# Lending Club Case Study

Exploratory Data Analysis

# Summary

- Problem Statement
- Data Summary
- Data Cleaning
- Data conversions
- Derived Columns
- Outliers
- Univariate Analysis
- Bivariate Analysis
- Correlation

# Problem Statement

**Problem**:

- There is 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:
- 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
- Need to EDA to understand how consumer attributes and loan attributes influence the tendency of default.

# Data Summary

- Loan.csv file contains 39717 rows and 111 columns.
- There are two types of attributes Loan Attribute and Customer attributes.

# Data Cleaning

- There was no header or footer rows present which need to be deleted.
- Deleted the rows which have "loan_status" as "Current". Lender who are still paying loans, they can fully pay the loan or can be charged off. These rows will not help us make decision.
- Deleted the columns which are having all the values as Null
- Deleted the columns 'member_id' and 'url'
- Deleted the columns which are having values as text/description as these columns will not contribute to EDA
- Deleted the columns not available during loan approval process, like 'earliest_cr_line', 'last_pymnt_amnt' etc.
- Deleted the columns which are having more than 40% of values as null.
- Two columns were having null values still .The percentage of the null values was very less, so dropping the rows - 4.484537418669155 %.

# Data Conversion

- Removed months from the column ''term'' and converting the data type into int from object
- Removed "%"  from "int_rate" and converting the data type into float
- Converted 'issue_d' to date
- Converted 'emp_length' to integer

# Derived Columns

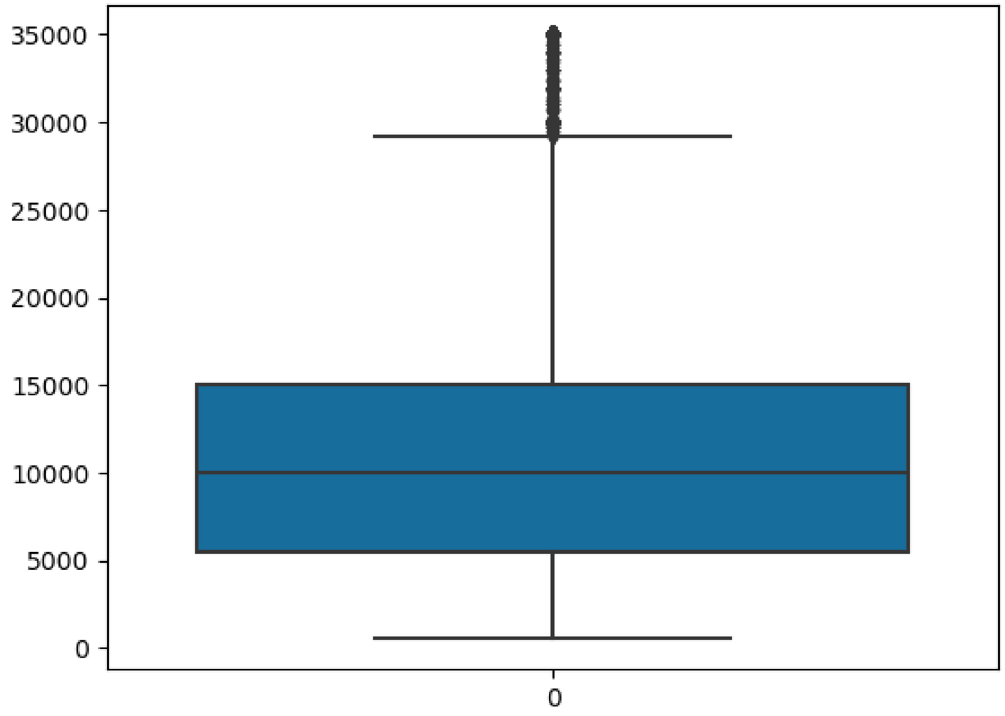
- Derived columns for issue month and issue year from "issue_d"

# Univariate Analysis

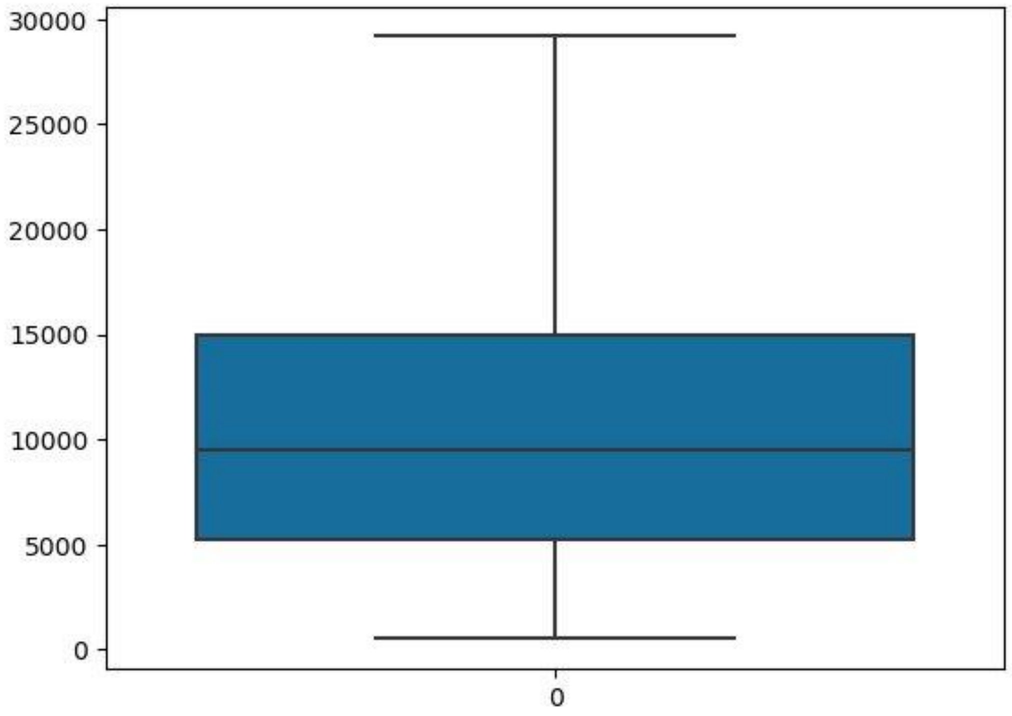
# Univariate Analysis

- Used Box Plot to analyse the distribution and removed outliers.
- Below are the before and after removing outliers Box Plot for the column 'loan_amnt'.

