Lending Club Case Study

SAURABH DUGDHE

Problem Statement

You work for a **consumer finance company** which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

1. If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
2. 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

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Objective

Using Exploratory data analysis, find out what factors drive the defaulter rate in the loan applicants.

EDA Approach

1. Data Understanding
2. Data Cleaning and Manipulation
3. Univariate Analysis
4. Segmented Univariate Analysis
5. Bivariate/Multivariate Analysis

Data understanding

We have been provided 2 files

1. Loan.csv – This files has the dataset of loan applications from the period 2007-2011
2. Data_Dictionary.xls – This is metadata for the loan.csv file. This contains meanings of all the columns present in the dataset.

Upon looking into the dataset we can conclude below:

1. Dataset contains 39717 rows and 111 columns
2. ‘loan_status’ is the target variable for the analysis
3. Multiple columns have null values
4. There are some similar columns like ‘purpose’ and ‘title’

Data Cleaning/Manipulation

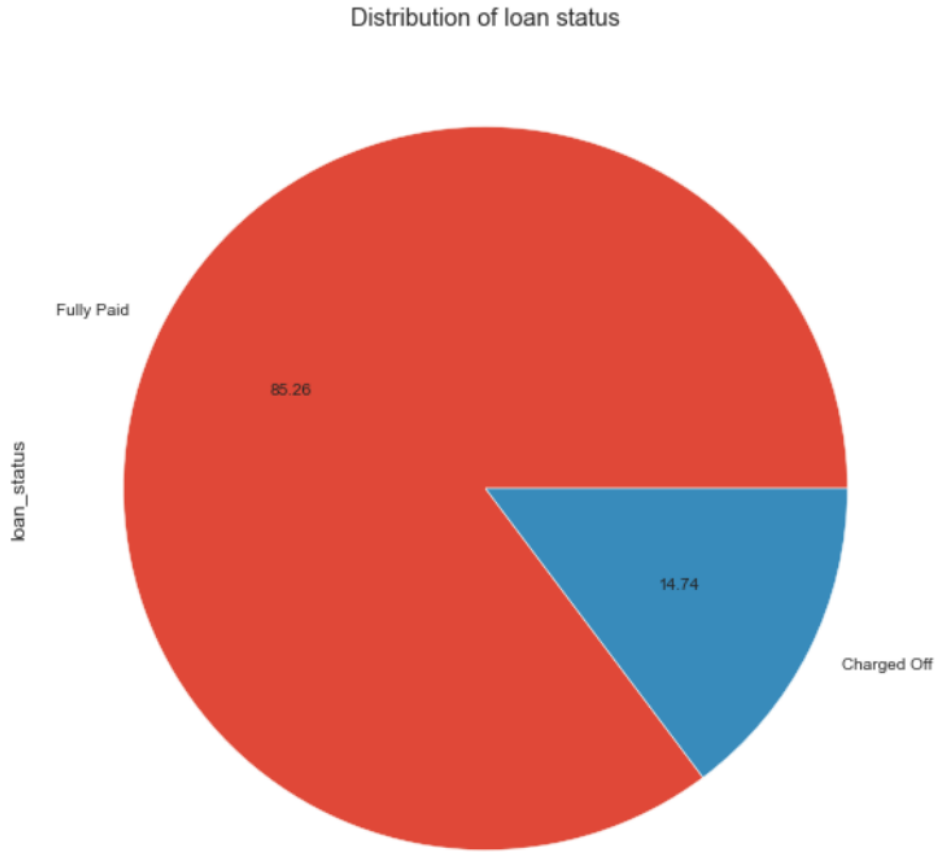
1. Drop unnecessary columns
    1. Removed columns that do not help us in identifying defaulter status of an applicant. E.g. **url, desc, zip_code, id, member_id, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_int, total_rec_late_fee, total_rec_prncp**
    2. Removed columns that had only one unique value through out the table. E.g. **pymnt_plan, policy_code**.
    3. Removed columns with mean, min, max are all same.

2. Standardize columns
    1. % sign is removed from **int_rate** and **revol_util** and type-casted to float
    2. '**term**' column is standardized by removing keyword 'months' from it.
    3. '**emp_length**' is standardized as per the requirements in data dictionary and has values ranging from 0-10 and removing "+", "<" from the values.
    4. Standardized **home_ownership** column by removing 'None' value as it is not mentioned as a valid value in the data dictionary.
    5. Renamed **emp_title** to **job_title** as the data dictionary defines it to be job title that is supplied during application.
    6. Standardized top 15 entries of **job_title** as there are multiple values pointing to the same applicant. E.g. Us Army, U.S. Army, us army are basically the same applicant
    7. Excluded value "Current" from '**loan_status**', as it does not imply the defaulter status.

Data Cleaning/Manipulation

3. Handle null values
    1. Dropped columns with more than 60% of null values
    2. Replaced null values in '`emp_length`' and '`revol_util`' with corresponding median values
4. Derived metrics
    1. '`issue_month`' and '`issue_year`' are derived from issue date i.e '`issue_d`'
5. Outlier treatment
    1. Outliers are treated by removing 5% of data from columns '`annual_inc`' and '`revol_bal`'

