Problem

For the financial institution, the **driver variables** from the dataset needs to be identified that will indicate loan default possibilities for a new application.

Analysis Approach

1. Sanitize the dataset and identify the variables that have no data or only one data sample in all rows. Discard these variables as they provide no insight in the ongoing analysis
2. The remaining variables can be classified as:
    a. Transactional
    b. Profile information

    We need to use the Profile variables for analysis and trend identification

3. For any variable, the ratio of defaulters to Non-defaulters are taken for analysis. This is done to give an insight into the relative percentage of defaulters instead of absolute numbers which is not the correct indicator always due to the variation of sample sizes of number of loan applications for each variable.
4. In all analysis we have ensured that only factors with total number of loan applications more than 30 are considered to ensure that the indications are a normal approximation of the population
5. The mean of defaulter ratio is calculated for each variable across all its subcategories. This ratio gives an indication of the defaulter indication strength of the particular variable compared to other Driver variables. [Used in Inference Section]

Observations

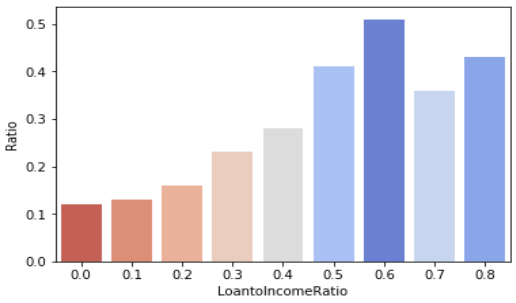
Profiling variables identified for analysis –

1. LOAN_AMNT
2. VERIFICATION_STATUS
3. TERM
4. PURPOSE
5. INSTALLMENT
6. ADDR_STATE
7. GRADE
8. DTI
9. EMP_LENGTH (no trend)
10. DELINQ_2YRS
11. HOME_OWNERSHIP
12. EARLIEST_CR_LINE (no trend)
13. ANNUAL_INC
14. OPEN_ACC
15. PUB_REC

Loan Amount

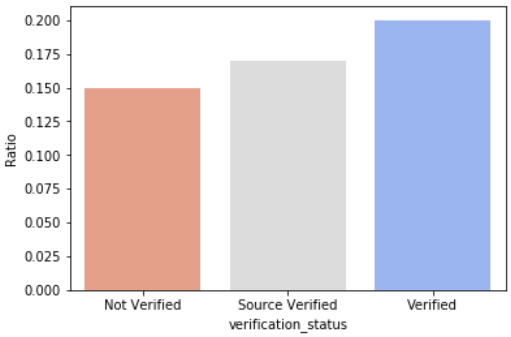
Loan amount to Annual income was used for analysis.

Loan amount up to 30% of annual income has comparatively low risk of default; loan amount 30 - 40 % has moderate range of risk; above 40 % has high risk



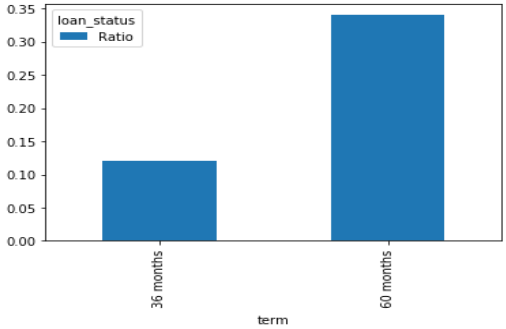
VERIFICATION_STATUS

Percentage of defaulters with Source verification completed is significantly higher than defaulters with source verification. This is an unexpected insight.



