LENDING CLUB CASE STUDY

Sanjay Biswas

Vivek Mathpal

CONTENTS

Introduction

Data Preparation

Univariate Analysis

Multivariate Analysis

Observations and Recommendations

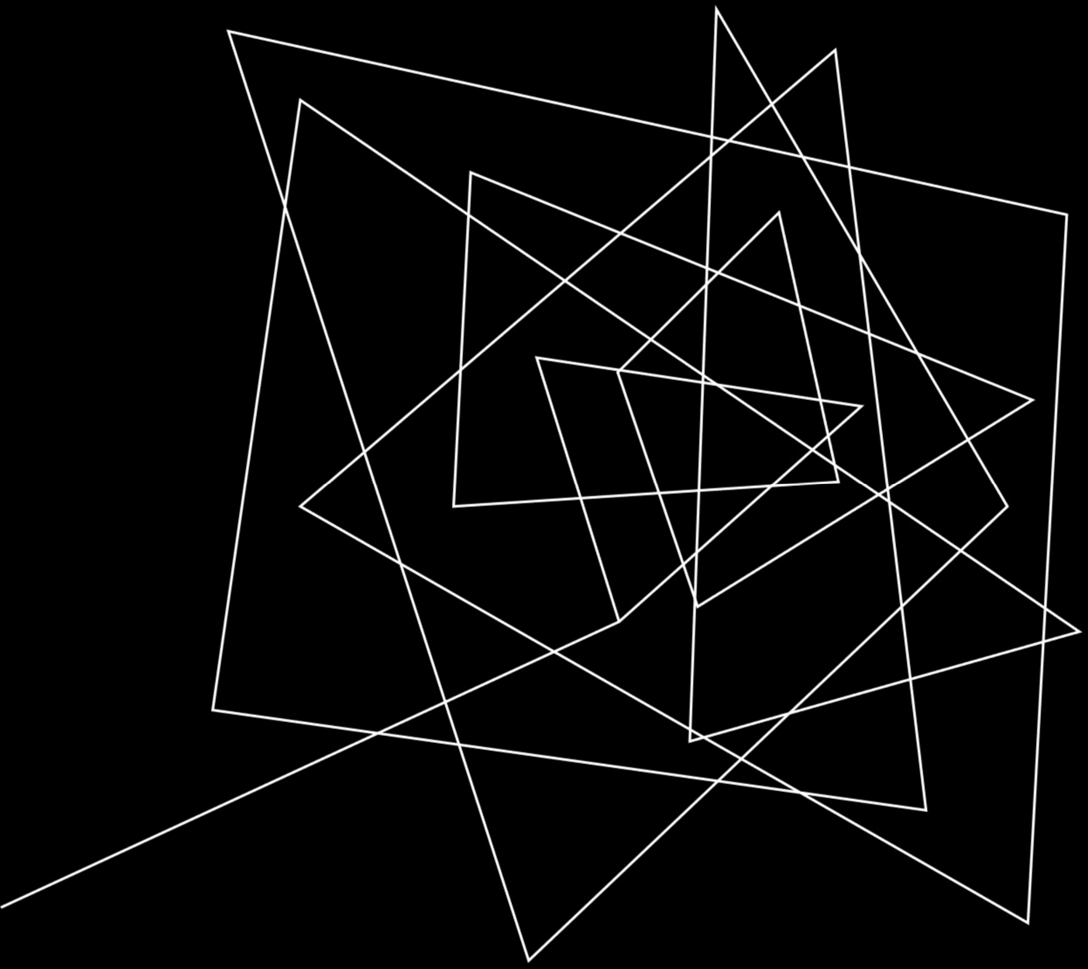
INTRODUCTION

Lending Club is one of the largest online marketplace for peer-to-peer lending. It provides an online platform to borrowers and lenders where they can transact smoothly and quickly.

Lenders Club wants to conduct a study identifying the factors which contribute most to the defaults in loan payments.

As a data analyst, we need to analyze the historical data and look for factors which leads to borrowers defaulting the loan.

For analysis, we will be using python and packages used are numpy , pandas , seaborn & matplotlib.

