# Analysis of Lending Club Records

Paymahn Moghadasian and Lauren Slusky

COMP 4710: Data Mining

Wednesday, December 3, 2014

## Table of Contents

## INTRODUCTION

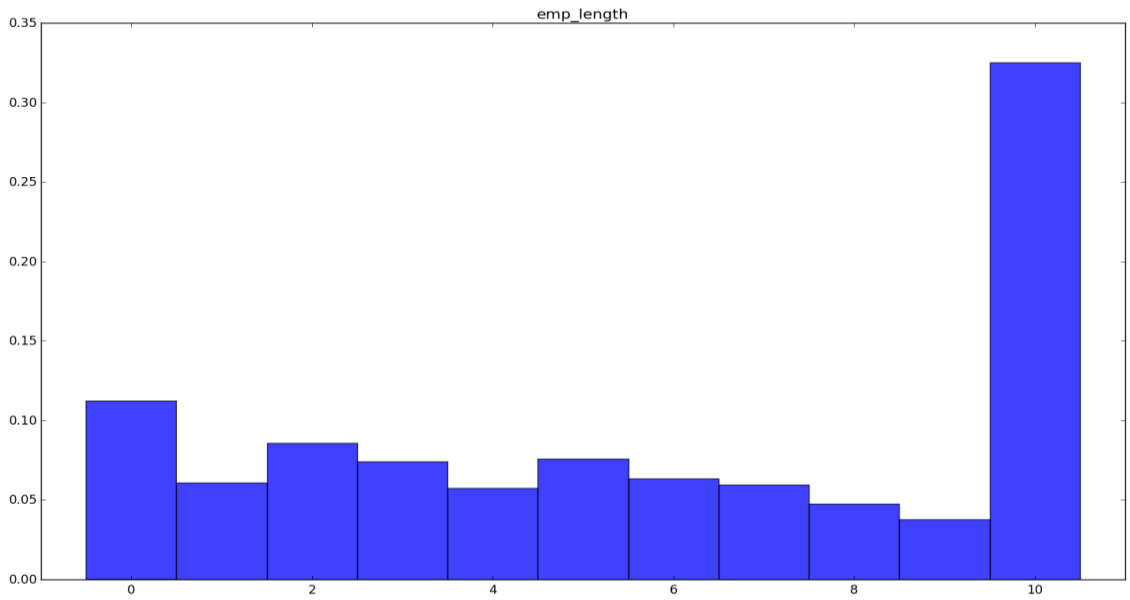
Following what we have been learning in class, we were interested in finding our own data and applying a few of the techniques we've discussed. After some looking, we found this 'Lending Club' data. The data consists of a mass amount of information. The tables of data that we included in our repository are a list of records of accepted requests for loans, a list of rejected loan requests, and counts of where the loans were requested, by state. Because of having access to so many dimensions and so many transactions we were able to run a variety of tests and answer some very interesting questions.

The structure of this report is as follows. First, we will give you an overview of the data and interesting trends we found. Next, we analyzed different aspects of the data, including predicting Grades, Loan Status and Risk. The last analysis we ran on the data was some association mining using the Apriori algorithm. Each of these analysis sections will consist of an overview of our analysis technique and our results. Next is a discussion regarding future work we would be interested in doing. Lastly, we will conclude with some final remarks.

## INTERESTING GRAPHS

### EFFECTS OF EMPLOYMENT LENGTH

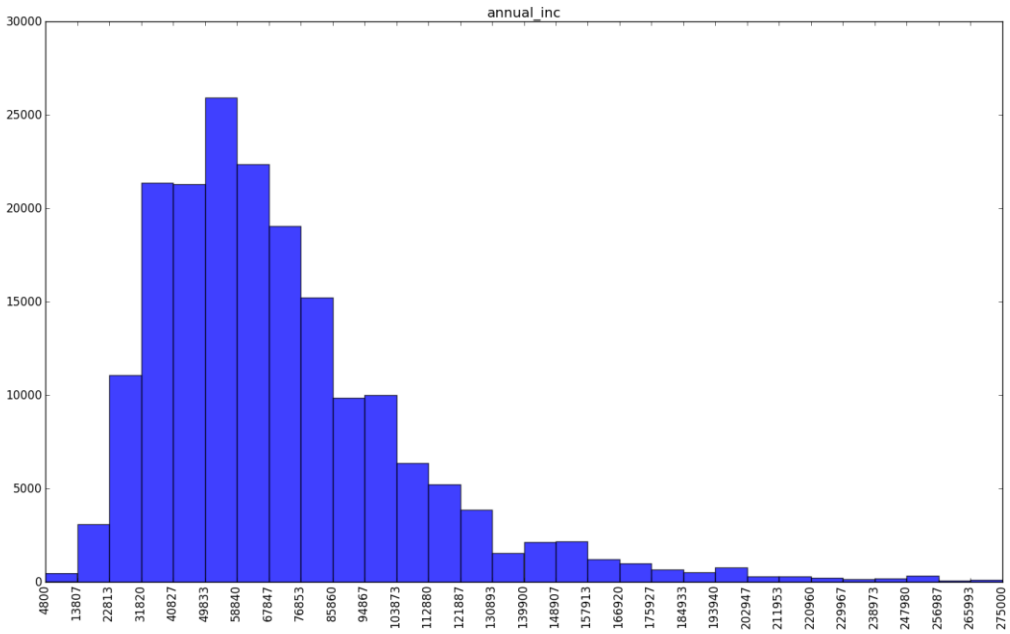
We were initially interested in whether loan applicants would have a normally distributed value for employment length. Turns out they don't. The following are graphs of employment length shown as percentages (as opposed to absolute counts) for accepted and rejected loans respectively.

