LENDING CLUB CASE STUDY

SUBMISSION

By:

1. Nishant Chaturvedi

Business Understanding

When lending 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:

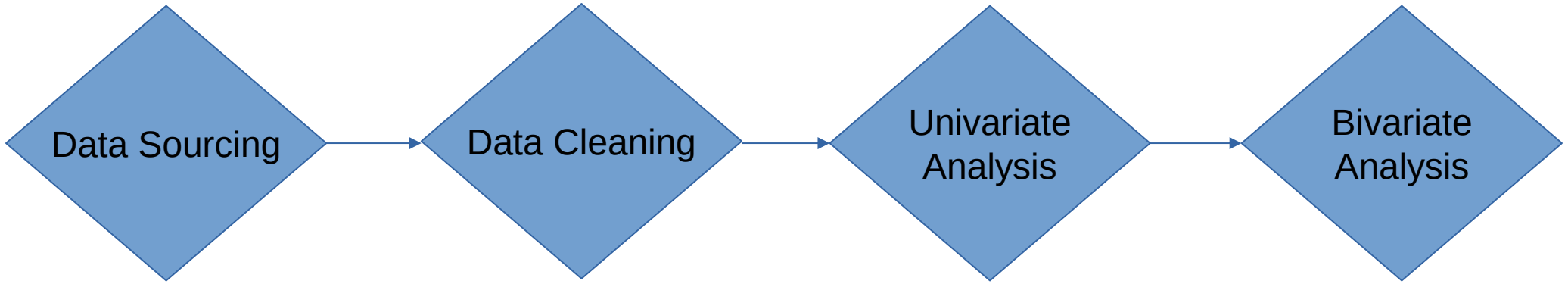
- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

Problem statement

Lending 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.

In this case study, we use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default

Workflow Diagram



Approach for solving the case study

1. Identify the Loan primary, consumer and loan secondary attributes.
2. Bin the continuous variables into categorical variables for each of the 3 segments.
3. Establish a Pearson correlation matrix and heat map for all the numeric variables within a given segment
4. To establish the level of dependency with respect the loan status (Fully paid or Charged off),
5. For the chosen variables, perform the graph visualization to understand the various aspects of the same.

Data Sourcing

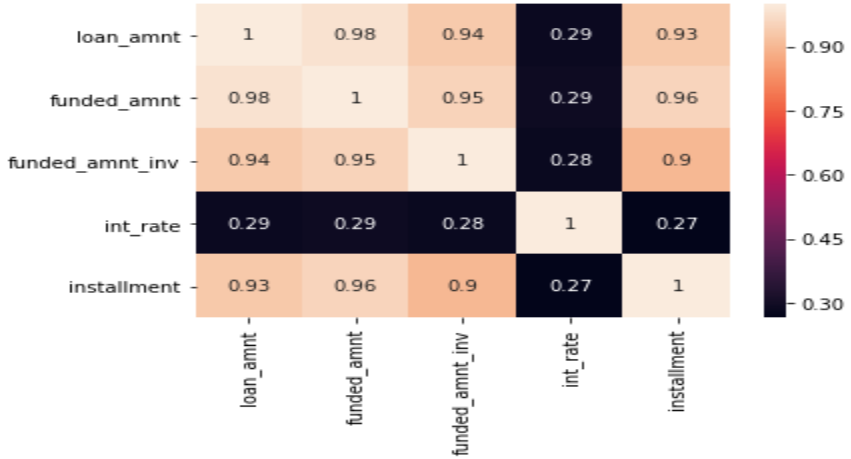
- The given data file is Private data type.
- Banking, telecom, retail, and media are some of the key private sectors that rely heavily on this type of data to make decisions.

Data Cleaning

- Columns Considered after data cleaning:
  1. **Reference Attribute**: loan_status.
  2. **Loan Primary Attributes**: loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade,
  3. **Consumer Attributes**: emp_length, home_ownership, annual_inc, verification_status, purpose, addr_state, dti, issue_d
  4. **Loan secondary attributes**: earliest_cr_line, inq_last_6mths, open_acc, revol_bal, revol_util, total_acc, total_pymnt_inv, total_rec_prncp, total_rec_int, last_pymnt_d, last_pymnt_amnt.
- Steps followed:
  1. Dropping columns which have more than 60% data missing.
  2. Dropping columns which have all zero values as they will be useless.
  3. Applying the IQR technique to remove the outliers.
  4. Dropping irrelevant columns.

Analysis of Loan Primary Attributes

Heat Map



Correlation matrix

	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment
loan_amnt	1.00	0.98	0.94	0.29	0.93
funded_amnt	0.98	1.00	0.95	0.29	0.96
funded_amnt_inv	0.94	0.95	1.00	0.28	0.90
int_rate	0.29	0.29	0.28	1.00	0.27
installment	0.93	0.96	0.90	0.27	1.00

