Lending Club Case Study

VINAY JOSHI

Table of Content

1. Introduction

- Welcome to the presentation on Exploratory Data Analysis of Loan Default Risk.
- In this presentation, we delve into the realm of risk analytics within the consumer finance sector, exploring how data-driven insights can mitigate financial risks associated with loan approval.
- By analyzing past loan applicant data, we aim to uncover patterns and factors that influence the likelihood of loan default.
- Through this exploration, we gain valuable insights into the decision-making processes of lending institutions and the role of data in minimizing financial losses.

2. Understanding of the Business Requirements

- Our work revolves around a consumer finance company specializing in various loan offerings for urban customers.
- When assessing loan applications, the company faces two primary risks:
  - Approving loans to individuals who may default, resulting in financial loss
  - Rejecting loans to potential customers who could contribute positively to the company's business
- This dual risk underscores the importance of informed decision-making based on applicant profiles and historical data.
- Our task is to utilize Exploratory Data Analysis (EDA) techniques to discern patterns and trends within our loan dataset, thereby identifying indicators of default risk.
- By doing so, we empower the lending institution to make more informed decisions, minimize losses, and optimize its lending portfolio.

3. EDA: Methodology

- Exploratory Data Analysis (EDA) serves as the cornerstone of our investigative approach, enabling us to uncover insights and patterns within our loan dataset.
- The methodology encompasses a systematic exploration of the data, comprising several key steps:

Data Cleaning

Identifying and treating data quality issues

- Remove columns with:
  - Only 1 unique value for all the rows
  - More null values than the actual values. In other words, null values for more than 50% of the rows
  - Columns that aren't useful for our analysis like Desc, title etc.
- Clean up the rows:
  - Remove duplicate rows
  - Check for missing values across the row and remove the row if significant
  - Remove the current loans as they aren't needed for the analysis
  - Check for columns with null values. Either remove or fix the null values based on the nature of the data
- Standardize the data
  - Identify data formatting issues and remove unnecessary characters and strings like %, month etc.
  - Remove the outliers in numerical data

Univariate Analysis

Examination of individual variables to understand their distributions, central tendencies, and variability.

Analyze the variables:

- Continuous Variables: That can take on an infinite number of values within a certain range.
  - Income
  - Loan amount
  - Interest rate
- Categorical Variables: Represent distinct categories or groups
  - Nominal Variables: Variables that represent categories without any inherent order or ranking
    - Loan Purpose
  - Ordinal Variables: Variables that have categories with a specific order or ranking
    - Grade
    - Verification Status

Segmented Univariate Analysis

Segmented univariate analysis is an extension of univariate analysis where we examine the distribution of a single variable across different segments or groups within the dataset. This technique allows us to gain deeper insights into how the distribution of a variable varies across different categories or conditions.

It Is performed on the Categorical values

Bivariate Analysis

It is a method used to explore the relationship between two variables within a dataset. Unlike univariate analysis, which focuses on analyzing a single variable at a time, bivariate analysis examines how two variables are related to each other. This analysis helps us understand the nature and strength of the relationship between the variables and identify any patterns or trends present in the data.

Types of Variables:

- Continuous vs. Continuous
- Categorical vs. Categorical
- Continuous vs. Categorical

Visualization

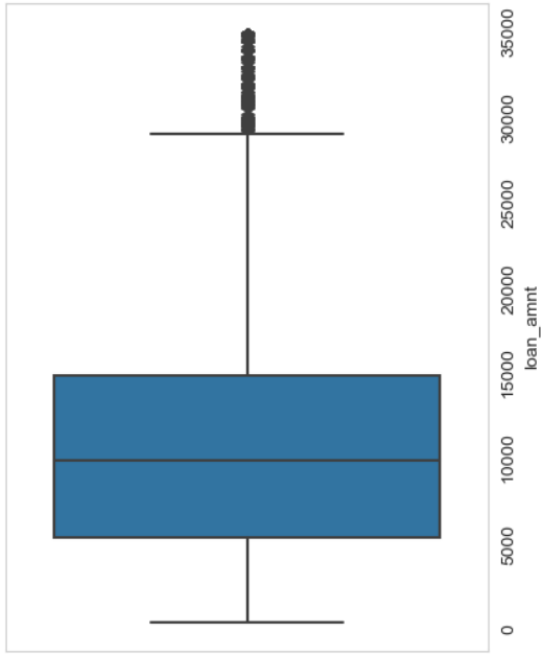
Visualization techniques play a crucial role in data analysis by providing intuitive and insightful representations of the underlying patterns and relationships within the dataset. Here's a brief overview of some common visualization techniques:

- Histograms
- Box Plots (Box-and-Whisker Plots)
- Bar Charts
- Heatmaps
- Line Charts
- Scatter Plots
- Pie Charts

