EECE 5642

Data Visualization

Final Project Report

Lending Club Loan

Yao Jin

Rui Ma

Yan Liu

4/23/2018

Content

1. Abstract

This report describes the detailed procedures of analyzing and visualizing the Lending Club Loan dataset. It introduces the background and motivation of the organization Lending Club. And then it explains the data structure and content of the dataset. Similar real world designs and works from other authors are analyzed for literature review. The goal of the project is to find out what factors have significant influence on the interest rate. Design principles and novelty analysis are discussed in the report. Visualization results are presented and discussed in the report. Significant features are identified and summarized in the conclusion.

2.Introduction

2.1 Background

Lending Club is a US peer-to-peer lending organization, which is the world's largest peer-to-peer (P2P) lending platform[1]. It reduces the cost of lending and borrowing for individuals with advanced data analytics. The function of peer-to-peer borrowing is to match people who have money with people who want to borrow money.

2.2 Data Source & Data Process

The dataset of Lending Club Load was downloaded from Kaggle. The link to our data source is: https://www.kaggle.com/wendykan/lending-club-loan-data/data

It contains complete loan data for all loans issued from 2007 to 2015, with **890k** records and **75** features (such as Current Loan Status, Latest Payment Information, Credit Scores, Number of Finance Inquiries, Address including Zip Codes, and States). The challenges for the visualizing the dataset is its volume, and data complexities. The dataset contains a mixture of numeric and categorical features.

To address those challenges, following steps were taken:
(1) Preprocess the dataset with programming language Python/R to remove missing values and regulate the format of the dataset;
(2) Analyze the pre-processed dataset to select the most significant features to visualize.

13 Features are selected for visualization and analysis, they are presented below,
Loan Status -- Status of the loan (fully paid, current being paid, charged off, etc)
Loan Purpose -- Purpose of the loan (debt consolidation, car, credit card, etc)
Loan Amount -- Amount of a loan, in USD
Installment -- Monthly payment of a loan
Annual Income -- Annual income of the borrower

Debt to Income Ratio (DTI) -- Borrower's debt amount divided by annual income
Interest Rate -- Interest rate of a loan
Home Ownership -- The housing status of the borrower
Loan Grade -- Borrower's credit score, ranked from A (the best) to G (the worst)
Pay-off Term -- The number of payments (monthly) on the loan, either 36 or 60
States -- Loan's issued state
Year -- Loan's issued year
Verification Status -- Whether a borrower is source-verified or not

2.3 Real-world Applications

Real-world similar designs are listed below.

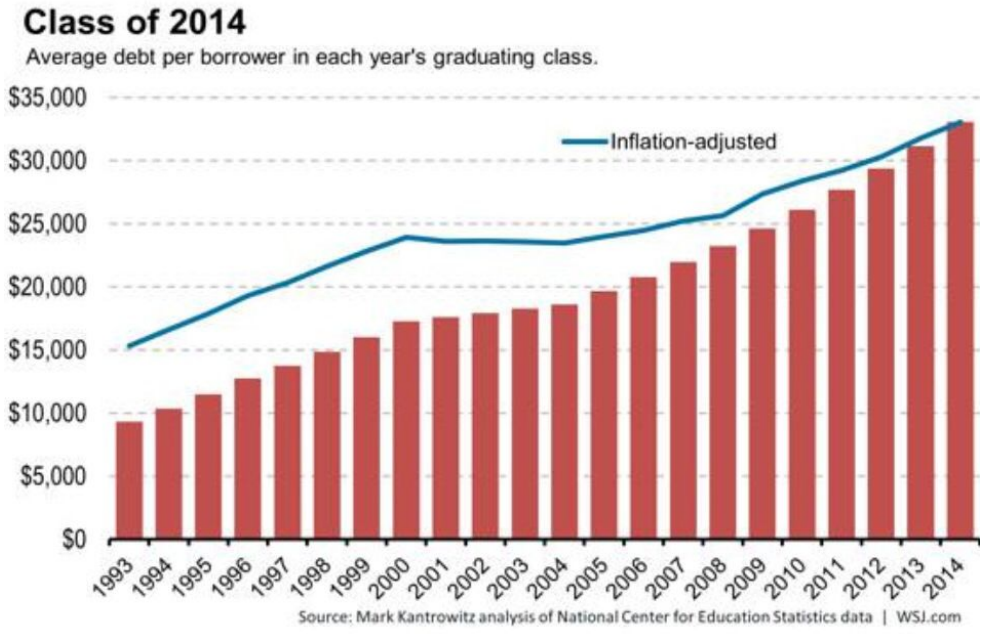

Fig. 1 Average Debt Per Borrower in Each Year's Graduating Class[2]

Fig. 1 is a bar plot which illuminates that the average debt per borrower is growing by years.

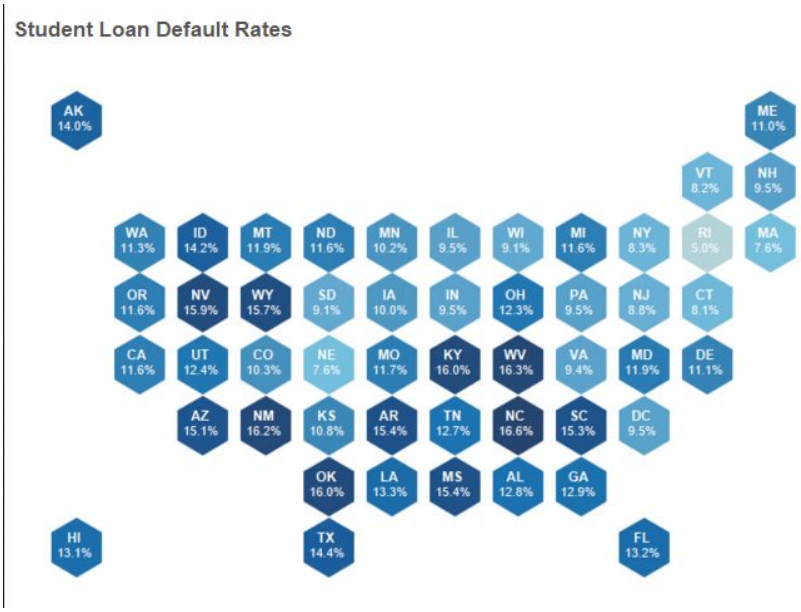

Fig. 2 Student Loan Default Rates[3]

Fig. 2 displays the variation in student loan default rates by states.

