Lending Club - Data Exploration

Atharva Tere, Nilima Sahoo, Piyush Shinde, Saheli Saha[1]

Abstract
Peer-to-peer lending, also abbreviated as P2P lending, is the practice of lending money to individuals or businesses through online services that match lenders with borrowers. P2P Lending began a decade ago and it is growing exponentially after it[6]. By cutting banks out of the process, borrowers typically got a lower interest rate than they would have paid on a credit card or a loan without collateral. And individual lenders earned higher returns than they would have received by parking their money in a savings account or a certificate of deposit. LendingClub, whose name is often in the first for search of P2P lending, helped pioneer the business model of "marketplace" lending. LendingClub is the largest online lender for personal loans in the United States, having facilitated more than $28 billion in loans since it was founded in 2007[10]. Main goal of this project is to analyze the lending club data to understand what are the reason behind the inclination of seeking loan from lending club instead of banks in USA, different factors that are effecting the loan amount and different group of people opting for lending club loan.

Keywords
Cartogram — Principal Component Analysis — tSNE— Linear Regression — Time Series Analysis

[1] *Data Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
[2] *United States*

Contents

Introduction

Private loans are a very common thing in US. There were a total worth of USD 7.8 Billion in 2014-15. By cutting banks out of the process, borrowers typically got a lower interest rate than they would have paid on a credit card or a loan without collateral. And individual lenders earned higher returns than they would have received by saving money in Savings account or in mutual funds. LendingClub is the largest online lender for personal loans in the United States, having facilitated more than $28 billion in loans since it was founded in 2007. [2]

Lending club is a peer-to-peer lending company which connects investors with customer who are seeking loan. According to this article[1], getting a loan approved from Lending Club has been tedious for borrowers. According to a TransUnion report [9] LendingClub has grown from 1% of the total personal loan market in 2012 to 10% in 2017 while the category has expanded more than 2x in that time period. TransUnion put total US personal loan balances outstanding at $47 billion in 2012 expanding to $112 billion in 2017, of which LendingClub has a 10% share. By seeing this statistics our curiosity increased to know about the reasons for it's exponential growth and why people opt for it. Our project focuses on the analysis of the data and the visualization to understand the trend of customers and various factors that are affecting Lending Club. We are trying to find answers of the trend of the loan, how it is varying with the region, factors effecting the interest rates, loan term changing status. Additionally our data provides us with 26 attributes such as loan applicants, loan status, region of the loan applicants- these attributes helped to

create rich visualization for deeper understanding of the data.

1. Data Set Description

The source of this data set is Kaggle. Previously the data was comprised of more than 50 variables but after cleaning the data we took 26 variables with 887351 observations. The different variables are: member_id, loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, purpose, zip_code, addr_state, state_full_name, delinq_2yrs, inq_last_6mths, open_acc, pub_rec, total_acc, total_rec_late_fee, tot_cur_bal. A data dictionary has been included in the Appendix for variables with their details.

2. Background and existing work

This data set has been analyzed from different perspectives to see which factor or factors have caused most number of customers. Following is the visualization done previously. [3]

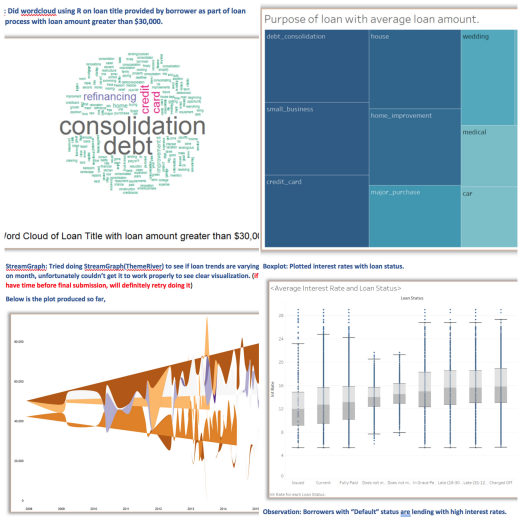

Figure 1. Effect of attributes on Employee Resignation

We can see that some of the visualizations like the steam graph are not giving out any meaningful insights.

The world cloud does need a lot of cleaning and the tiles have used area as the visual encoding and it is not very intuitive. Also, both the images in the top row are giving the same information and are redundant.

The whisker plot (bottom right) might have overlapping points which could mislead the interpretation. One must user jitter to avoid that.

Here are few more existing visualizations. In Figure 2, [8] we have distribution of Interest rate by Grade. The color choice for this plot is not explainable. To the right of the graph is a tree map showing various purposes of loans. Since tree maps are good for hierarchical data, purposes does not have any correlation with each other. Tree map for purpose is a bad choice of visualization. Moreover, the color choice is random and does not make much sense.

In Figure 3, the plot above shows variation of interest rate by loan terms and by Grade. It is a good visualization and gives us basic insights but the choice of color is poor for the graph in the lower half.

