Lending Club Case Study

Group Facilitator – Purnachander Gajjalla

Group Member – Pramod Khandare

Data Understanding :

- Number of loans: ***39717***
- Number of attributes considered for each application: ***111***
- Number of numerical attributes: ***27***
- Number of categorical attributes: ***84***
- Identified target column - ***loan_status***
    - This variable identifies if a loan is fully paid, defaulted or currently active .
    - Values of various attributes associated with currently active loans will be either null or incomplete. **Hence current active loans are not considered for this study.**
- Identified columns containing one or more garbage values - ***68***
    - Values such as ‘’, ‘?’, ‘-’, ‘NA’, ‘na’, ‘N/A’, ‘n/a’, ‘NONE’, ‘None’, etc are considered as Garbage values
- Identified irrelevant columns - ***18***
    - Columns whose values are not measurable or incomplete at the time of loan application are considered as irrelevant columns.
    - For example: Borrower’s outstanding principal, total payment, recoveries,
- Identified columns with mixed data types
    - Mixed data type columns having same values specified in two different types like string and numeric, in source CSV file.

Data Cleaning :

- Filtered out currently active loans
- Dropped 58 columns with high percentage of garbage values
    - Of these 58, 54 columns are with 100% garbage values.
- Dropped 18 irrelevant columns
- Handled columns with mixed data types
    - Of the 4 columns with mixed data type, column **collections_12_mths_ex_med** has been identified as irrelevant one.
    - After observing values of other three columns, they are converted to float type.

	col_name	n_categories	categories
0	chargeoff_within_12_mths	3	[0 '0' 'NA']
1	pub_rec_bankruptcies	6	[0 1 2 '0' '1' 'NA']
2	tax_liens	3	[0 '0' 'NA']

- Handled columns wrongly mapped to object data type.
    - Removed percentage symbols from columns **int_rate** and **revol_util** and converted them to numeric
    - Converted **term** column to numeric.
    - Converted **earliest_cr_line** column into datetime format

Data Cleaning :

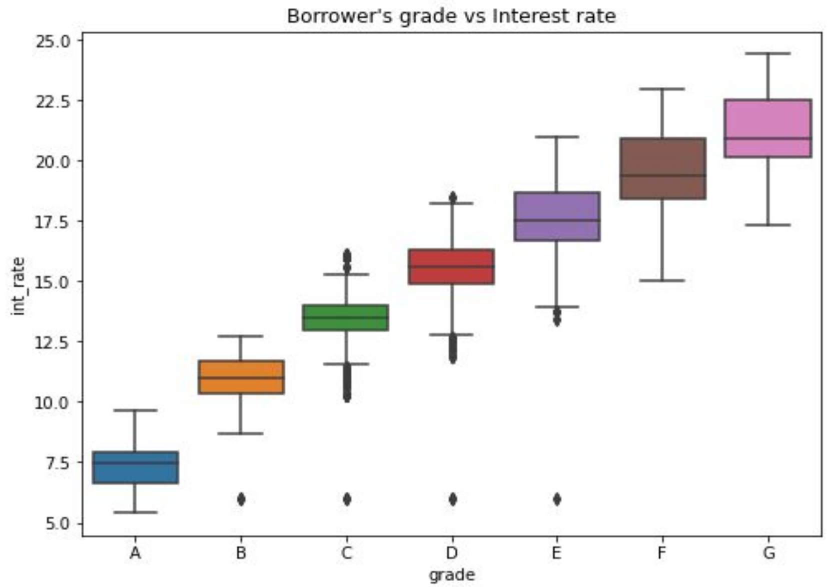
- Dropped columns such as **pymnt_plan**, **application_type** due to presence of just one value.
- Handled columns wrongly mapped to numeric data type
    - Converted numeric columns such as **id, member_id, policy_code** to object type
    - Dropped columns having just zero value: **acc_now_delinq, delinq_amnt**
- Handled columns with high number of categories.
    - Dropped columns with higher categories: **emp_title, title, url, zip_code**
    - Dropped column **member_id** and set **id** column as index, since both represented a unique row
    - Dropped **policy_code** column since it has just one value
- Handled columns with low garbage values.
    - Dropped columns: **chargeoff_within_12mths**, and **tax_liens**, since they contained just one value (i,e 0) apart from low garbage values
    - Replaced remaining garbage values across all columns with null (**np.nan**)

Data Cleaning :

- Interpreted missing values in **home_ownership** column
    - Values marked as **'NONE'** are considered as missing in this column.
    - Missing values in column **home_ownership** are replaced with it's mode i.e **RENT.**
- Interpreted missing values in **emp_length** column
    - Generally, experience of an employee is associated with employees annual income.
    - Based on the annual income (**annual_inc**) of the borrower, **emp_length** values are interpreted.
- Interpreted missing values in **revol_util** column
    - A significant difference in distributions of **revol_util** values across **fully paid** and **charged off** loans is found.
    - Hence missing values in this column are interpreted as **median revol_util value of the loan status group** to which it belongs.
- Interpreted missing values in **pub_rec_bankruptcies** column
    - A high correlation is found between number of public records (**pub_rec**) and public bankruptcies (**pub_rec_bankruptcies**).
    - Hence missing values of this column are interpreted as mode of **pub_rec_bankruptcies** column based on the corresponding **pub_rec** column.

Data Analysis :

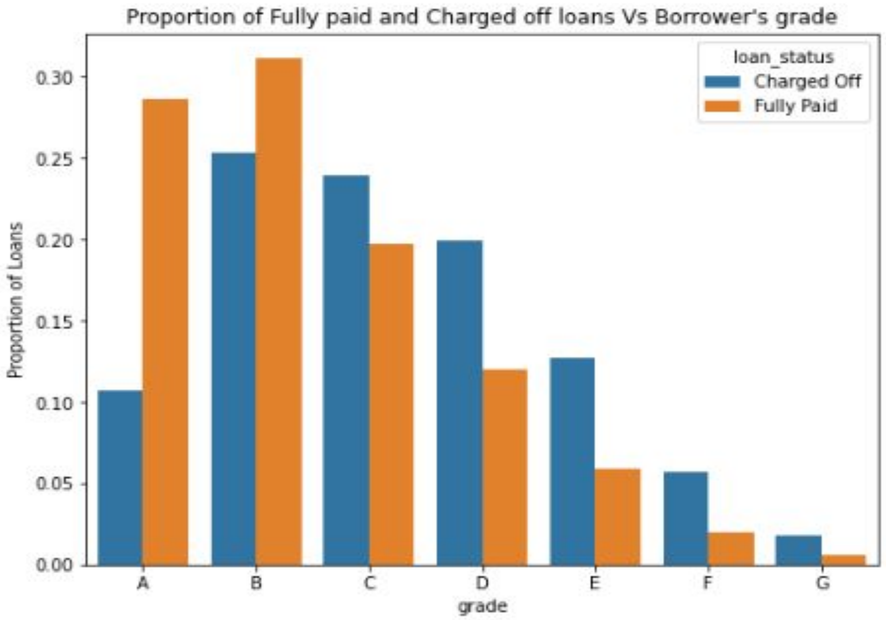
- Relation between ***Grade*** and ***Interest rate***
  - The **grade** of a borrower is based on past history of loans.
  - It generally indicates worthiness of a borrower.
  - Usually a borrower with high grade gets low interest rate offered and vice versa.



