Exploratory Data Analysis

Group 21

Daksh Varshneya(IMT2012012)
Praseeda Kalkur(PH2016008)
Ujjwal Jain(IMT2013056)
Udbhav Vats(IMT2013055)

Data Cleaning.

1) As part of data cleaning tasks we have deleted the below columns as they dont have any relevance in our analysis.

- Desc – description provided by the user – very few people have given the purpose
- id – id assigned by lending club
- last_credit_pull_d – The most recent month Lending Club pulled credit for this loan. Not many rows with data
- member_id – unique id for a member
- policy_code - code given by lending club for the type of loan
- pymnt_plan – only 10 records have a payment plan
- title – Loan title provided by the borrower. Very few borrowers have provided title.
- zip_code – Redundant attribute, used state attribute for descriptive analysis.
- Url – url for Lending Club page with listing data. We need login to access the data, which is not required for analysis.

2) Date attributes – Attributes like issued_date, last_payment_date had values in the format Jan-2011 etc. We were unable to extract year from these attributes for analysis as the data was not in date format. We formatted these date columns for analysis.

3) Data Outliers – Attribute - 'pub_rec' which corresponds to the number of derogatory public records on the credit file has an outlier value of 86 which is found by plotting a box plot for that particular column.

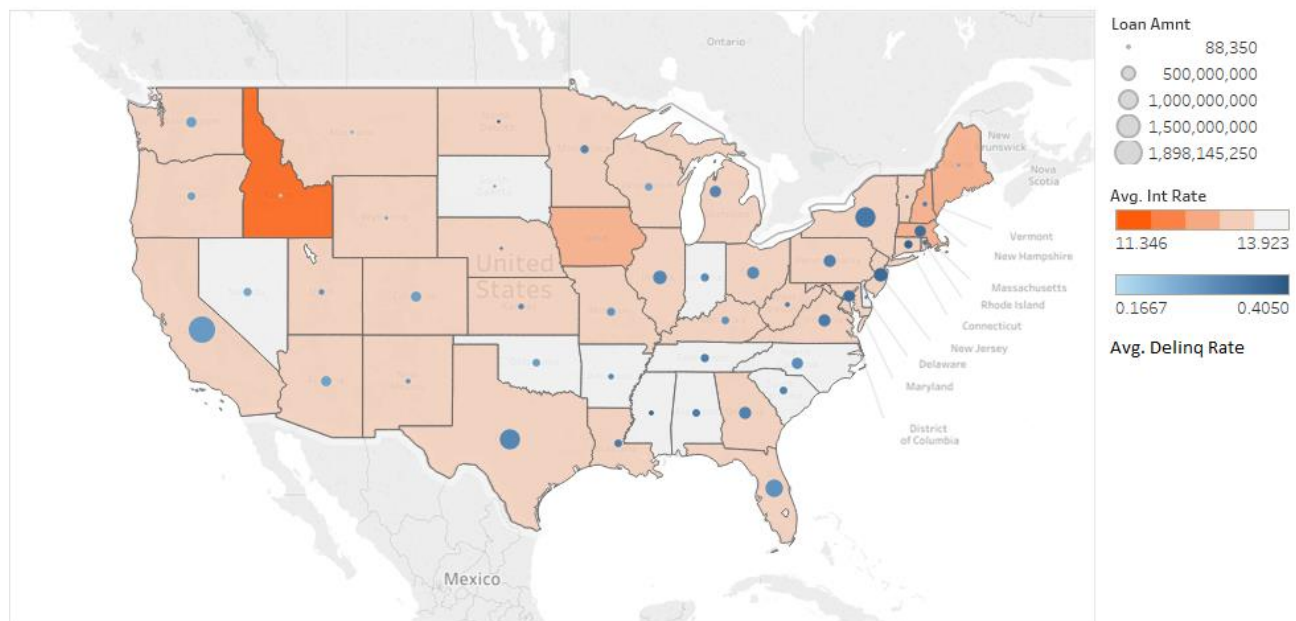Number of derogatory Public Record
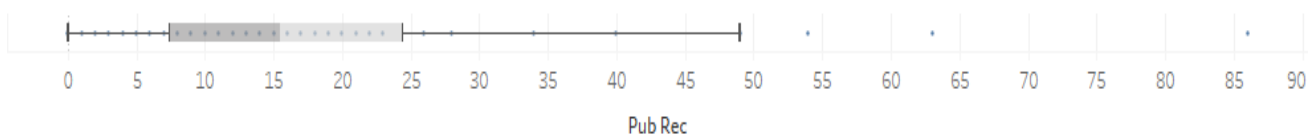
Pub Rec

Pub Rec.

Exploratory Analysis of Data

1) Geographical Analysis of Loans

Average Interest rate, Sum of Loan Amount and Average Delinq Rate By Geography


Map based on Longitude (generated) and Latitude (generated) and Latitude (generated). Details are shown for Addr State. For pane Latitude (generated): Color shows average of Int Rate. For pane Latitude (generated) (2): Color shows average of Delinq 2Yrs. Size shows sum of Loan Amnt.

This map shows that Lending club has maximum revenue from California State (shown by circle size) followed by New York and Texas. The company could focus on these states to increase their revenue. The colour of the state depicts the average interest rate charged by Lending Club grouped by states of America. The lowest rate charged is in Idaho State and not in California (leading state in terms of amount of money issued in Loan). It also shows the number of defaulters in US grouped by states with the help of intensity of the colour of the circle. Again, the number of defaulters is the highest in the state of California.