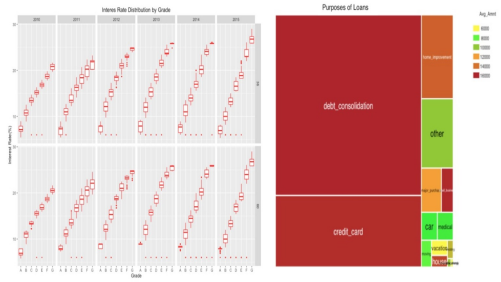

Figure 2. Interest Rate Distribution by Grade

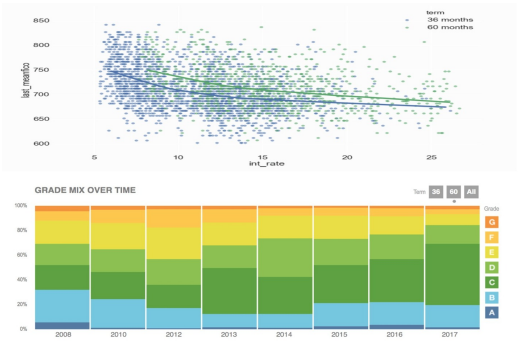

Figure 3. Interest Rate Distribution by Grade

Apart from these, the current visualizations are not very comprehensive and does not give a complete view of current situation. These are the things inspired us to dig further to get meaningful insights.

3. Research questions and working hypothesis

This data set has been analyzed from different perspectives at first we built the hypothesis and from there we tried to understand whether the hypothesis is holding true or not. Our research questions are:

- What are the components influencing the interest rate of the loan?
- How are the borrowers who default a loan different from the borrowers who do not?
- What are the factors that contributes on the loan amount

Following hypothesis are part of our analysis for this project.

- Interest rates varies in different states.
- Interest rates changes with loan term.

- Interest rates varies with purpose of loan application.
- Profession makes a difference on getting loans.
- Loan amount varies with Status of loan application.
- Loan amount varies with Annual Income of each person applying for loans and how interest rate varies with annual income.
- Grades has an impact on Interest Rate.
- Interest rate varies with employment length.
- Interest rates varies with delinquency.
- Interest rates varies with purpose of application for the loan.

4. Analysis Process

We applied a few exploratory data analysis techniques to see how various factors have affected the interest rate and loan amount. We visualized each feature to see which feature have a great impact on the measure under consideration (could be interest rate or loan amount for example) or it does not. We used several packages in R to have visually appealing graphs. This report describes different visualizations and insights that we have come up with to reach major conclusions.

4.1 Analysis of Data

We went through all the variables in the dataset and tried to analyze which are the features which leads people to take and post loan in Lending Club. There were around 50 features in the dataset with more than 8 million data points. By doing data wrangling, we made the data tidy by context i.e. we kept all the data which are relevant to our hypothesis and **filled missing values with median and other techniques (bivariate regression analysis).** Following are the processes we went through to analyze the data.

4.2 Visualization Methods

We made various plots like box plots, scatter plots, jitter plots, wordmap, histograms and heatmap to visualize the data. We used hexbins to create scatter plots where there were too many data points to visualize in the traditional scatter plots. The packages we used for visualizing these graphs are ggplot2[11] (to make visualizations), Maps [11](to plot the heat map), RColorBrewer(another package to enhance ggplot2 colors), wordcloud [5] (to make wordcloud) and tm [7] (to clean or make the data tidy for making the wordcloud) and hexbin[4]. We also made sure to use the best practices discussed in the class while creating the visualizations so that they deliver a clear and crisp message.

4.3 Failed Experiments

We tried to visualize various factors using different visualization techniques. But, few techniques were not appropriate for the kind of data we had. Following are such cases where our visualization experiments failed.

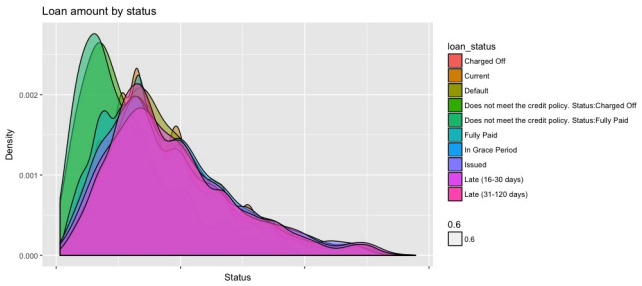

Figure 4. Loan Amount by status of loan application

In the figure above, we have the density plot of Loan Amount by status of loan application. This experiment failed as there are millions of data points, the density plot does not show the results properly. Visualization of various status is overlapping here which does not clarify the results.

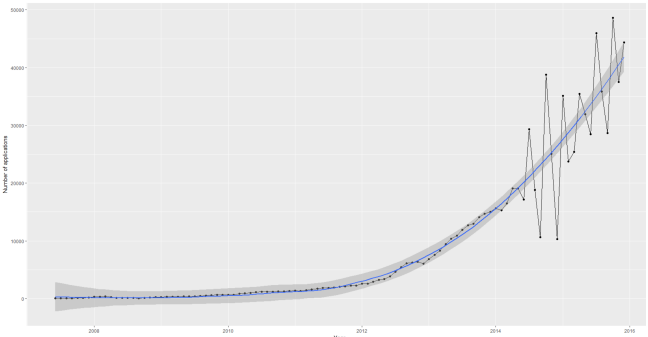

Figure 5. Time Series - Trend component Analysis

Figure 6. Time Series - Seasonal component Analysis

We also tried to a trend and seasonal component analysis of the time series data of number of loan applications, but could not find any actionable insights. We plan to dig deeper in the future.