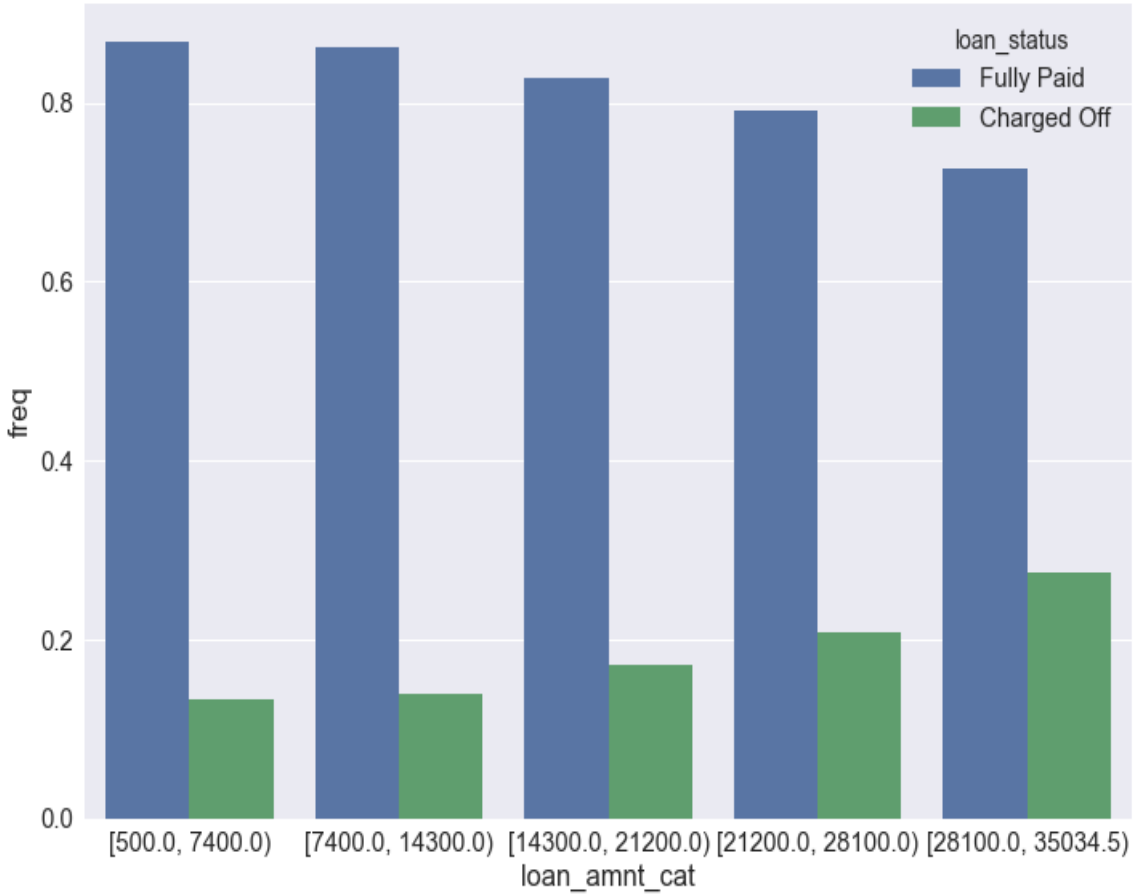
Observations

Based on correlation and heat map between loan primary attributes below attributes will be used for further analysis against loan status

- *loan_amnt*
- *int_rate*
- *term*
- *grade*
- *sub_grade*

Analysis of Loan Primary Attributes (cont.)

1. Loan Amount v/s Loan status

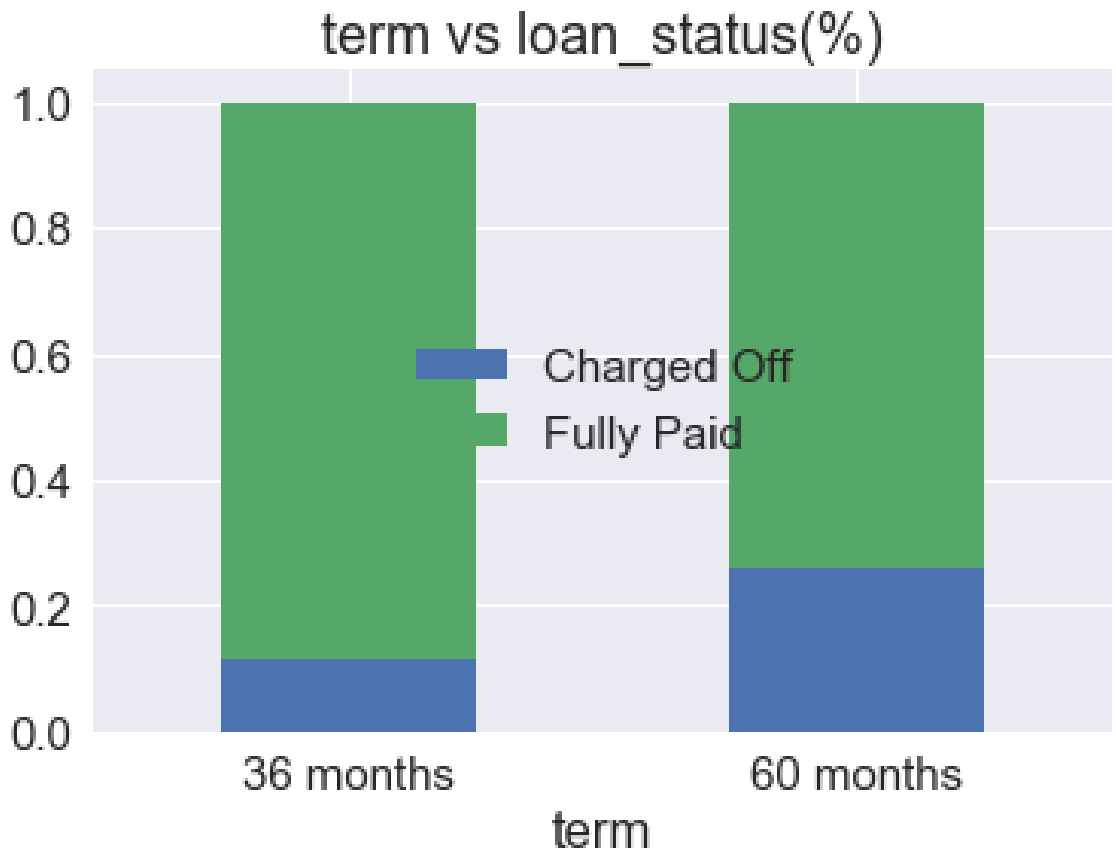


Observations

Probability of consumer defaulting is relatively higher in category of loan amount between $28100 and $35000

Analysis of Loan Primary Attributes (cont.)

