Background

LendingClub is an American peer-to-peer lending company. P2P loans are loans made by individuals and investors – as opposed to loans that come from a bank. People with extra funds offer to lend that money to borroweres in need of cash. A P2P service (such as a website) matches lenders and borrowers so that the process is relatively easy for all parties.

Objective

Loan default prediction is a common problem for lending companies. This analysis focusses on using the Lending Club dataset which is available on Kaggle. The objective is to make predictions about loan default and whether investors should lend to a customer or not based on their risk profile. Data is from 2007-2018 where most of the loans from that period have already been repaid or defaulted on.

The goal of this analysis is to identify credit-worthy customers that were not recognized by traditional credit scores, and predict the possibility of current loans becoming bad. This will help LendingClub minimize default risk and make higher profits in future.

1. Problem Statement

To develop a basic understanding of risk analytics how data is used to minimise the risk of losing money while lending to customers.

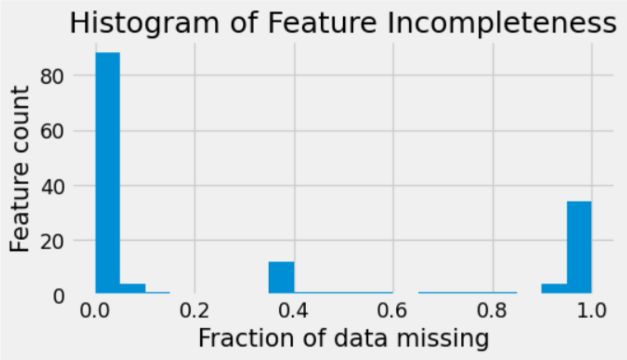
lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Defaulters cause large credit losses to the lenders. The customers labelled as 'charged-off' are the 'defaulters'.

2.1 Data Description

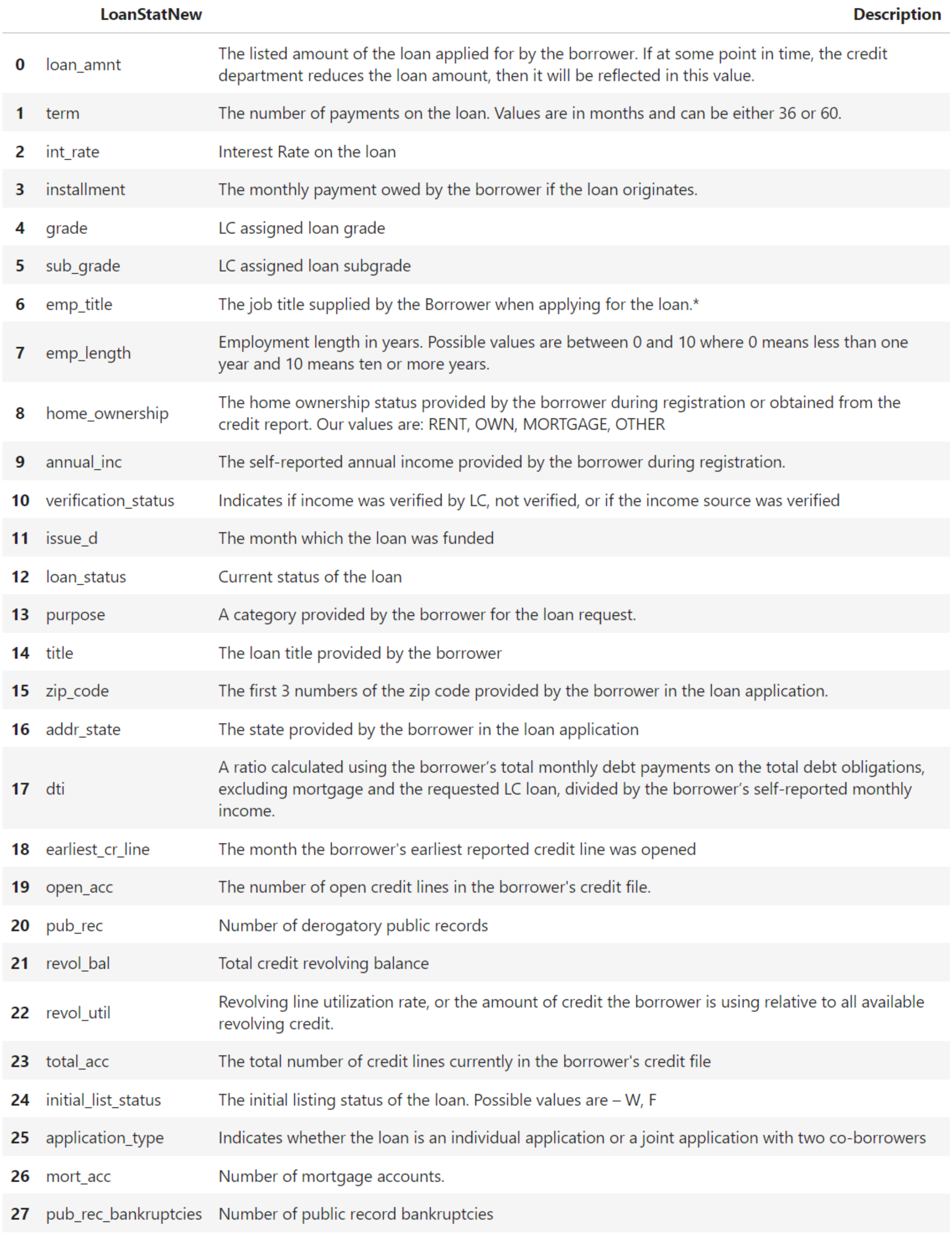
The dataset used in this project contains 2260701 rows and 151 columns of the lending Company from 2007 to 2018.

However, not all features are valuable and crucial in this analysis. Due to the missing fractions in the data there is a need to eliminate several aspects from the dataset, such as the borrower’s zip code, member ID etc.

Interest rate, Instalment, grade, subgrade, loan status, annual income, purpose etc. are the main features. This dataset is both qualitative and quantitative, which contains all discrete, continuous, ordinal, and nominal values.



