LENDING CLUB CASE STUDY

SUBMISSION

Names

- Pramod Yadav
- Avishek Sen Gupta

Lending Club

Introduction

The **Lending Club** is a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant **is likely to repay the loan**, then not approving the loan results in a loss of business to the company
- If the applicant **is not likely to repay the loan**, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

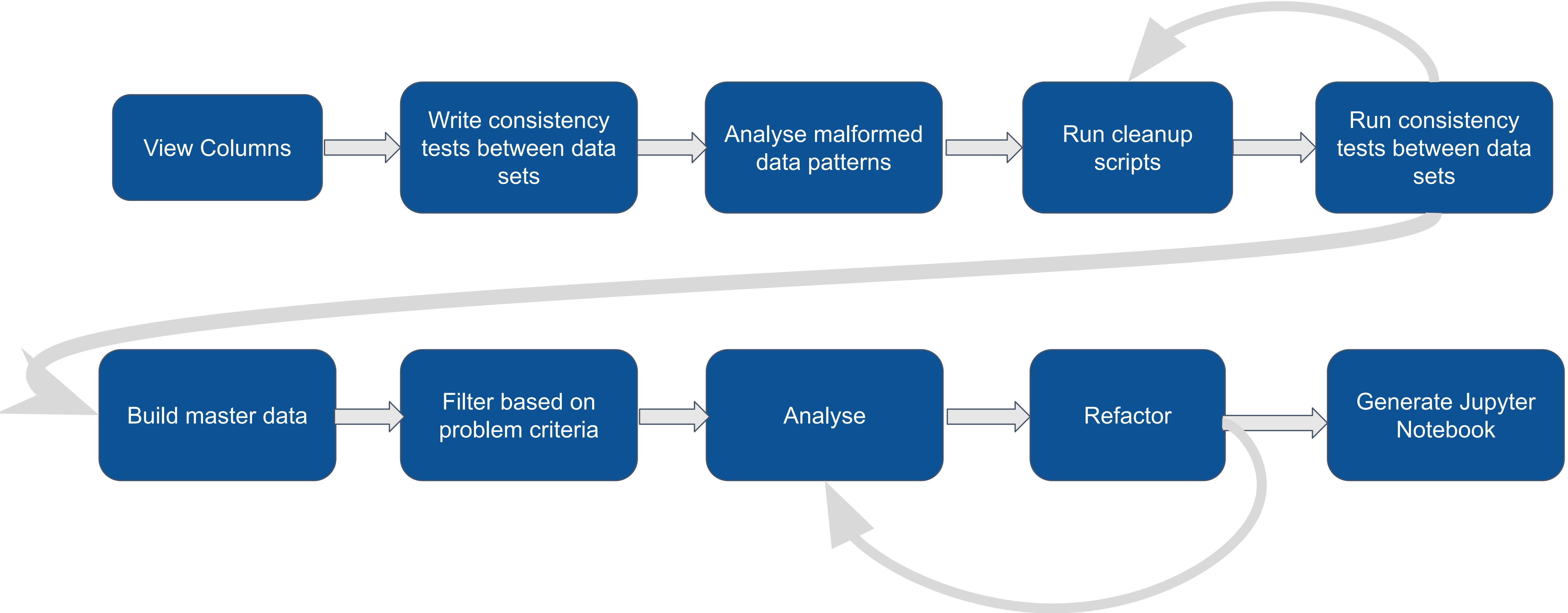
Business Understanding and Domain

- Like most other lending companies, lending loans to **'risky' applicants** is the largest source of financial loss (called **credit loss**).
- The credit loss is the amount of money lost by the lender when the **borrower refuses to pay or runs away with the money owed**. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as **'charged-off'** are the **'defaulters'**.

Objective

The company wants to understand the **driving factors (or driver variables) behind loan default**, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its **portfolio and risk assessment**.

Problem Solving Workflow



Data Understanding: CRISP-DM #2

Data understanding was achieved in several ways:

- A **data dictionary** was provided, which contained explanations for all the columns that are present in the original loans data set. This also directly helped in understanding what the valid values were in a particular columns, and later helped in the **Data Preparation** step.
- There was a specific session which was conducted where information about these columns were **clarified to a further degree**, eg: **the ranking of the loan grades**, etc.
- A basic understanding of the loan industry also helped in doing exploratory data analysis based on some initial intuition, eg: **loan amount could be relevant** to determining the possibility of default, etc.