2. Term v/s Loan status

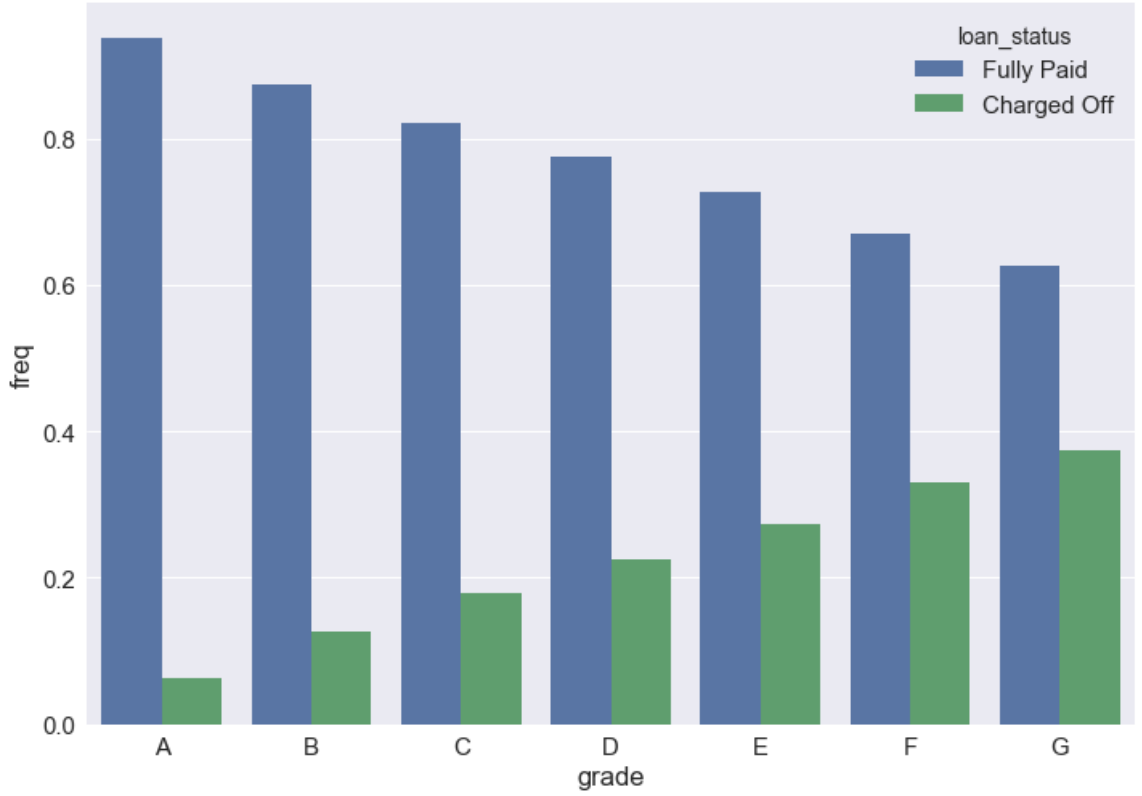


Observations

Probability of consumer defaulting is relatively higher in 60 months term

Analysis of Loan Primary Attributes (cont.)

3. Grade v/s Loan status

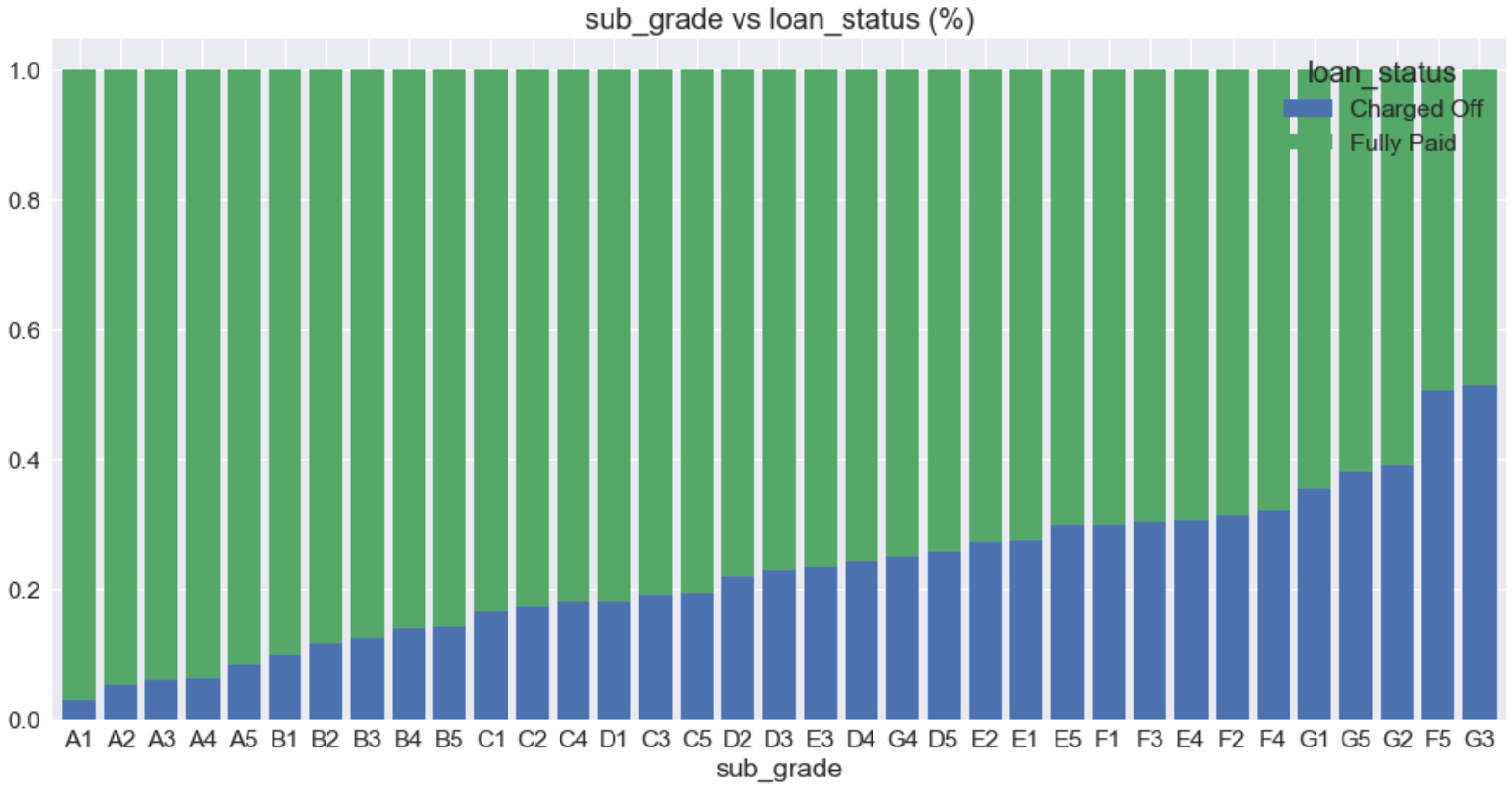


Observations

Probability of borrower defaulting increases as the grade lowers, the grade G shows the highest probability of default

Analysis of Loan Primary Attributes (cont.)