2) Purpose wise Distribution of Avg. Interest Rate v/s Avg. Loan Amount

Purpose wise average interest rate v/s average loan amount
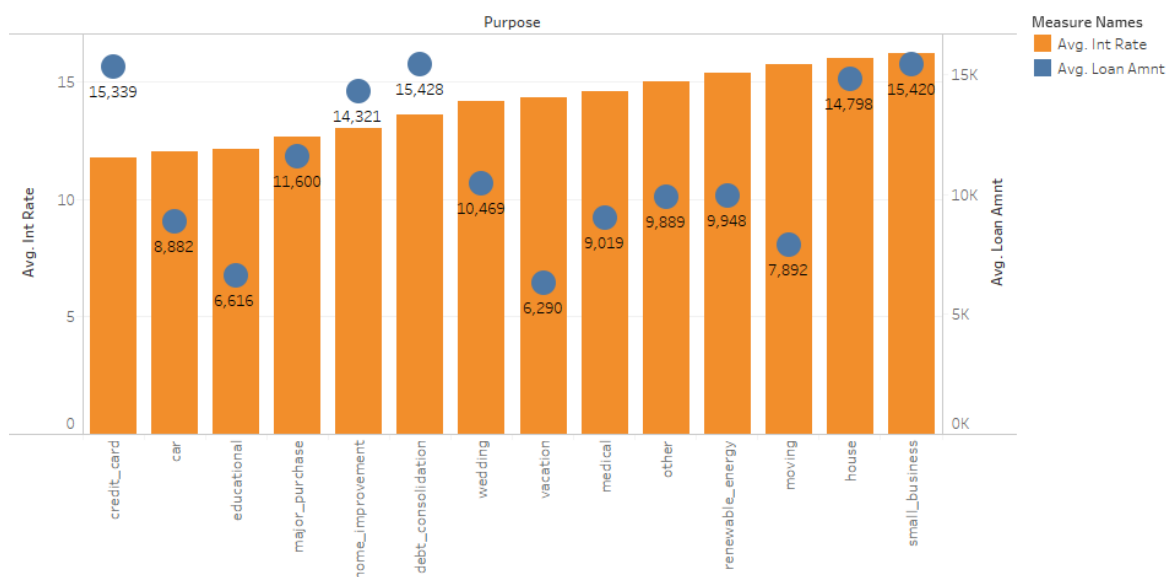
Avg. Int Rate and Avg. Loan Amnt for each Purpose. Color shows details about Avg. Int Rate and Avg. Loan Amnt. For pane Average of Loan Amnt: The marks are labeled by Avg. Loan Amnt.

This chart shows the purpose that the lender has specified for taking loan against average interest rate and average loan amount. It appears that interest rate has no correlation with the purpose of taking loan. But least amount of loan is taken for education and vacation.

3) LC Assigned Grade v/s Avg. Delinquency

Variation of Grade with interest rate and delinquency

Average of Int Rate for each Sub Grade broken down by Grade. Color shows average of Delinq 2Yrs.
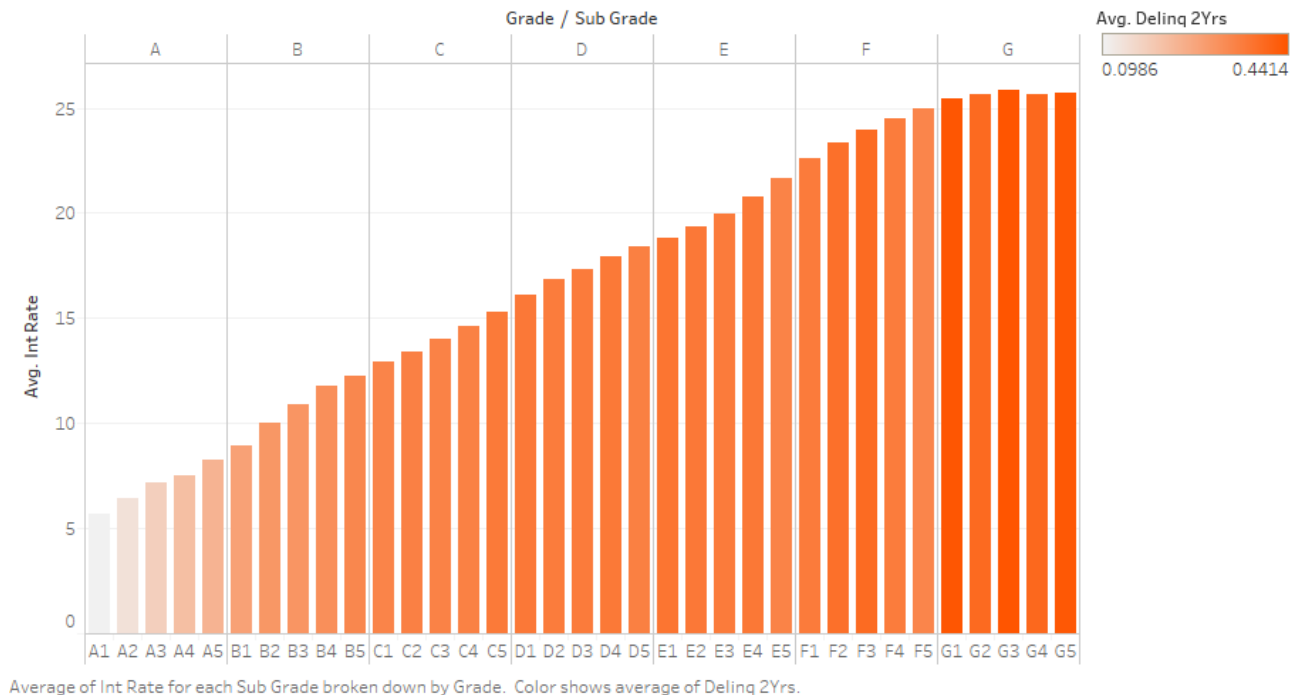
Lending club has an attribute grade(A to F) and within a grade, subgrade(1 to 5) assigned with each loan application. This chart shows that grades A has the least average interest rate and grades G has the highest interest rate. For all grades except G grade the interest rate increases from 1 to 5 i.e. A1 subgrade has lower interest rate and it keeps increasing to subgrade A5. Also the average amount of delinquency is the lowest for grade A.
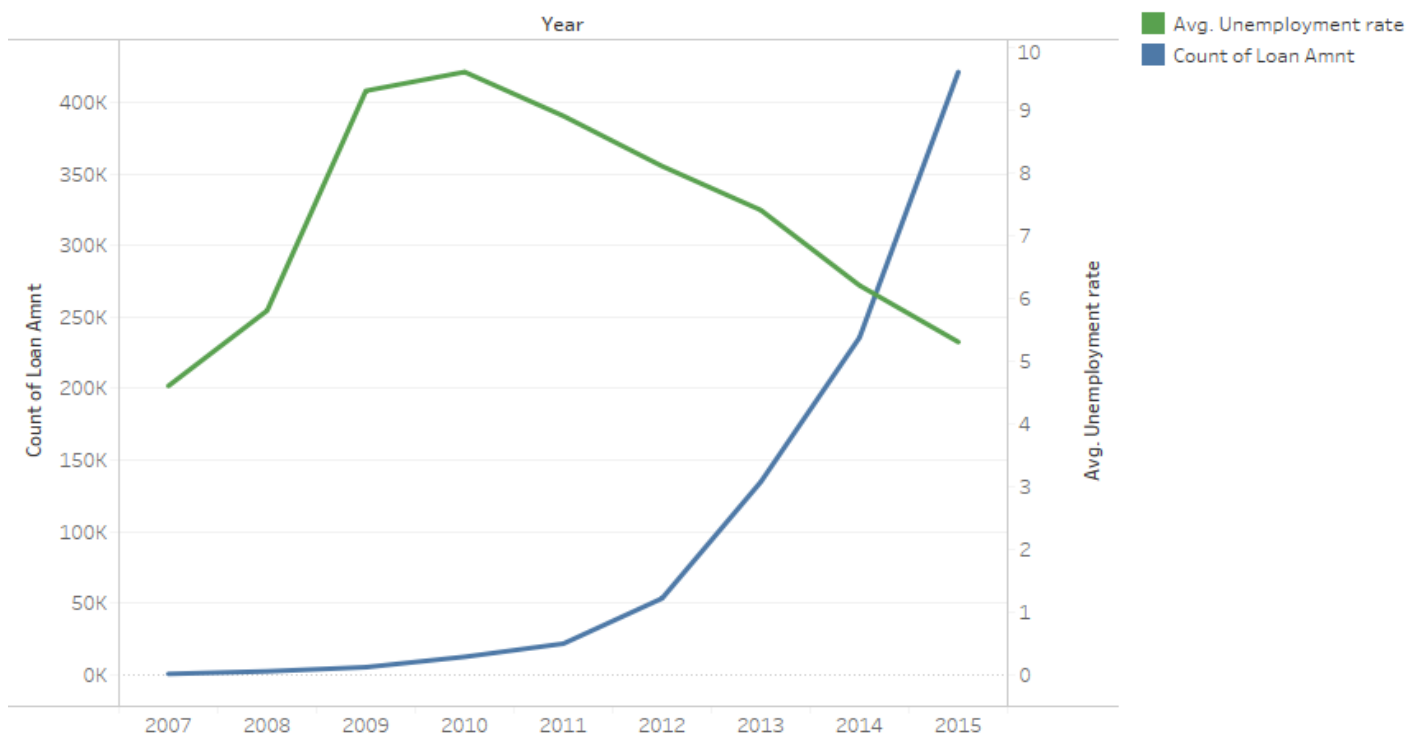
4) Debt-to-income ration v/s Unemployment Rate(Foreign data)