**Before**

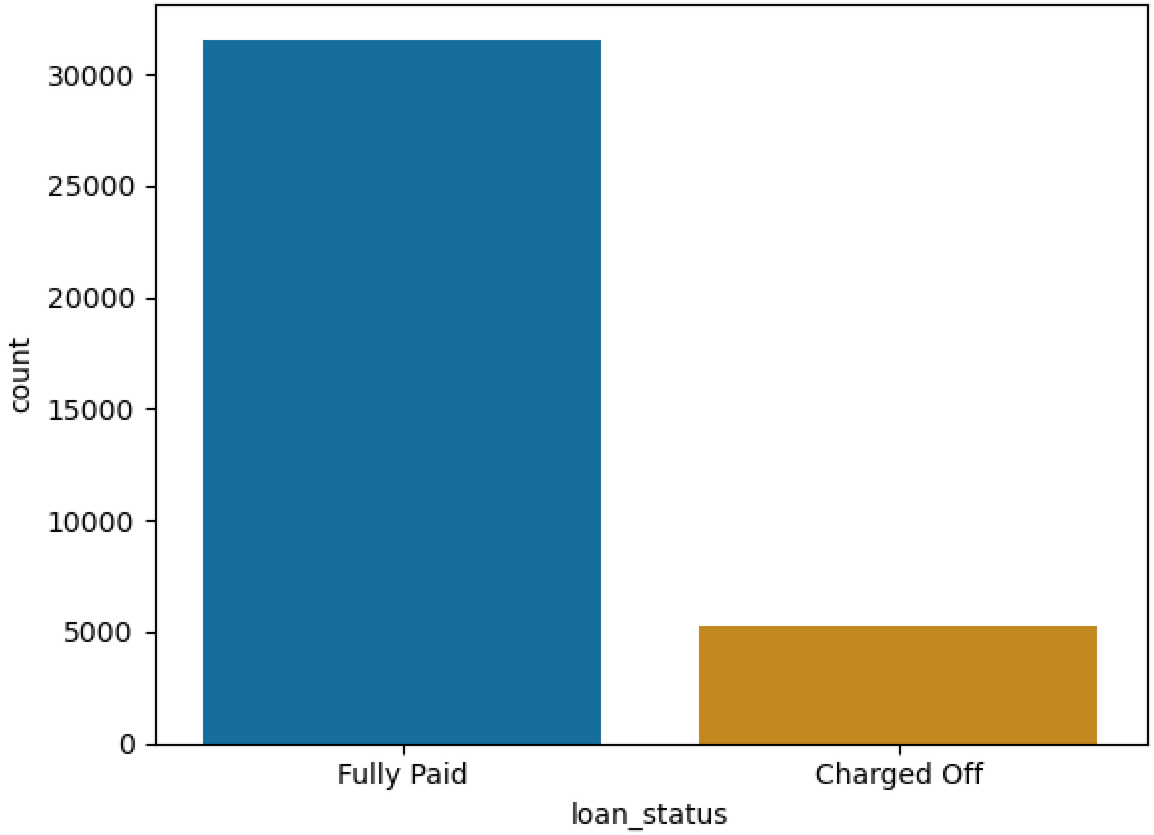


# After



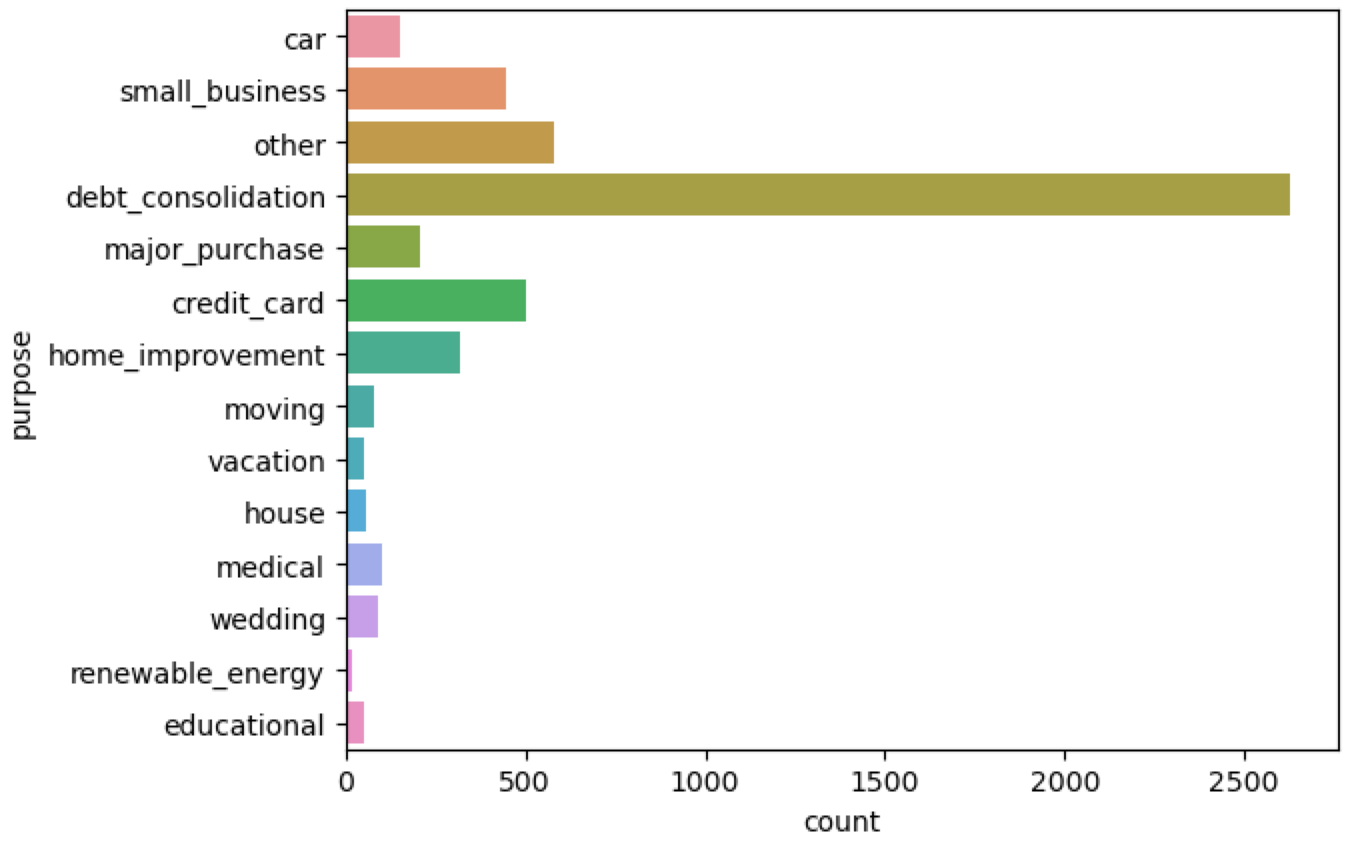
# Loan Status

- The number of fully paid loan is more than the number of charged off loan.



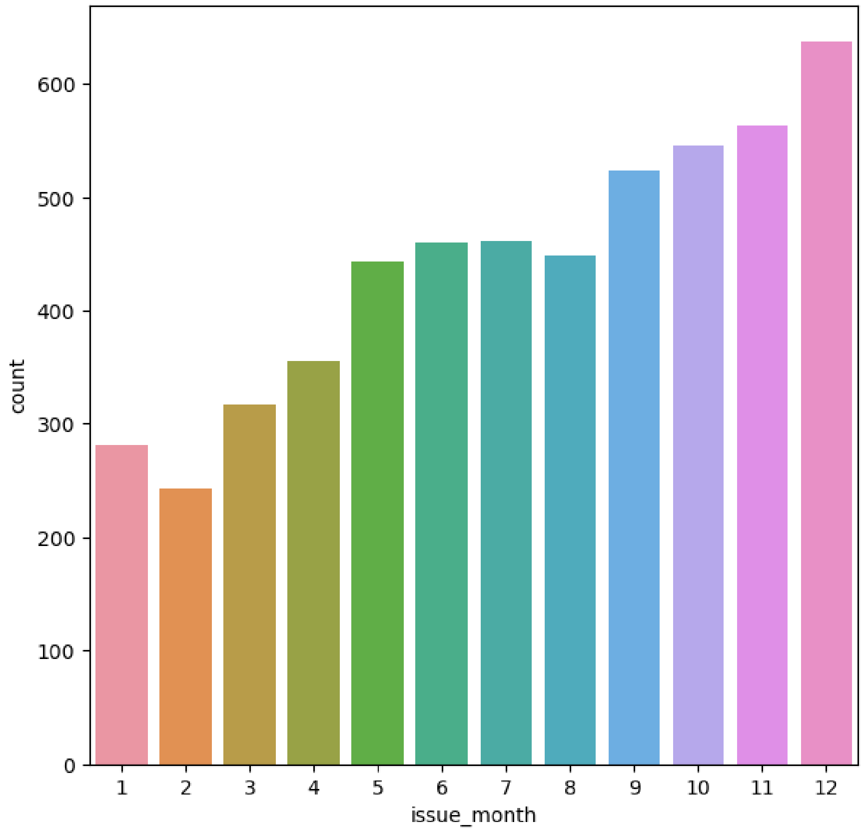