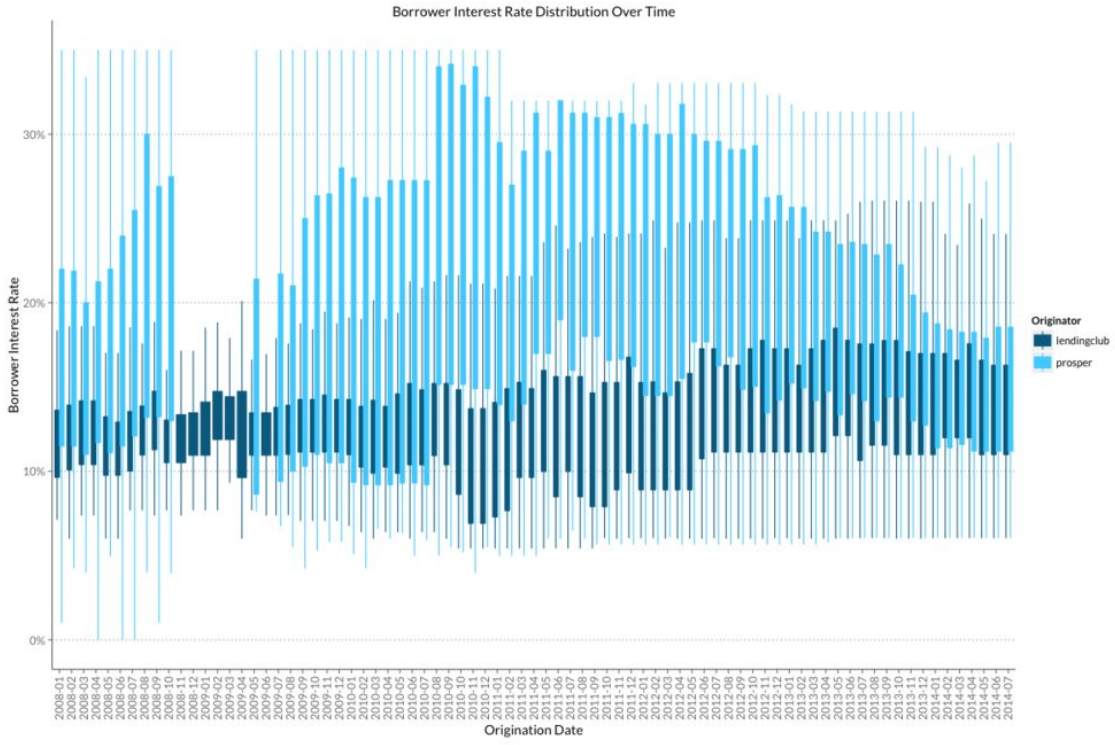

Fig. 3 Borrower Interest Rate Distribution Over Time[4]

Fig. 3 compares the distribution of borrower's interest rate over time of two lending platform: Lending Club and Prosper.

2.4 Contributions of the Work

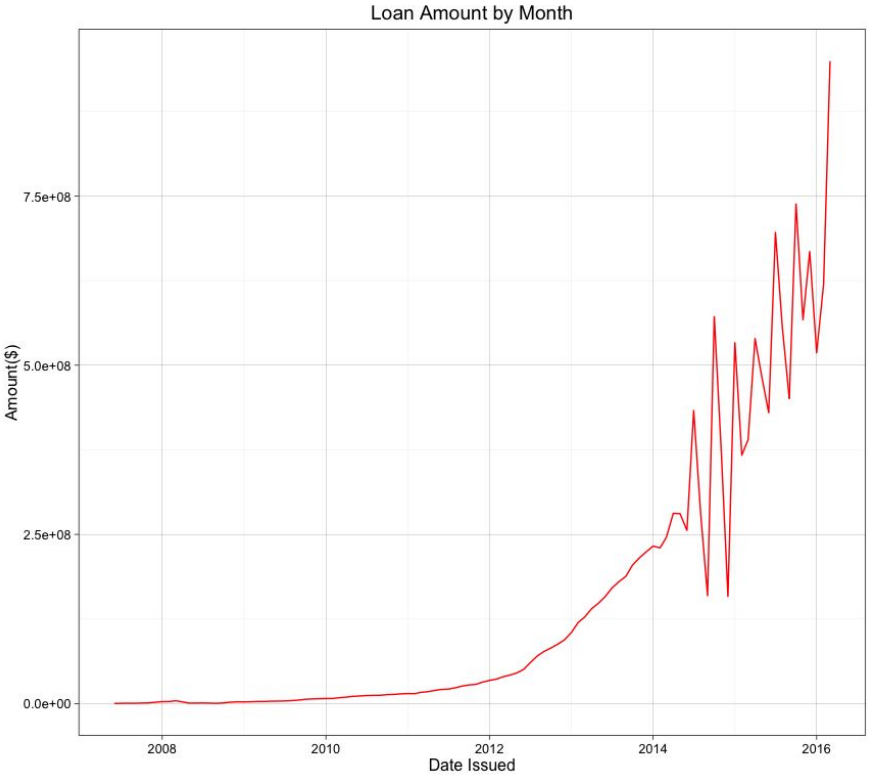

Fig. 4 Loan Amount over Time[5]

Fig. 4 displays the total loan amount of Lending Club by months from year 2008 to 2016.

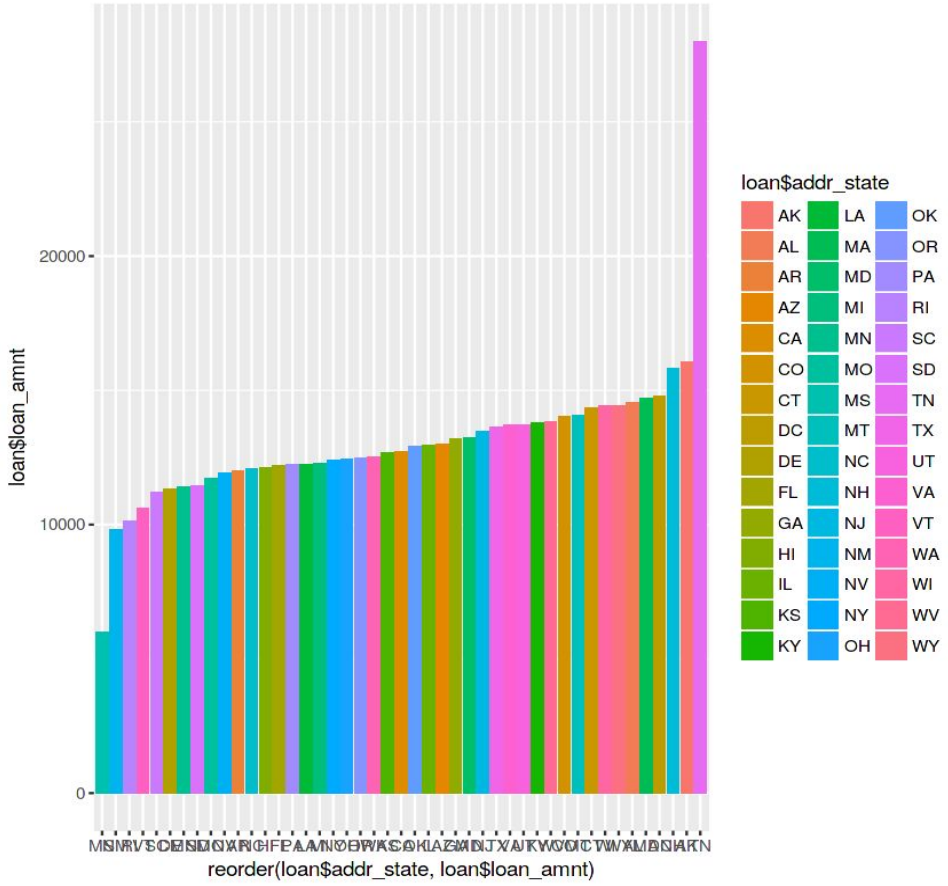

Fig. 5 Total Loan Amount by State[6]

Fig. 5 is a bar plot that displays the total loan amount of each state. It can be seen that it is difficult to read horizontal axis. the color use make the graph hard to read as well. This graph can be represented in a state heatmap form to display the results clearer.