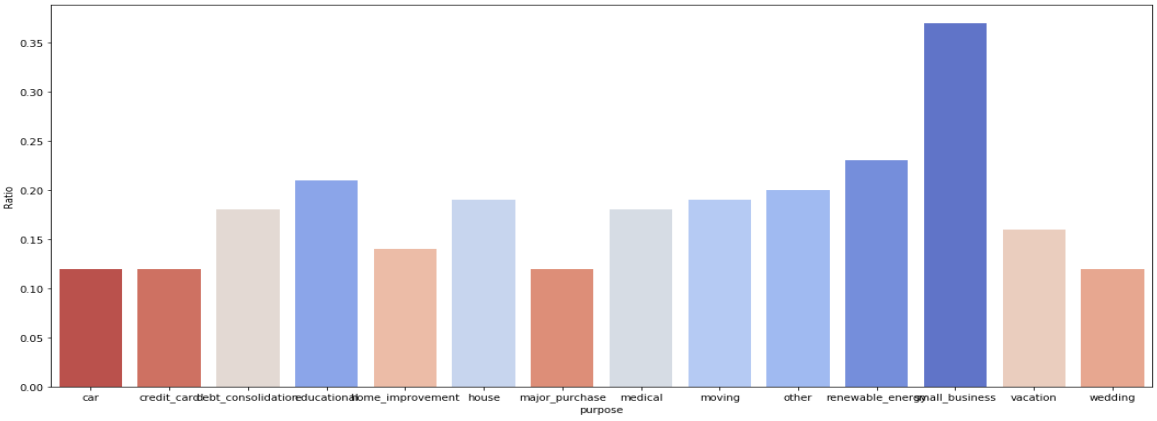
TERM

60 months is more risky



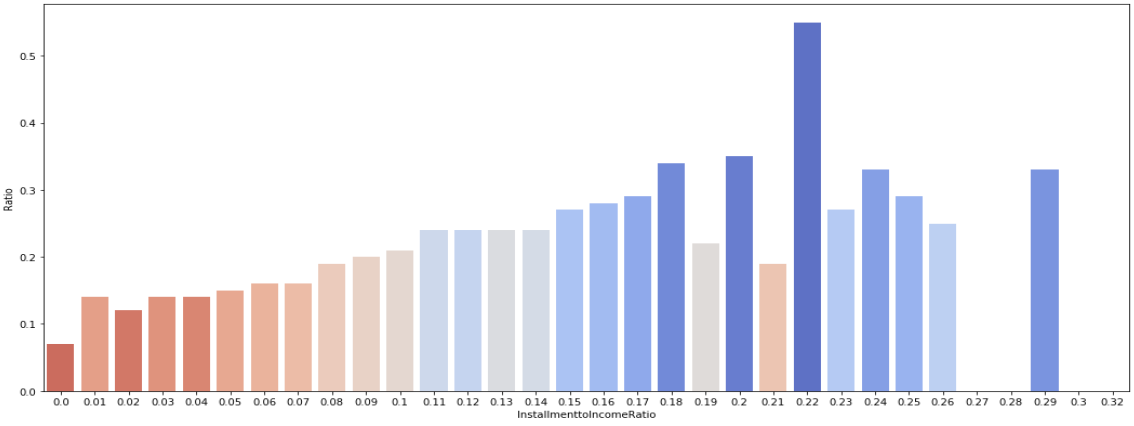
PURPOSE

car/credit_card/major_purchase/wedding/improvement/vacation - low risk;
small business/renewable energy/other/education/ - high risk;
consolidation/moving/house/medical - moderate risk



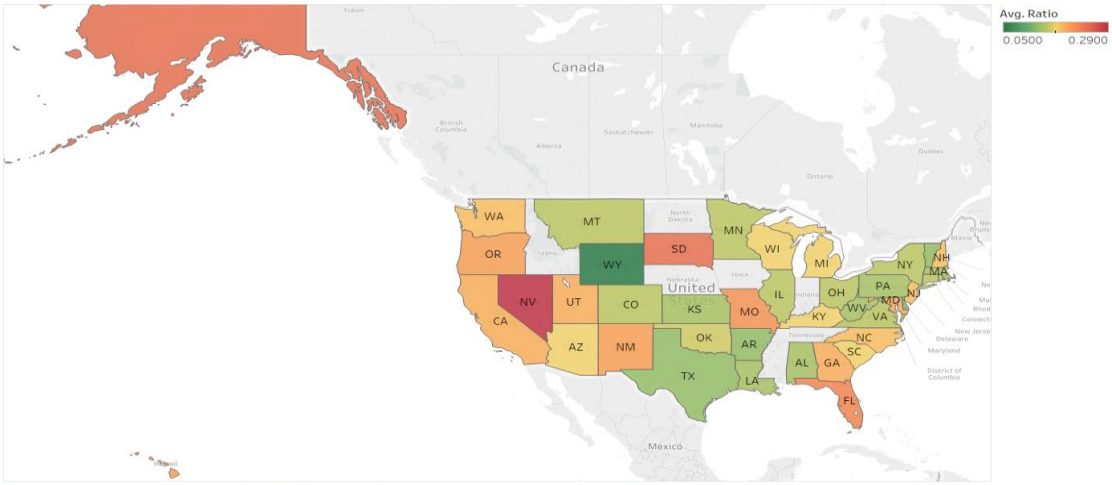
INSTALLMENT

Installment to Annual Income used for analysis
installment amount up to 10% of annual income has comparatively low risk of default; 10 - 15 % has moderate range of risk; above 15 % has high risk of default



ADDR_STATE

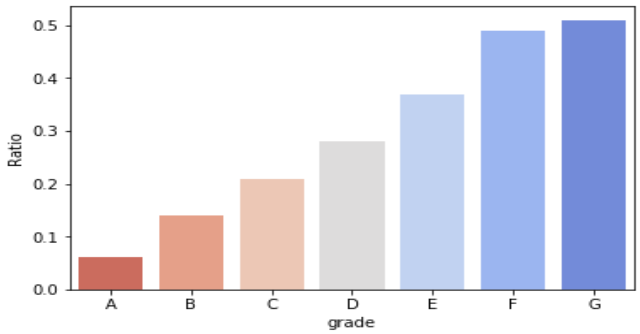
Analyzing the data we find that the eastern states are consistently showing high percent of defaulters; the south western states are showing moderate risk. The below map shows the average performance of the states. We have chosen only those states that has at least 30 samples to analyze



Map based on Longitude (generated) and Latitude (generated). Color shows average of Ratio. The marks are labeled by State.