Year wise Average DTI ratio vs Unemployment Rate
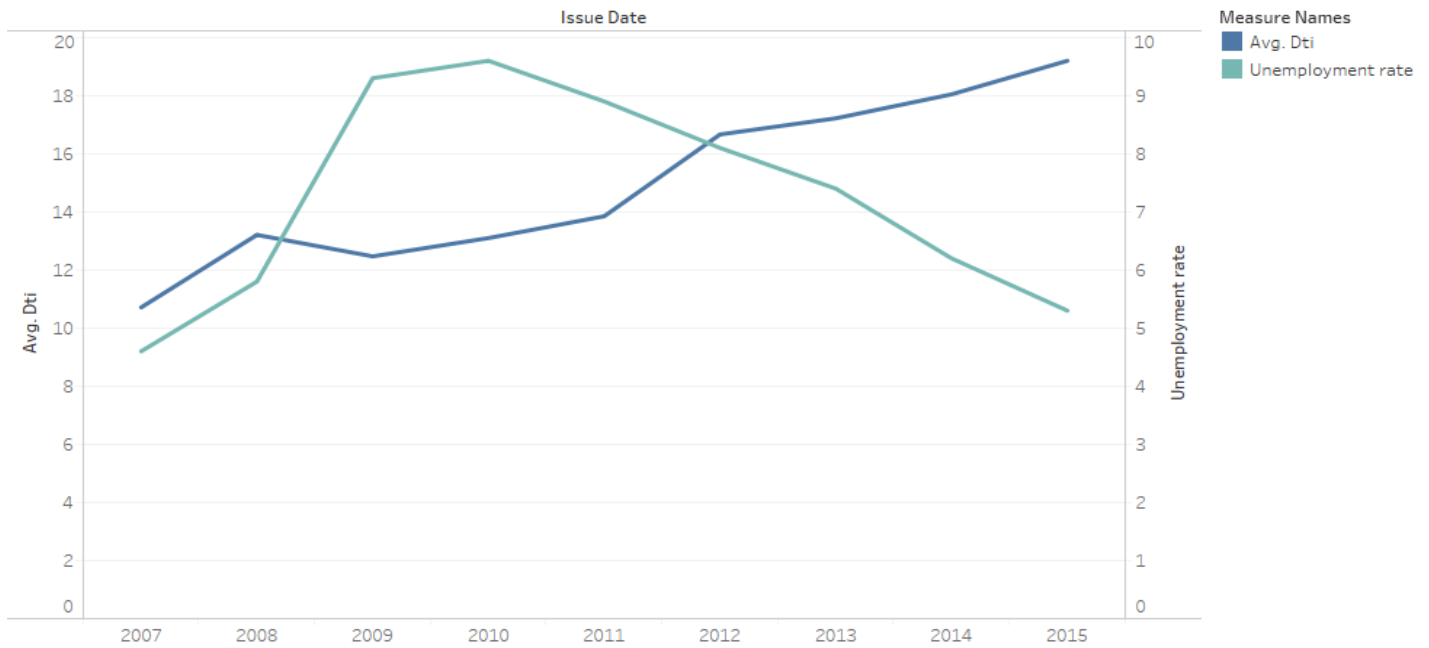
The trends of Avg. Dti and Unemployment rate for Issue Date Year. Color shows details about Avg. Dti and Unemployment rate.