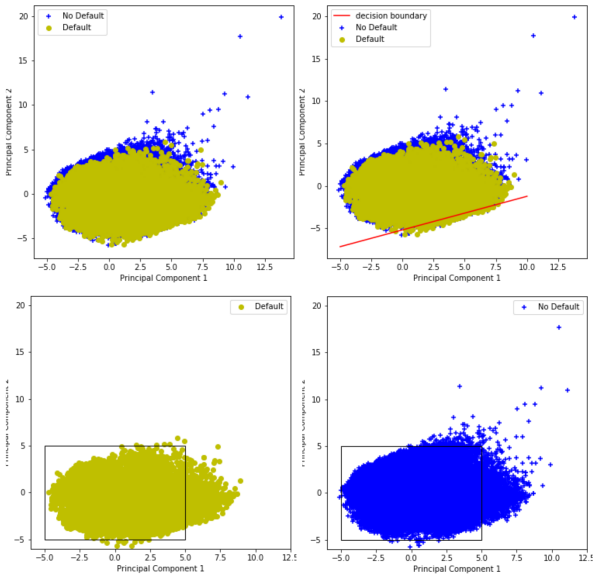

Figure 7. Principal Component Analysis of 2 most importance features

Another failed experiment is Principal component analysis and using TSNE on the extracted data. We tried to do principal component analysis on the given data and plotted the first two principal components. The plot came out to be misleading originally, when we looked at the data combined since there were overlaps. We decomposed it into multiple scatter plots as shown in the bottom 2 plots which confirmed the overlap. This clearly states that the data is not linearly separable and one must try some different technique for classification.

5. Feature Engineering

The dataset was huge. It had more than 70 features with 887,379 entries.

The aim was to create a model to predict factors affecting loan interest rate. Various tools and techniques has been used to find insights about the data. We used Python, SAS, R as our primary programming and scripting languages.

5.1 Feature Selection

We first checked whether the dataset had missing values. Around 20 features had more than 75% missing values. And two more features had 51% and 28% missing values. We dropped these 22 features since they weren't even relevant as factors that might affect loan interest rate.

Out of the remaining 50 features, 17 had less than 10% missing values. 8 of these features were again irrelevant with regards to predicting loan interest rate. We dropped them to be left with 9 features (<10% missing values) that required filling of missing values. 7 out of the 9 features had exactly 29 missing values. If the 29 entries were common to all the 7 features, dropping these 29 entries would not be an issue since we would then be dropping approximately 0.0003% values and at the same time handling missing values of 7 features. We checked and found the 29 indices to be common to the 7 features, so we decided to drop those 29 entries. We now had 2 features that required handling missing values with the total number of entries now reduced to 887,350 (887,379 - 29).

These two features were namely - annual income (4 missing values) and total current balance (70,276 missing values). We checked the value counts of annual income and the top 3 annual incomes, with respect to counts were 60000.00 (34281), 50000.00 (30575), 65000.00 (25498). It would not matter if we either dropped the 4 entries, or fill them with any one of the top counts. We decided to keep the entries and fill them with values '60000.00'. This left us with the last feature requiring missing values handling which was total current balance. The missing values of total current balance had to be filled since dropping them would result in a significant loss of data (approximately 8%). Total current balance was a continuous feature and not a categorical one, so it could not be filled by checking the counts of values alone.

We decided to run Bivariate Linear Regression on total current balance to fill it's missing values. To execute Bivariate Linear Regression we would require two variables. The first being total current balance. The second would be the feature that had the highest correlation with total current balance. We checked the correlation between total current balance and the other features and found annual income to have the highest correlation (0.42).

To proceed with Bivariate Linear Regression on total current balance and annual income, we followed the following steps:

1. Split null and finite valued entries into two sets. Finite valued being training set while null valued being testing set.
2. Use the annual income and total current balance columns of training set (X-inc and X-bal) to train the bivariate regression model. This model would be used to predict null values of total current balance from annual income in testing set (Y-inc).
3. Merge both the datasets, once the predicted values were filled.

We followed the above steps to fill the missing values of total current balance. To verify the aptness of the filled values in total current balance we checked it's correlation again with annual income. The new correlation was 0.43 which was very close to the previous correlation. We thence concluded the missing values were ideally filled.

Finally we had the complete dataset with no missing values and no features that ideally would not affect loan interest rate. We had 23 features (including loan interest rate).

Our next step, was to run one hot encoding on the categorical features (from the 22 features) and drop a column from each to avoid dummy trap. The result left us with 290 features (excluding interest rate). When the number of features are so high, it's advisable to run feature selection techniques. We decided to execute scikit-learn's variance threshold. It's a

feature selection technique that eliminates the features having variance less than a threshold pre-decided by the user. We decided to eliminate the features with <0.8 variance. This would reduce the number of features and would also filter the relevant features. Executing scikit-learn's variance threshold algorithm with a threshold of 0.8 on our dataset of 290 features, filtered out 18 features. We again checked their correlation. Loan amount was highly correlated to funded amount. Since loan amount is technically more important in predicting loan interest rate than funded amount, we decided to drop funded amount.

We finally had 17 features (excluding interest rate) with 887,350 entries to create a model to predict loan interest rate.

5.2 Linear Regression Model in SAS

We proceeded by generating a simple linear regression model on interest rate in SAS. The first linear regression model had very low parameter coefficients of total current balance and annual income. We took the log transforms of both total current balance and annual income and generated the linear regression model once again. This time the Variance Inflation values of installment amount and loan amount were slightly high, thereby hinting at presence of multicollinearity in the model.

We then decided to create a new variable - 'number of installments' by taking the ratio of loan amount to installment amount. The features installment amount and loan amount were then dropped. We generated the simple linear regression model. This time there was no issue with either parameter estimates, P values (all <0.05 so significant) and variance inflation values (all ≤ 5.2). The final model had an R^2 value 76%.

The statistics of the final linear regression model on interest rate can be seen in figure 8.