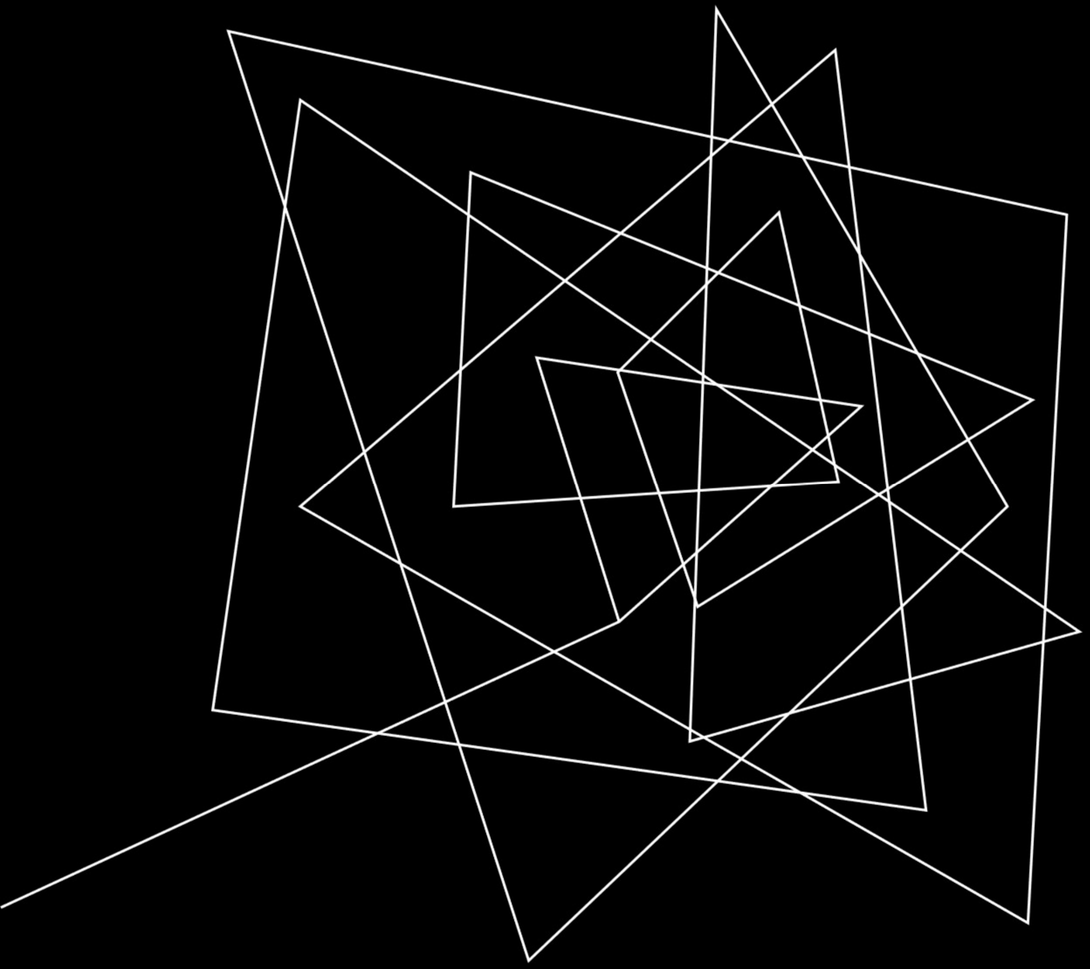


DATA PREPARATION

Understand and Cleaning of the data

Data Preparation & Cleaning

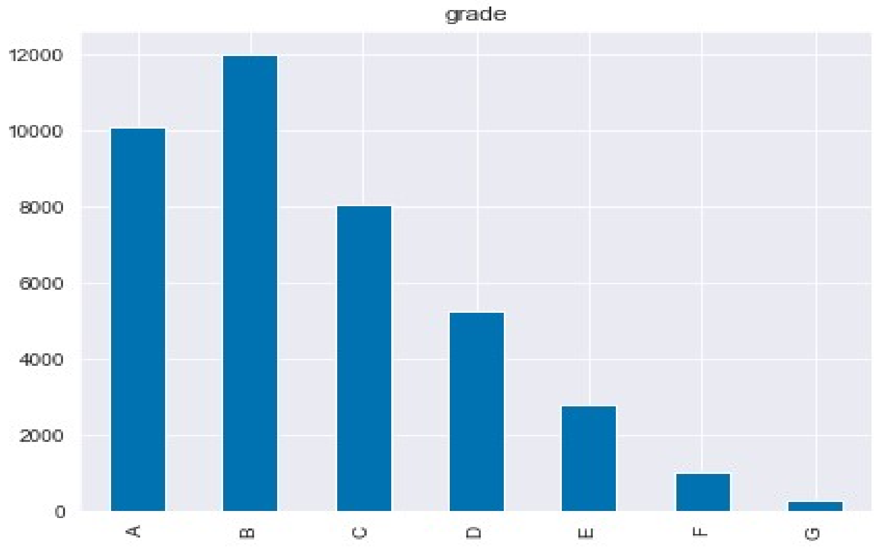
- The data provided has a total of 111 rows and 39,717 rows in the data sets
- Majority of the data types in the data set are float (74) and others being integer (13) and object (24)
- While inspecting the data set we found, there are lot of columns with missing values. Of the total fields in the data set, almost 51% of the data was missing.
- There were columns, which had all the rows as "NA", we can completely get rid of these by dropping them from data set. The threshold for dropping column was any column with more than 40% NA values.
- On further analysis, we observed there are columns with data on customer behaviour after loan has been granted, these will be redundant for the analysis needed. (funded_amnt','funded_amnt_inv','emp_title','url','desc','title','pymnt_plan','zip_code','addr_state','delinq_2yrs','earliest_cr_line','open_acc','pub_rec','revol_bal','revol_util','total_acc','out_prncp','out_prncp_inv','total_pymnt','total_pymnt_inv','total_rec_prncp','total_rec_int','recoveries','collection_recovery_fee','last_pymnt_d','last_pymnt_amnt','last_credit_pull_d','application_type','tax_liens').
- In addition, upon removing the "NA" values from the columns, we still had some "NA" values, which we dropped by removing some of the rows.

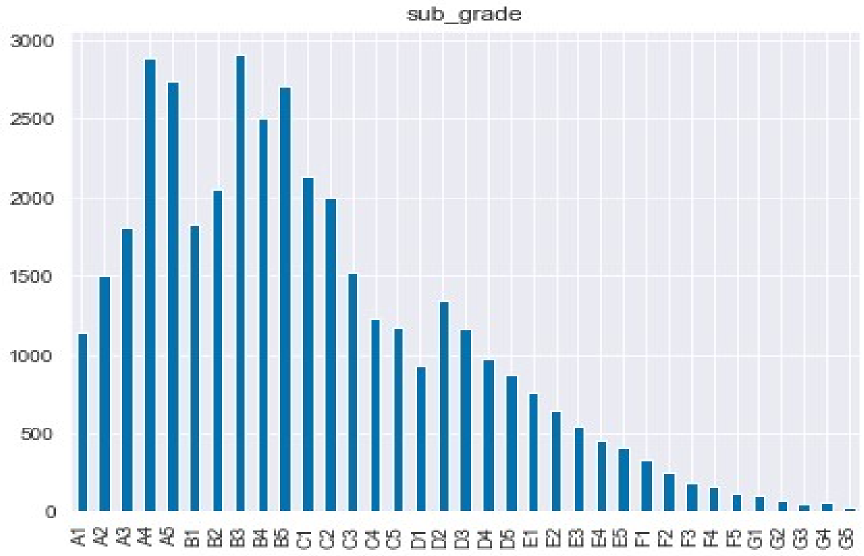


UNIVARIATE ANALYSIS

Observations

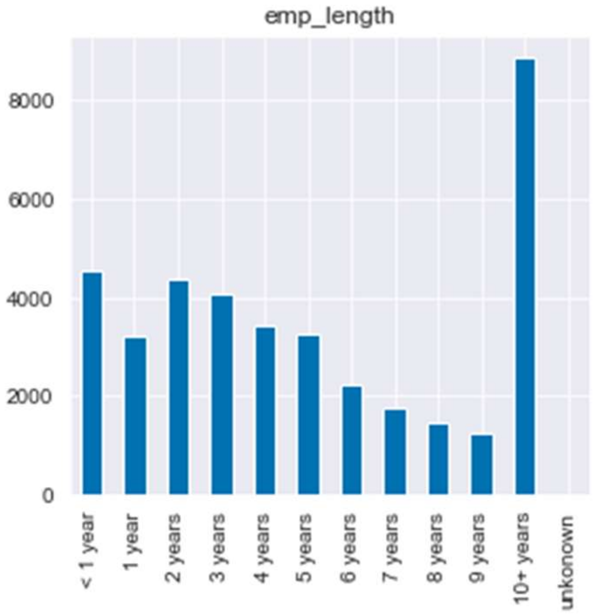
- Even though loan application acceptance decreases with decrease in grade, however grade B has a higher frequency than A. This could be because grade A are more affluent and fewer people need a loan.
- Withing grades A & B, number of loan application acceptance increases with decrease in subgrades, while for C and lower, the acceptance decrease with decrease in sub-grades. A reason could be that for grades A & B, loans are accepted irrespective of subgrades and more applications are received with decreasing sub grades, whereas for C & below, sub grades have an impact on application acceptance.

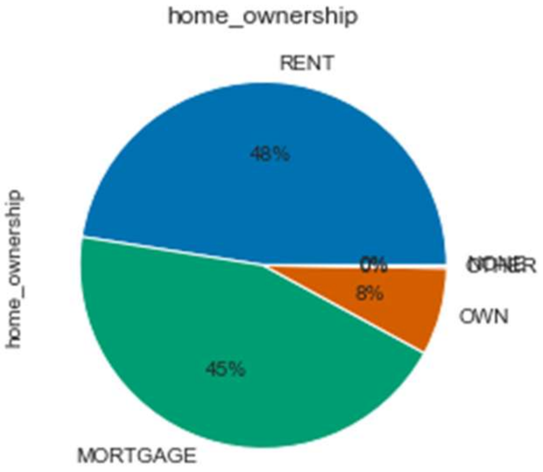


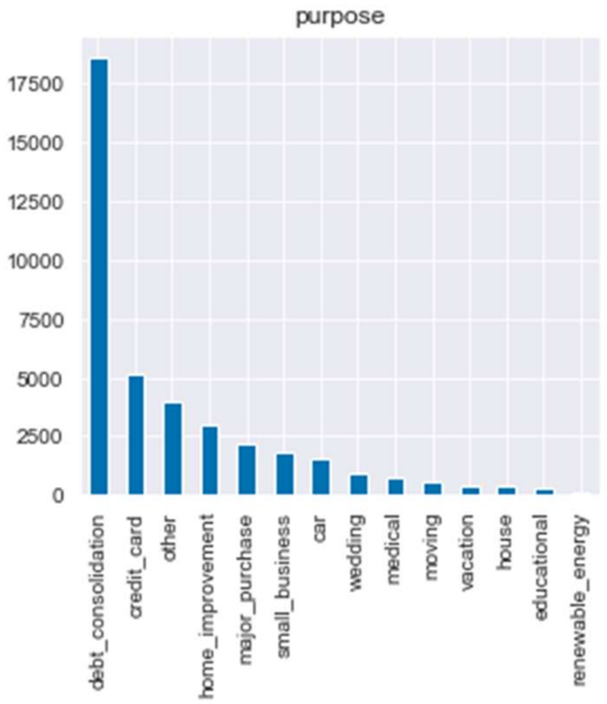


Observations

- Loan applications accepted generally decreases with increase in years if employment, indicating more need of loan in early years of employment. 10+ years has a peak possibly because all 10+ years are bucketed into this category
- Customers with rented and mortgaged house apply and get accepted for loan the most
- Most loans are for debt consolidation