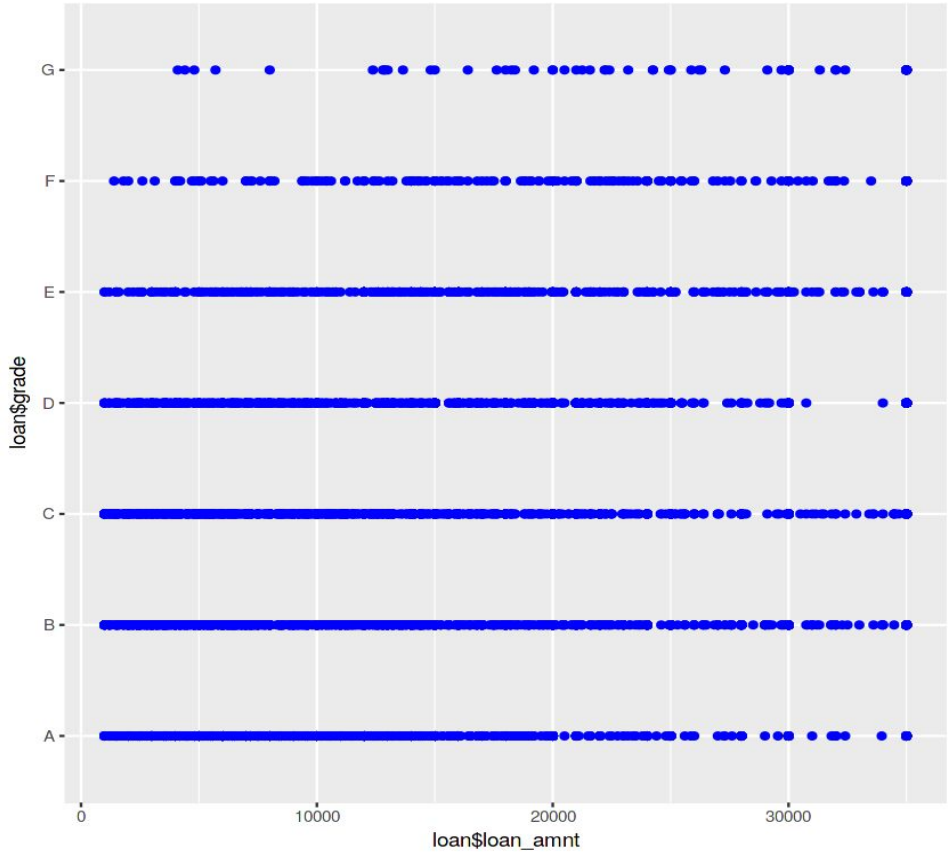

Fig. 6 Loan grade v.s. Loan Amount[6]

Fig. 6 is a scatter plot between loan grade, a categorical feature, and loan amount, a numeric feature. It is also difficult to obtain useful information from this graph, as the loan amount of each grade varies in large range. This graph can be represented using box-plot to display the results better.

2.5 Motivations

By visualizing some features of the data, a good understanding of characteristics for these features can be obtained. Another issue is to figure out what factors have dominant influence on the interest rate of the loan.

2.6 Novelty of the Work

Random forest model is created to visualize each feature contribution to interest rate.

2.7 Tools and Programming Languages

Data Visualization Tools: Tableau, Jupyter Notebook.
Programming Languages: Python, R.

2.8 Design Principles

a. Show the data in details with different aspects
b. Present the plots in clear, detailed form
c. Provide the viewers a thorough understanding of the data
d. Maximize the data-ink ratio
e. Keep the lie-factor between the interval (0.95, 1.05)

3. Visualization Techniques & Results

3.1. General Visualization

Some general features are visulaized to gain a comprehensive knowledge of the dataset.

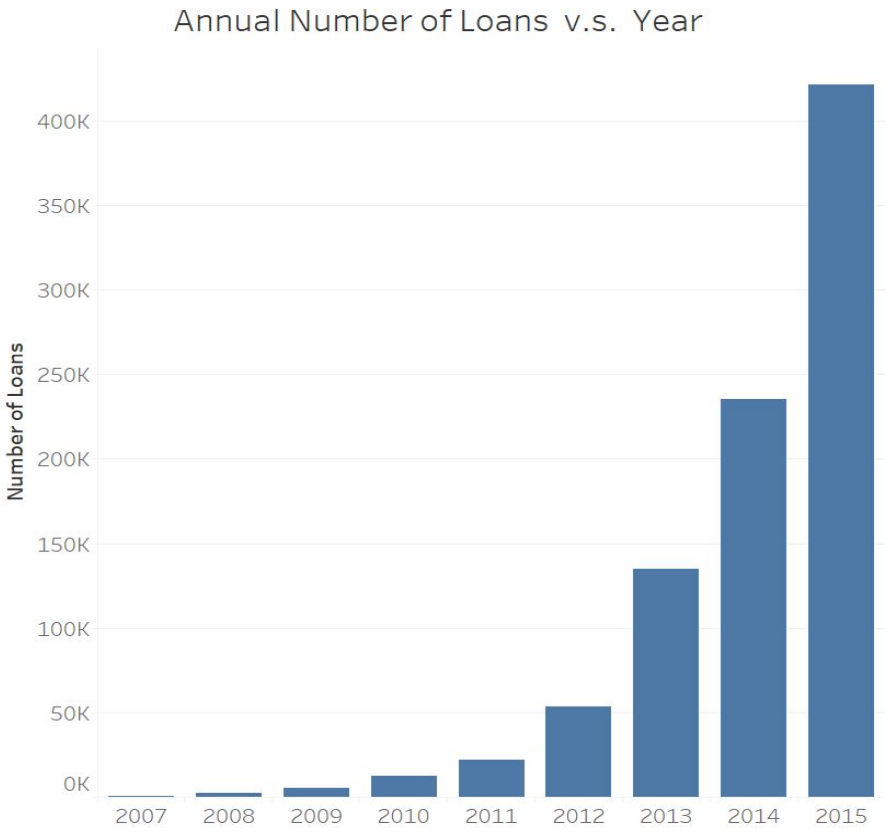

Fig. 7 Annual Number of Loans v.s. Year

Fig. 7 displays the annual number of loans in terms of years. It shows the growing pattern of Lending Club. From this plot, it can be seen that the annual number of loans increases every year, from hundreds in 2007 to more than 400,000 deals in 2015.