4. Visualization

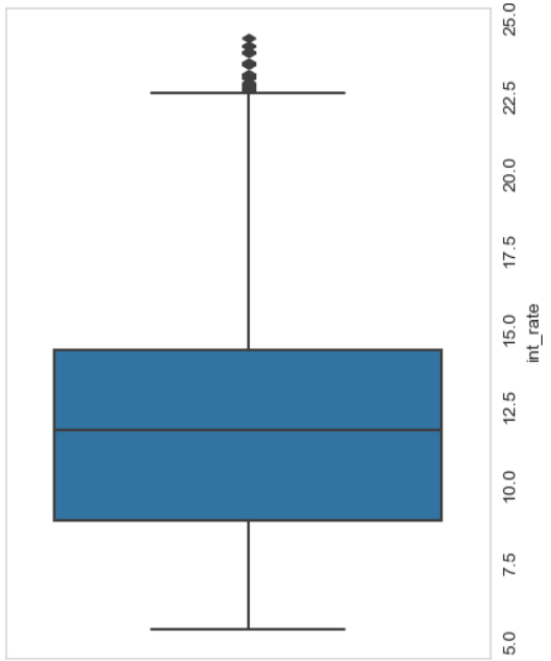
4a. Univariate Analysis: Numerical Data

Loan Amount



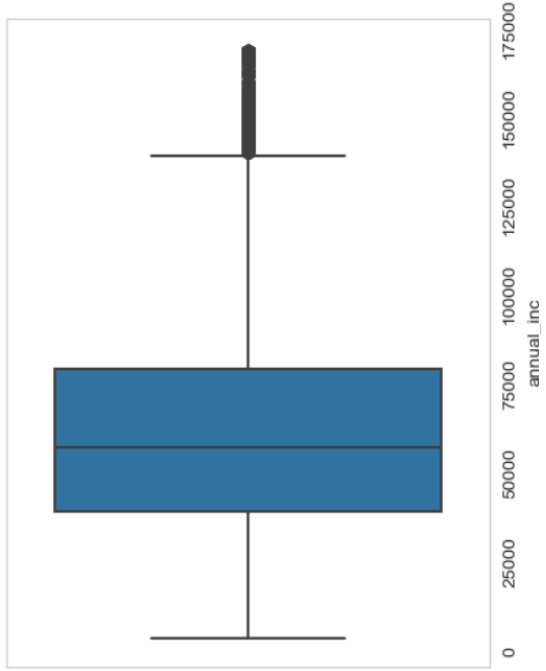
Percentile	Value
25%	5,500
50%	10,000
75%	15,000
100%	35,000

Interest Rate



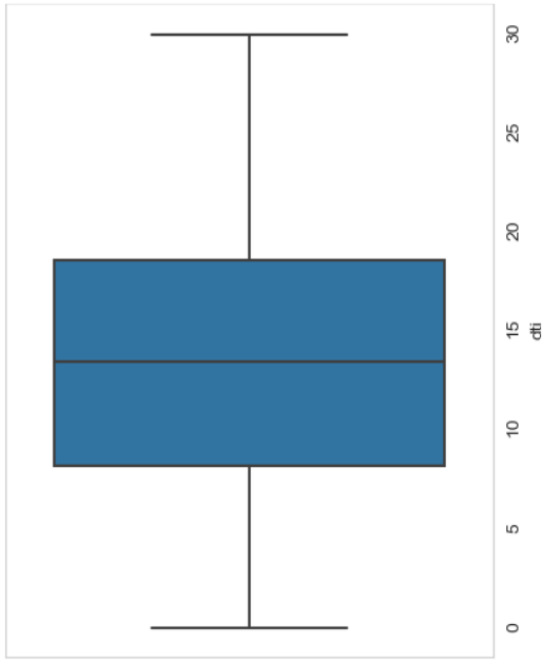
Percentile	Value
25%	8.94%
50%	11.83%
75%	14.42%
100%	24.4%

Annual Income



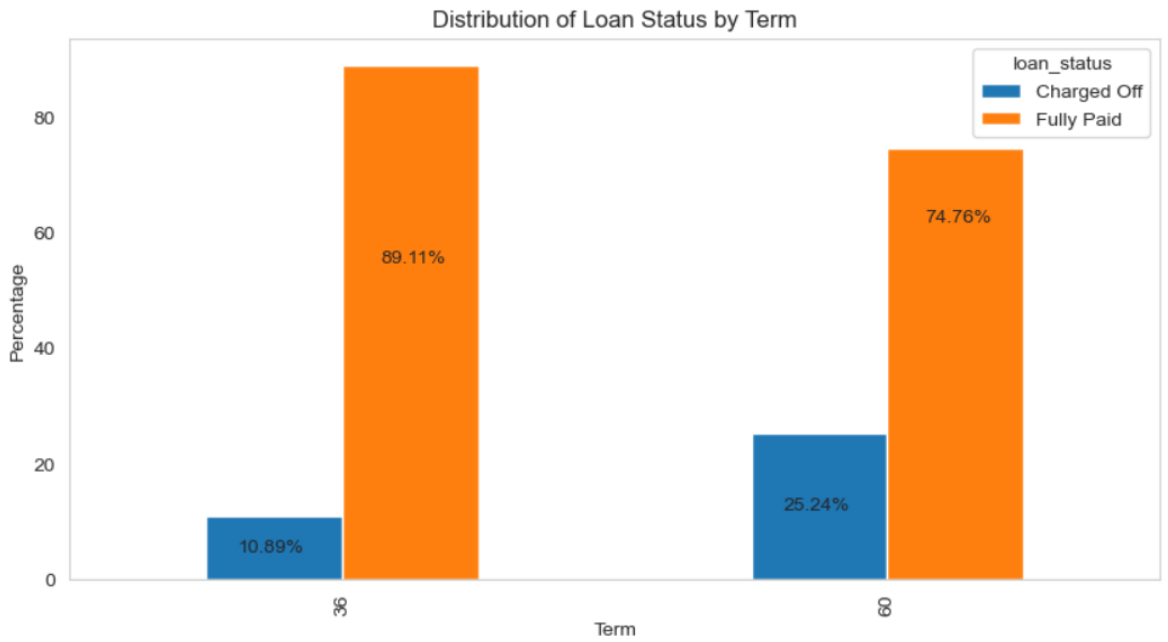
Percentile	Value
25%	40,000
50%	58,000
75%	80,000
100%	1,70,000

