gets very high during the Financial crisis of 2007–2008 – meaning people are more in debt. And after 2010, as the crisis ends, the ratio gets lower. This graph is a better representation of US economy than the first one, as a ratio tells a better story than just an amount. As the recovery of US economy, the average Debt-To-Income Ratio gets lower. In 2012, in general the debt-to-income ratio gets below 10.

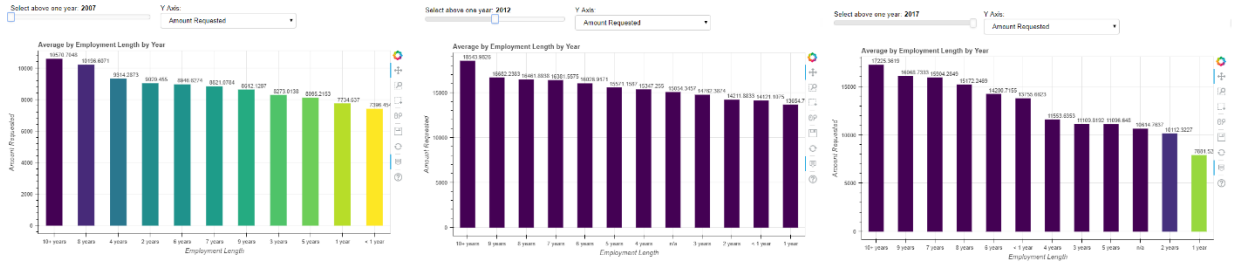
For further analysis, I include only the data from 2013 to 2017 Q3 and setting y_range=(0,10) for a better visualization:



As we can see, as our economy further stabilizes, the declined Average Amount Requested stabilizes, and the Average Debt-To-Income Ratio remains mostly between 1.5 to 2.

The third (bottom left) chart:

First, for Average Amount Requested (Declined) by Employment Length by year. For space consideration, below is three screenshots for year 2007, 2012 and 2017, respectively.



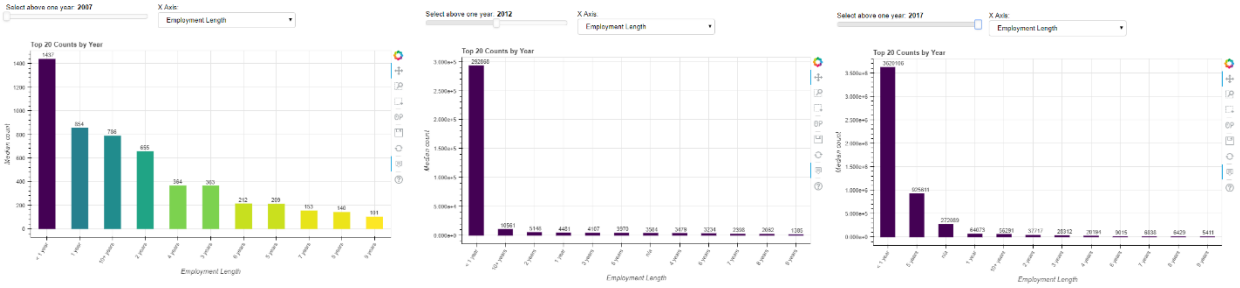
As we can see, the trend of the amount requested is consistent with what we interpret from the first chart. And it further shows that, the longer the employment, the larger the amount requested and declined. But since we don't find the rate of declined amount vs. total amount, or income vs. employment length in this data, the only analysis we can make is that people with longer employment length generally require more loan. Isn't it weird? Shouldn't the longer employed people have more savings and require less loan? Two possible theses: 1, since America is a credit enabled country, people with longer employment are granted bigger loan amount and generally use more; 2, people with longer employment have more responsibilities to spend money on, such as mortgage/children's school tuition, that they need to get more loan.

Now let's that a look at the Average Debt-to-Income ratio by Employment Length by year.