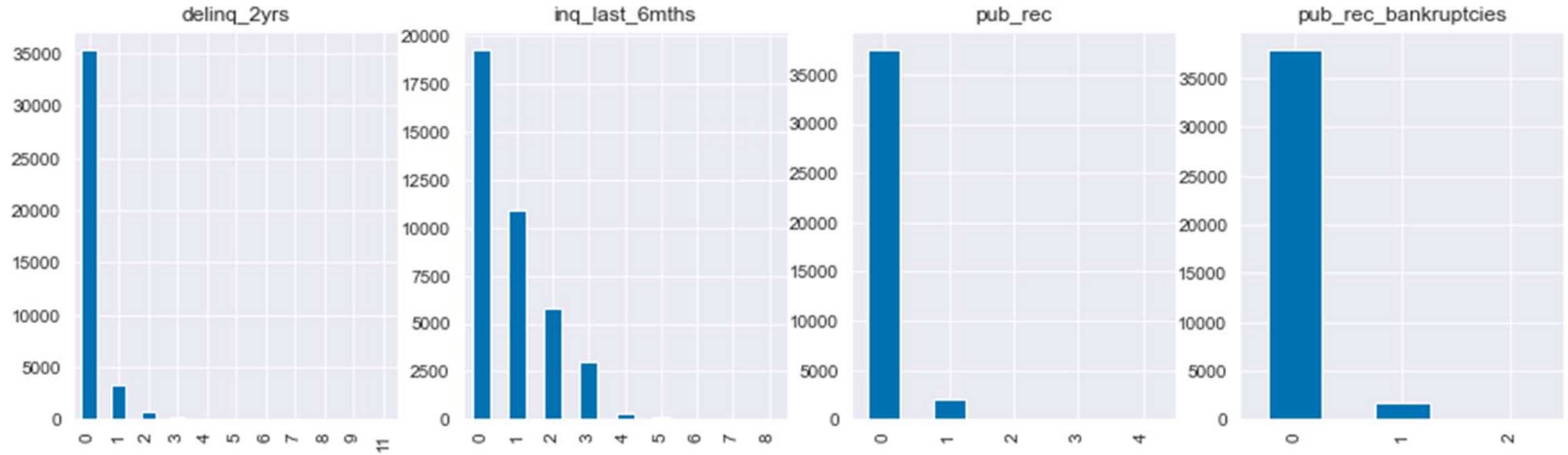




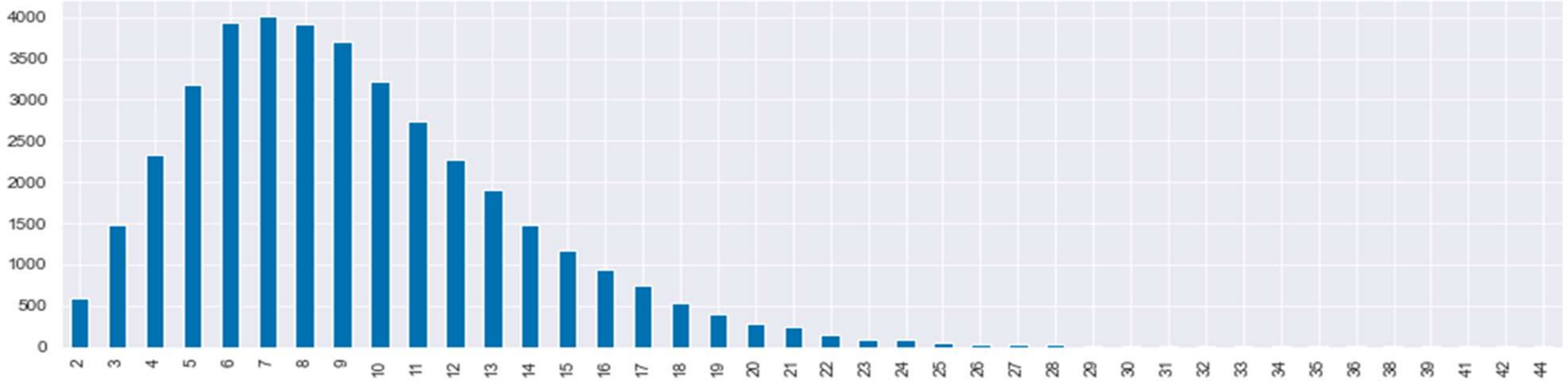


Observations

- Loan applications are almost always being accepted for customers with no derog records