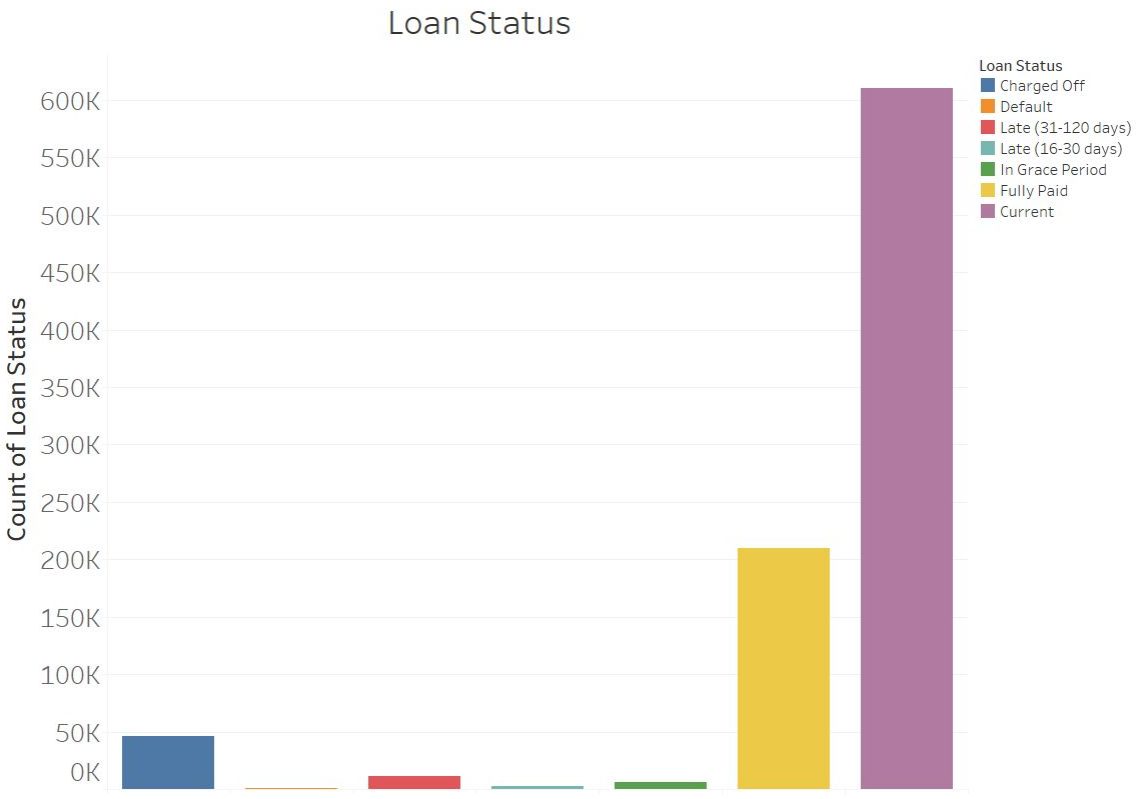

Fig. 8 Loan Status

Figure 8 displays the distributions of loan status. it can be seen there are 7 types of loan status: charged off, default, late(31-120 days), late(16-30 days), in grace period, current, fully paid.

Each loan status is explained below,
Charge off -- Loan for which there is no longer a reasonable expectation of further payments, typically occurs when a loan is no later than 150 days past due.
Default -- Loan has not been current for 121 days or more.
Late(31-120 days) -- Loan has not been current for 31 to 120 days.
Late(16-30 days) -- Loan has not been current for 16 to 30 days.
In grace period -- Loan is past due but within the 15-day grace period.
Current -- Loan is up to date on all outstanding payments.
Fully paid -- Loan has been fully repaid, either at the expiration of the 3- or 5-year year term or as a result of a prepayment.

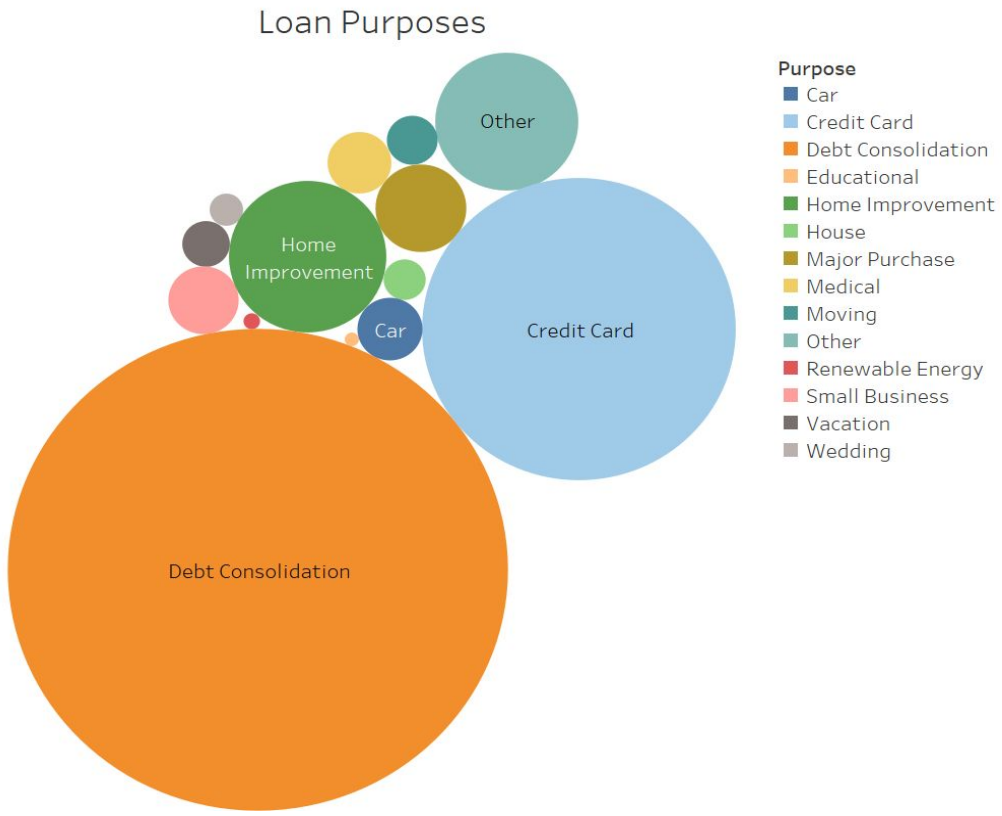

Fig. 9 Loan Purposes

Fig. 9 shows the loan purposes of the lending club such as for car, credit card, debt consolidation, home improvement, education and so on. It can be seen that debt consolidation is the No.1 popular purpose of loan, and the second is for credit card, and the third one is home improvement. Note that there are some small bubbles in different colors with no words on it because the words cannot show on such small sizes. Nevertheless, clicking the circle can present a more detailed figure, and it will show the details of the name and data for this purpose. Link for interactive display: https://public.tableau.com/profile/rui.ma4337#!/vizhome/LoanPurpose/Dashboard8

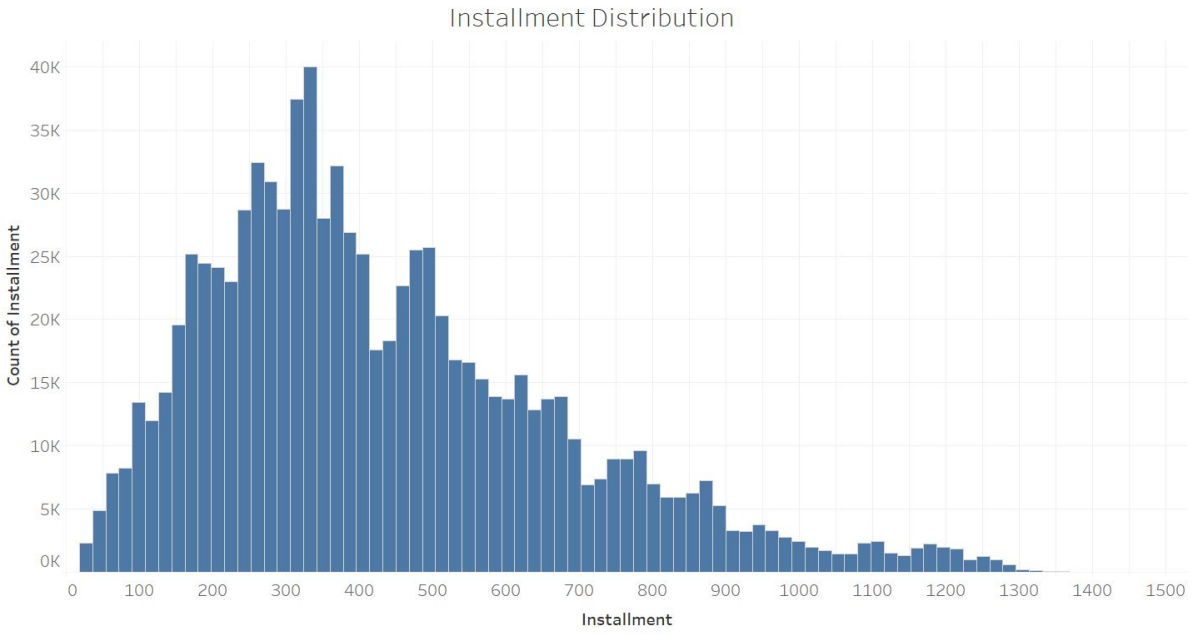

Fig. 10 Installement Distribution

Fig. 10 shows the distribution of installment. It is a skewed Gaussian distribution. According to this figure, the most installment lies between $200-$500.