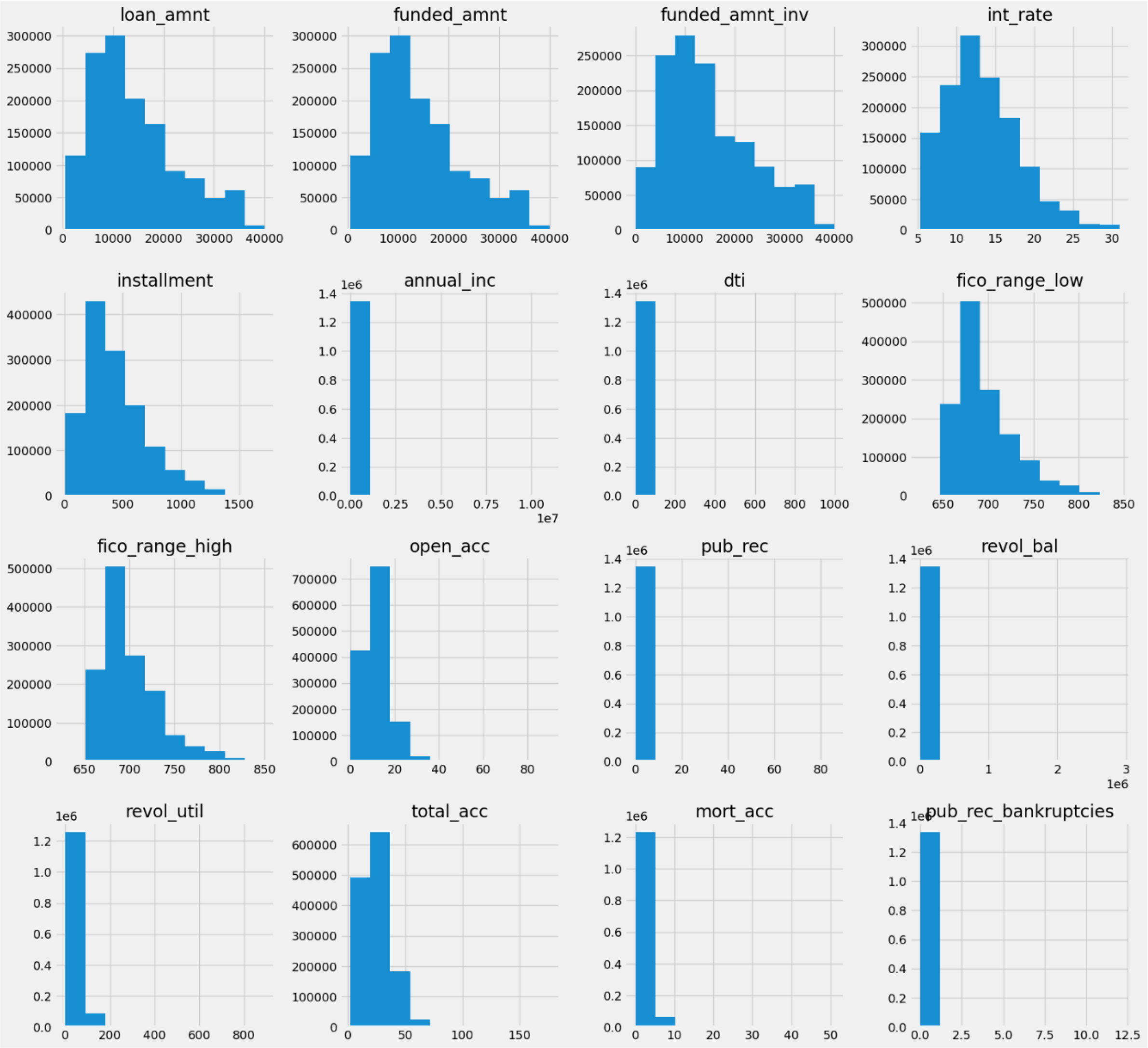
2.2 Data Dictionary



	LoanStatNew	Description
0	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
1	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
2	int_rate	Interest Rate on the loan
3	installment	The monthly payment owed by the borrower if the loan originates.
4	grade	LC assigned loan grade
5	sub_grade	LC assigned loan subgrade
6	emp_title	The job title supplied by the Borrower when applying for the loan.*
7	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
8	home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
9	annual_inc	The self-reported annual income provided by the borrower during registration.
10	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
11	issue_d	The month which the loan was funded
12	loan_status	Current status of the loan
13	purpose	A category provided by the borrower for the loan request.
14	title	The loan title provided by the borrower
15	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
16	addr_state	The state provided by the borrower in the loan application
17	dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
18	earliest_cr_line	The month the borrower's earliest reported credit line was opened
19	open_acc	The number of open credit lines in the borrower's credit file.
20	pub_rec	Number of derogatory public records
21	revol_bal	Total credit revolving balance
22	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
23	total_acc	The total number of credit lines currently in the borrower's credit file
24	initial_list_status	The initial listing status of the loan. Possible values are – W, F
25	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
26	mort_acc	Number of mortgage accounts.
27	pub_rec_bankruptcies	Number of public record bankruptcies

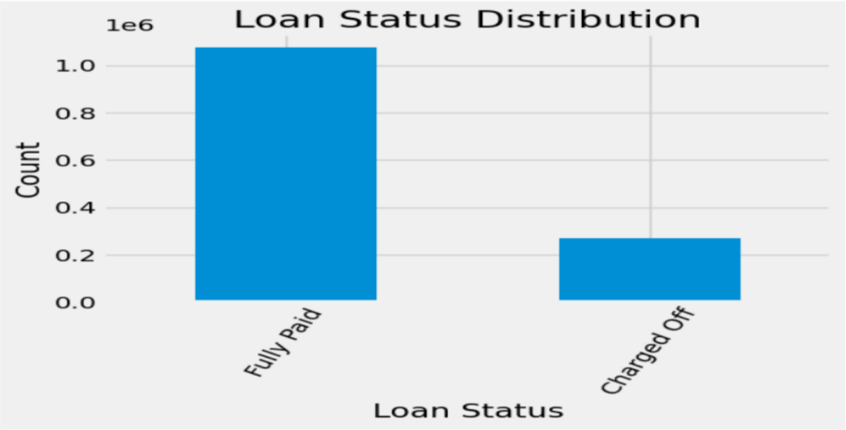
3 Methodology

This project mainly tackles the problem using visualization, categorization, and statistical techniques to enhance the interpretability of complex datasets and draw conclusions based on features that are correlated.



