upGrad

LENDING CLUB CASE STUDY SUBMISSION

Group Members:
Nithin John Jacob

PROBLEM STATEMENT

A consumer finance company which specialises in lending various types of loans to urban customers, The company makes a decision for loan approval for each application 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.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

- **Loan accepted**: If the company approves the loan, there are 3 possible scenarios described below:

1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
2. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
3. **Charged-off:** Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan.

- **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.).

LOAN DATASET

Loan Accepted → Default

Loan Accepted → Non-Default

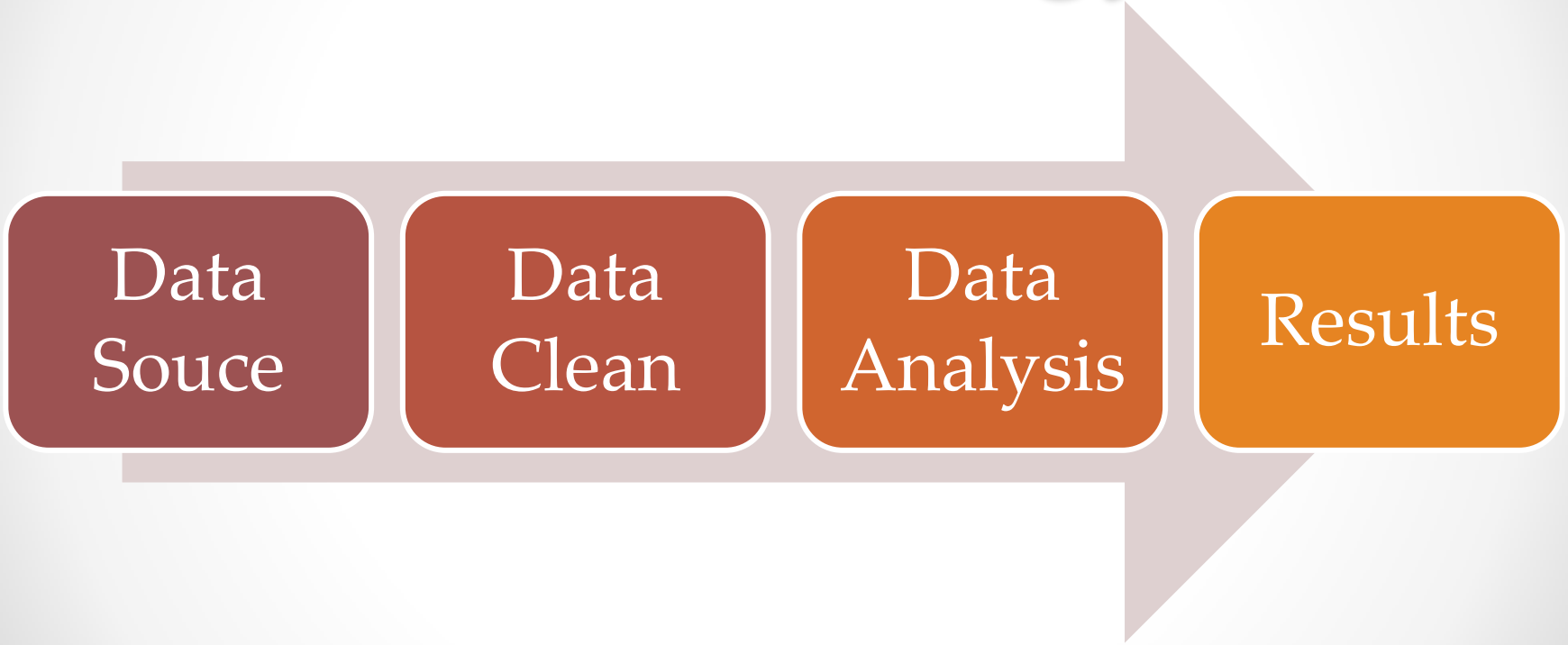
Loan Rejected

(Not considered in dataset)

Business Objectives

- The company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.
- Identification of risky loan applicants using EDA is the aim of this case study.

