LENDING CLUB CASE STUDY ASSIGNMENT

NAME: AKASH GARG
DIVYA PEACHI

BUSINESS UNDERSTANDING

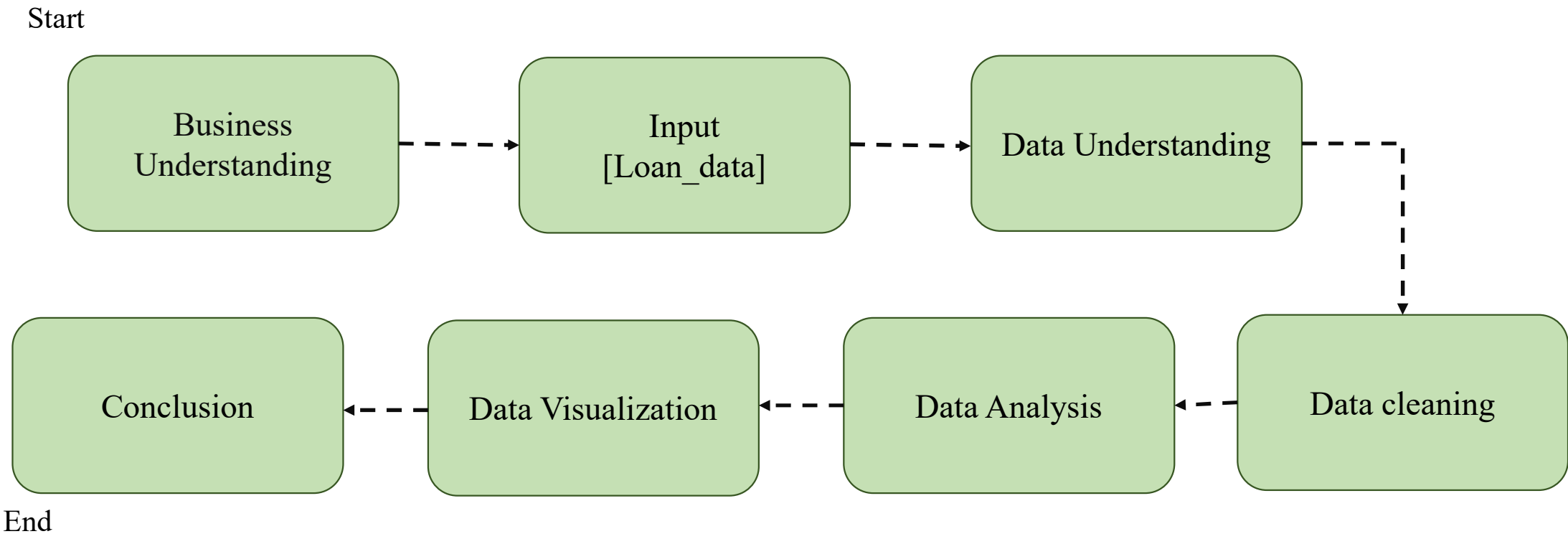
- Urban residents are the primary target market for the company's consumer financing. It is up to the corporation to decide whether or not to provide the loan after receiving the application.
- There are two types of risks involved in making the decision –
  - ✓ It is detrimental to the company's bottom line if the loan is not approved because the applicant is unlikely to repay the loan.
  - ✓ Accepting a loan from someone who isn't likely to pay it back, or who has a history of default, could put the company in jeopardy of incurring losses.
- In order to determine if a person is likely to default or not, we need to look for patterns.
- When someone requests for a loan, the corporation can make two types of decisions:
  - ✓ An application for a loan accepted: Following the company's approval of the loan, the following are the three probable outcomes if the company approves the loan:
    - Current
    - Fully Paid
    - Charged-off
  - ✓ Loan rejected

BUSINESS OBJECTIVE

- Largest source of financial loss comes from making loans to "high-risk" customers. Lenders suffer the most when borrowers fail to make payments as agreed. When a customer's account gets charged-off, they are said to be delinquent or defaulters.
- To limit the quantity of credit losses, a corporation must be able to detect these hazardous loan applications. The goal is to identify those candidates.
- To put it another way, the organization seeks to identify the elements that serve as strong predictors of loan default in order to better manage risk. This information can be utilized in the company's portfolio and risk analysis.

METHODOLOGY

The flow chart described below depicts the major steps to approach problem solving methodology



UNDERSTANDING THE DATA

- The data has total 39717 rows and 111 columns
- The data is in three different forms of data types:
  - Where 74 columns from total 111 columns are float
  - 13 columns from total 111 columns are int
  - 24 columns from 111 are object
- The size or the memory usage of the input data is 33.6 MB
- By understanding the data we recognized that our goal is to compare the values of loan status across all other variables or attributes.