4. Sub-Grade v/s Loan status

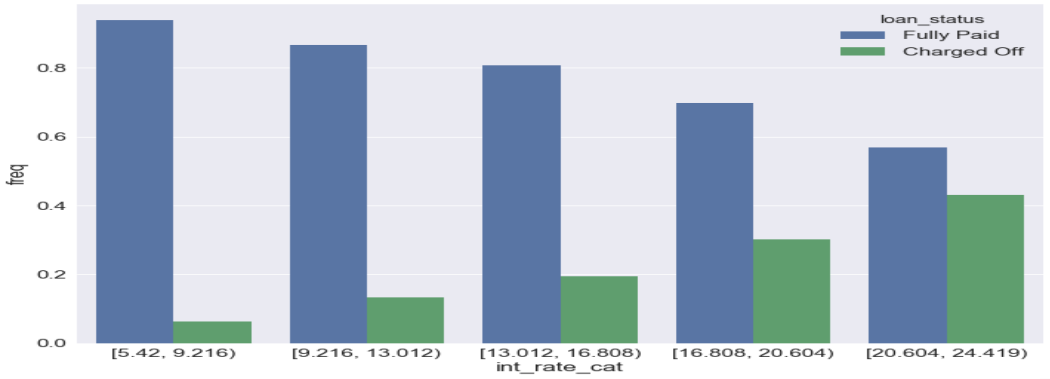


Observations

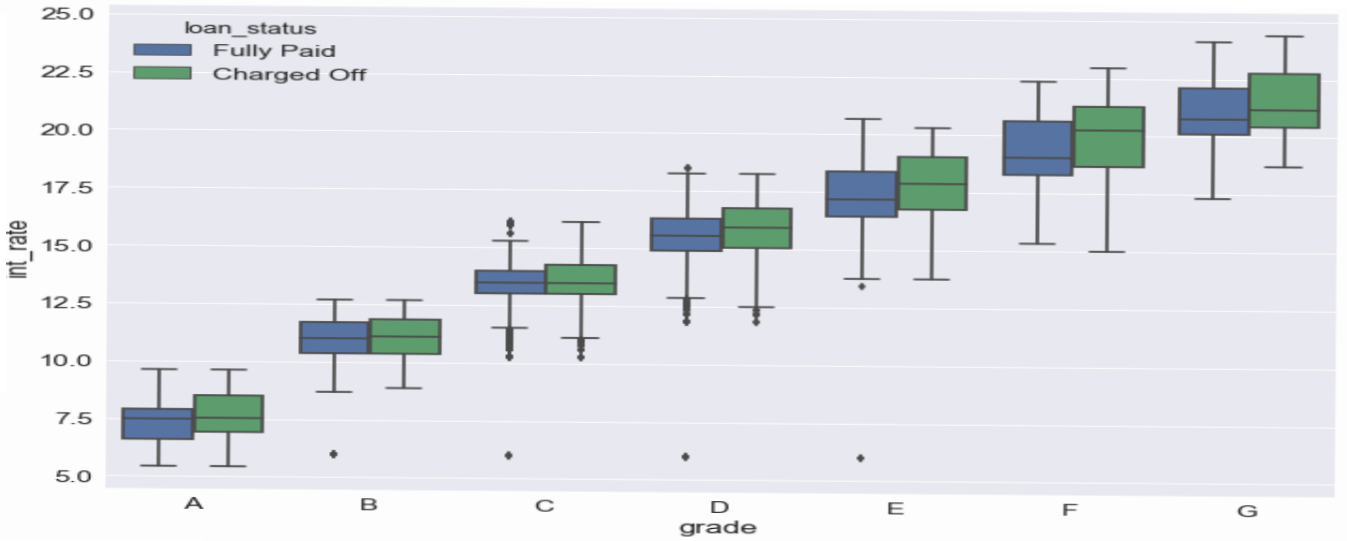
1. *Sub Grade F5 has 50% chances of a borrower defaulting*
2. *The plots show an increase in frequency of Charged off's for lower sub grades*

Analysis of Loan Primary Attributes (cont.)

5. Interest rate v/s Loan status



Interest rate v/s grade

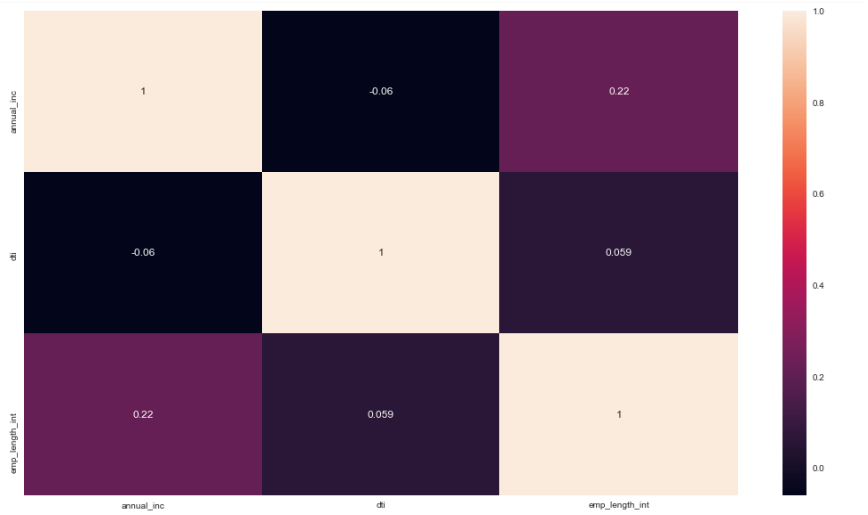


Observations

1. *Loans with interest rate in the bracket of 20% to 25% have highest frequency of default*
2. *Interest rates increases as the loan amount increases, higher interest rates have higher number of charge off's*
3. *Higher interest rates are assigned to lower grades*

Analysis of Consumer Attributes

Heat Map



Correlation matrix

	annual_inc	dti	emp_length_int
annual_inc	1.00	-0.06	0.22
dti	-0.06	1.00	0.06
emp_length_int	0.22	0.06	1.00