Methodology



Data Source

Import

- Loan Dataset
- Metadata

Explore

- 3917 rows
- 111 Columns

Meta-Data

- Description of each column

Data Clean

Rows & Columns

- Remove Empty Rows & columns
- Remove single valued Columns
- Remove other irrelevant columns

Missing Values

- Remove Rows having more than 50% missing values
- Remove columns having more than 5000 missing values
- Fix the missing value for each column by mean/mode/ other values/ ignore

Invalid Values

- Change Percentage columns to numeric from text
- Change date-time columns

Data Clean

Standardise Numerical

- Remove Outliers by Box plots and 95/98/99percentile

Standardise Categorical

- Title font
- Remove extra spaces

Filter Data

- Remove all data with loan status as ‘current’

Derived Metrics

- Create year, month, employee length columns

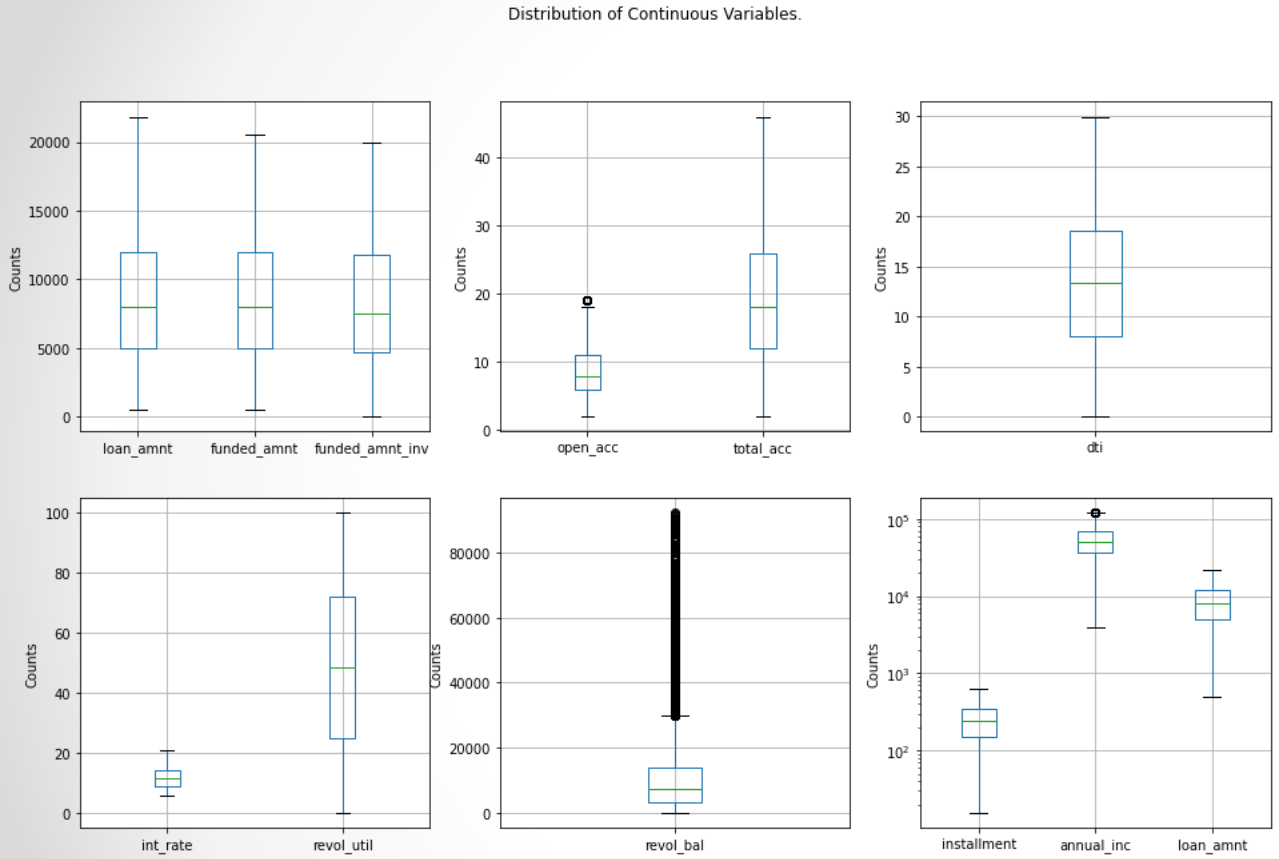
Data Analysis

Univariate
- Distribution plots
- Count plots

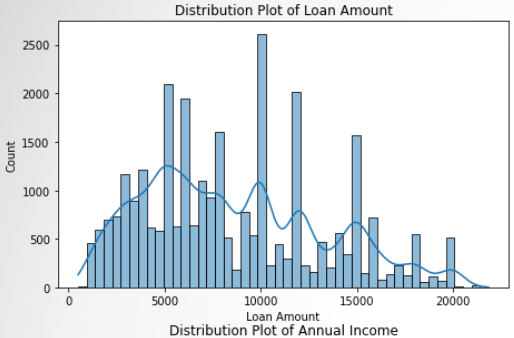
Segmented
- Box plot
- Bar plots

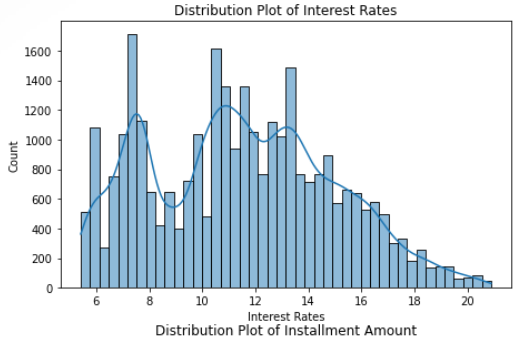
Bivariate
- Plots of data by segments

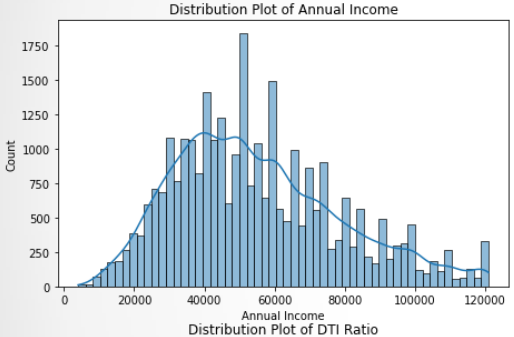
Results

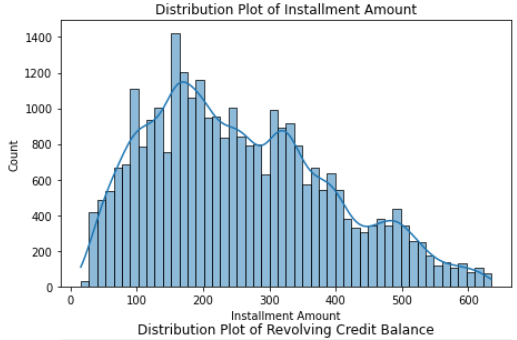


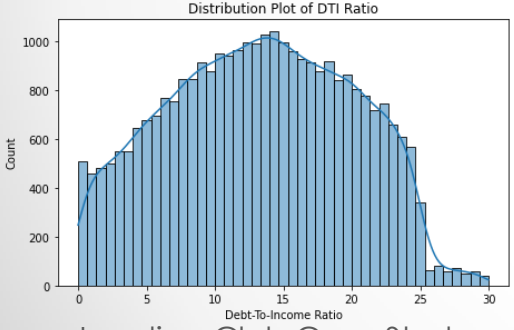
- 0.5% to 10% of total values were removed from each column.
- All columns except revol_bal has no outliers now.