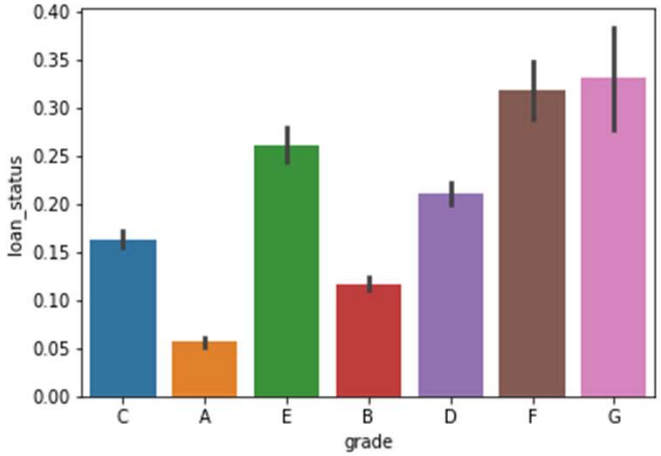
DATA CLEANING AND MANIPULATION

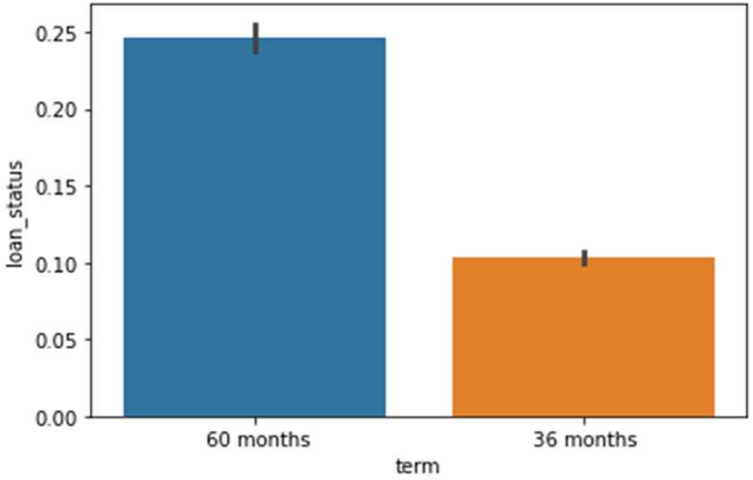
- For each column in the data frame, the number of unused rows is first determined. As a result, any column in which the empty value is greater than 50% is deleted or dropped.
- Each column's missing value rows are omitted.
- Rows with only one unique value in several columns are discarded since they are considered to be garbage.
- Many superfluous or irrelevant columns, such as unique ids, customer behavior variables, and so on, have been omitted from our analysis as it doesn’t play any crucial role.

