LENDING CLUB CASE STUDY

GROUP FACILITATOR : SUMITKUMAR NAYAK
TEAM MEMBER : KARTHIK VARMA VEGESNA

Background

- ✓ You work for a **consumer finance company as an analyst** which specializes 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

Objectives

- ✓ Loan defaults are largest source of loss for consumer finance companies which makes risk analytics prime objective.
- ✓ 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 utilize this knowledge for its portfolio and risk assessment.

Expectations

- ✓ Using past data, derive driver variables of loan defaults by employing efficient EDA techniques
- ✓ Identify key drivers of loan default using various graphing and visualization techniques and efficient use of existing python libraries. Arrive at well substantiated conclusions and recommendations.

BRIEF ANALYSIS APPROACH

Data Cleaning

✓ Removed missing values, duplicate rows, mismatched columns, outliers and columns with significant portion of NA values.

Data Preparation

✓ Deleted columns which are not helpful for prediction, created new columns like investor profit percentage and converted data into analysable data types (Ex: dates etc.)

Data Analysis

✓ Performed univariate, segmented univariate and bivariate analysis with python libraries like seaborn, pandas etc

DATA CLEANING AND PREPARATION

- ✓ Finding prevalence of NA values in data and removing columns with proportion greater than 20% NA values
  Ex: tot_current_bal, tax_liens etc

- ✓ Replaced some NA values with 0 so that it is amenable to analysis removed duplicate rows.
  Ex: emp_length, int_rate and revol_util

- ✓ Columns which are interpreted as objects due to presence of certain characters and properties were cleaned using Regex and converted to required datatype columns
  Ex: loan_amnt,funded_amnt,int_rate,term etc.

- ✓ Converted columns which are related to dates to an python date object which is much useful for analysis
  Ex: last_pymnt_d, issue_d etc.

- ✓ Removed outliers from columns such as annual_inc which has some incomes which are too high.

- ✓ Removed unwanted columns which are not useful for analysis in business perspective like member_id, url, desc etc.

- ✓ Added new columns such as investor profit percentage and loan income percentage which can provide further business insights

DATA ANALYSIS APPROACH

- ✓ Identified Numerical and categorical columns as Analysis methodology varies
- ✓ Identified important metrics in consumer loan analytics by referring to online resources.

UNIVARIATE ANALYSIS

- ✓ Used countplots, distplots and boxplots of seaborn library to analyse individual columns with focus on count,density,median and quartiles

