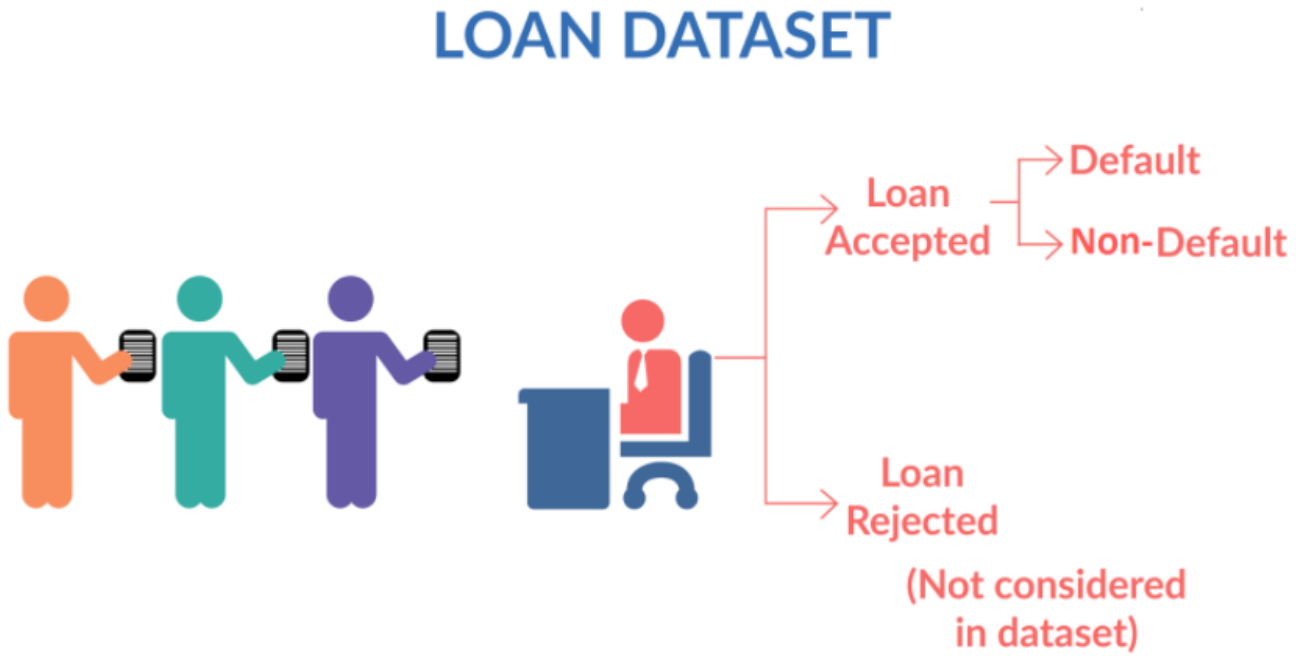
Problem Statement



Understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.

Analysis Approach

Data Cleaning

- Identifying missing values in each column
- Removing columns with more then 80% missing values

Data Analysis

- Identifying important variables
- Univariate Analysis
- Binning for Continuous Variables
- Segmented Univariate Analysis

Final Recommendations

Data Cleaning

- Looking at the dataset and identifying the columns
- Identifying percentage of missing values in each column
- Many columns have 100% missing values
- Removing columns having more then 80% of missing values from the dataset
- Visualizing and identifying slightly higher number of missing values (final column numbers - 53)
- Identifying missing values in rows
- Identifying rows having more then 10% missing values – 0%