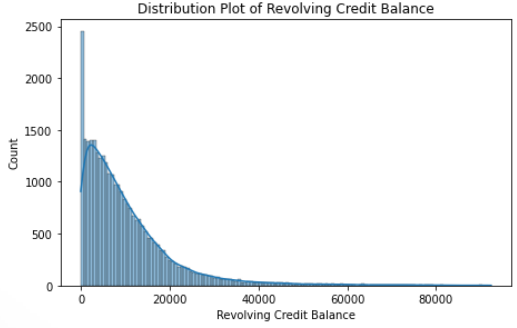






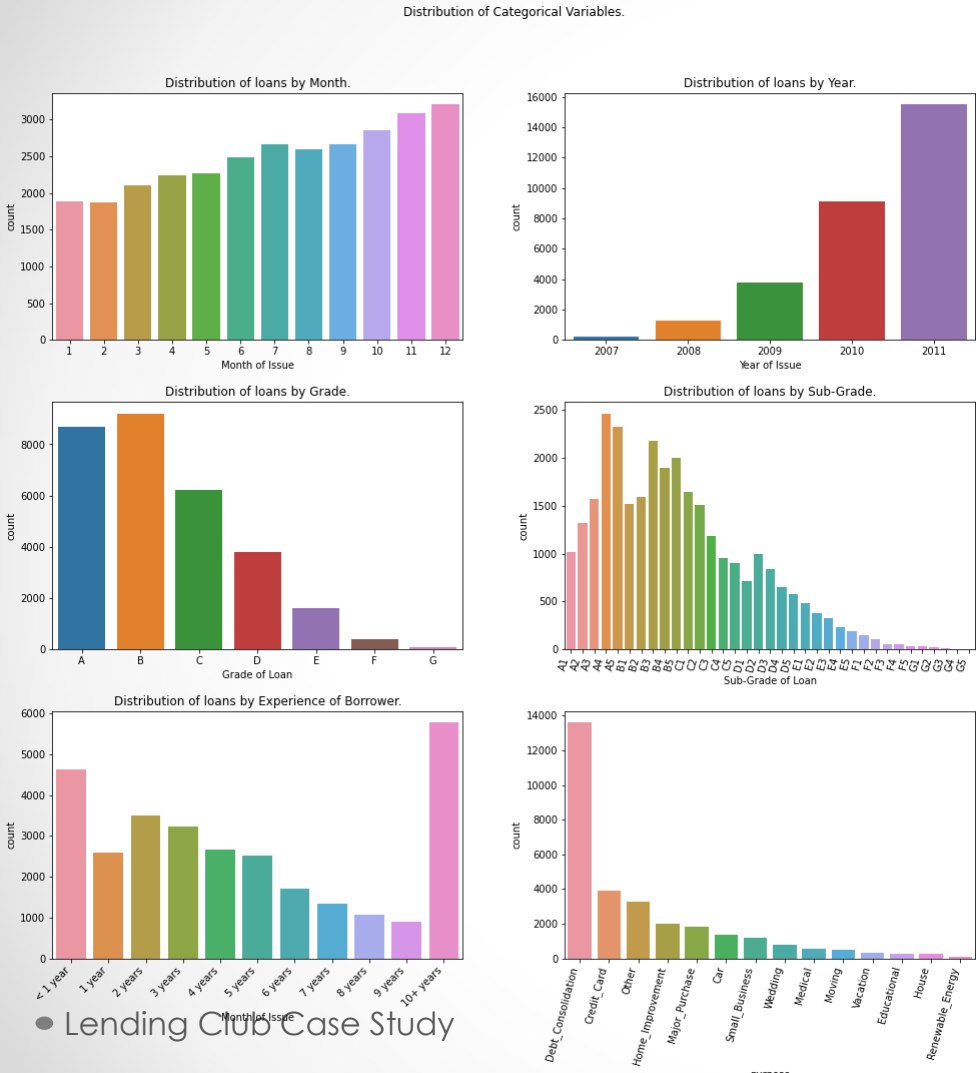






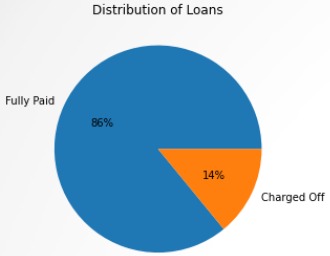
Distribution Plots

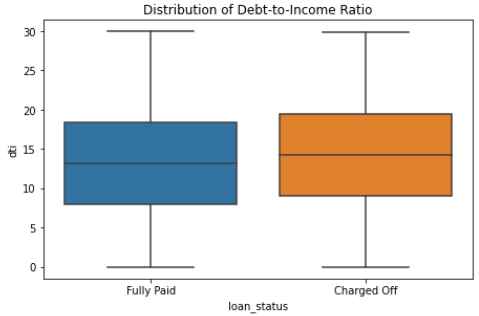
- Most plots are bell curved.
- Common interest rate offered are from 6-8% and 10-15%.
- Instalment amount is skewed towards 150 region.
- Annual income spike at around 50000.
- Most people don't use revolving credit