- **Insights**
  - Clearly above figure shows the relationship between borrower's grade and the interest rate at which loan is borrowed.
  - Borrowers with high grades like 'A', 'B' are likely to get loans at cheaper interest rates and vice versa

Data Analysis :

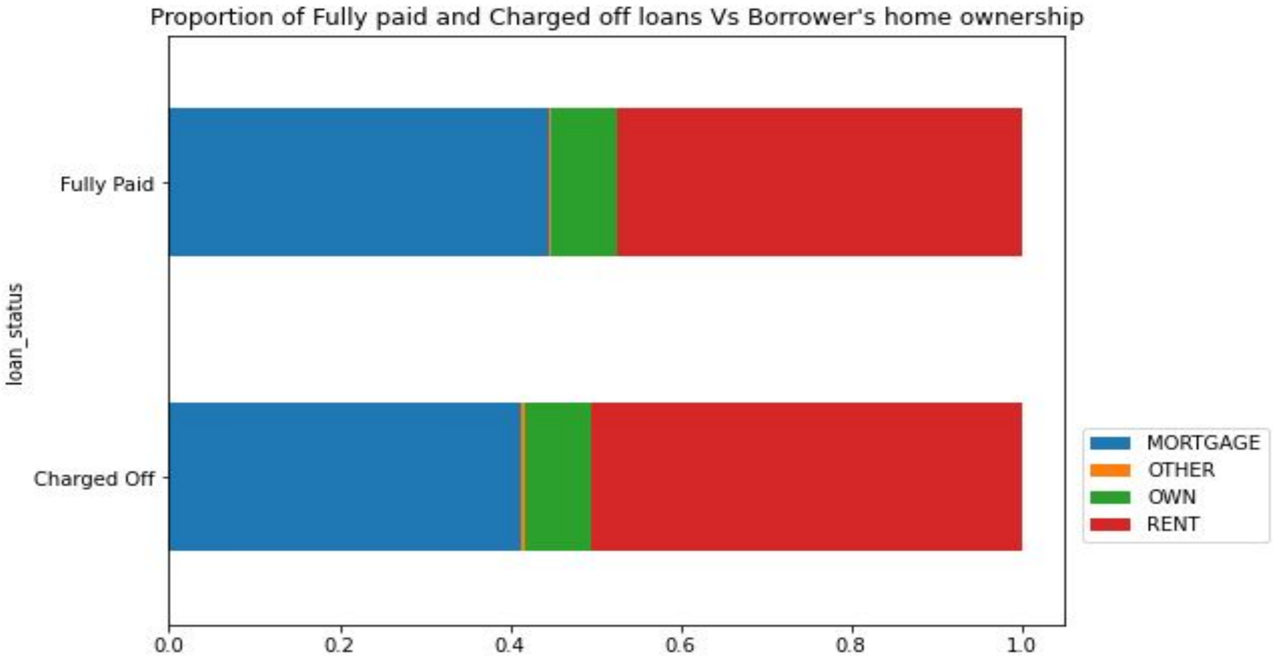
- Relation between ***Grade*** and ***Loan Status***



- **Insights**
  - The proportion of defaulted loans increased with decrease in the borrower's grade.
  - Meanwhile, the proportion of fully paid loans decreased with decrease in the borrower's grade.
- **Recommendations**
  - Borrowers with grade 'A' are highly unlikely to default. Targeting these borrowers will reduce the risk.
  - Borrowers with grade 'C' and above are more likely to default.

Data Analysis :

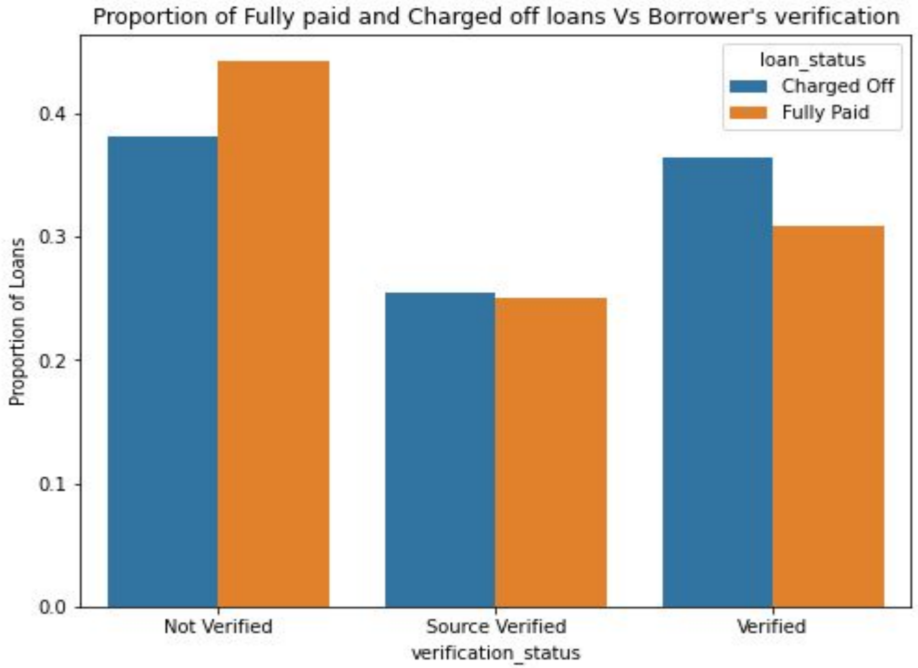
- Relation between ***Home Ownership*** and ***Loan Status***



- **Insights**
  - Borrowers staying in rented homes are slightly more likely to default.
- **Recommendations**
  - Home ownership is a weak indicator and should be used in combination with other indicators of default.

Data Analysis :

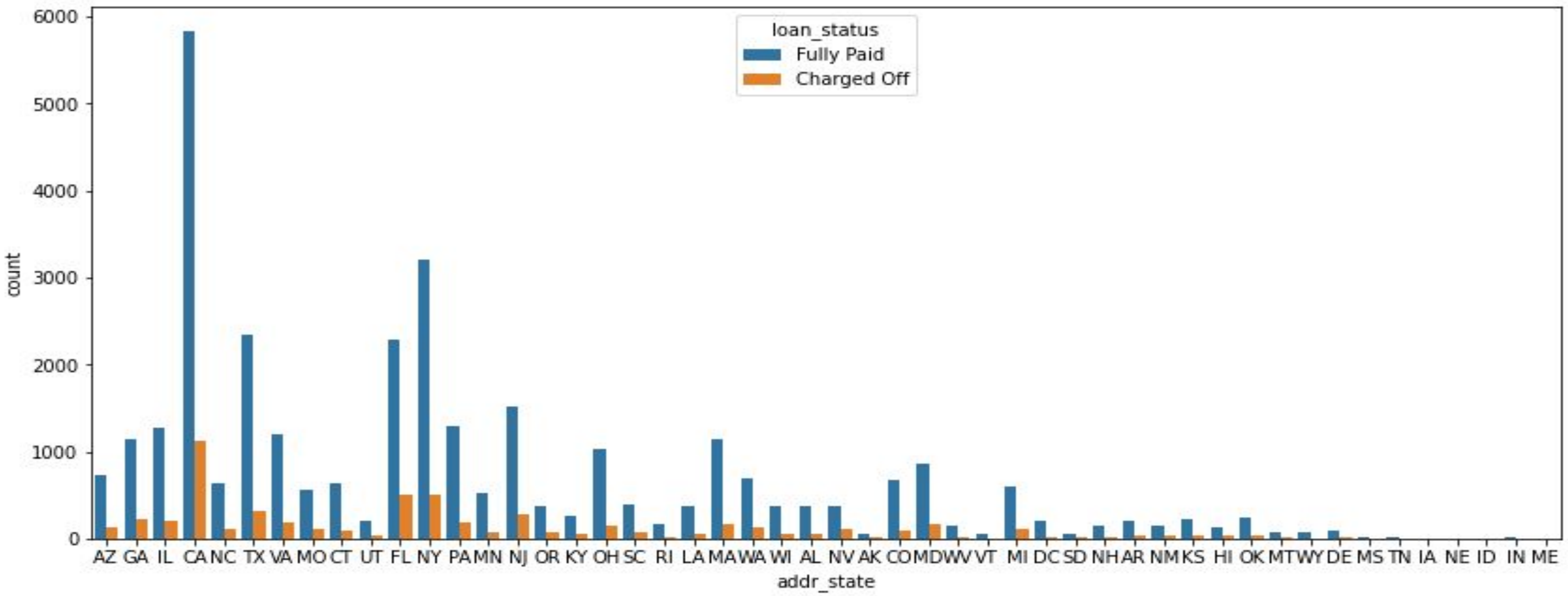
- Relation between ***Verification Status*** and ***Loan Status***



- **Insights**
    - Verification of income source is more important than just verification of job profiles.
- **Recommendations**
    - Verifying the income source details of borrowers reduces the risk of default.