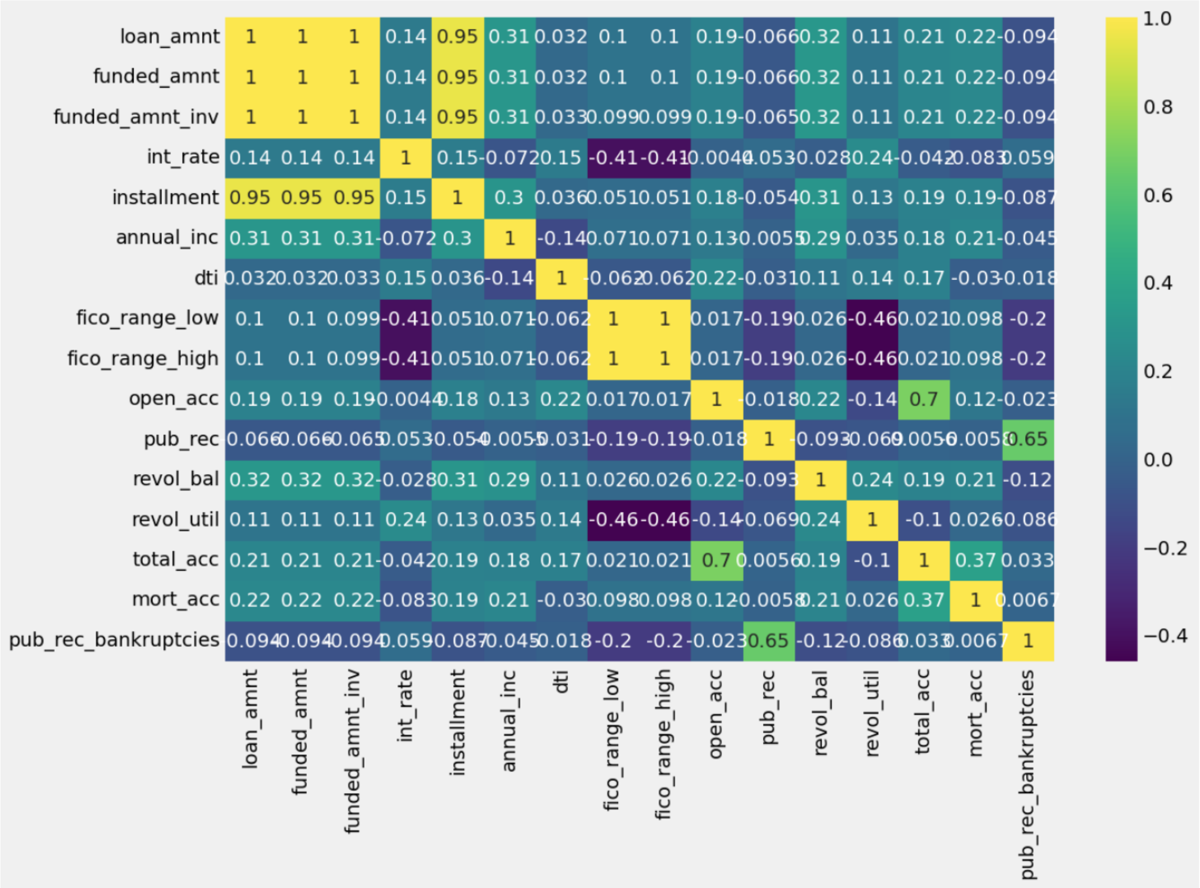
4 Target Variable

Since the aim is to identify the variables that influence the tendency of default, 'Loan Status' is our target variable. The graph below shows that around 22% of total loan counts have defaulted.



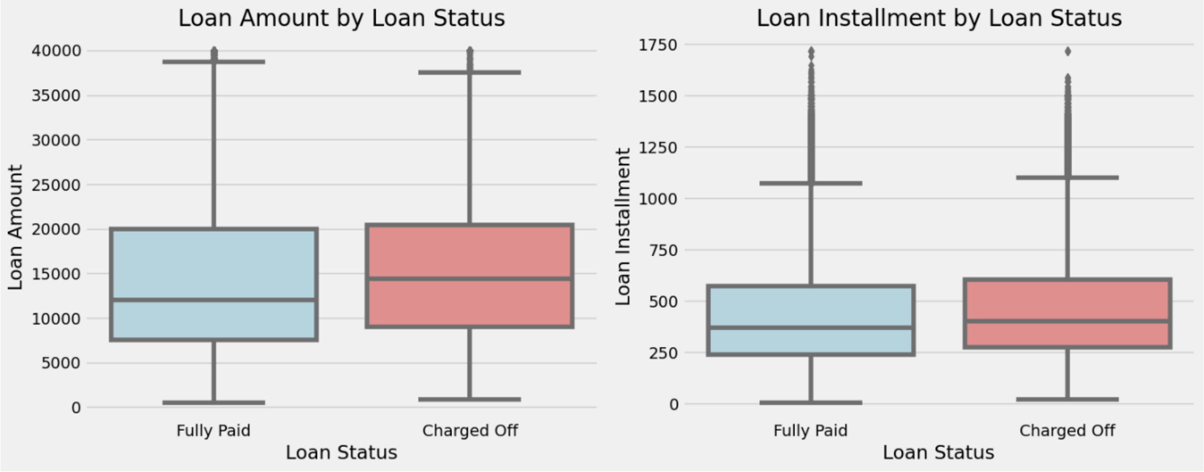
5 Correlation Matrix

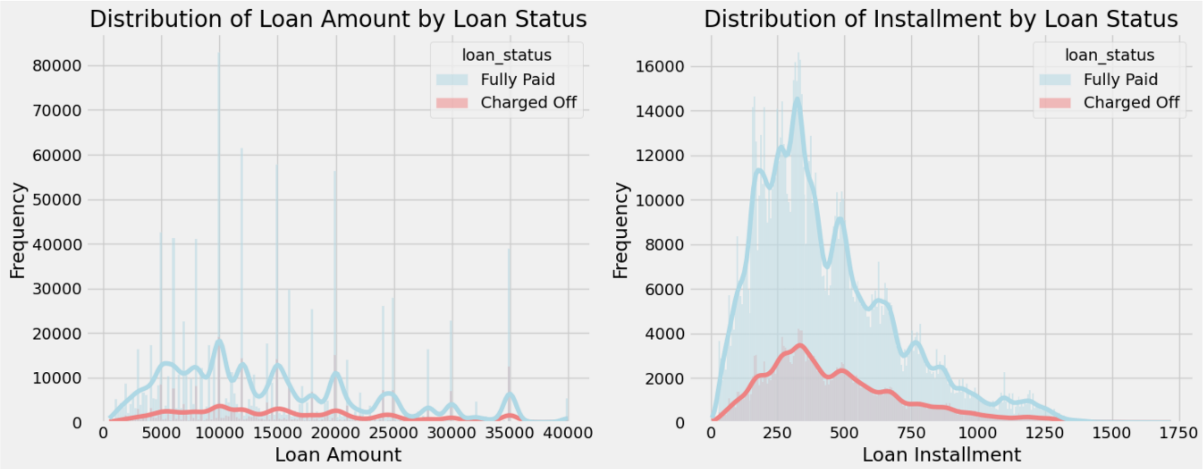
There is high correlation between loan amounts and instalment. Moderate correlation between loan amounts and annual income.



6 Loan amount and Instalment:

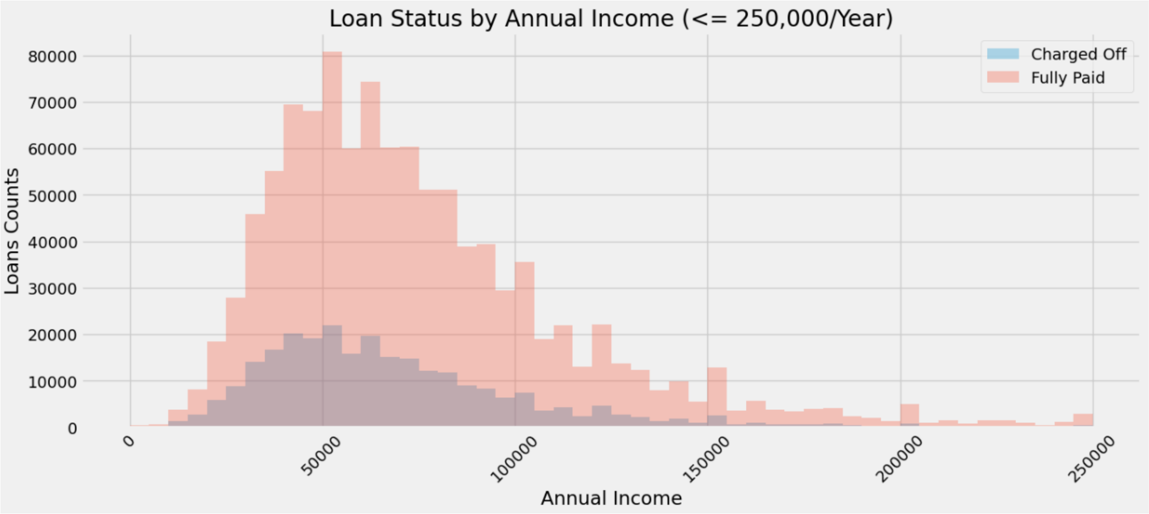
Count: 1,345,310; Mean: $14,420; Minimum loan amount: $500; Median: $12,000 and Maximum loan amount: $40,000