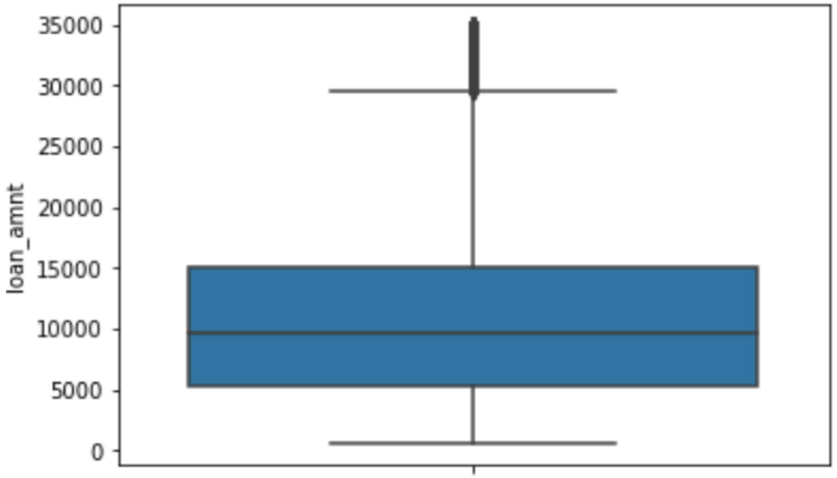
SEGMENTED UNIVARIATE

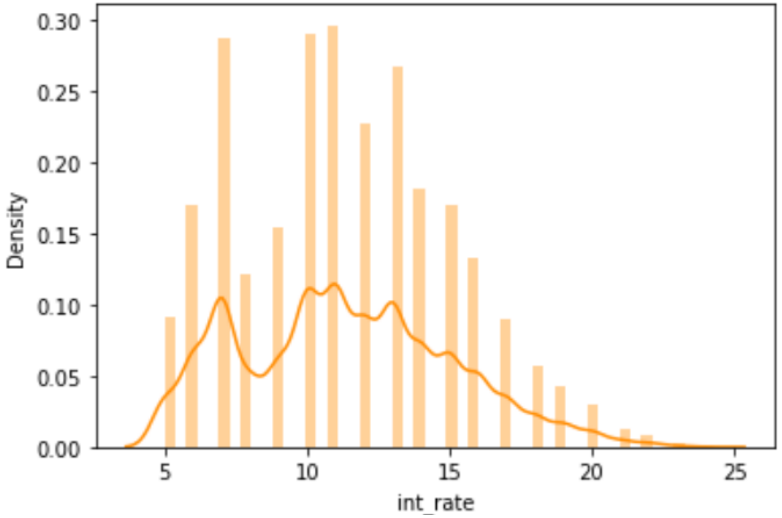
- ✓ Through binning continuous data and using categories as basis in categorical compared key metrics across categories and bins using seaborn plots

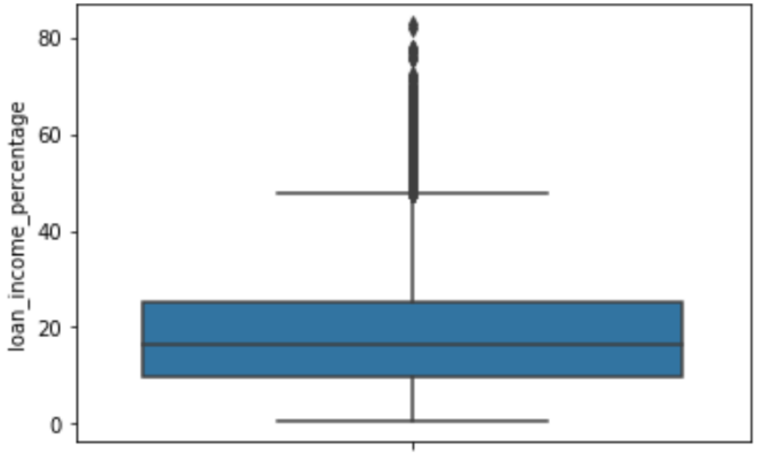
BIVARIATE ANALYSIS

- ✓ Used correlation matrices,histograms, boxplots etc to compare two columns and special focus on charged_off percentage column which is derived to gain insights into risk metrics