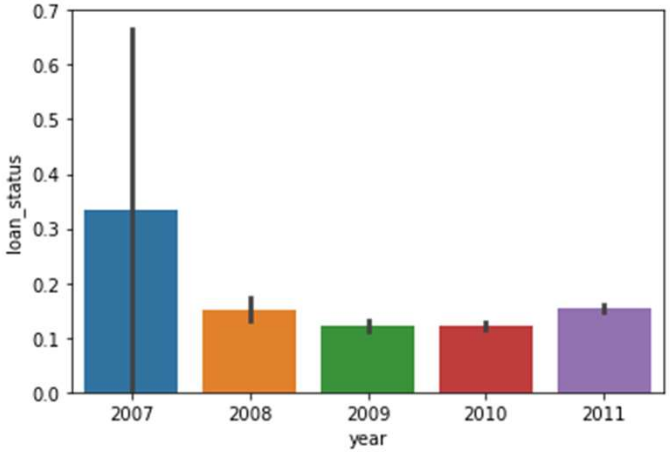
ANALYSIS

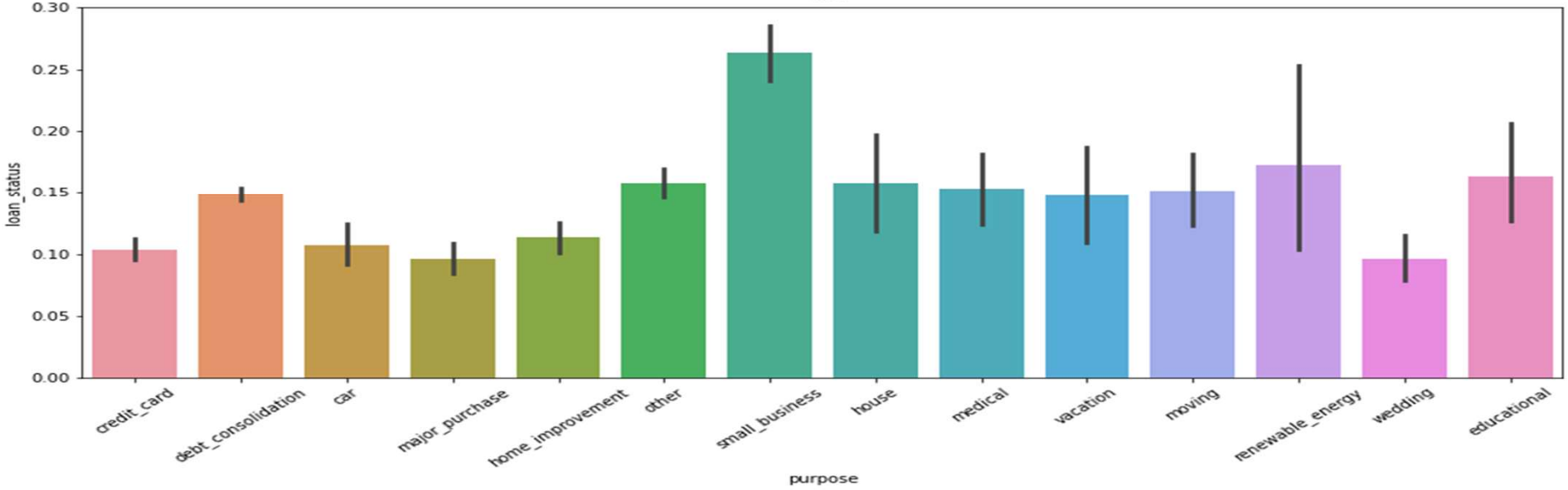
UNIVARIATE ANALYSIS

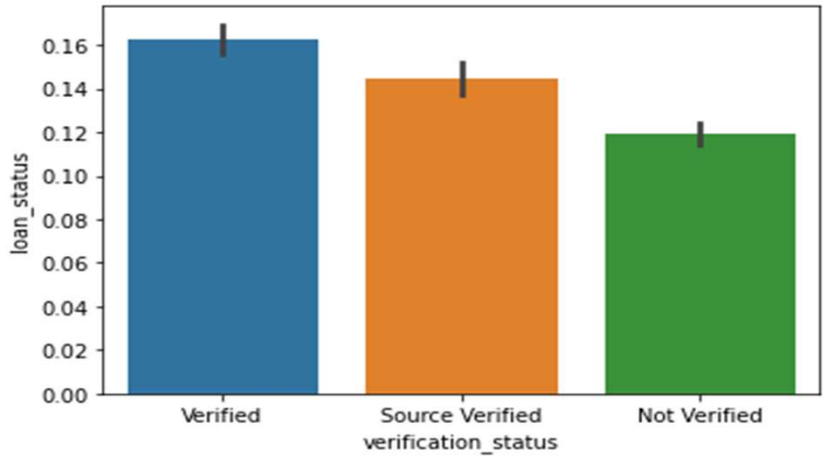
- Bar graphs are utilized here to show the default rates for a variety of variables.
- December is the month when the most defaults occur.
- Approximately 14% of all loans have been repaid.

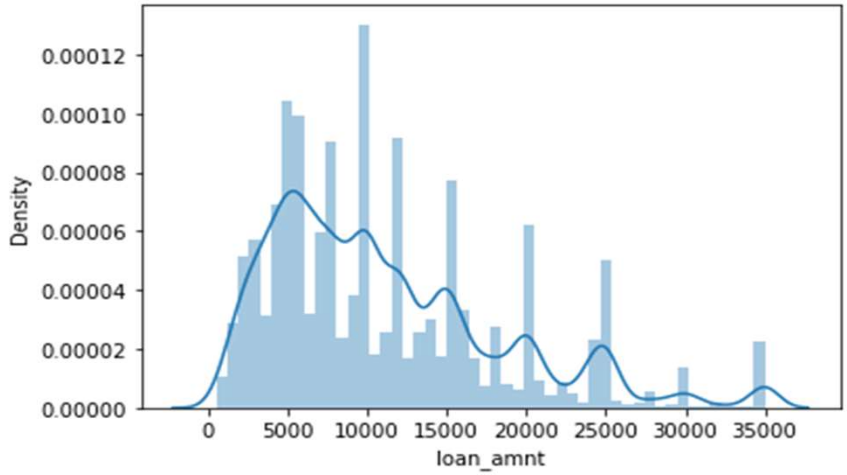

loan_status
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
C
A
E
B
D
F
G
grade