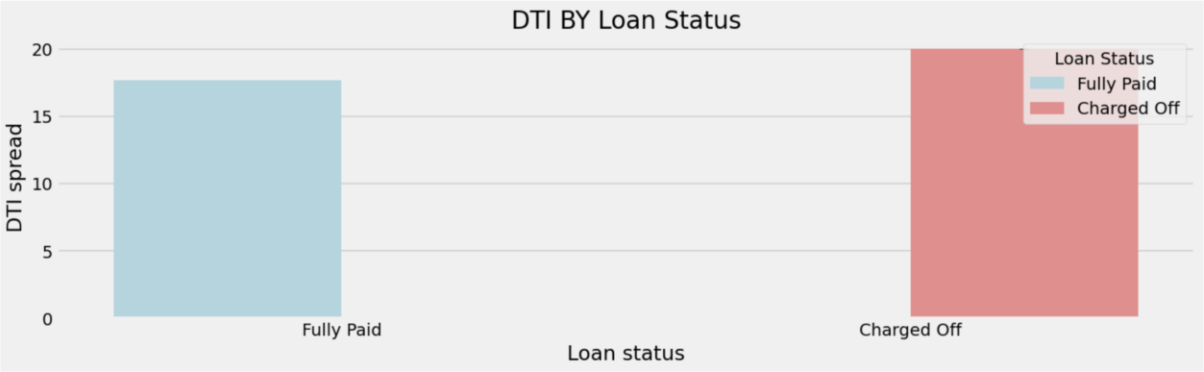
7 Annual Income

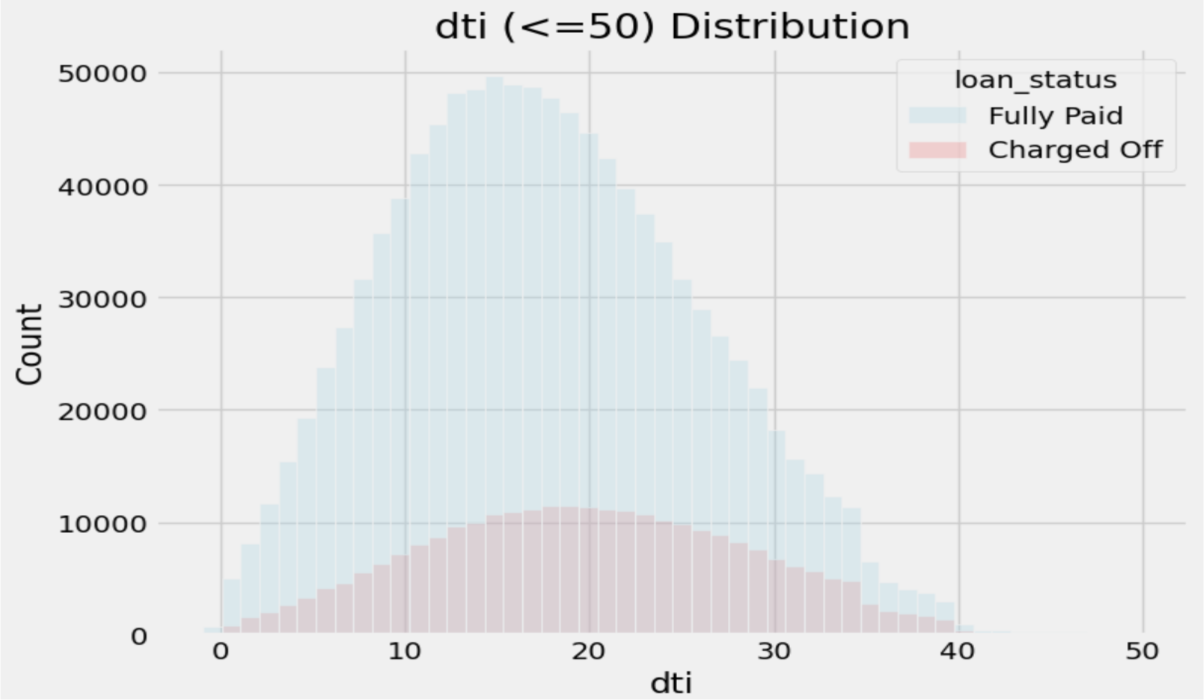
High interest rate loans are more likely to default, especially when the avg annual income is <70k.



8 Debt to Income Ratio analysis

Charged off tend to have higher debt-to-income ratios.