## Count of Loan Amount vs Unemployment Rate by Year

The trends of count of Loan Amnt and Avg. Unemployment rate for Year (unemployment-rate-in-the-us) Year. Color shows details about count of Loan Amnt and Avg. Unemployment rate.

We have used the unemployment data in America between 2007 and 2015 and analysed the pattern of lending. We see that though unemployment has decreased i.e. though more people are getting jobs the numbers loans taken has not reduced. It has increased among the years. Even debt-to-income ratio has increased over the years.

5) Amount of Loan Sanctioned Profession wise

## Sum of Loan Amount for each Profession(minimum 2000)
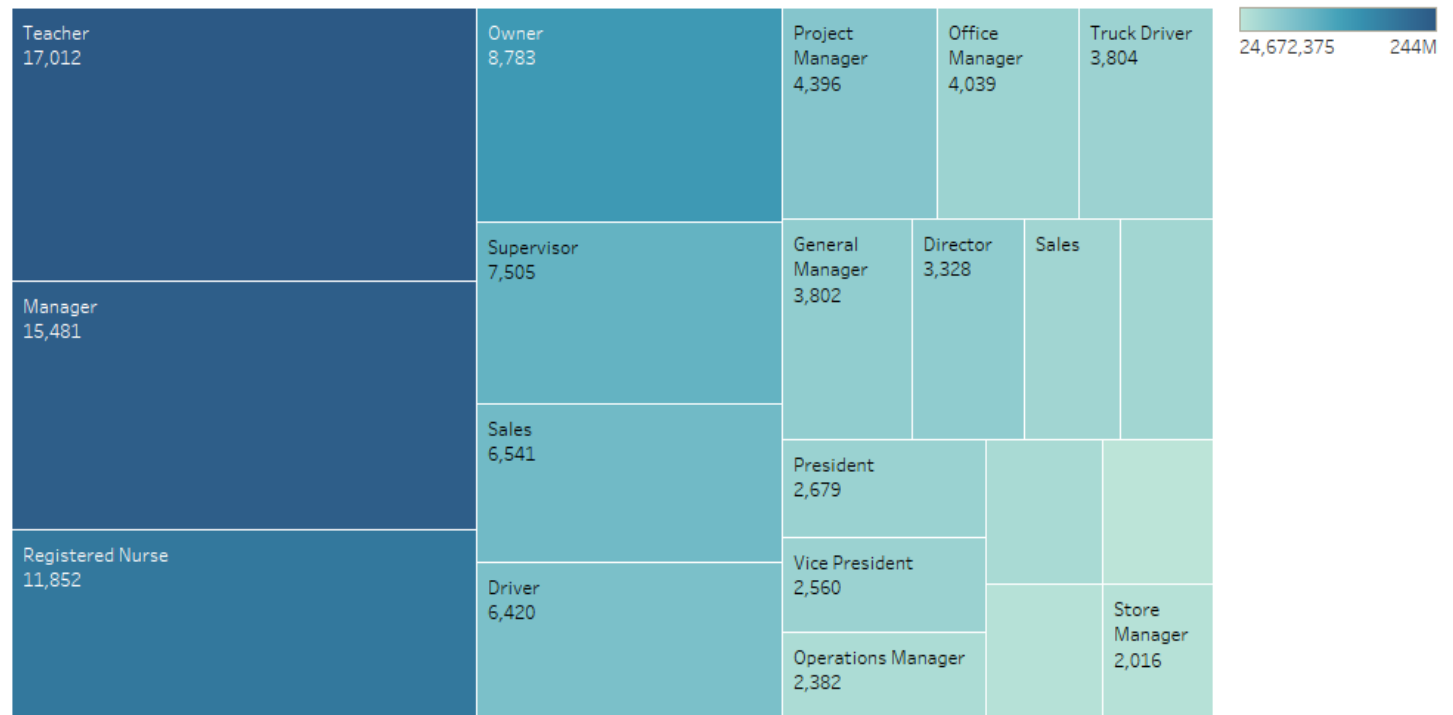
Emp Title and count of Emp Title. Color shows sum of Loan Amnt. Size shows count of Emp Title. The marks are labeled by Emp Title and count of Emp Title. The view is filtered on count of Emp Title, which includes values greater than or equal to 2,000.

The above treemap depicts the variation in number of loans and the total amount of money sanctioned by Lending Club, grouped by profession. The colour tone variation corresponds to the total amount of money sanctioned and the area of rectangles correspond to the number of loans sanctioned. We see that 'Teacher', 'Manager' and 'Registered Nurse' are the distinct highest loan seeking professions according to the data.