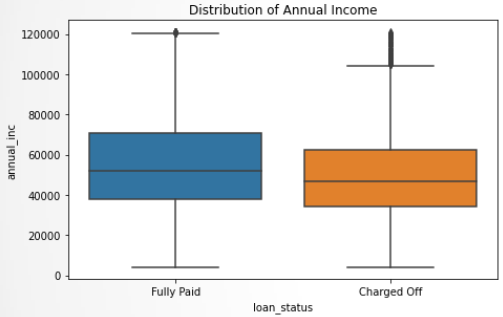


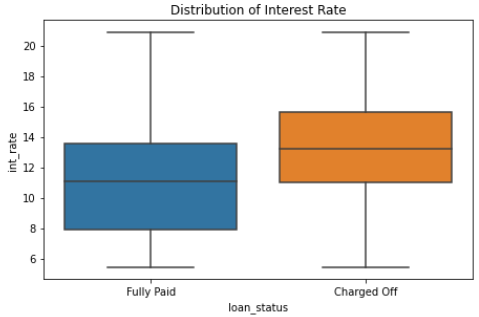
Results

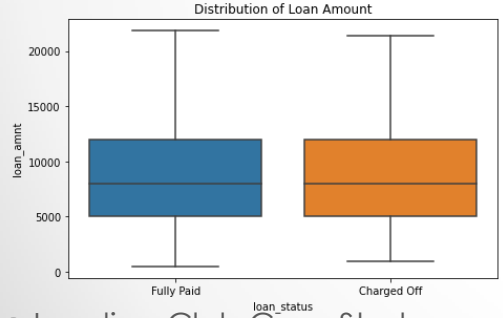
- More loans taken towards the end of a year.
- Loans taken increase every year from 2007-2011
- Most loans are graded A, B or C.
- Most loans are graded A4, A5, B3 or B5.
- Most loans to people with more than 10 years or less than 1year experience.
- The most common reason for loan is Debt Consolidation. The second is credit card.