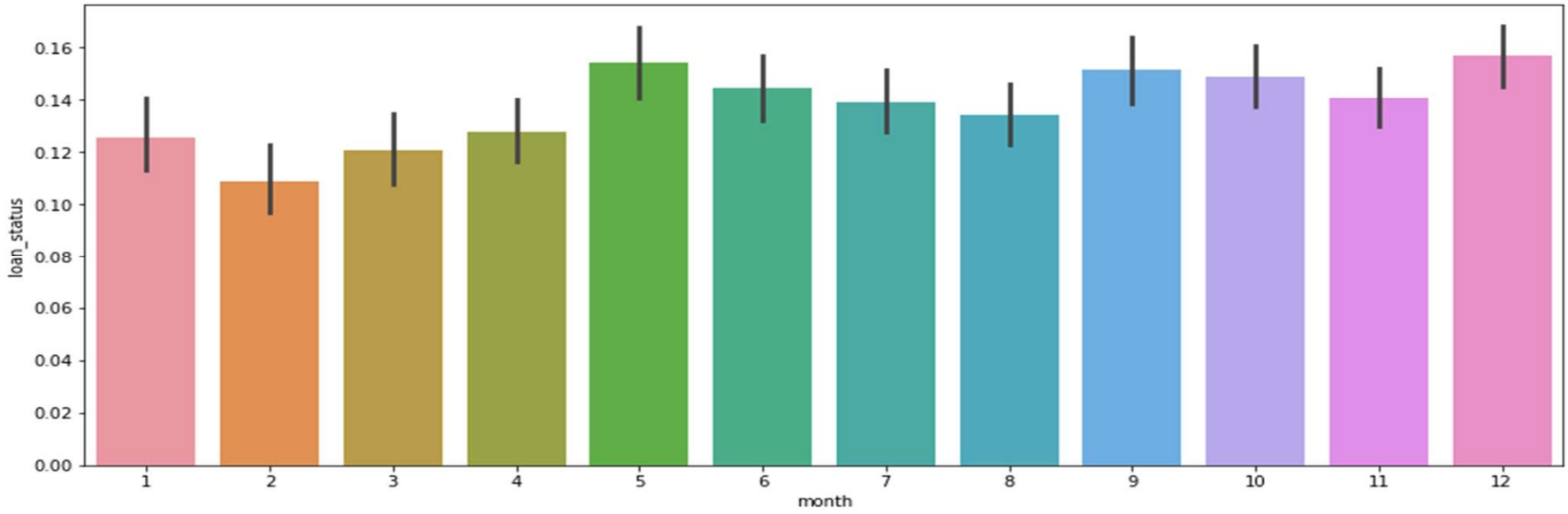

loan_status
0.25
0.20
0.15
0.10
0.05
0.00
60 months
36 months
term


loan_status
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2007
2008
2009
2010
2011
year


loan_status
0.30
0.25
0.20
0.15
0.10
0.05
0.00
credit_card
debt_consolidation
car
major_purchase
home_improvement
other
small_business
house
medical
vacation
moving
renewable_energy
wedding
educational
purpose

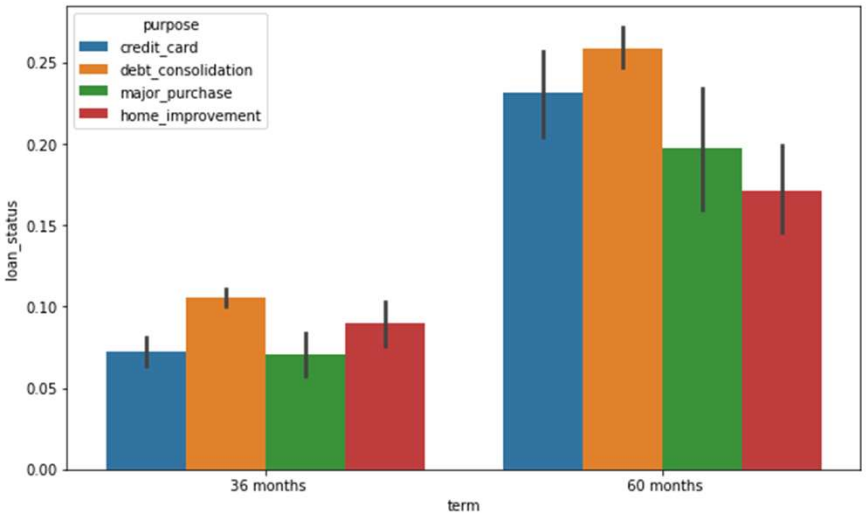

loan_status
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Verified
Source Verified
Not Verified
verification_status

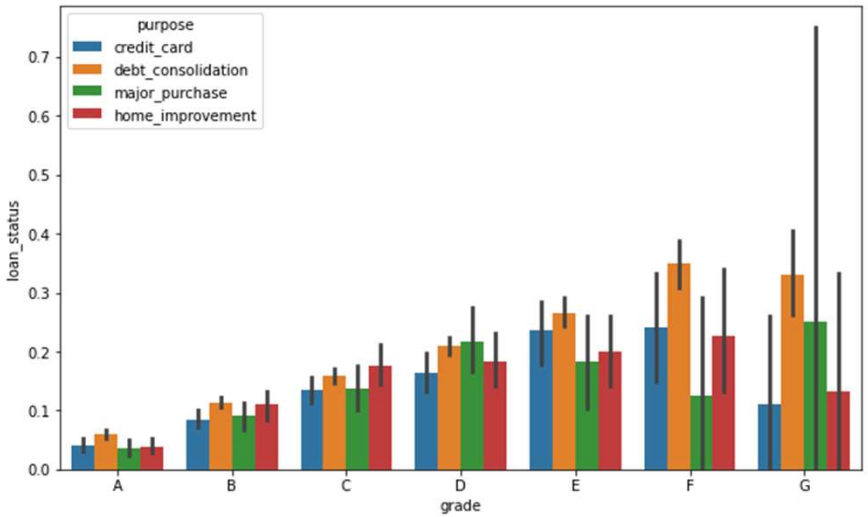

Density
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
25000
30000
35000
loan_amnt