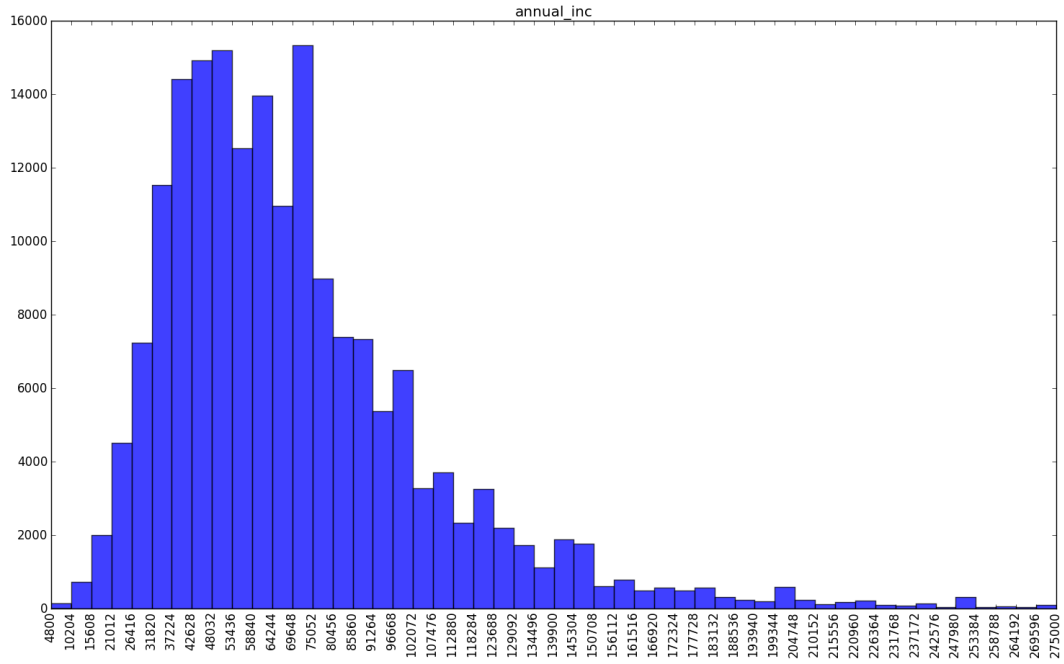



It's important to note that the data provided on employment length was categorized as, less than 1 year, 1 year, 2 years… 9 years, 10+ years. It seems to be pretty apparent that there is some correlation between employment length and acceptance/rejection rate. The proportion of applicants with less than 1 year of employment is much higher for rejected loans. Similarly accepted loans have a much higher proportion of applicants with a long employment history.

**INCOME DISTRIBUTION**

We decided to create a histogram of incomes just for fun. The generated histogram was not that great originally but we looked into the stats a bit more and found that some outliers were mucking things up. The highest reported income was approximately 7 million. These are the graphs generated with the top 1000 incomes removed:



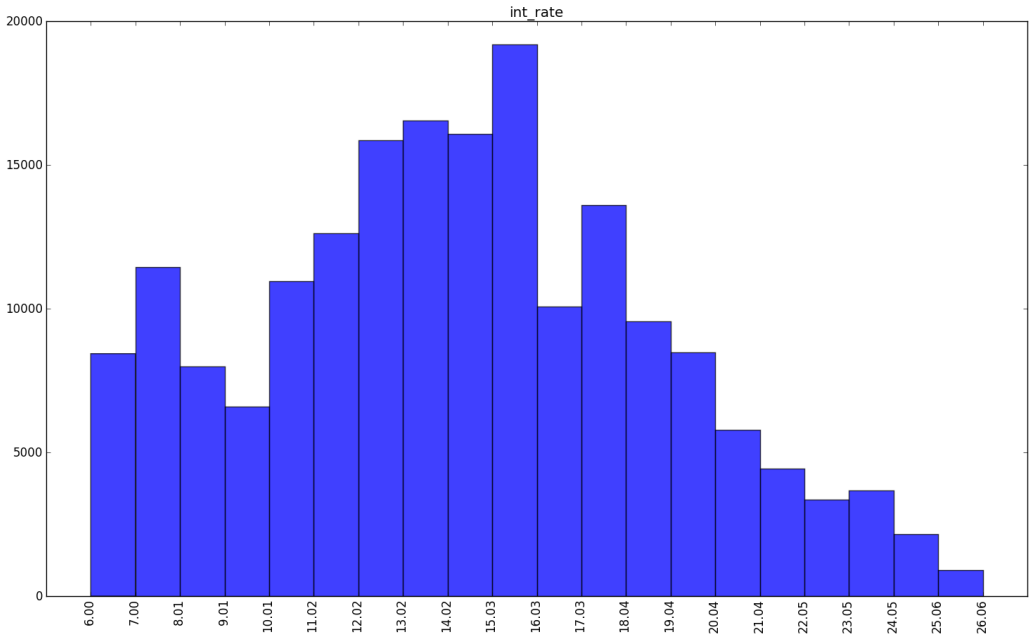




This returned both interesting and expected results. Something we found interesting and unexpected was that wealthy people also apply for loans from the lending club. However, on the flip side, most of the people who were applying for loans were more middle and/or lower class (so we suspect based on income). The distribution is a skewed normal distribution with a long tail for higher incomes.