BASIC LOAN CHARACTERISTICS







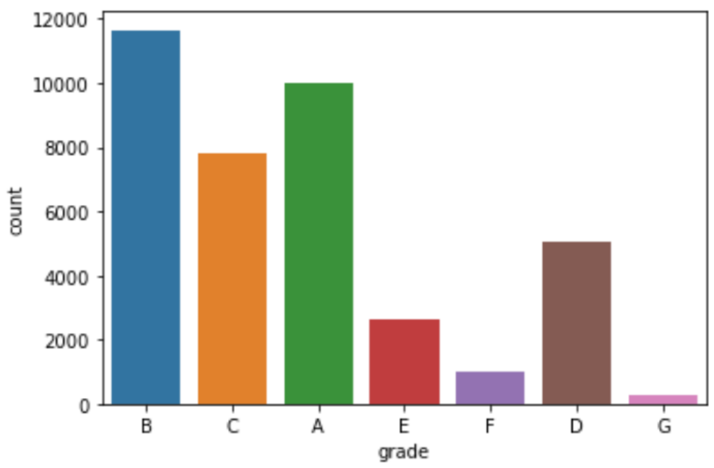
Inference

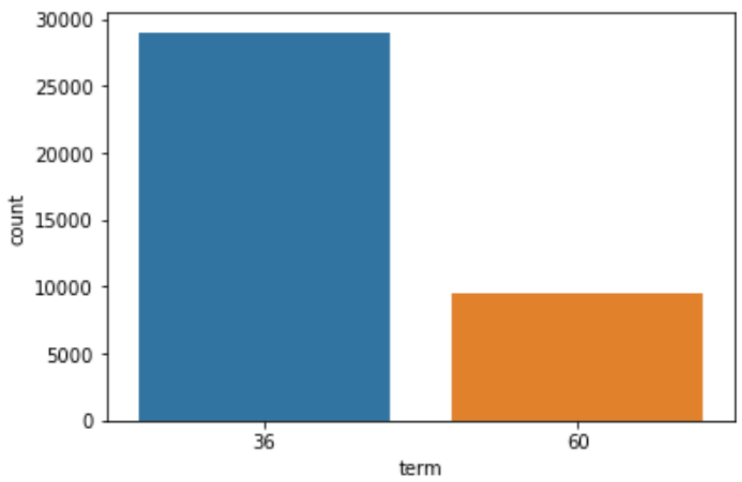
✓ Median Loan amount is 10000 and highest loan is just 3 times of median value(There are no significant outliers)

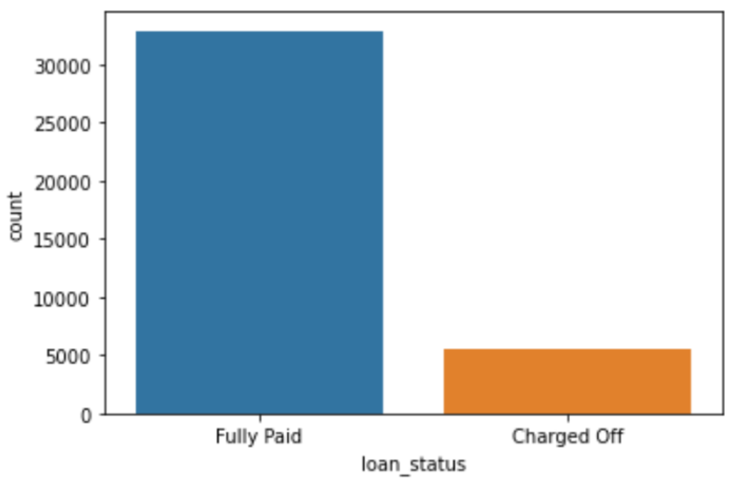
✓ Most of loans are around 20% of annual income, but there exists some outliers even to the range of 80% of annual income which needs to be monitored

✓ Interest rate distribution is not unimodal which may imply that there is underlying interest rate metric for different group of people depending on risk