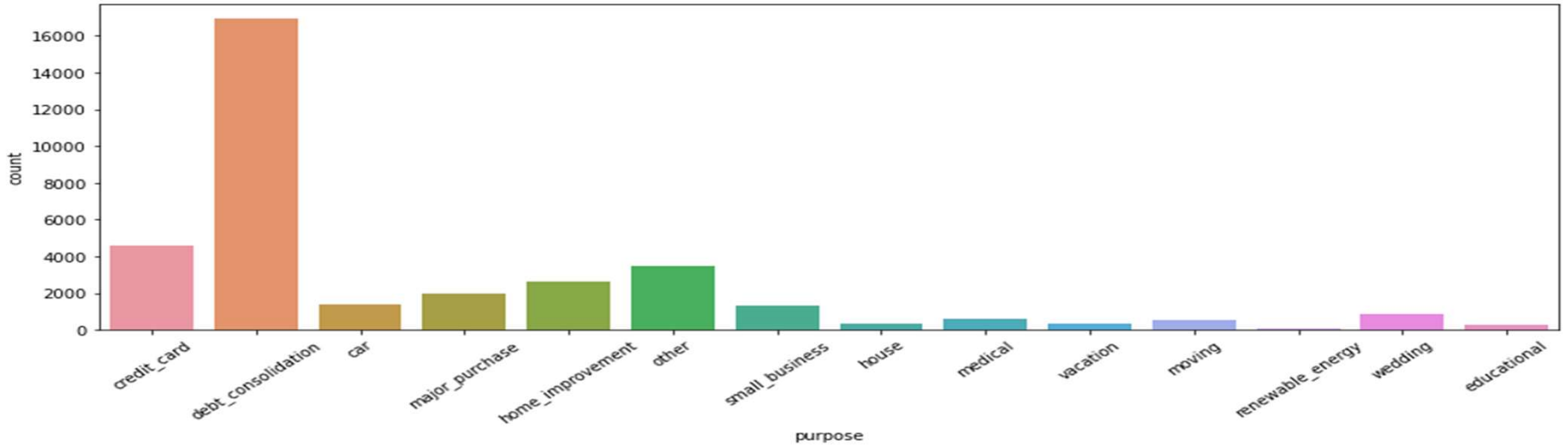

loan_status
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
6
7
8
9
10
11
12
month

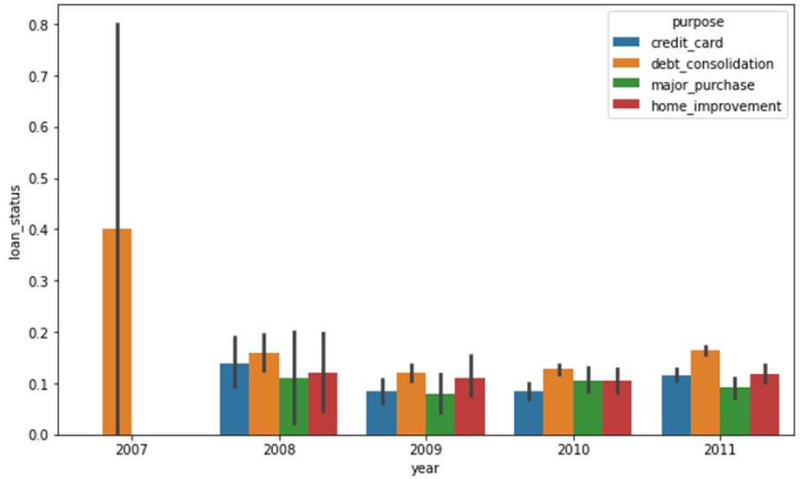
BIVARIATE ANALYSIS

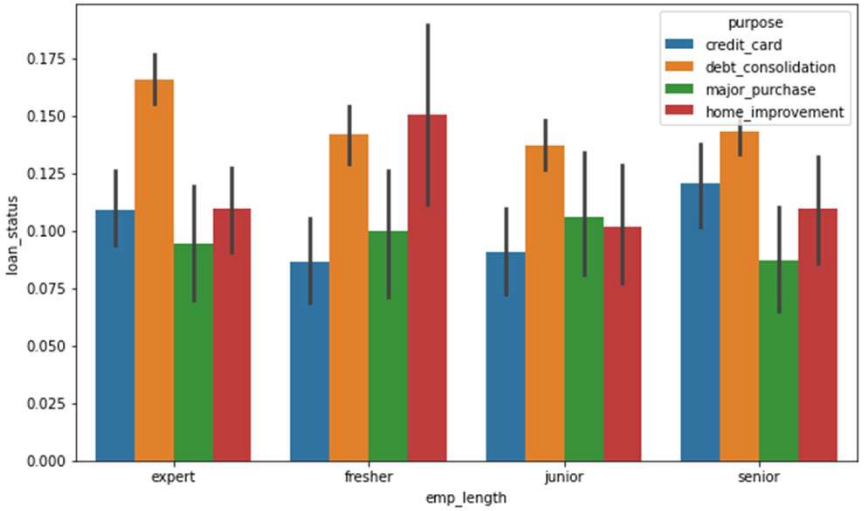
- Two column variables are used in this section of the analysis in order to see how defaulter behavior changes with the variables.
- To see which states in the US have the most defaulters, a heatmap is employed.
- There are a number of factors that contribute to loan defaults, and these are called driving variables.
- In the end, the top five driving factors have been identified and listed.