6) Defaulting Parameters v/s Loan Status

### Average Delinq, avg inquiries vs Loan Status
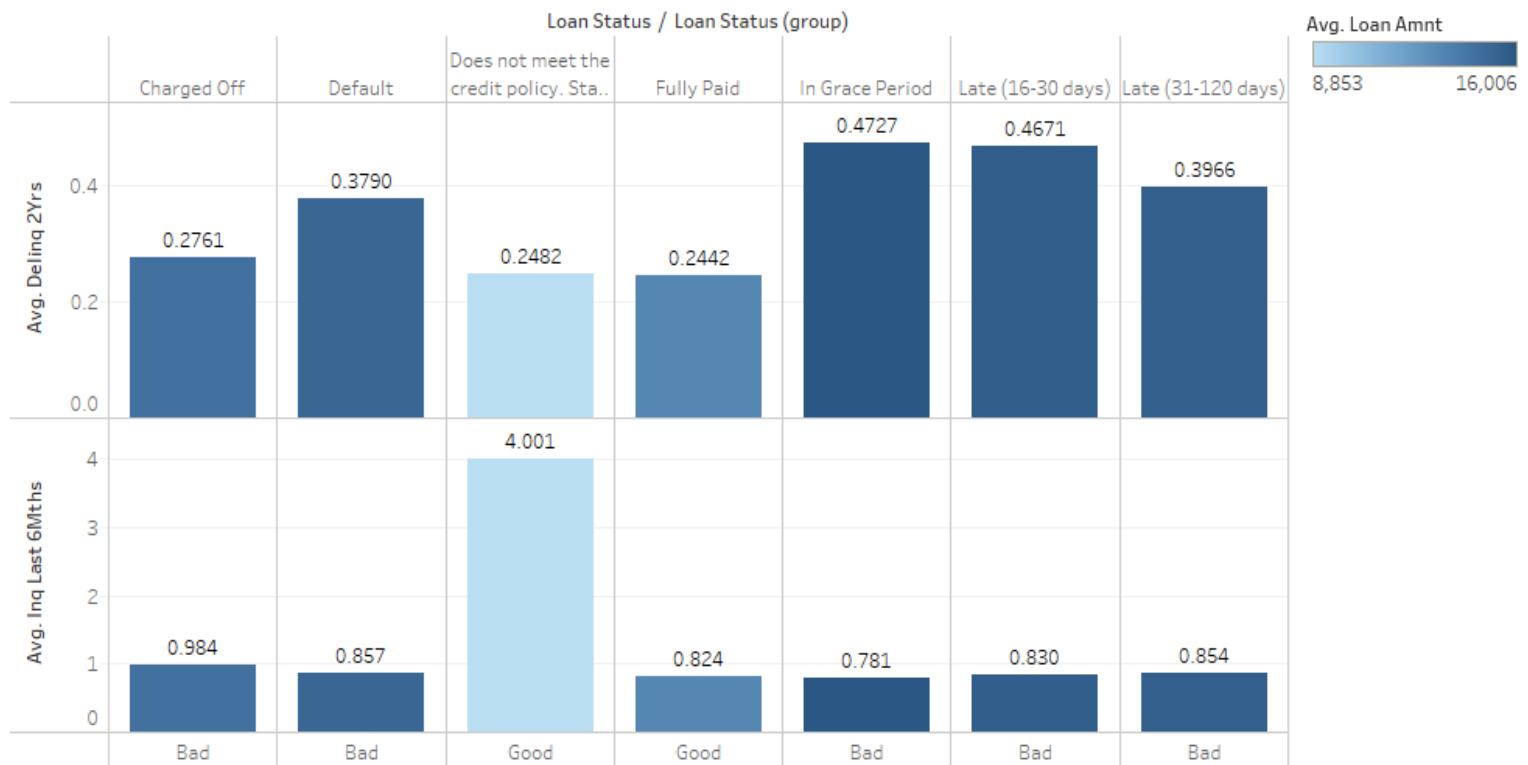
Average of Delinq 2Yrs and average of Inq Last 6Mths for each Loan Status (group) broken down by Loan Status. Color shows average of Loan Amnt. For pane Average of Delinq 2Yrs: The marks are labeled by average of Delinq 2Yrs. For pane Average of Inq Last 6Mths: The marks are labeled by average of Inq Last 6Mths. The view is filtered on Loan Status (group) and Loan Status. The Loan Status (group) filter keeps Bad and Good. The Loan Status filter excludes Does not meet the credit policy. Status:Charged Off.

Delinquency is a condition which arises when the borrower fails to adhere to the payment date. Lending Club provides data on the number of times a borrower has achieved such a condition in the past 2 years. We would expect the average number of delinquencies for defaulters to be high which is well depicted in the graph above. Also, the colour variation shown above depicts the average loan amount borrowed which is higher in defaulters' case than the non-defaulters.

Number of Inquiries correspond to the number of times a user's credit file has been opened by a business (lending or non-lending). This could be a bad or a good indicator and could affect the credit score of the borrower whichever way the lending agency wants to model it as. We find that the average number of enquiries is much higher for people who fully pay off the loan amounts sanctioned to them in comparison to the defaulters.
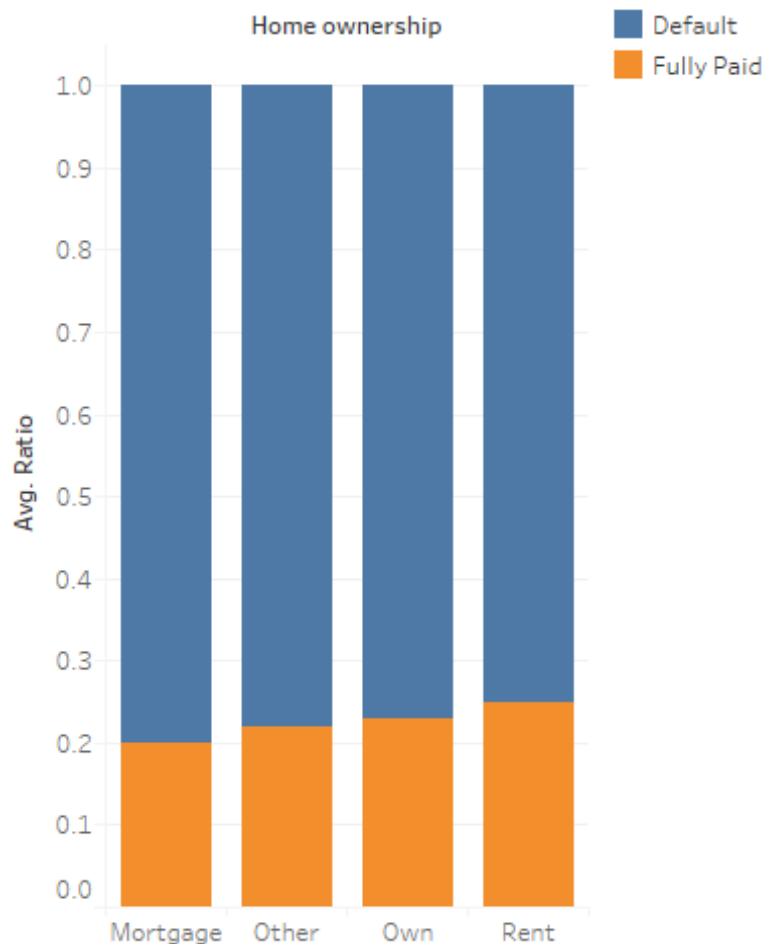
7) Impact of Home Ownership on Loan Status