**INTREST RATE DISTRIBUTION**

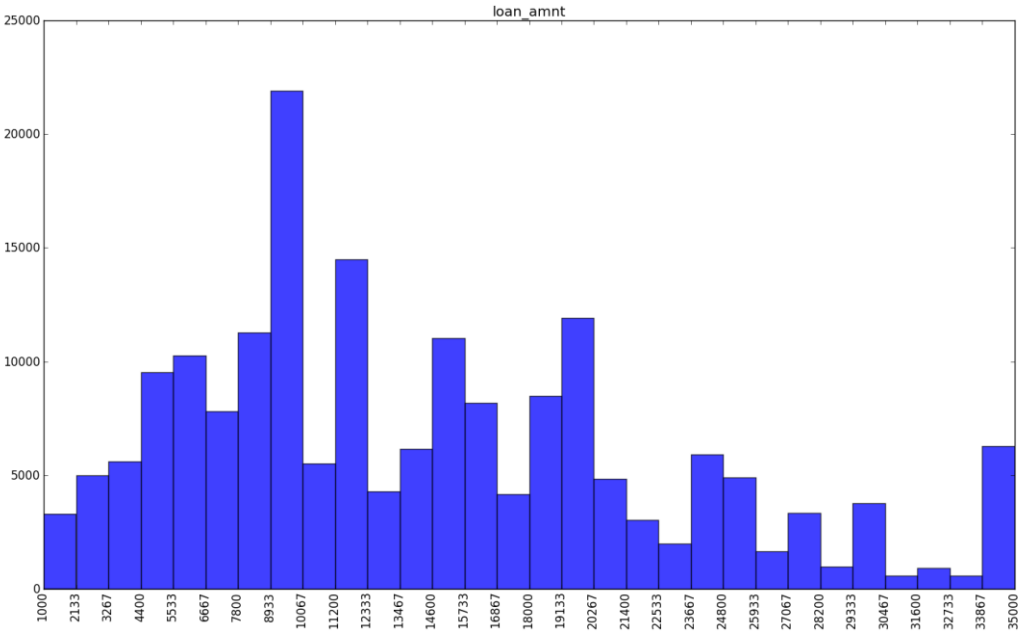
Here is the graph we generated for the different interest rates. What we can see is it is not normally distributed but rather it looks more like a Poisson distribution. We can explain this, potentially, by looking at the numbers we see of the interest rates. Since the average

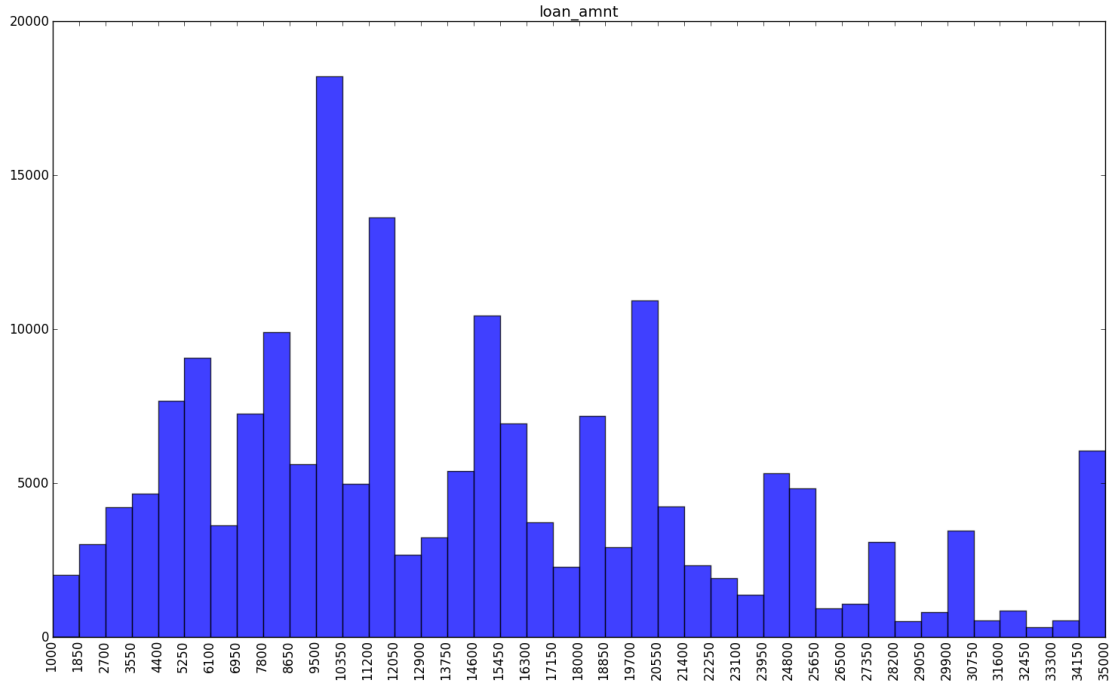
interest rate is so close to the minimum of the domain we see a spike at that end of the graph but a tail as we go toward higher values because there is still a possibility of higher interest rates.