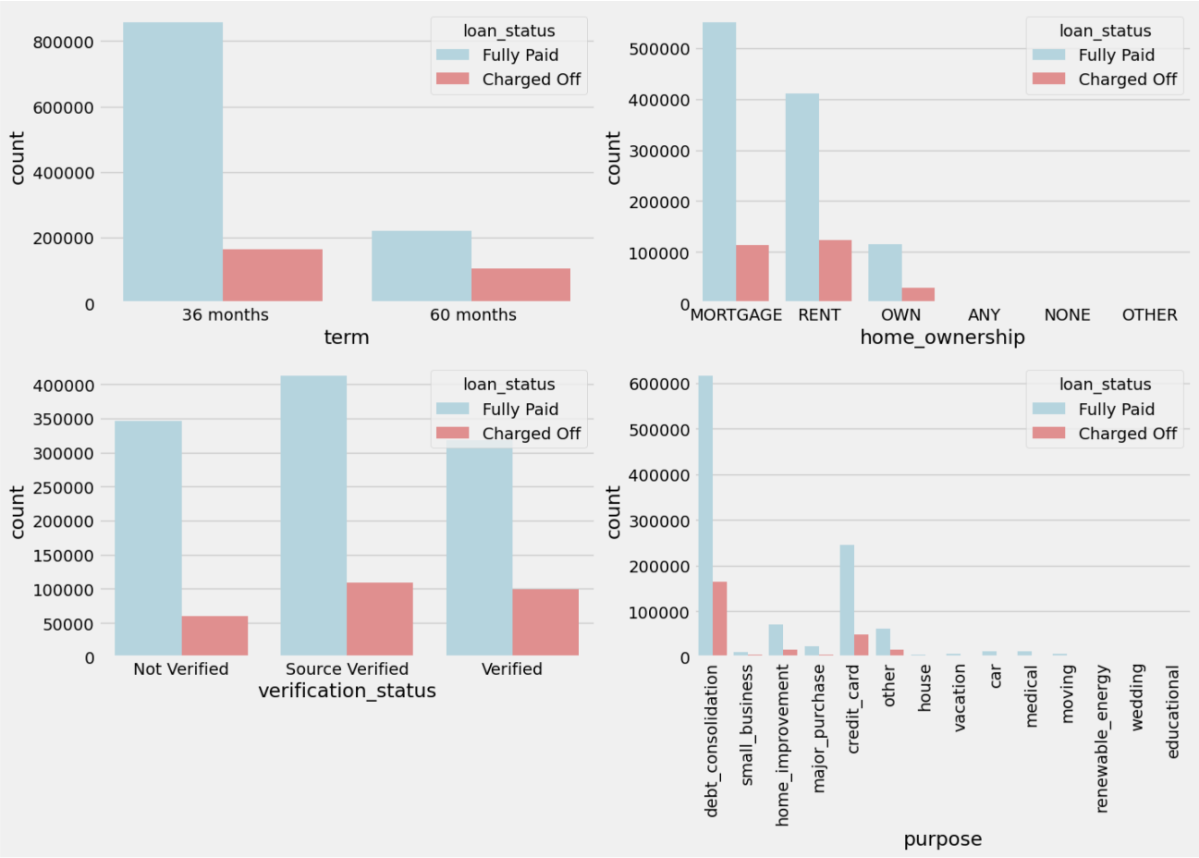




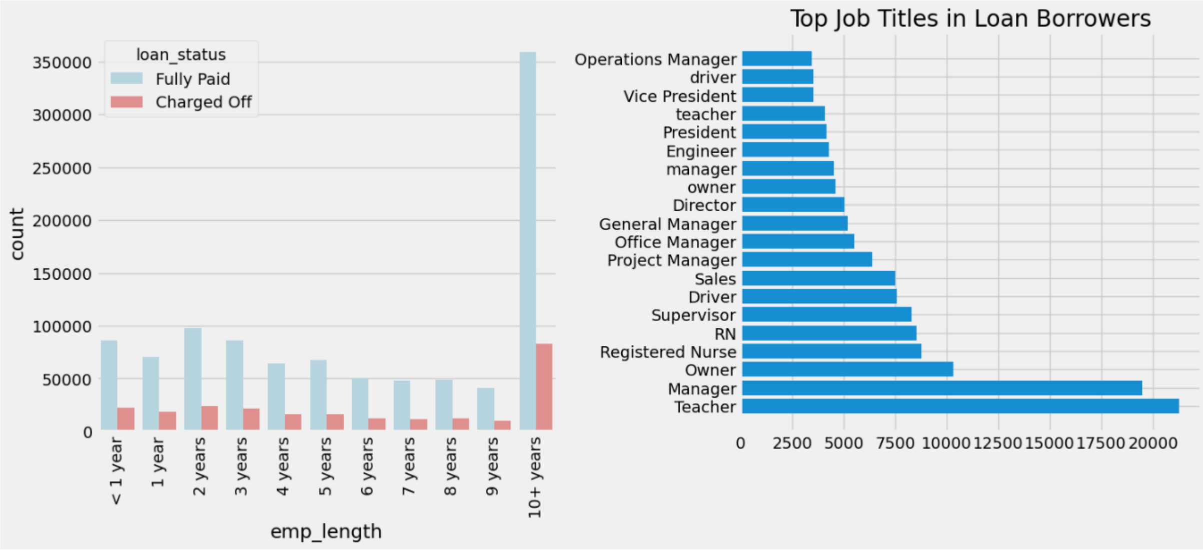
9 Categorical variables for default prediction

In addition to visualization, the analysis also implements categorization when analysing the data. For example, debts were divided into good debts and bad debts, whereas income was classified into low income, medium income, and high income. Data categorization separates data into subsets that share similarities in certain aspects. It is different from classification, which assigns data into different unique classes within a system. The advantage of grouping is that it can save the volume of computation needed and draw conclusions on a feature basis. Nevertheless, categorization requires analysts' personal judgment based on experience and sensitivity to the datasets.

- 36-month term loans have a higher default rate
- Rent and Mortgage holders are more likely to default
- Verification status doesn’t seem to impact default likelihood
- Credit card and debt_consol are more likely to default

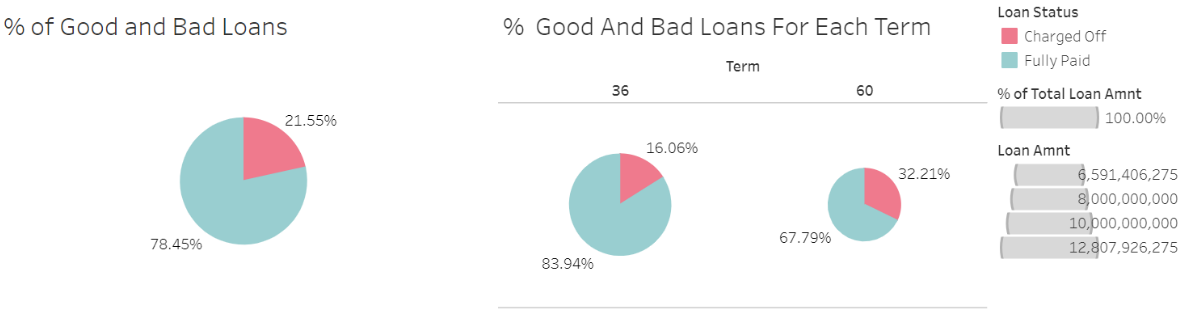


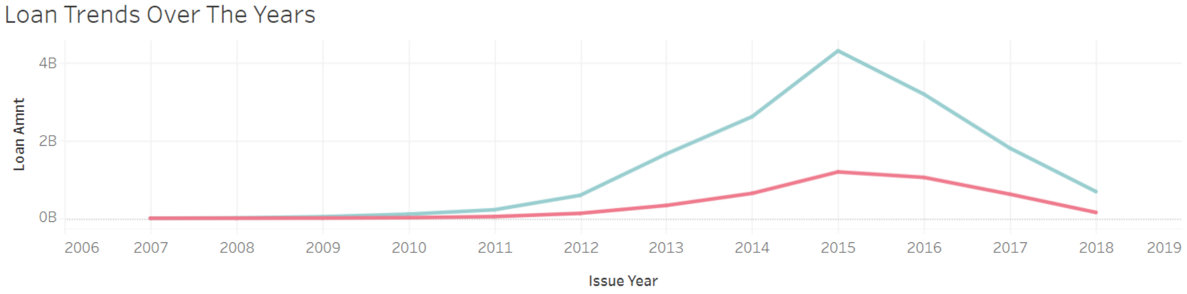
- Employment length > 10 years are the highest defauters
- Teacher, Manager, Owner and Nurse are the top borrower categories.