Home Ownership/ Loan Status vs Number of Loan Transactions
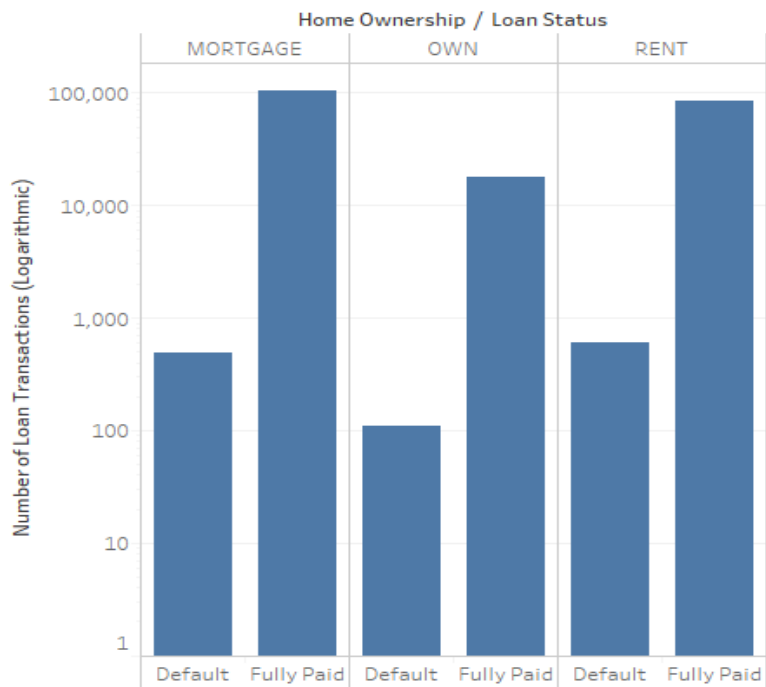
Count of Loan Amnt for each Loan Status broken down by Home Ownership. The data is filtered on Loan Status (group), which keeps Bad, Good and On Going. The view is filtered on Home Ownership and Loan Status. The Home Ownership filter keeps MORTGAGE, OWN and RENT. The Loan Status filter keeps Default and Fully Paid.

Holme Ownership, Loan Status vs Default Count

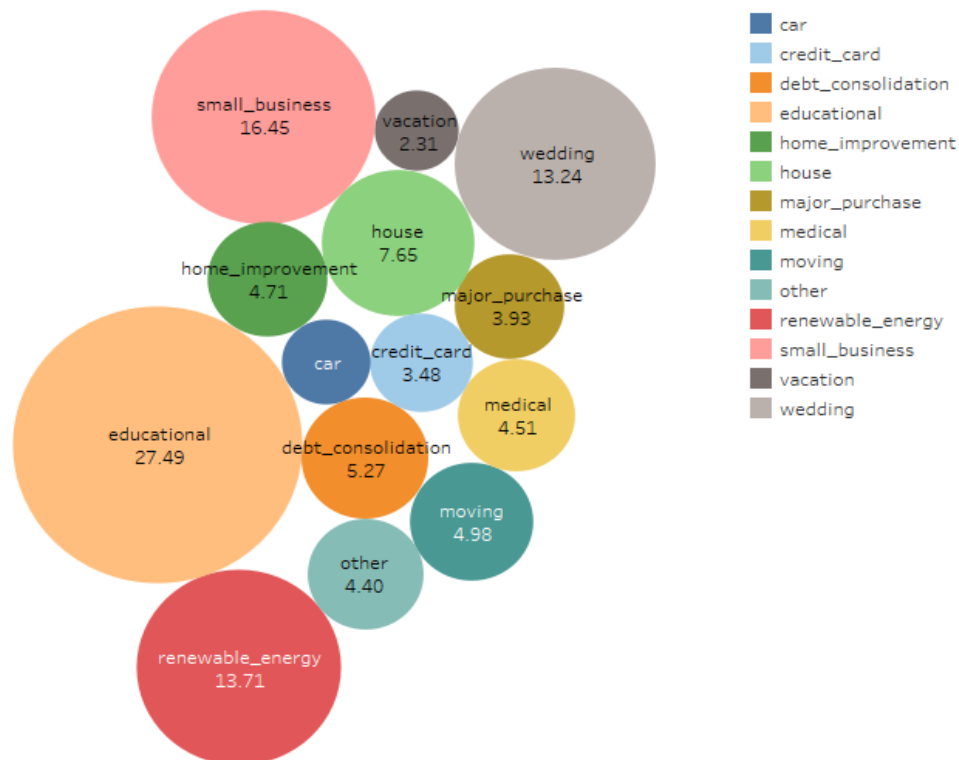Average of Ratio for each Home ownership. Color shows details about Loan status.

