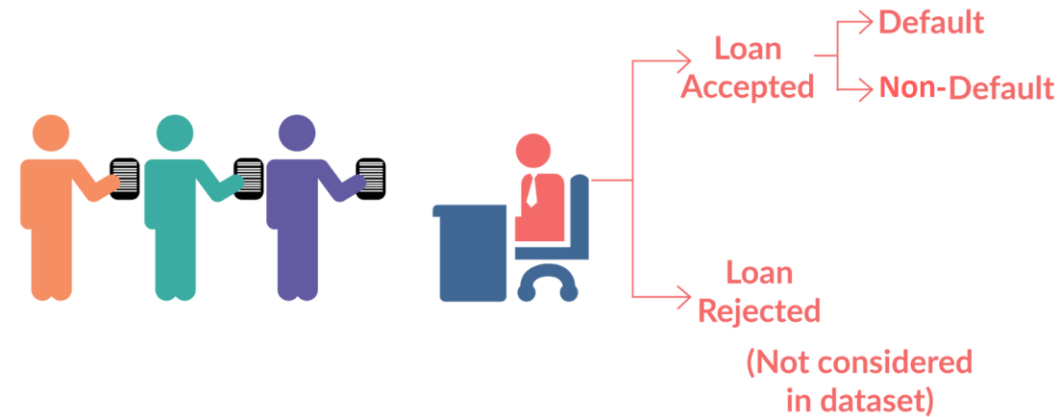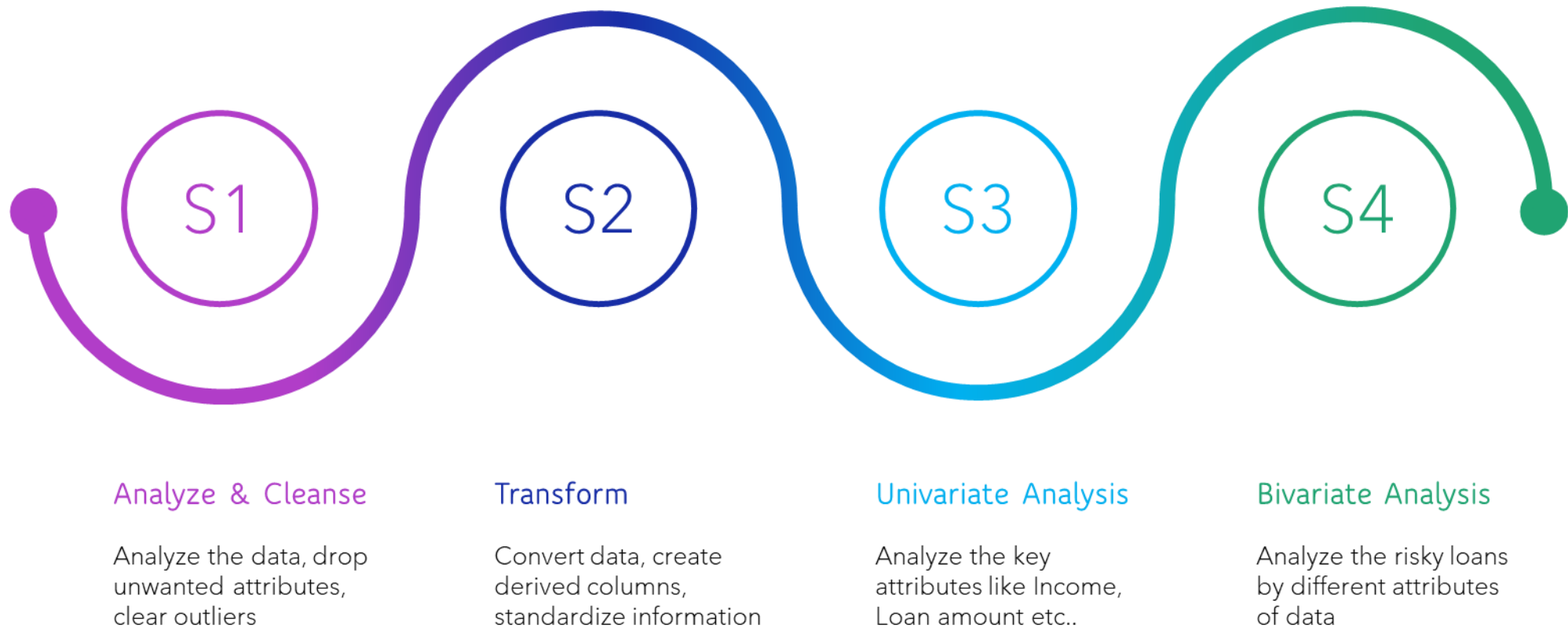# Lending Club Case Study

Ajay Rawat

Ravichandra Kolluru

Problem Statement

Use historic loan data to identify risky applicants whose loans can be risky what action contribute to reduced risk like reduced loan amount

Approach

S1 — Analyze & Cleanse

Analyze the data, drop unwanted attributes, clear outliers

S2 — Transform

Convert data, create derived columns, standardize information

S3 — Univariate Analysis

Analyze the key attributes like Income, Loan amount etc..

S4 — Bivariate Analysis

Analyze the risky loans by different attributes of data

Libraries used

Pandas

Numpy

Matplotlib

Seaborn

Plotly

Analyze & Cleanse

- Identify the columns that have nulls in all rows or have only one value for all rows as they can't support analysis (60+ columns)

- Drop columns with high volume of Nulls (next_pymnt_d, mths_since_last_record)

- Drop columns with unique values as they can't help analysis (id, member_id, url)

- Drop Columns with more than 50% of unique values.('emp_title','title','desc')

- Remove % sign from column and convert column type to Float

- Loans with status of Current as still active and a decision/conclusion can't me made of the effectiveness for the analysis of the same and so excluded

- Outliner analysis done for Loan amount and income amount, but could not find a reason to remove any data even if outlier because it can hide certain information suppress good or bad news

Transform

- Remove percentage from interest rate and revol util to convert to float and create an integer column

- Populated the nulls in employee tenure with mode for the 2.7% of rows missing values

- Date columns Issue Date and Earliest Credit line are converted to dates, created year column for both and month column for Issue Date

