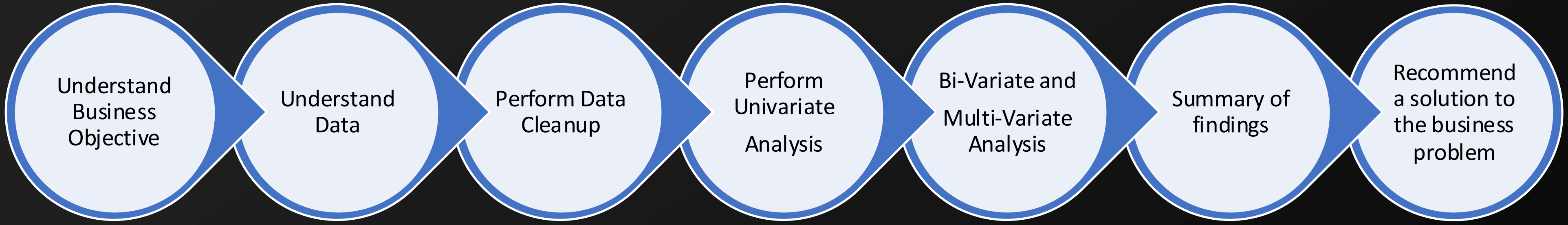
LENDING CLUB CASE STUDY

LendingClub is a financial services company, which specializes in lending various types of loans to urban customers. While approving a loan application, the company wants to assess the likelihood of a customer to repay the loan vs defaulting on it essentially establishing the **creditworthiness**. While make such a decision, the company wants a balance between potential loss of business if a ‘good’ customer is denied a loan, vs extending the same to a ‘bad’ customer leading to a financial loss.

Exploratory Data Analysis (EDA) was performed on the dataset in a multi-stage process. The dataset contains loan details for all loans issued through years 2007 to 2011.

This presentation describes the key findings from the EDA and makes recommendations for improving the business outcomes.



Submission By:

Ramprasad Gutulla

Vidyadhar Watve

- Dec2024

upGrad

upGrad and IIITB Machine Learning & AI Program - October 2024

Key observations about the Data

- The dataset contains **39717 rows** and **111 columns**

- **COLUMN analysis**
  - **54** columns are found to be blank or **NULL** values
  - **09** columns are found to have a **single value** and are not useful for analysis
  - **04** columns appear to have **some valid values**, but majority (>50%) are NULL
  - **10** columns carry data of **no significance** for the analysis e.g. ID, Title, URL
  - **09** columns carry information after a loan is granted, so has **no bearing** on establishing creditworthiness
  - **03** columns have fewer missing values and need to be imputed or dropped *(emp_length, revol_util, pub_rec_bankrupt)*

- **ROWS analysis**
  - **0** duplicate rows were found
  - **1140** rows belong to 'current' or active loans. These do not contribute to historical data

- **Improper data**
  - **03** columns need conversion of data type (E.g. interest rate to numeric, dates to be of type *datetime.)*

Conclusion: Data cleanup is necessary to drop invalid and unwanted data and to convert it to a consumable format.

Summary of cleanup performed on the data

Original Dataset: **39717 rows** and **111 columns**

Cleaned up Dataset: **38557 rows and 27 columns**

Dropping of invalid and unusable column data reduced the columns to 111-54-13-19 = 25

- 54 columns with **all NULL** values are dropped
- 13 columns with **majority NULL** values or **single value** are dropped
- 19 columns with **no significance** to creditworthiness and statistically insignificant are dropped

Standardizing Data (Column datatype changed, 2 columns added)

- Column carrying date information converted from Object to Datetime for easier analysis
- Further the Datetime was split into additional 2 columns for Month and Year

Dropping rows that are not useful for analysis

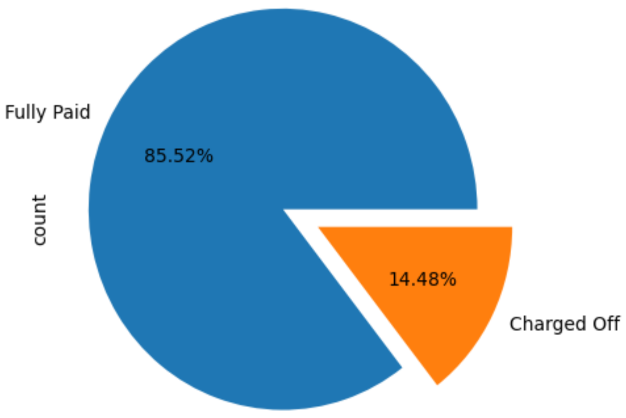
- **1140** rows corresponding to active loans were dropped as there is no default

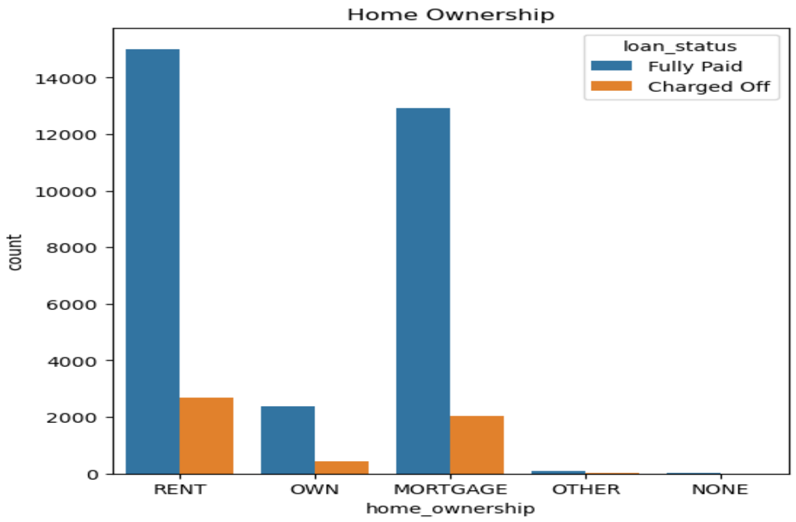
Outlier Treatment

- **20** Rows with outliers (above 95th quartile) were dropped so that they do not skew the analysis