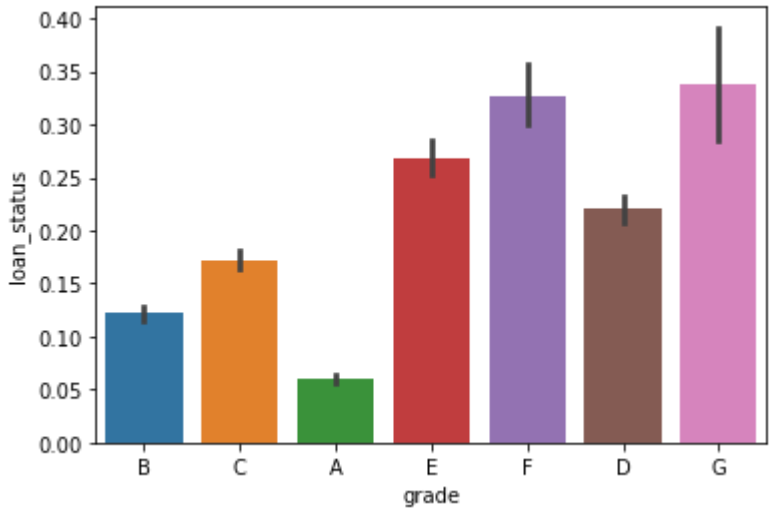
Data Analysis

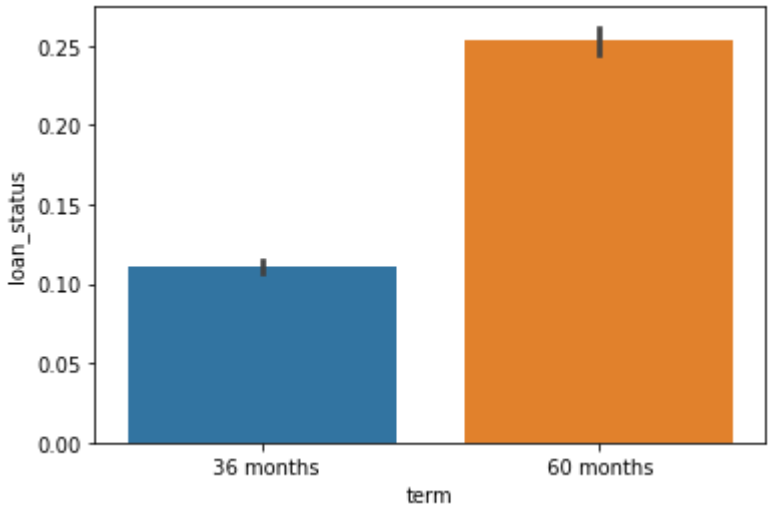
- Removing some of the unrequired variables (based on understanding) remaining columns 29

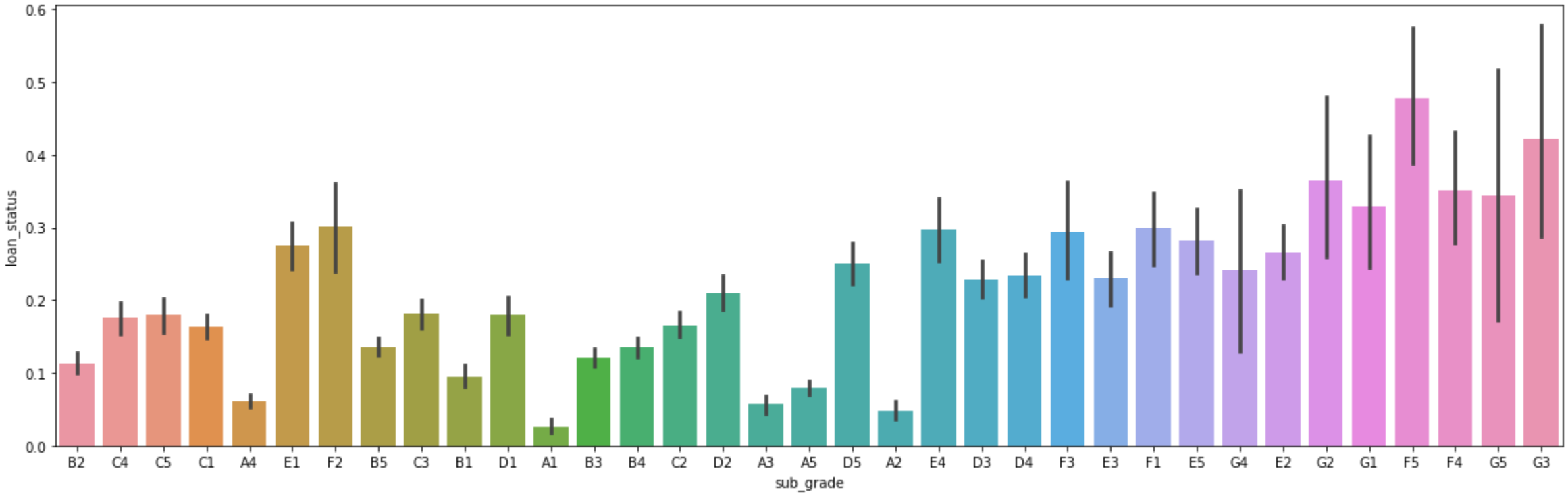
Univariate Analysis:

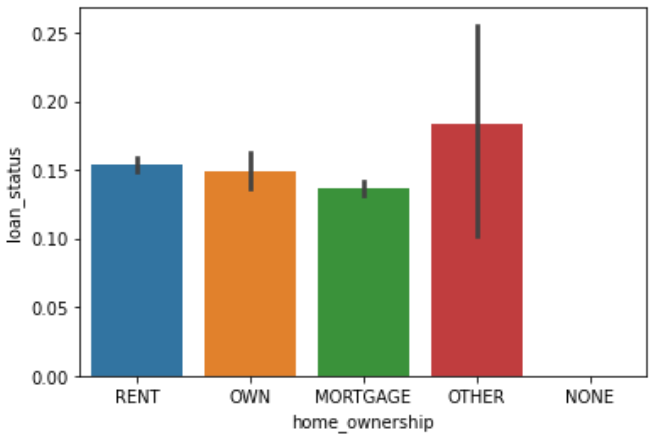
- Considering loan default rate to be the constant variable mean default rate – 0.15
- grade of loan goes from A to G, the default rate increases
- Default rate for 60 month term is mire then 36 month term
- Default rate increases with increase in sub grade i.e. default rate for G2 is higher then G1, for F5 is higher then F1 etc.
- Default rate is not much varying with home ownership
- Verified loans default higher then not verified loans
- Small business loan default the most
- default rate had continuously decreased from 2007 to 2010 then suddenly increased in 2011
- since default values for every month lie in between 12 to 17% there's not much variation across months