The above plots attempt to show the impact of home ownership on number of defaulting loan transactions. Intuitively, one would expect this attribute to have a large impact on 'loan status' but this isn't true from the data's perspective.

From the second plot, it is evident that the ratio of defaulting by types of ownership for past observations is identical which clearly suggests that there is no relationship between home ownership and loan status. This means that there is an equal chance of a person defaulting irrespective of its home ownership.

8) Average Collection Fee for each Loan Purpose

Average Collection Fee for each loan Purpose

small_business
16.45

vacation
2.31

wedding
13.24

house
7.65

home_improvement
4.71

major_purchase
3.93

car

credit_card
3.48

educational
27.49

debt_consolidation
5.27

medical
4.51

moving
4.98

other
4.40

renewable_energy
13.71

Legend:
- car
- credit_card
- debt_consolidation
- educational
- home_improvement
- house
- major_purchase
- medical
- moving
- other
- renewable_energy
- small_business
- vacation
- wedding

Purpose and average of Collection Recovery Fee. Color shows details about Purpose. Size shows average of Collection Recovery Fee. The marks are labeled by Purpose and average of Collection Recovery Fee.

There is a strong correlation between charge off rates and if the loan purpose is either educational or small business. These loans tend to be a lot riskier.

9) Interest rate vs Grade

Interest Rate vs grade
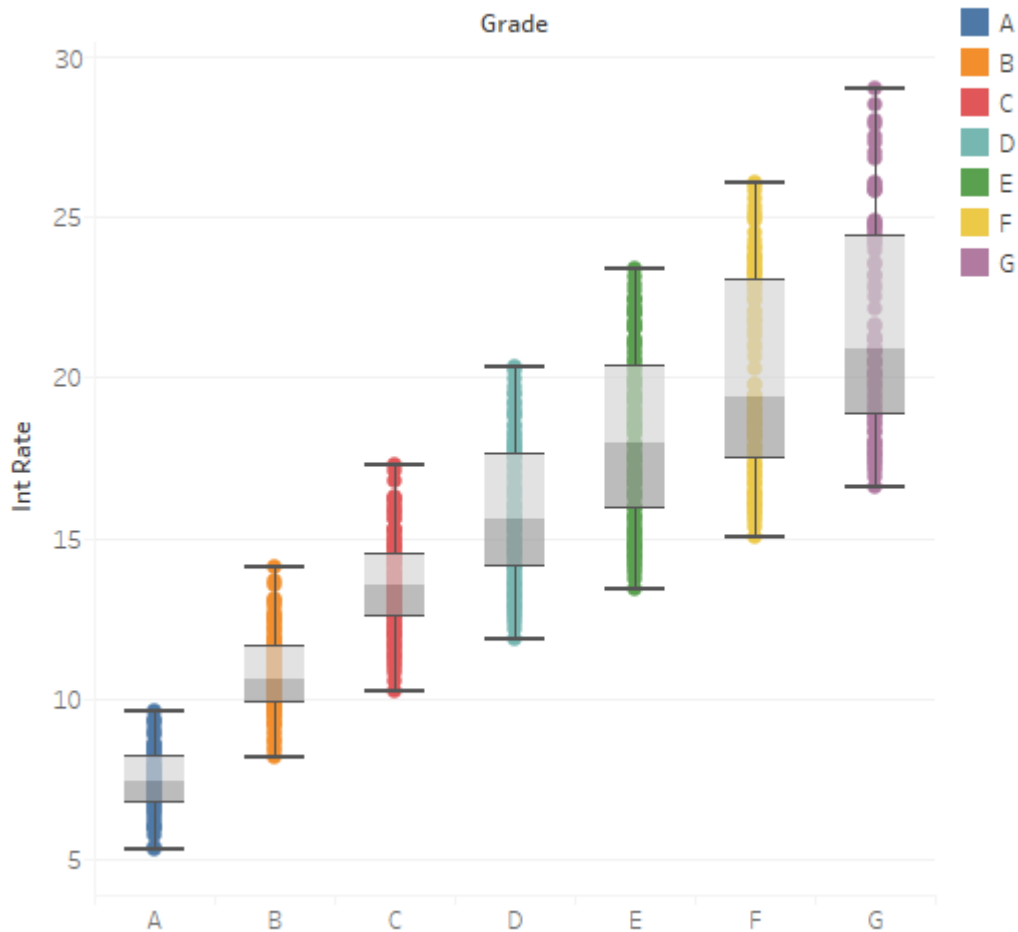
Int Rate for each Grade. Color shows details about Grade. The view is filtered on Exclusions (Grade,Int Rate), which keeps 620 members.

The above box plot shows that the interest rate is least for Lending Club grade A and maximum for grade G. Box plot also indicates that the 90% of interest rate of grade A loans is below 10% while 90% of the loans with grade G pay upto 28% interest rate.