Data Analysis :

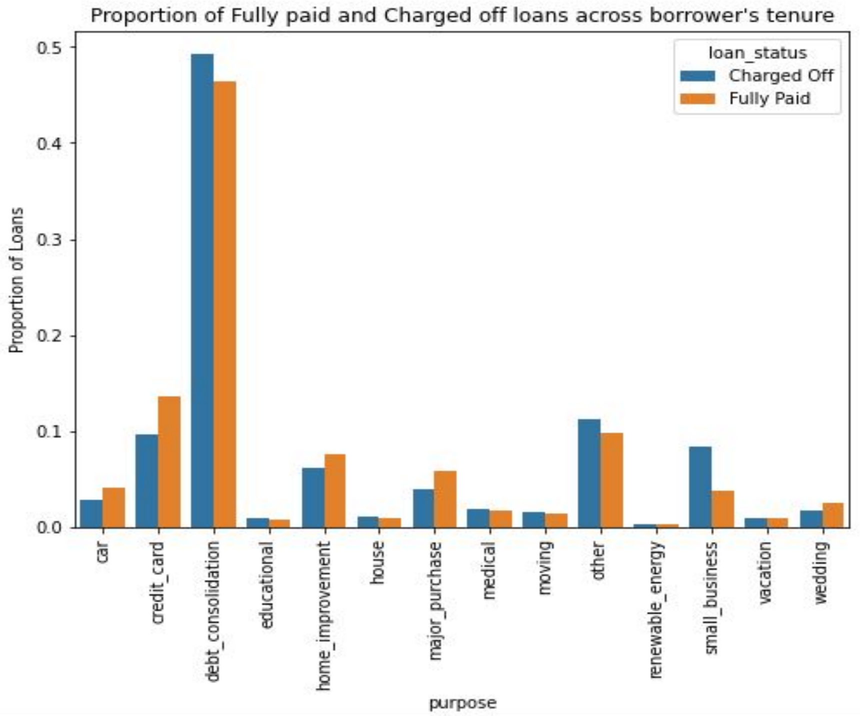
- Relation between ***Borrower's State*** and ***Loan Status***



- **Insights**
  - More number of charged off loans are associated with high cost of living states like California, Texas, Florida, New York, and New Jersey
- **Recommendations**
  - While lending loans to borrowers residing in higher cost of living states, consider other attributes such as income verification for reducing risk of default.

Data Analysis :

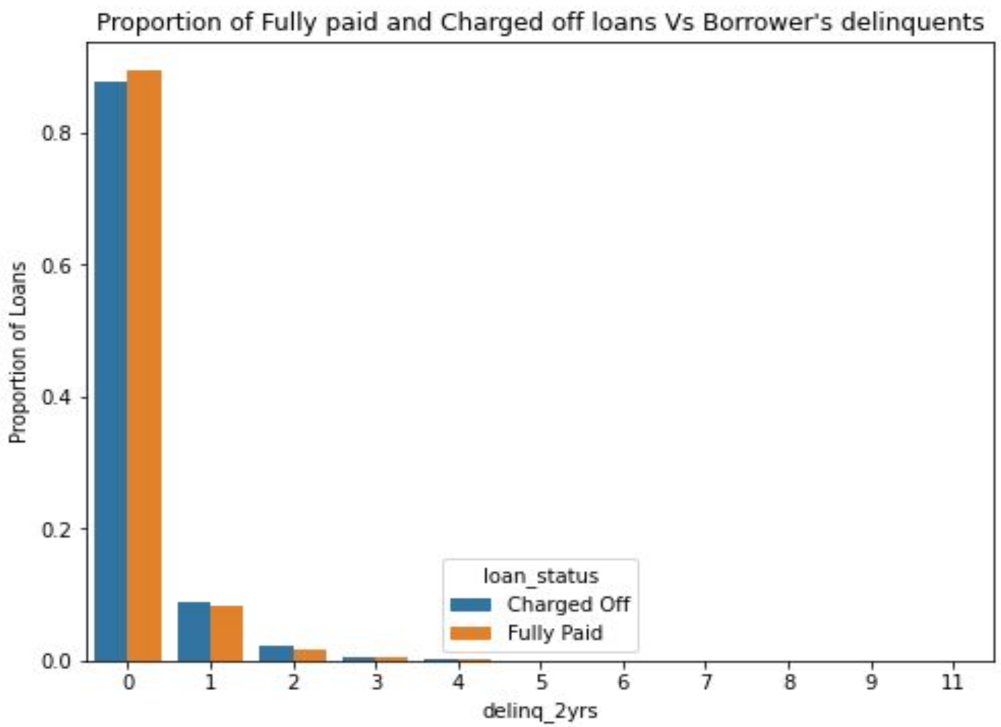
- Relation between ***Loan Purpose*** and ***Loan Status***

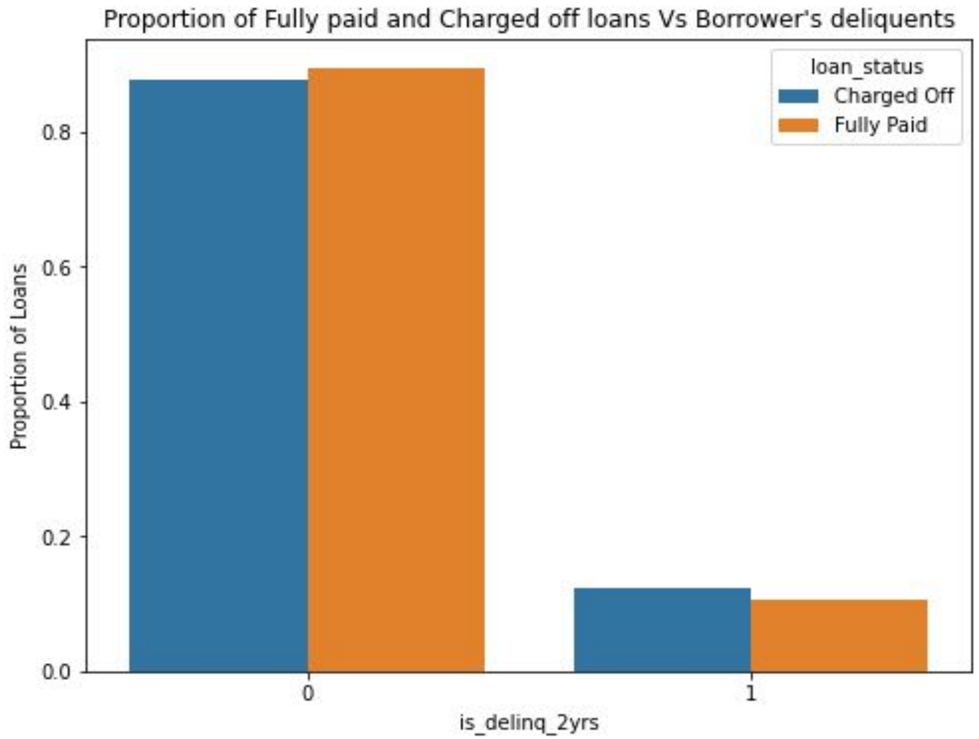


- **Insights**
  - Borrowers taking loans for purpose of **debt consolidation**, **small business** and **other** are more likely to default.
  - On other hand, borrowers taking loans for purpose of **car, credit card, home improvement, major purchase** and **wedding** are more unlikely to default.
- **Recommendations**
  - Encouraging loans for home improvement, car, credit card is a good idea.