Root MSE	2.12910	R-Square	0.7639
Dependent Mean	13.24644	Adj R-Sq	0.7639
Coeff Var	16.07297		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	66.56429	0.05936	1121.46	<.0001	0	.	0
ln_annual_inc	1	-0.50380	0.00514	-98.06	<.0001	-0.06074	0.69357	1.44182
emp_length_10__years	1	0.01691	0.00493	3.43	0.0006	0.00181	0.95107	1.05144
grade_B	1	-0.44953	0.00581	-77.42	<.0001	-0.04640	0.74083	1.34983
grade_C	1	0.45109	0.00555	81.21	<.0001	0.04608	0.82665	1.20970
home_ownership_MORTGAGE	1	-0.05232	0.00824	-6.35	<.0001	-0.00597	0.30095	3.32284
home_ownership_RENT	1	0.16324	0.00812	20.12	<.0001	0.01826	0.32291	3.09688
inq_0_0	1	-1.04135	0.00652	-159.77	<.0001	-0.11794	0.48836	2.04768
inq_1_0	1	-0.54022	0.00708	-76.26	<.0001	-0.05488	0.51393	1.94580
loan_amnt_inst	1	-1.04660	0.00088359	-1184.5	<.0001	-1.39771	0.19113	5.23200
purpose_credit_card	1	-0.92875	0.00724	-128.35	<.0001	-0.08953	0.54701	1.82811
purpose_debt_consolidation	1	-0.41346	0.00619	-66.78	<.0001	-0.04640	0.55130	1.81390
term__36_months	1	-15.74088	0.01137	-1384.2	<.0001	-1.64643	0.18811	5.31602
ln_tot_cur_bal	1	-0.04808	0.00263	-18.26	<.0001	-0.01322	0.50759	1.97010
total_acc	1	-0.00826	0.00021132	-39.09	<.0001	-0.02232	0.81617	1.22524
verification_status_Source_Verif	1	0.18514	0.00569	32.57	<.0001	0.02042	0.67713	1.47682
verification_status_Verified	1	0.65711	0.00592	111.08	<.0001	0.07041	0.66235	1.50978

Figure 8. Statistical representation of the Linear Regression Model in SAS

The final regression model suggested that interest rate was driven by -

- annual income,
- employment length (>10 years),
- loan grades (B and C),
- type of home ownership (Mortgage and Rent),
- number of inquiries in credit file (0 and 1),
- number of installments,
- purpose of loan (credit card and debt consolidation),
- total current balance,
- total number of accounts,
- verification status (Verified and Source Verified), and
- term (36 months) [most important]

6. Exploratory Data Analysis, Insights & Results

To begin with, in the following graph, we tried to see the distribution of loan applicants through out the country. We plotted the heat-map in R using Map package. We collected latitude and longitude of all the states in USA (using map package in R) and then calculated the frequency of applicants divided by the population of respective state and plotted it on a log scale since its a ratio. We observer that Idaho (ID) and Iowa (IA) has the lowest ratio of loan applicants to population while North Dakota (ND), Nebraska (NE) and Maine (ME) have pretty low ratios too.

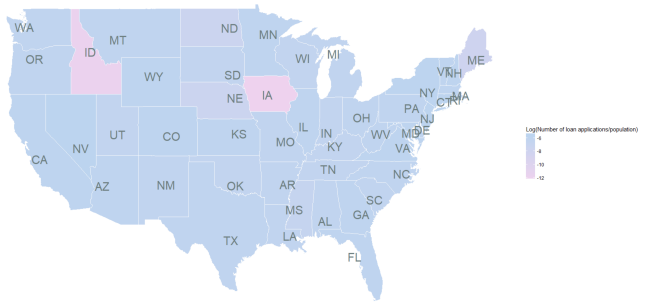

Figure 9. Distribution of Loan Applicants

There are many reason for this uneven distribution. First of all California has lending club headquarter and cost of living is high in West Coast. Likewise New York, Florida are also highly populated. Moreover, if we go through the wordmap of professions of people applying for loans are mostly Managers, CXO and high pay-grade employees and it is evident that, most of the high post professionals are in West Coast. However, population could be a major factor, which could make the data biased and it is our future work. These are some of the reason for this distribution. Further we will analyze other aspects of lending club data.

- Variation of interest rate across different states -

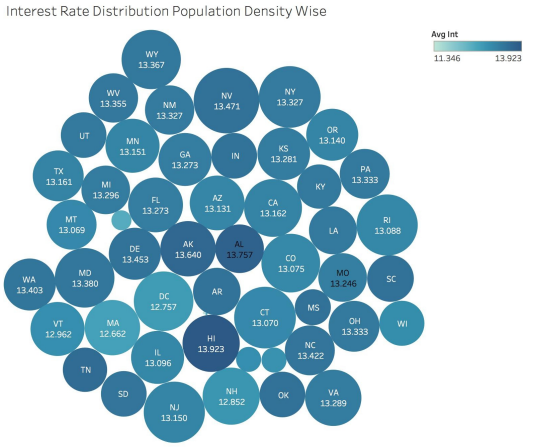

Figure 10. Grade Analysis in terms of Loan Applicants, Loan Amount and Interest Rate

In Figure 10 we have plotted a cartogram, population density wise to analyze the distribution of interest rate in all the states. We chose Dorling Cartogram as a method of visualization because, it combines statistical information with geographic location. We have considered population density instead of population to normalize the population and to get more accurate insight. As we can see the interest rate is varying from 11.346 to 13.923. Most of the states where the population and the loan applicants are high the interest rate is around 13 % where as states with less population and loan applicants the interest rate is low. So we can connect the similarity here with figure 6. Although there are some other factors on which are determining the interest rate. This has been discussed in details in section 5.2.