10) Average income vs Average Loan amount per year

Avg income vs Avg loan amount by Year
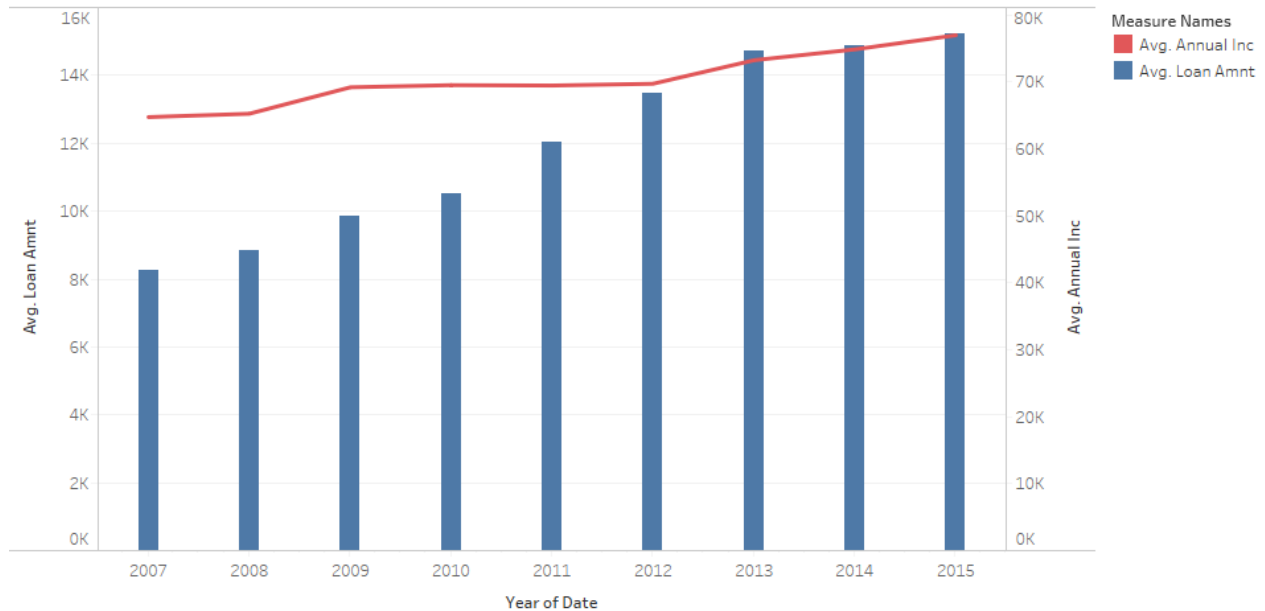
The trends of Avg. Loan Amnt and Avg. Annual Inc for Date (Data (statistic_id200838_average-household-income-in-the-united-states-1990-2015)) Year. Color shows details about Avg. Loan Amnt and Avg. Annual Inc.

The bar chart shows that there is an exponential increase in the average loan amount taken over the years as the income of individuals is increasing at the steady rate.

11) Home Ownership vs Average Interest Rate

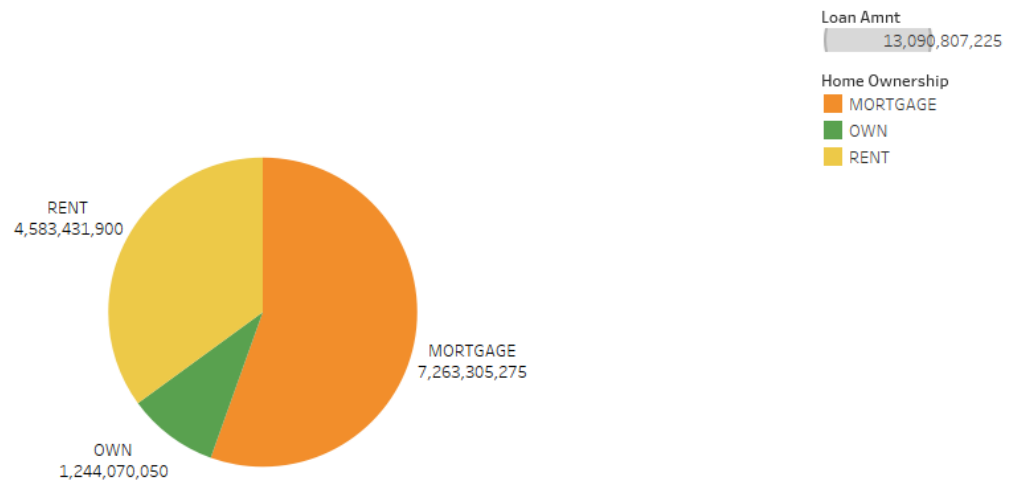### Home Ownership vs Avg. Interest Rate
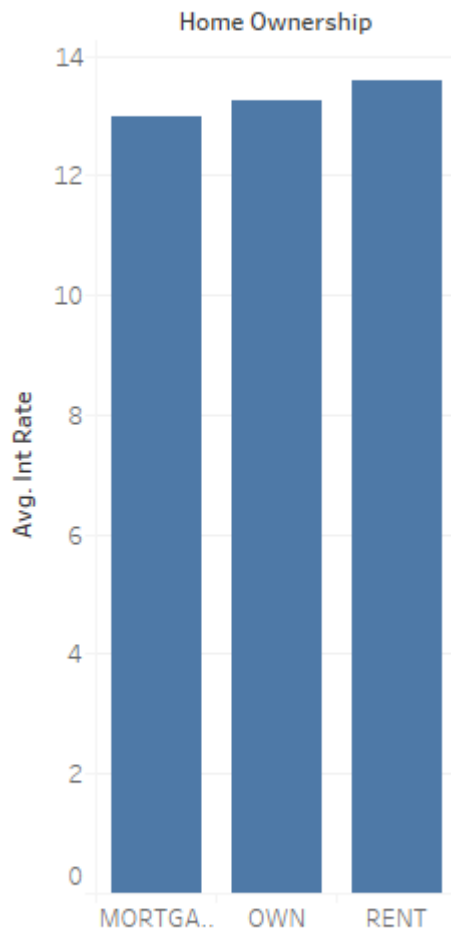
Average of Int Rate for each Home Ownership. The view is filtered on Home Ownership, which keeps MORTGAGE, OWN and RENT.

The above bar graph depicts that average interest rate has no correlation with home ownership status of an individual.

12) Distribution of Loan Amount according to Loan Ownership

Home Ownership vs Loan Amount1

Loan Amnt
13,090,807,225

Home Ownership
- MORTGAGE
- OWN
- RENT

RENT
4,583,431,900

MORTGAGE
7,263,305,275

OWN
1,244,070,050

Home Ownership and sum of Loan Amnt. Color shows details about Home Ownership. Size shows sum of Loan Amnt. The marks are labeled by
Home Ownership and sum of Loan Amnt. The view is filtered on Home Ownership, which excludes ANY, NONE and OTHER.

The above pie chart tells us that the maximum number of people who are taking loans have already mortgaged their houses. This tells us that taking loan is a vicious circle.