Univariate Analysis

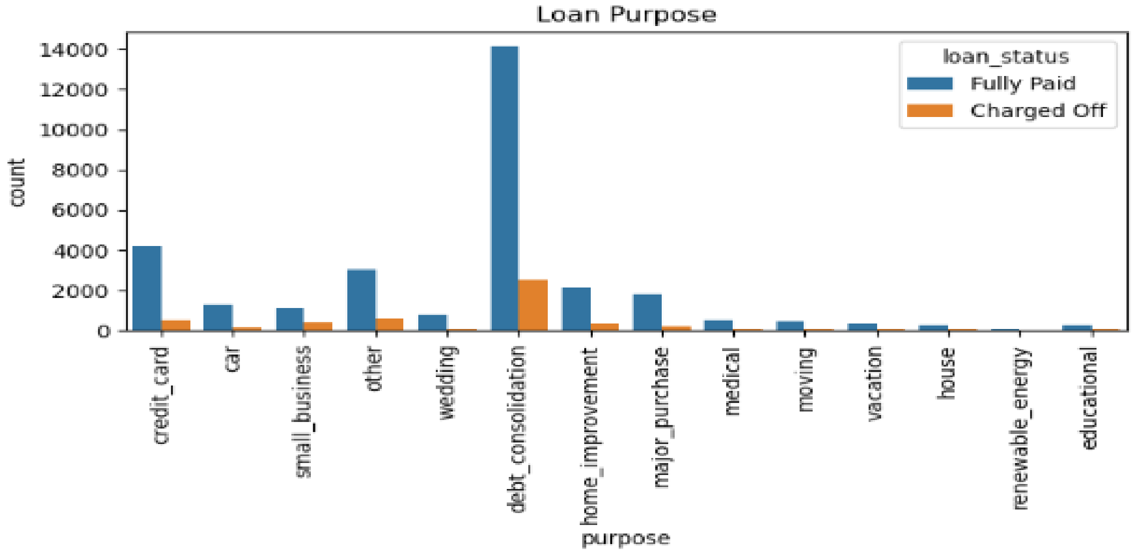
Performed on Unordered Categorical Variables, Ordered Categorical Variables, Quantitative Variables





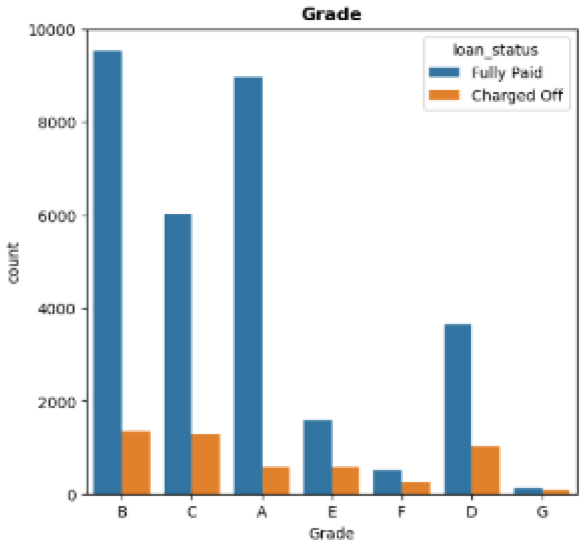
Inferences:

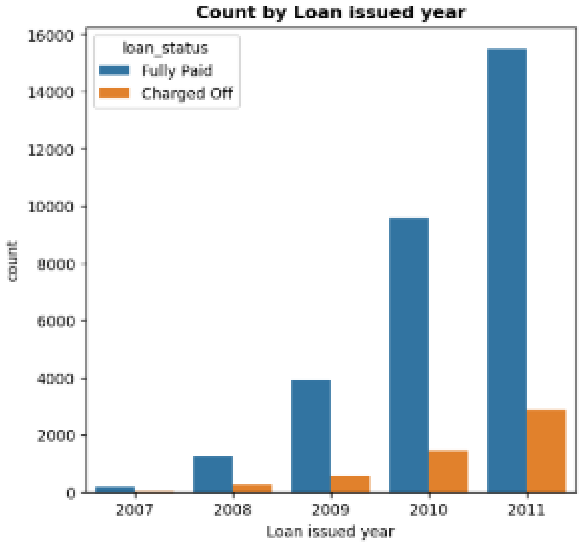
- 14.5% of overall loans get *charged-off* which is very concerning
- Majority of Loans are taken for debt consolidation
- Majority of loans are taken who doesn't own a home



Univariate Analysis

Performed on Unordered Categorical Variables, Ordered Categorical Variables, Quantitative Variables