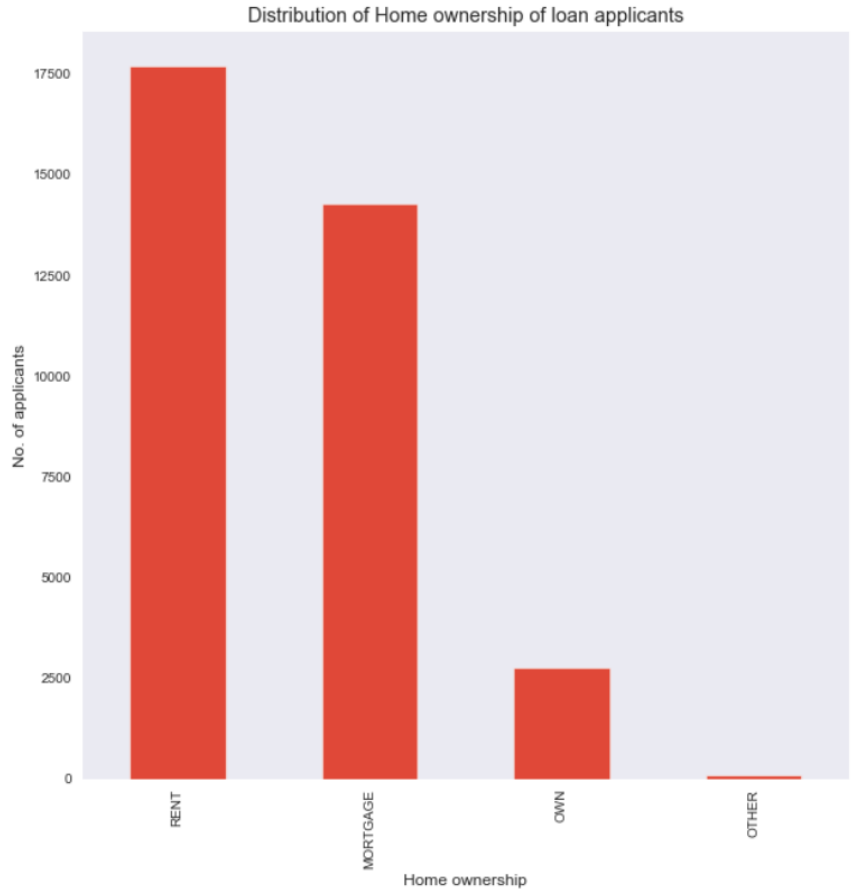
Univariate Analysis



Distribution of target variable

The dataset consists of **86.26%** of **Fully paid** the loan and **14.74%** have been **charged off**.

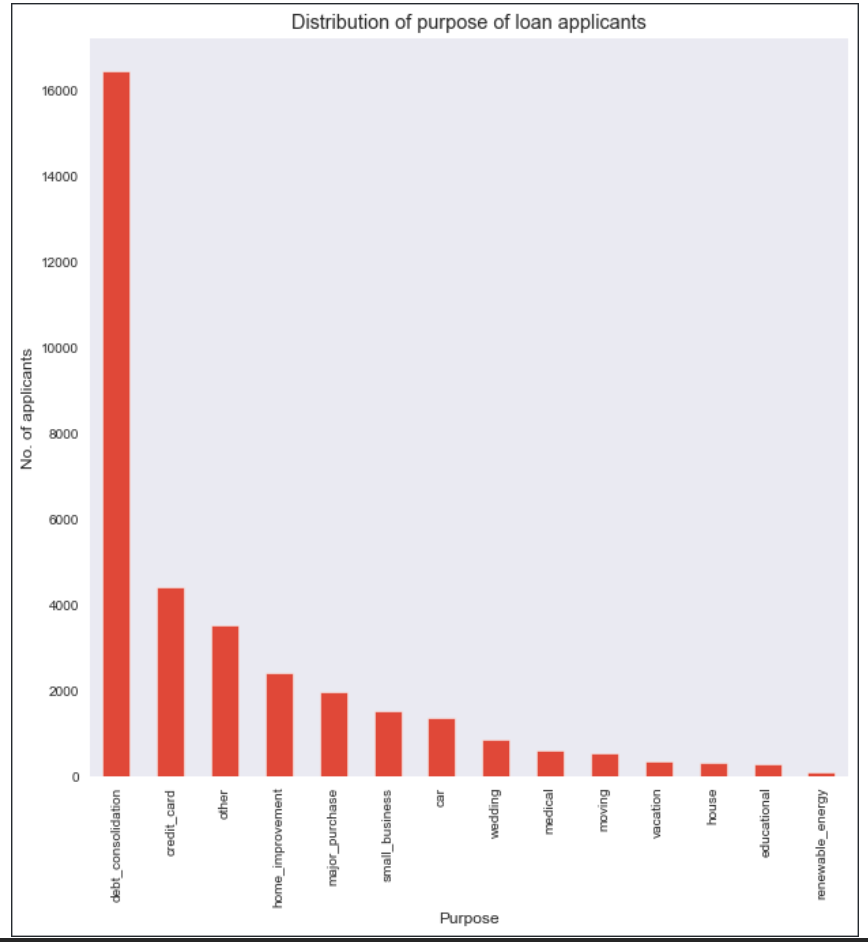
Univariate Analysis



Distribution of Home Ownership

Applicants with rented housing are more likely to request for a loan than the ones who own a property.