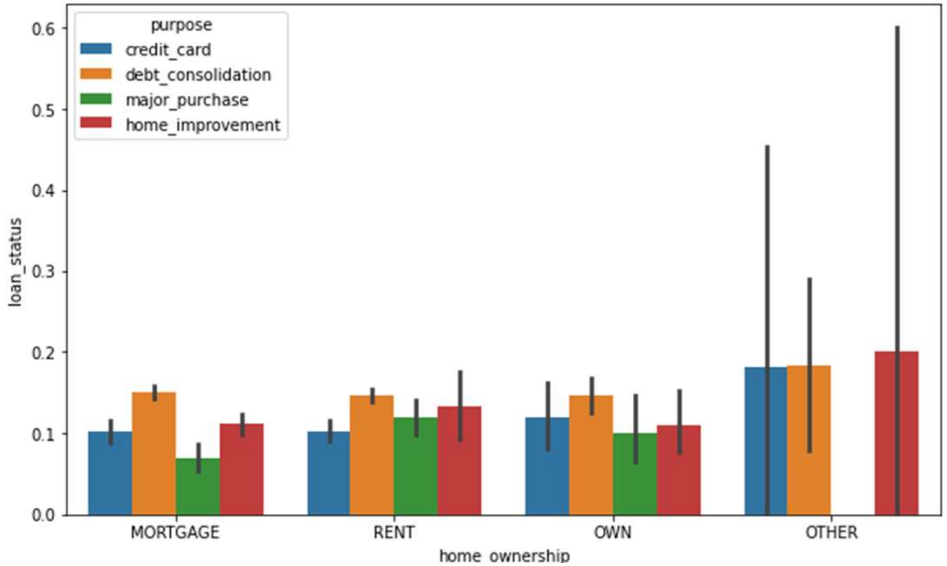

purpose
credit_card
debt_consolidation
major_purchase
home_improvement
loan_status
0.00
0.05
0.10
0.15
0.20
0.25
36 months
60 months
term


purpose
credit_card
debt_consolidation
major_purchase
home_improvement
loan_status
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
A
B
C
D
E
F
G
grade


count
0
2000
4000
6000
8000
10000
12000
14000
16000
credit_card
debt_consolidation
car
major_purchase
home_improvement
other
small_business
house
medical
vacation
moving
renewable_energy
wedding
educational
purpose


purpose
credit_card
debt_consolidation
major_purchase
home_improvement
loan_status
year
2007
2008
2009
2010
2011


purpose
credit_card
debt_consolidation
major_purchase
home_improvement
loan_status
emp_length
expert
fresher
junior
senior

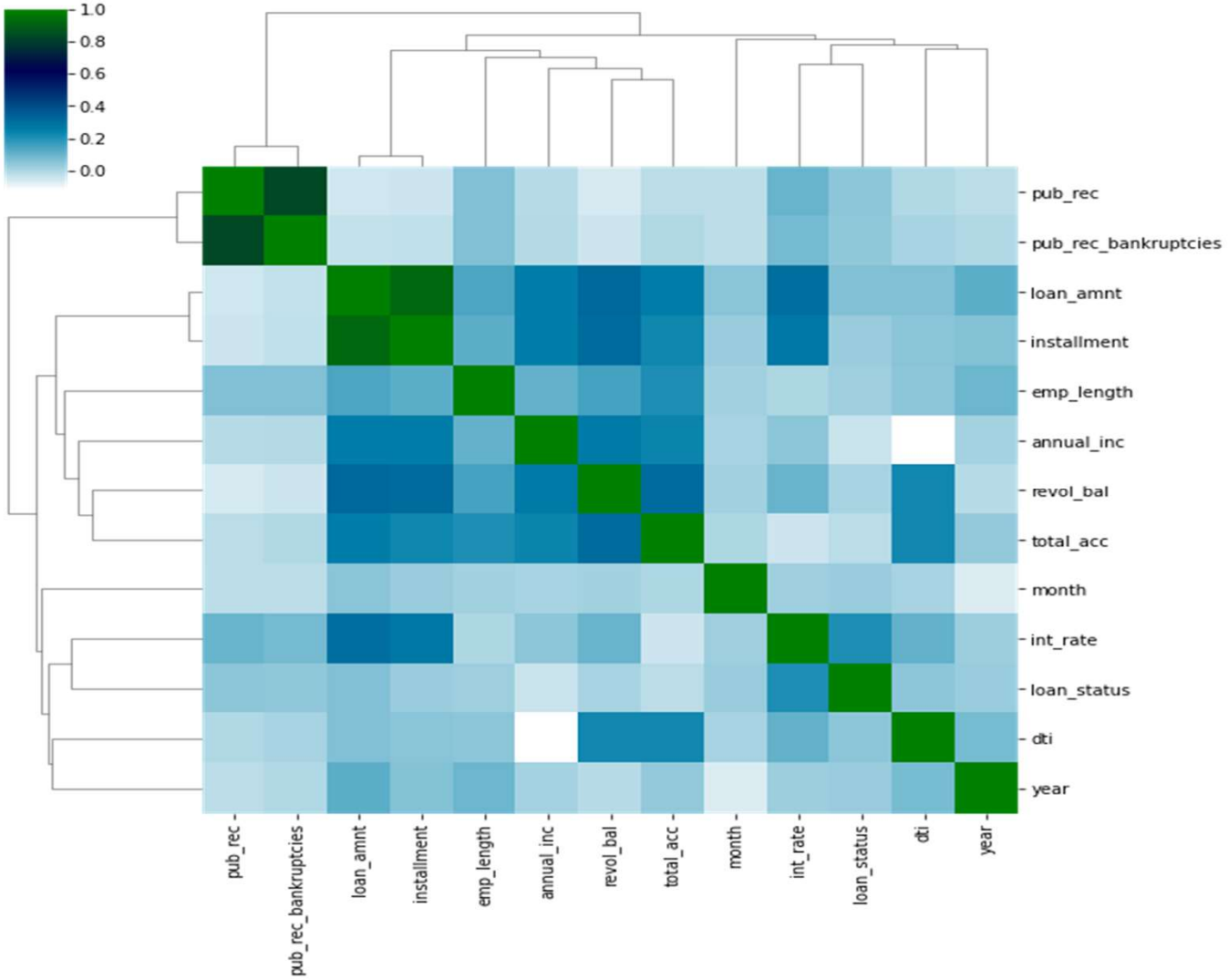

purpose
credit_card
debt_consolidation
major_purchase
home_improvement
loan_status
home_ownership
MORTGAGE
RENT
OWN
OTHER