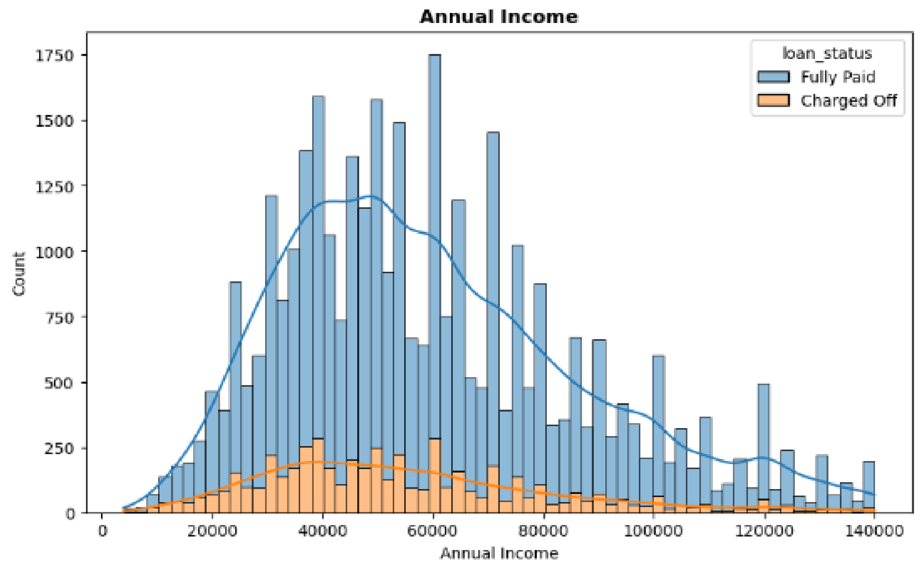


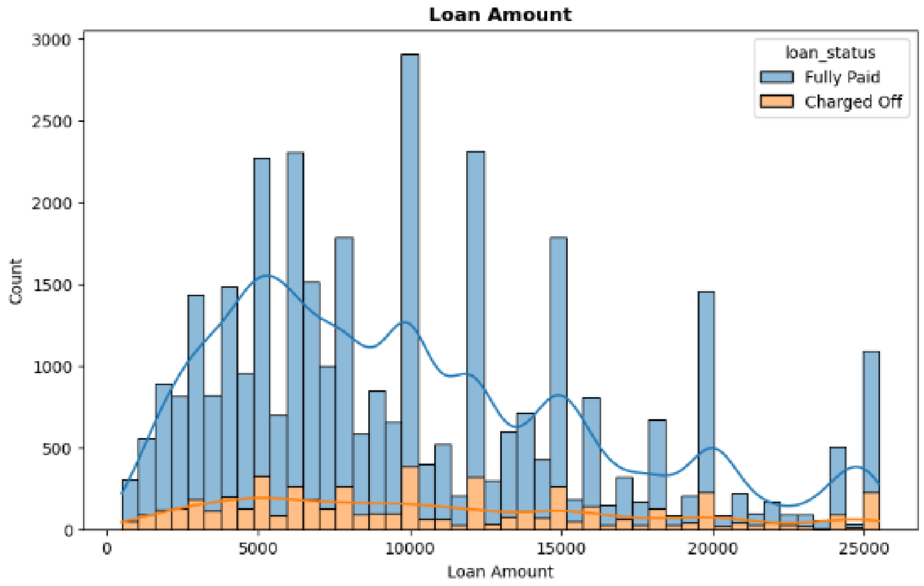
Inferences:

- Loans issued are decremental from grades A-G
- Loans issued are decremental from experience 0 to 9 but 10+ has a spike probably because all other experiences fall under this category. Going by this trend we can assume more the experience less the possibility of taking loan
- Loans issued are almost doubled every year

Univariate Analysis

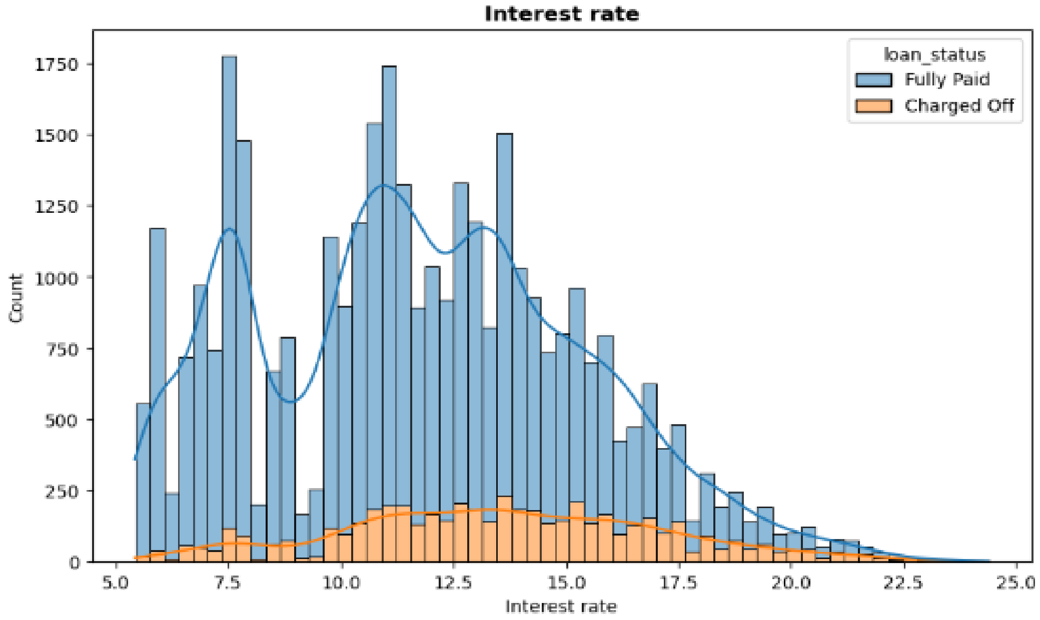
Performed on Unordered Categorical Variables, Ordered Categorical Variables, Quantitative Variables