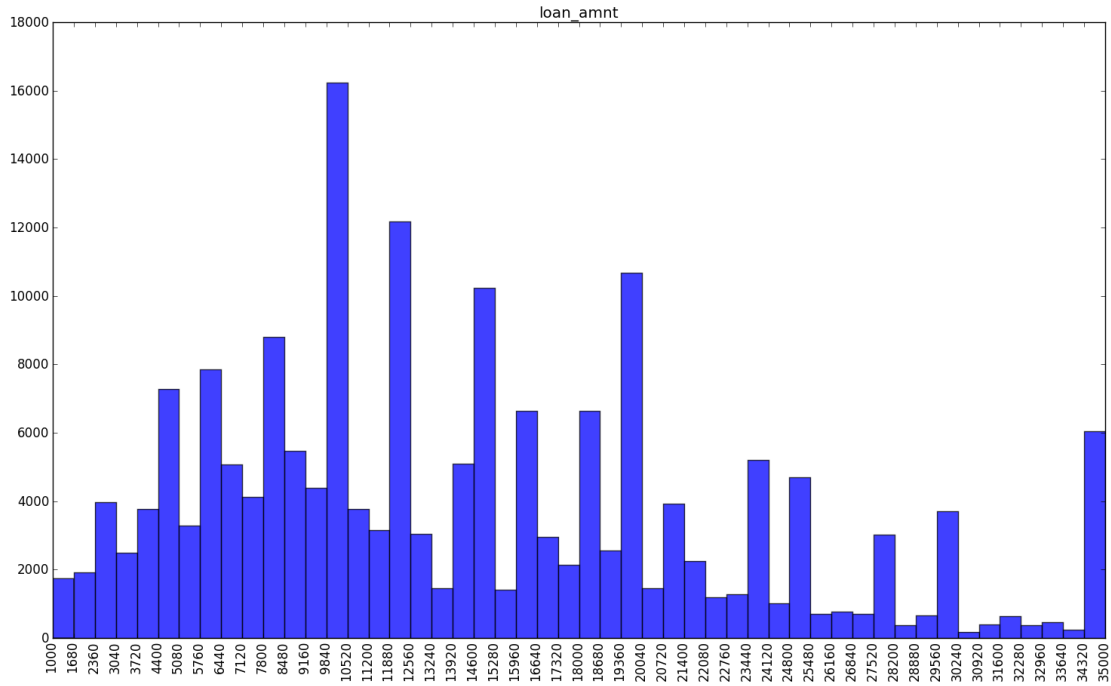


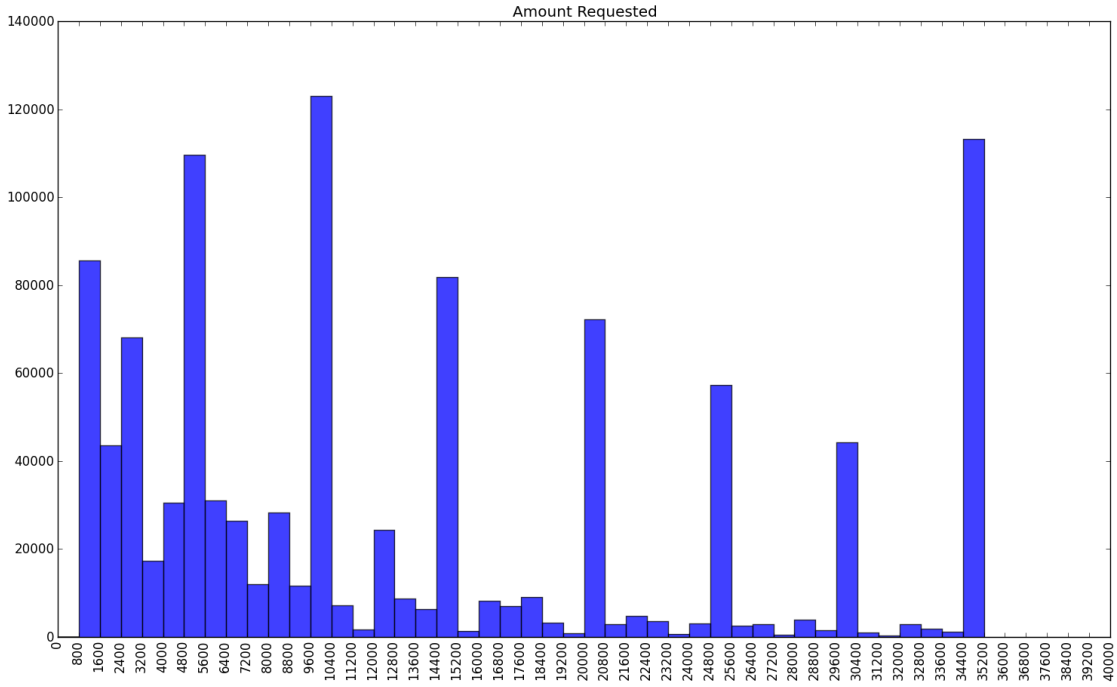

**REQUESTED LOAN AMOUNTS**

The next four graphs speak to the different counts of requested loan amounts that we generated for the accepted requests with 30, 40 and 50 bars respectively. The final graph of this section is for the rejected loan amounts. What we found to be interesting was that there are some spikes of requests for multiples of $5000. This could speak to the fact that people are most comfortable requesting nice, 'even' amounts. After discussing this, we agreed that we would be more likely to loan money to someone asking for $5000 instead of $4634.87. We see this same trend in the rejected loans.