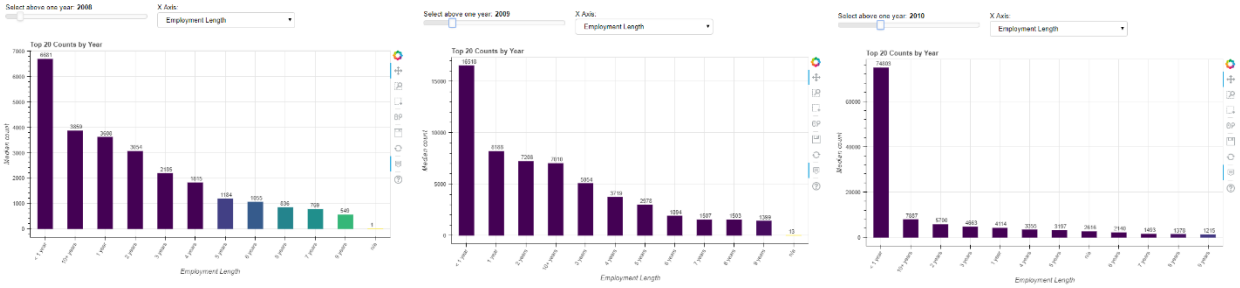
As we can see, the years from 2007 to 2010 varies some, as we were in the time of Financial crisis, but generally people who are employed one year or less, and who are longer employed have higher ratios, that is, are more in debt. For people with longer employment, this is consistent with my analysis on the amount requested by employment length chart. For people who are employed one year or less, this makes sense, as many of them have a large amount of remaining school loan/car loan and earn the fewest income.

The fourth (bottom right) chart:

First, for Top 20 counts by Employment Length by year. Below is three screenshots for year 2007, 2012 and 2017, respectively.