Data Analysis :

- Relation between ***Borrower's Delinquents*** and ***Loan Status***

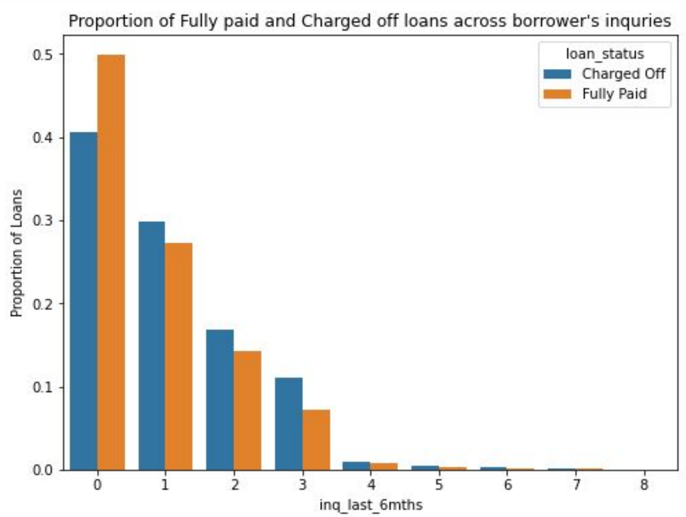


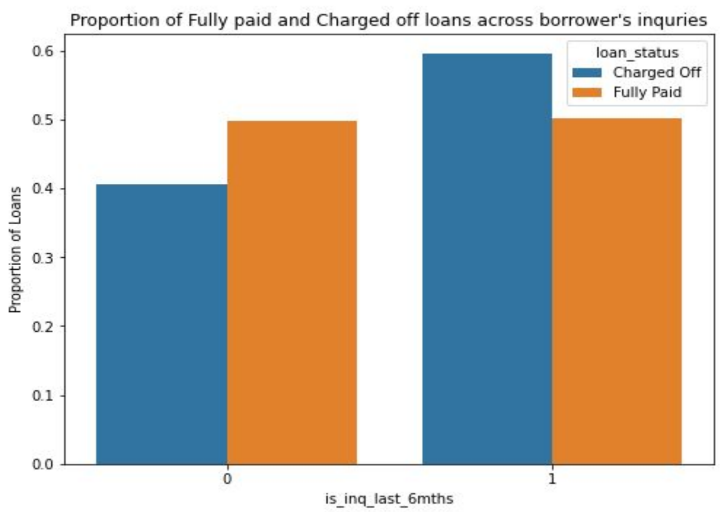


- **Insights**
  - The borrower's who were delinquent in last 2 years likely to default. However the difference from non delinquent users is not subtle
- **Recommendation**
  - There is a chance of getting default if the borrower has a record of delinquency in last 2 years.

Data Analysis :

- Relation between ***Borrower's Inquiry in last 6 months*** and ***Loan Status***

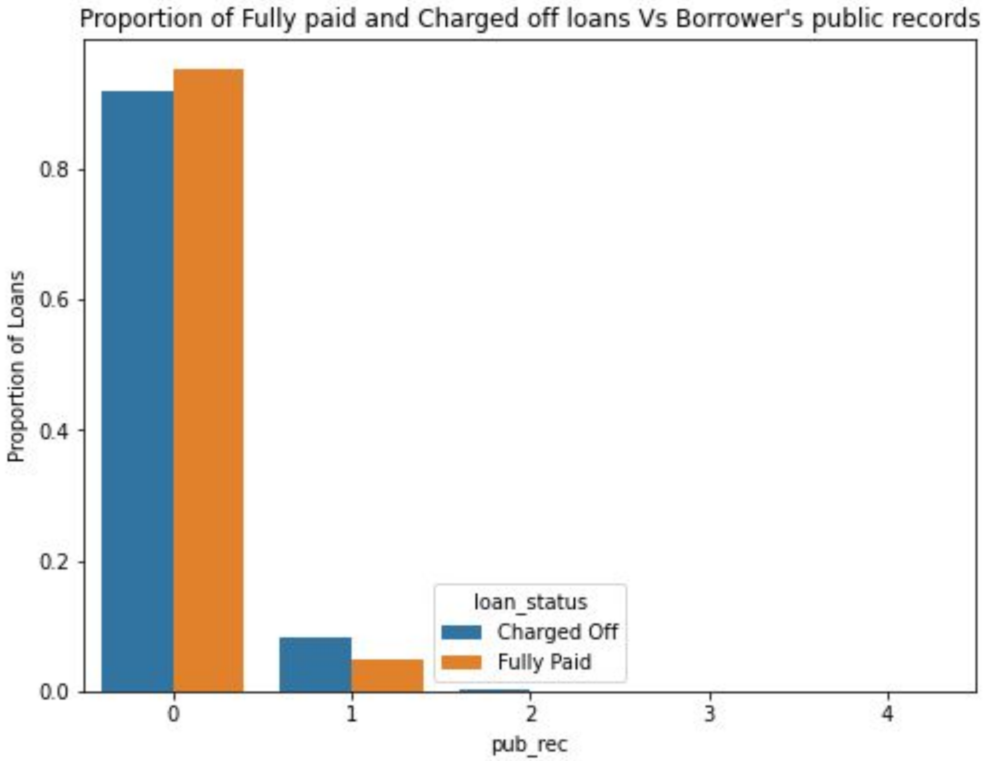


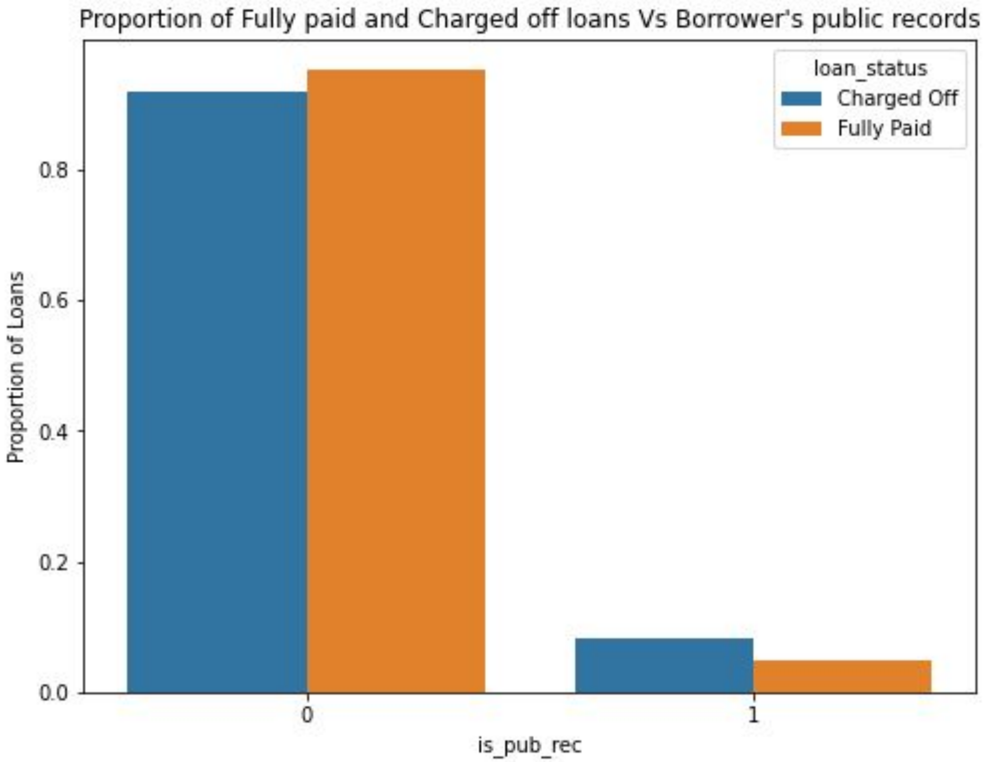


- **Insights**
  - The borrower's who have done any inquiries in last 6 months, about their credit history, are more likely to default.
- **Recommendations**
  - Look for borrowers who have not done any inquiries in last six months for reducing the risk of default.