The analysis is mostly focused on **Exploratory Data Analysis**, and some conclusions have been made, and some advice suggested. However, more rigorous **hypothesis testing** as well as potential **data transformations** (on data which could be normal but is skewed) **need to be made before drawing definitive conclusions** around **Driver Variables**.

Data Preparation: CRISP-DM #3

This is a summary of the steps we have taken to clean the data:

1. Checking the **null values** across the column and row.
   a. Dropped the column which have 100% missing values. 54 columns have missing values such as *annual_inc_joint,tot_coll_amt*
   b. Dropping **column which have more than 60% of missing values**
2. Checking the **duplicate rows** in the data
3. Dropping columns which have **constant values** Such as: *pymnt_plan,initial_list_status*
4. Dropping **customer behaviour columns** which we will not have initially while granting the loan Such as:*last_pymnt_d,application_type*
5. Checking **highly correlated columns** which we can drop such as : *funded_amnt,funded_amnt_inv*
6. Dropping **rows which aren't helpful for analysis** like "*Current*" in *loan_status*, because they contain data that would not be useful for inference or prediction.
7. **Correcting data types** of columns such as: *int_rate,emp_length*
8. Creating **derived columns**:
   a. Created Year and Months Column from *Date_of_Issue*
   b. Interest rate Category from Interest Rate Column

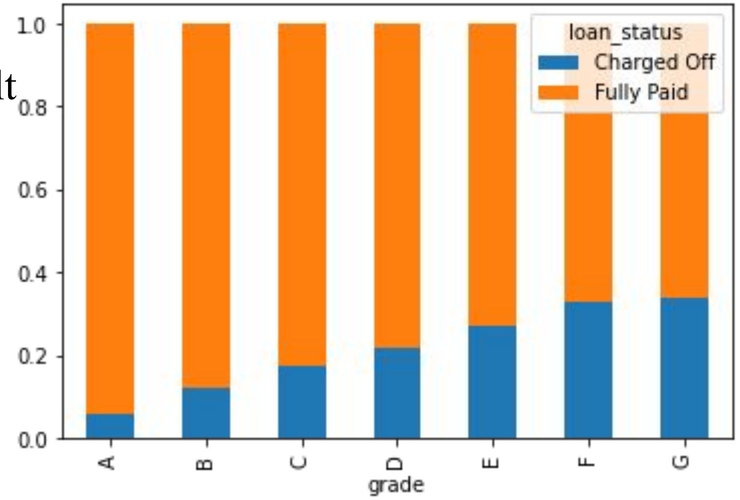
Data Analysis

Exploratory Data Analysis was divided into the following categories:

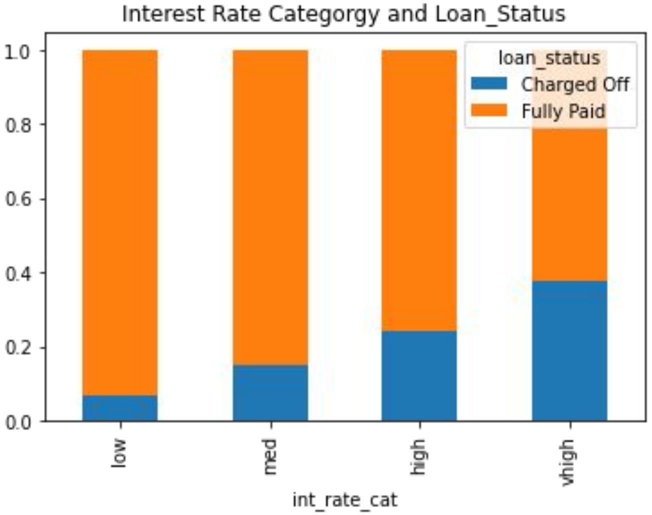
- Univariate Analysis: Checking **Distribution and Frequencies** of data
- Segmented Univariate Analysis: **Segmenting the data** as per loan_status column and performing univariate analysis
- Bivariate Analysis : Using **2 columns** at a time to see the relationship among the variables.