RECOMENDATION

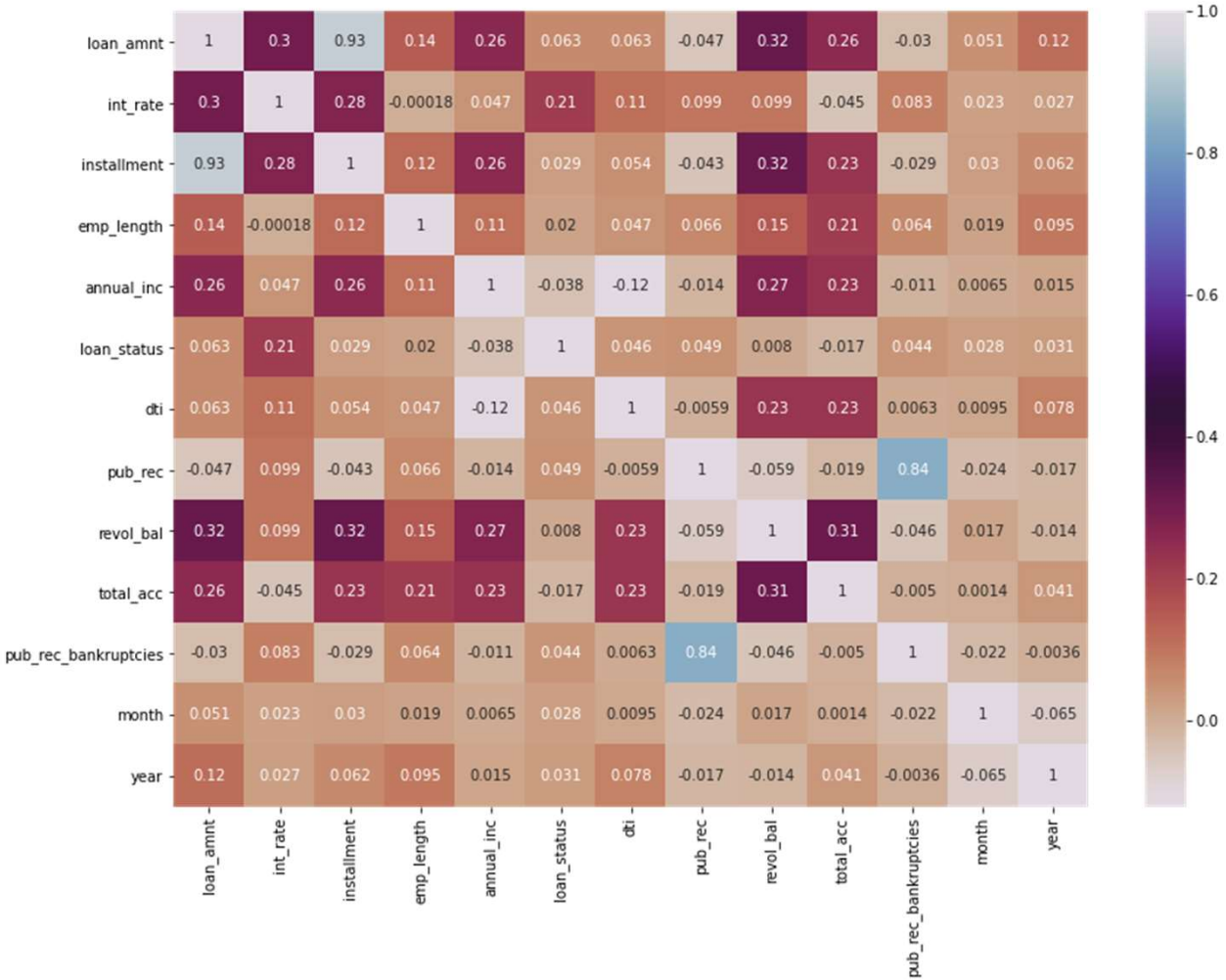
- Don't give out loans to those who've had a negative history in the past. At the very least, stop approving loans with excessive interest rates.
- Minimize the amount of approvals for small businesses.
- Increase interest rates on loans with a dti larger than 20 percent of the total
- Stop admitting applicants whose incomes are greater than 30% of the state's average.