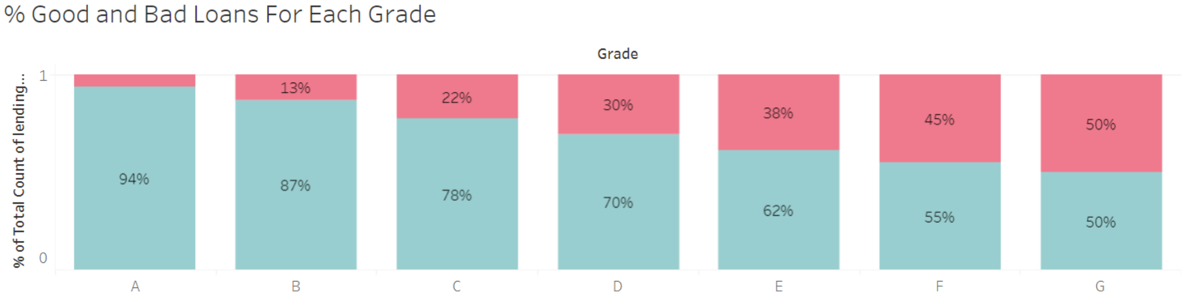


10 Loans Type, Quality, Analysis

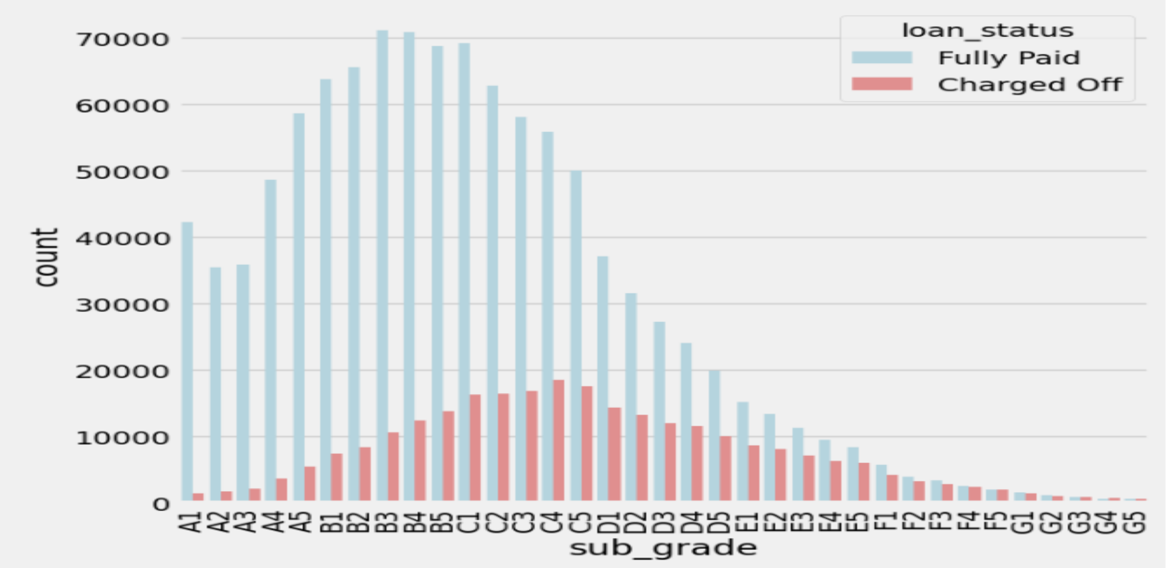
The below graph shows that Long term loans have higher default rate. 2014-2016 had the highest borrowings as well as defaults. Loan grade ‘C’ and above are riskier.







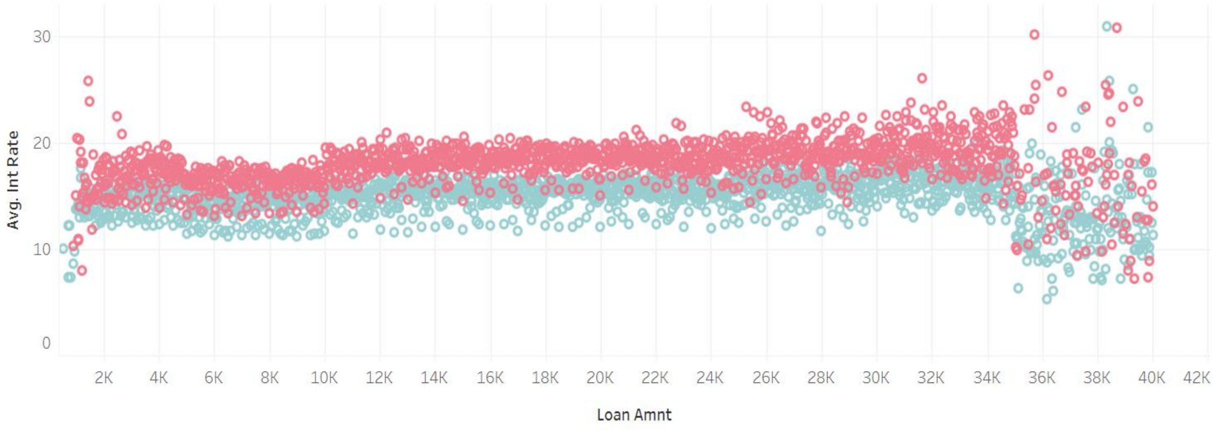
Drilling down into the subgrades, C4 has the highest default rate.



11 Interest Rate and Annual Income Analysis

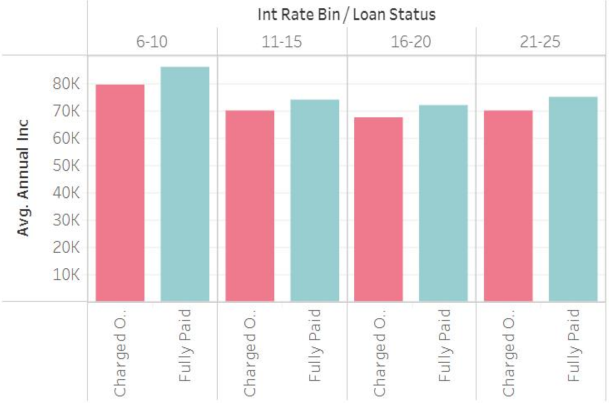
Higher interest rate loans are more likely to default, specially where the annual income is <$70k. Borrowers with multiple credit lines are more likely to default.

Intrest Rate Vs. Number of Loans Approved



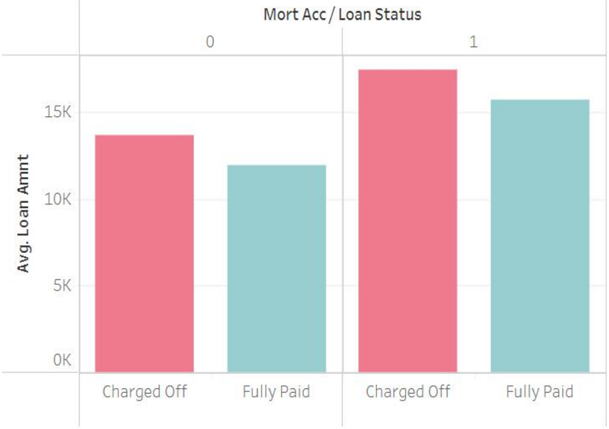
Loans with higher interest rates are more likely to default

Int Rate vs Income



High interest rate loans are more likely to default, specially when the avg annual income is <70k