Let's dive a bit deeper in this. What you see here is a histogram faceted by the purpose. Since we are trying to compare the distribution pattern here, the histograms are created on a free y-scale. On x-axis, we have the loan applicant's credit score grade while the color of the histogram represents the interest category.

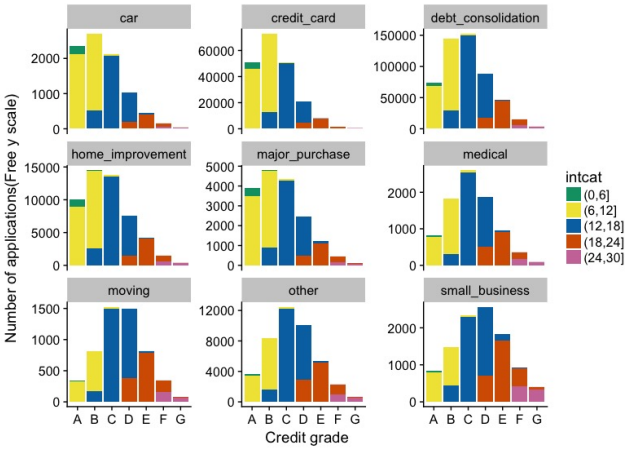

Figure 11. Distribution of loan applications across credit grades and interest

What we see here is that for purposes like buying a car or paying off the credit card bill or for debt consolidation, people who have better grades are dominating the distribution - this could be interest to preserve the grades by paying off existing loans with help of the loan given by lending club. But as we see, for medical purposes - there are considerable people with low credit grades who are taking loans with higher interests. This could be because traditional banks are not ready to loan money to this set of people, and same could be the case with people who are taking loans for moving or to start a small business.

Next we made another wordcloud of title of people getting loans approved or which professionals are granted for loans.

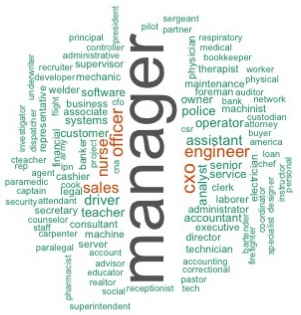

Figure 12. Professionals getting more loans

The insights we got from the wordcloud is interesting as we saw managers are mostly approved for loans. Next come officers, CXO, engineers, sales analyst and nurse. We could conclude the fact that, most people applying for loans are highly paid professionals with high annual income which is obvious because, most of the Lending Clubs lenders would feel comfortable to lend money to people who could pay it off on time. Word clouds are not generally considered the best way to visualize text data, but it seems to be a good option in this case because we have keywords instead of text data and hence there is no noise. Also, our emphasis here is on the frequency of word occurrence, which is best highlighted by a word cloud.

Next insight we tried to find out is variation of Loan amount with Status of the loan application i.e. does the amount of money granted for loan varies for people not in a fully paid status.

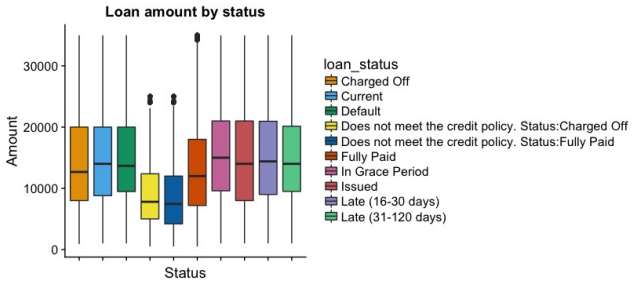

Figure 13. Variation of Loan Amount with Status

From the graph we could see that people who does not meet the credit policy of Lending club and not getting a higher amount of loans approved. People with any other status are on the same range on an average. They are getting loans approved in amount ranging between 8000 to 20000.

To understand the effect of various factors on the interest rates of lending money, we first investigated the variation in the interest rate data. More than 50% of the interest rates lie between 10 to 16.5. Then we tried to map each factors with interest rate to see how interest rate varies.

- Variation of interest rate with Grades -

The below graph shows how interest rate varies with Grades. These grades ranging from A to G in a scale where good being A and bad being G are assigned by Lending Club to differentiate between borrowers.

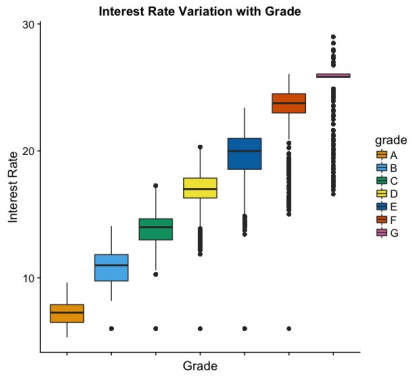

Figure 14. Variation of Interest rate with Grade

As we can see the major chunks of the grades have distinct interest rates. Grade G has the highest interest rate while grade A has the lowest interest rate. We also tried to look into other factors like Loan Applicants, Loan Amount and Interest Rate are variation in terms of Grade.