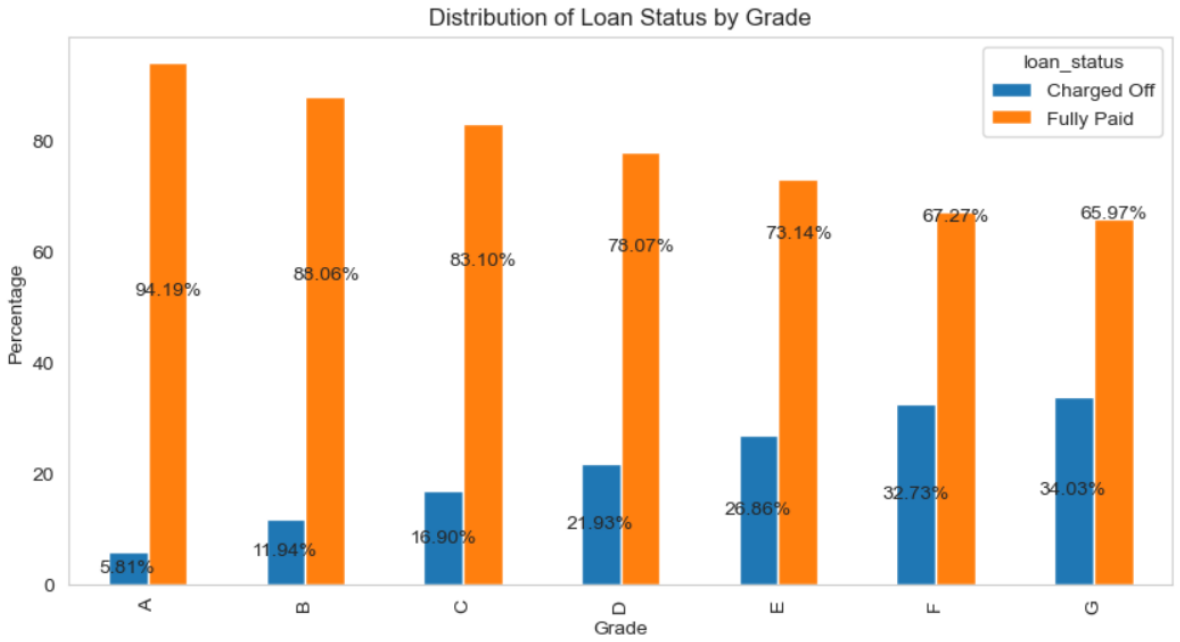
DTI



Percentile	Value
25%	8.24
50%	13.45
75%	18.59
100%	29.99

4b. SUA: Distribution of Loan Status by Term & Grade

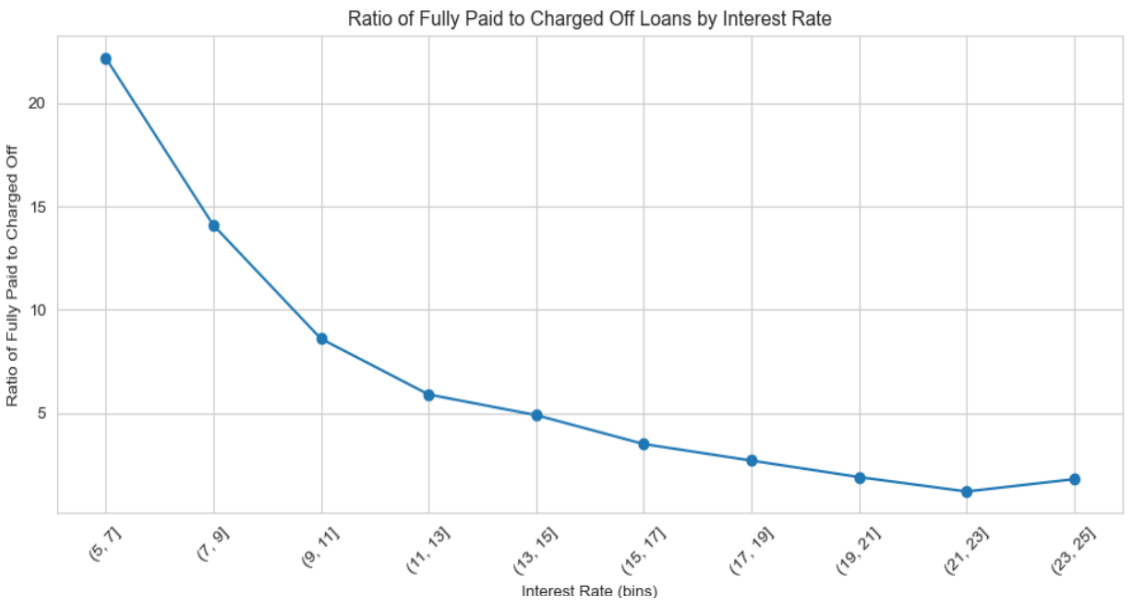


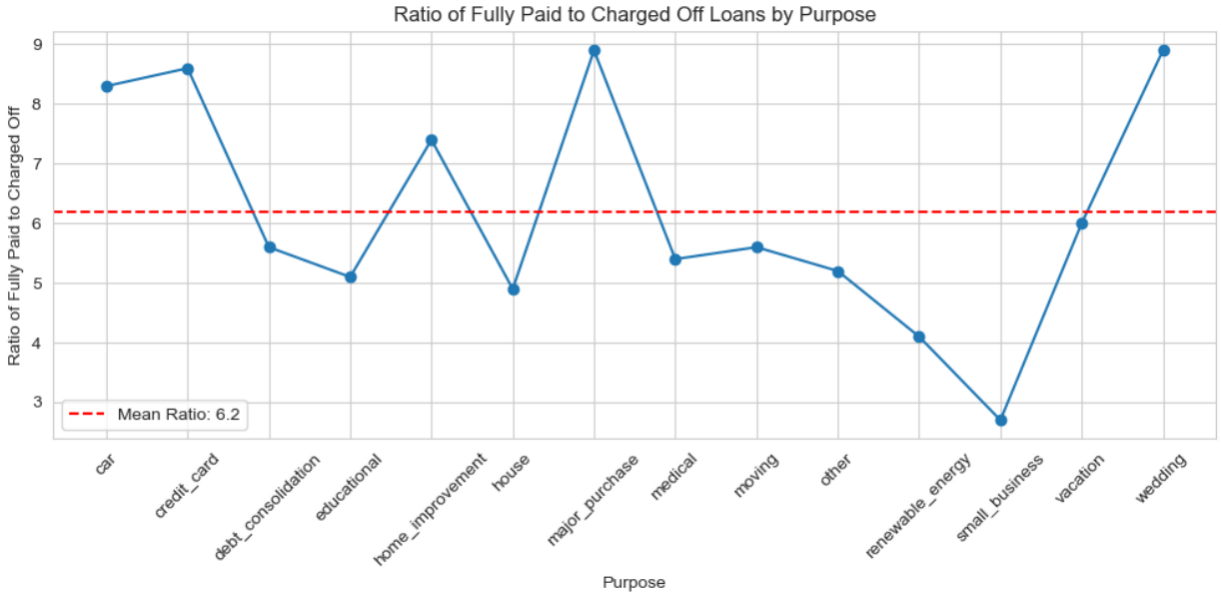


- The probability of Charged Off loans is higher in the longer term loans