BASIC LOAN CHARACTERISTICS







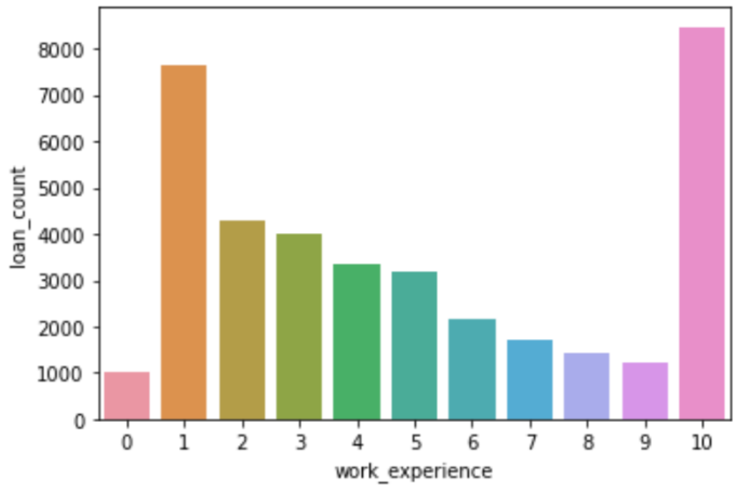
Inference

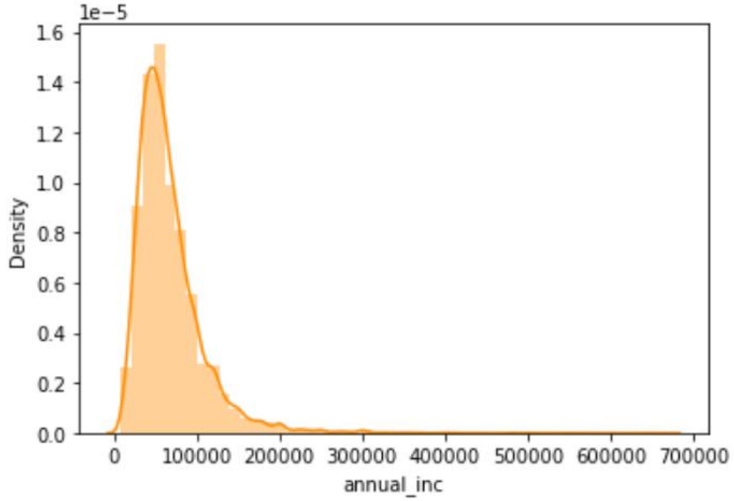
✓ Most of the loans are A and B grade loans which are assumed safe and very few are F and G which are risky.

✓ Most loans are of 36 months term.