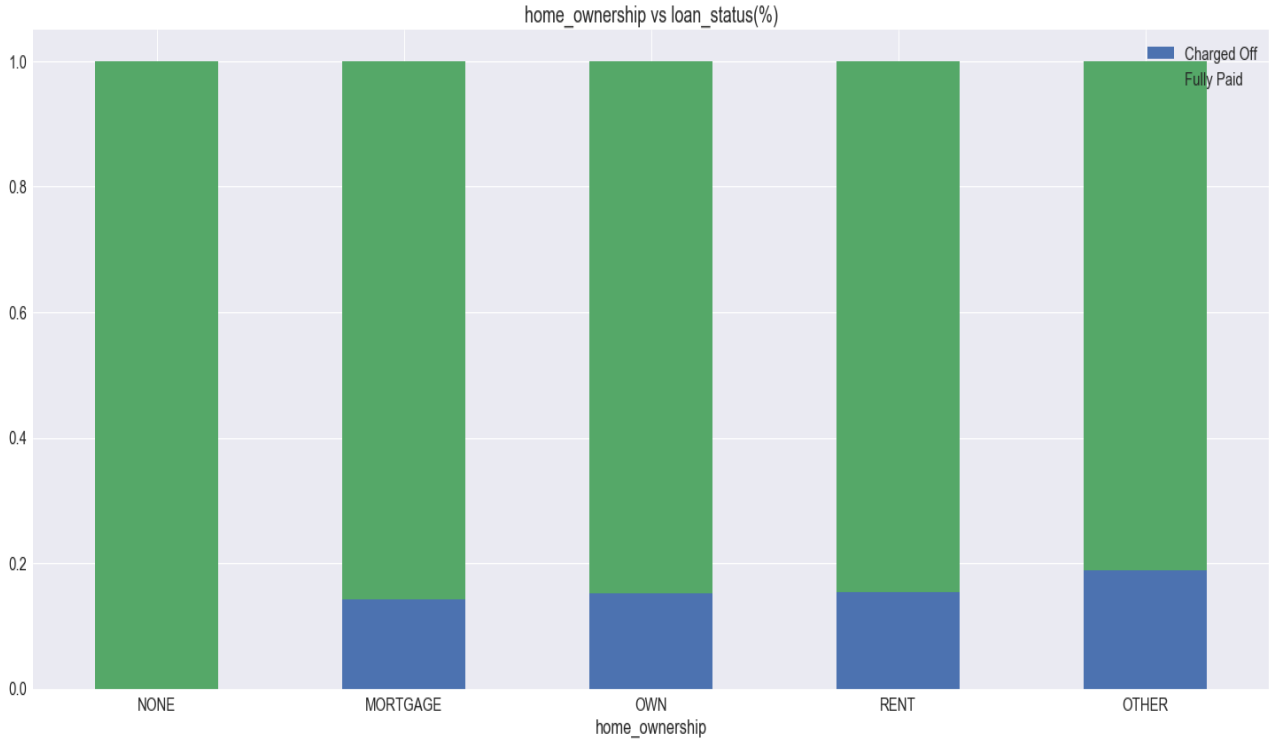
Observations

Observing the correlation between consumer attributes and heat map below attributes will be used for further analysis against loan status

- *emp_length_int*
- *home_ownership*
- *verification_status*
- *purpose*
- *addr_state*
- *issue_d*
- *dti*

Analysis of Consumer Attributes (cont.)

1. Home ownership v/s Loan status

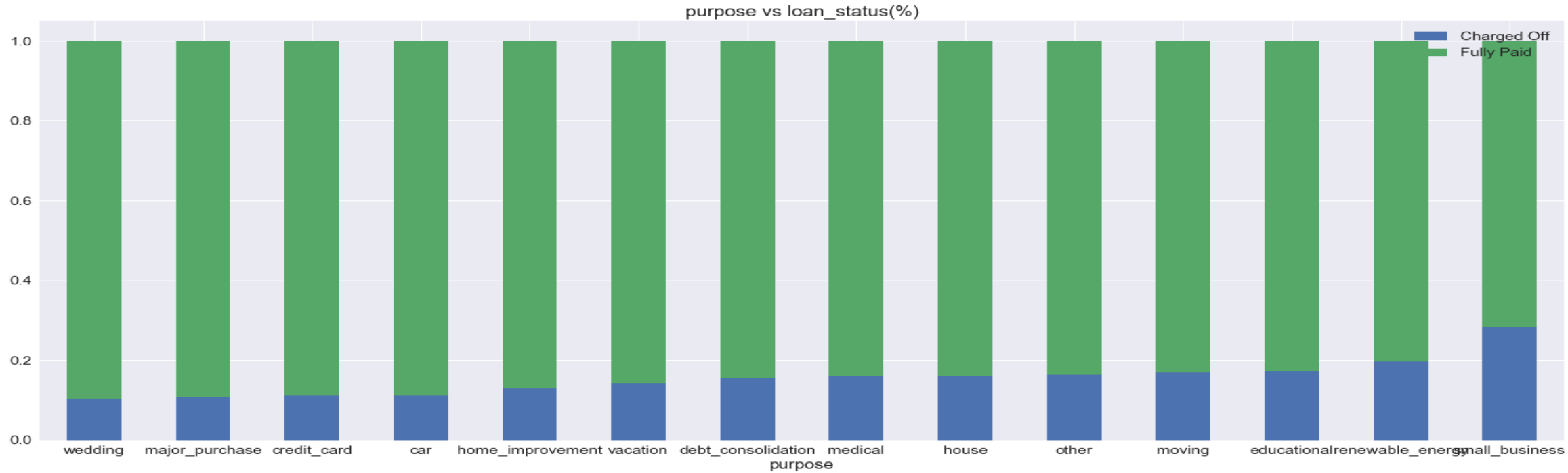


Observations

Borrowers with home ownership status as "other" have a higher default rate

Analysis of Consumer Attributes (cont.)