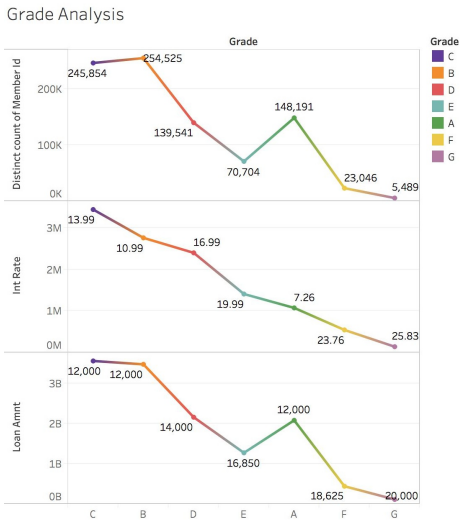

Figure 15. Grade Analysis for Loan Amount, Interest Rate, Number of members in each grade

Here we are trying to understand the division of grades, that is why we have taken member id count and median of loan amount and interest rate. As we can see colors are divided in terms of different grades. In the graph we can see a very similar pattern specially for loan amount and member id. For group B and C average loan amount is quite high whereas number of customers are are quite close for this 2 groups, but interest rates are 13.99 and 10.99, so we can conclude there are other factors effecting the interest rate. For groups D, E, A, F and G the pattern is very similar for count of member id and loan amount, but similar to other groups interest pattern is different for the grades. Specifically for Grade A interest rate is low where count of member ID and average loan amount is pretty high.

Figure 16. Grade Analysis in terms of Loan Applicants, Loan Amount and Interest Rate

For further analysis we have included term variable and plotted median loan amount and interest rate to understand the flow. We can see people are mostly preferring long term loan that is 60 months. Grade A and B are the most popular loan

type specially grade A, as we can see interest rate is really low, where as grade G type is the least taken loan type. Here we can see median loan amount is less and interest rate is very high around 25.80.

- Variation of interest rate with Employment Length -

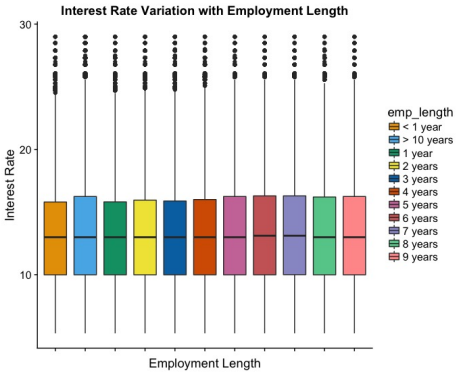

Figure 17. Variation of Interest rate with Employment Length

Employment length is in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. As we can see from the above plot, employment length is independent of interest rates. Hence, we can infer that employment length does not affect interest rates, but we can observe majority of interest rates between 10 to 16.

- Variation of interest rate with Delinquencies Incidences -

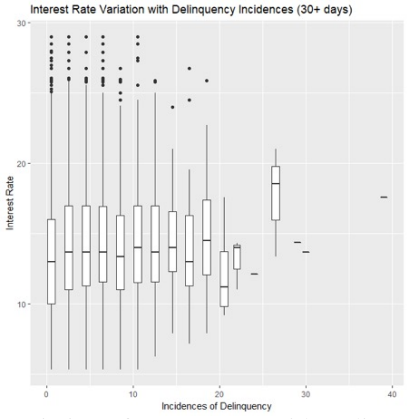

Figure 18. Variation of Interest rate with Delinquency Incidence

Delinquency is failure to pay an outstanding debt for a certain amount of time. Here we have the data with 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years As we can see most delinquency incidences are less than 11 for interest rates between 10 to 16.5. Of course there are few outliers but, with increasing amount of delinquency, interest rate also increases. i.e. people who have a history of not paying outstanding dept on time are getting loans at a quite high interest rate in later times.

- Variation of interest rate with purpose of Loan -

We tried wordcloud for variation of loan amount with purpose for applying loans.

Figure 19. Variation of Loan amount with purpose

We could clearly see from the wordcloud that, debt consolidation is the most popular reason for people to get approved loans. Then follows credit card bill payment, home improvement and major purchase. We chose word cloud for visualizing this factor because, word cloud shows the most occurring words in a block of sentence and by using word cloud, we got the opportunity to use Text Mining techniques to find word frequency. We can observe purpose of loan affects interest rates. The median of interest rates for small businesses, renewable energy and for buying a house is above 15%, while interest rates for cars and credit cards are the least.

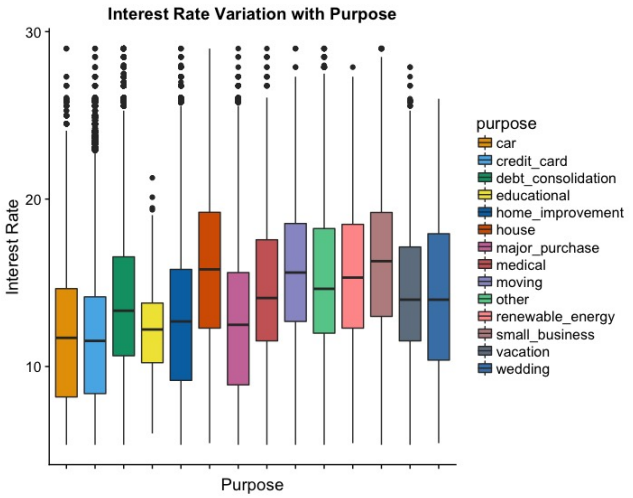

Figure 20. Variation of Interest rate with Purpose

Lets see How the Loan Applicants, Loan Amount and Interest Rate are varying in terms of Purpose.