✓ Majority of loans are fully paid and 1/7th of total are charged off

BASIC CLIENT CHARACTERISTICS

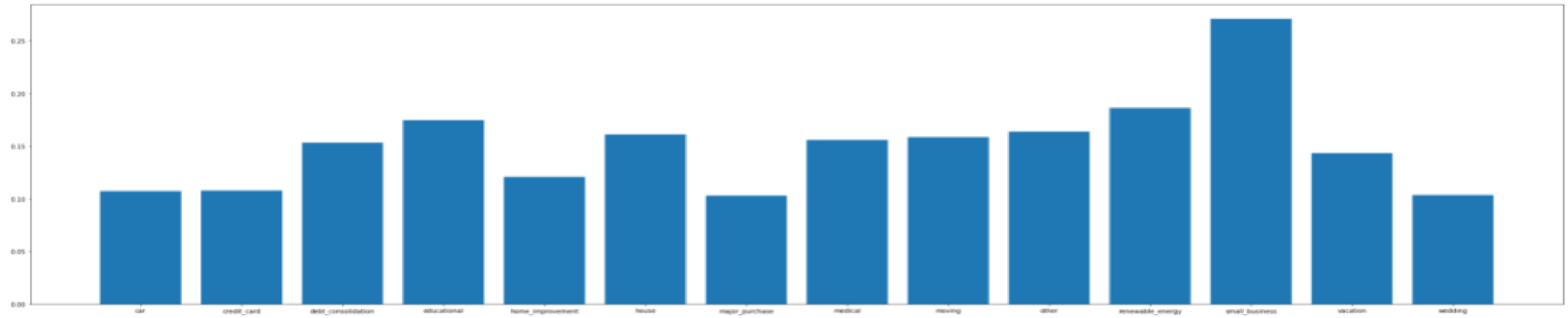


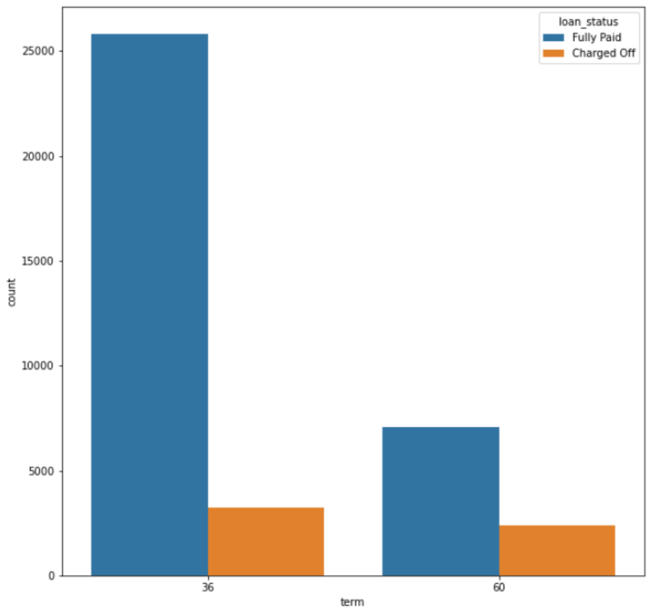


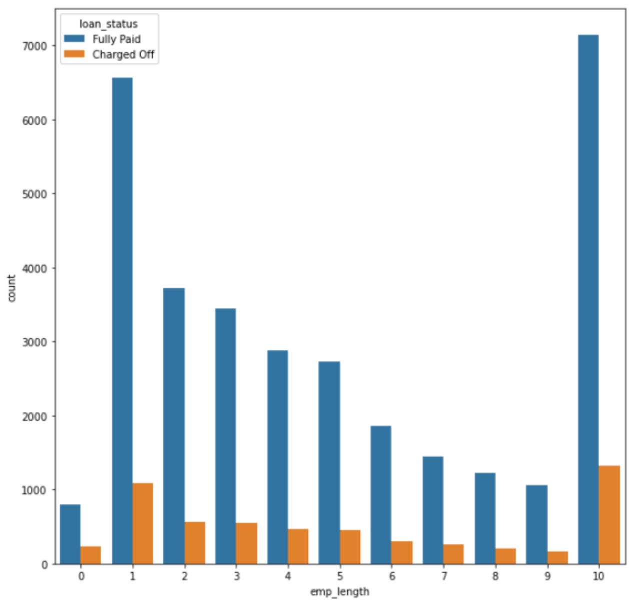
Inference

- ✓ Very few clients have own houses whereas share of rent and mortgage categories are significant and similar
- ✓ Most loans are granted for people with 10/10+ years of experience and 1/<1 year experience.(0 refers to unknown work_experience)
- ✓ Most clients have annual income around 1 lac and most below 2 lacs and it tapers of to the right with high incomes(Removed extreme income values in data cleaning step)