- The ratio of Fully Paid to Charged off loans decreases drastically with the High Risk loans (As we move from Grade A to Grade G, the risk increases)

4b. SUA: Ratio of Loan Status by IR & Purpose

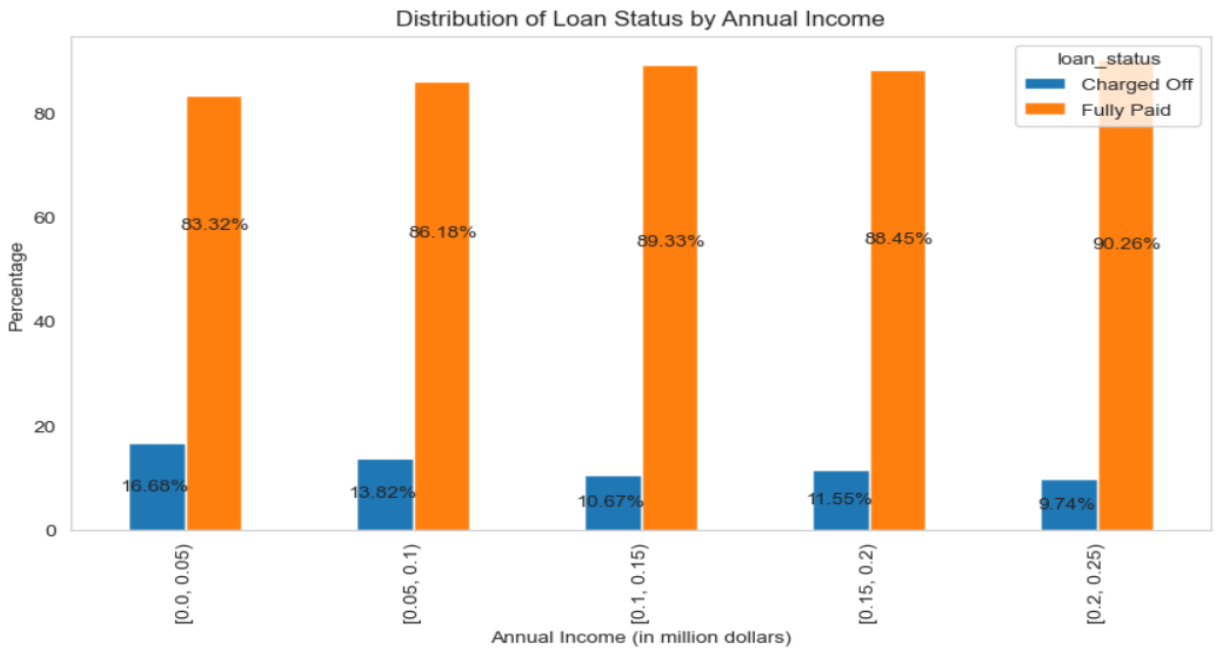


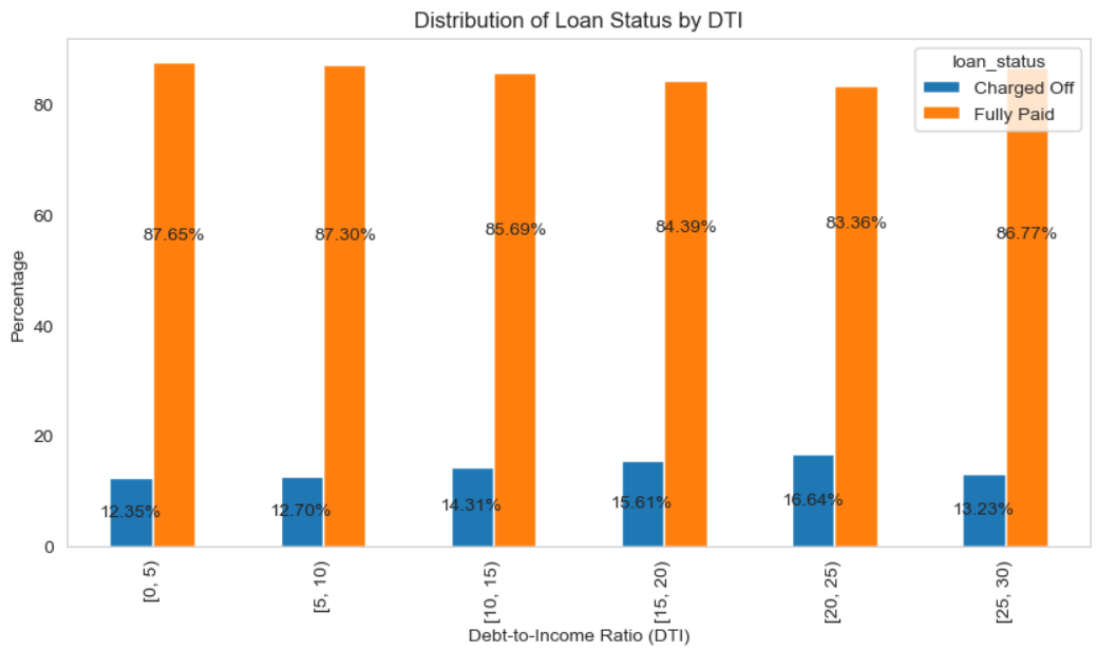


- The ratio of Fully Paid to Charged off loans decreases drastically with increase in the interest rate