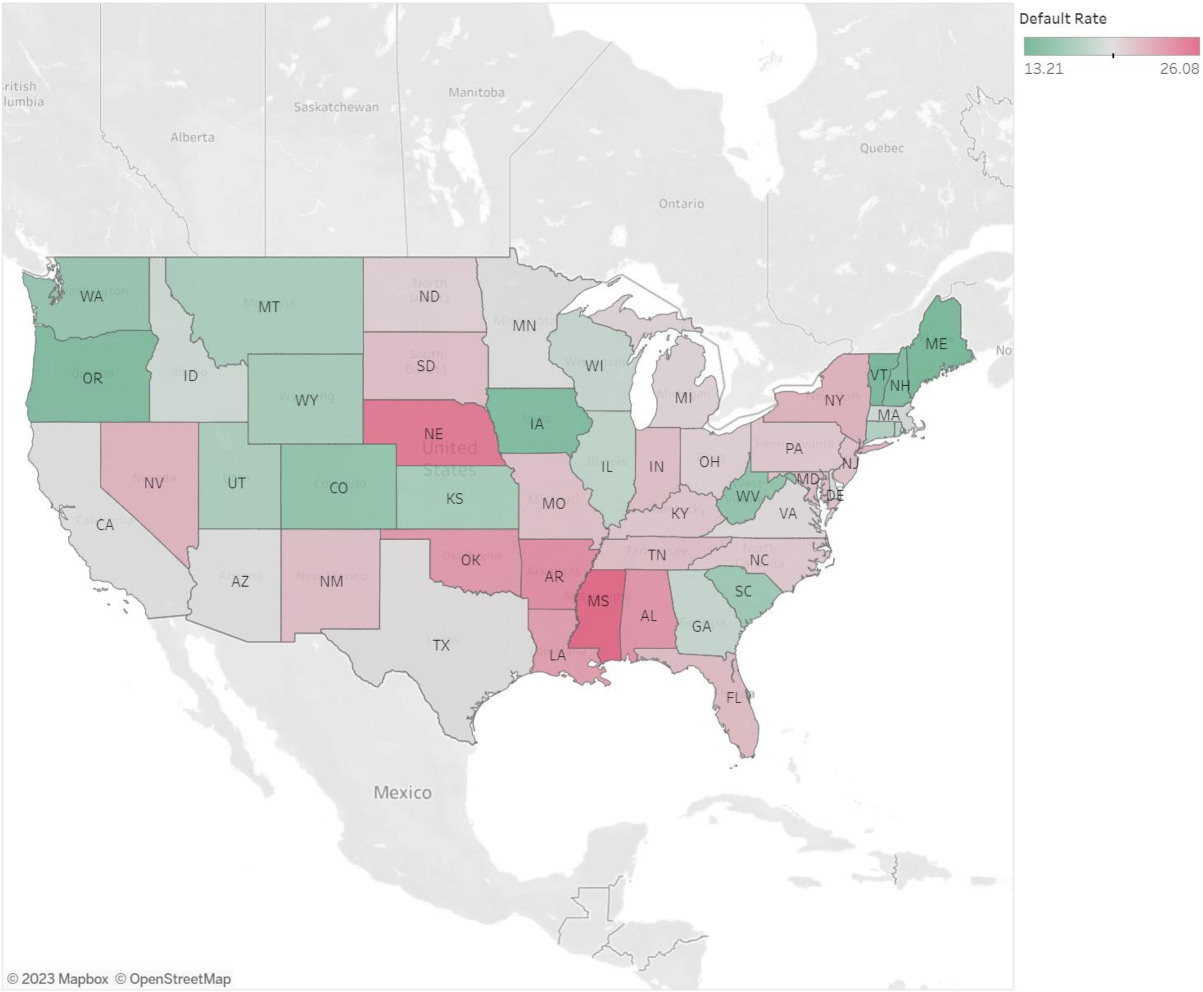
Mort Acc



Borrowers with multiple credit lines have a higher liklihood to default.

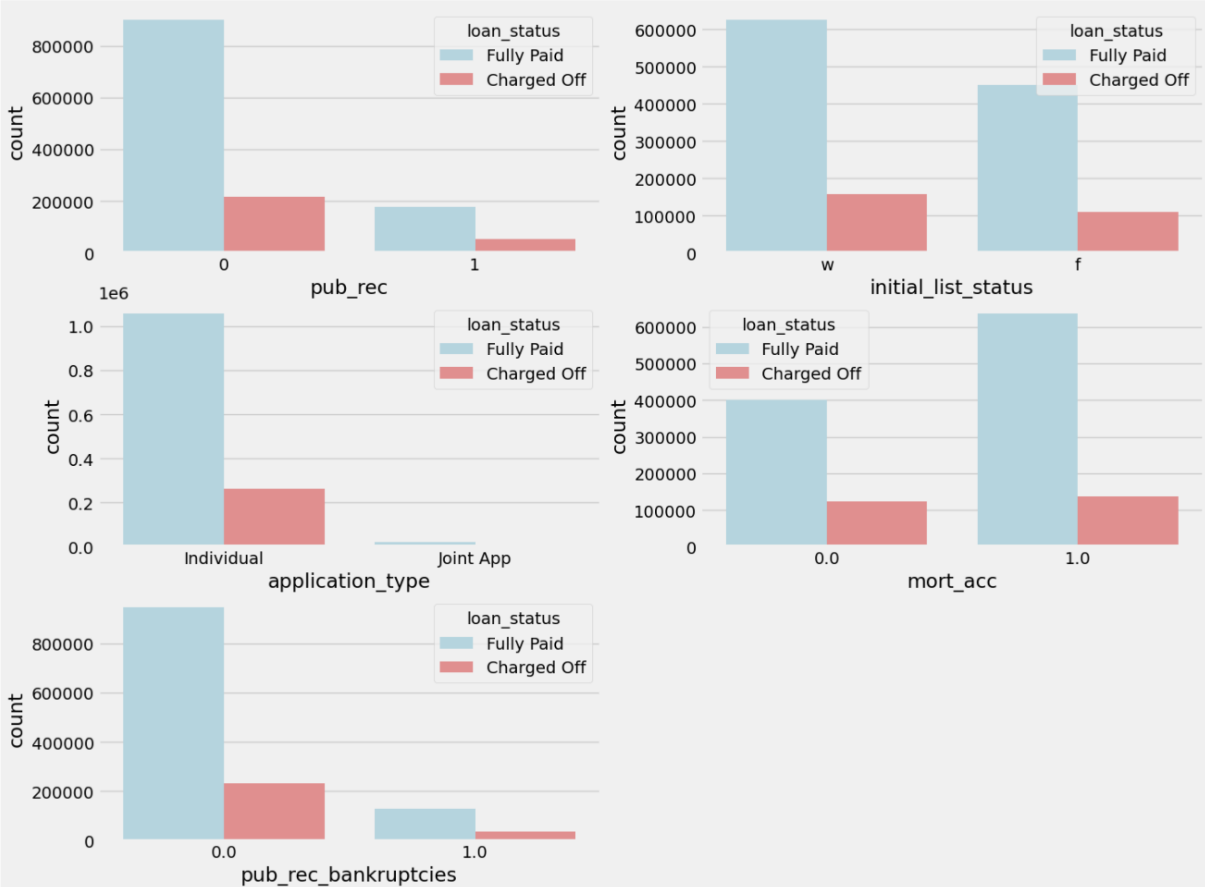
12 Loan Default by State

Western states have a lower default rate as compared to eastern, southern and mid central states.



13 Other Variables

Other variables such as public records, public bankruptcies and initial list status have a very high positive correlation to the loan default.