# Purpose

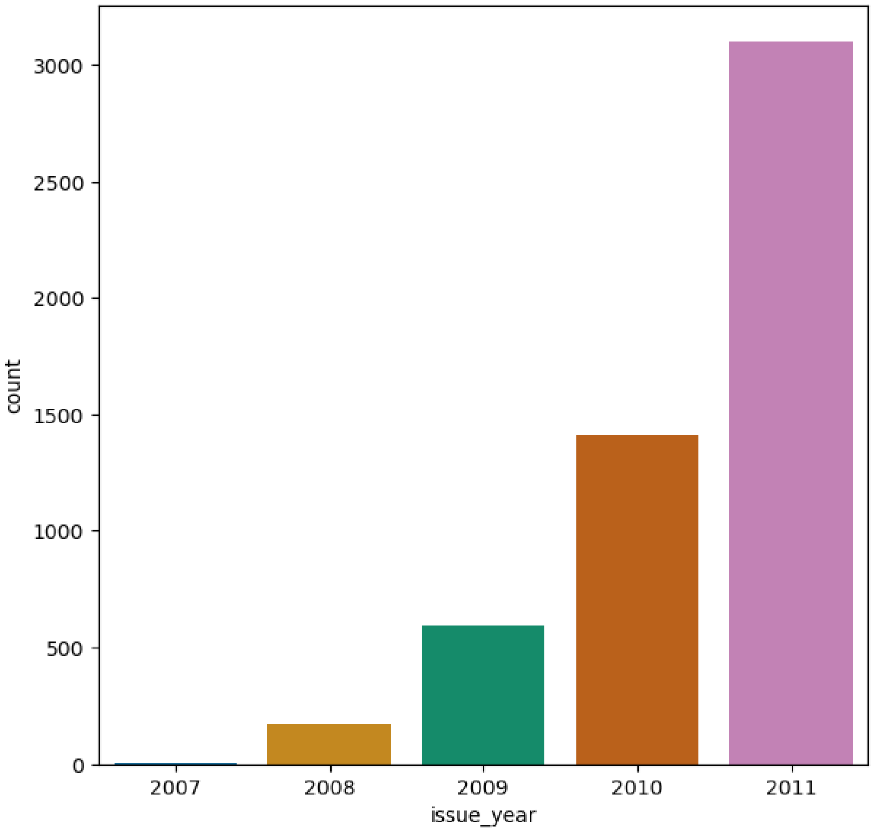
- Analysed the values of 'purpose' based on the value of 'loan_status' as charged off. It is clearly visible that when the 'purpose' field with value 'debt_consolidation' is having most 'Charged Off' loans