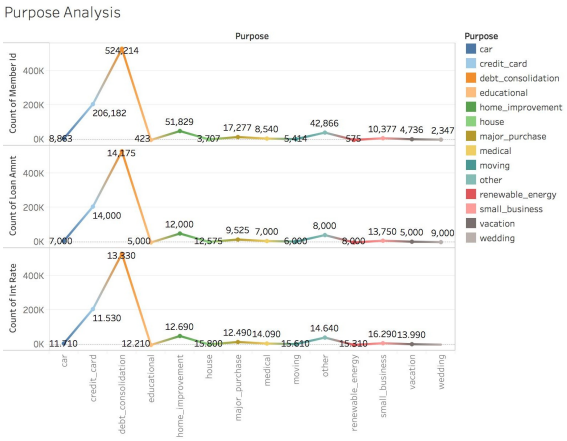

Figure 21. Purpose Analysis in terms of Loan Applicants, Loan Amount and Interest Rate

This graph is depicting the variation of Loan applicants, loan amount and interest rate for different purposes, we are also trying to find any dependency on purpose for these 3 features. We can see a clear dependency on the purpose, very similar pattern we can see here for all three variables. Most of the people has taken loan for Debt consolidation and then credit card and depending on that interest rate for these groups are high, but there are groups with very less number of people like other, moving or small business here the interest rate is high based on different scenarios. Finally we can conclude that our word cloud visualization for purpose is giving us good result.

Now let's move on to the next question.

- Variation of interest rate with loan term -

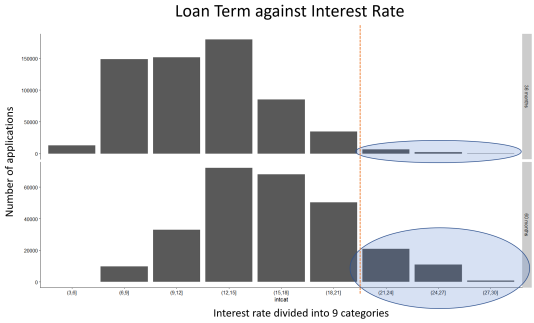

Figure 22. Variation of Interest rate with loan term

As we can see in the graph above, the number of applications to the left of the red line where the interest rate is greater than 20 significantly increases for long term loans. Since the number of long term applications are less than short term applications, the y-axis was kept free so we could compare the distributions.

- Is a time series analysis possible?

In our initial hunch, we found that we could not do time series analysis on the data we have, but after cleaning for null values and outliers, the cleaned data looks usable for ARIMA forecasting.

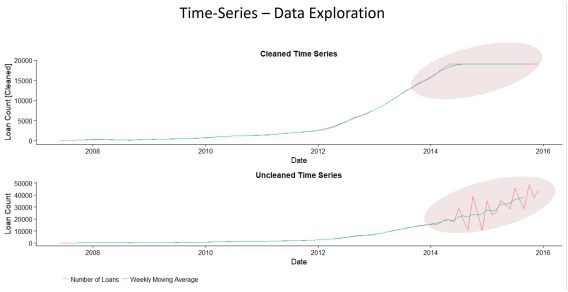

Figure 23. Time series: Moving average over week

We can see that the moving averages fit the cleaned time series much better while the original time series has higher variance due to outliers.

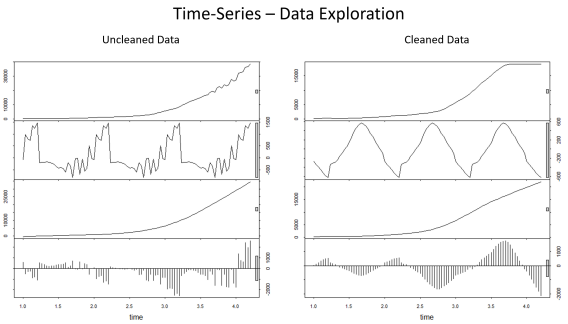

Figure 24. Decomposed time series

As we can see, the trend line (third row of visualizations) fits the time series better and the seasonality (second row of visualizations) is much more distinct in case of the cleaned data.

6.1 Classification and dimensionality reduction

Dimensionality reduction is a technique of feature extraction where we extract few most impacting features from an initial set of measured data and builds derived features intended to be informative and non-redundant. Principal Component Analysis is a type of dimensionality reduction technique. Here, few most important variables are taken into account. The new variables obtained after PCA are called Principal Components and they are orthogonal to each other which indicates that they are independent of each other.

In Figure 23, we have a pair plot of PC1 and PC2 with respect to Loan Status. It shows that, the Principal components are overlapping showing they are not independent of each other. We do not see any data with Loan Status 0 (blue dots) in the Principal Components PC1 and PC2.

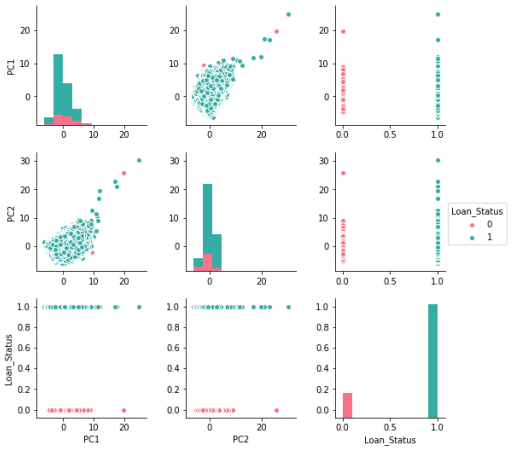

Figure 25. Principal Component Pairplot

We also tried to do tSNE on the derived data. t-Distributed Stochastic Neighbor Embedding(tSNE) is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two or more dimensions. While performing tSNE with complete dataset, the system was crashing due to high volume of data. So we sampled 2000 data points and performed tSNE on it.