Data Analysis :

- Relation between ***Borrower's Public Records*** and ***Loan Status***

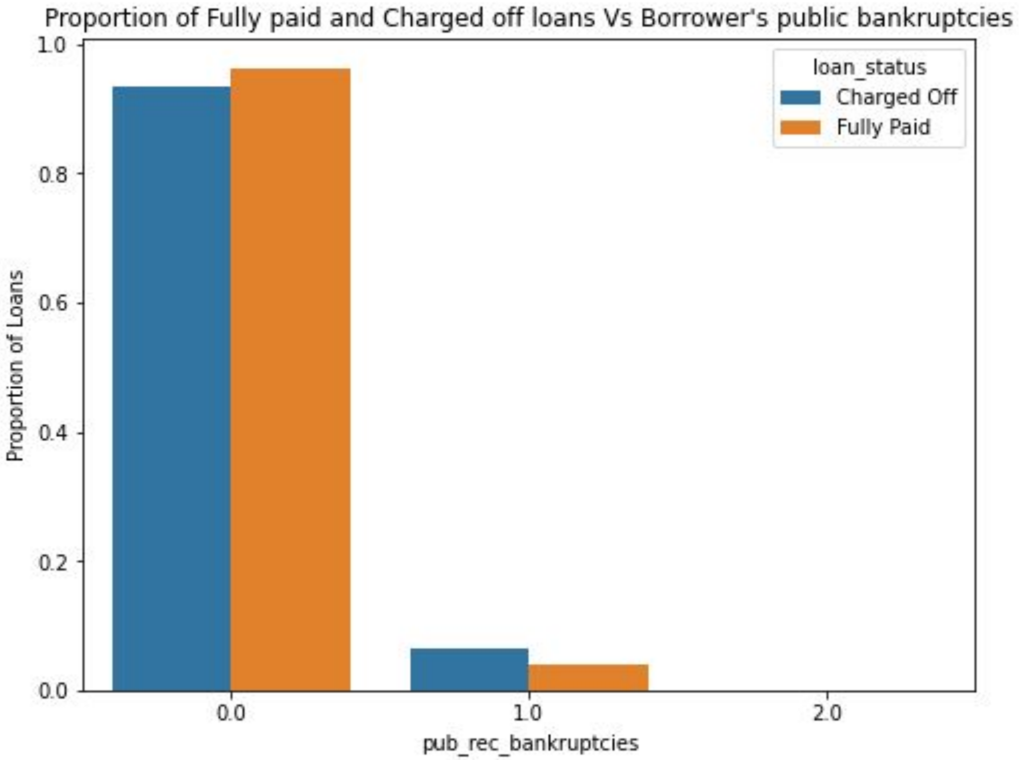


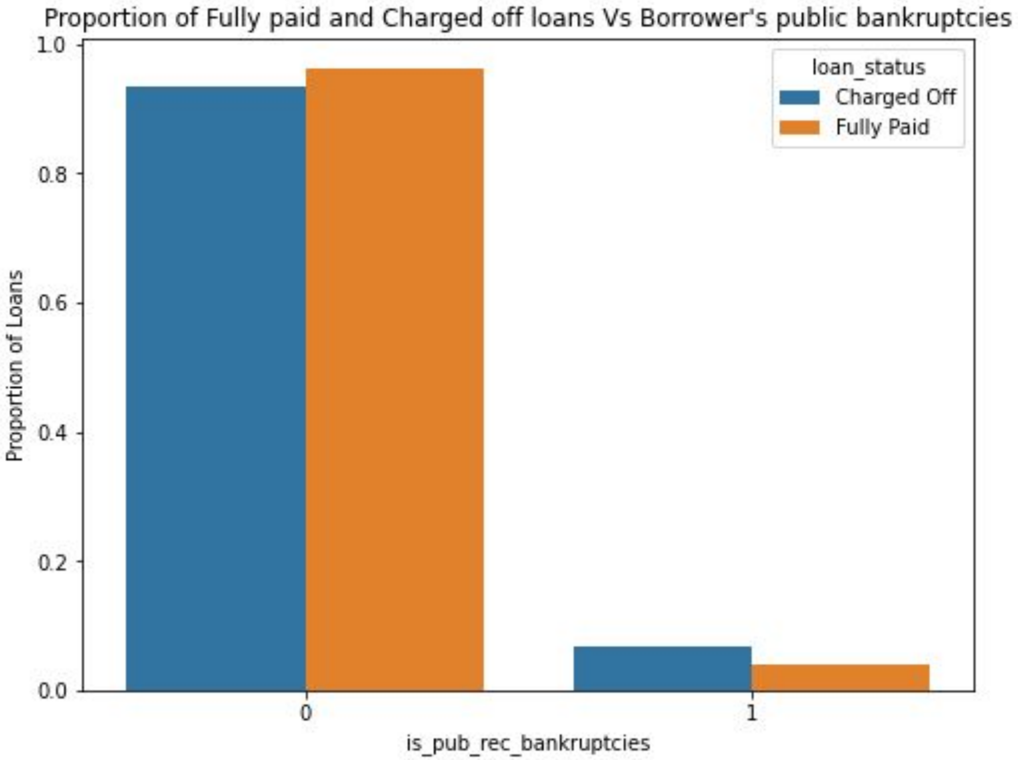


- **Insights**
  - Loans are more likely to get charged off if there are any public records reported.
- **Recommendations**
  - Check for any public records present before approving loan as the chances of loan getting charged off are very high

Data Analysis :