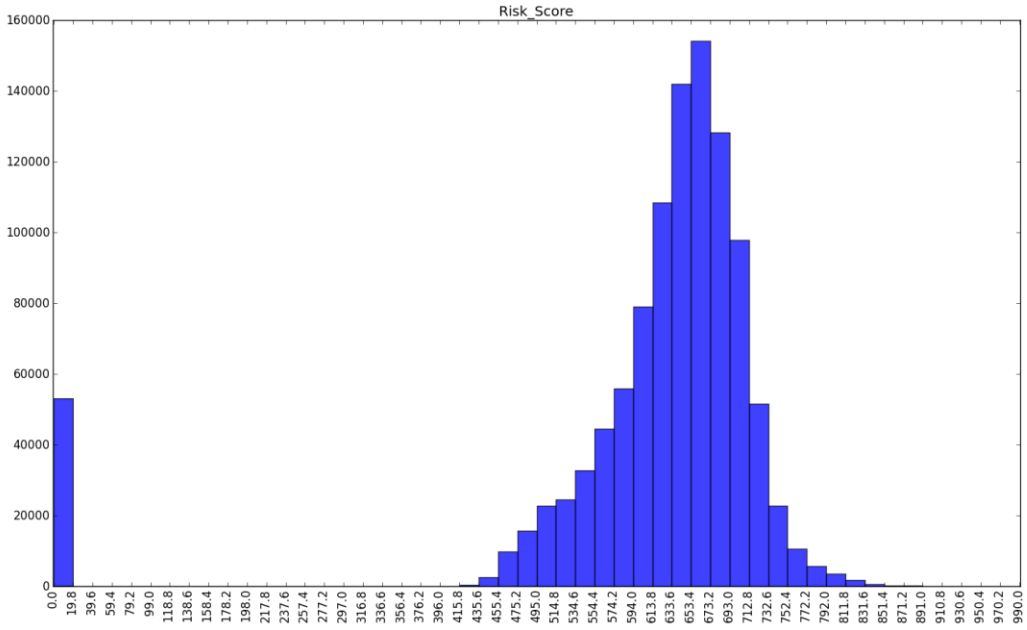

loan_amnt
20000
15000
10000
5000
0
1000
1850
2700
3550
4400
5250
6100
6950
7800
8650
9500
10350
11200
12050
12900
13750
14600
15450
16300
17150
18000
18850
19700
20550
21400
22250
23100
23950
24800
25650
26500
27350
28200
29050
29900
30750
31600
32450
33300
34150
35000


loan_amnt
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
1000
1680
2360
3040
3720
4400
5080
5760
6440
7120
7800
8480
9160
9840
10520
11200
11880
12560
13240
13920
14600
15280
15960
16640
17320
18000
18680
19360
20040
20720
21400
22080
22760
23440
24120
24800
25480
26160
26840
27520
28200
28880
29560
30240
30920
31600
32280
32960
33640
34320
35000