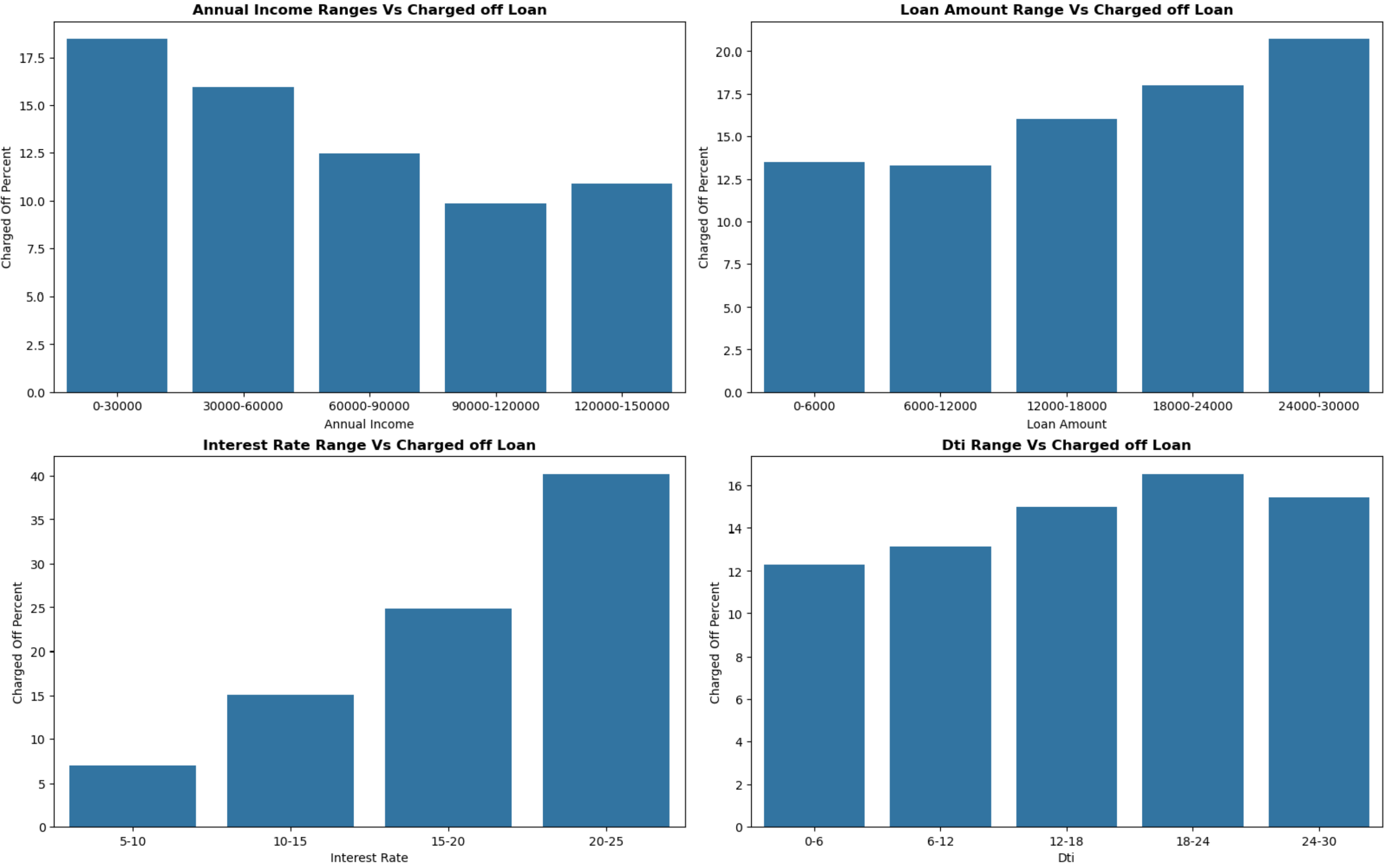
Inferences:

- Majority of Loans applicants has annual income between 30k and 80k
- Majority of Loans issued are for value less than 15k
- Interest rate of majority of loans is between 7.5 to 15



Bi-Variate Analysis

Performed on Unordered Categorical Variables, Ordered Categorical Variables, Quantitative Variables