FACTORS INFLUENCING LOAN DEFAULT



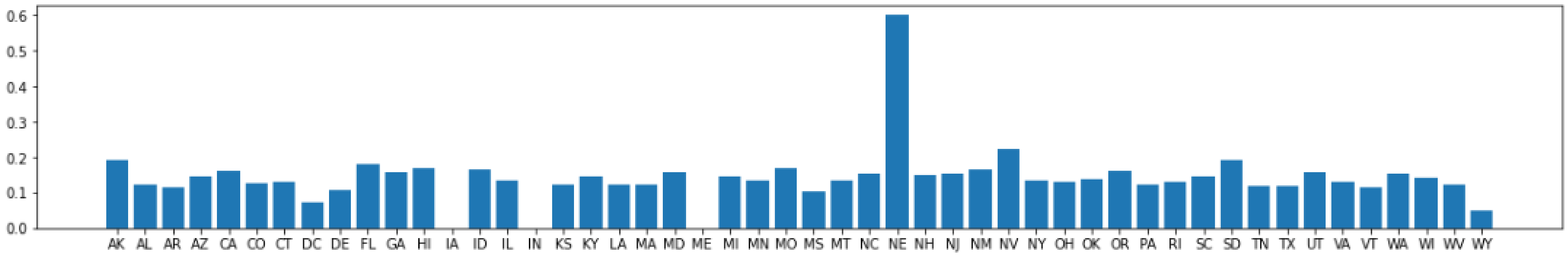


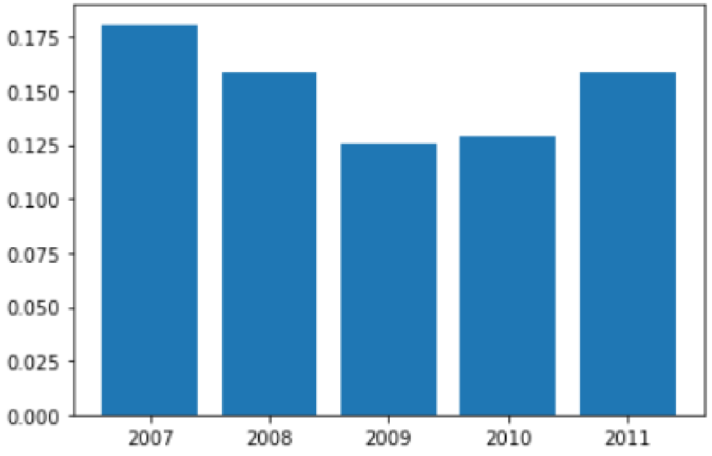


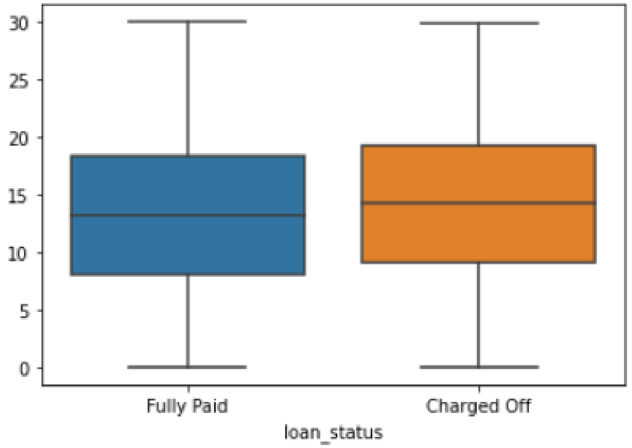
Inference

- ✓ Chances of loan defaulters are higher for Smaller Business loan type.
- ✓ People with 10+ years of experience were given more loans and Charged off proportion is also significant.
- ✓ More loans are issued for the 36 months tenure and proportion of Charged Off loans are more in 60 months tenure.