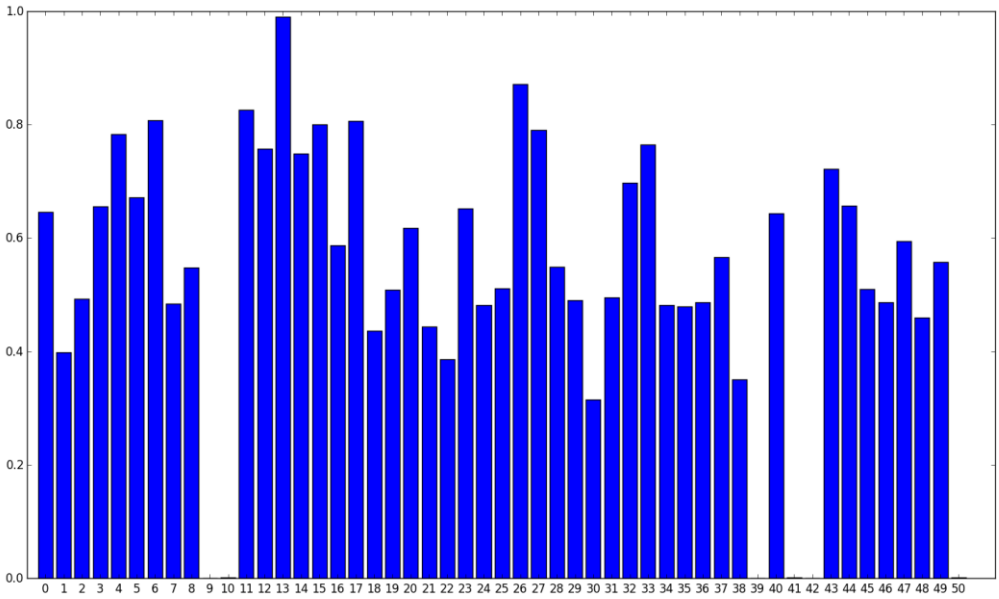

**RISK SCORE DISTRIBUTION**

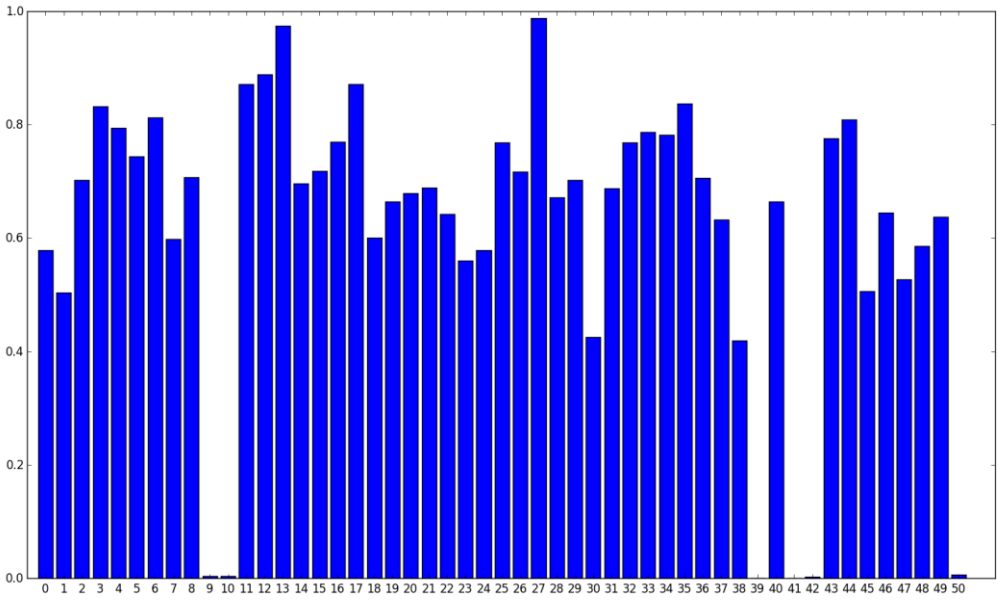
We didn't find that much to be interesting when analyzing the distribution of risk scores. We saw a relatively normal distribution which was to be expected. Most people fall in the same range of risk scores and then there are the few that trail off at either end. As you can see by the graph, there were a few people who were almost outliers because they had little to no risk score but because of the number of them, it isn't really an outlier.



**STATEWIDE APPLICATIONS**

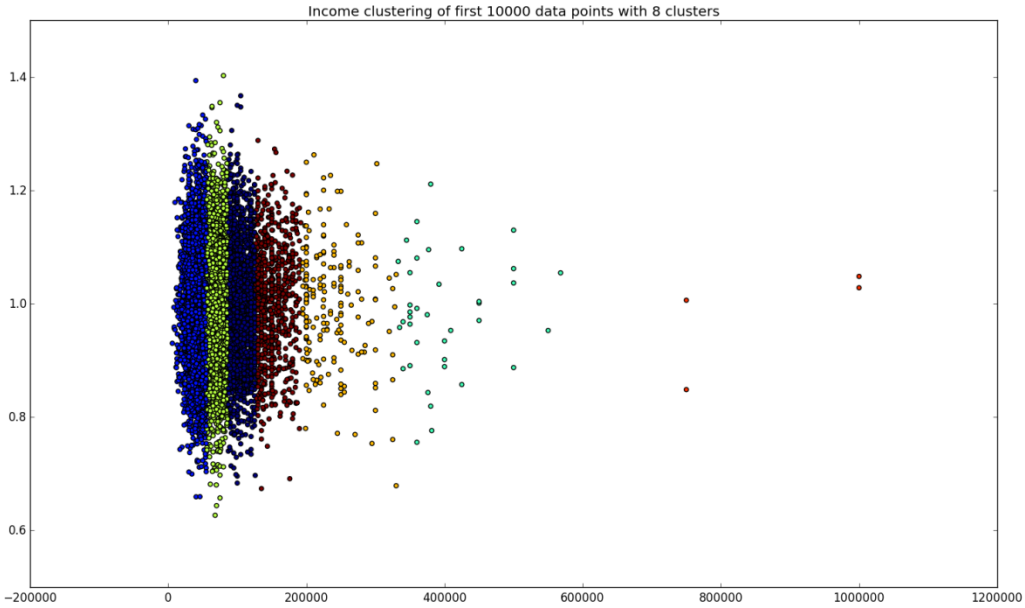
Here are the numbers of applications per state normalized for population for accepted and rejected loans respectively. The legend of state name to state number has been included as **Appendix i. T**he following states have significantly fewer applications per person compared to other states: Florida, Georgia, South Dakota, Texas, Utah, and North Dakota. Illinois has significantly more applications per person compared to other states. We couldn't really say why these trends appeared. Utah, Georgia and Texas aren't considered to be particularity progressive states and that may explain why they have fewer applications per person. It might be that the oil boom in the Dakotas explain why they don't have many applications. We couldn't explain this trend for Florida or Illinois.



## CLUSTERING THE LOAN AMOUNTS

For use in a few different analyses, we clustered the different loan amounts. We did this using kmeans. What we found is that using 10 clusters was a reasonably good number. Increasing the number of loan amount clusters actually made the accuracy much worse in some of the analyses. While having more specific and detailed clusters could be a good idea and help with analysis, it can also lead to, possibly, insufficient data in each cluster.