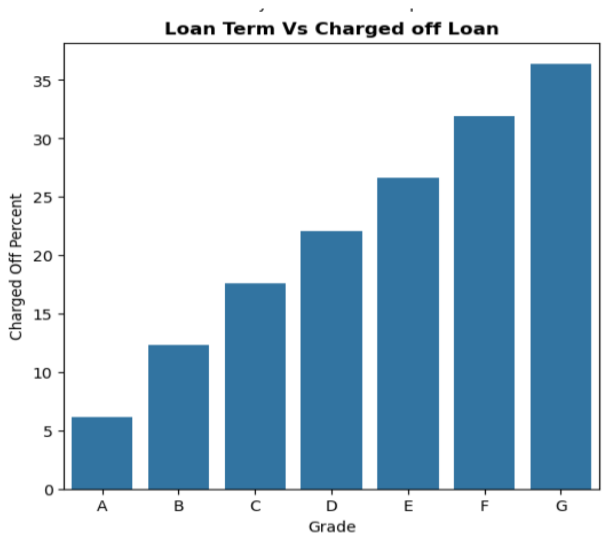


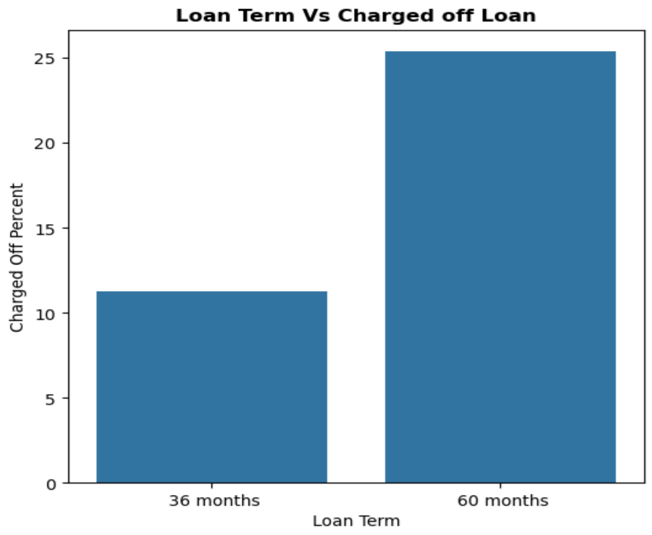
Inferences:

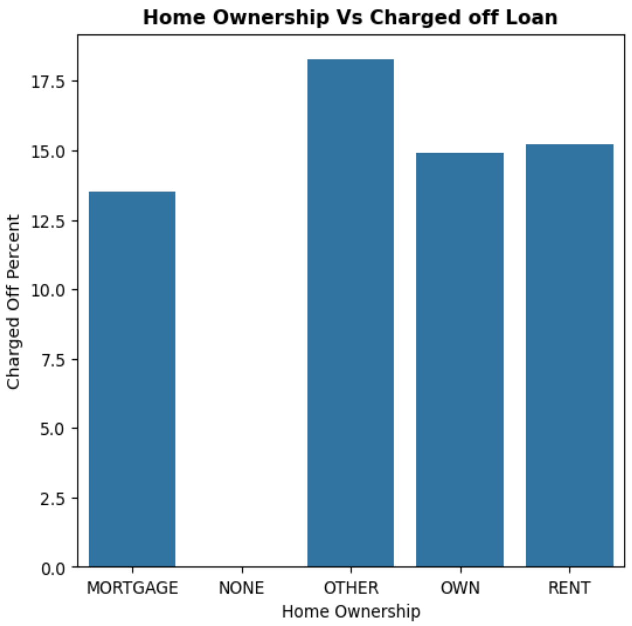
- Lower the annual income the percentage of charged off loan are higher
- Higher the loan amount/interest rate/Revolving line utilization the percentage of charged off loan are higher
- Dti and Total credit revolving balance doesn't seem to be a considerable factor in loan defaulting

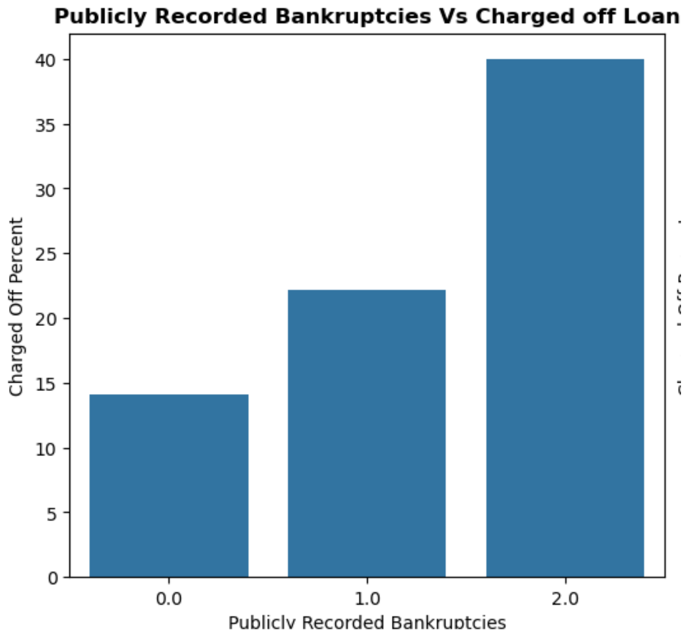
Bi-Variate Analysis

Performed on Unordered Categorical Variables, Ordered Categorical Variables, Quantitative Variables