# Issue Year and Month

- Analysed the distribution of the years and months when the charged off loan was issues.

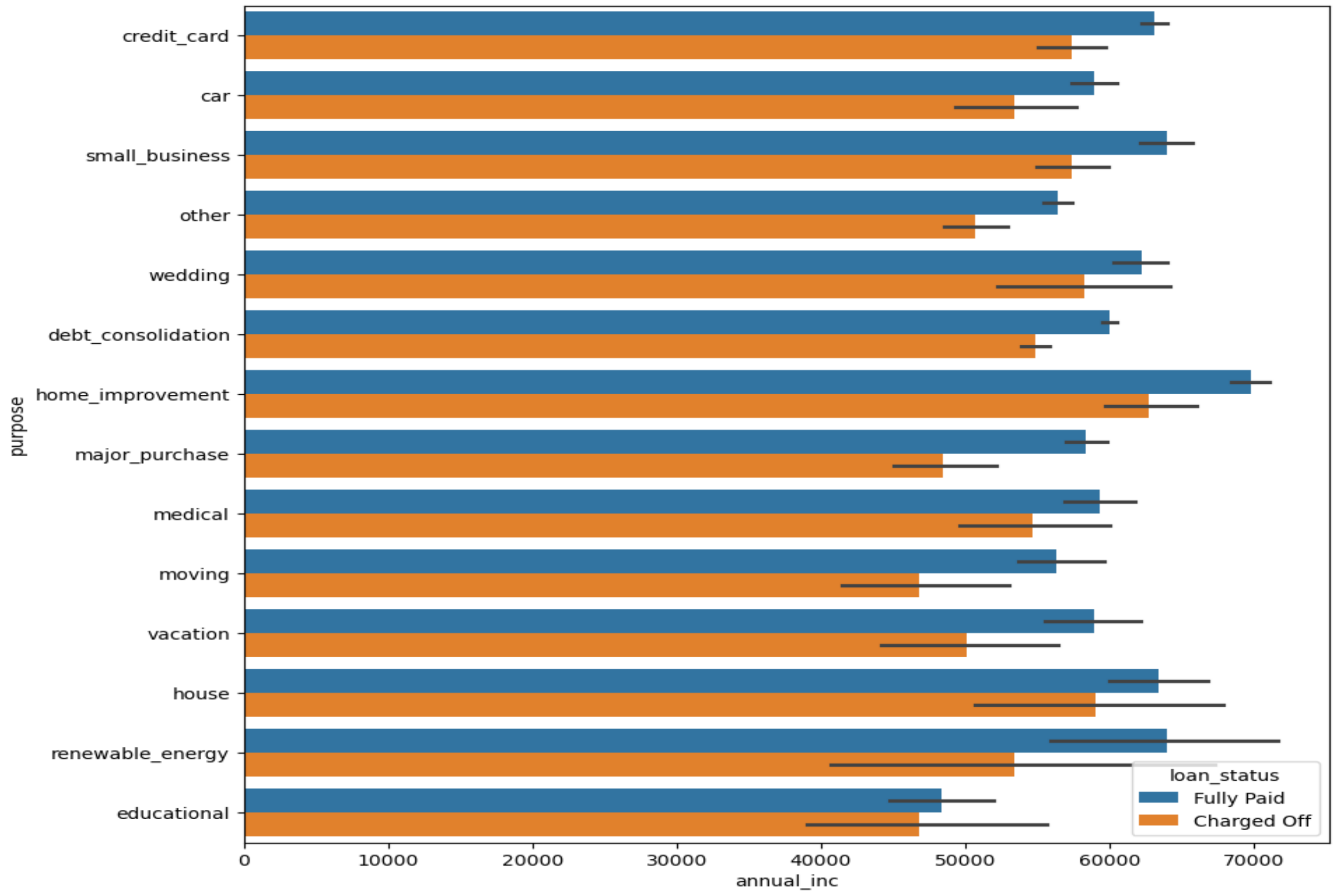


# Bivariate Analysis

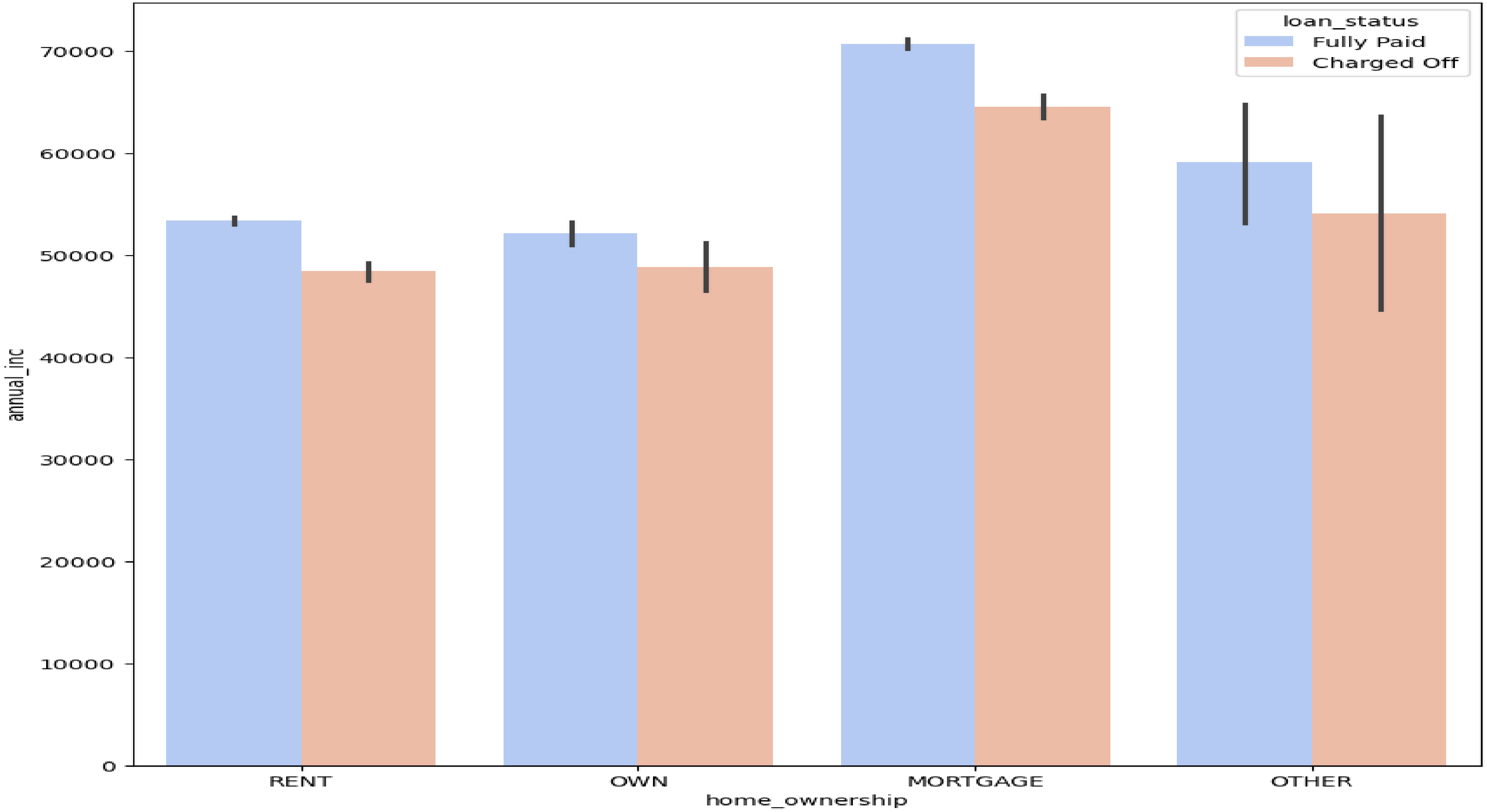
# Annual Income vs loan purpose

- Applicants with higher salary mostly applied loans for "home_improvment", "house", "renewable_energy" and "small_businesses"