- The ratio of Fully Paid to Charged off loans varies based on the Purpose. A larger majority of people who took the loan for Major Purchases, Credit Card, Cars and Wedding paid their loans back while the loans taken for Small business did the worst.

4b. SUA: Distribution of Loan Status by Annual Income & DTI

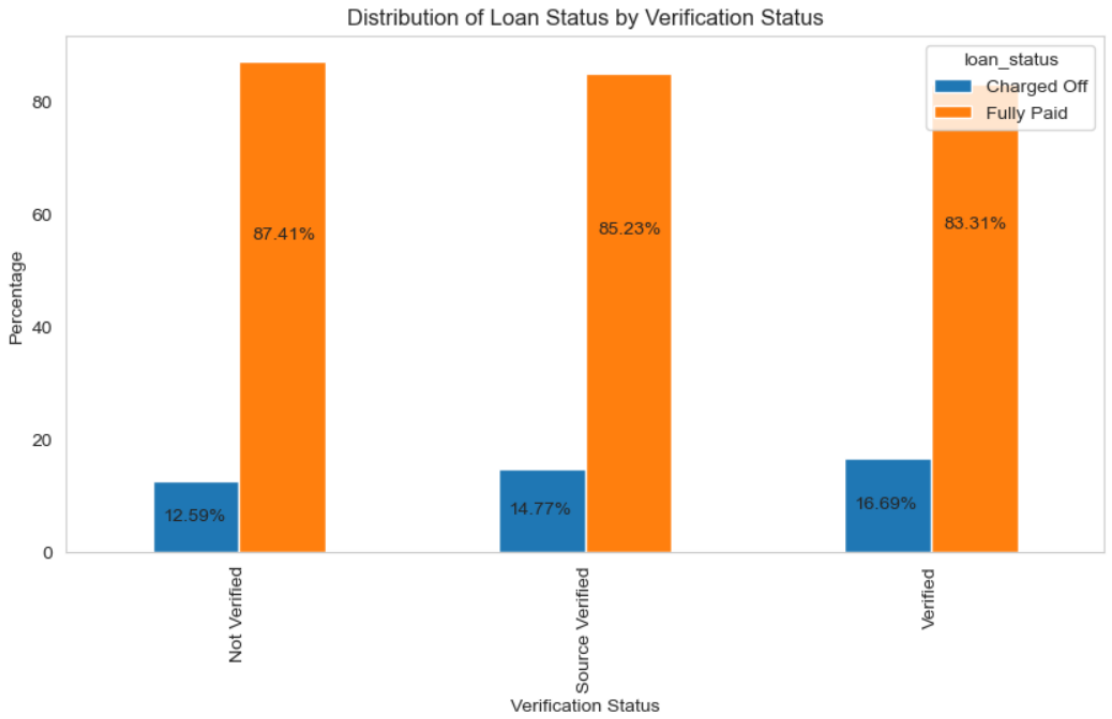


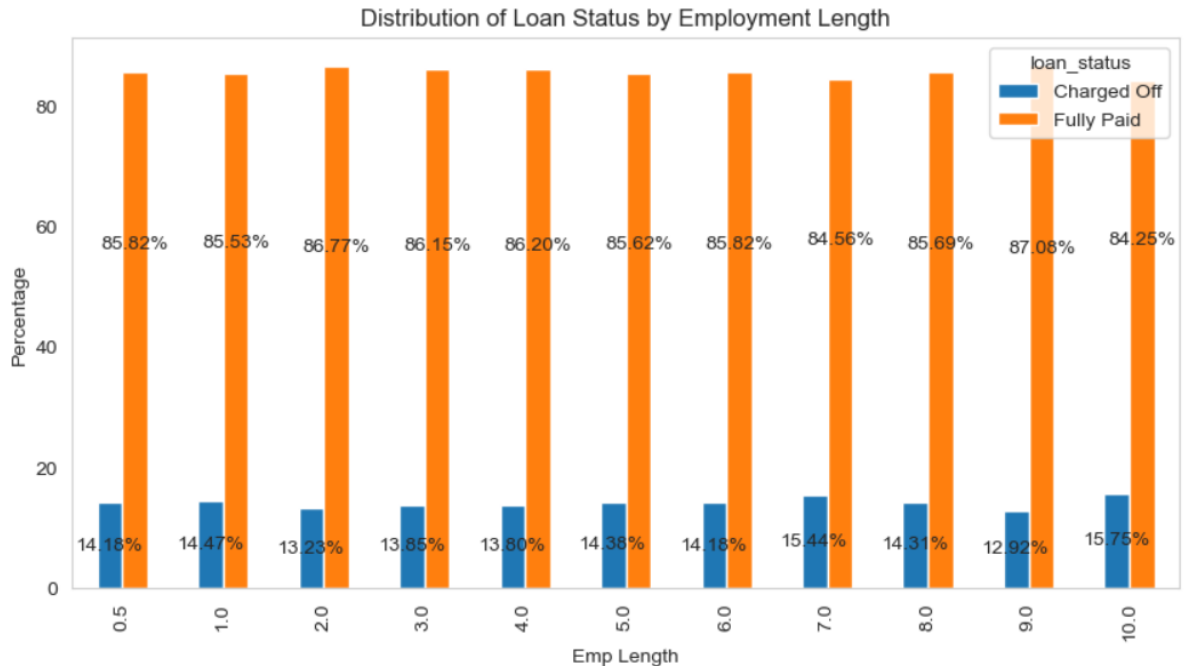


- The ratio of Fully Paid to Charged off loans shows borrowers with Higher salary are more likely to pay their loans back