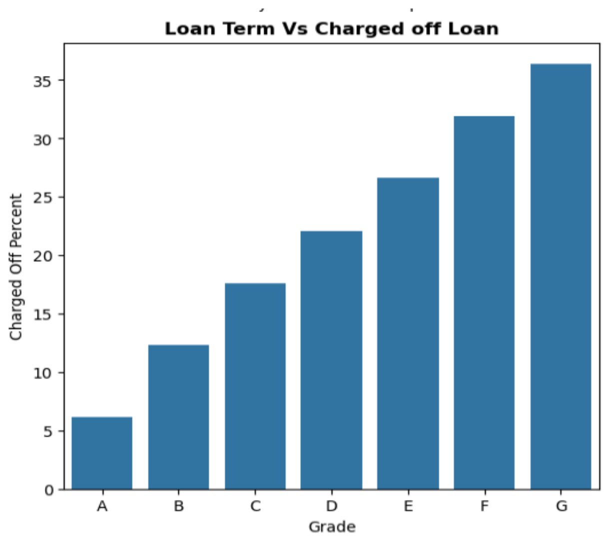


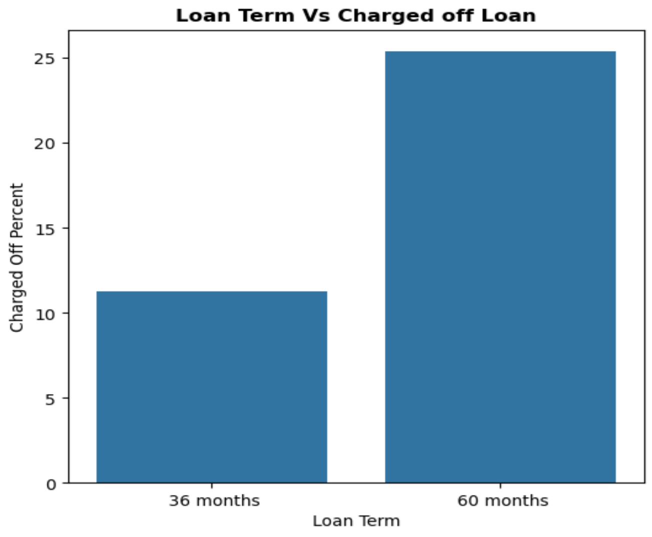
Inferences:

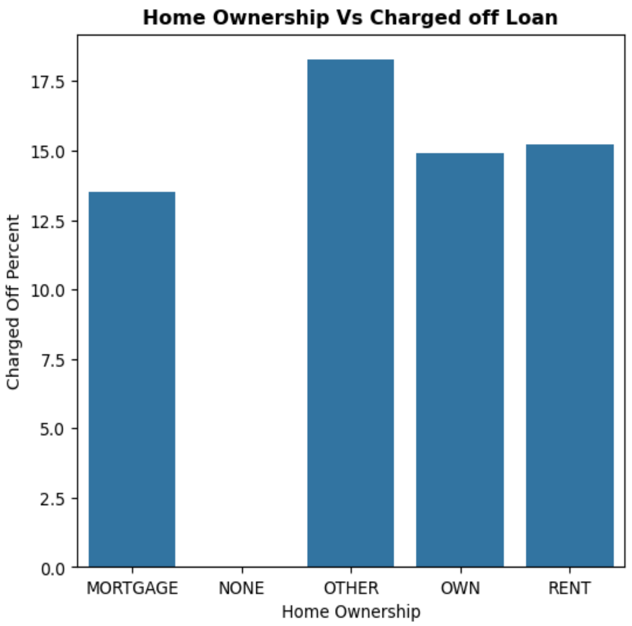
- Higher the grades, subgrades, loan term higher the percentage of charged off loans
- Home ownership other has more percentage of charged off loans
- Loan term with 60 months have a greater number of charged off loans
- More the number of inquiries in last 6month, derogatory public records , and public recorded Bankruptcies the percentage of charged off loans are higher
- Loans taken for small business more percentage of charged off loans
- State/Employment length/number of open accounts doesn't seem be a considerable factor in loan defaulting

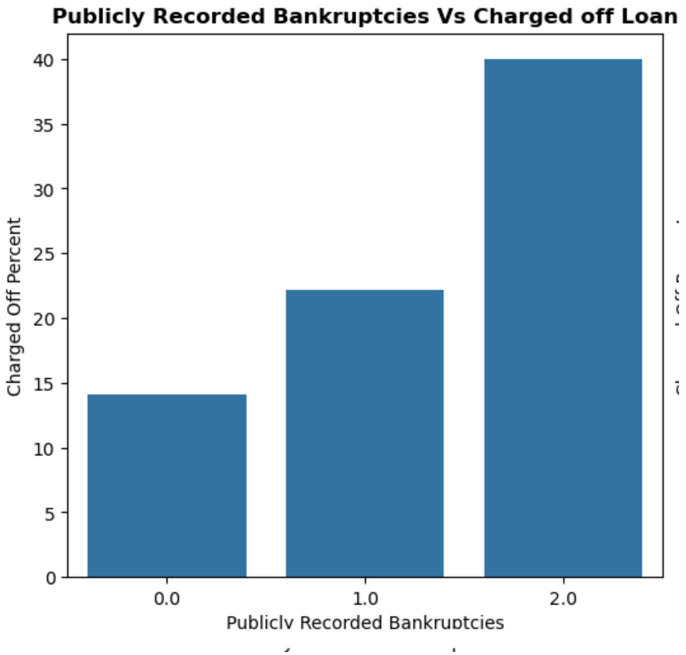
Bi-Variate Analysis

Performed on Unordered Categorical Variables, Ordered Categorical Variables, Quantitative Variables