Findings

Lower grades imply a higher percentage of default

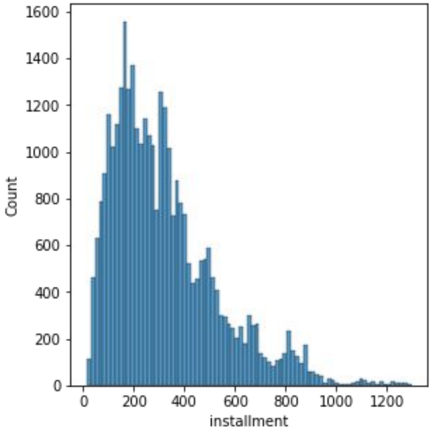


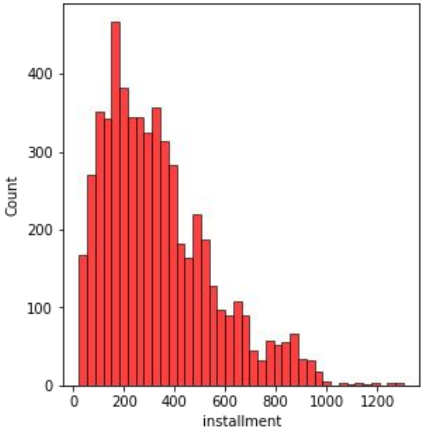
A higher interest rate implies a higher percentage of default.



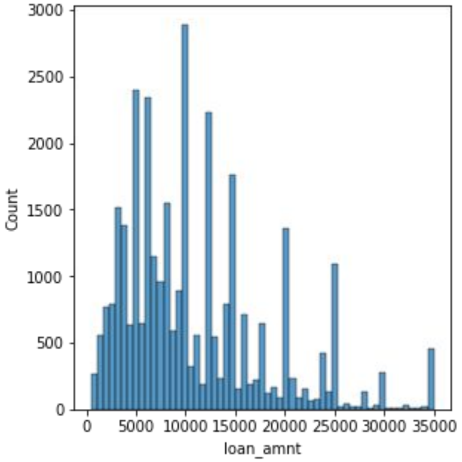
Findings

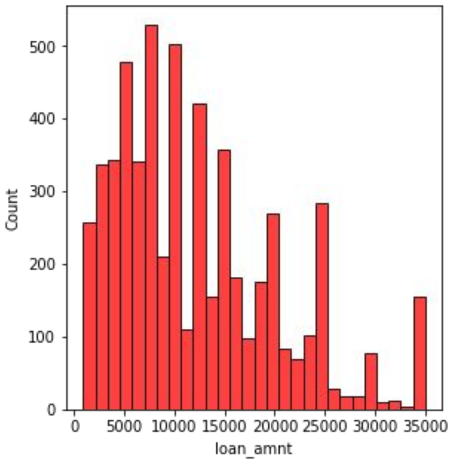
For Defaulters, the installment size distribution is right skewed with a peak of around 160, indicating that lower-valued loans around $160 are more likely to default.

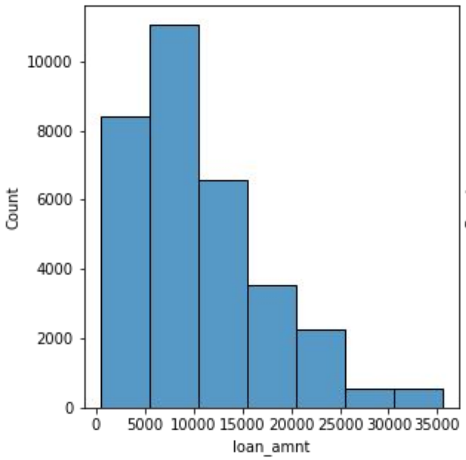


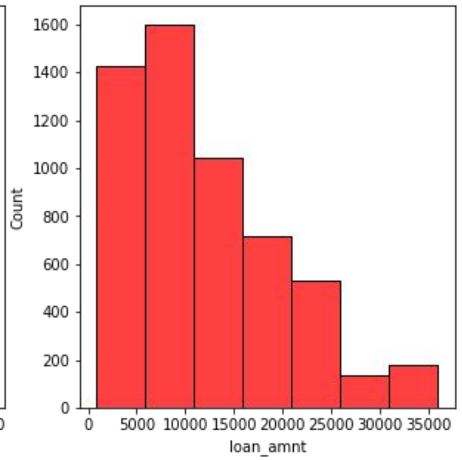


There are interesting spikes when plotting counts by loan amounts, at 5000, 10000, 12500, 15000. This suggests people usually borrow at these amounts. Taking a larger bin width of 5000 smoothes the histogram, and gives us a similar right skewed distribution for Defaulters.