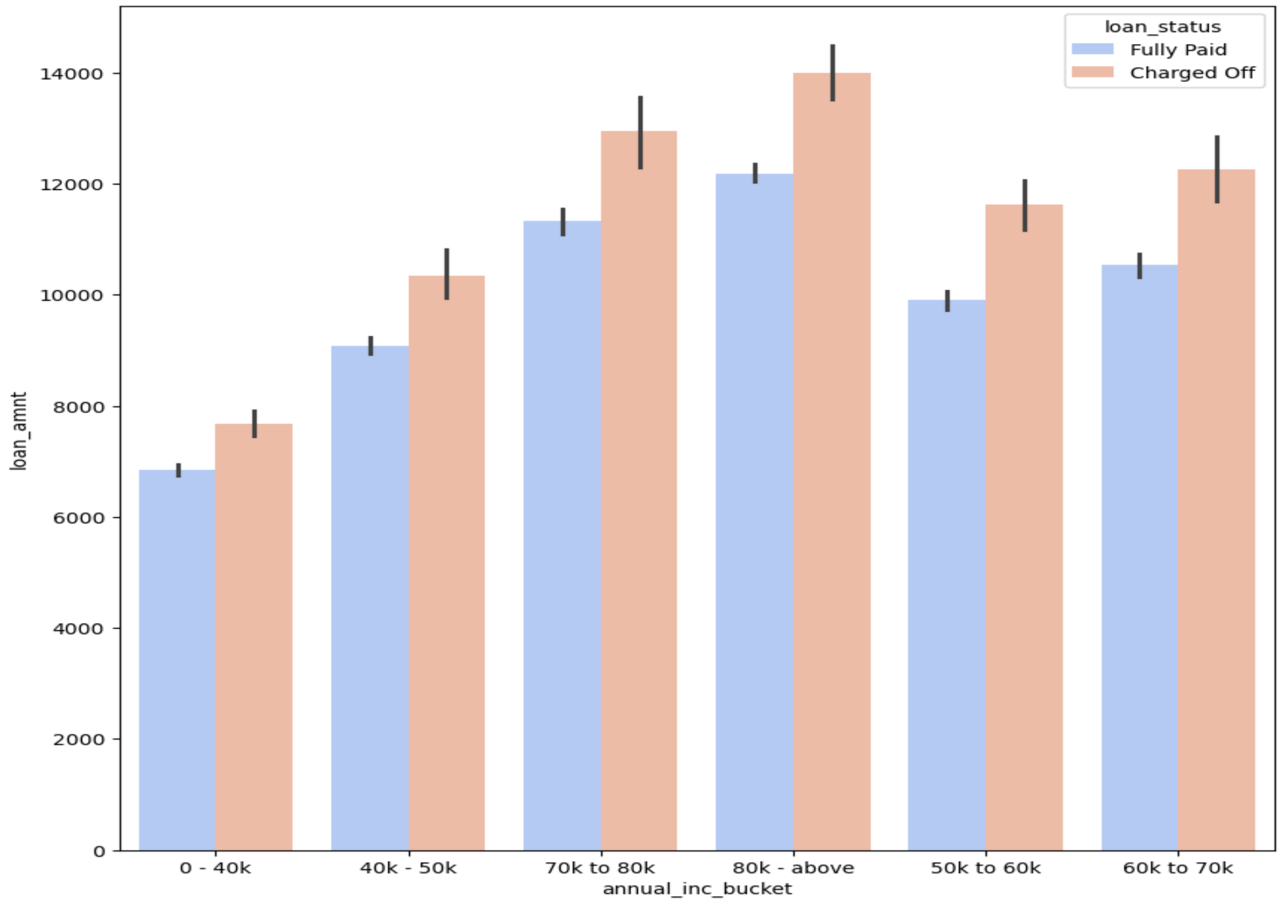
# Annual Income vs Home Ownership

- Applicants with higher salary mostly have home ownership status as 'MORTGAGE'



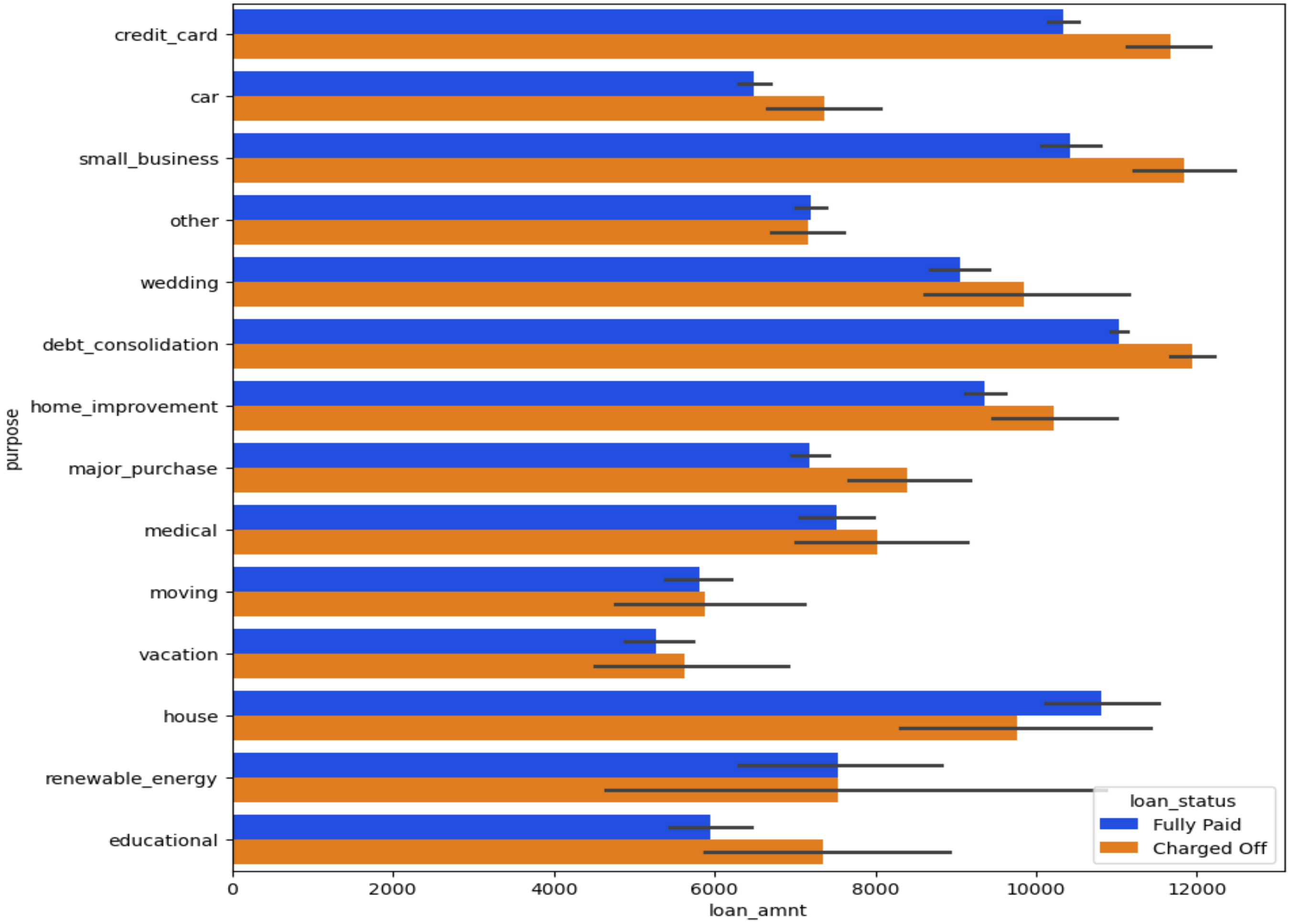
# Annual Income vs Loan Amount

- Across all the income groups, the loan_amount is higher for people who defaulted.