CONCLUSION

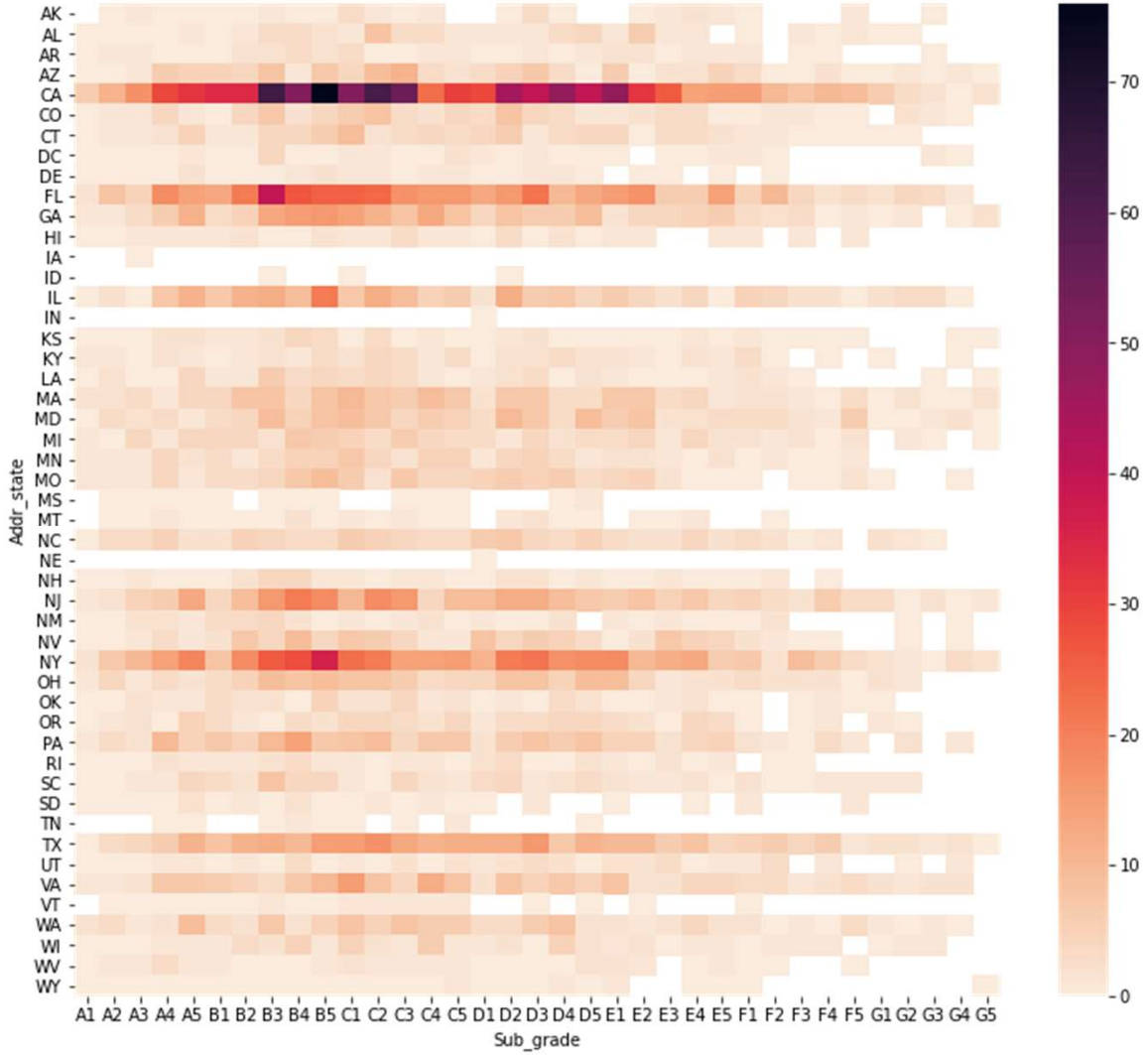
- The following columns are important driving factors, 'sub_grade', 'pub_rec_bankruptcies', 'addr_state', 'grade' and 'int_rate','revol_util','revol_bal'. Since grade and subgrade are dependent variables we can add the 'term' column also as a driving factor for defaulting
- As per the heatmap we can conclude location of the borrower has an important role in identifying the defaulting behaviours, but there are several other column variables which have quite and influence on the defaulters behaviour.
- The loan_amnt and its correlation with the few variables
  - int_rate has 0.3 with loan_amnt
  - annual_inc has 0.26 with loan_amnt
  - revol_bal has 0.32 with loan_amnt
  - total_acc has 0.26 with loan_amnt
  - The loan amount given to the customers without considering the annual_income(where few loan applicants have taken the loan amount 50% more than there annual income).This scenario has to be taken care.
  - There is a positive correlation of loan status with the term and interest rate (as compared to other variables) and the same was also observed during univariate analysis



- The loan amount given to the customers without considering the annual_income(where few loan applicants have taken the loan amount 50% more than there annual income).This scenario has to be taken care.
- There is a positive correlation of loan status with the term and interest rate (as compared to other variables) and the same was also observed during univariate analysis



- int_rate has 0.3 with loan_amnt
- annual_inc has 0.26 with loan_amnt
- revol_bal has 0.32 with loan_amnt
- total_acc has 0.26 with loan_amnt



As per the heatmap we can conclude location of the borrower has an important role in identifying the defaulting behaviours, but there are several other column variables which have quite and influence on the defaulters behaviour.