- The ratio of Fully Paid to Charged off loans shows borrowers with Higher DTI are more likely to default on their loans

4b. SUA: Distribution of Loan Status by Verification Status & Employment Length

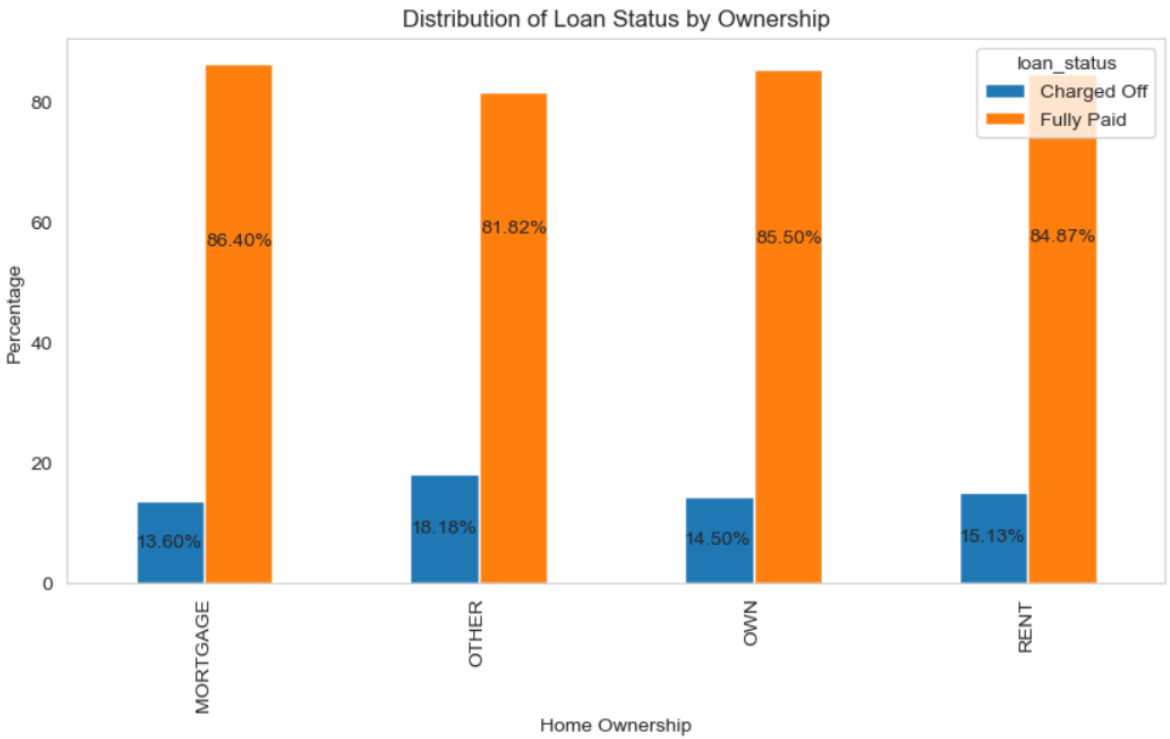




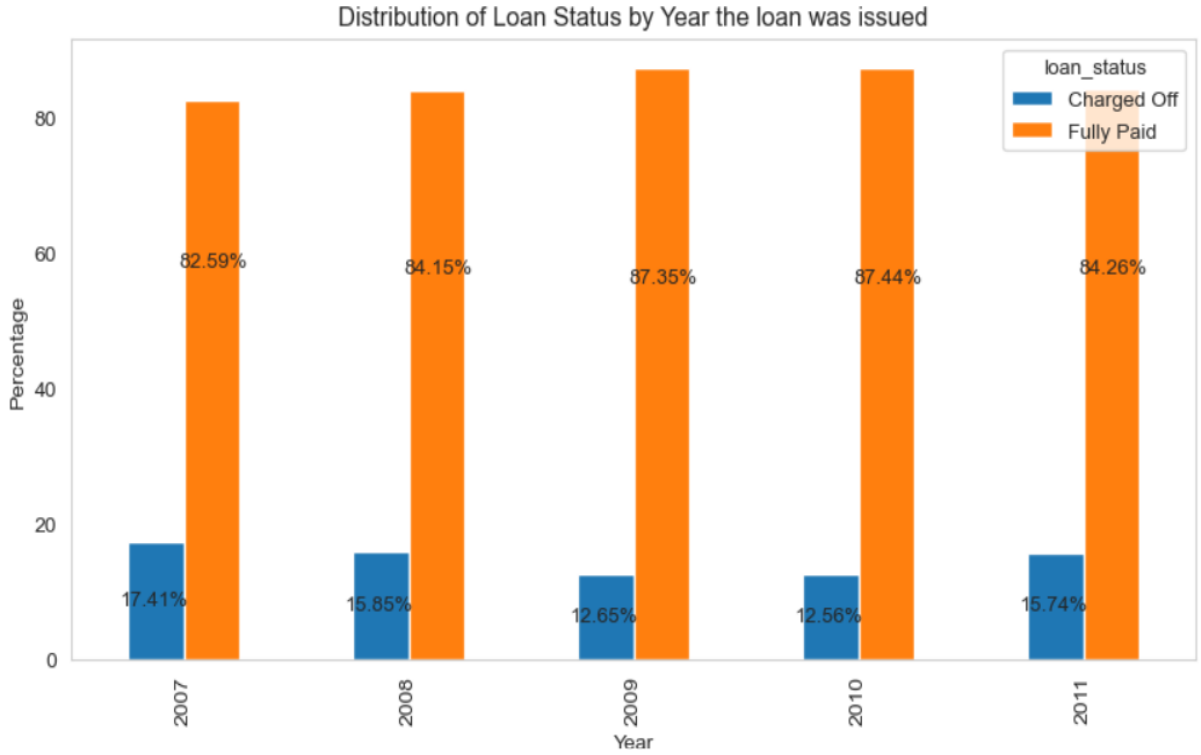
- The ratio of Fully Paid to Charged off loans shows little change for verification status

- The ratio of Fully Paid to Charged off loans shows no significant difference for various emp length.