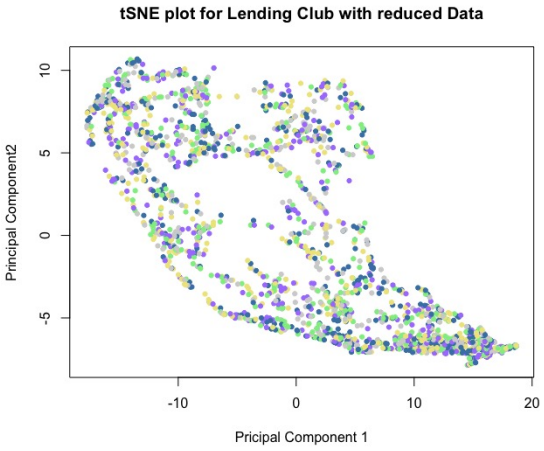

Figure 26. tSNE Plot with the Reduced Data

Figure 26 shows the graph of it and it is evident that the data is overlapping and not independent.

7. Conclusion

Throughout the visualization, we saw that credit grades, term, annual income, delinquency and purpose have impact on the interest rate. From the word cloud visualization we can see maximum customers are of manager level and they have applied for loans mostly with a purpose of debt consolidation and credit card. Through a regression model to predict the interest rate we found out features affecting it. For predicting loan status for future customers we tried to apply classification model and used PCA for feature selection. This did not work because the extracted Principal components are not linearly separable. Although these visualizations are showing a potential relationship over here, we have to deep dive and test if a correlation really exists. We also saw that the employment length does not have an impact on interest rates - this is against the general notion and hence is a valuable insight. The visualizations also did show that there are considerable amount of loans which are delayed on repayment and hence need to be analyzed further to see how can we figure out such applications and mark them risky. Purpose wise and grade wise distribution of loan amount and interest rate are giving us good insights for determining the interest rate. Combining the cartogram and map we are getting a detailed view of loan applicant and interest rate distribution. We have chosen colors very cautiously for visualization, so that everyone including color blind can differentiate and analyze the plots. We have not restricted ourselves with the by default colors (Both in R and python), manually tweaked colors to get nice and interpretable visualization.

8. Future work

In the visualizations and exploratory data analysis, we found that few states like, California, Texas and Florida have higher rate of loans. So, we would like to do some more analysis on how population of the state matters .

We would be fitting few models to see different perspectives of Lending Club loan giving strategies and what are the main reasons for people to get more loans. We have only used Logistic Regression for classification and linear regression. We would like other classification and prediction model to improve and verify our model.

We have already research work using visualization but there are lot more to be done. We will do more experiments to get interesting insights.

8.1 Predictive Forecasting

We can use the current time series data to forecast the number of applications in future and visualize the same to see if it makes sense. There are a number of techniques to do this - we will be trying ARIMA (auto regressive integrated moving average) and multiplicative time series models.

9. Acknowledgements

We would like to thank our Professor YY for his support and advice all throughout. We would also like to thank our Assistant Instructor as well for his technical guidance and feedback.

10. Appendix

10.1 Data Dictionary

Table 1. Description of attributes used in analyses

Attribute Name	Description
Member_ID	Unique ID for each member applying for loan
loan_amnt	Loan amount for which the application is made
funded_amnt	Total amount funded
funded_amnt_inv	Total amount funded by investors
term	Loan term (36 or 60 months)
int_rate	Interest rate in %
installment	Loan installment per month
grade	Grade assigned by Lending Club based on credit score
sub_grade	Sub-grade assigned by Lending Club based on credit score
emp_length	Number of years for which the loan applicant has been employed
home_ownership	Home ownership status for loan applicant (rented, own property etc.)
annual_inc	Annual income of the loan applicant
verification_status	Verification status of applicant's income
issue_d	Loan issue date
loan_status	Current loan status (Fully paid, charged off, default etc.)
purpose	Purpose for which the applicant applied for loan
zip_code	Zip code of the applicant's address
addr_state	State abbreviation of the applicant's address
state_full_name	Full state name of the applicant's address
delinq_2yrs	Delinquency in past 2 years
inq_last_6mths	Number if inquiries in past 6 months
open_acc	The number of open credit lines in the applicant's credit file
pub_rec	Number of derogatory public records
total_acc	Total number of credit lines in the applicant's credit file
total_rec_late_fee	Late fees received till date
tot_cur_bal	Total current balance of all accounts

Please see the table above for a detailed information about each of the attribute used in our analyses.

10.2 Color Choice for the Graphs

Color choice for the graphs has been done according to color blind friendly color palette. We made a Color blind friendly Color pallette in R and are using it in the graphs.

References

[1] *Loan approval with Lending Club*.

[2] *Private Loans: Facts and Trends*.

[3] Vishnu Vardhan A. *Existing Visualization*.

[4] Dan Carr, ported by Nicholas Lewin-Koh, Martin Maechler, and contains copies of lattice functions written by Deepayan Sarkar. *hexbin: Hexagonal Binning Routines*, 2016. R package version 1.27.1.

[5] Ian Fellows. *wordcloud: Word Clouds*, 2014. R package version 2.5.

[6] Wikipedia Foundation. *tm: Peer-to-peer Lending*, 2017. R package version 0.7-1.

[7] Kurt Hornik Ingo Feinerer and David Meyer. *tm: Text Mining*, 2008. R package version 0.7-1.

[8] Shu Liu. *Data Visualization of LendingClub Loans*, July 2016.

[9] Peter Renton. *The Opportunity on the Borrower and Investor Side*, December 2017.

[10] Nerd Wallet. *tm: LendingClub Personal Loans*, 2017.

[11] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2009.