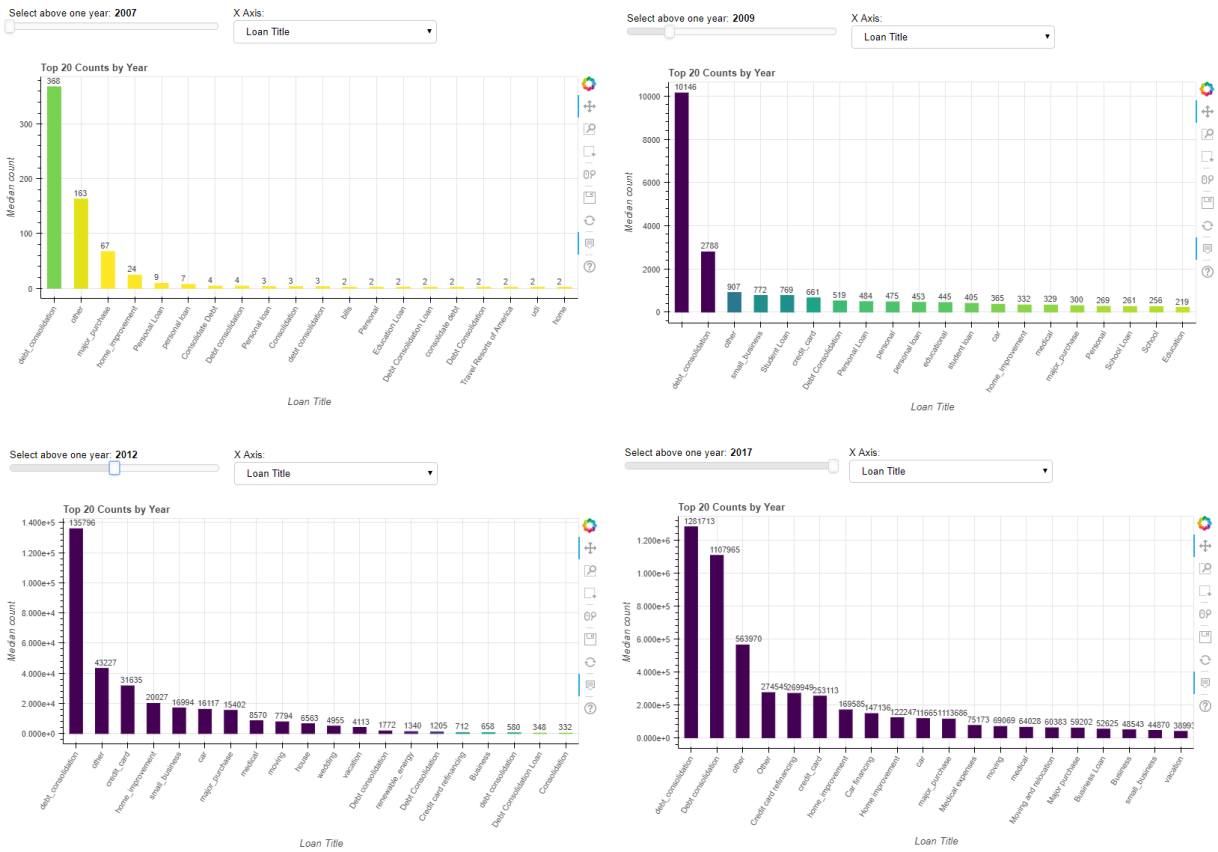


As we can see, people employed 1 year or less are the dominant participants. This is consistent with my thesis, as many of them have a large amount of school loan/car loan remaining and earn the fewest income.



We can take a further look and look at the 2008, 2009, 2010 charts. One interesting finding is that, the year 2008 and 2009 charts are more similar to year 2007 chart, whereas the year 2010 chart is more like the later charts (as the year 2012 and 2017 charts shown). In year 2007, 2008, 2009 charts, people employed longer than 10 years also count a large amount, as many middle-aged/in middle-management people got laid off during Financial crisis, and as a result, more of them applied loan (but declined).

Now let's look at Top 20 counts of Loan Title by year.



As I stated in the beginning, the values in 'Loan Title' column in the first data file from 2007 to 2011 are very unorganized, thus the such smaller count of loan titles in the first few years. This shouldn't be considered as a valid part of analyzed data.

Since year 2010 the data is structured better, and we can clearly see that the dominant reason that people apply loan is to have their debt consolidated, followed by home improvement, and credit card/credit card financing, and car/car financing. This makes sense, as home, credit card, and car are the major biggest purchases in one's life.

Conclusion

Our conclusion from the charts are:

- In financial crisis, the average debt-to-income ratio gets higher; as the economy stables, it becomes lower. And people with longer employment, that is, are middle-aged/in middle-management, appear to require loan more frequently, as they get laid off.

- People with longer employment generally require more amount of loan as their credit as well as spending increase.
- People with one year or less employment generally have the highest debt-to-income ratio as they have a large amount of remaining school loan/car loan and earn the fewest income.
- The dominant reason for people to apply loan is debt consolidation. The next major reasons are home improvement, credit card/credit card financing, and car/car financing, as those are mainly biggest purchases in one's life.

We can call the above findings "common sense", but "common sense" cannot be taken as facts without valid supports. Since our data amount is very big enough and also cross a solid span of time, thanks to my charts, the findings are now facts.

Further work

In terms of building on what we have we plan to conduct further analysis including variables that we could not include in this phase of our EDA. Other ameliorations include creating a robust model which would allow us as small investors to invest wisely on individuals that fit the criteria of a creditworthy individual. This could generate a sizeable income which is always welcome. Eventually this could also be marketed for the layman investors who are in need of a tool to explore the Lending club data on their own.

References

1. www.lendingclub.com
2. http://kldavenport.com/lending-club-data-analysis-revisted-with-python/
3. http://www.dealingdata.net/2016/08/03/PoGo-Series-Making-a-Choropleth-Map/#bokeh