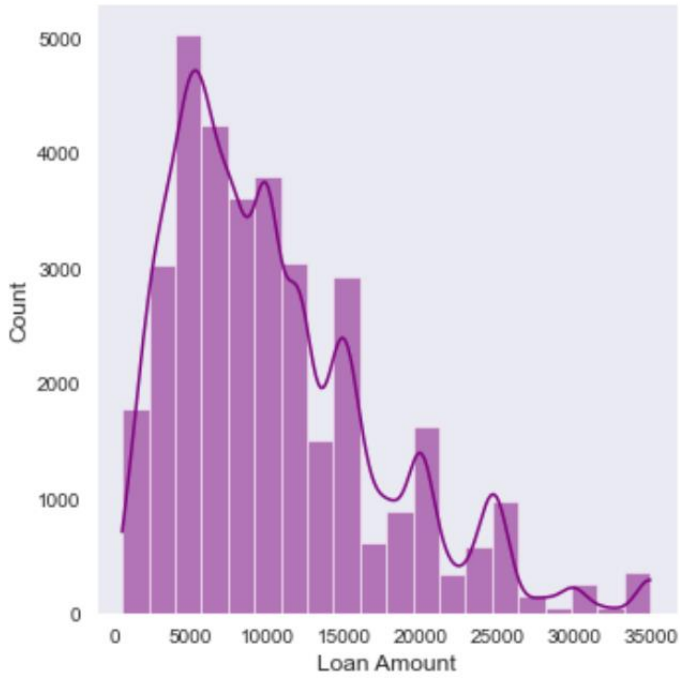
Univariate Analysis

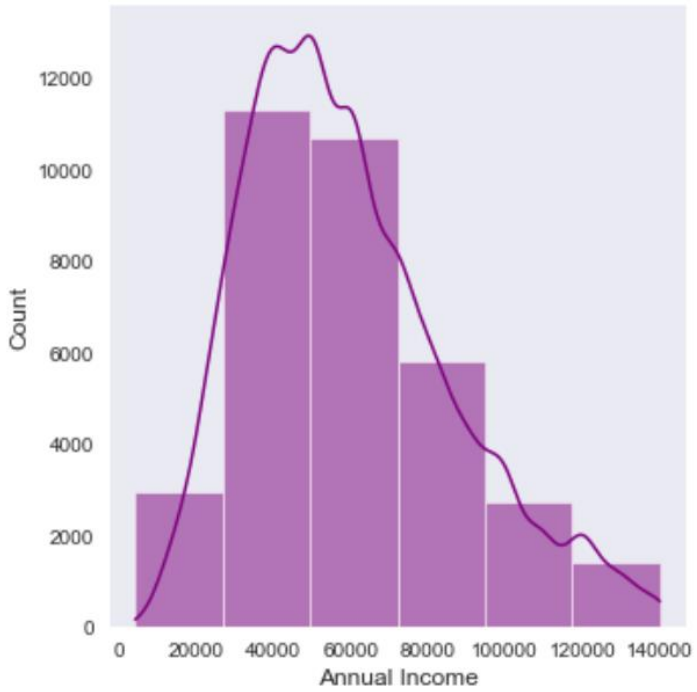


Distribution of purpose

Debt consolidation is the major reason driving people to apply for loan, followed by **credit card** payments. The least driving reason is education and renewable energy.

Univariate Analysis



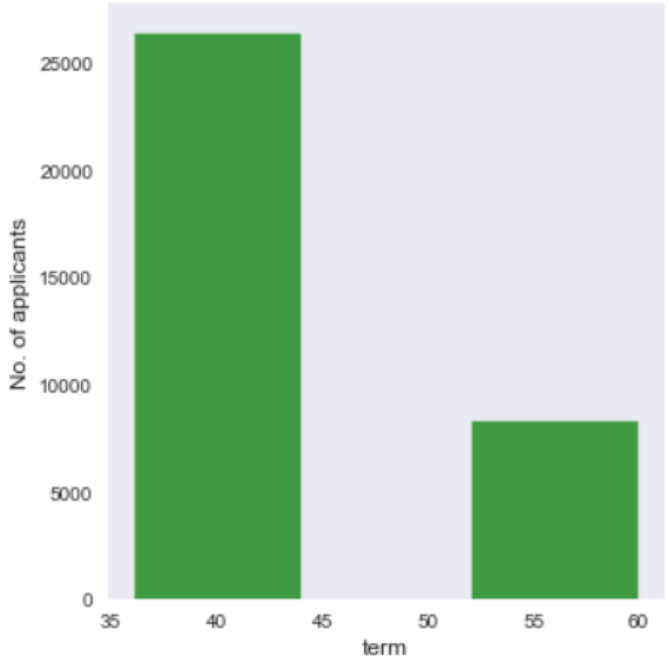


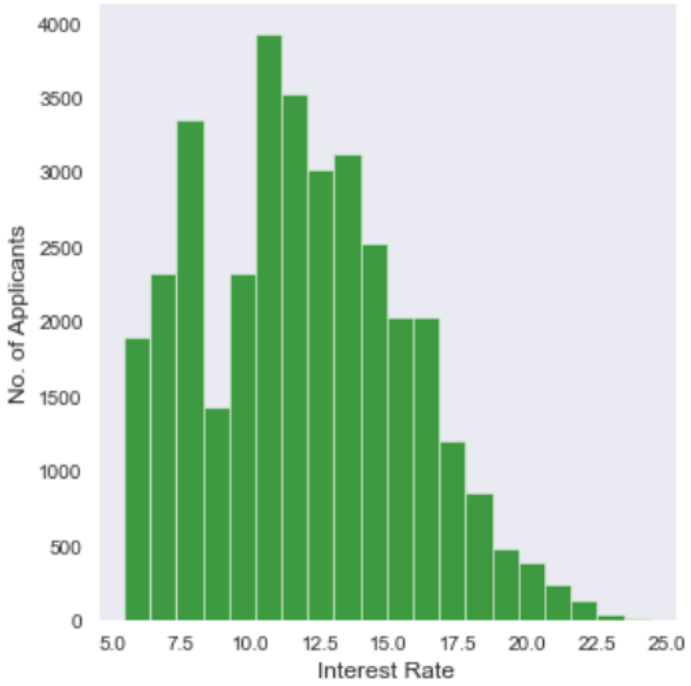
Distribution of loan amount and Annual income

Highest number of loan applicants tend to apply a loan of amount in the range of 5000-10000