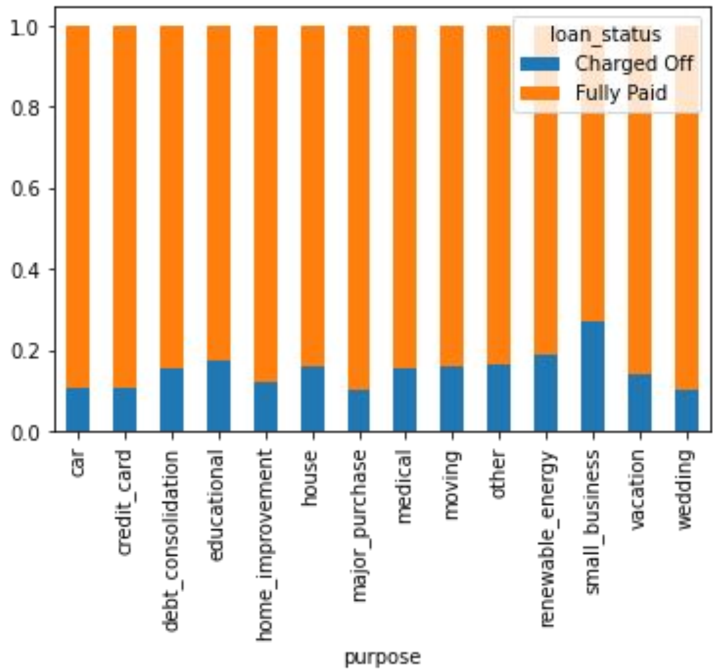


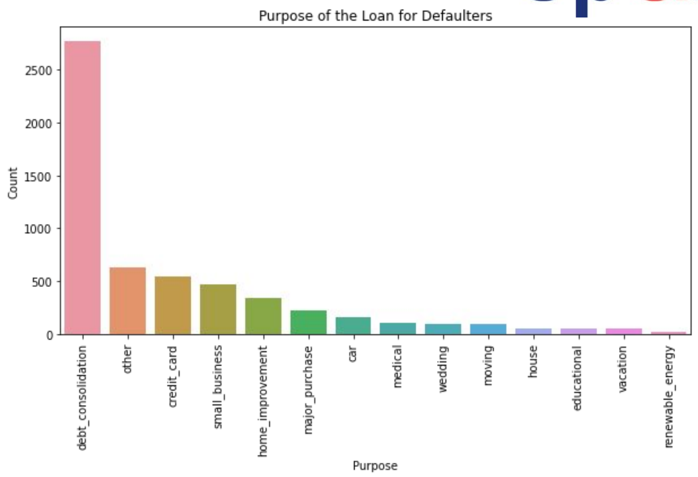




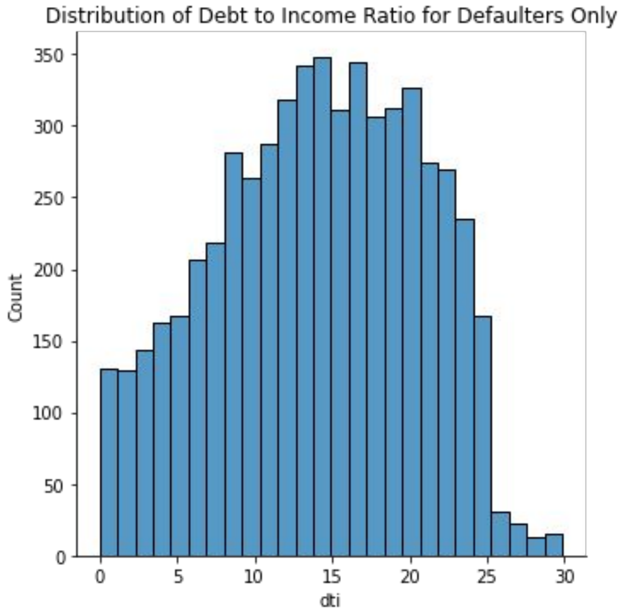


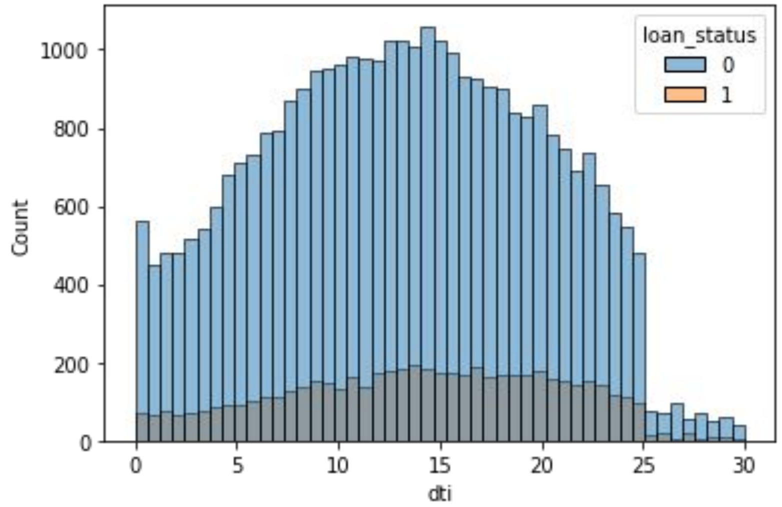
Findings





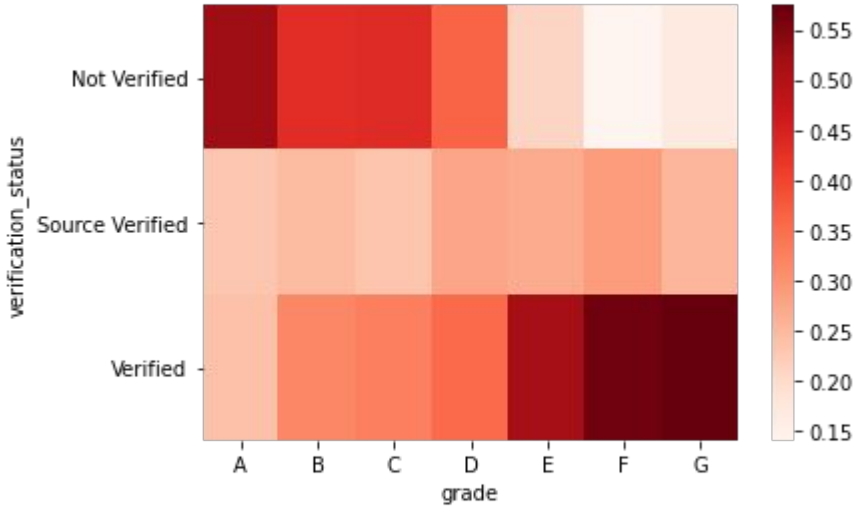
No specific trend in default percentages noted based on loan purpose. However, among Defaulters, a large percentage state debt_consolidation as the purpose of the loan.





Debt-to-Income Ratio is somewhat normally distributed across Defaulters with a possible left skew. Median DTI for both Defaulters and Non-Defaulters is around 15.

Findings



Verification Status Heatmap: Possibility of Default when using Third Party as verification indicates a higher chance of loan repayment.

Annual Income and Loan Amount:

Defaulters earn less median income and request slightly higher loan amounts, than non-Defaulters.