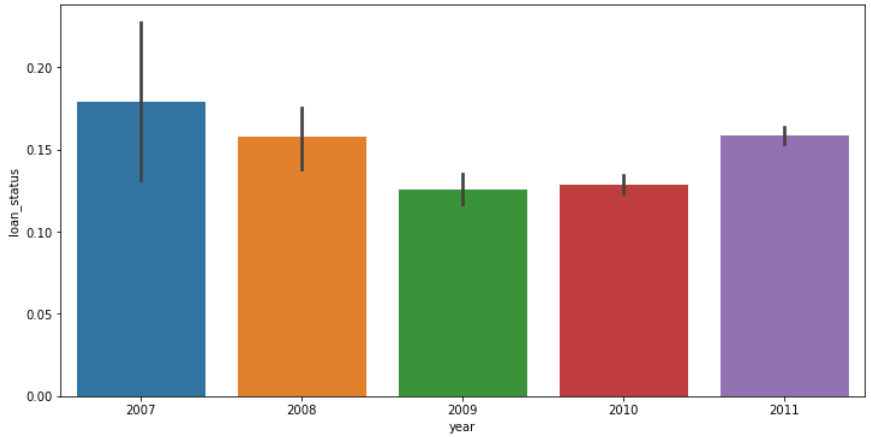
Key Findings

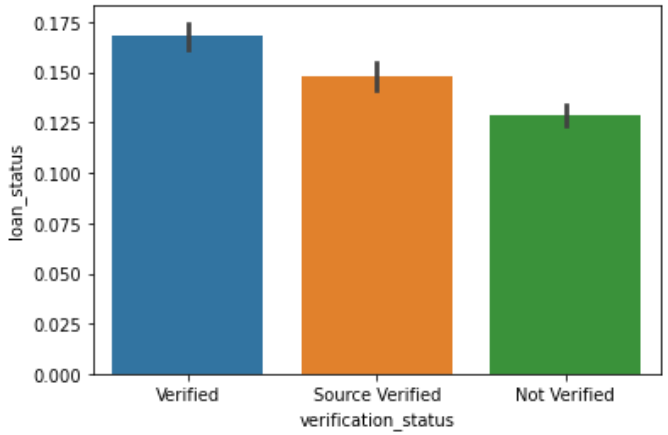


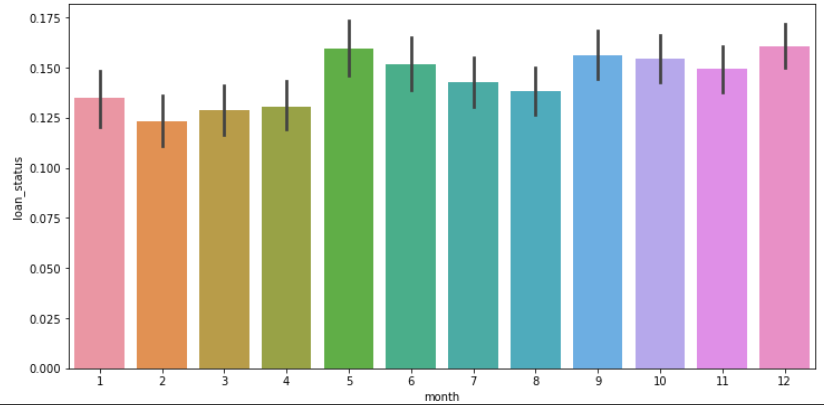


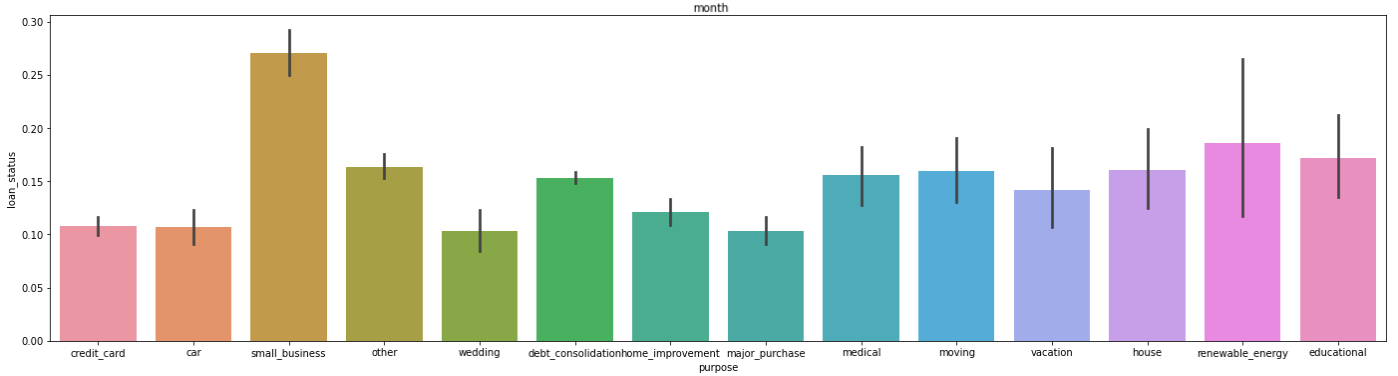



loan_status
0.25
0.20
0.15
0.10
0.05
0.00
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership


loan_status
0.20
0.15
0.10
0.05
0.00
2007
2008
2009
2010
2011
year


loan_status
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Verified
Source Verified
Not Verified
verification_status

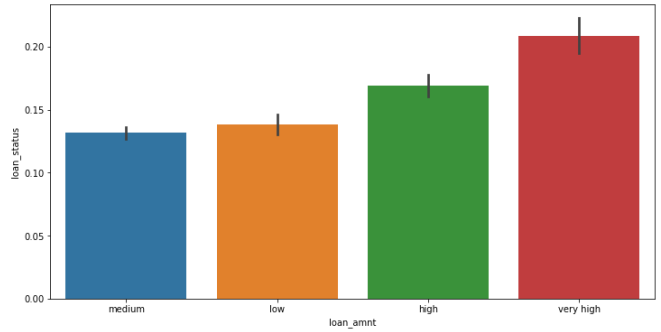

loan_status
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
1
2
3
4
5
6
7
8
9
10
11
12
month

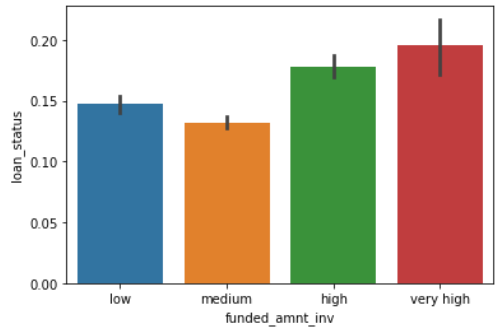

loan_status
0.30