4b. SUA: Distribution of Loan Status by House Ownership & Loan Issued Year

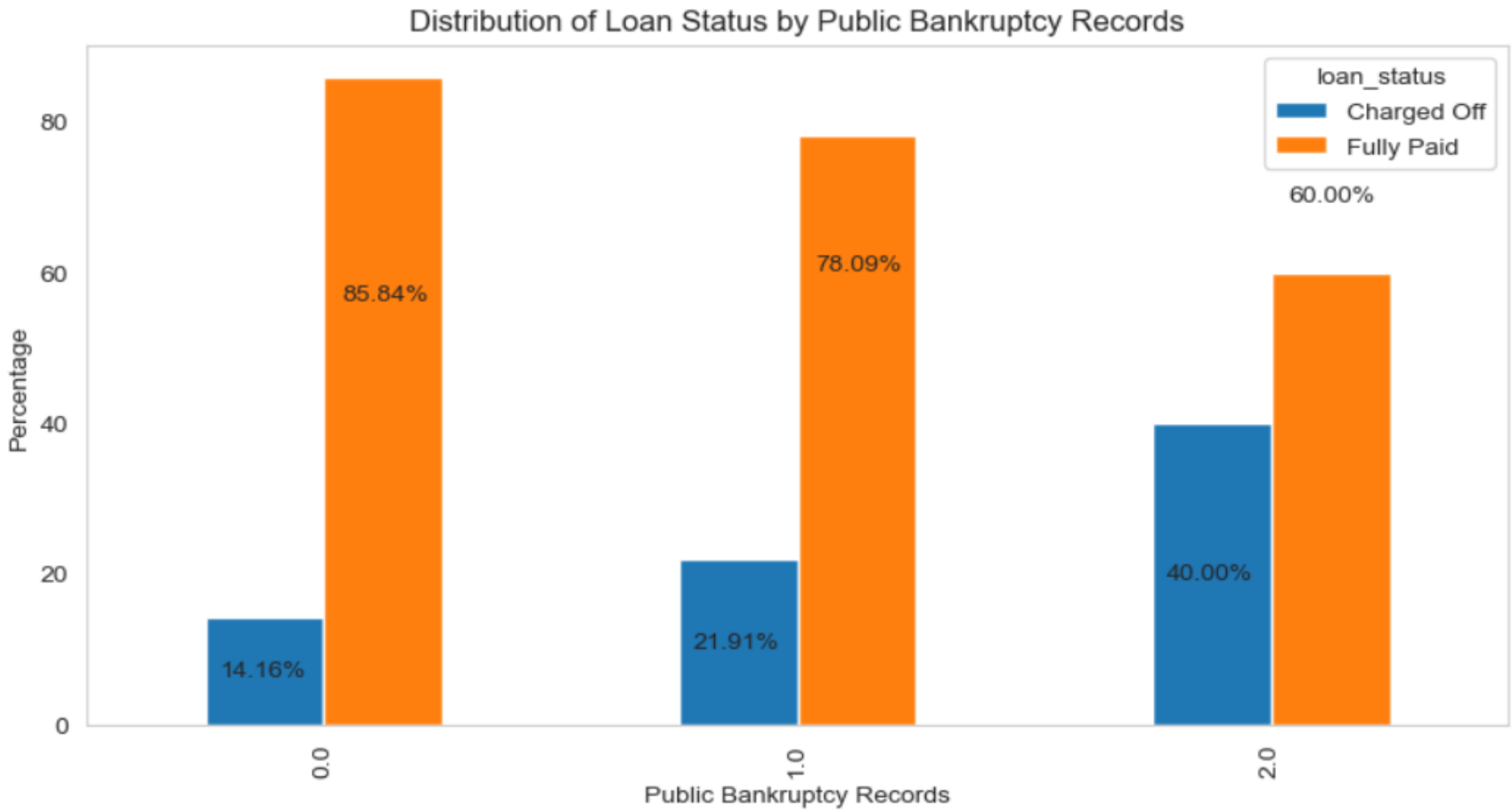


- The ratio of Fully Paid to Charged off loans shows no significant difference. A line chart to confirms the same (please refer to the python file)