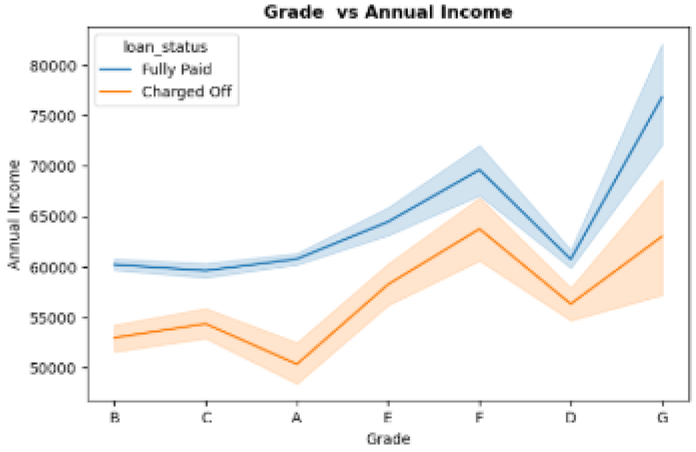


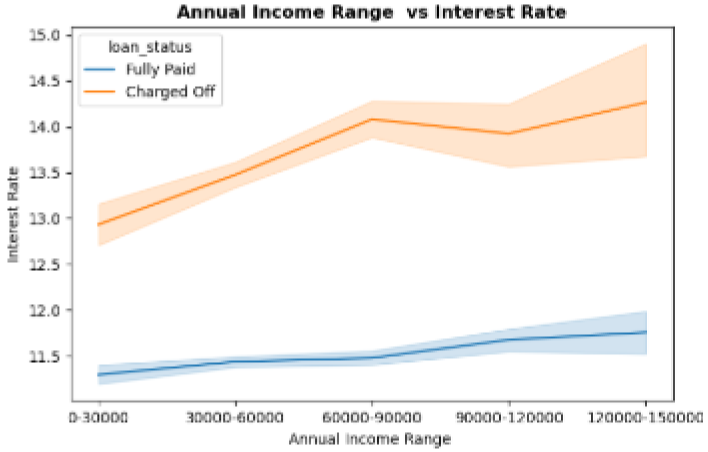
Inferences:

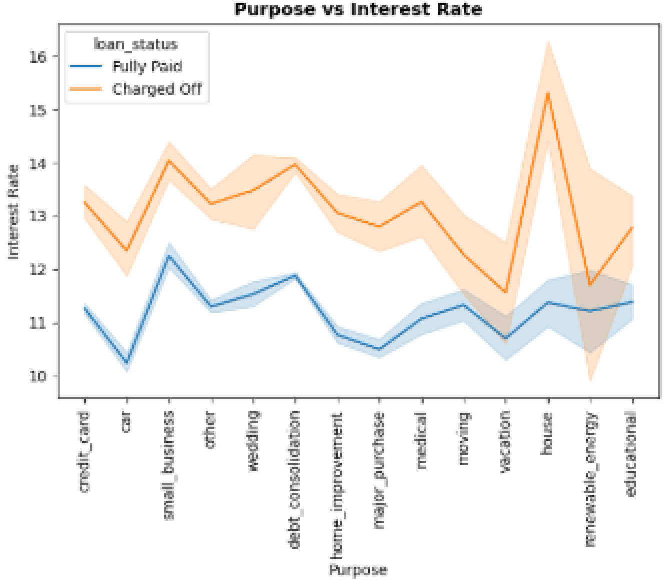
- Higher the grades, subgrades, loan term higher the percentage of charged off loans
- Home ownership other has more percentage of charged off loans
- Loan term with 60 months have a greater number of charged off loans
- More the number of inquiries in last 6month, derogatory public records , and public recorded Bankruptcies the percentage of charged off loans are higher
- Loans taken for small business more percentage of charged off loans
- State/Employment length/number of open accounts doesn't seem be a considerable factor in loan defaulting

Bi-Variate Analysis

Performed on Unordered Categorical Variables, Ordered Categorical Variables, Quantitative Variables