- Relation between ***Borrower's Public Bankruptcies*** and ***Loan Status***

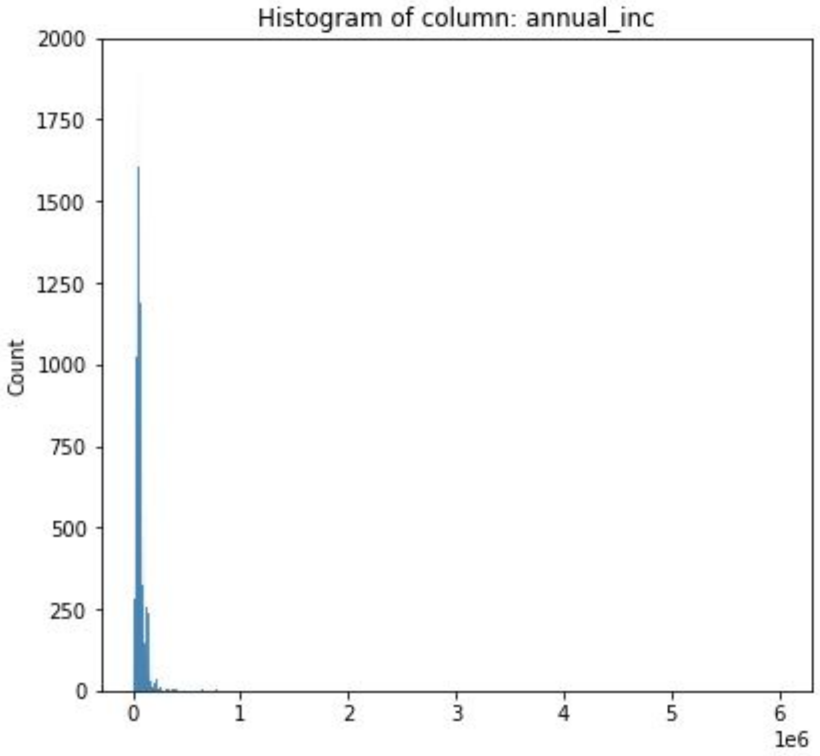


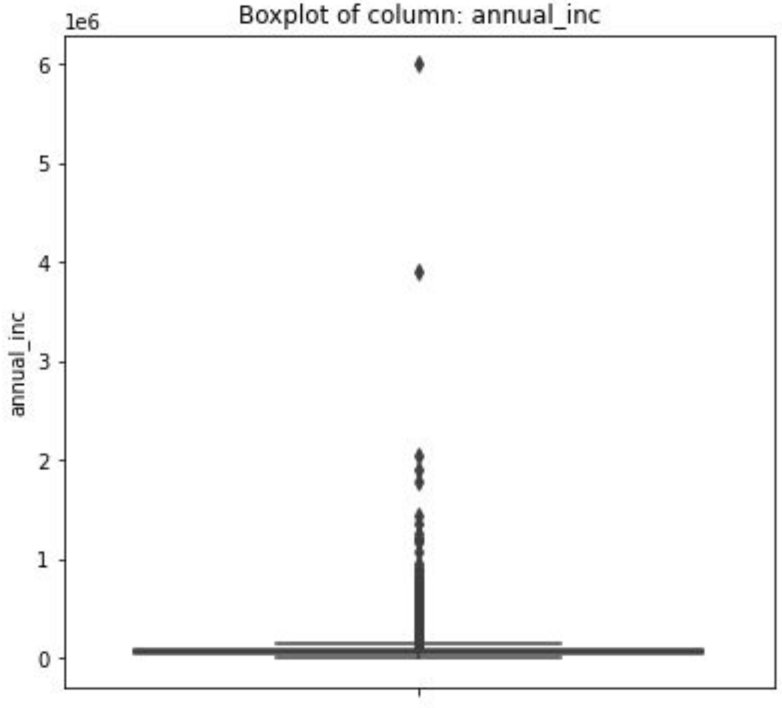


- **Insights**
  - Fully paid loans are more when the bankruptcies value is Zero
- **Recommendations**
  - Don't approve loan for the user having record for bankruptcies

Data Analysis :

- Relation between *Annual Income* and *Loan Status*

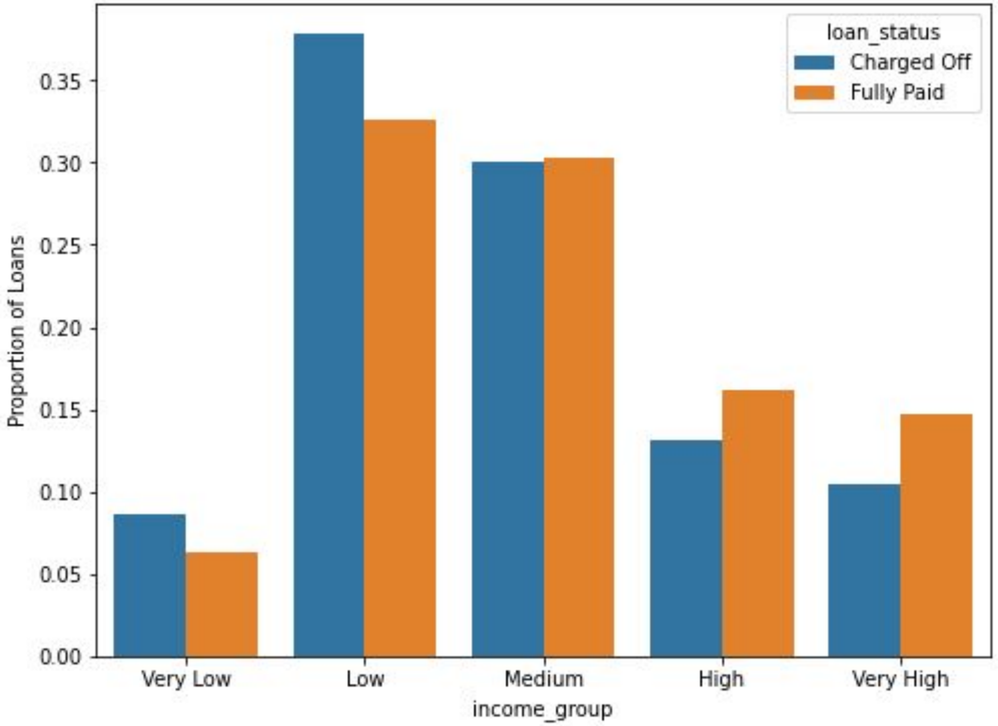




- **Insights**
  - Annual income of borrowers is highly skewed towards lower income values.
- **Recommendation**
  - Bin annual incomes of borrowers into various income groups and further analyze the distribution of loans within each income group.

Data Analysis :

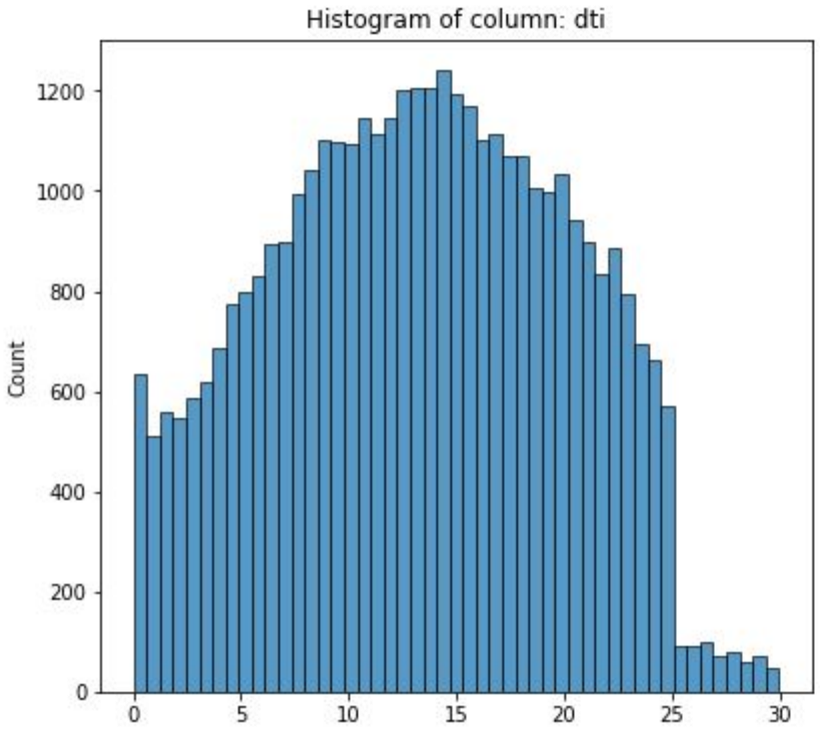
- Relation between ***Annual Income*** and ***Loan Status***