Major loan applicants fall in the annual salary bracket of 30000 - 80000

Univariate Analysis



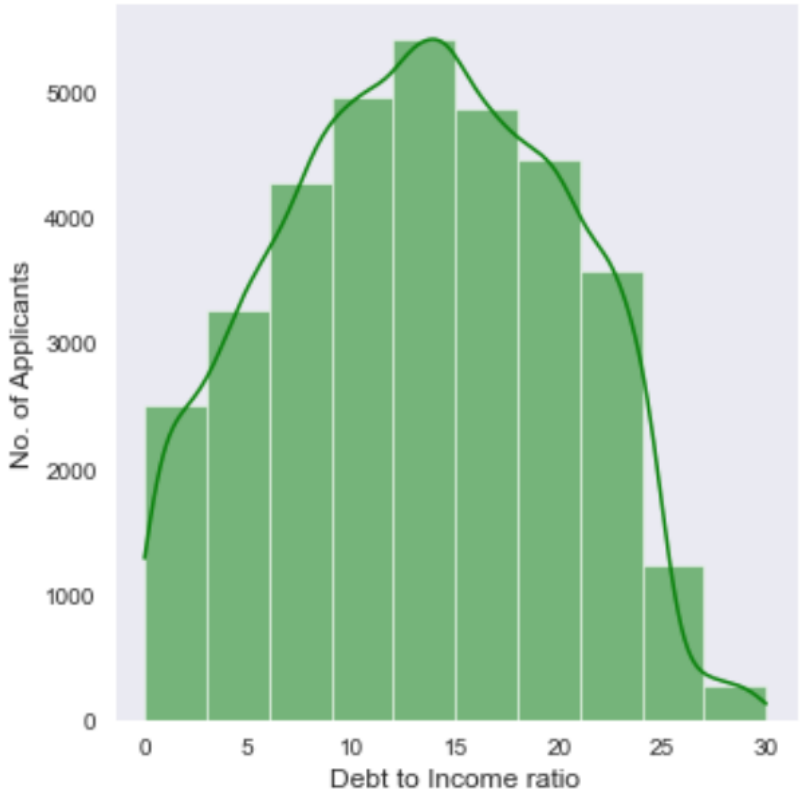


Distribution of Term and Interest rate

Majority of people apply for a loan of term of **36 month.**

Most applicants fall in the interest range of **10%-12%.**

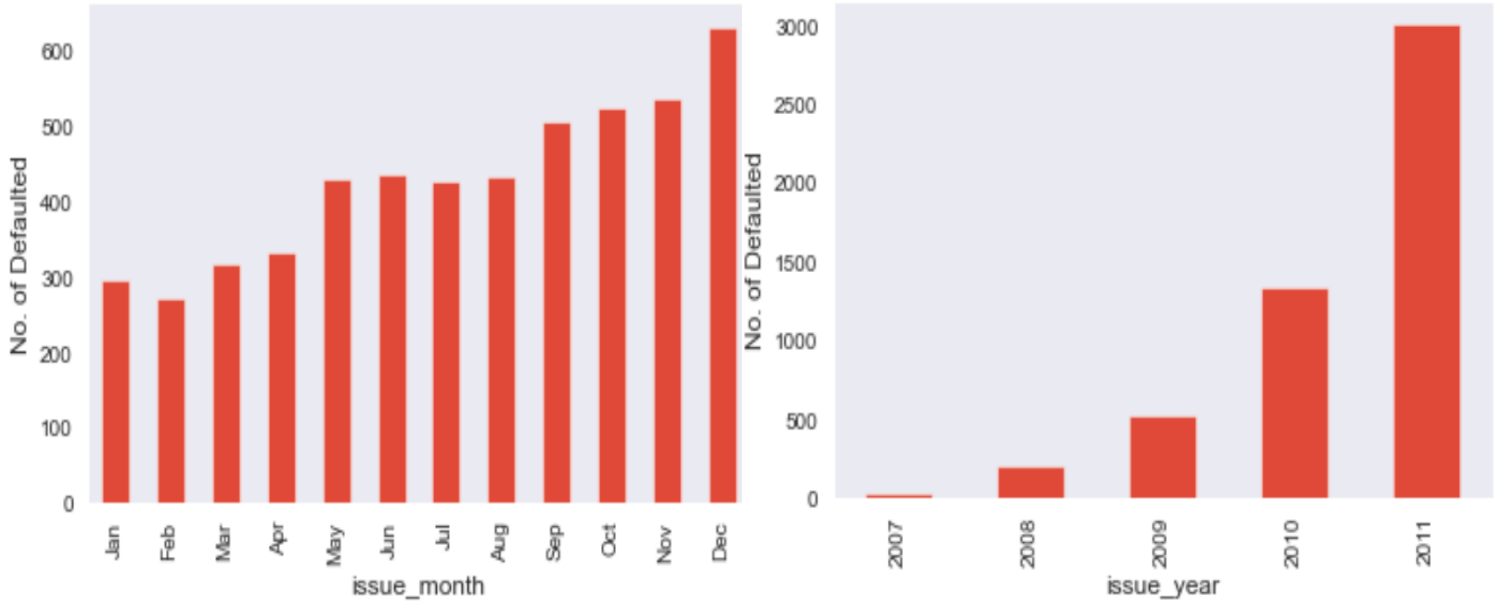
Univariate Analysis



Distribution of debt to income ratio

Majority of Debt to income ratio of applicants lies between 10-20%.

Segmented Univariate Analysis

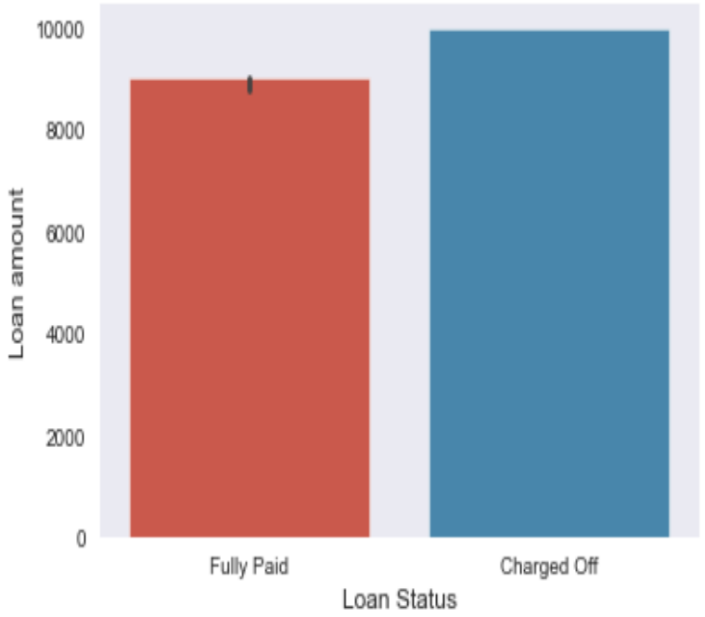


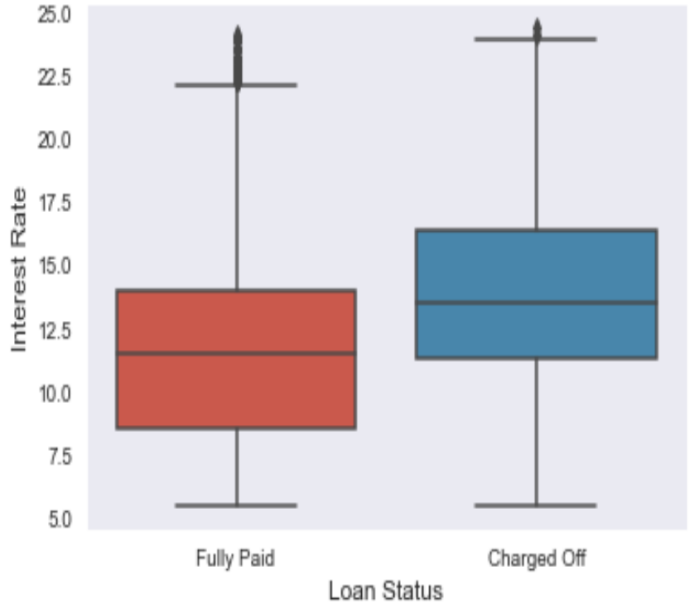
Defaulting rate as per issue month and year