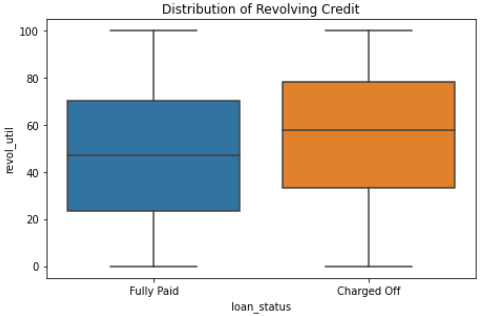






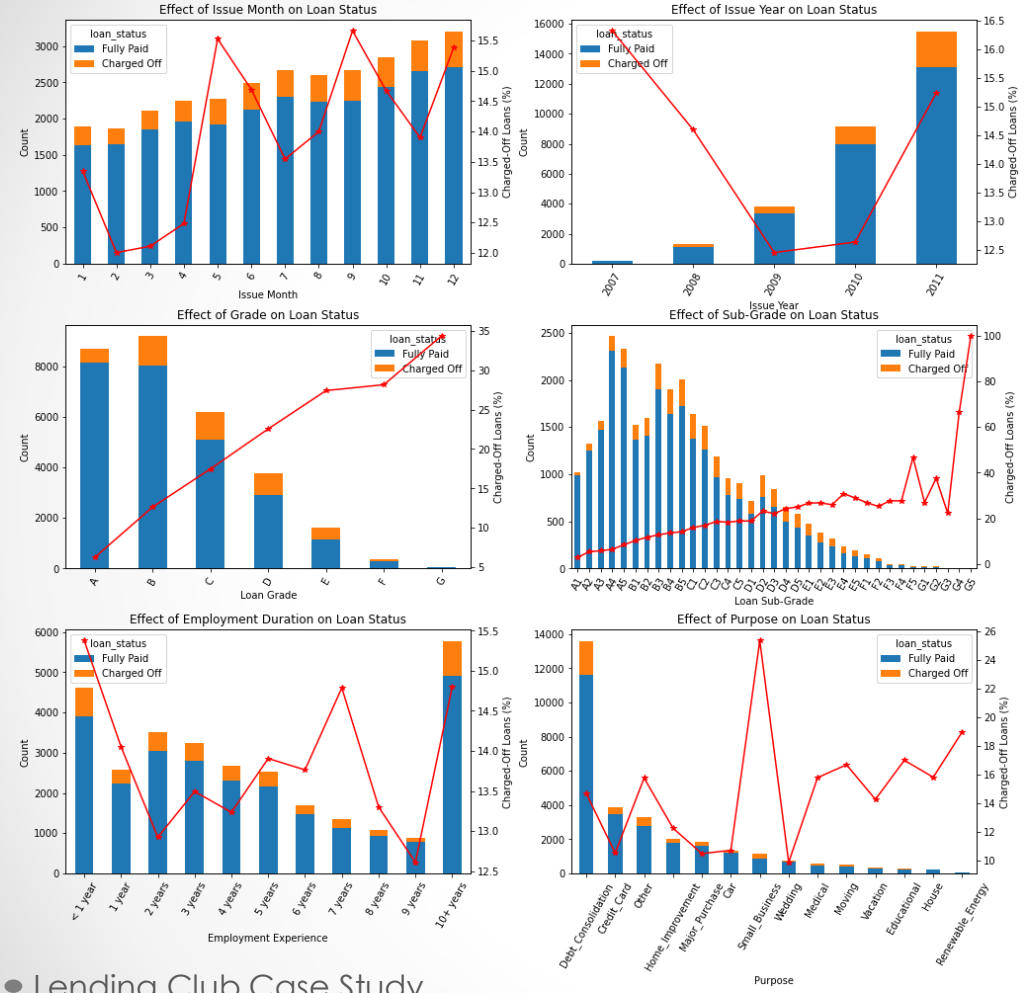






Box Plots

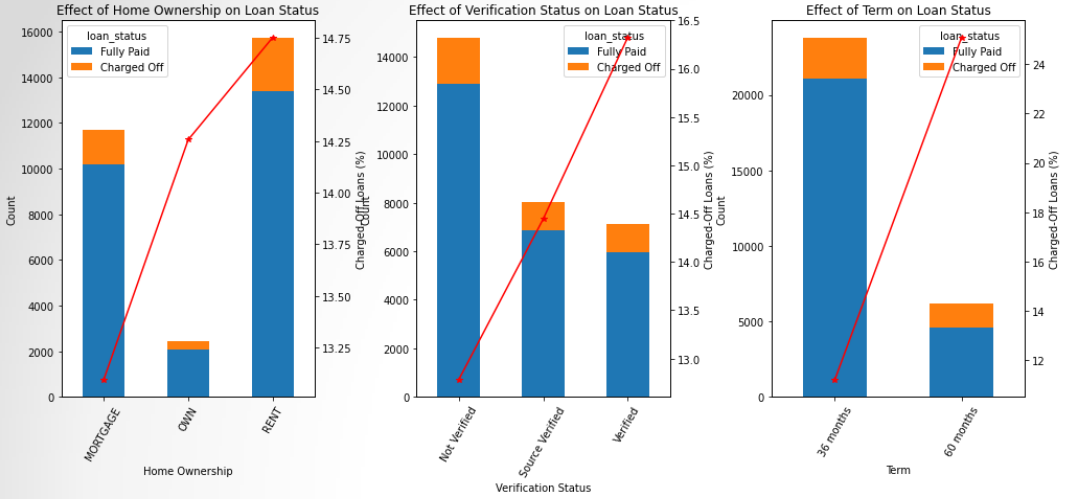
- 14% of Loans are defaulted.
- Higher debt to Income Ratio tend to default.
- Higher annual income tend to pay back fully.
- Higher interest rate tend to default.
- Loan amount doesn't have any impact on default.
- Higher revolving credit tend to default.