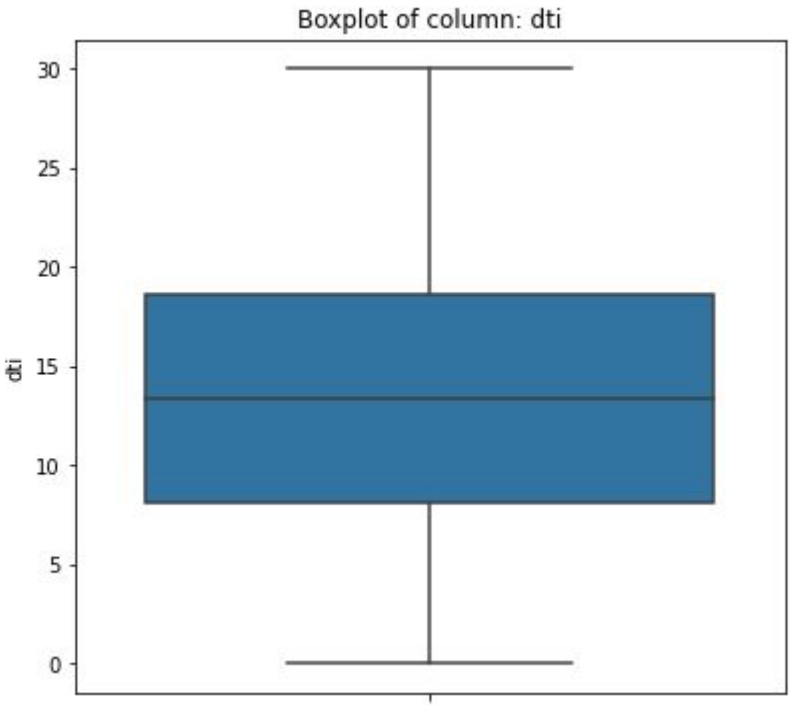


- **Insights**
  - Borrowers belonging to **Very Low** and **Low** income groups (annual_inc <= 50000) are more likely to get default.
  - On other hand, borrowers of **High** and **Very High** income groups (annual_inc > 75000) are less likely to default.
- **Recommendations**
  - More defaulters from the low annual income group

Data Analysis :

- Relation between ***Debt to Income*** and ***Loan Status***

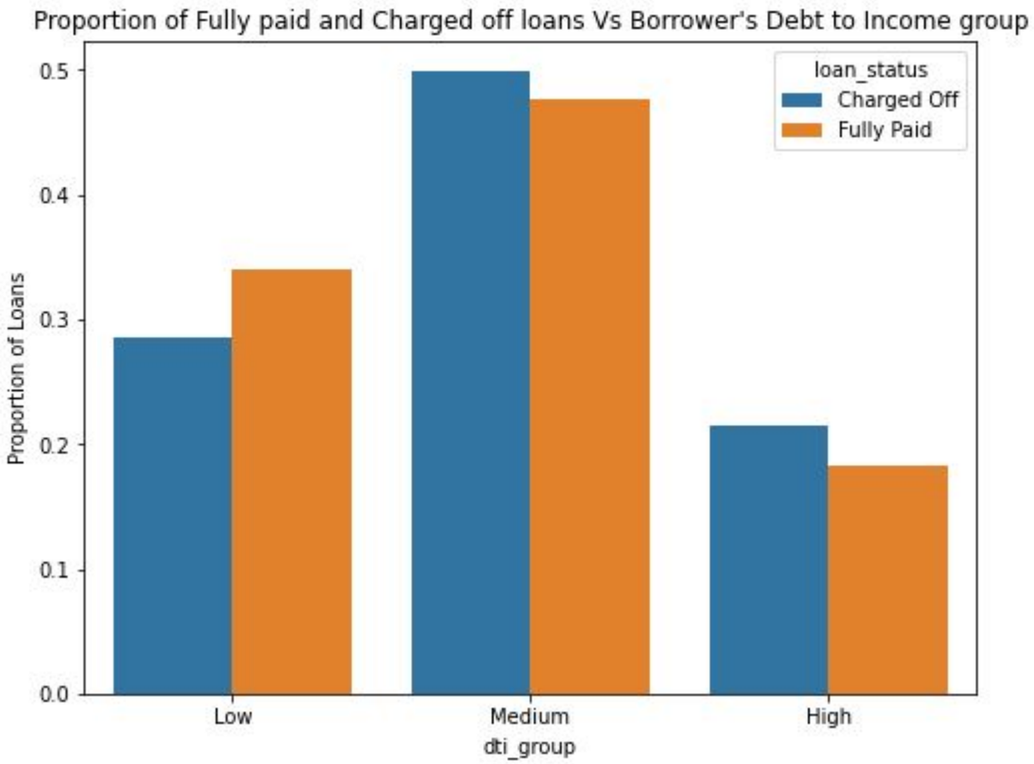




- **Insights**
  - Debt to Income of borrowers ranges from 0 to 30
- **Recommendation**
  - Bin debt to income values of borrowers into various debt to income groups and further analyze the distribution of loans within each debt to income group.

Data Analysis :

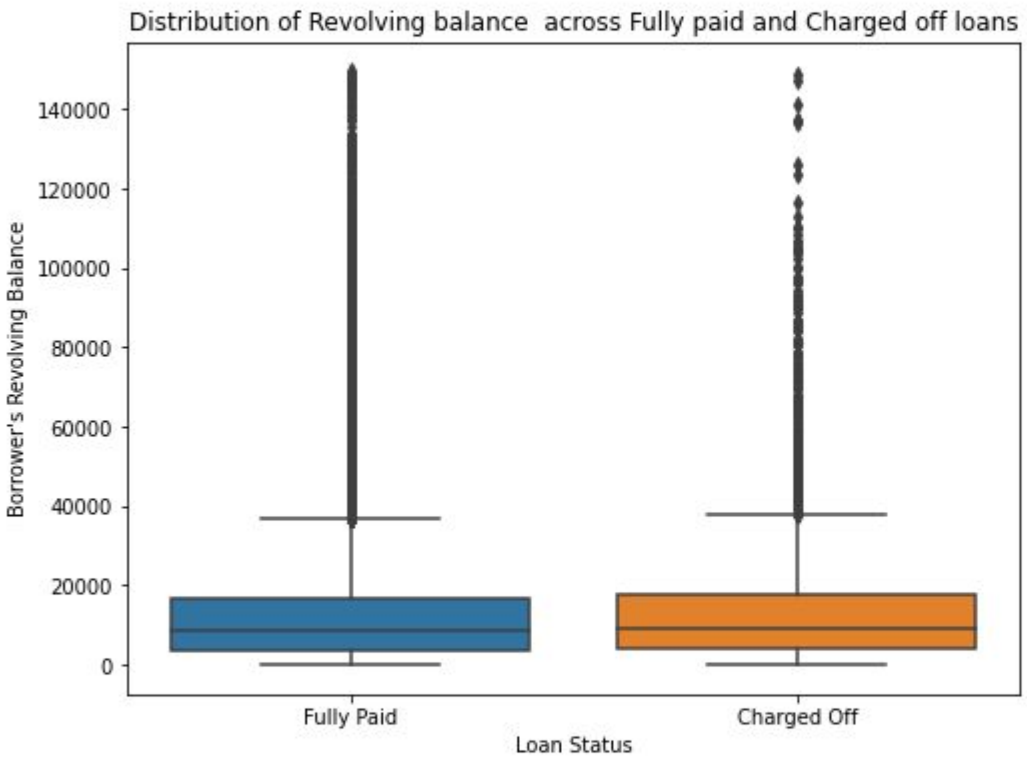
- Relation between ***Debt to Income*** and ***Loan Status***



- **Insights**
  - Borrowers having **Low** debt to income group (dti <= 10) are less likely to get default.
  - On other hand, borrowers having **High** debt to income group (dti > 20) are more likely to default.
- **Recommendation**
  - Target borrowers with low debt to income ratio reduces the risk of default.