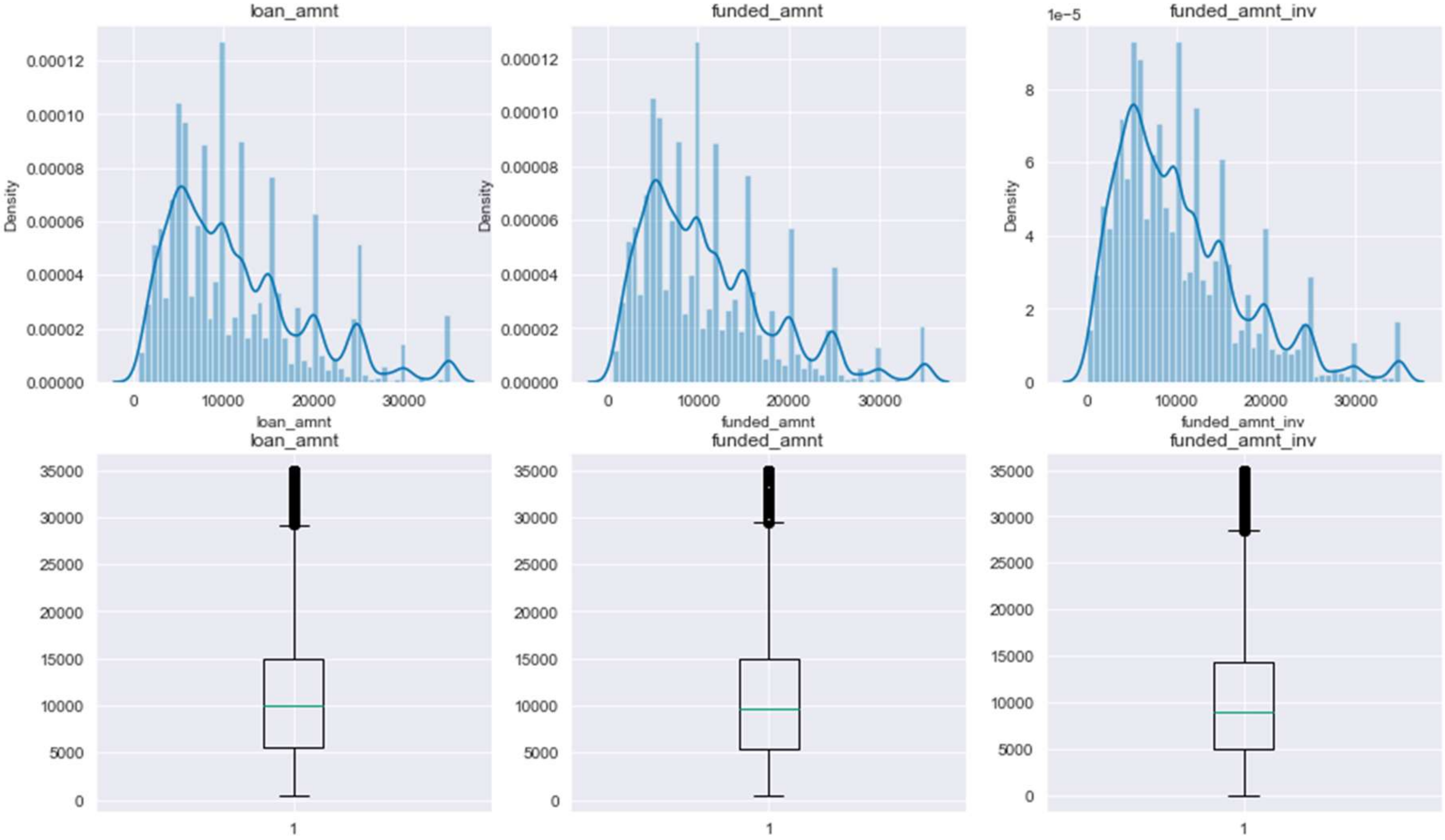


- Loan acceptance have a tendency towards 7 open accounts on credit file



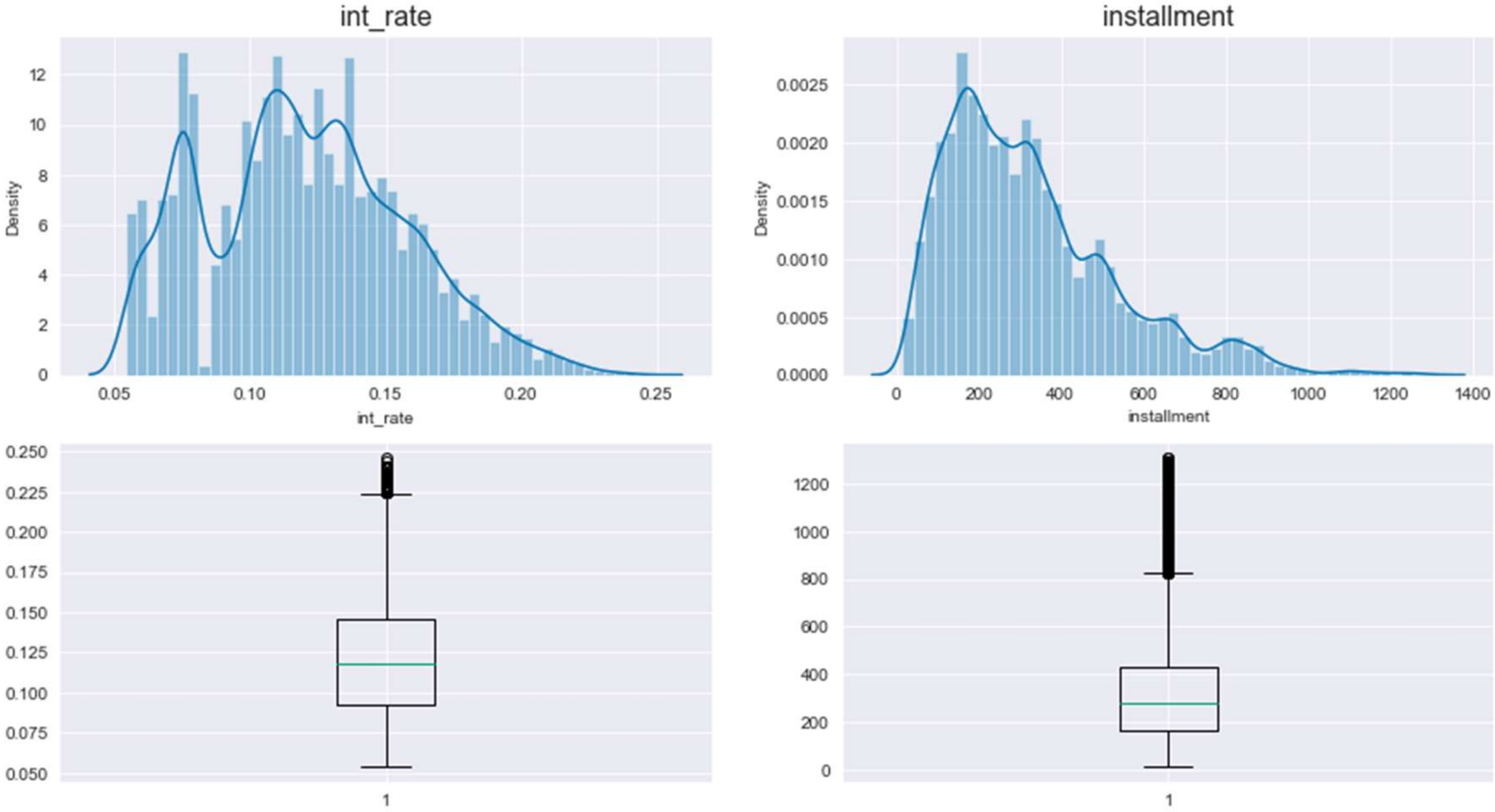
Observations

- There are peaks at intervals possibly due to rounding off
- Most loan amounts are in the range of IQR(5000-15000), however funded_amnt_inv is skewed towards