14 Statistical Techniques

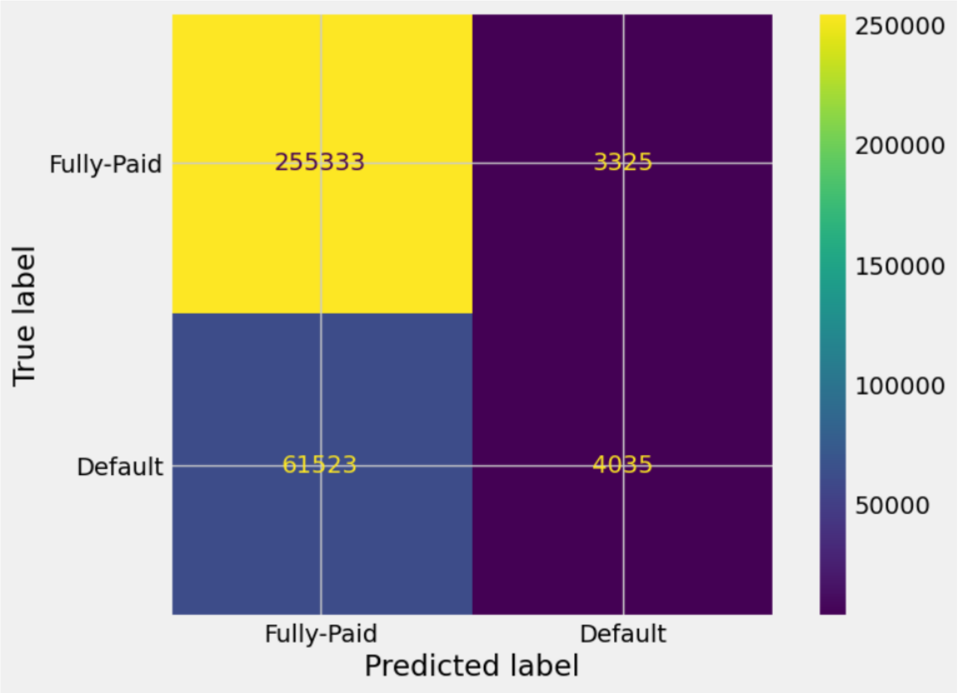
The statistical model for this project is based on a conservative approach and evaluation metrics. The loan default prediction is a problem of binary classification.
Logistic regression is a statistical model that is primarily used for modelling binary dependent variables. It is simple to implement and effective while having shortcomings such as poor performance with non-linear data and highly correlated features.
Random forest is a machine learning method for classification, regression, and other tasks by constructing a multitude of decision trees at training time. It can operate with a massive volume of data while reducing errors and the impact of outliers.

14.1 Linear Regression Model

The model predicted 80% accuracy, meaning 20% of the loans are predicted to default.

Cross-Validation Score is: [0.80 0.796 0.80 0.796 0.797]
Mean Accuracy: 0.798
Standard Deviation of Accuracy: 0.0021

14.2 Confusion Matrix



14.3 Decision Tree Classifier

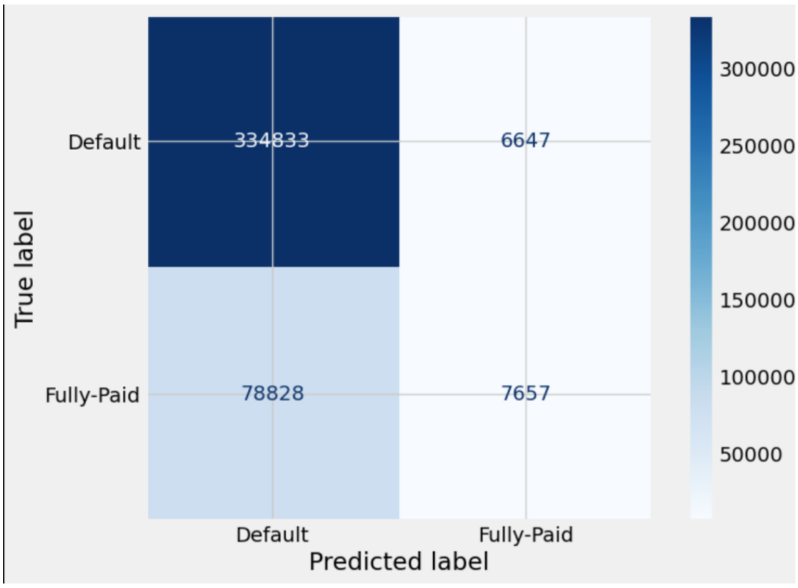
This model predicts 70% accuracy, meaning 30% of the loans are predicted to default.

Accuracy is: 70.14

Crossvalidaton score is: 70.18

14.4 Random Forest classifier

This model predicts 80% accuracy, meaning 20% of the loans are predicted to default.



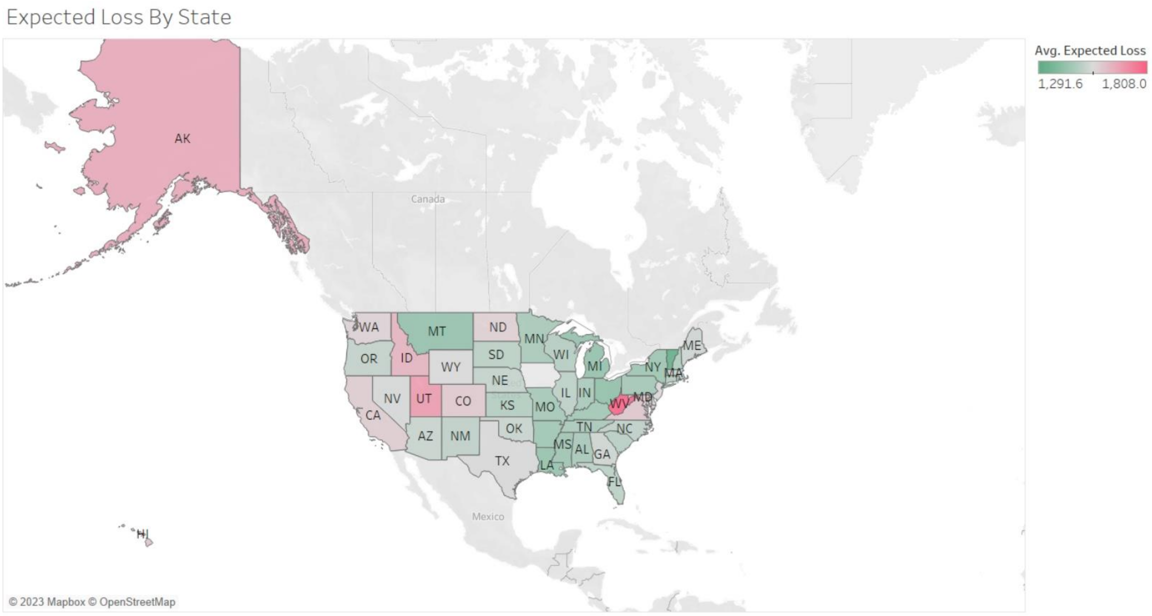
15 Conclusion

Lending Club is a peer-to-peer lending company, where its head office is located in San Francisco, California. It was the first peer-to-peer lender to register with the Securities and Exchange Commission to issue securities and trade loans in the secondary market. The company's core product is unsecured personal loans. The borrowers use to raise car loans and cards.
However, Lending Club needs partner banks to provide the credits. This may cause pressure on its profitability. In the study, the data type of the set which appeared at the beginning of the paragraph is both qualitative and quantitative. The method this study used is visualization, categorifications as well as statistical techniques.

This study mainly talks about various aspects as well as risk analysis of the loan. The study analysed bad loans from various aspects by using different charts and analysis.

The most important thing the company needs to pay attention to is the potential risk of some current loans becoming bad loans in the future. In the future, they could predict the possibility of current loans which had the possibility to become bad loans. If they could predict this element, they will avoid many problems in the future, more importantly, they will make more profits in the future.

Expected Loss by State Using Predictive Modelling



Map based on Longitude (generated) and Latitude (generated). Colour shows average of Expected Loss. The marks are labelled by Addr State.

The expected loss for each state is the summation of probability of default, times the payment gap.

Payment Gap is defined as the difference between total amount of the loan and the amount already paid at a specific point in time.

The probability of default is obtained by matrix transformation based on the parameters estimated from a training set, with variables as annual income, funded amount, home ownership, borrower's grade and verification status. The results, based on the model assumptions, show that the states of WV, UT, ID, AK are at a higher risk of loss, whereas the mid-east states present a much more optimistic loan repayment.