- Calculated the length of credit for row at the time of load issue date

Univariate Analysis

- Purpose – Most of the loans are for debt-consolidation

- Verification status – Non-Verified status has highest number of loans

- Home Ownership – Less number of home owners take loan compared to Rent or Mortgage

- Issue Month – Increase in number of loans processed end of the year. Decrease in number of loans during May 2008-Dec 2008 in line with financial and real estate crisis

- Years of Credit line – More applications for loan from people with 5-17 years of credit line

- Grade – A, B & C grades highest number of loans

- Sub Grade – A3, A4, A5, B3, B4, B5 have the highest number of loans

- Interest – Most of the loans fall under 5-16% interest

Univariate Analysis – Visual Analysis
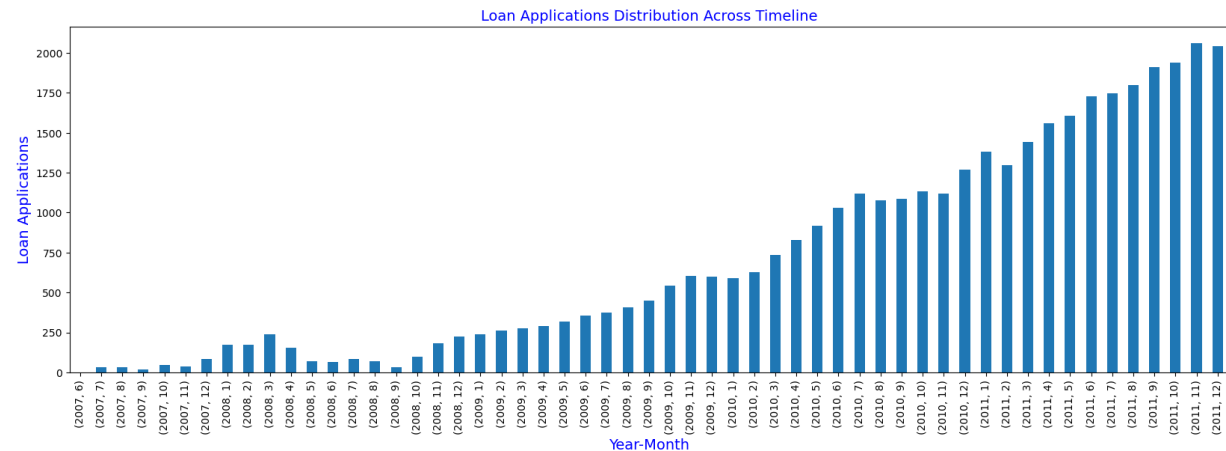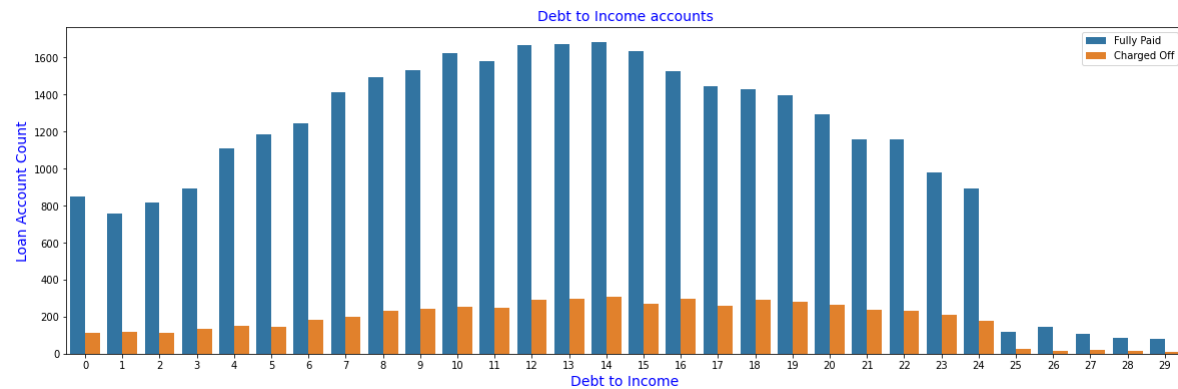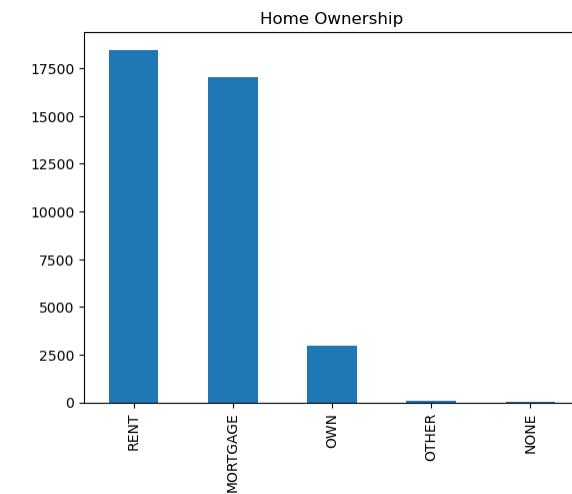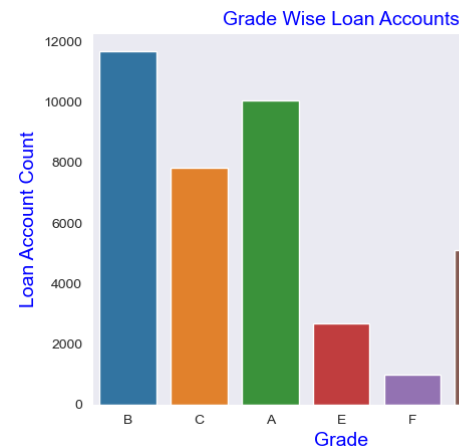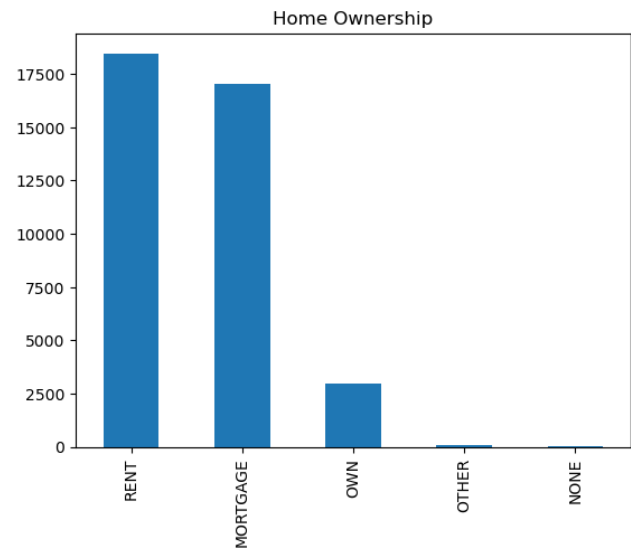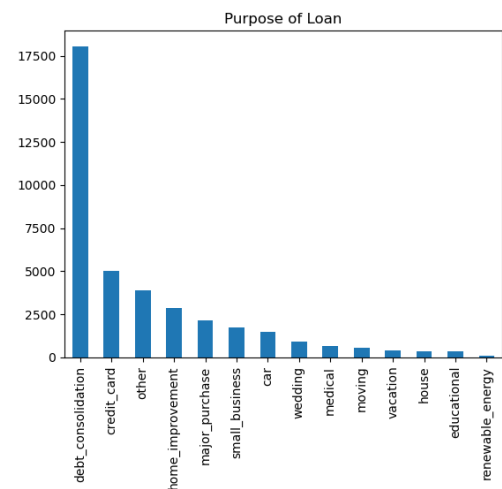
Bivariate Analysis