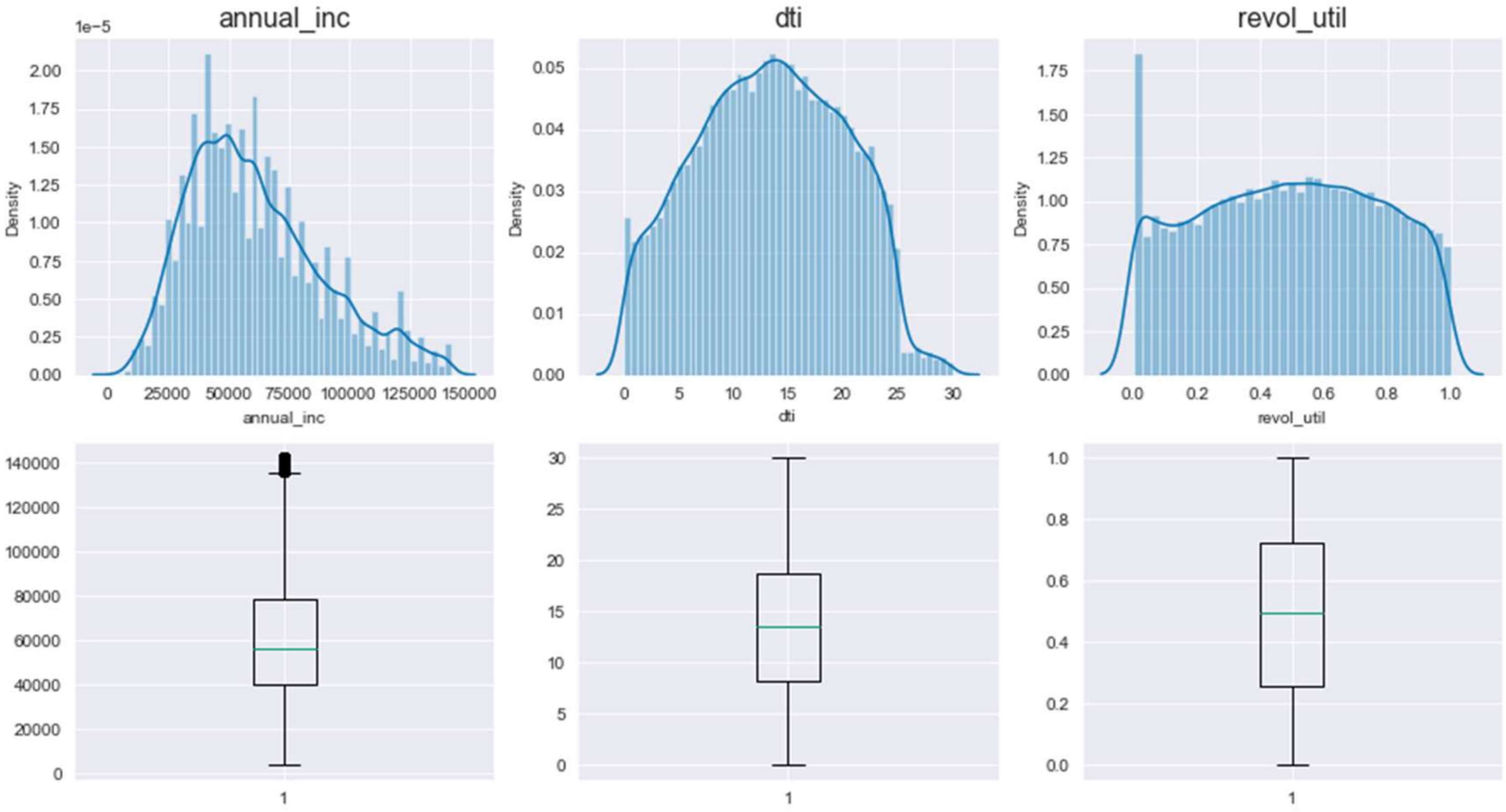
Observations

- Very few loans given at 8% int rate. What could be the reason? Further study and drill down needed to make any conclusion



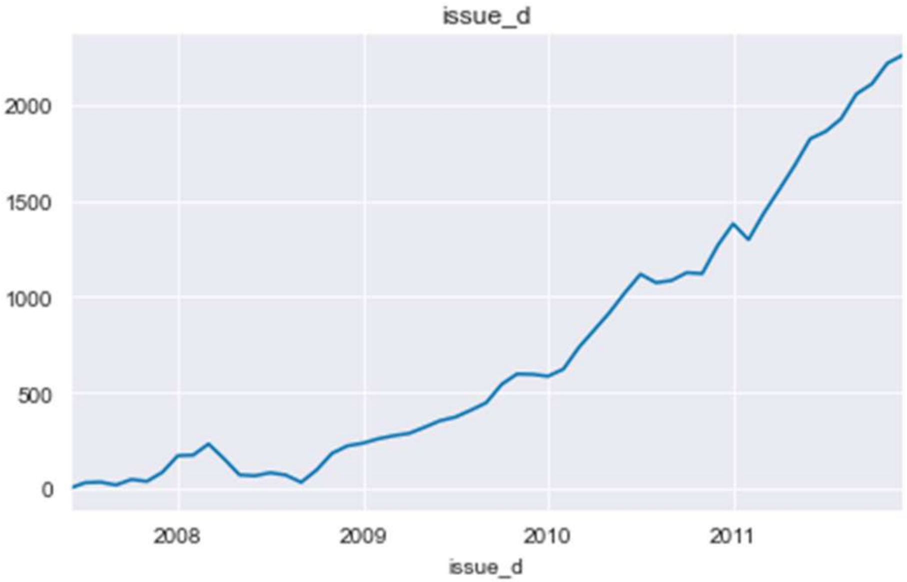
Observations

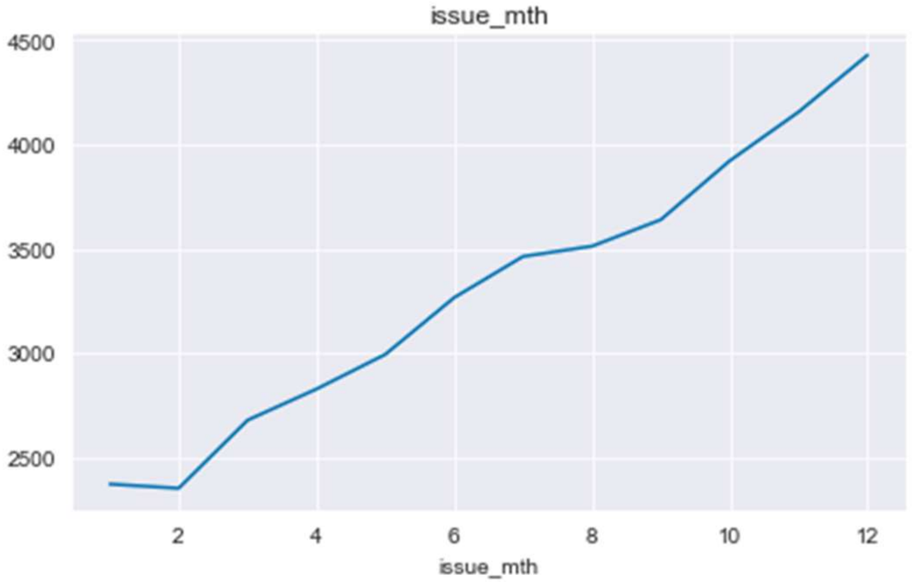
- Spikes in annual income because of round offs
- Maximum loans are given for dti and revol_util towards zero
- revol_util seems to have a uniform distribution, so no relationship with loan data other than the spike at zero