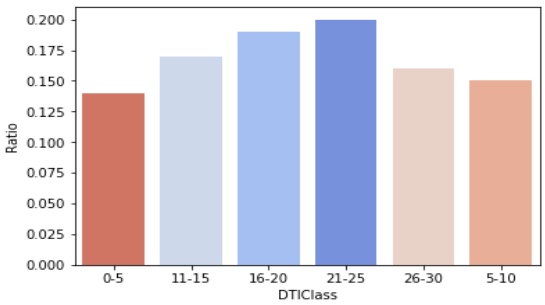
GRADE

Applicant having Loan grade A-C are low risk, D- moderate risk, E-G - high risk



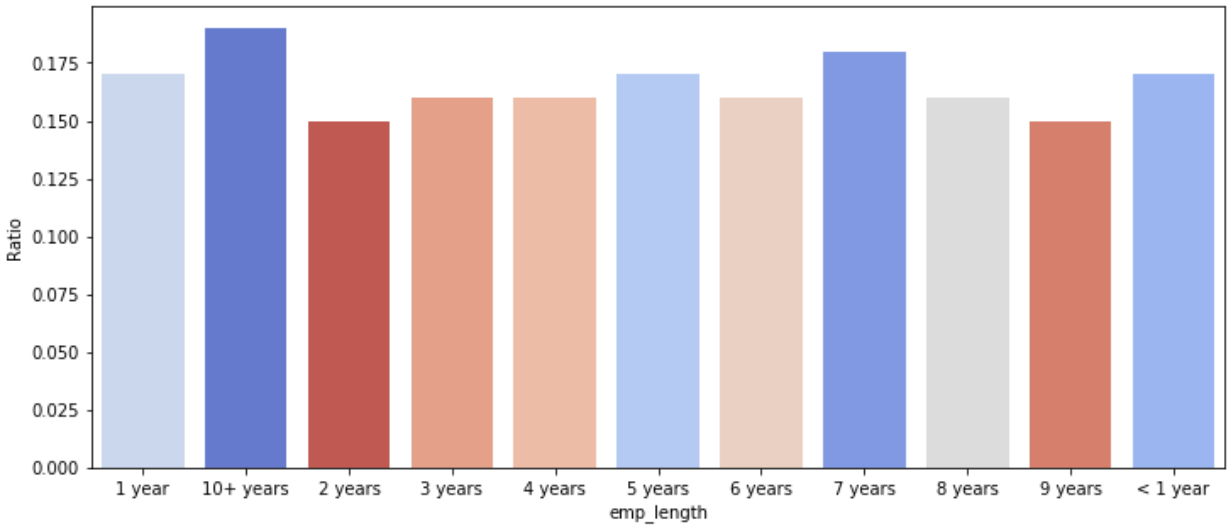
DTI

16- 25 has high risk; 11-15 & 26-30 has moderate risk and 0-10 has low risk



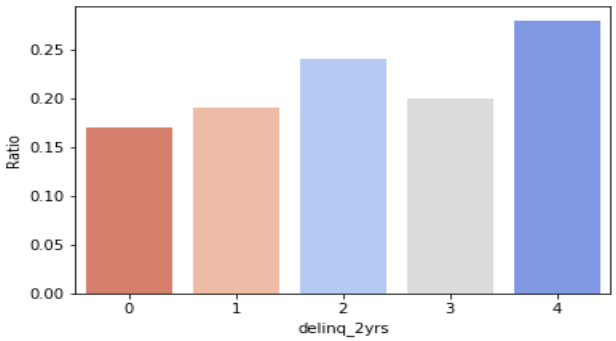
EMP_LENGTH (no trend)

No significant trend available across all Employment Length categories



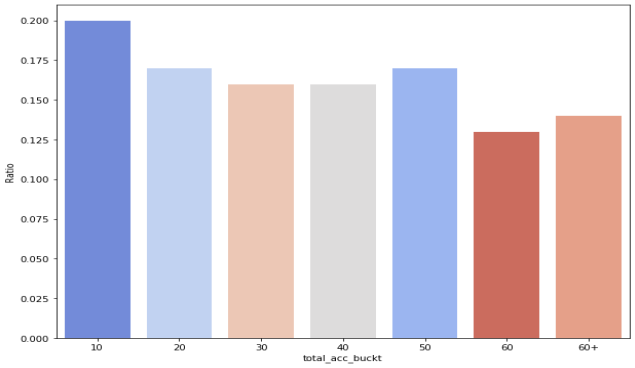
DELINQ_2YRS

delinquency of less than 2 times is comparatively less risky but >=2 times has higher defaulter rate