- Purpose – small_business have highest risk of default

- Verification Status – Does not seem to have major variance in risk, but non-verified has less risk which may not be logical (but data inference is that)

- Home Ownership – Home ownership seem to have no impact on risk

- Grade – A Grade has lowest level of risk and risk increased by the Grade

- Debt to Income – Higher the debt to income, higher the risk of default

- Interest – As the interest rate increases the amount of risk increases significantly

- Years of Credit – Most loans to applicants with 10-12 years of credit history, with higher number of years of credit with lower risk of default

- Loan Amount to Income Ratio – Risk increases with Loan amount to Income ratio

- Loan Term – Increase in risk with 5 year loans compared to 3 year loans.
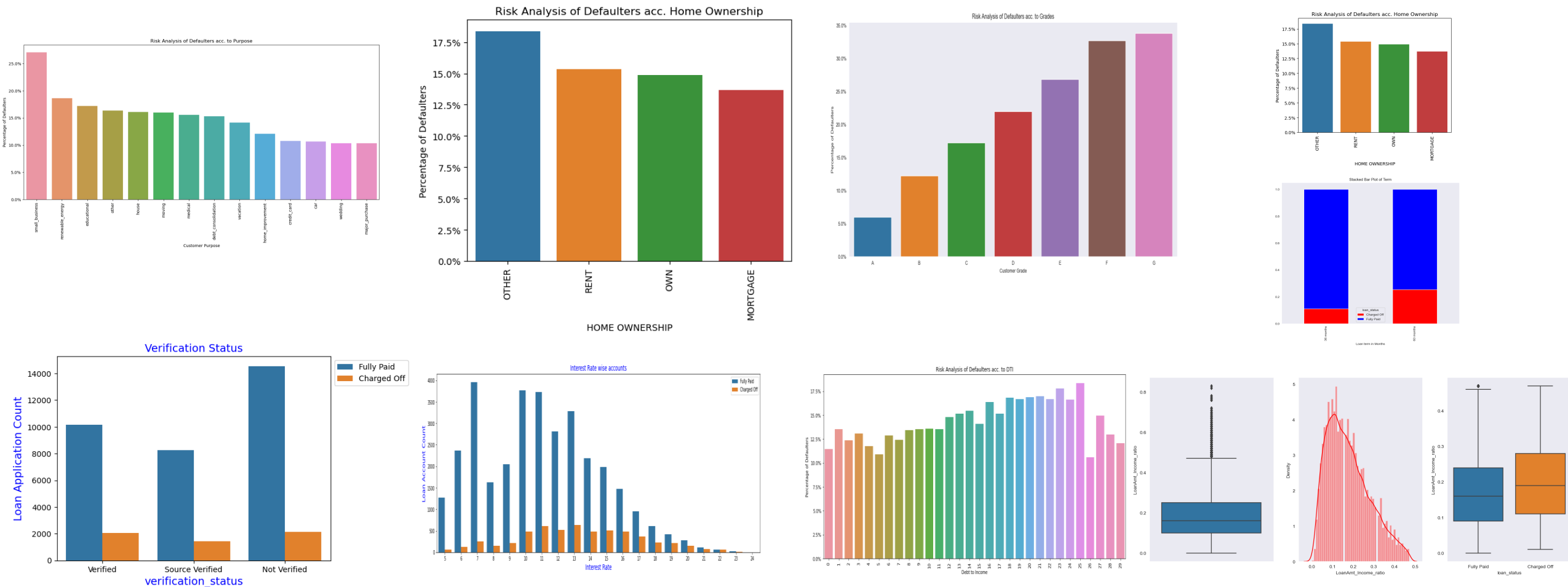
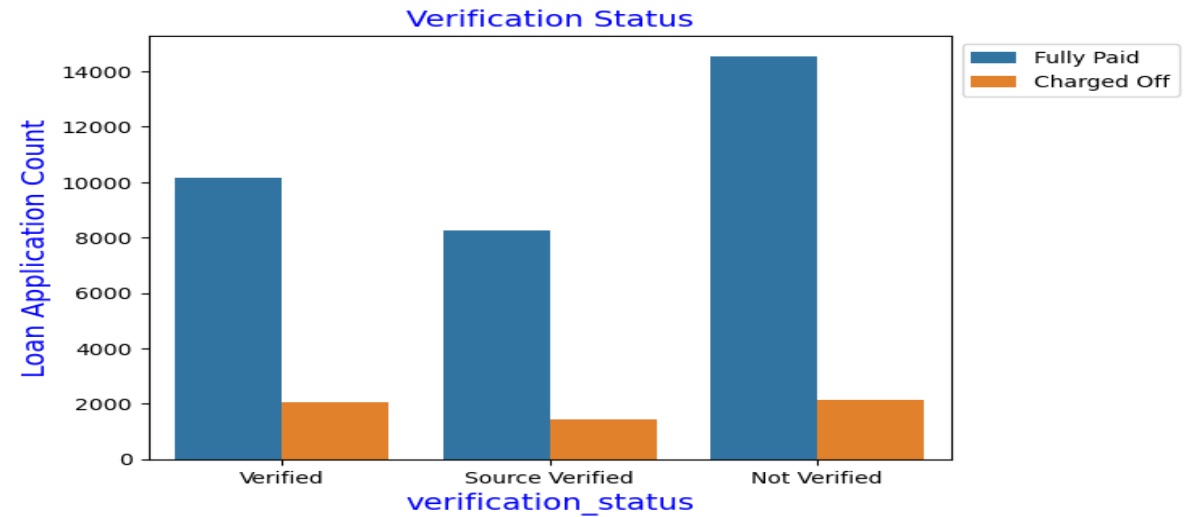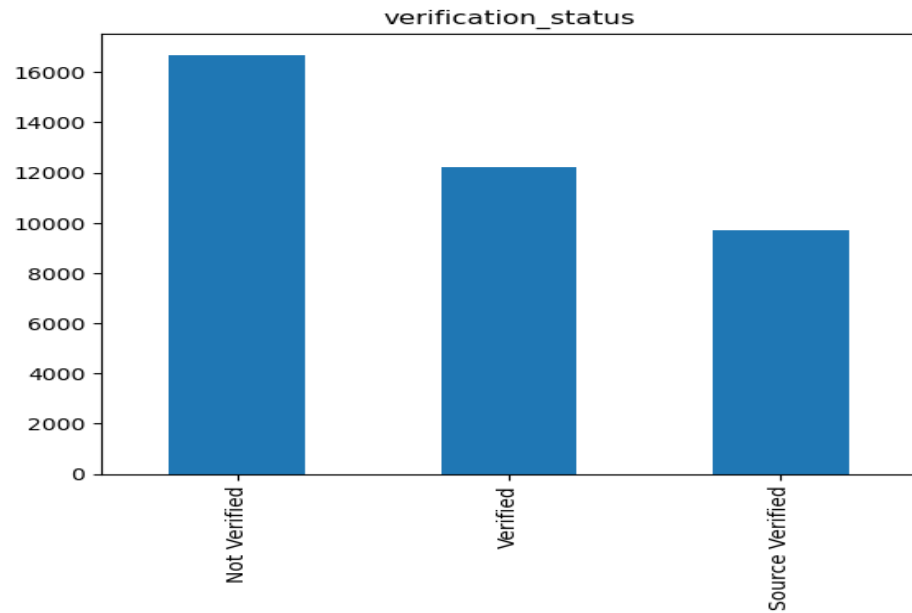# Bivariate Analysis – Visual Analysis