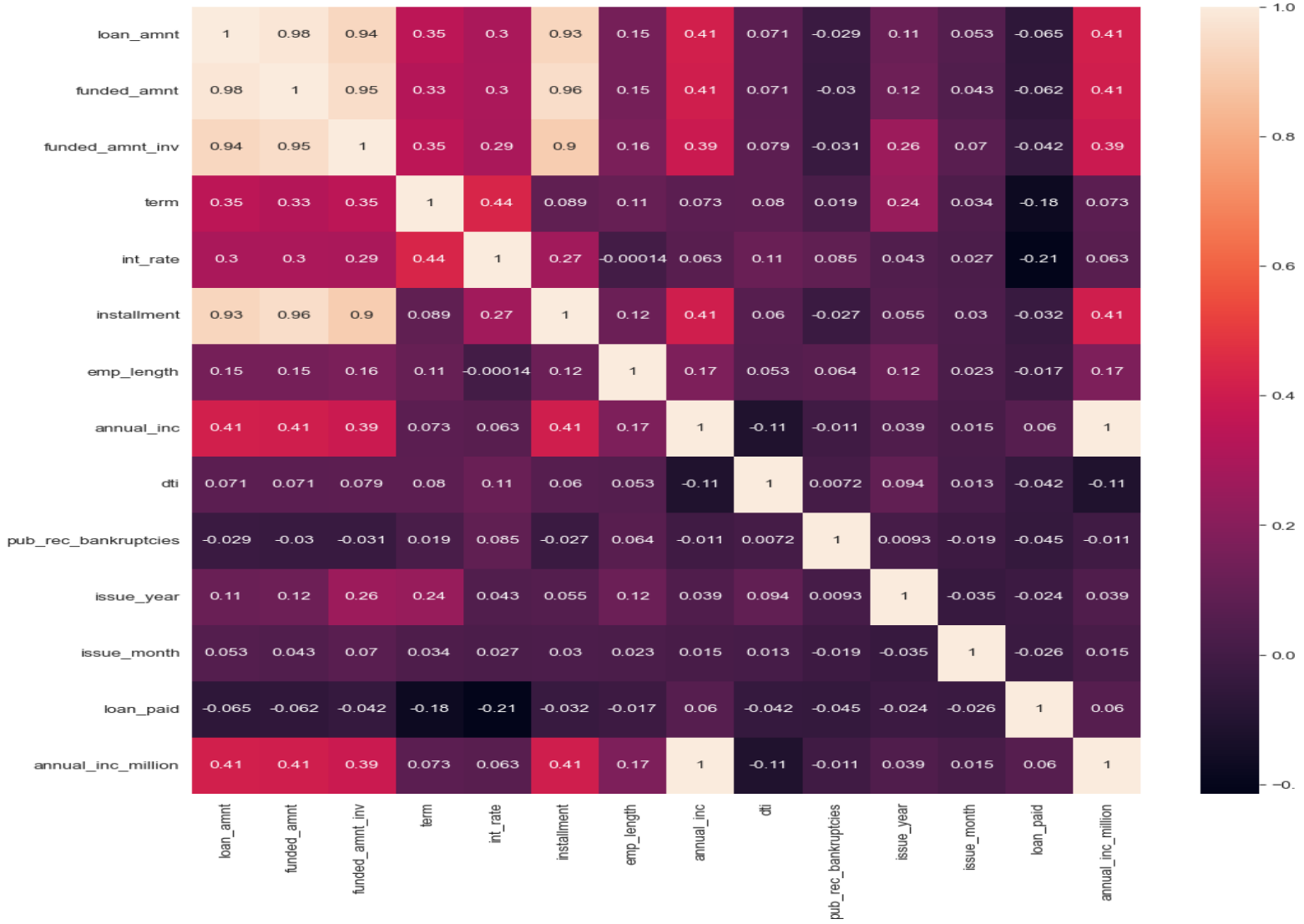
- The ratio of Fully Paid to Charged off loans shows variation by issued year. It was highest in 2007. It came down during the tough economic conditions of 2008, 2009.. Again went up in 2011. I don't want to see the absolute number is the no of loans given were way too high in 2011 as compared to other years.

4b. SUA: Distribution of Loan Status by Bankruptcy



- The ratio of Fully Paid to Charged off loans for borrowers with One or more bankruptcy record is very low. Avoid such borrowers.

4c. Bivariate Analysis: Heat Map



1. No very strong correlation found (+ve as well as -ve)

2. Loan being fully paid is:
   a. Negatively Correlated to:
      i. Interest Rate
      ii. Term

3. DTI is:
   a. Negatively Correlated to:
      i. Annual Income

3. Interest is:
   a. Positively Correlated to:
      i. Term
      ii. Funded Amount

4. Funded Amount is:
   a. Positively Correlated to:
      i. Annual Income
      ii. Term

5. Recommendations

For avoiding loan defaults:

1. Implement stricter criteria for approving loans:
   a) With longer terms (60 months) and higher interest rates (> 11%)
   b) With lower grades (e.g., Grade D to Grade G)
   c) For certain purposes like Small Business, Renewable energy etc.
   d) Issued in years that are difficult for the economy, like recession of 2008,2009

2. Implement stricter criteria for approving loans where borrower:
   a) Is from the state NE
   b) Has high DTI (>18.59)
   c) Has 1 or more Bankruptcy record

3. Debt consolidation accounts for 47.18% of all the loans, but has a high risk of default as well. Run campaigns to attract more such customers, but implement strict evaluation criteria to approve the loan as well

4. Explore strategies to incentivize borrowers to choose shorter terms and lower interest rates, potentially through promotional offers or discounts.

5. Continuously monitor and adjust lending practices based on changing market conditions and borrower behavior to optimize the lending portfolio and minimize default risk.

6. Recommend investigating the underlying reasons for the drop near 9-10% in the loan distribution curve and an unusual rise in the no of loans taken by borrowers with employment length > 10 years and assessing its impact on loan default rates.

5. Recommendations

For giving loans to people who will pay back based on our analysis:

1. Conduct targeted marketing or outreach campaigns for
   a) Larger states like CA, FL, NY, and TX to attract potential borrowers and expand the customer base
   b) Borrowers borrowing for Car, credit card, home improvement, major purchases and wedding

2. Tailor loan offerings or promotional incentives for borrowers from states like WY, ME, IN and LA after analyzing the exact reason of low default rates in these states

3. Analyze seasonal trends in loan applications and adjust staffing levels or operational capacity to accommodate higher demand during peak periods:
   a) Offer special year-end promotions or incentives to capitalize on increased demand and attract more borrowers
   b) Make sure that the staff doesn't overlook policies and compliances in the pressure of increased application volume