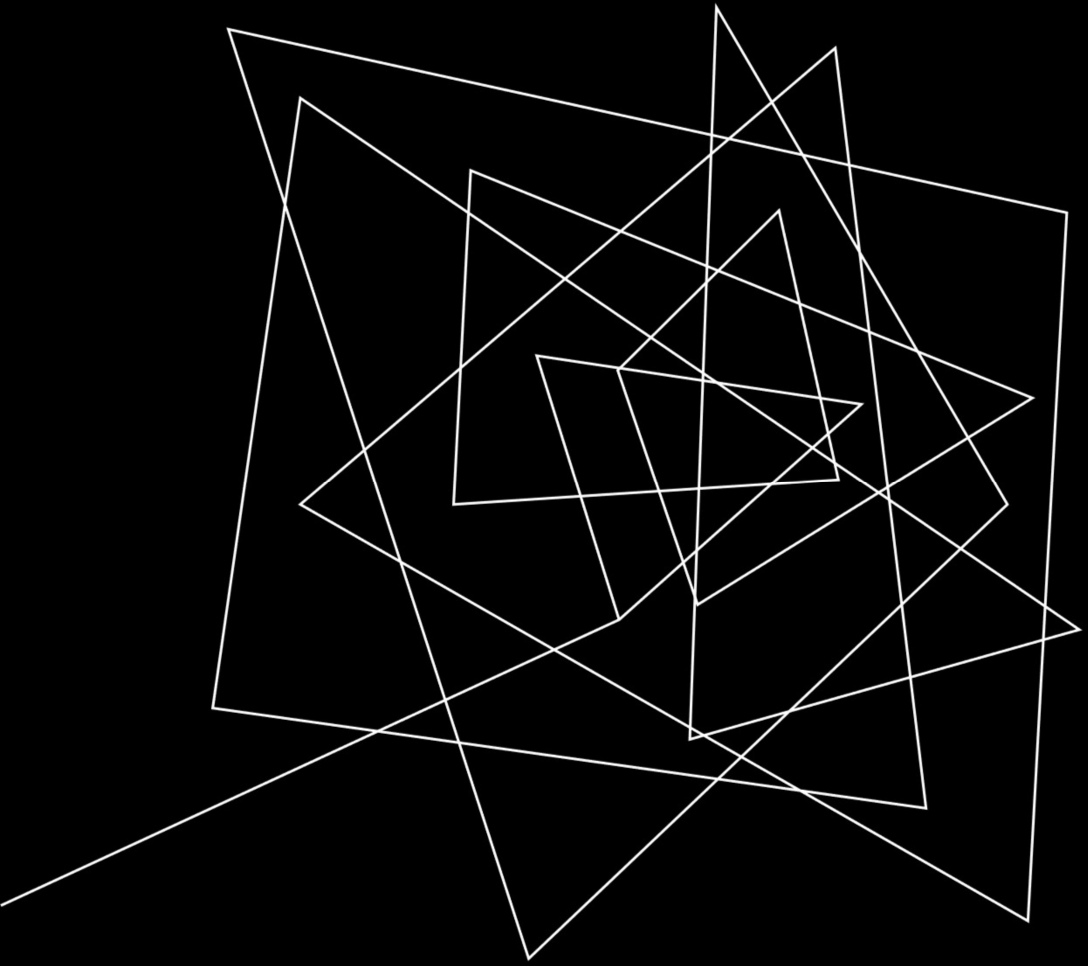


Observations

- Almost linear growth in loans since 2009
- Almost linear growth in loans starting March throughout Dec
- Slope: Jan-Feb least number of loans issued| March-May, Jul-Sep growth rates of loan issuance are slower than May-Jul, Sep-Dec



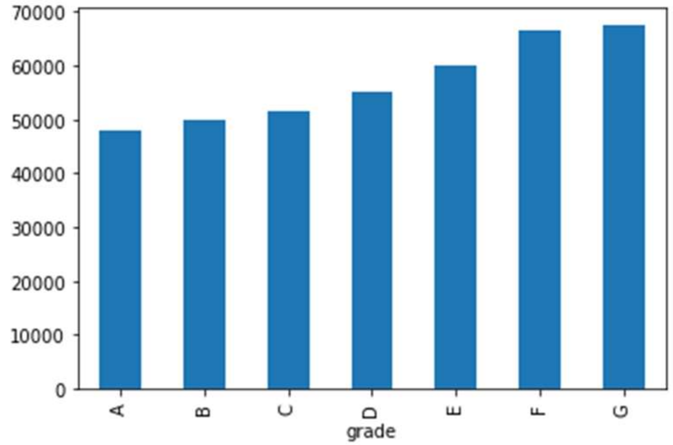


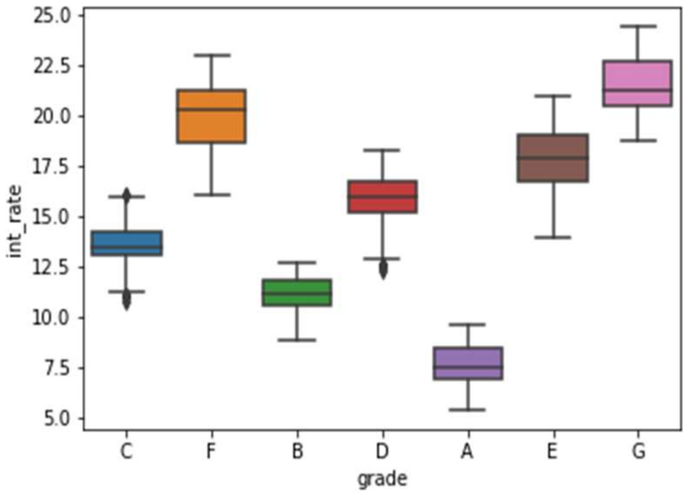


BIVARIATE ANALYSIS

Observations

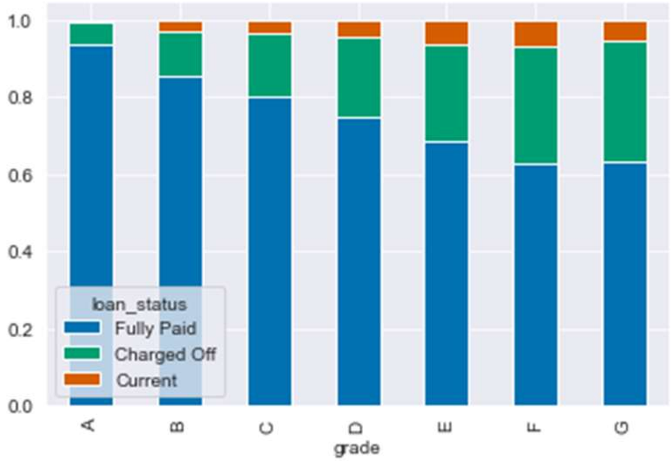
- The graph between the grades and the median annual income shows it is gradually increasing from A to G. It shows that borrowers with high income group tend to default on loan more.
- There is a large overlap in grade B&C, in which some of the applicants can be offered lower interest rates or vice versa. Similar is the case with E, F& G grades