1. Purpose Of Loan
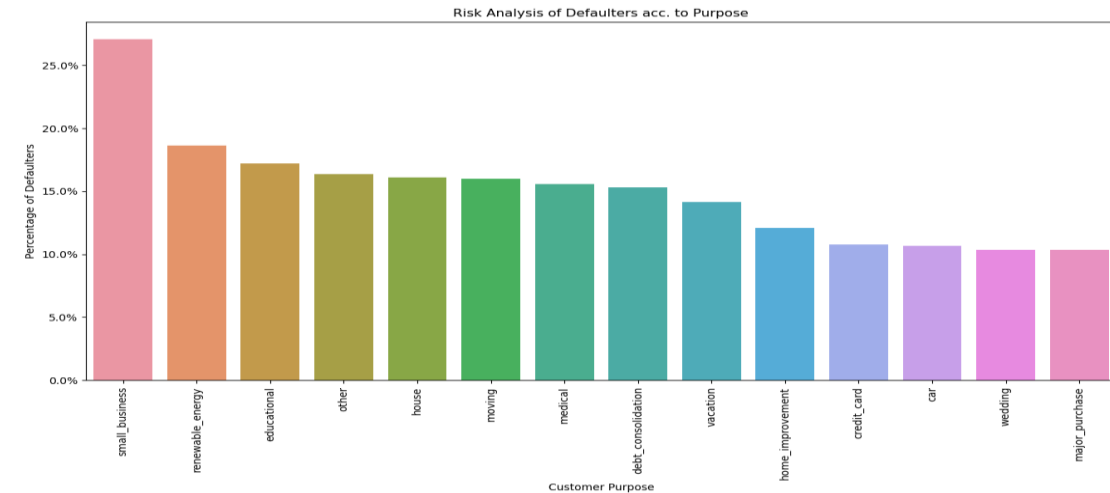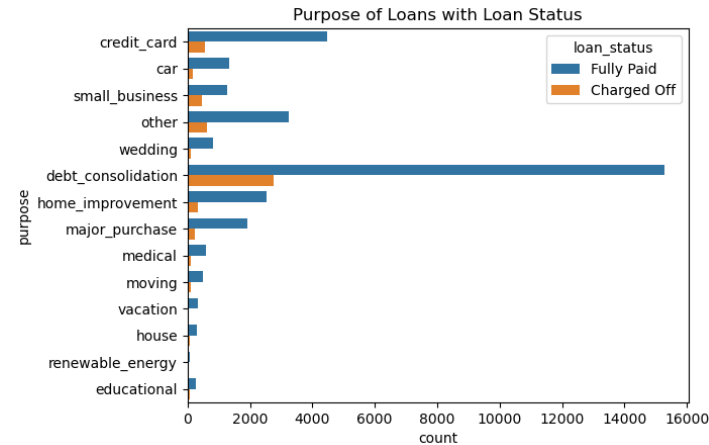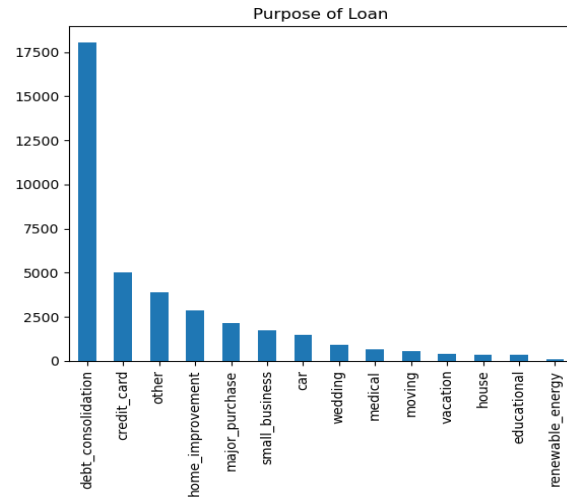
Observation:

1. Maximum Loans are taken for Debt Consolidation
2. Debt Consolidation also has highest number of Charged Off Loans
3. Highest percentage of defaulters are from Small Business then Renewable Energy and then Education.