2. Purpose v/s Loan status

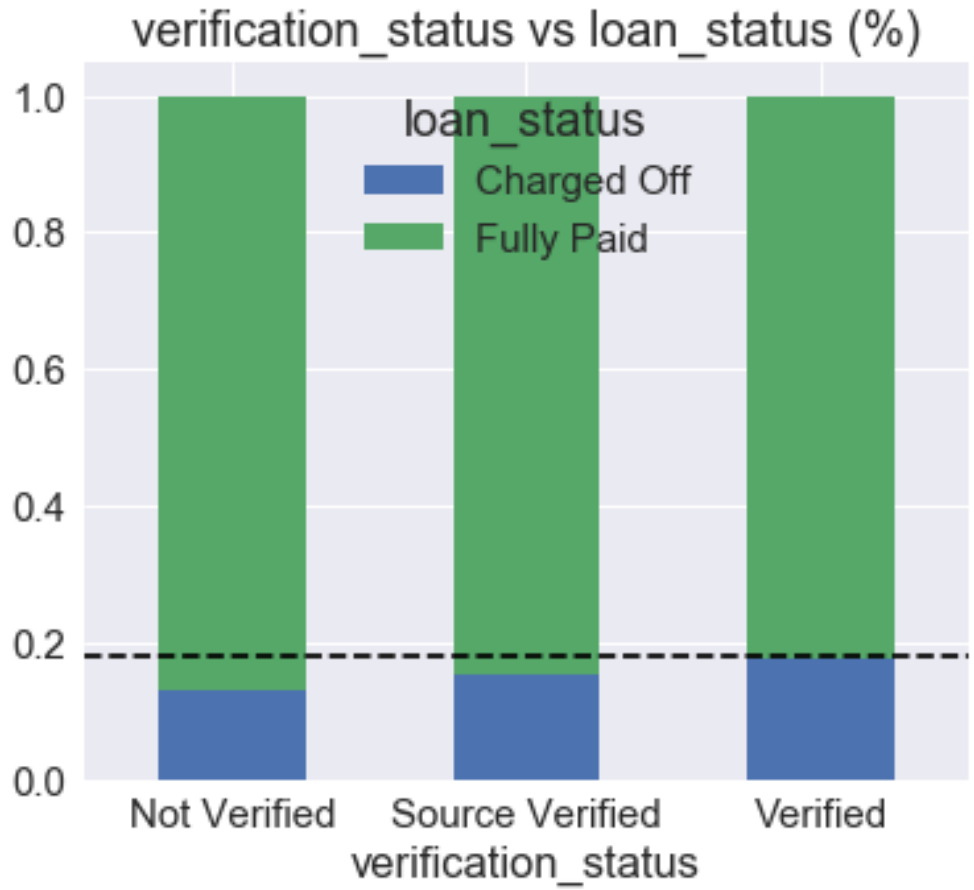


Observations

Borrowers with purpose as "small business" have a higher default rate

Analysis of Consumer Attributes (cont.)

3. Verification status v/s Loan status

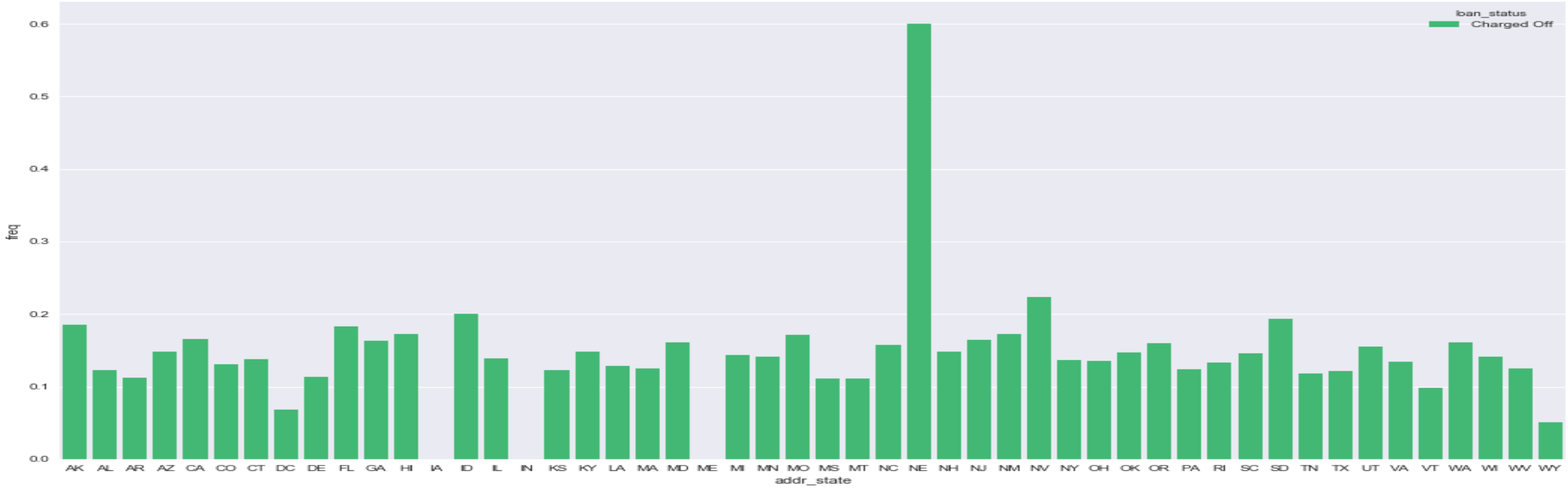


Observations

Relatively verified loans seem to have more defaults, although variation is very less

Analysis of Consumer Attributes (cont.)