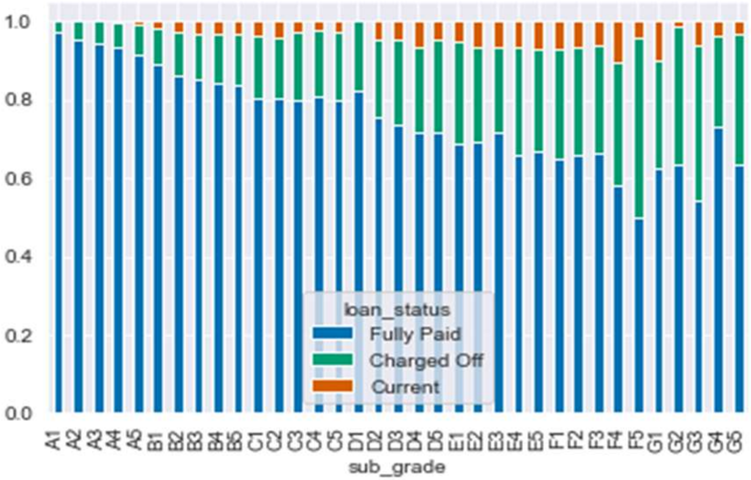


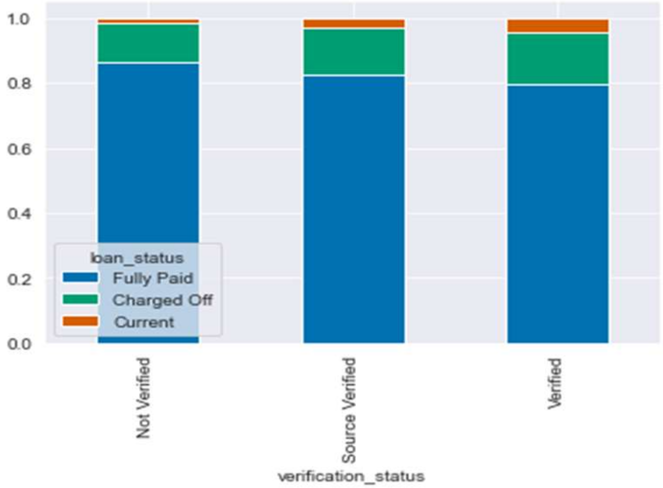


Observations

- Charge off frequency increases with decline in grades



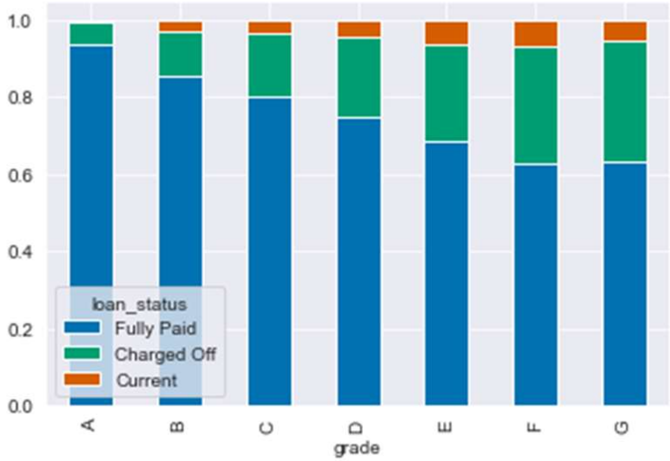


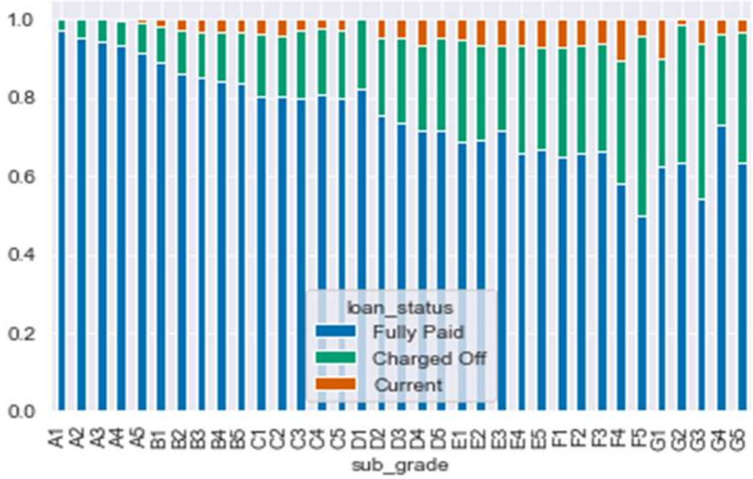


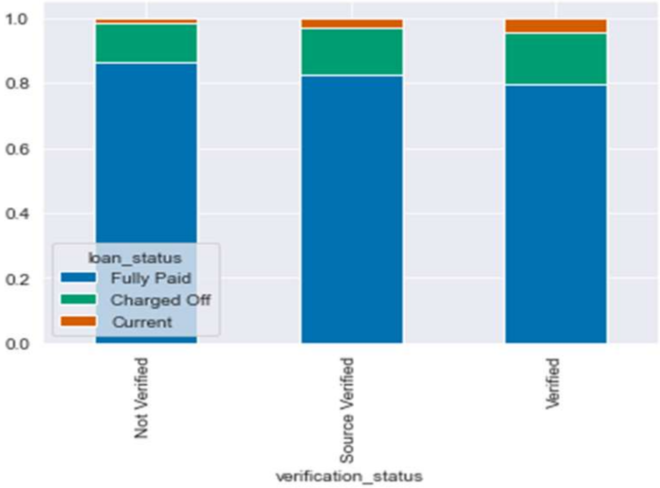
- Surprisingly non verified accounts have lower default rates

Observations

- Charge off frequency increases with decline in grades



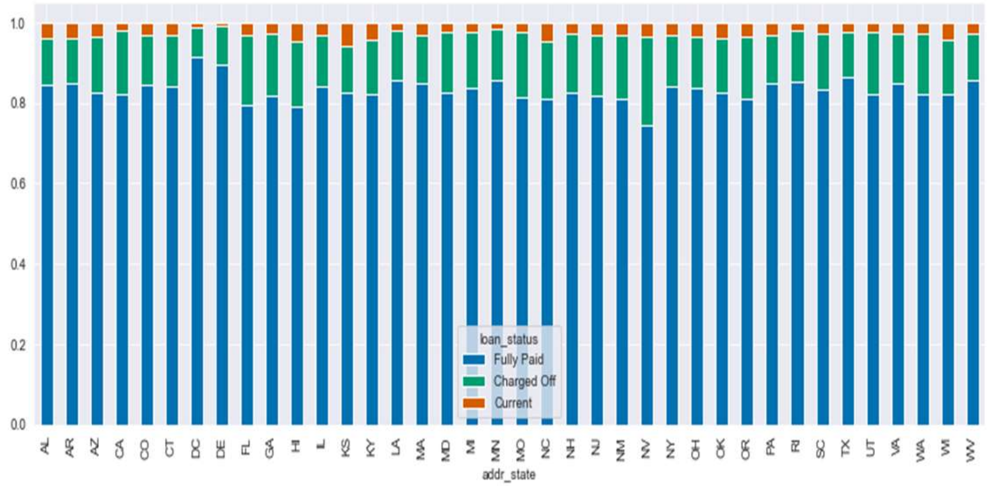




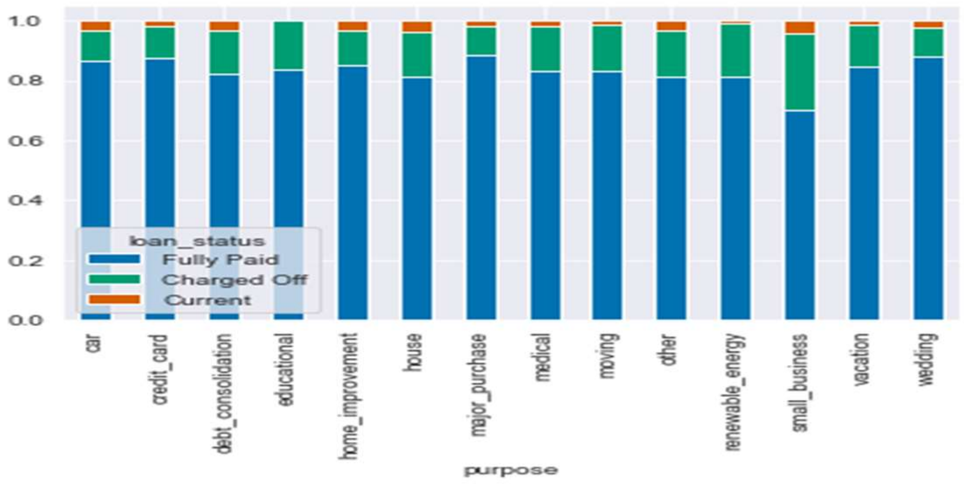
- Surprisingly non verified accounts have lower default rates