2. Loan Verification Status

Observation:
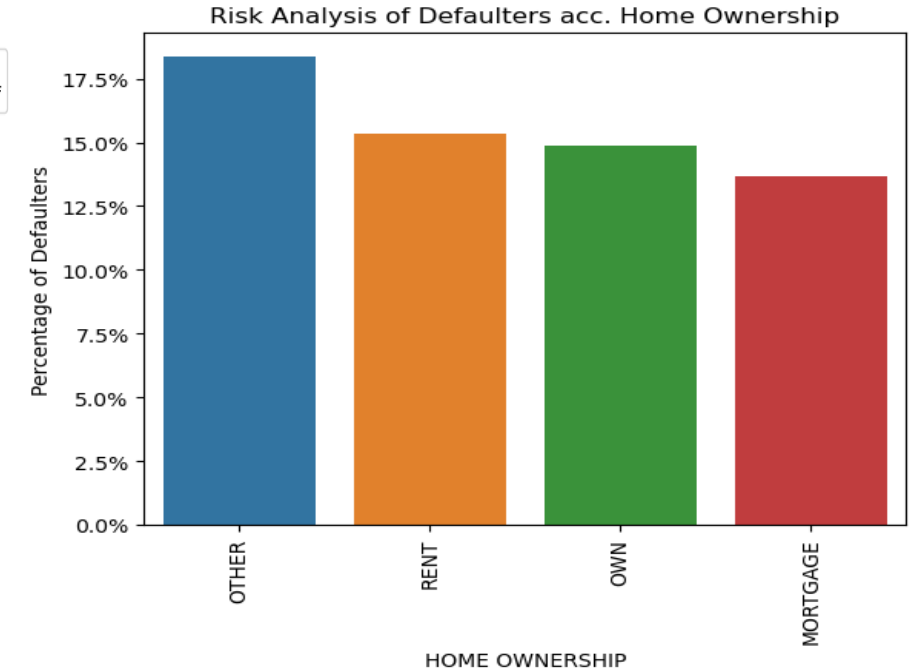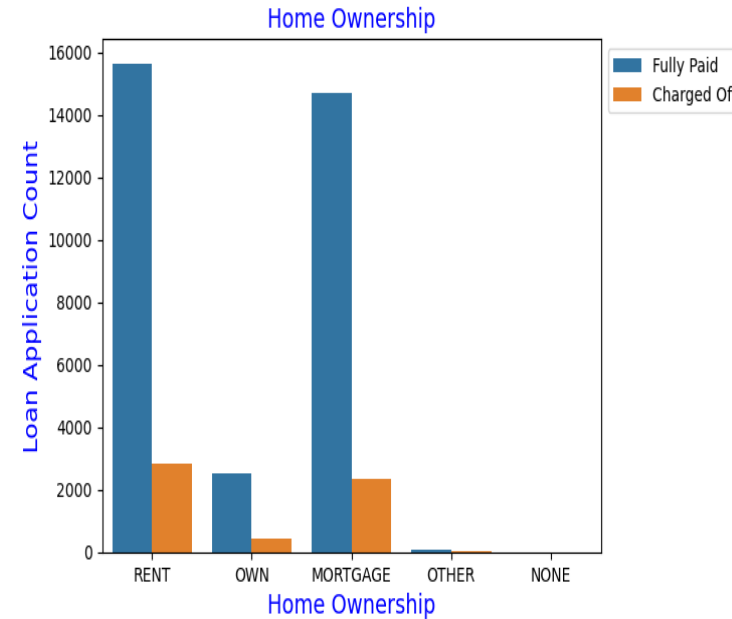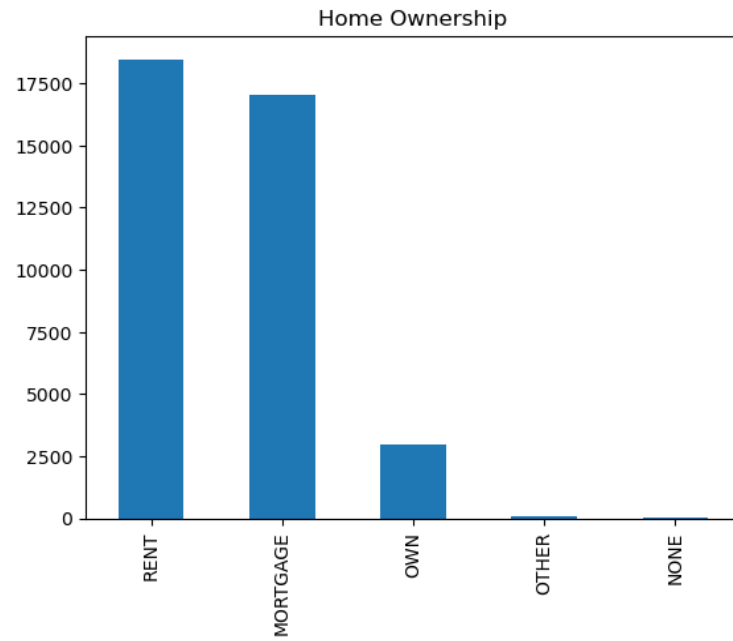
1. We cannot draw a conclusion of impact of verification status on loan status