## PREDICTING LOAN STATUS

We tried to predict the loan status of an accepted applicant using the following information: clustered loan amount, term length, interest rate, grade, employment length, home ownership status, annual income, and delinquencies in the last 2 years, and number of open accounts. We were able to achieve 70% accuracy with a depth of 12. Even though we achieved this level of accuracy, the classifier wasn't great because most applications had a status of "Current" which was correctly classified ~63000/65000 times. Most incorrect classifications also happened to be for the label "Current". This classifier did not work out as well as we had hoped. Here is a table diagraming our results:

| True Labels | Charged Off | Default | Issued | Fully Paid | Current | Late (16-30 days) | Late (31-120 days) | In Grace Period |
|---|---|---|---|---|---|---|---|---|
| Charged Off | 53 | 0 | 0 | 191 | 5182 | 7 | 7 | 3 |
| Default | 2 | 0 | 0 | 1 | 137 | 0 | 0 | 0 |
| Issued | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Fully Paid | 118 | 0 | 0 | 452 | 19520 | 5 | 21 | 14 |
| Current | 269 | 1 | 0 | 1136 | 63896 | 17 | 72 | 27 |
| Late (16-30 days) | 1 | 0 | 0 | 8 | 366 | 0 | 2 | 0 |
| Late (31-120 days) | 19 | 0 | 0 | 48 | 1655 | 2 | 3 | 2 |
| In Grace Period | 5 | 0 | 0 | 14 | 609 | 0 | 1 | 1 |

**PREDICTING GRADES**

The first attempt to predict grade involved using the following predictors: clustered loan amounts, term length, employment length, home ownership status, annual income, income verification status, DTI, delinquencies in the last 2 years, and number of open accounts. We only achieved 40% accuracy with a depth of 8. These results were not great. The results of this are in the following table:

| True Labels | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 3588 | 9471 | 970 | 54 | 2 | 0 | 0 |
| B | 2936 | 22819 | 5124 | 231 | 133 | 18 | 0 |
| C | 1035 | 12703 | 10053 | 439 | 616 | 102 | 0 |
| D | 474 | 6666 | 5747 | 428 | 601 | 103 | 0 |
| E | 84 | 1262 | 3440 | 304 | 776 | 179 | 0 |
| F | 14 | 246 | 1756 | 186 | 554 | 174 | 0 |
| G | 2 | 16 | 318 | 55 | 124 | 65 | 0 |

When we included interest rate as a predictor our results were much better. We had 99% accuracy and this did a very good job on all the grades. This lead us to thinking of there was some kind of correlation between grade and interest rate for a lending club. A quick Google search shows us that there is indeed a link between these two attributes. (https://www.lendingclub.com/public/how-we-set-interest-rates.action)
Here is this table:

| True Labels | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 14084 | 1 | 0 | 0 | 0 | 0 | 0 |
| B | 0 | 31258 | 1 | 0 | 0 | 2 | 0 |
| C | 0 | 2 | 24940 | 6 | 0 | 0 | 0 |
| D | 0 | 1 | 312 | 13705 | 0 | 1 | 0 |
| E | 0 | 2 | 0 | 48 | 5744 | 251 | 0 |
| F | 0 | 0 | 0 | 0 | 55 | 2875 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 115 | 465 |

**PREDICTING RISK**

Next, we tried to predict risk scores using the following predictors: clustered loan amounts, debt to income ratio, and employment length. We achieved an accuracy of 28% with a depth of 8. The risk scores were put into 10 clusters. This seemed to infer that there just wasn't enough information to predict the risk score, which makes sense. A Lending club wouldn't want to give up whatever proprietary algorithm they have for this. Here is a table outlining our results from this test:

| True Labels | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 9732 | 2120 | 7354 | 21 | 18431 | 644 | 160 | 4212 | 144 | 20609 |
| 1 | 1997 | 21324 | 156 | 25 | 392 | 508 | 31 | 782 | 63 | 1253 |
| 2 | 3835 | 1001 | 49024 | 5 | 25568 | 156 | 201 | 1255 | 874 | 10394 |
| 3 | 5533 | 2570 | 2317 | 25 | 5205 | 797 | 64 | 3453 | 46 | 8025 |
| 4 | 6187 | 1296 | 33453 | 12 | 39674 | 274 | 172 | 1988 | 515 | 21388 |
| 5 | 4435 | 2703 | 1057 | 9 | 2707 | 891 | 53 | 2938 | 46 | 5059 |
| 6 | 989 | 565 | 6234 | 7 | 3267 | 115 | 383 | 355 | 279 | 1850 |
| 7 | 7842 | 2519 | 3872 | 36 | 9511 | 775 | 98 | 4219 | 86 | 13211 |
| 8 | 2550 | 1078 | 32804 | 6 | 14341 | 136 | 228 | 737 | 958 | 5986 |
| 9 | 8495 | 1183 | 13720 | 9 | 31295 | 402 | 148 | 3103 | 249 | 25444 |

**ASSOCIATION MINING**

One of the algorithms that we chose to run our data through was the Apriori algorithm that was discussed in class. We generated frequent item sets for the following four columns: amount of funds requested, employee title, annual income and reason for requesting a loan. The reason we chose to look at these four columns was to answer some specific queries. We were interested in seeing possible correlations between the funds requested versus the annual income. The hope was to identity some kind of connection between how much money one made and if there was some kind of trend in how much money they may need. We were also hoping to see a connection between the amount of funds requested and the reason for requesting a loan. Some of the reasons for requesting a loan included buying a car, debt consolidation and weddings. It would have been very interesting to see if there was a consistent range for people requesting a certain amount and what they were putting the money toward.