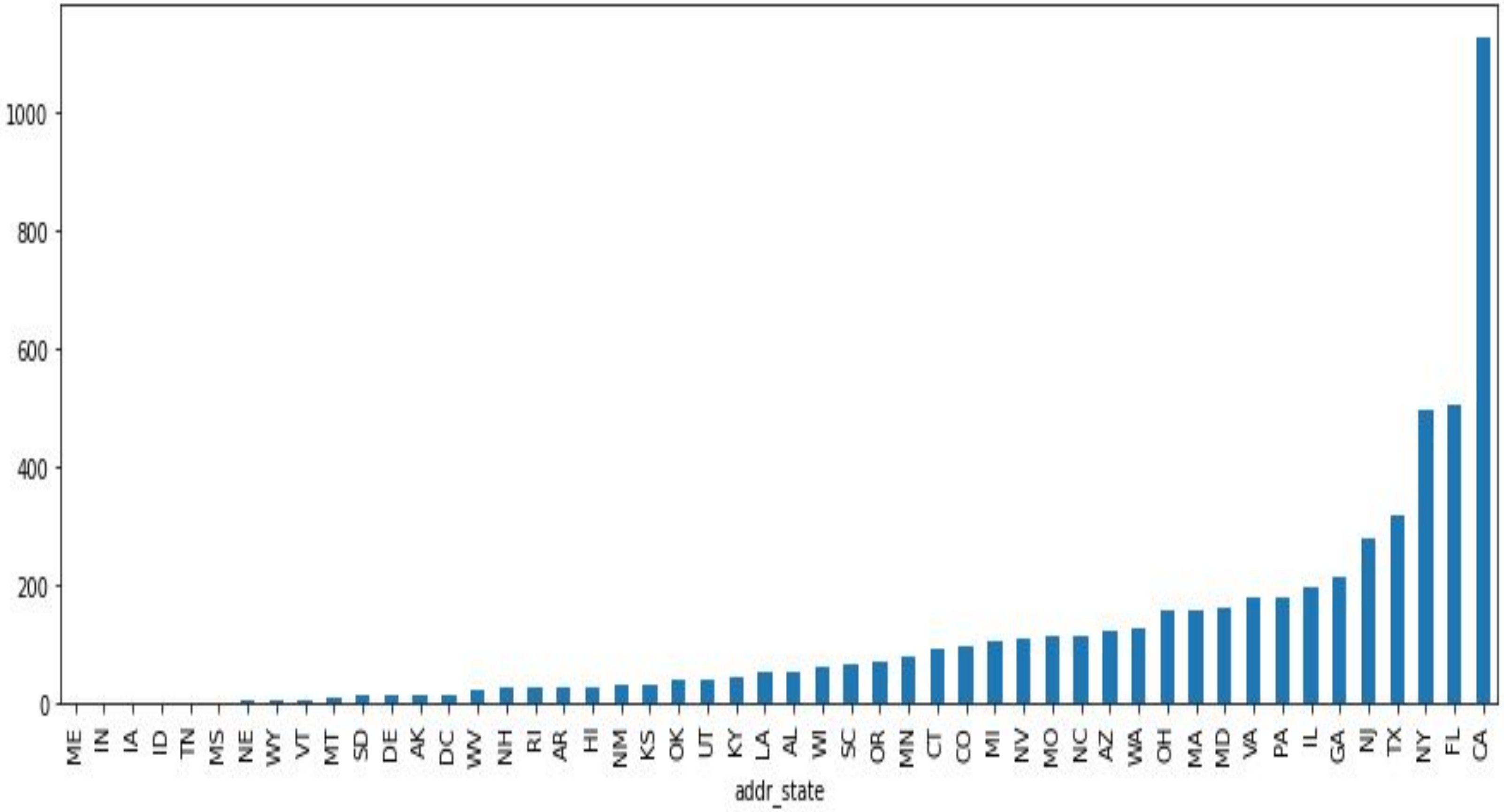
loan_status	count	mean	std	min	25%	50%	75%	max
0	32950.00	69862.50	66562.25	4000.00	41132.75	60000.00	84000.00	6000000.00
1	5627.00	62427.30	47776.01	4080.00	37000.00	53000.00	75000.00	1250000.00

loan_status	count	mean	std	min	25%	50%	75%	max
0	32950.00	10866.46	7199.63	500.00	5200.00	9600.00	15000.00	35000.00
1	5627.00	12104.39	8085.73	900.00	5600.00	10000.00	16500.00	35000.00

Findings

Defaulters originate mostly from:

- California
- Florida
- New York



Conclusions and Advice

Key Drivers:

- **Loan Grade / Sub-Grade**: Lower grades (G is the lowest) imply a higher percentage of defaulters.
- **Interest Rate**: A higher interest rate implies a higher percentage of defaulters.
- **Installment Amount**: The defaulters are normally distributed with a right skew across Installment Amounts. With a proper data transformation, the distribution can be made normal.
- **Loan Amount**: The defaulters are normally distributed with a right skew across Loan Amounts. With a proper data transformation, the distribution can be made normal.
- **Loan Purpose**: The most frequent purpose of loan cited is “debt_repayment” for Defaulters.
- **Debt-to-Income Ratio**: The defaulters are normally distributed across Debt-to-Income Ratio.
- **Verification Status**: Using a Third Party for verifying income source (Source Verified) seems to give a better repayment rate (lower default) across all grades of loans.
- **Annual Income and Loan Amount**: Defaulters earn less median income and request slightly higher loan amounts, than non-Defaulters.
- **State**: Defaulters originate mostly from California (high proportion), Florida, and New York.