3. Home Ownership
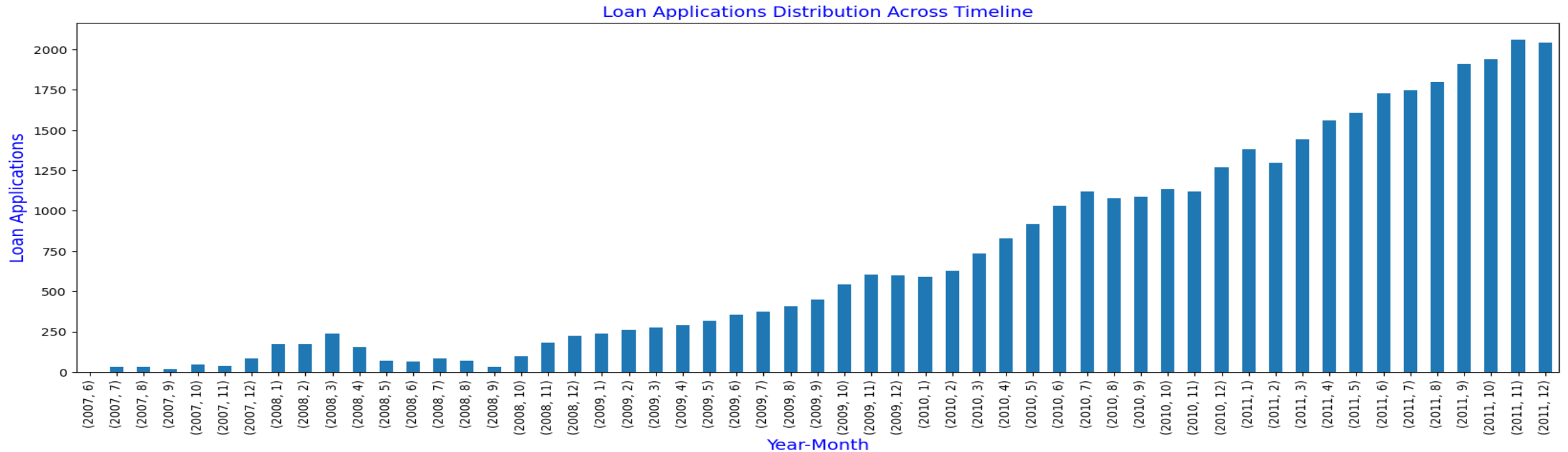

Observation:

1. 47% are applicants stay on rent 44% applicants stay in Mortgage 7% applicants own a house Other can have variety of data hence ignoring it .
2. Around 15% of applicant staying on rent default on their payment.

4. Loan Issued Timeline


Loan Applications Distribution Across Timeline

Observation:

1. There is an increase in the number of loans processed every Year.
2. There is decrease in Loan processed from May 2008 - Dec 2008. Probably due to Recession.
3. Highest number of loans are issued in November and December. Probably due to some promotion campaign during festive season

5. Impact of duration of Credit Line
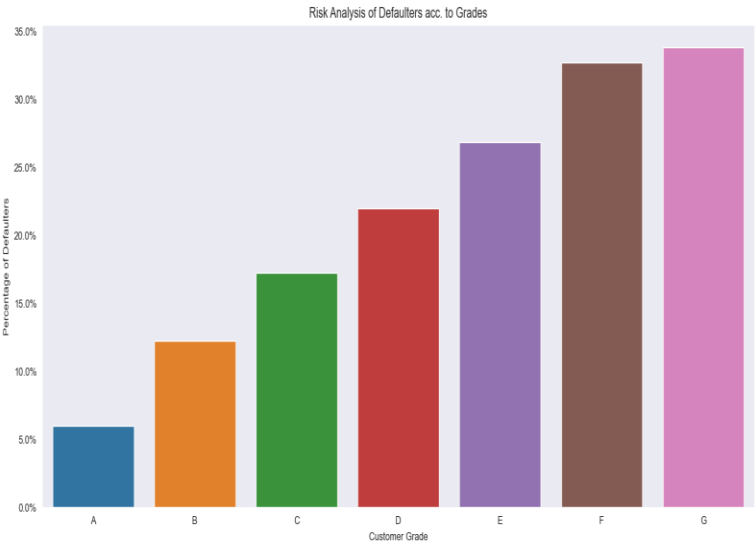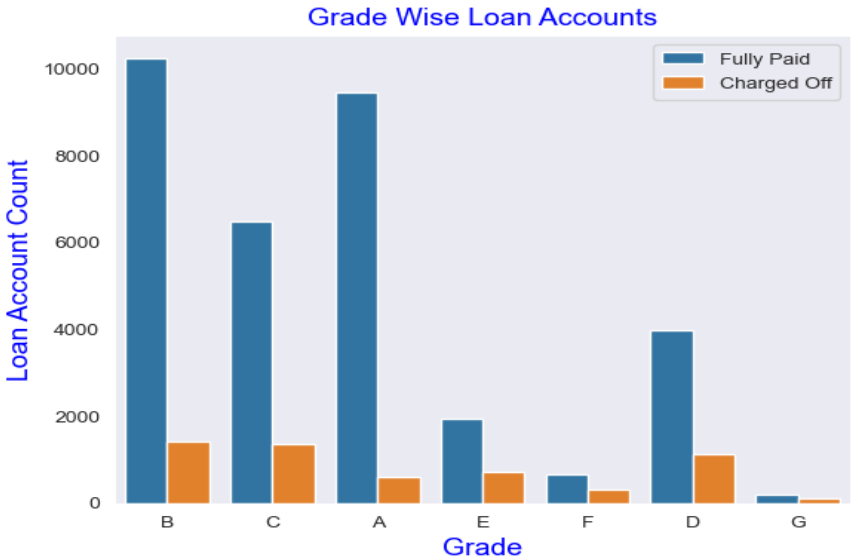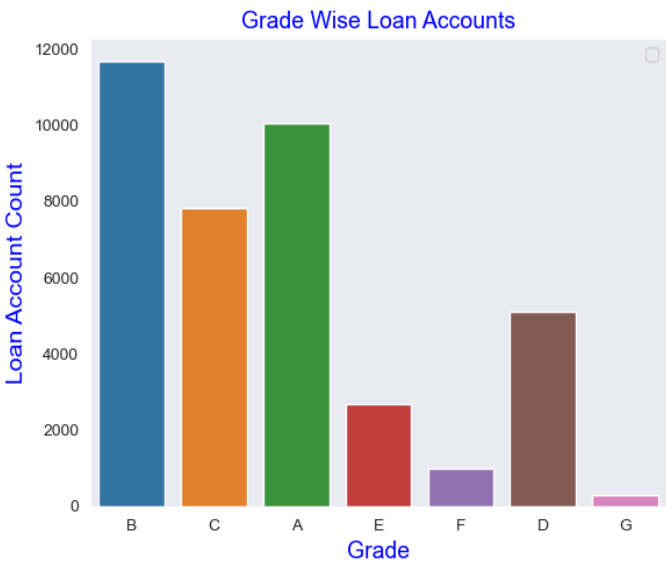
Years of Credit Vs Loan Status
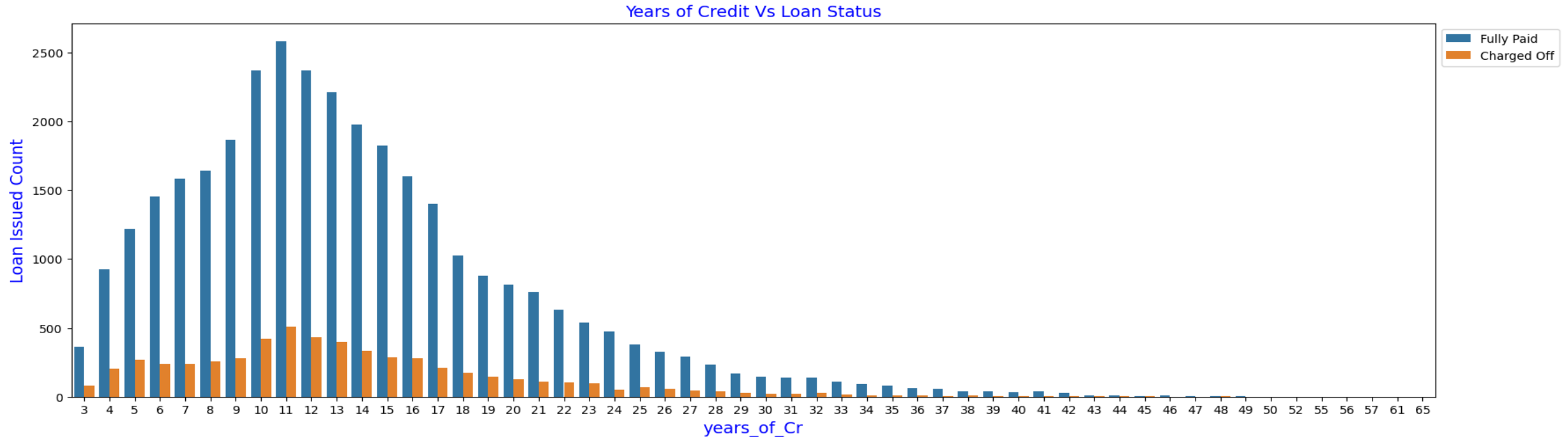
Observation:

1. Higher the number of Years from earliest credit line , lower the probability to default.
2. Maximum number of loans are issued to applications with 10-12 years of credit line history.
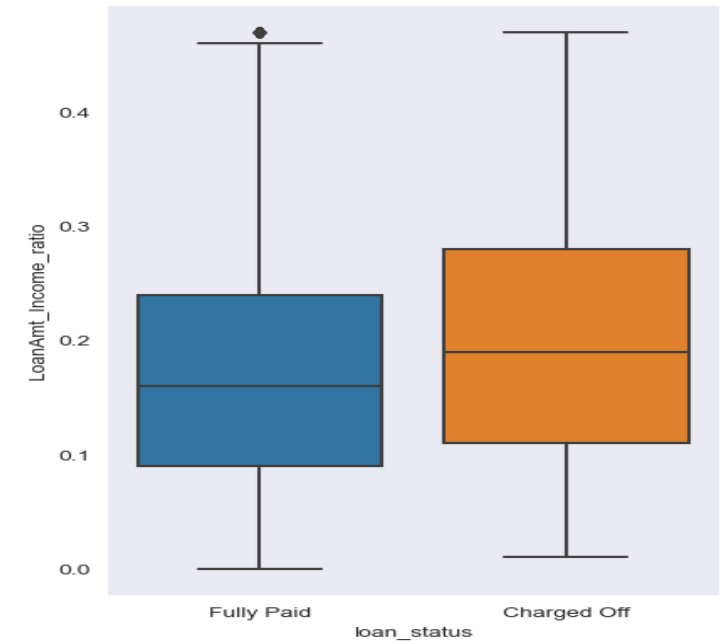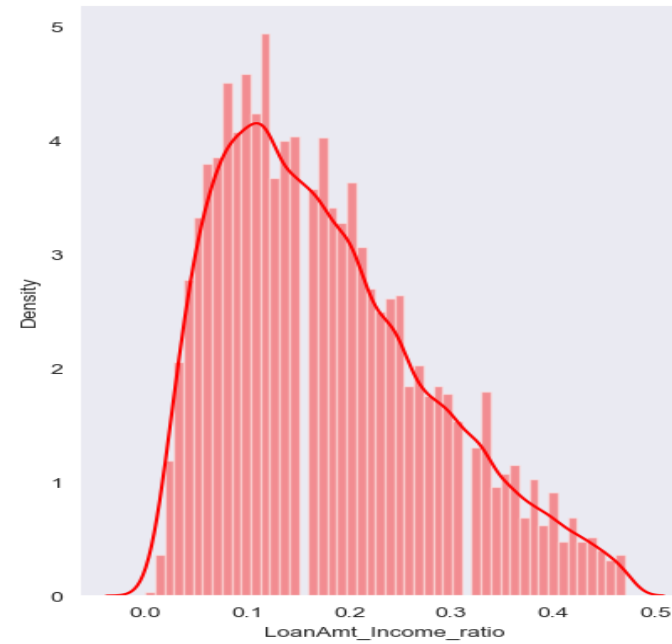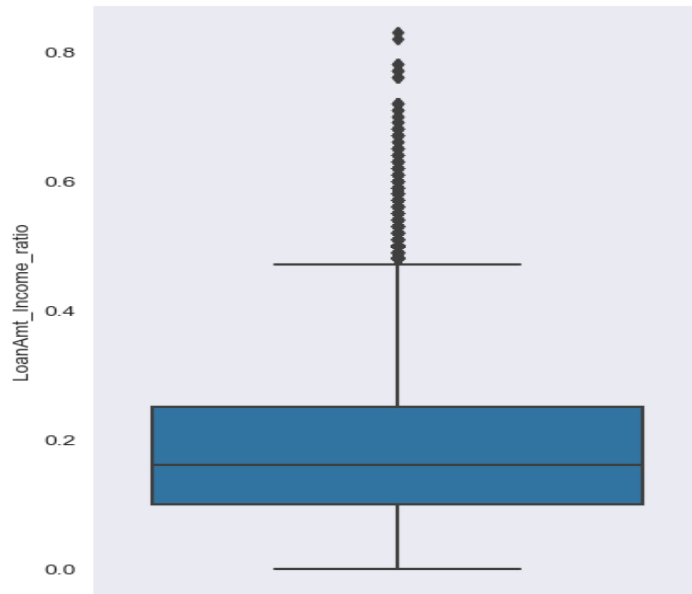
6. Analysis of Grade