1. Loans issued at the start of the year are less likely to default than the ones issued in the last quarter.

2. Loan defaulting rate is increasing every year

Bivariate/Multivariate Analysis

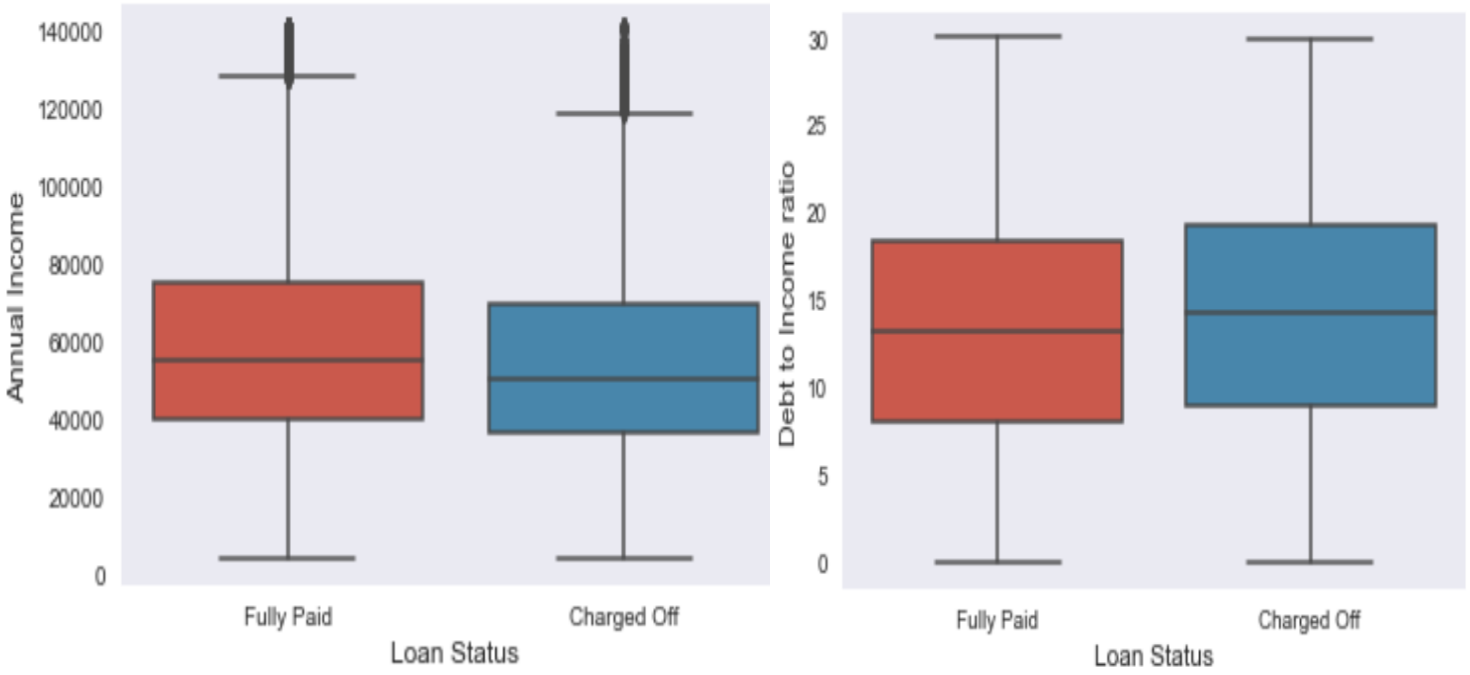




Impact of loan amount and interest rate on loan status

Loans with higher amount and higher interest rate tend to be defaulted.