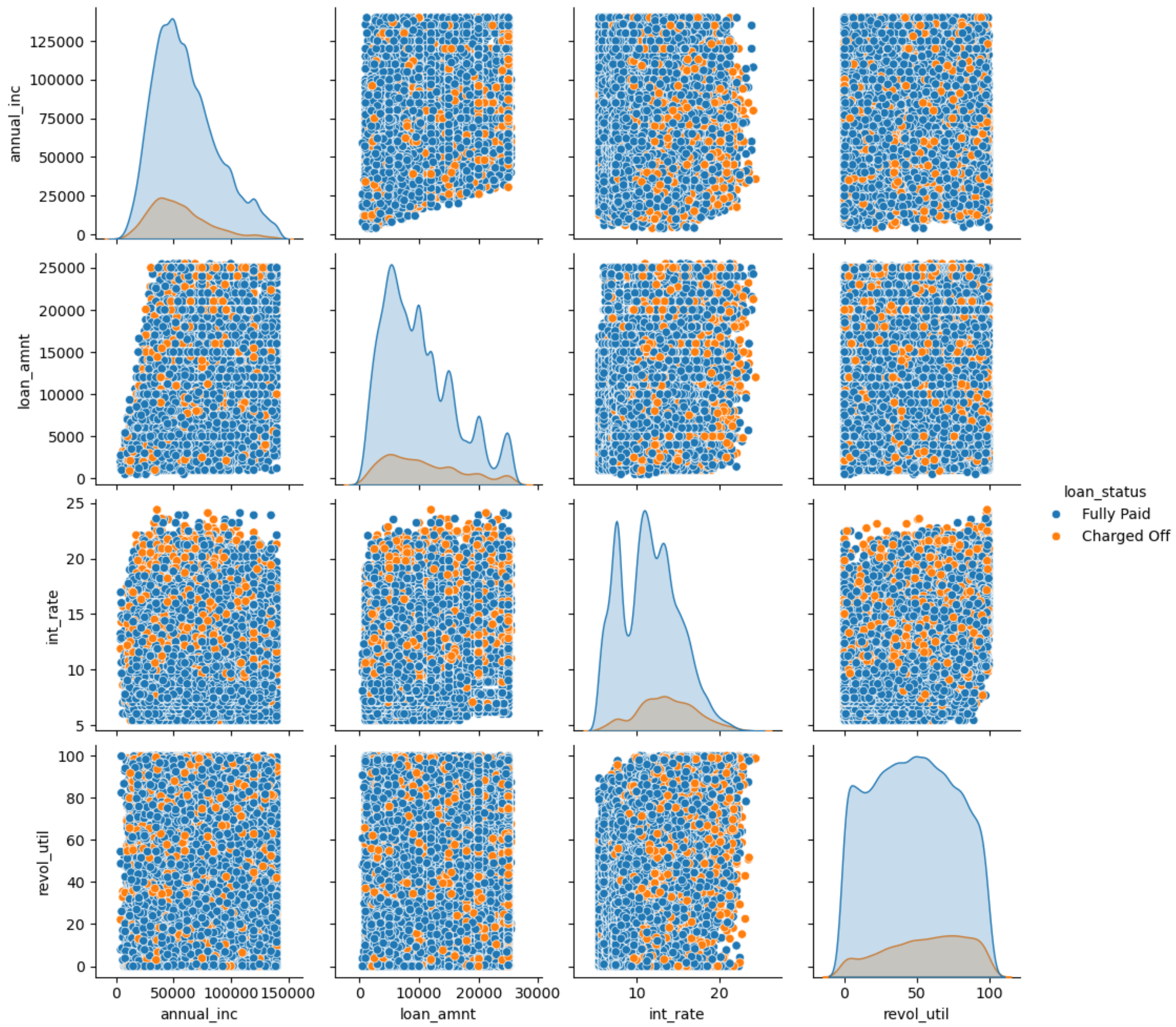




Inferences:

- Grades D,E,F,G has more defaulters below 60k annual income
- More defaulters with annual income more than 9k and higher interest rate
- Loan taken for purpose Small business, debt consolidation and home has more defaulters when the interest rate is higher

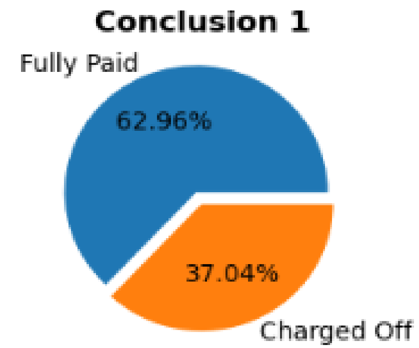
Multi-Variate Analysis

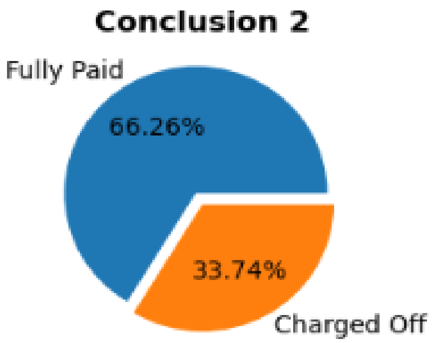


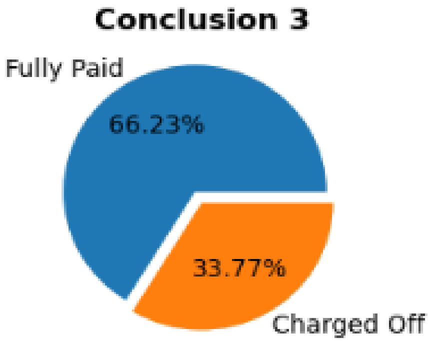
Inferences:

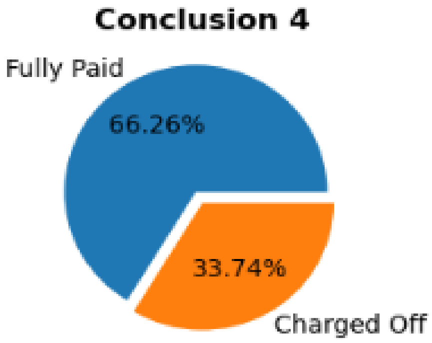
- Higher interest rate more the defaulters
- Loan amount > 15k and annual income < 75k seems to have more defaulters

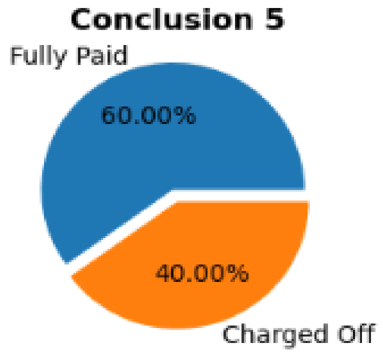
Conclusion











Strong indicators of default as depicted in the pie charts

Conclusion 1 : Applicants with loan purpose as small_business, educational, house and interest rate >= 15% has more chances of defaulting
Conclusion 2 : Any loan with interest rate above 18% has more chances of defaulting
Conclusion 3 : Applicants with Annual income <= 90k and Interest >= 17% has more chances of defaulting
Conclusion 4 : Applicants with grades E,F,G and annual income <=40k has more chances of defaulting
Conclusion 5 : Applicants with more than one public recorded bankruptcies has more chances of defaulting

Weak indicators Or In-conclusive indicators of Default

Seasonality: There is some seasonality with number of loan applications increasing during May-June & Dec, likely due to holidays and festivals. This is not a as prominent though.
Experience : Applicants with 10+y of experience are seen to have higher count of defaults. This is counter-intuitive
Given with higher experience and age it is expected that the person has reached more stability and likely having decent
Earnings. Hence experience alone does not help to establish creditworthiness.

Recommendation

Loan term length: Applicants applying for longer term loans (60 months) with higher interest rates have a more rate of default and hence should be scrutinized more with respect to their background and prior default.

Reason of loan: Debt consolidation appears to be a prominent factor associated with the default. There are likely elements of mis-use and fraud and hence the company should minimize extending loans for such purpose.

Verification of applications: Defaults are seen equally among verified and un-verified applications. The company should do a strong review and introspection on its customer verification process and introduce more compliance audits.

Income level: Lower income group has been observed to have more defaults. The company could set a lower limit on this band to ensure it does not exceed the repaying capability.

Mortgage: Limit loan exposure to this segment given high chance of default

Thank you