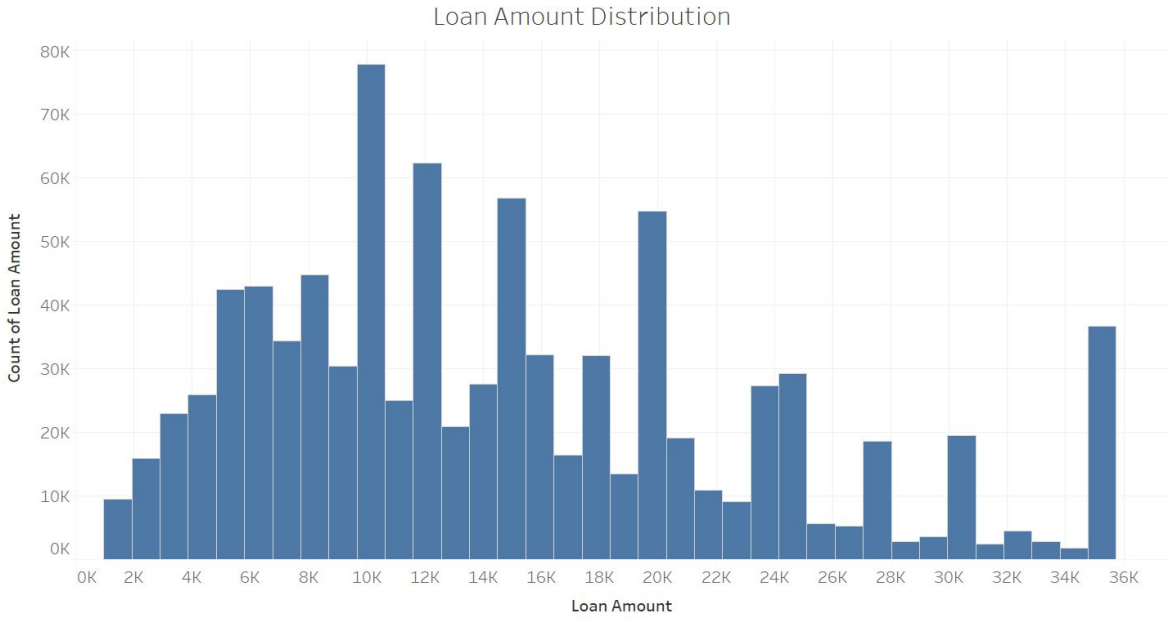

Fig. 11 Loan Amount Distribution

Fig. 11 is the distribution of loan amount. Similarly, it looks like a skewed Gaussian distribution. And the most loan amount lies between $6k-$9k.

3.2. Specific Visualization: Each Feature v.s. Interest Rate

This section is to identify the fators(i.e., features) that affect the loan interest rates.

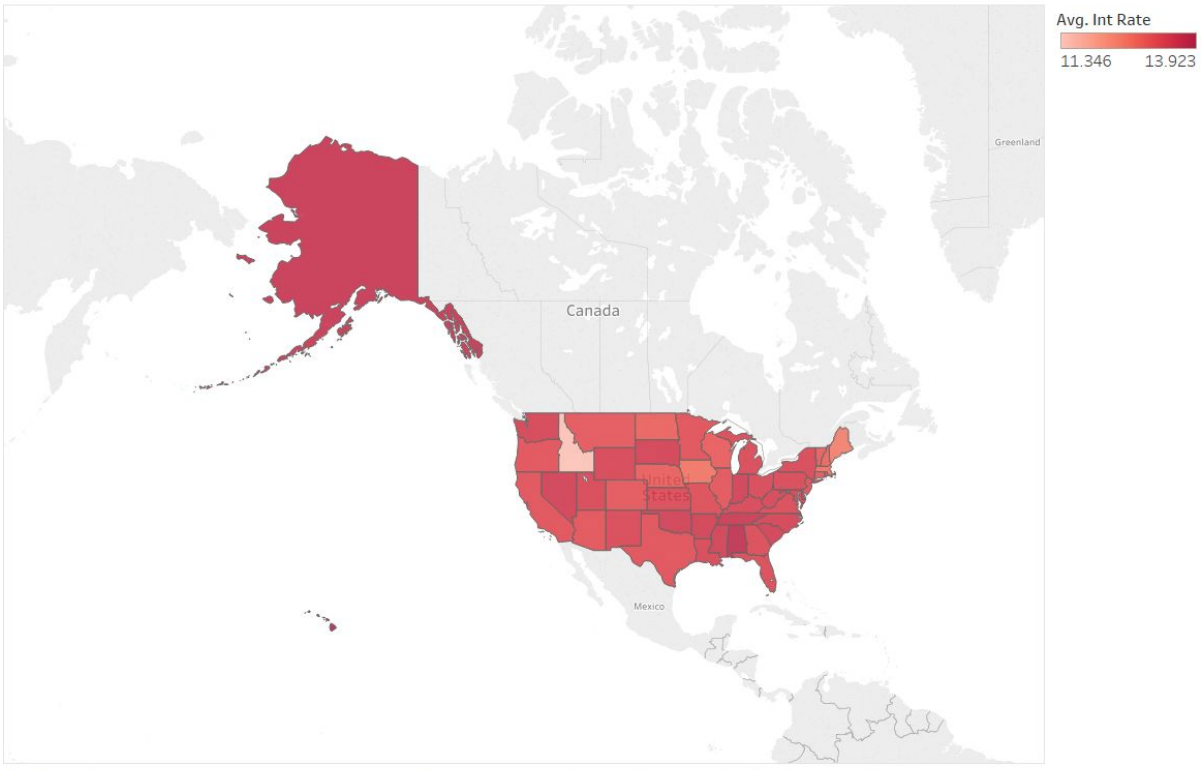

Fig. 12 Interest Rate by State

Figure 12 shows the average interest rate in terms of geological locations. The degree of interest rate for every state is reflected by the color. Darker color indicates a higher interest rate. It can be seen the display is interactive, it shows the exact interest rate once the cursor is on the display.
https://public.tableau.com/profile/rui.ma4337#!/vizhome/intr-state/intr-state