Bivariate/Multivariate Analysis

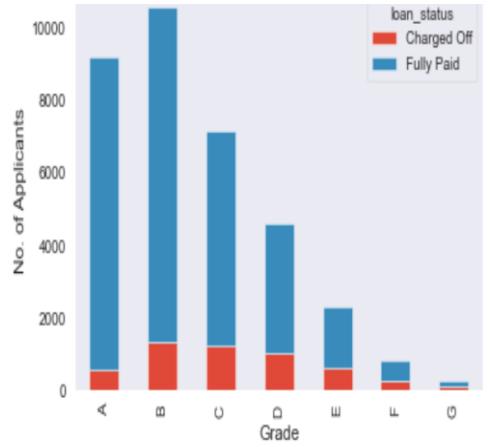


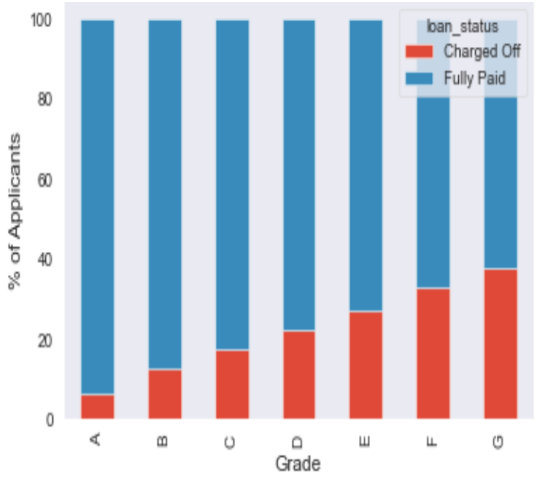
Impact of annual income and DTI

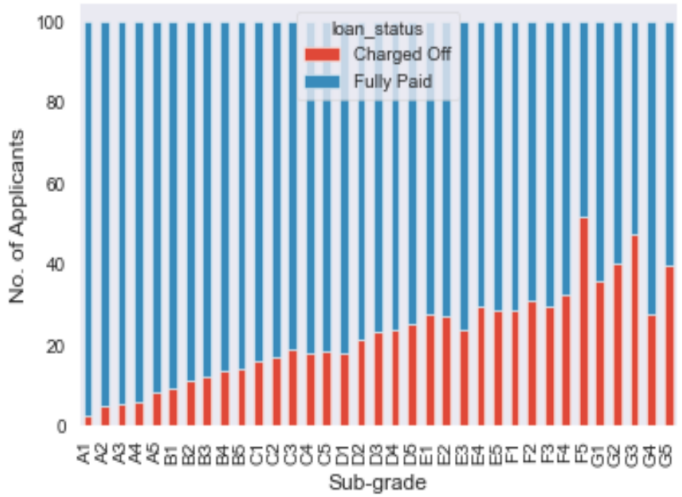
1. Applicants with **lower income** tend to **default** the loans. While, applicants with higher annual income will fully pay the loan back.

2. Applicants with **higher DTI** tend to **default** than the applicants with lower DTI.

Bivariate/Multivariate Analysis







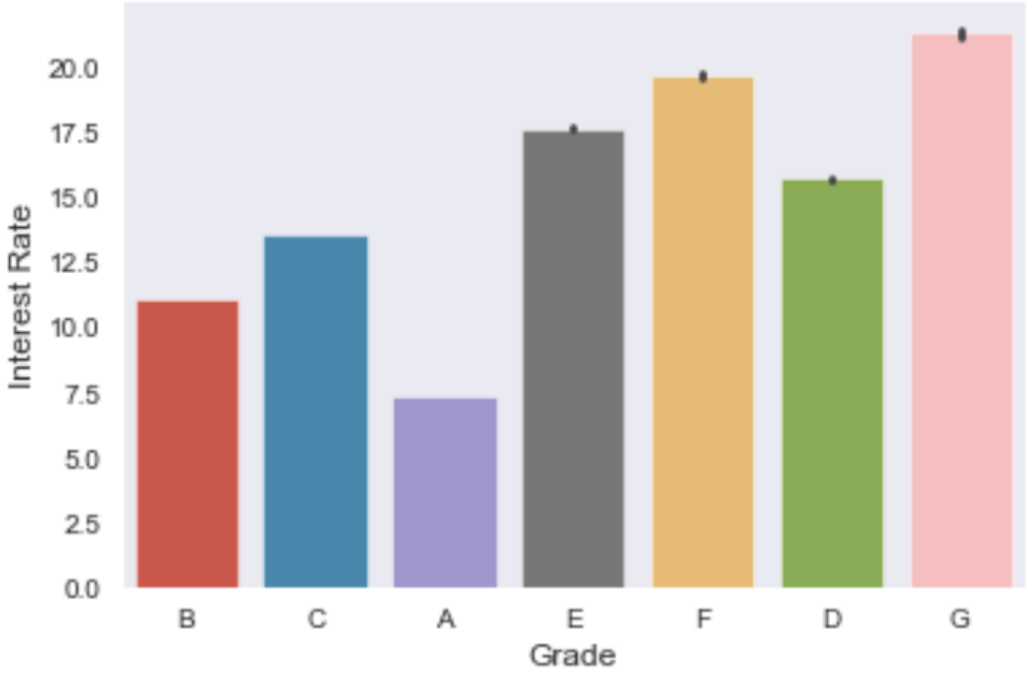
Impact of grade on loan status

1. Maximum number of people fall in group B followed by A. The least number of people have grade G.

2. **Grade A** has the **least percentage** of **defaulters**. While, **grade G** has the highest percentage of **defaulting** rate

3. Likewise, sub grades of **A** have **low defaulting** rate. While subgrades of **F & G** have **high defaulting** rate.