FACTORS INFLUENCING LOAN DEFAULT





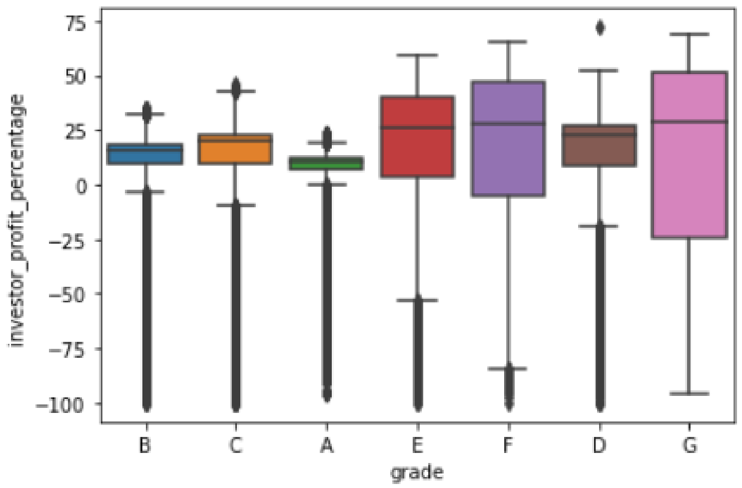


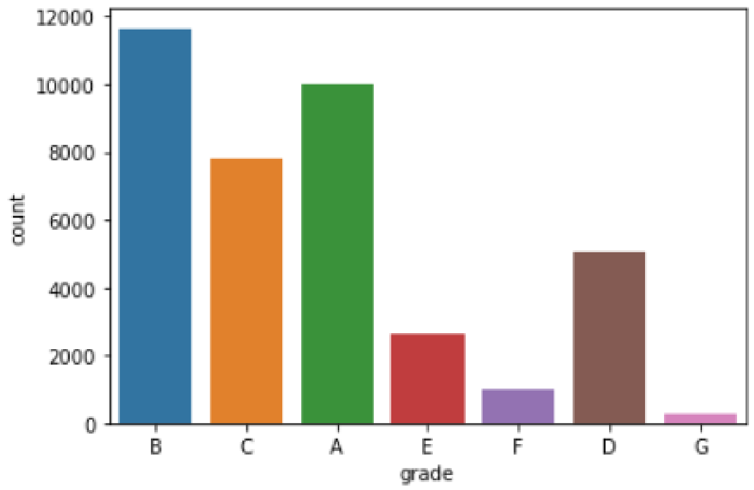
Inference

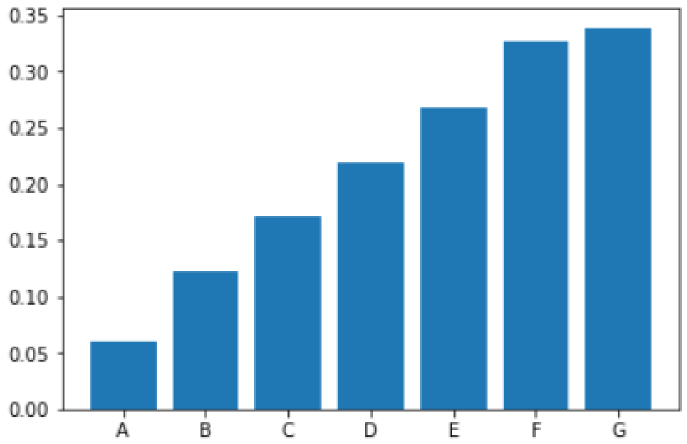
- ✓ 2007 is the highest Charged Off loan, and there is no noticable trend in the charged off loans percentage even though loan issue has increased.
- ✓ Although we expect debt to income ratio to be a good predictor of loan being charged off as high debt burden can be predecessor to default, its not a sensitive predictor as can be seen above(Although it can be seen that charged off loans have higher median dti than fully paid)
- ✓ Nebraska (NE) has the highest people of loan charged off percentage and it is important to compare only ratio's not absolute numbers.

*Charged of loans ratio is variable for Y-Axis if not mentioned

PERFOMANCE AND RELEVANCE OF GRADING







Inference

- ✓ It is interesting to note that there is no significance difference between median values of profit percentage of investors between loan grades. But Grade A,B,C,D loans have higher least values which implies they don't carry significant loss probability as compared to other grades. So grading system seems to work.

- ✓ Most of the loans are A and B grade loans which are assumed safe and very few are F and G which are risky.

- ✓ As the Grade increases, chances to Loan defaulter also increases implying good performance of grading system.

*Charged of loans ratio is variable for Y-Axis if not mentioned

KEY DRIVERS FOR LOAN DEFAULT AND RECOMMENDATIONS

- ✓ Performance of grading system is robust and it strict adherence to grading parameters is prescribed for minimising chances of loan default and it is one of the best tool to avoid losses.

- ✓ Although loan sanctions have increased dramatically over the years, charged off loan ratio is not trending upward which implies business expansion to more clients and guideline oriented aggressive loan sanction can help in increasing profits.

- ✓ Location factors do influence loan default and Nebraska has significant loan default risk which implies stringent loan approval process needs to be in place for this state.

- ✓ Customers with 10+ years experience, belonging to small businesses and loans of 60 months tenure are identified as risky categories which requires improved vigilance and approval measures in place

- ✓ Investors should not put too much emphasis on grading as median profits are same across all grades implying case to case need based loans are need of time and no consumer section should be deprived of loans without proper follow up.