Data Analysis :

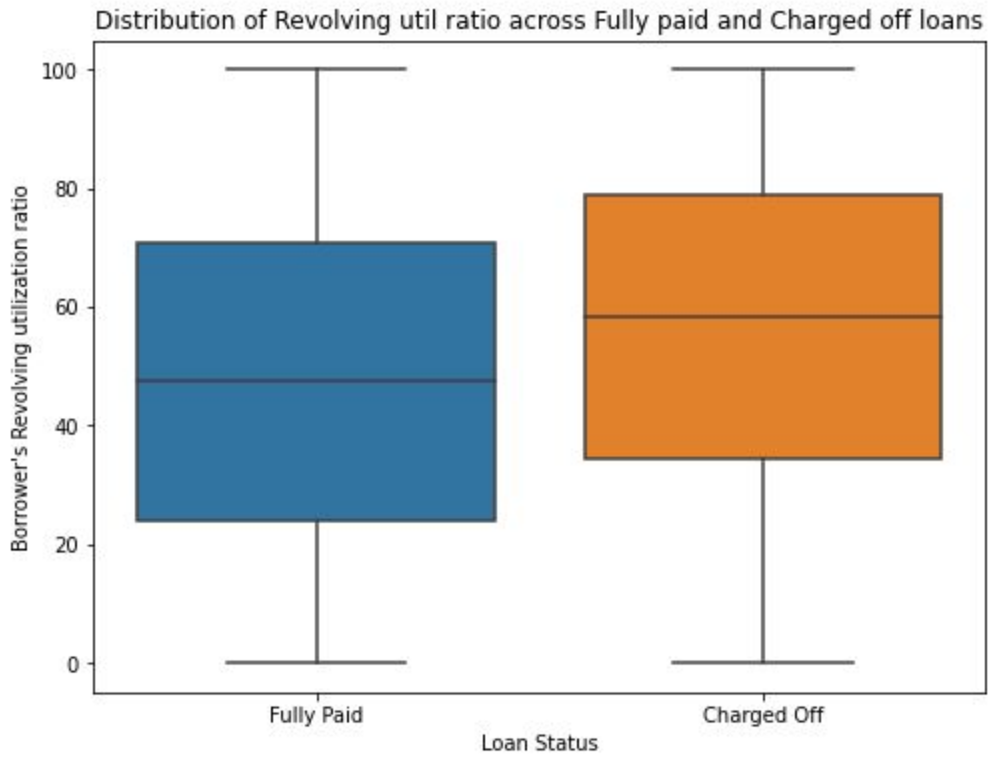
- Relation between ***Revolving Balance*** and ***Loan Status***



- **Insights**
  - Borrowers revolving balance is also highly skewed towards lower income values
- **Recommendation**
  - Bin revolving balance of borrowers into various groups and further analyze the distribution of loans within each group.

Data Analysis :

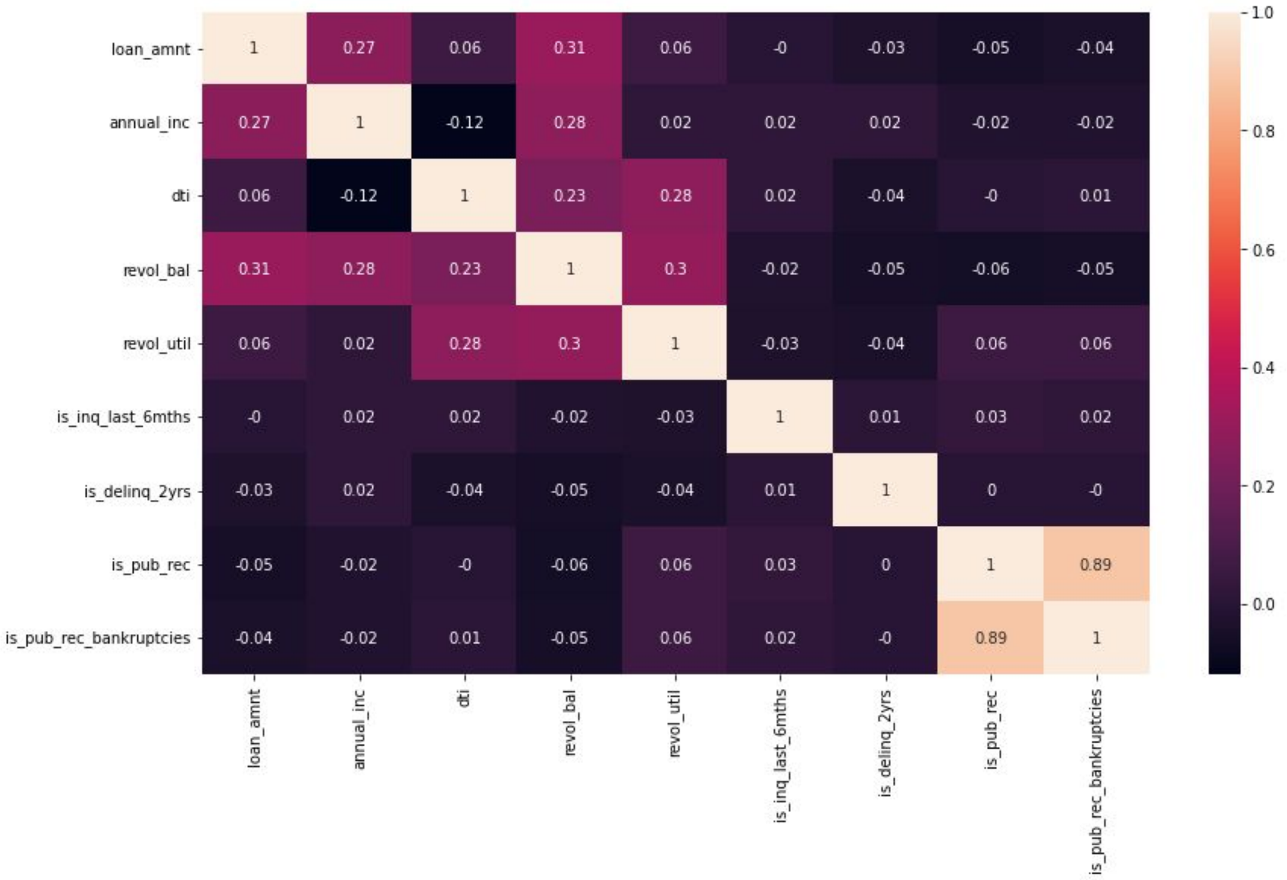
- Relation between ***Revolving Utilization*** and ***Loan Status***



- **Insights**
  - There is a significant difference in distributions of borrower's revolving utilization ratio.
  - A higher revolving ratio (> 60%) indicates higher risk of default.
  - A lower revolving ratio (< 40%) reduces risk of default.
- **Recommendation**
  - Target borrowers with lower revolving ratio for reducing risk of default

Data Analysis :

- Relation correlation across various numerical variables



- **Insights**
  - The variables representing number of public records and number of public bankruptcies of a borrower are highly correlated (correlation = 0.89)
- **Recommendation**
  - Due to presence of high correlation we can either drop one of the variables or derive a ratio of these two variables.

Recommendations :

- **Strong indicators of Default**
  - **Grade**
    - Borrowers with grade 'A' are highly unlikely to default. Targeting these borrowers will reduce the risk.
    - Borrowers with grade 'C' and above are more likely to default.
  - **Revolving utilization ratio**
    - Target borrowers with lower revolving ratio for reducing risk of default
  - **is_inq_last_6mths**
    - Look for borrowers who have not done any inquiries in last six months for reducing the risk of default
  - **Debt to income group**
    - Borrowers having Low debt to income group (dti <= 10) are less likely to get default.
    - On other hand, borrowers having High debt to income group (dti > 20) are more likely to default.
  - **Annual income group**
    - More defaulters from the low annual income group

Recommendations :

- **Weak indicators of Default**
    - **Home ownership**
        - Borrowers staying in rented homes are slightly more likely to default.
    - **Verification status**
        - Verifying the income source details of borrowers reduces the risk of default.
    - **Loan purpose**
        - Borrowers taking loans for purpose of debt consolidation, small business and other are more likely to default.
        - On other hand, borrowers taking loans for purpose of car, credit card, home improvement, major purchase and wedding are more unlikely to default.
    - **Borrower's address state**
        - More number of charged off loans are associated with high cost of living states like California, Texas, Florida, New York, and New Jersey
    - **Revolving balance**
        - Bin revolving balance of borrowers into various groups and further analyze the distribution of loans within each group.
    - **is_delinq_2yrs**
        - The borrower's who were delinquent in last 2 years likely to default. However the difference from non delinquent users is not subtle
    - **is_pub_rec**
        - Check for any public record present before approving loan as the chances of loan getting charged off are very high

Thank You