Bivariate/Multivariate Analysis

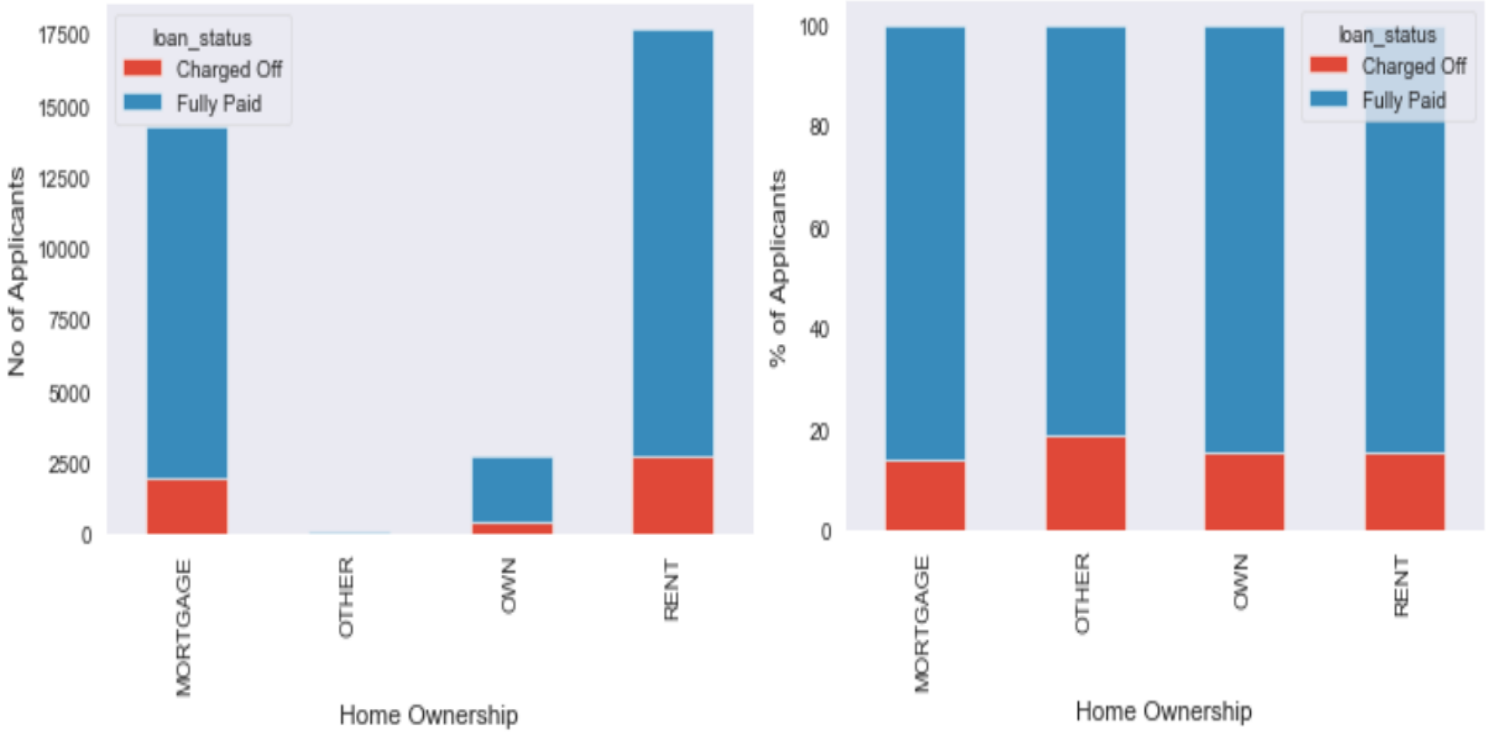


Variation of interest as per grades

1. Grades G and F have high interest rates, this could be the reason of higher defaulting rate for these grades.

2. Grade A has the least interest rate and hence the defaulting rate.

Bivariate/Multivariate Analysis

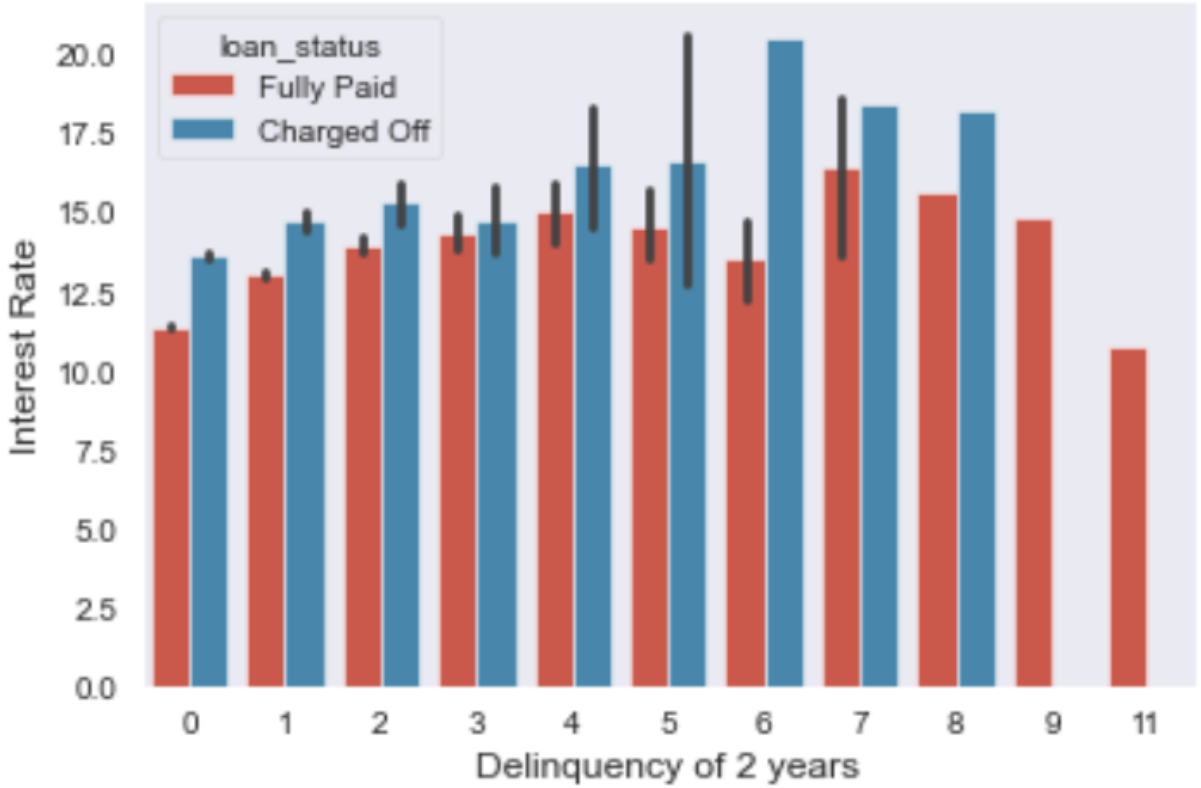


Impact of Home Ownership on loan status

1. Applicants from rent and mortgage categories are the highest.

2. Defaulting rate of other and owners is higher than that of rent and mortgage.

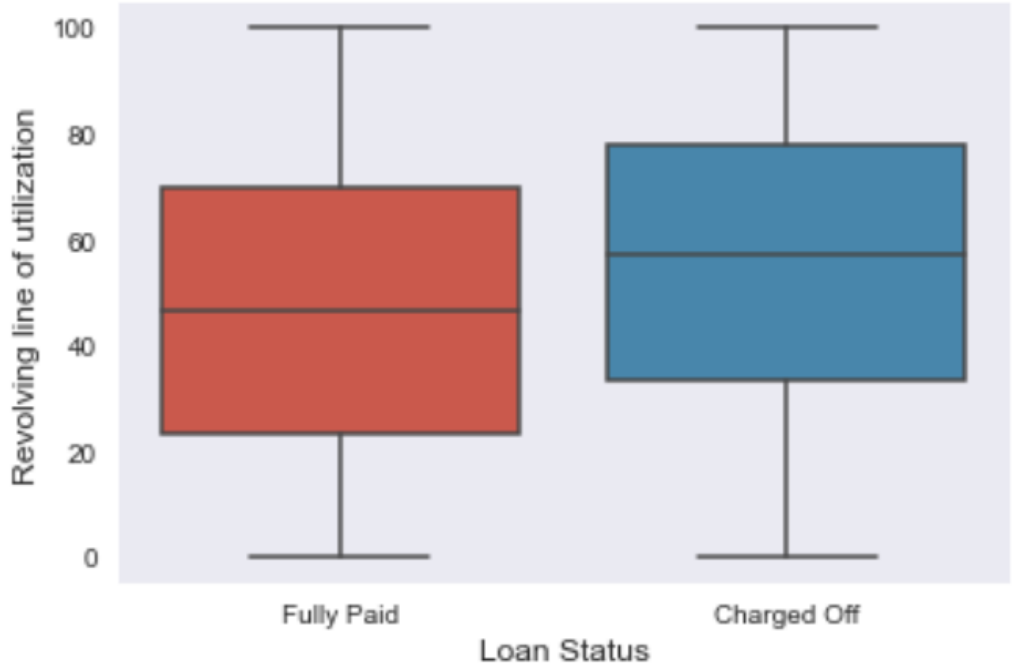
Bivariate/Multivariate Analysis



Variation of loan status on interest rate and delinquency of 2 years

Defaulted applicants with delinquency in past 2 years have higher interest rates.

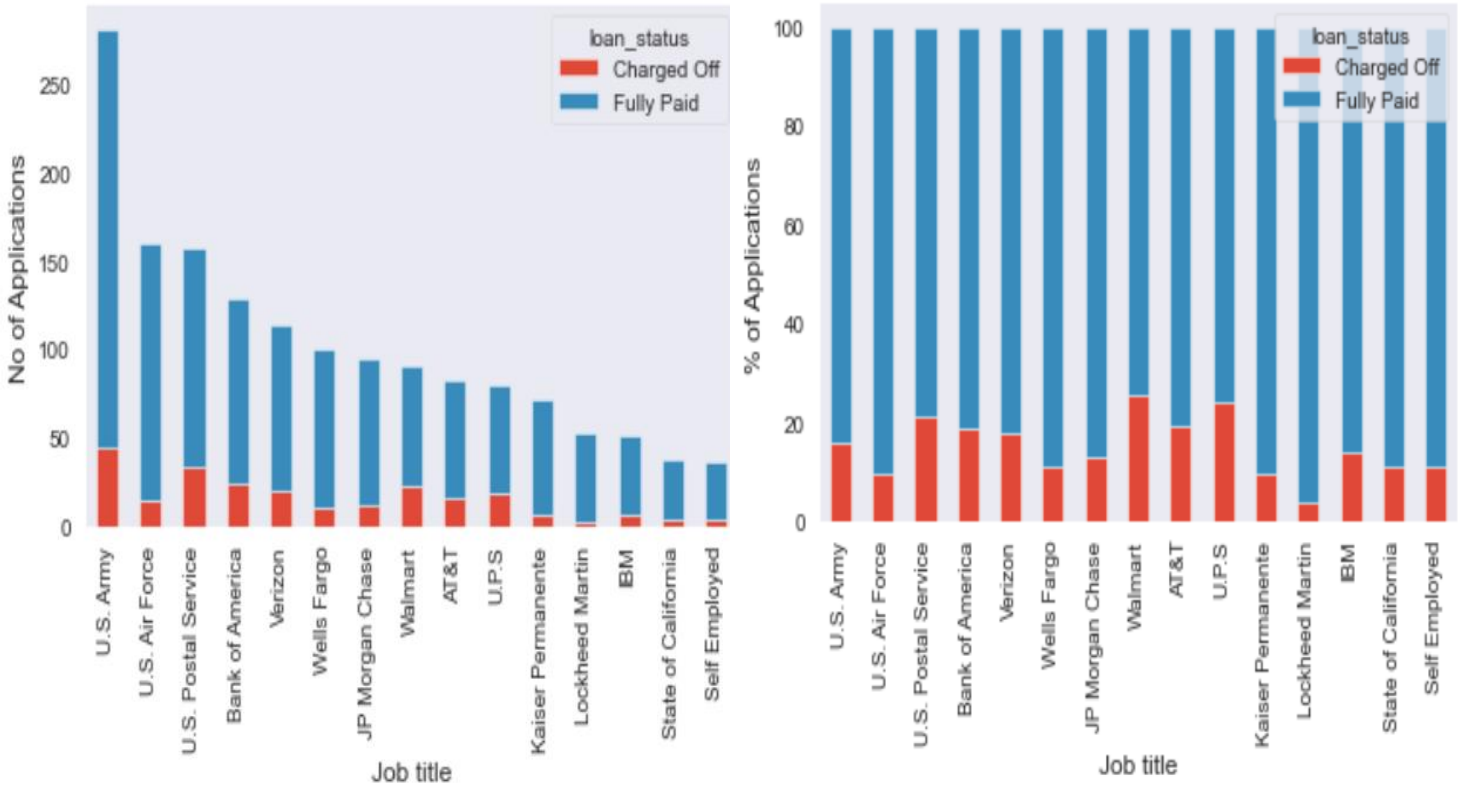
Bivariate/Multivariate Analysis



Impact of revolving rate of utilization on loan status

Applicants with revolving rate of utilization of more than ~50% are likely to default their loans

Bivariate/Multivariate Analysis



Defaulting rate as per job title of the applicant

1. We can see that **U.S. Army** is the most **frequent** loan applicant, while **self-employed** are the **least.**

2. Out of the top 15 applicants **Walmart** is the **top defaulter**, while **Lockheed Martin** generally pays the loan **fully.**

Conclusion

Following variables drive the defaulter rate.

1. **Loan Amount:** With increase in loan amount the defaulting rate increases.
2. **Interest Rate:** With the increase in interest rate, defaulting rate also increases.
3. **Annual Income:** Defaulting rate increases with lower annual income.
4. **DTI:** With higher DTI, the defaulting rate increases.
5. **Grade:** Defaulting rate increases as we move from grade ‘A’ to grade ‘G’.
6. **Revolving rate of utilization:** With increase in the revolving line of utilization, the defaulting rate increases

Recommendations/Proposals

1. Sanction loans of small amounts
2. Extremely high interest rates must be avoided
3. Decrease loans for applicants with low annual income
4. Approve loans for applicants with lower debt
5. Consider loaning often to grades A, B, C over F, G
6. Decrease loaning to applicants with high revolving line of utilization