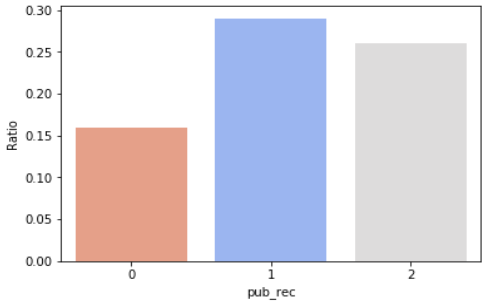
OPEN Credit ACCOUNTS

Up to 10 credit line accounts - high risk; >10 and <50 - moderate risk; > 50 - low risk



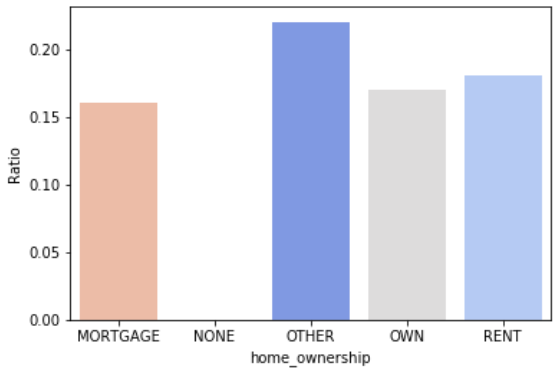
NUMBER OF PUBLIC RECORDS

Applicants with no public records is low risk where as the defaulter ratio spikes for applicants with public records.



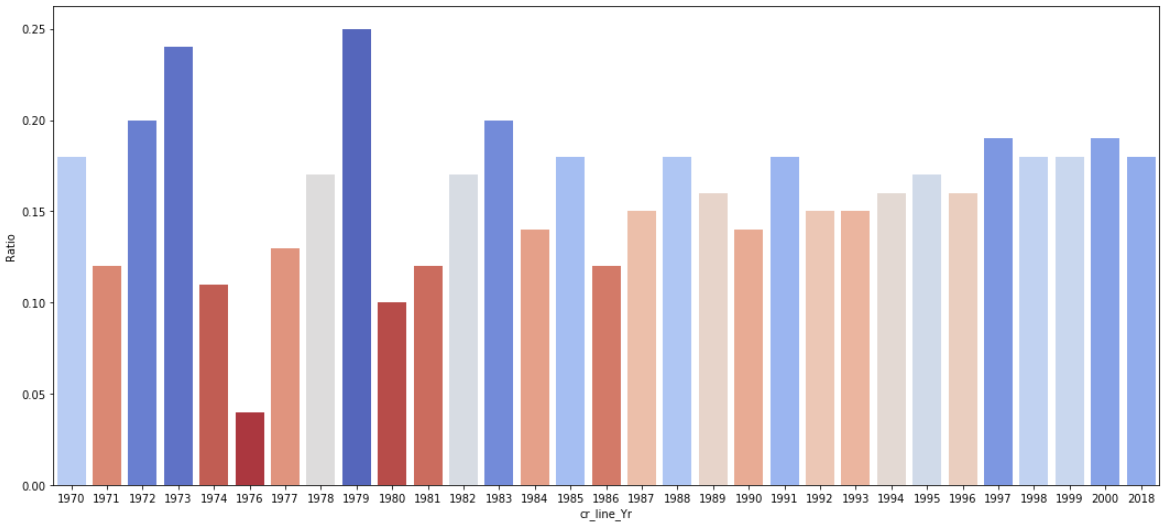
HOME_OWNERSHIP

Within all applicants who had mentioned their home ownership types, applicants with type 'OTHER' has high risk of defaulting. Applicants with defined Home ownership have moderate risk



EARLIEST_CREDIT _LINE (no trend)

The provided data do not give any substantial trend indications for default indication



Inference

From the above we have the following list of Driver variables and the mean defaulter ratio for each of them across all of their individual sub categories:

Driver Variables	Defaulting indicator value
GRADE	0.3
LOAN_AMNT	0.29
PUB_REC	0.24
TERM	0.23
DELINQ_2YRS	0.22
INSTALLMENT	0.21
VERIFICATION_STATUS	0.18
PURPOSE	0.18
ADDR_STATE	0.18
DTI	0.17
OPEN_ACC	0.16
HOME_OWNERSHIP	0.15