0.25
0.20
0.15
0.10
0.05
0.00
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
purpose

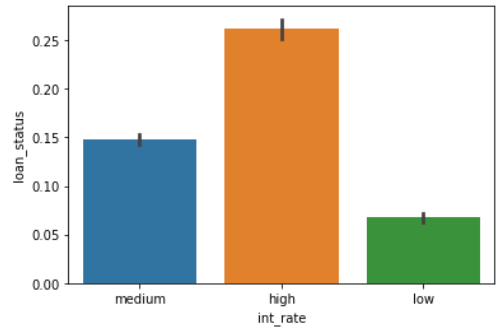
Continuous Variables (Binning required)

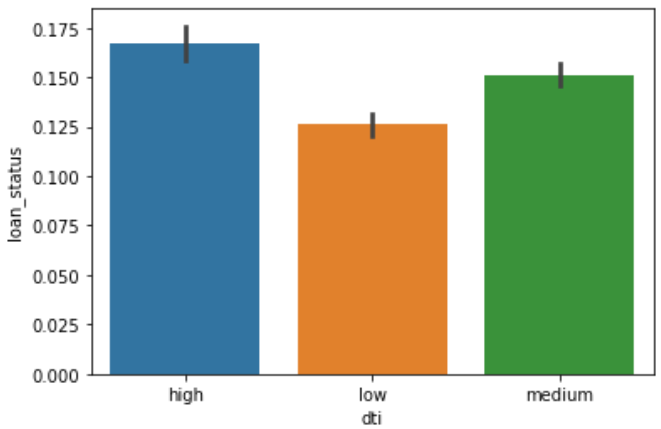
Key Findings

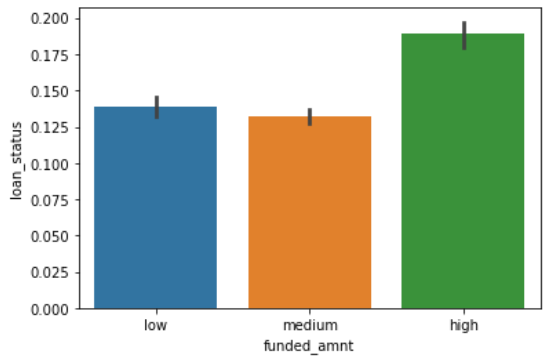
- Default rate increases as loan amount increases
- Default rate increases as funded amount investment increases
- Default rate increases as interest rate increases
- Higher dti (debt to interest rate) corresponds to higher interest rate
- higher funded amount corresponds to higher default risk rate
- higer installments corresponds to higher default rate
- low annual income corresponds to higher default rate