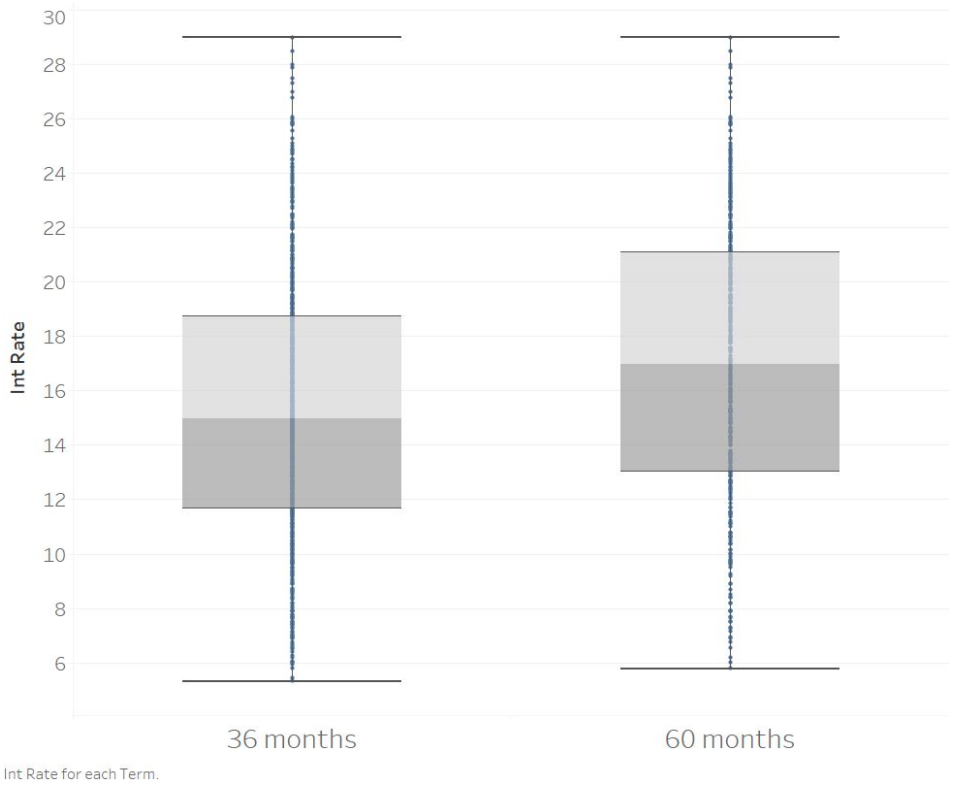

Fig. 13 Interest Rate by Pay-off Term

Figure 13 shows interest rates vary with the loan pay-off term. Longer pay-off terms usually indicates higher interest rate. It can be seen the median interest rate for 36 months pay-off term is around 15%, whereas the 60 months pay-off term is around 17%.

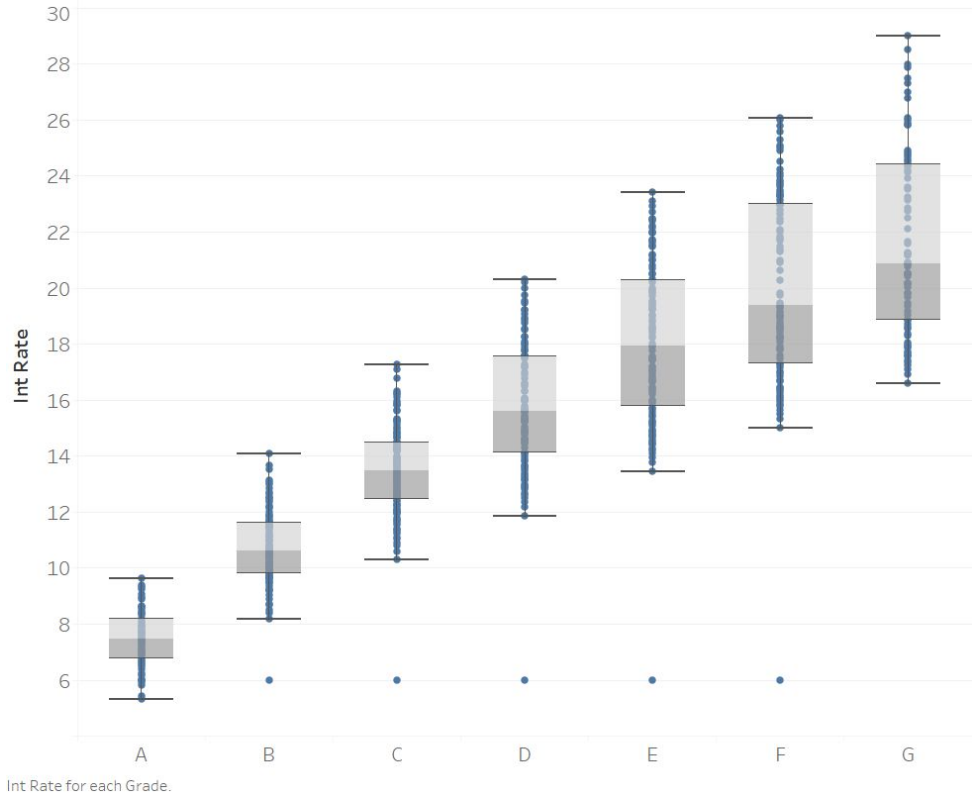

Fig. 14 Interest Rate by Loan Grade

Figure 14 shows interest rate is proportional to borrower's credibility. Loan grade is an internal credibility rating for every borrower, which can be considered as credit score in common sense. It can be seen from the picture that highest credibility A has an average interest rate of 21%, whereas lowest credibility G has an average interest rate of 7%.

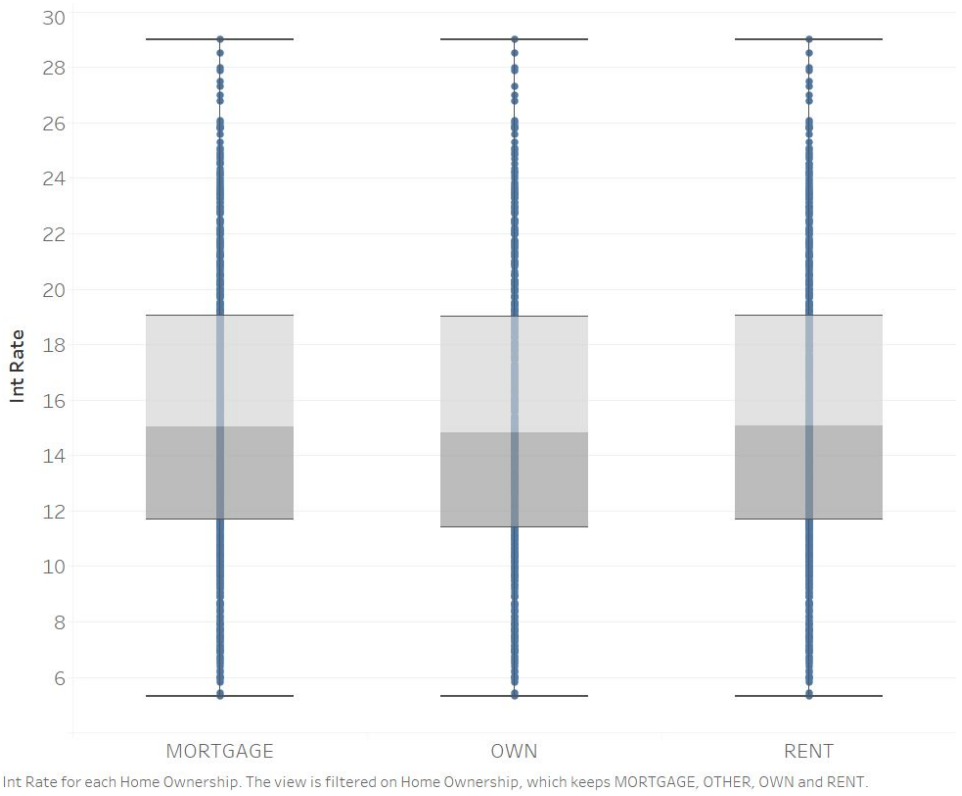

Fig. 15 Interest Rate by Home Ownership (Minor Effect)

Figure 15 shows there is not an obvious relationship between the types of house borrowers' posses and interest rate. In common sense, borrowers on mortgage usually indicate they are in debt or they are already in loan, thus have a higher interest rate. However, it can be seen the median of average interest rate is around 15% for all three types of house possessions.