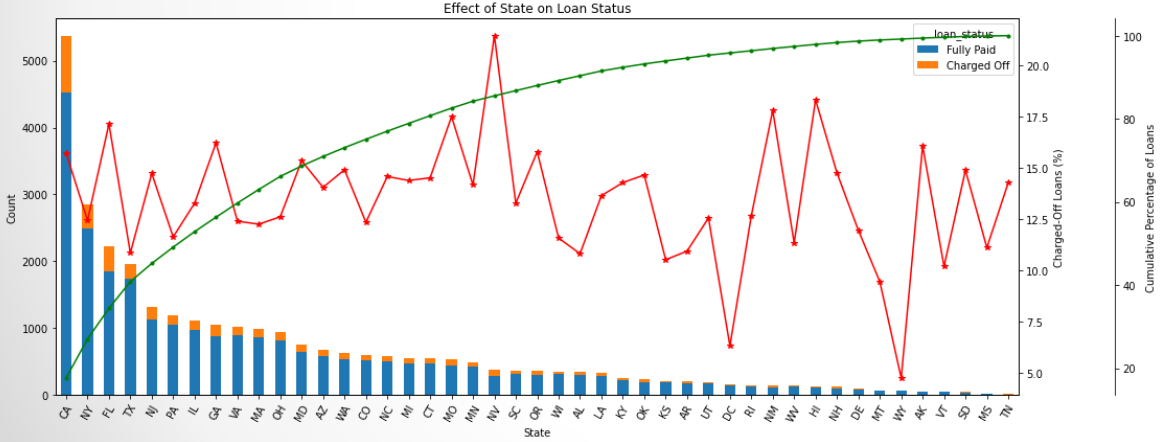


Bar Plots

with Percentage of Charged-OFF

- The default rate spikes at May, September and December.
- The default rate is higher towards 2011.
- Lower grade loans has higher tendency to default.
- This effect is seen in subgrades also.
- Less work experience are ore likely to default.
- Loans taken for small business has higher chance of default.

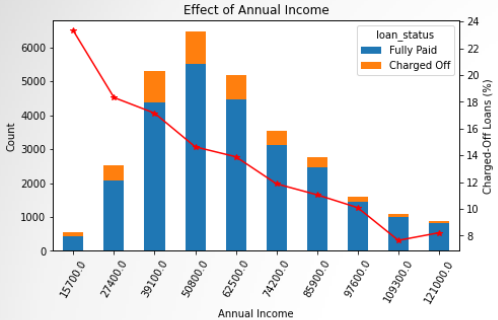


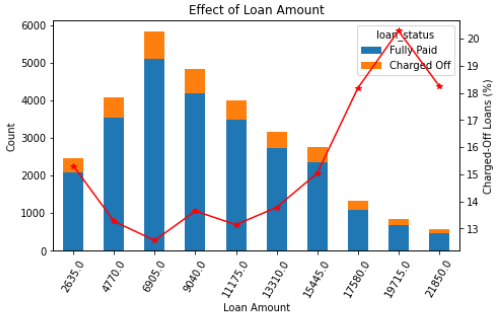


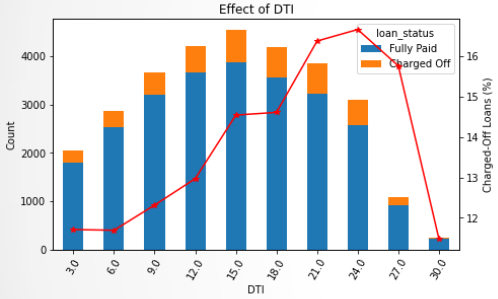
Bar Plots

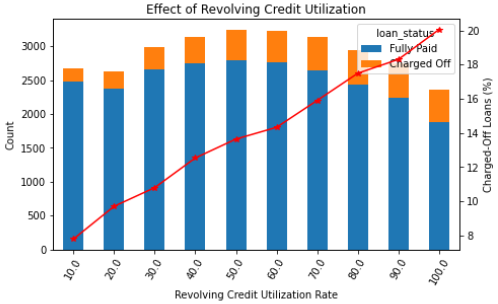
with Percentage of Charged-OFF

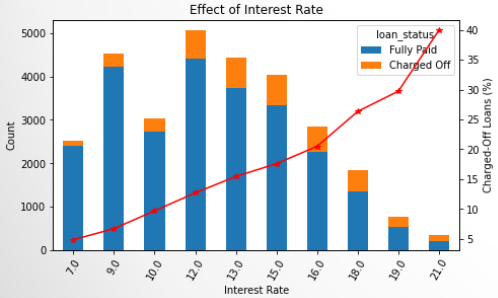
- People who own homes are less likely to default compared to who say on rent.
- Verified people have higher chance for default.
- Loans for 60 months are much more likely to default than ones with 36 months duration.
- 80% of loans come from 16 states. Loans from state NV seems to have higher probability of default.