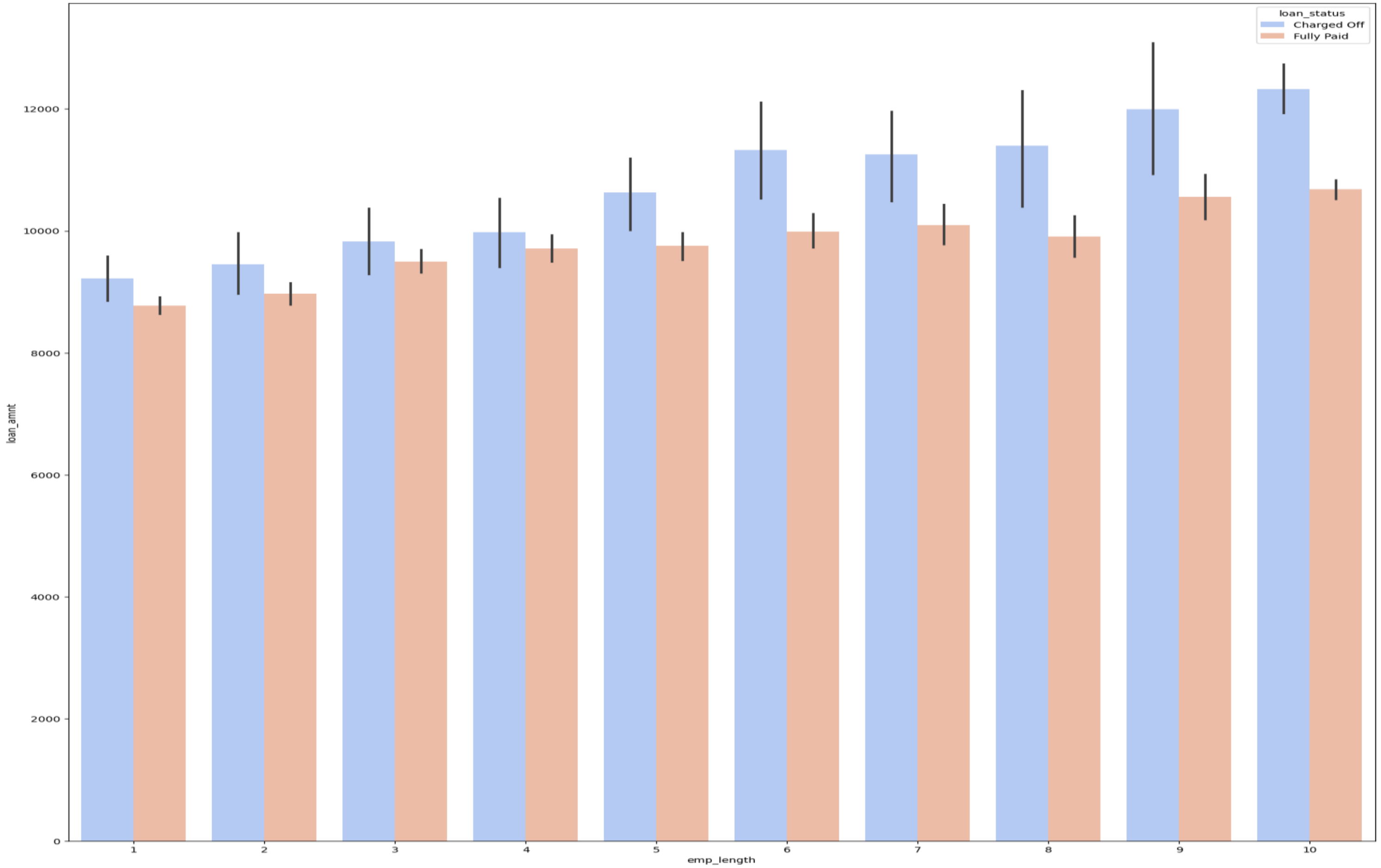
# Loan Amount vs Purpose

- Most of the purposes, the loan_amount is higher for people who defaulted.