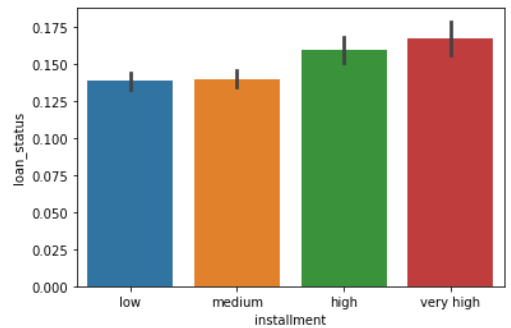

loan_status
0.00
0.05
0.10
0.15
0.20
medium
low
high
very high
loan_amnt

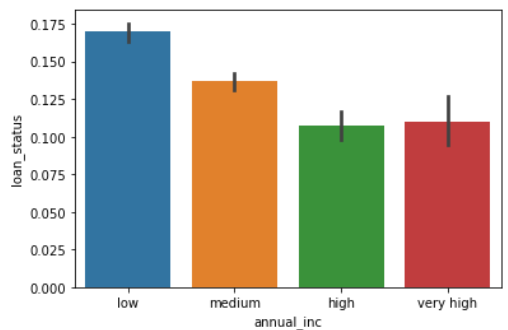

loan_status
0.00
0.05
0.10
0.15
0.20
low
medium
high
very high
funded_amnt_inv


loan_status
0.00
0.05
0.10
0.15
0.20
0.25
medium
high
low
int_rate


loan_status
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
high
low
medium
dti

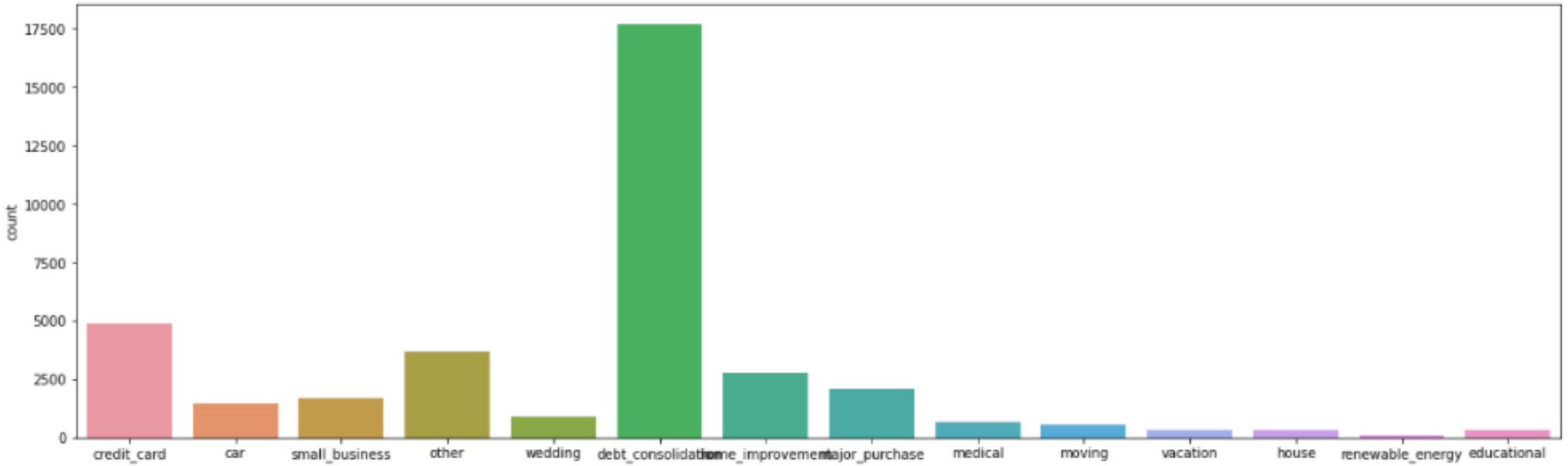

loan_status
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
low
medium
high
funded_amnt