Observations

- NV has the highest default rate followed by FL

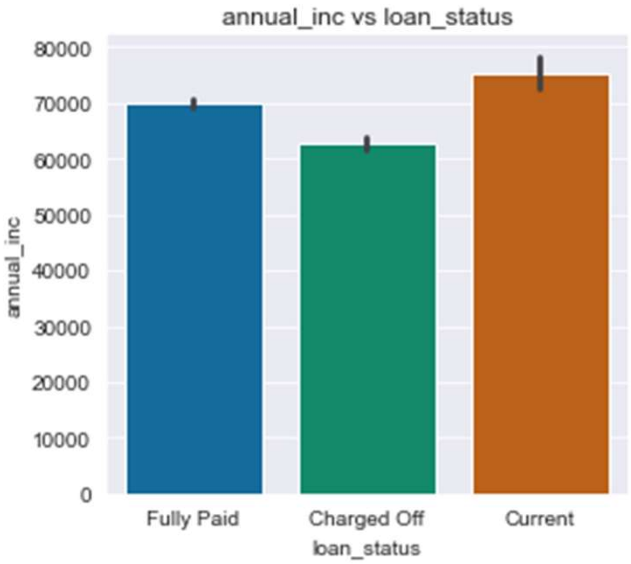


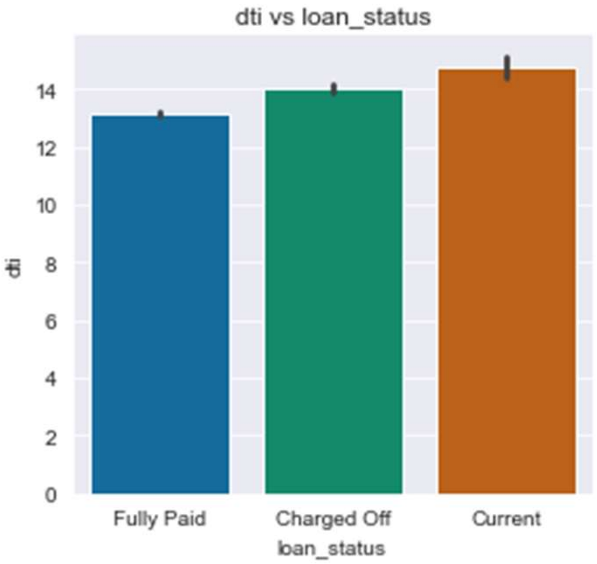
- Purpose with small_business have a very high rate of delinquency

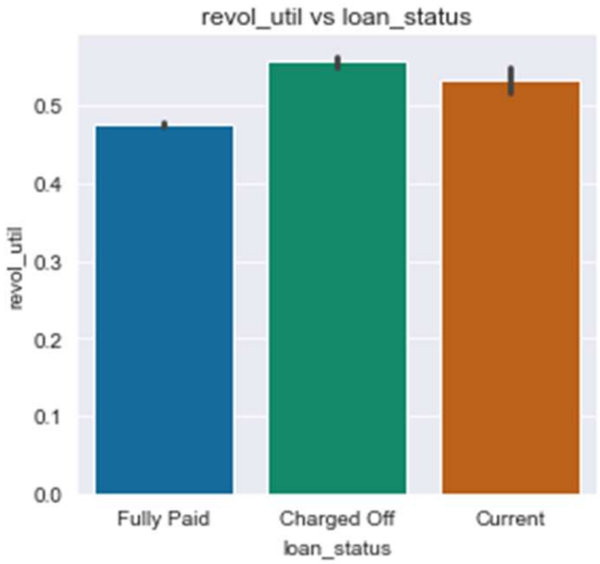


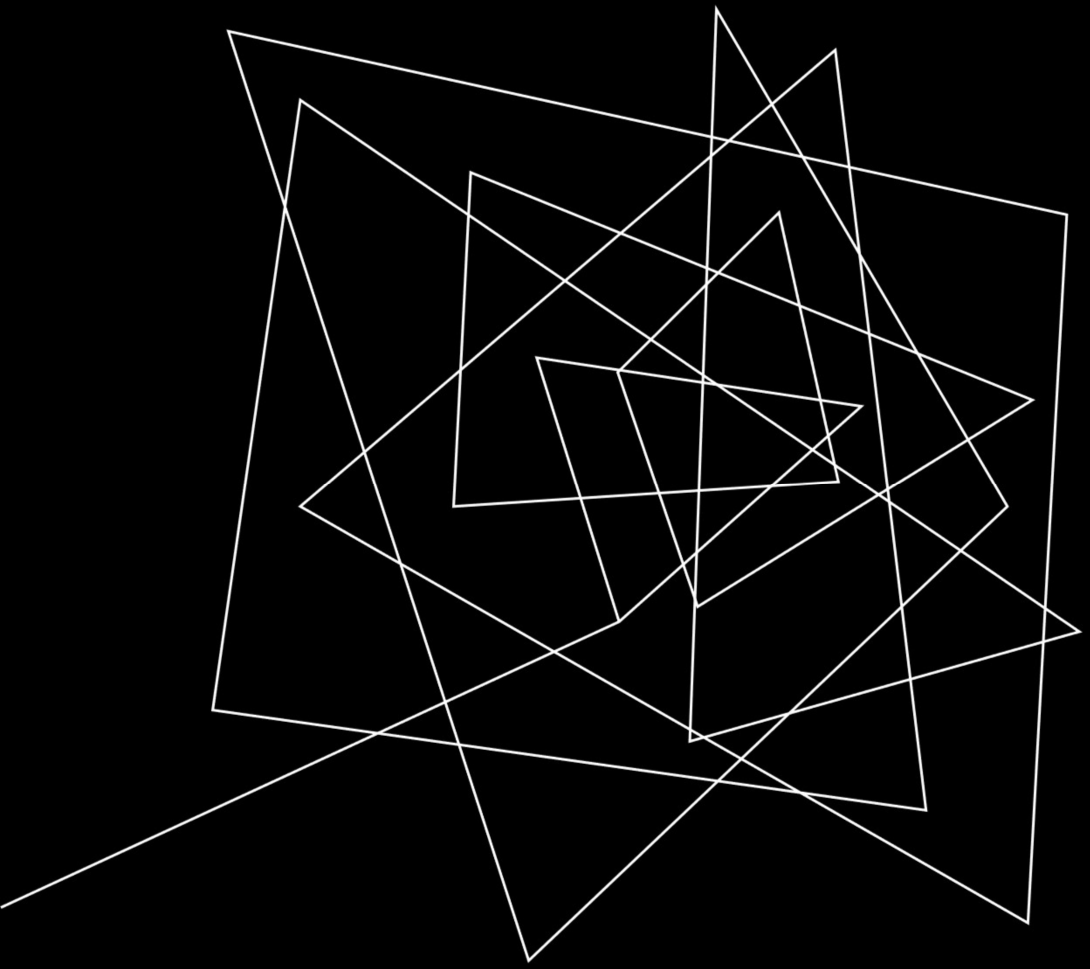
Observations

- Delinquency is high for lower annual income
- Higher DTI for Charged off customers opposed to Fully paid
- Charge off frequencues higher for high revol_util


annual_inc vs loan_status


dti vs loan_status


revol_util vs loan_status



OBSERVATIONS & RECOMMENDATIONS

Conclusions

- More marketing & sales efforts on employees with <5 years of experience and customers with rented and mortgaged houses
- Further investigation needed on the income verification, because the non-verified accounts are performing better than those verified.
- Higher interest rates should be charges to borrowers with debt-to-income ratio.
- More sample is needed for more robust state comparison.

THANK YOU