Since there were a lot of different values for loans and annual income, we put these values into ‘buckets’ to better cluster our data. This was done differently than the loan clustering done for the above analyses that used kmeans. This clustering was just a simple rounding down to the closest $1000. These were the rules produces for a support threshold of 0.007 and confidence threshold of 0.50 (there is one other trial included with a higher support threshold of 0.02 in the results folder of our repository):

frozenset(['loan_amnt: 8000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.565885
frozenset(['loan_amnt: 22000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.691867
frozenset(['annual_inc: 65000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.598986
frozenset(['annual_inc: 80000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.592853
frozenset(['annual_inc: 52000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.614810
frozenset(['loan_amnt: 24000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.659382
frozenset(['annual_inc: 85000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.612210
frozenset(['annual_inc: 45000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.601355
frozenset(['annual_inc: 42000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.620170
frozenset(['loan_amnt: 10000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.572176
frozenset(['loan_amnt: 35000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.639993
frozenset(['annual_inc: 50000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.598390
frozenset(['loan_amnt: 20000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.657361
frozenset(['annual_inc: 120000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.543070
frozenset(['loan_amnt: 21000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.647215
frozenset(['loan_amnt: 6000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.520426
frozenset(['annual_inc: 35000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.568299
frozenset(['annual_inc: 30000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.555216
frozenset(['loan_amnt: 15000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.627975
frozenset(['loan_amnt: 18000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.658306
frozenset(['loan_amnt: 19000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.649942
frozenset(['annual_inc: 55000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.614524
frozenset(['annual_inc: 62000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.612568
frozenset(['annual_inc: 70000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.599592
frozenset(['annual_inc: 38000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.588340
frozenset(['loan_amnt: 12000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.617928
frozenset(['loan_amnt: 14000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.616846
frozenset(['annual_inc: 48000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.591201
frozenset(['loan_amnt: 25000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.679849
frozenset(['annual_inc: 40000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.592650
frozenset(['loan_amnt: 9000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.596503
frozenset(['annual_inc: 100000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.581114
frozenset(['annual_inc: 72000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.601660
frozenset(['loan_amnt: 16000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.641770
frozenset(['loan_amnt: 7000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.544356
frozenset(['loan_amnt: 30000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.670480
frozenset(['loan_amnt: 11000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.626680
frozenset(['annual_inc: 90000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.599677
frozenset(['loan_amnt: 13000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.618872
frozenset(['loan_amnt: 28000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.646322
frozenset(['annual_inc: 75000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.594414
frozenset(['annual_inc: 110000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.593426
frozenset(['loan_amnt: 17000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.660852
frozenset(['annual_inc: 60000']) ---> frozenset(['purpose: debt_consolidation']) conf: 0.613691

As we can see, the most frequent item sets were always related to debt consolidation. Even though there were many other reasons for requesting a loan it seems as if across the board, it was most frequent. These were not the results we were hoping for. There was a lot of interest in possibly connecting different amounts to more frequent reasons but debt consolidation just took over any chance of finding those results.

## FUTURE WORK

Having so much data leaves room for a lot of further work. There are a lot of interesting questions we could answer using all of the data we have. One step would be to try and find a better classifier for predicting loan status. It would be very useful to know, given my own information, if there was a high or low chance I would be accepted for a loan, given this data.

Another area that would be interesting to explore would be a state-by-state breakdown of loan applications. It's very interesting that some states have so few applicants for the lending club while others were booming. Further clustering and classifying of the data would allow us to break down this information and see if they have a lot of lavish over-priced weddings in Illinois or if just everyone is in debt.

It would also be great to have a chance to generate some more interesting rules using some of the algorithms we discussed in class. We were hoping to see a lot more information from the associated mining than we got, and the next step would be do go through the frequent item sets with some of the enhancing techniques for interesting rules and see what's generated.

## CONCLUSION

In this paper we have discussed different types of analyses that we compiled using our lending data. We explored three main techniques of analysis, clustering, classification and association mining. All three techniques have their merits and are used for very different types of analyses. Like any large data set, we had the opportunity to answer ask and answer many different questions like, most frequent loan amounts, what are loans most often used for, and what state do most loans come out of. We think we gained a lot of knowledge from our interesting trend histograms. They broke down the data in various ways and allowed us to observe the information in many different perspectives. We found very interesting results when doing our prediction of grades. Since we don't know much about lending clubs it was very informative to learn about the connection between grades and interest rates. In the future, it would be very interesting to have a chance to run more analyses on frequent patterns in requesting loans.

**APPENDICES**

**Appendix i.**

| State Name | # on Graph | State Name | # on Graph | State Name | # on Graph | State Name | # on Graph | State Name | # on Graph |
|---|---|---|---|---|---|---|---|---|---|
| Alabama | 00 | Georgia | 10 | Massachusetts | 20 | New Mexico | 30 | Tennessee | 40 |
| Alaska | 01 | Hawaii | 11 | Michigan | 21 | New York | 31 | Texas | 41 |
| Arizona | 02 | Idaho | 12 | Minnesota | 22 | North Carolina | 32 | Utah | 42 |
| Arkansas | 03 | Illinois | 13 | Mississippi | 23 | Ohio | 33 | Vermont | 43 |
| California | 04 | Indiana | 14 | Missouri | 24 | Oklahoma | 34 | Virginia | 44 |
| Colorado | 05 | Iowa | 15 | Montana | 25 | Oregon | 35 | Washington | 45 |
| Connecticut | 06 | Kansas | 16 | Nebraska | 26 | Pennsylvania | 36 | West Virginia | 46 |
| Delaware | 07 | Kentucky | 17 | Nevada | 27 | Rhode Island | 37 | Wisconsin | 47 |
| District of Columbia | 08 | Louisiana | 18 | New Hampshire | 28 | South Carolina | 38 | Wyoming | 48 |
| Florida | 09 | Maryland | 19 | New Jersey | 29 | South Dakota | 39 | Maine | 49 |
| | | | | | | | | North Dakota | 50 |