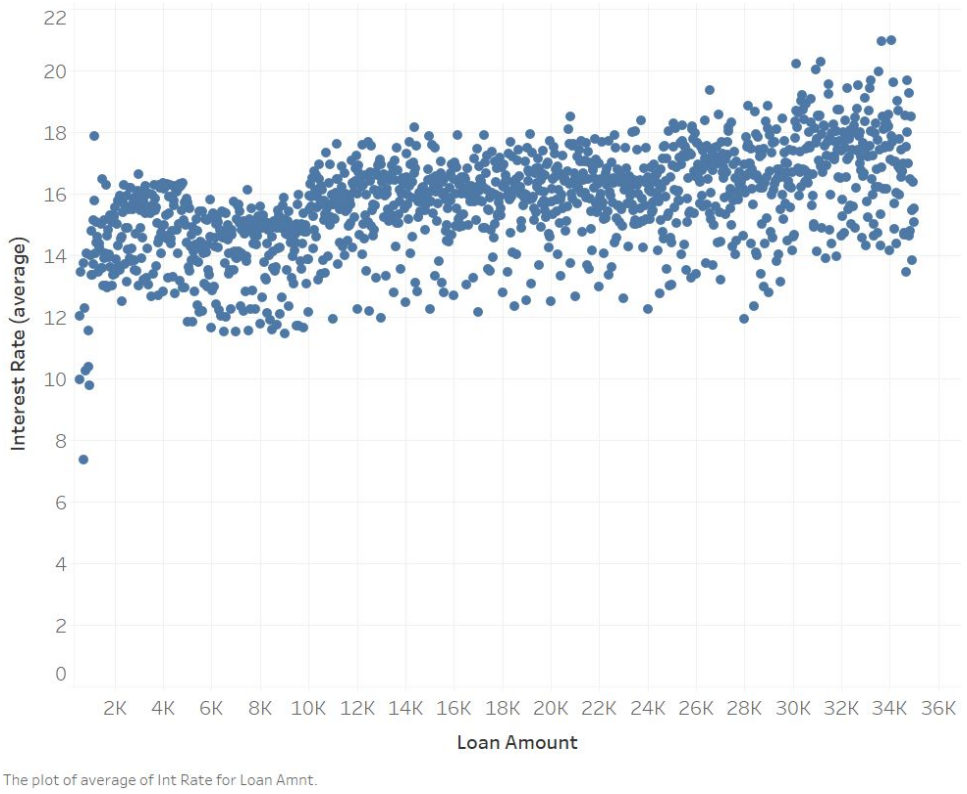

Fig. 16 Interest Rate v.s. Loan Amount

Figure 16 shows interests rate rises as loan amount increases. Interest rates for loan amount 2k lie between 13% and 16%.interest rates for loan amount 36k is between 15% to 18% range.

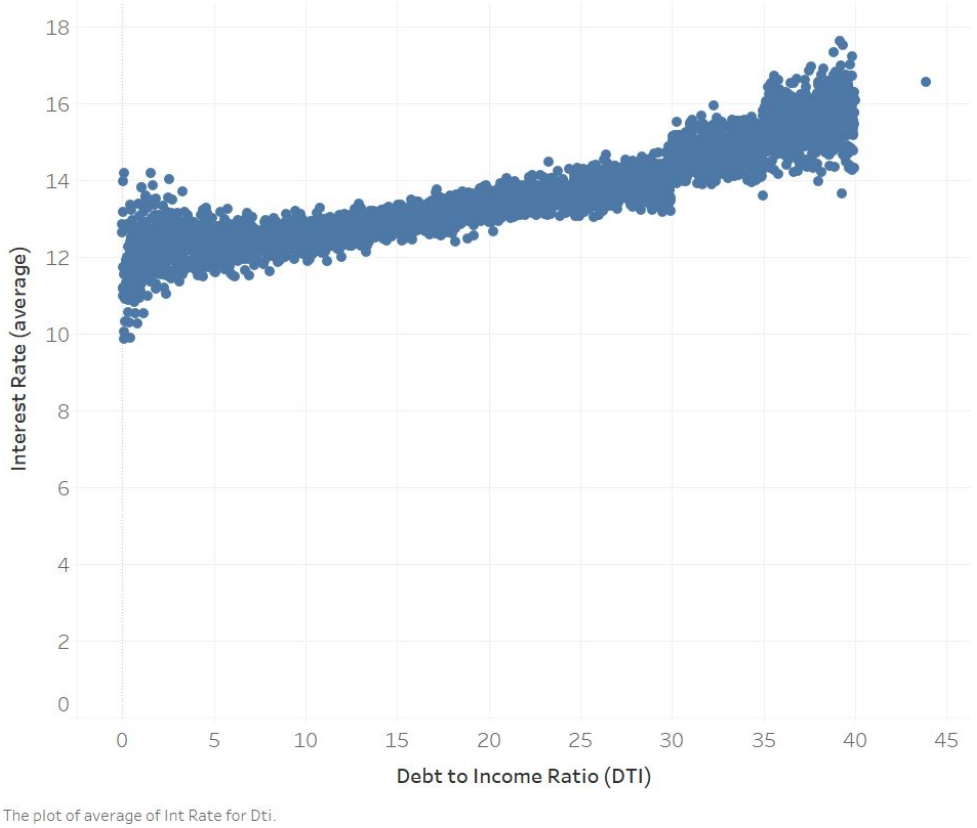

Fig. 17 Interest Rate v.s. Debt to Income Ratio (DTI)

Figure 17 shows higher interest rates apply to higher debt to income ratio. Higher debt to income ratio indicates the borrowers spend more than they earn in a period of time. Interest rates for borrowers that have a DTI of 0 is around 10% to 14%. Borrowers with a DTI of 40 have a interest rate around 14% to 18%.