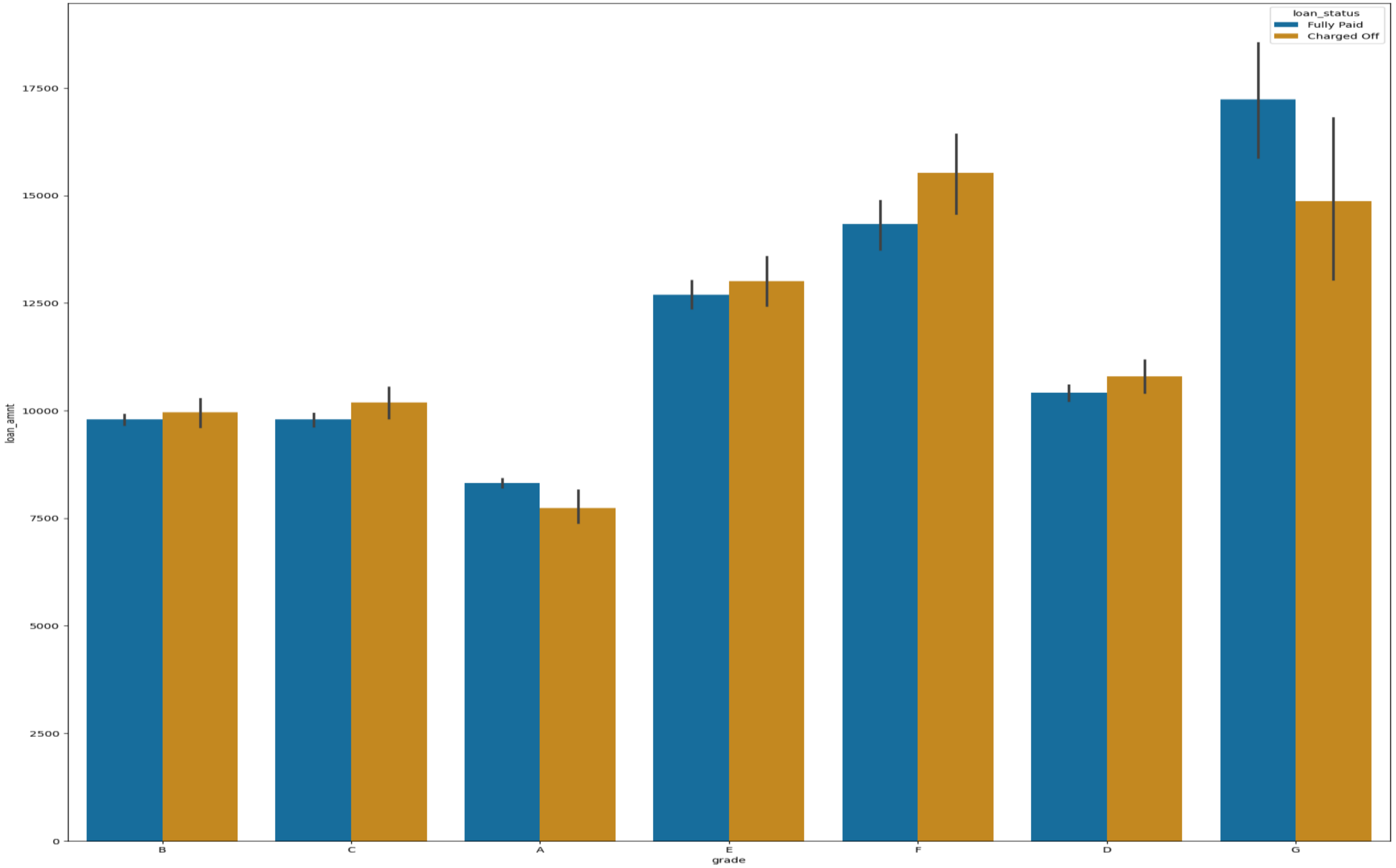
# Loan Amount vs Employment Length

- Across all the employment length groups,
  the loan_amount is higher for people who defaulted.



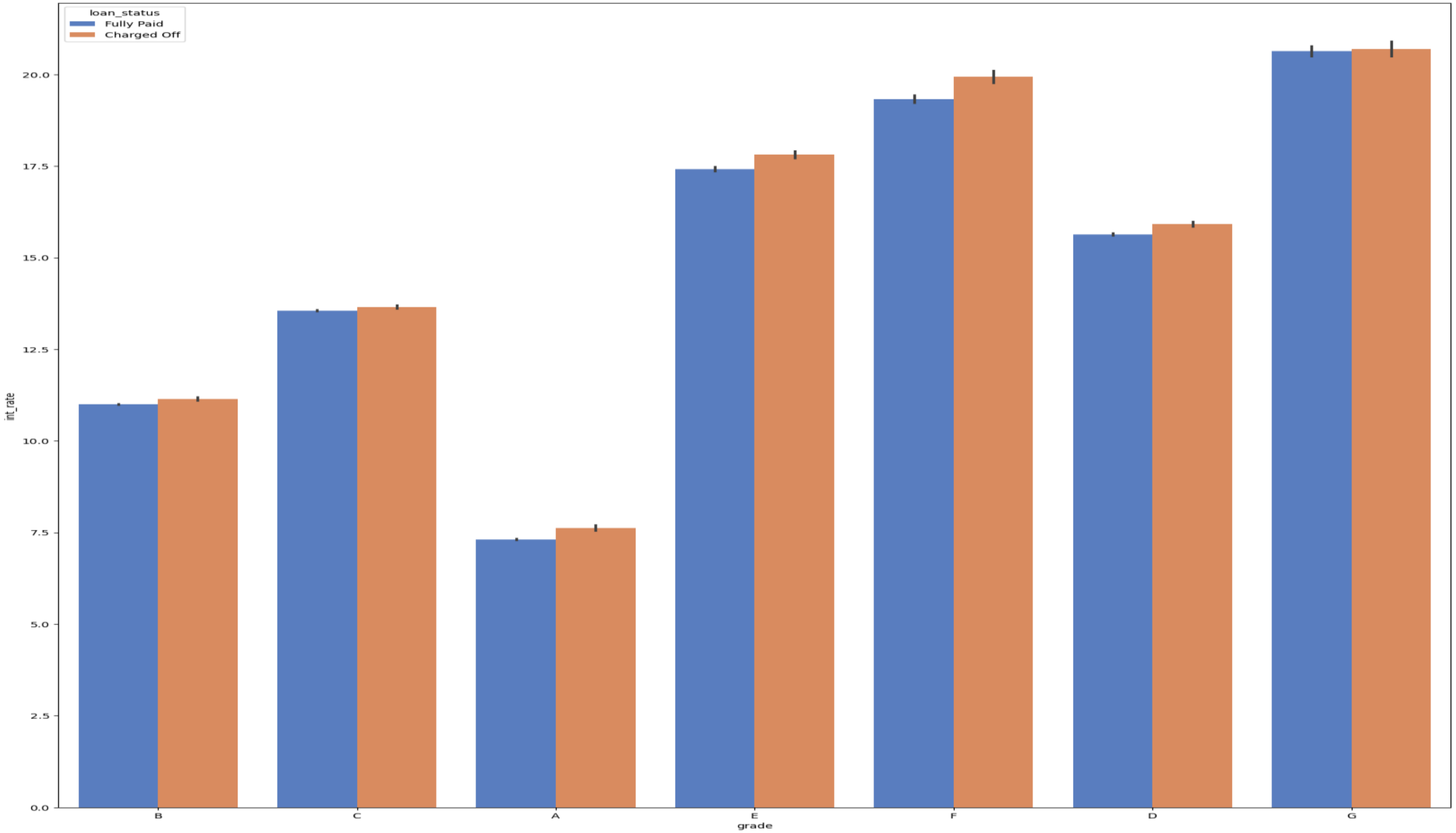
# Loan Amount vs Grade

- People who defaulted more when grade is F and loan amount is between 15k-20k



# Grade vs Interest rate

- People who defaulted more when grade G and interest rate above 20%



# Correlation

# Correlations

**Strong Correlation:**
- loan_amt has a strong correlation with funded_amt
- loan_amt has a strong correlation with funded_amt_inv
- funded_amt has a strong correlation with funded_amt_inv

**Negative Correlation:**
- loan_amnt, funded_amount, funded_amount_inv have negative correlation with pub_rec_bankrupticies
- annual income has a negative correlation with dti