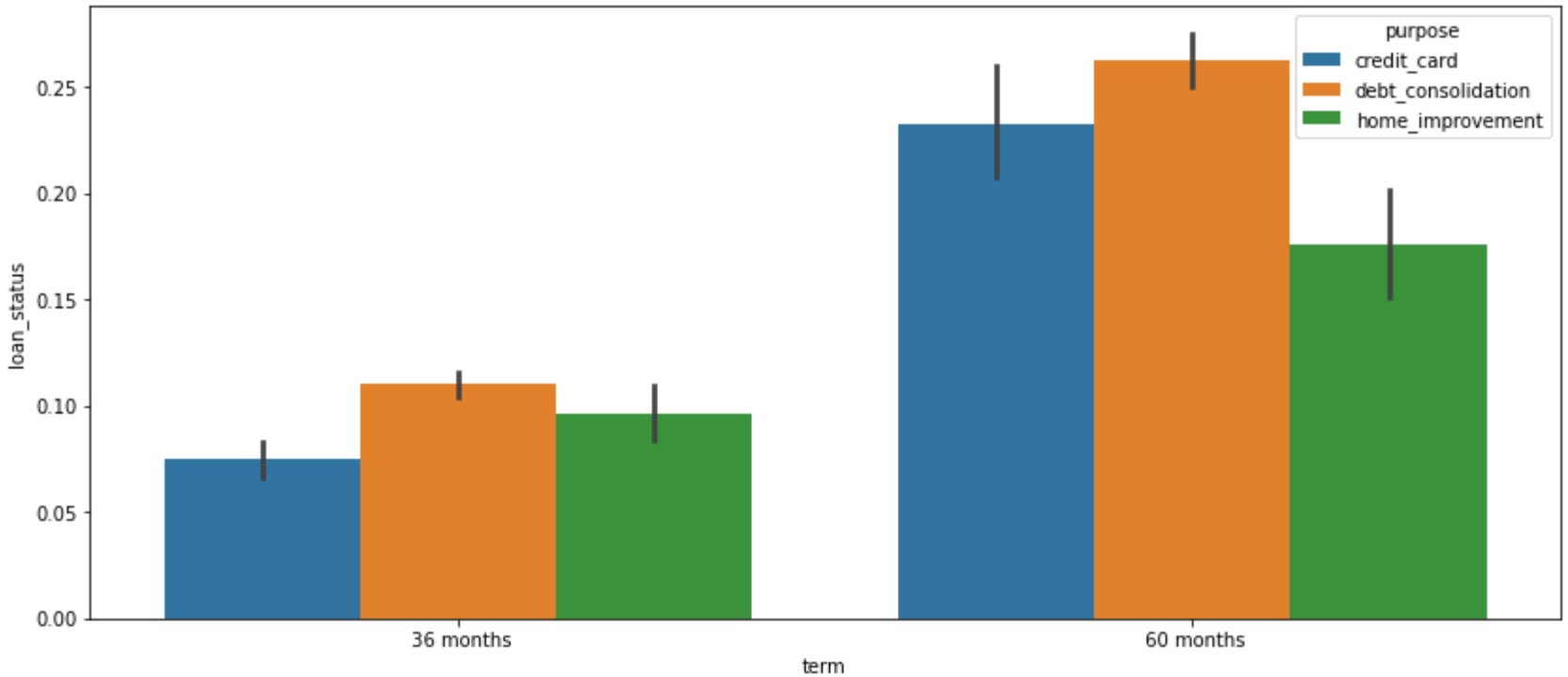

loan_status
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
low
medium
high
very high
installment


loan_status
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
low
medium
high
very high
annual_inc

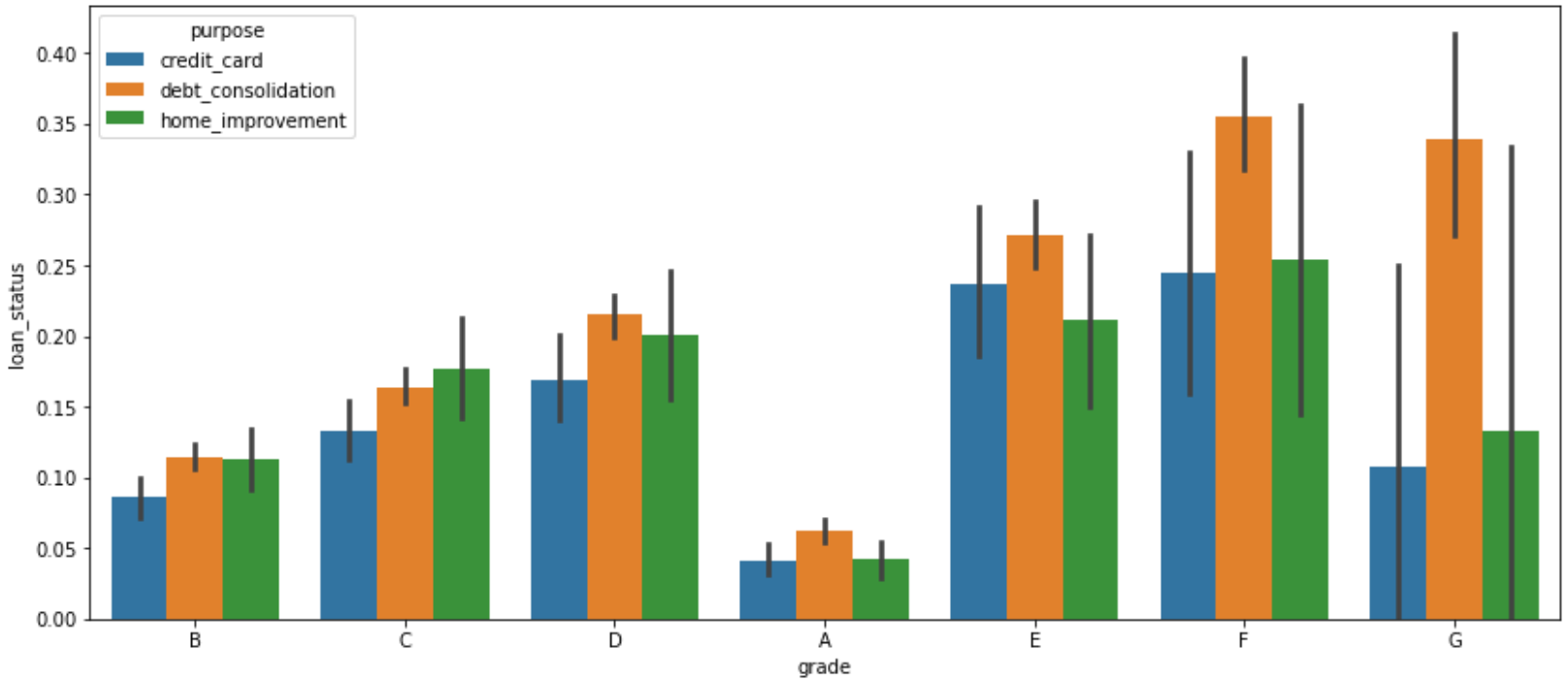
Segmented Univariate Analysis (w.r.t. Purpose of loan)