Observation:

1. Grade B applicant received the highest number of Loans followed by A Grade G has highest percentage of defaulters.
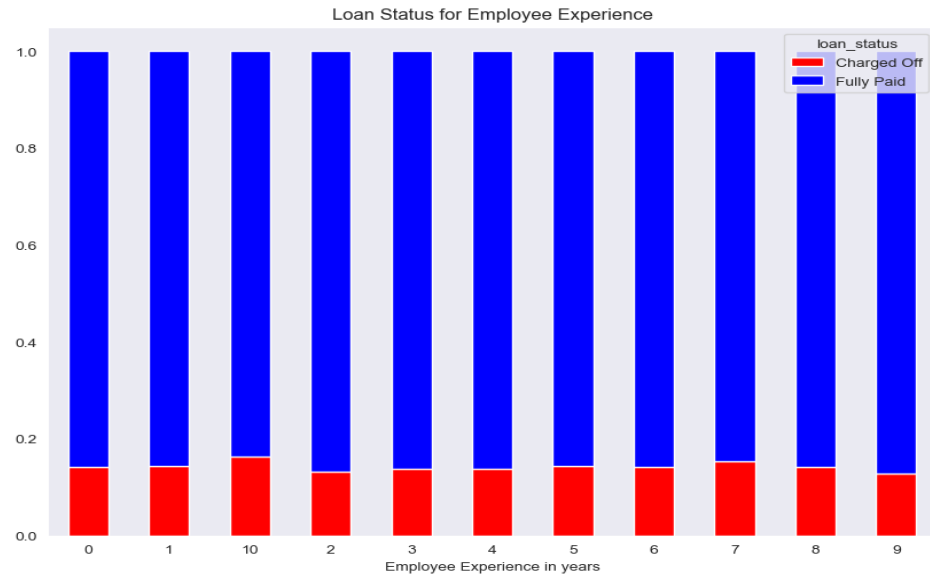
7. Loan amount to Income Ratio

Observation:

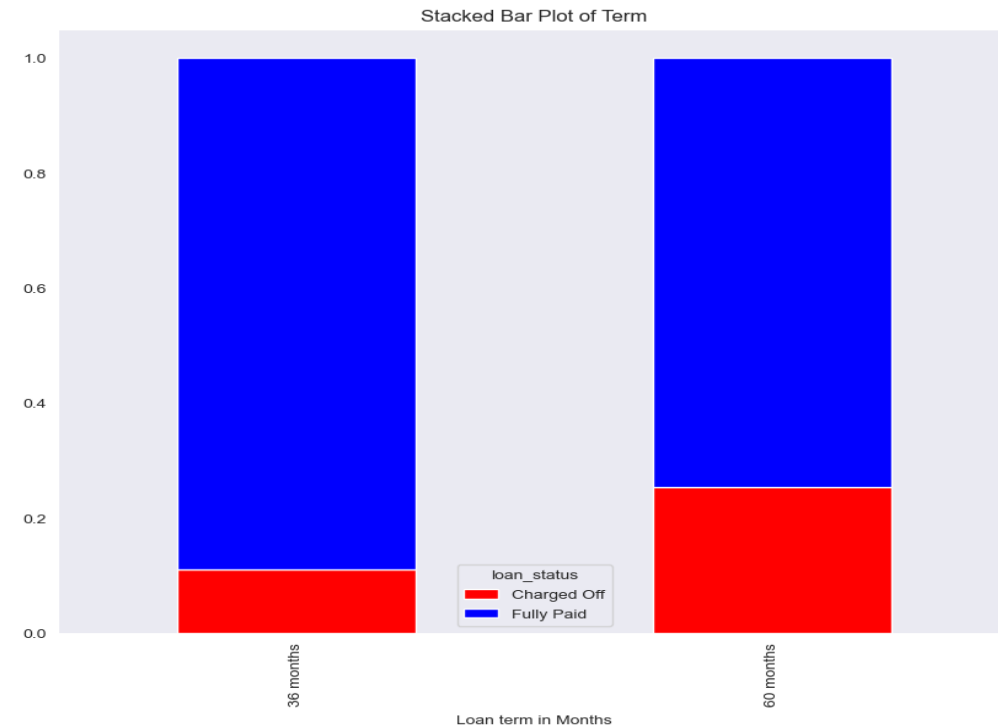1. Borrowers with high loan amount to income ratio are likely to default on their loans more.

8. Employee Experience

Loan Status for Employee Experience

9. Term Distribution

Stacked Bar Plot of Term

Observation:

1. Applicants with Professional Expirence more than 10 years are more prone to default.
2. Loan with 60 months term is more likely to default than loan with 36 months term.

Thank you