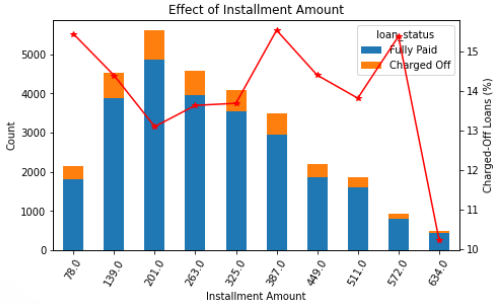










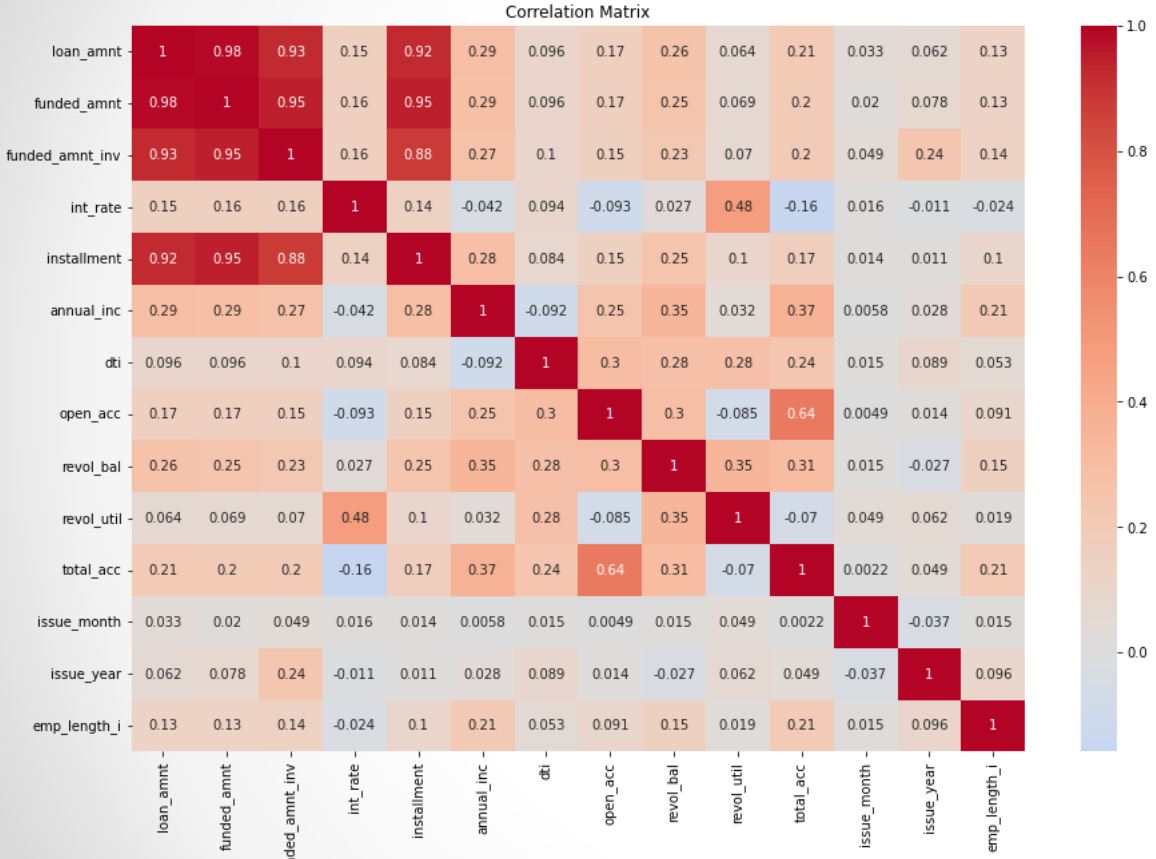


Bar Plots

with Percentage of Charged-OFF

- Higher average income are less likely to default.
- Higher loan amount are more likely to default.
- Higher DTI are more likely to default.
- People who use more revolving credit are more likely to default. It is a strong indicator of default.
- Higher interest rate are likely to default.
- The trend is not clear for installment amount.

Correlation Plot



- Loan amount, funded amount funded amount by investor shows a high level of correlation.
- Installment amount has good correlation with loan amount.
- Interest rate is correlated with revol_util.

Recommendations

Safe

- The **purpose of Weddings ,major purchase, car and credit**
- The lo**an amount** ranges between **5000 and 12000**
- **Annual income greater than 100000**
- **DTI** is **less than 12%**
- High **grade** like **A and B** is recommended.
- Home owners.
- Short **term** loans of **36 months**

Risky

- The purpose is **Small Business** as the percentage of a loan being charged off is high 25%.
- Higher the loan amount, the higher the chances of loan being charged off.
- **DTI** is **more than 12%.**
- Higher interest rate. Greater than18% is very risky.
- Low grades loans **E,F and G**.
- Rented Homes
- Long **term** loans of **60 months**

Thank You