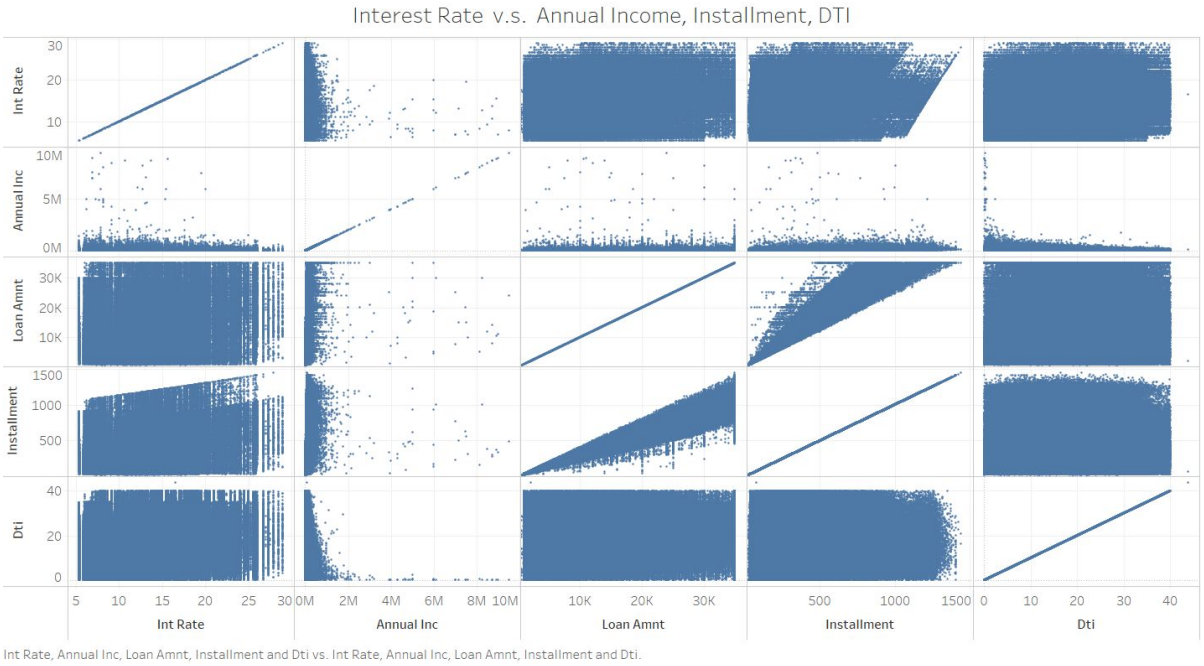

Fig. 18 Scatter plot: interest rate v.s. annual income v.s. loan amount v.s. installment v.s. DTI

Fig. 18 displays scatter plots among the 5 numeric features: interest rate, loan amount, installment, and dti. It is very difficult to observe any correlations, because each feature varies in pretty large scale.

In order to study the underlying correlations among these numeric features, pandas' built-in corr() function is used to generate a correlation matrix. Then Tableau is used to generate the correlation heatmap as shown in Fig. 19.

Correlation Heatmap

	Int Rate	Annual Inc	Loan Amnt	Installment	Dti
int_rate	1.000	-0.073	0.145	0.133	0.164
annual_inc	-0.073	1.000	0.333	0.326	-0.177
loan_amnt	0.145	0.333	1.000	0.945	0.044
installment	0.133	0.326	0.945	1.000	0.030
dti	0.164	-0.177	0.044	0.030	1.000

Fig. 19 Correlation Heatmap

Fig. 19 displays the Pearson correlation coefficient among the 5 numeric features, ranging from -1 to +1, where blue indicates positive correlation between two features and yellow indicates negative correlation; the deeper the color is, the higher the correlation. it can seen that the loan amount is highly correlated with installment (monthly payment), with a correlation coefficient of 0.945, which means that the higher the loan amount is, the larger the monthly payment will be. As for the interest rate, it can be seen it is not highly correlated with the other 4 numeric features.

3.3 Random Forest Model

Random Forest Model is created to identify which features carry significant effect on loan interest rate. However, in terms of analyzing feature importance,correlation score cannot be simply used as the importance rank, and correlation scores do not include categorical features.

To analyze the feature importance, label encoder is used to transfer each categorical feature (grade, term, states, verification status, home ownership) into numerical, then a random forest regressor is trained using Python's scikit-learn package. Finally model's built-in attribute "feature_importances_ " is used to get the entropy scores for each features, as shown in Fig. 20.

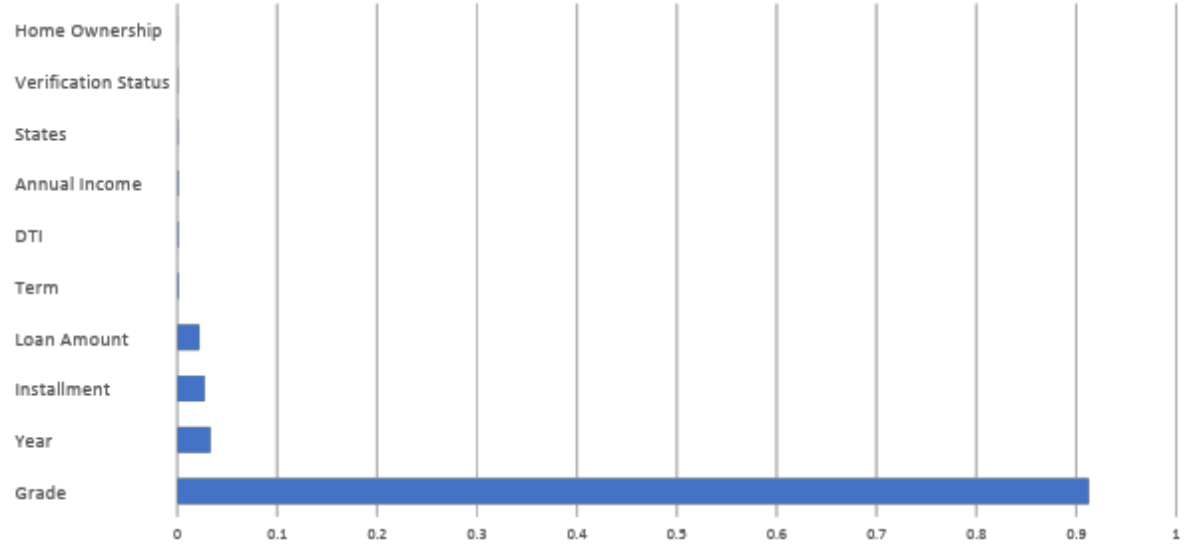

Fig. 20 Feature Importance: Random Forest Regressor

Fig. 20 displays each feature's contribution to loan interest rate. It can be seen the grade (credit score) has dominant effect on the interest rate, with a contribution of over 90%. The other three lesser important features are the year issued, installment, and loan amount.

4. Conclusion

From the visualization results and analysis of random forest model, it can be seen there are **four** major factors that affect the loan interest, which are: Loan Grades (91.26%), Loan Issued Year (3.32%), Installment (2.73%), Loan Amount (2.20%). Loan Grades feautre has the dominant influence towards the loan interest.

5. Division of the Work

Rui Ma: Cleaned the raw dataset; generated the scatter plots and correlation heatmap; built the random forest model.
Yan Liu: Analyzed and selected features; generated the bar plots, bubble plot and map plot; drew conclusions from plots.
Yao Jin: Generated box plots for numeric features; draw conclusions from generated graphs.
The presentation materials and final project report were finished by all team members.

References

[1] https://en.wikipedia.org/wiki/Lending_Club
[2] https://www.pinterest.co.uk/pin/510666045221085593/
[3] https://brightplanet.com/2016/04/data-visualization-student-loan-default-rates-by-institution/
[4] https://www.orchardplatform.com/blog/orchards-top-data-visualizations-of-2014/
[5] https://nycdatascience.com/blog/student-works/data-visualization-lending-club-issued-loans/
[6] https://www.kaggle.com/rgupta09/lending-club-loan-data-visualization