One of the most important factors affecting default is the purpose of the loan – since each purpose behave in a very different way then other., debt consolidation, credit card, home improvement are the three main purposes for which loan has been taken hence these are plotted against various variables.





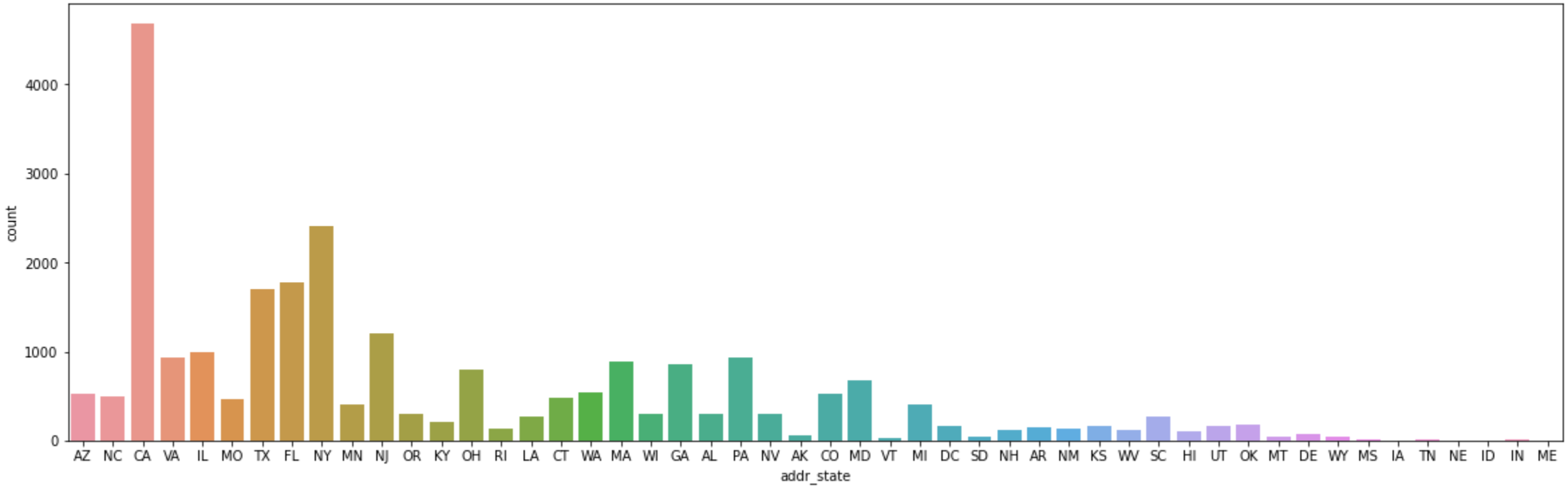
Default rate is higher for 60 month term for all three purposes



Same analysis is done for following variables:

- Home ownership
- Year
- Month
- Employment length
- Loan Amount
- Annual Income
- Interest rate
- Debt to interest rate (dti)
- Installment

Segmented Univariate Analysis (w.r.t. Address of loan)