4. Address state v/s Loan status

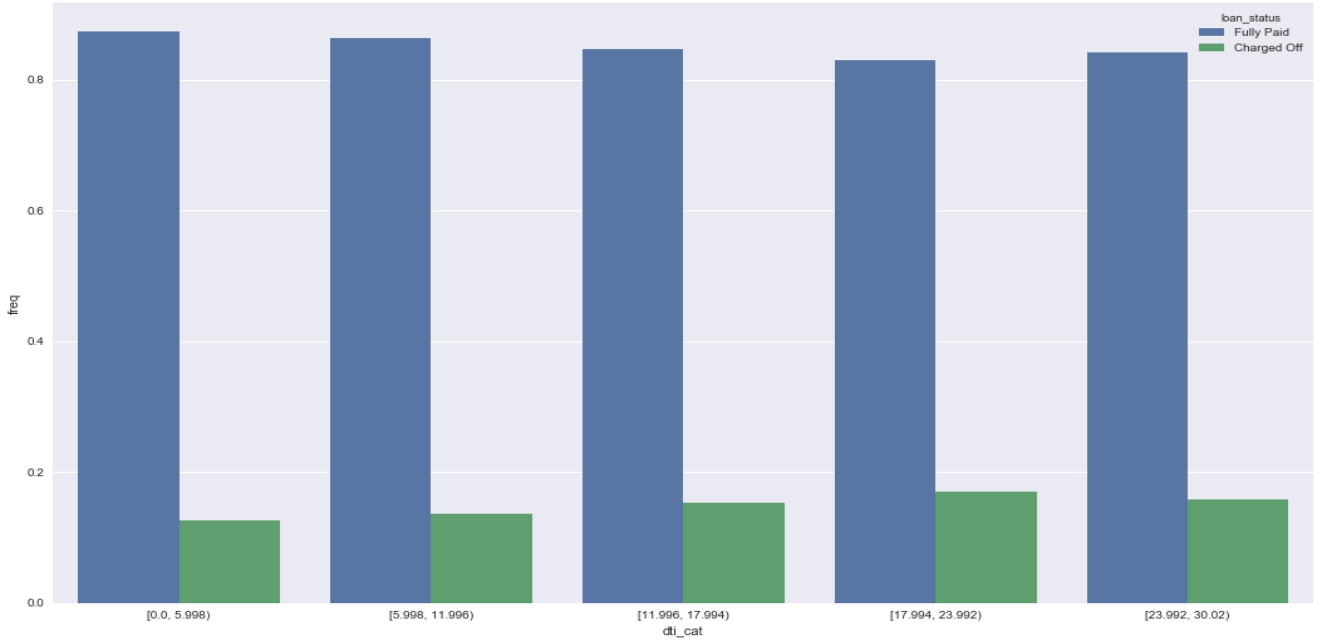


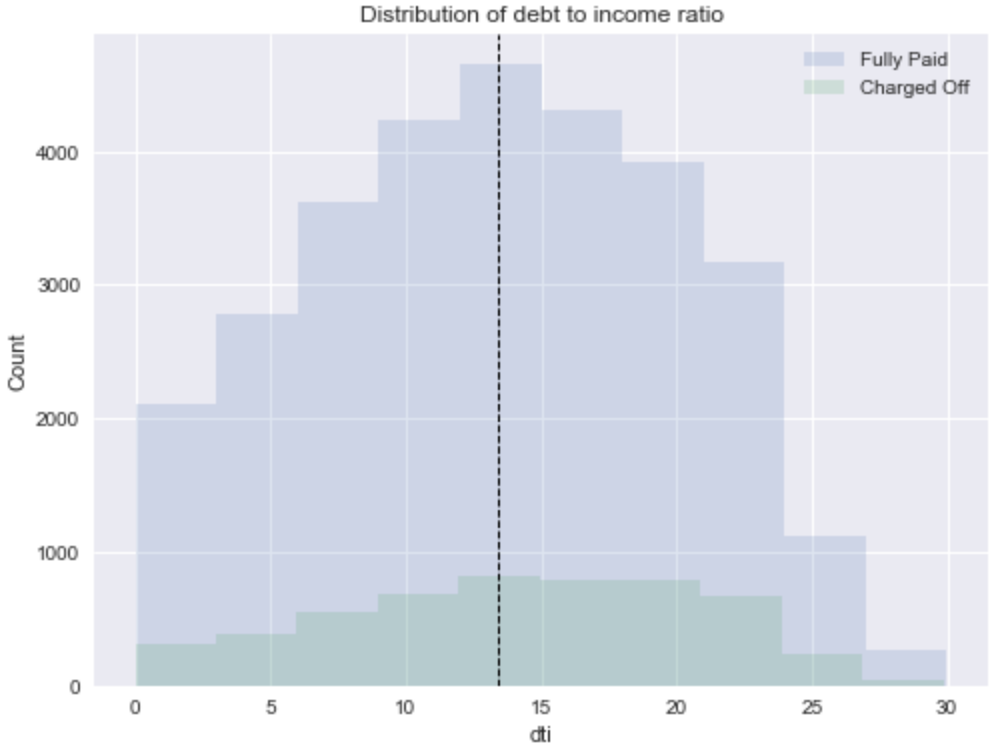
Observations

Nebraska (NE) followed by Nevada (NV) and Idaho(ID) shows higher incidences of borrowers defaulting.

Analysis of Consumer Attributes (cont.)