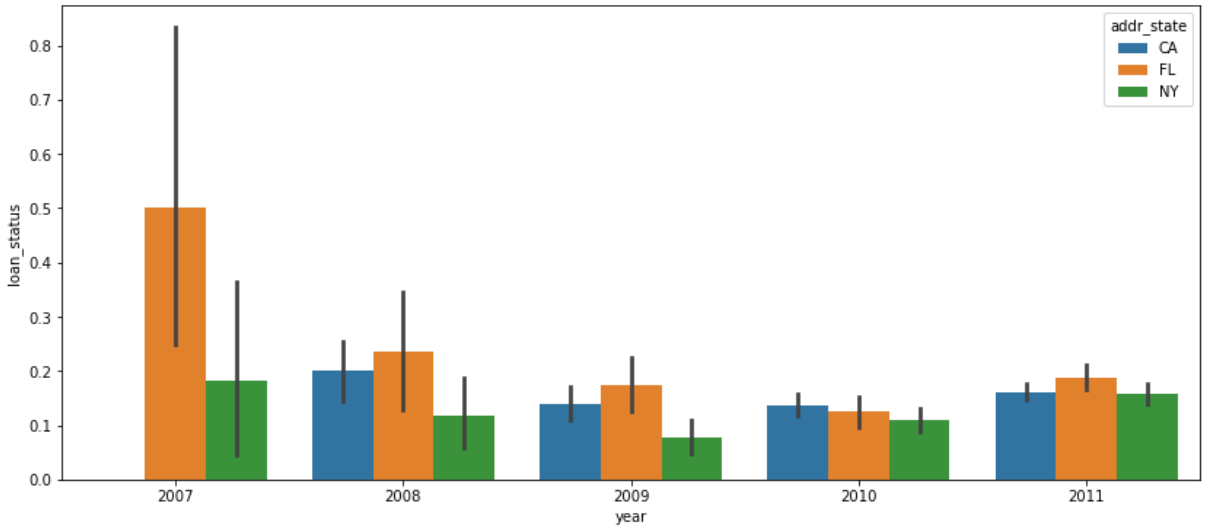


Loan rate is extremely high in the states of CA, NY and FL, hence extracting out these three states and plotting the same variables against these three states

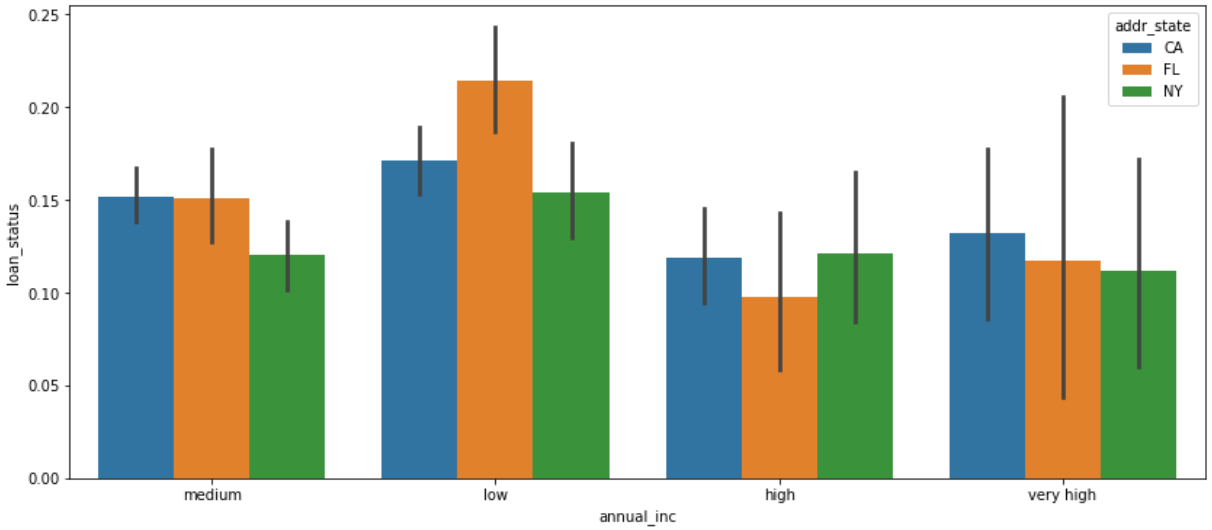
Analysis is done on following variables

- Term
- grade
- Home ownership
- Year
- Month
- Employment length
- Loan Amount
- Annual Income
- Interest rate
- Debt to interest rate (dti)
- Installment

Some of the key findings



- For the year 2007 FL had extremely high default rate



- FL has extremely high default rate for low annual income group

Key Findings

- There is 6% increase in default rate if we go from higher to lower income group
- There is 7% increase in default rate if we go from lower to higher funded amount investment
- There is 6% increase in default rate if we go from low to high debt to interest rate
- There is 3% increase in default rate if we go from low to high installment
- There is 19% increase in default rate if we go from low to high interest rate
- There is 6% increase in default rate if we go from higher to lower annual income