5. DTI v/s Loan status



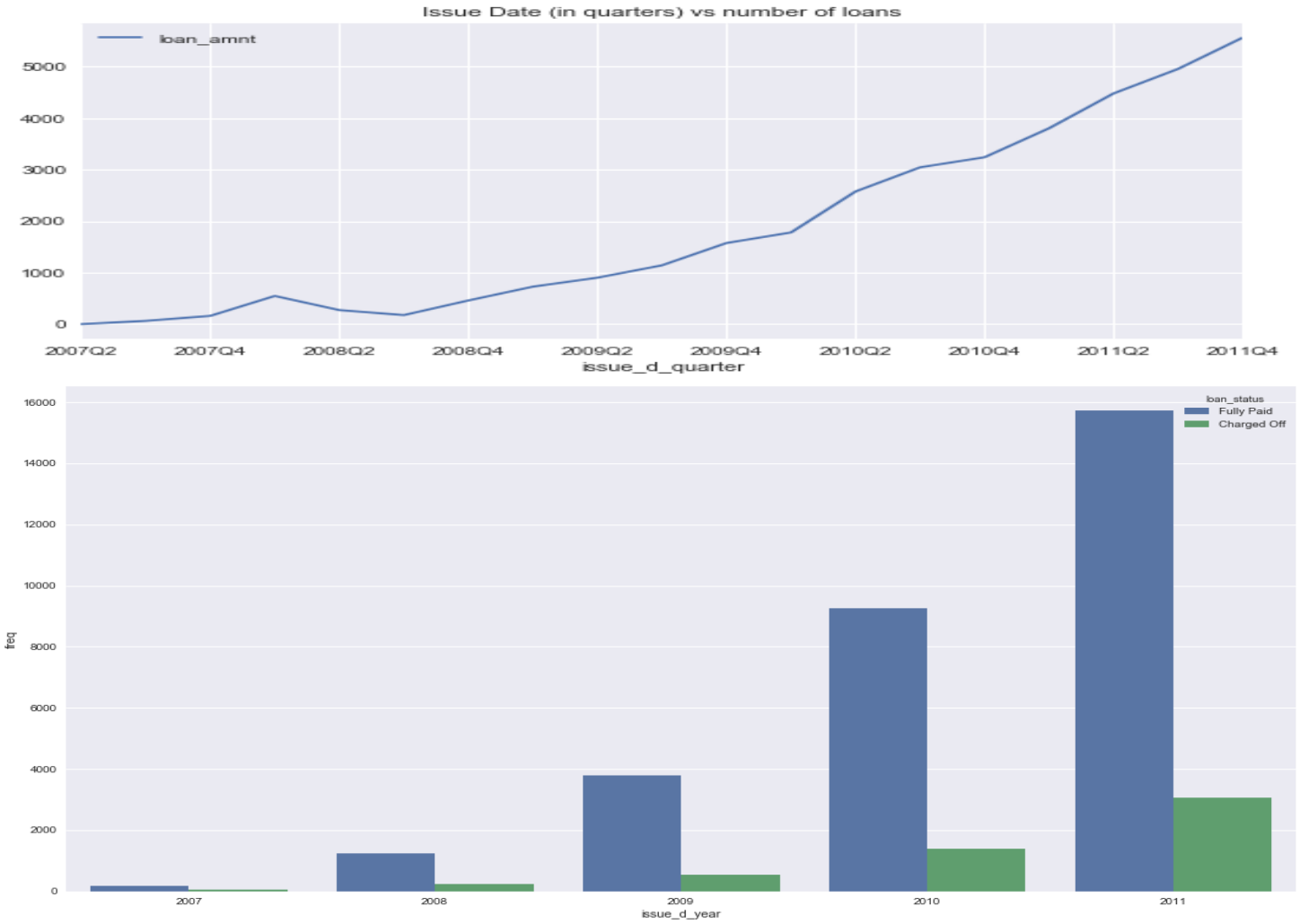


Observations

Probability of a borrower defaulting is high for borrowers with high dti

Analysis of Consumer Attributes (cont.)

6. Issue date v/s Loan status

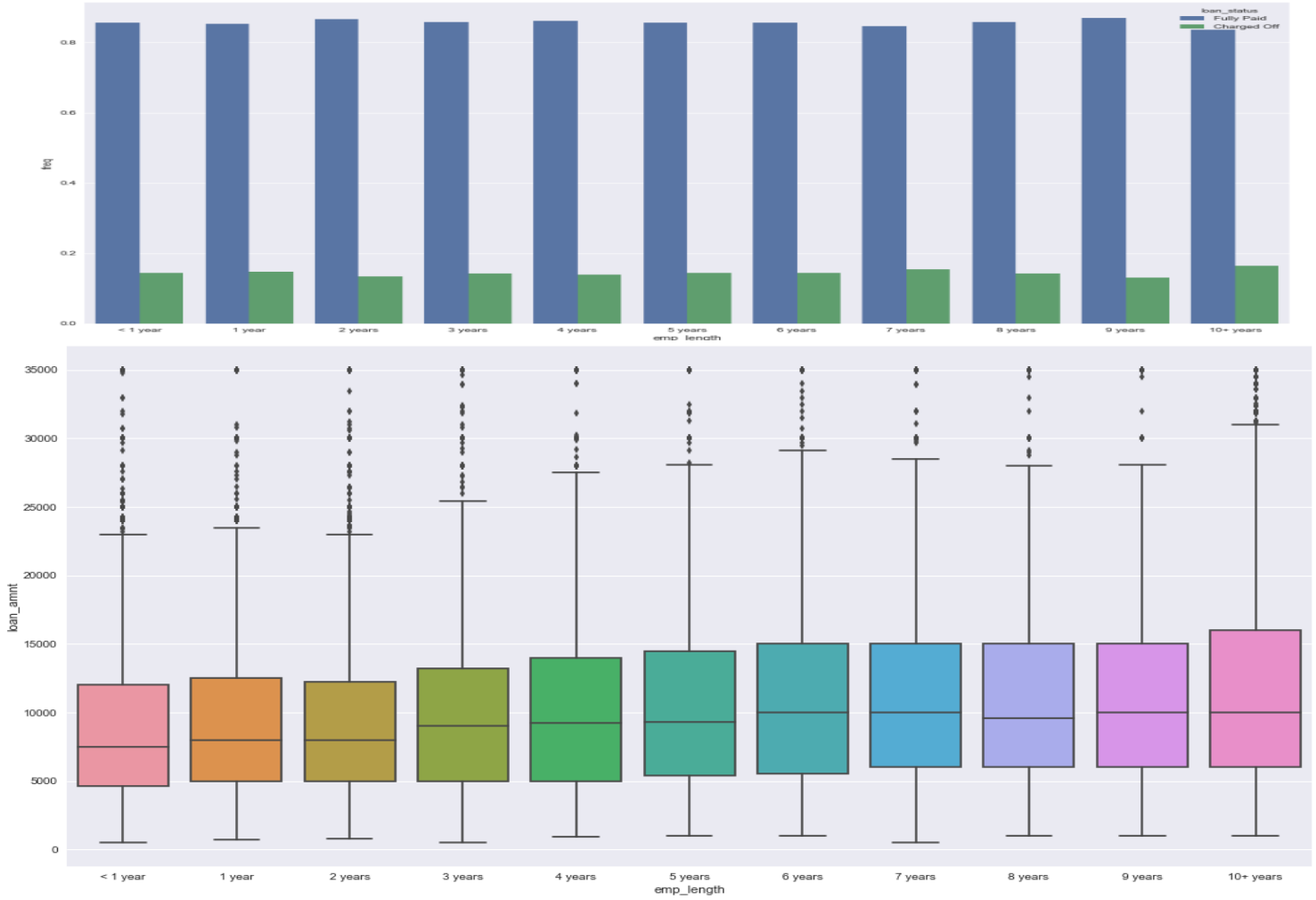


Observations

- *The first plot Shows an exponential increase in the number of loans issued over by quarters by years*
- *In the second plot ,there is a spike in the number of credit lines opened in borrower account in the Year 2K*

Analysis of Consumer Attributes (cont.)

7. Emp length v/s Loan status



Observations

- *Employees with higher work exp. (10+ years) tend to default more than those with lower work exp.*
- *One reason for borrowers employed longer defaulting more could be because they are taking larger loan amounts*

Summarize

After exploratory data analysis we have identified below attributes will influence the tendency of default

consumer attributes
emp_length_int
home_ownership
verification_status
purpose
addr_state
issue_d
dti

loan attributes
loan_amnt
int_rate
term
grade
sub_grade

Thank You