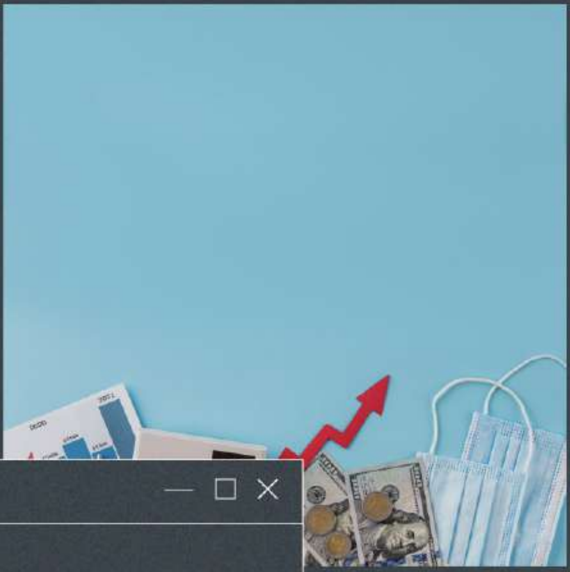
Credit Assessment for Loan Approval

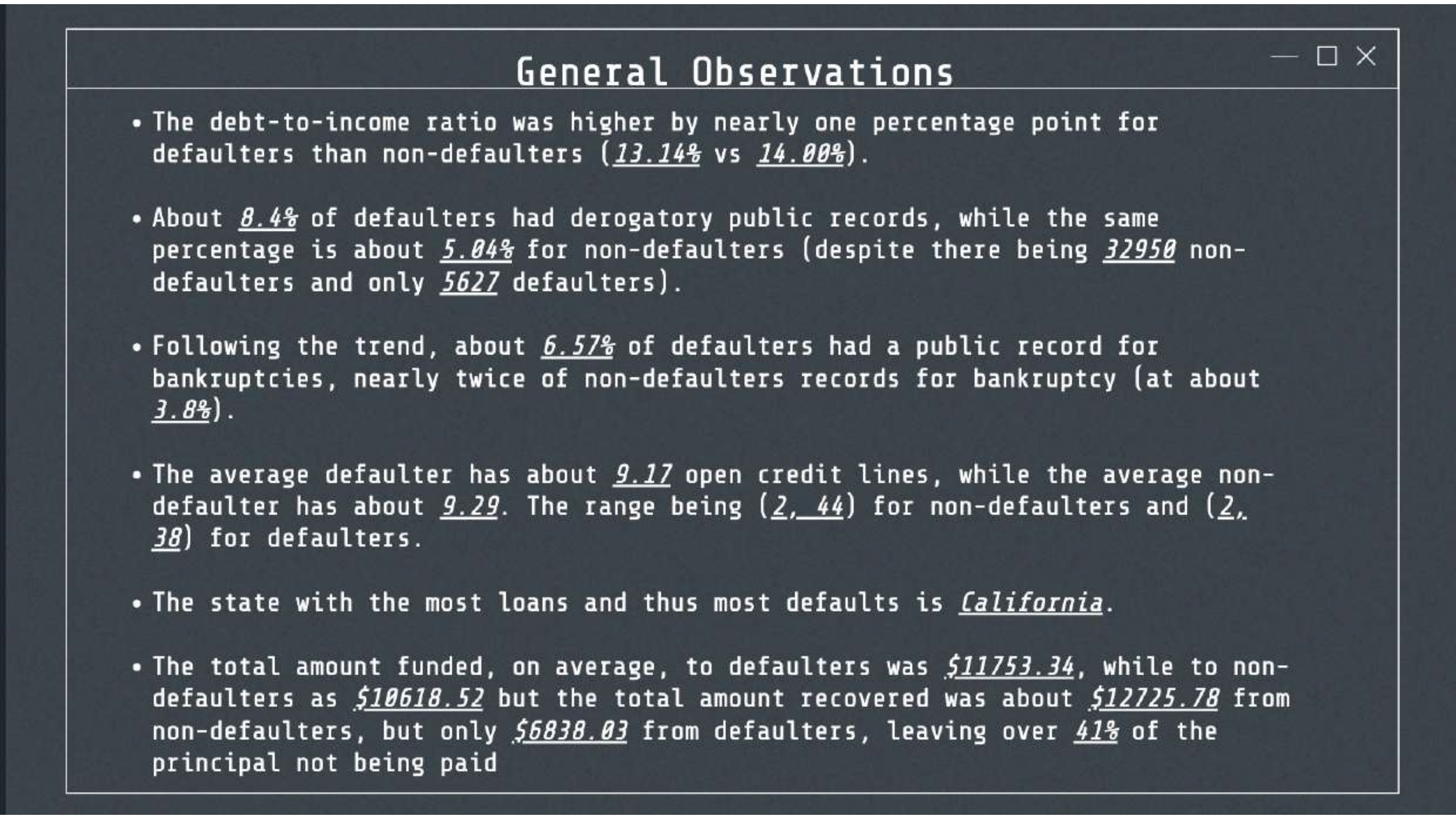
Credit Assessment

The goal of analysing the data was to find certain underlying patterns that might emerge between defaulters and non-defaulters, and to recognise potential indicators for the said patterns.



EDA-Approach

The main approach was to start, apart from data cleaning and engineering, was to get simple statistics to have a rudimentary understanding of the dataset (general observations). The next step was to analyse potential indicators/variables and find any underlying patterns in them.



General Observations

- The debt-to-income ratio was higher by nearly one percentage point for defaulters than non-defaulters (*13.14%* vs *14.00%*).

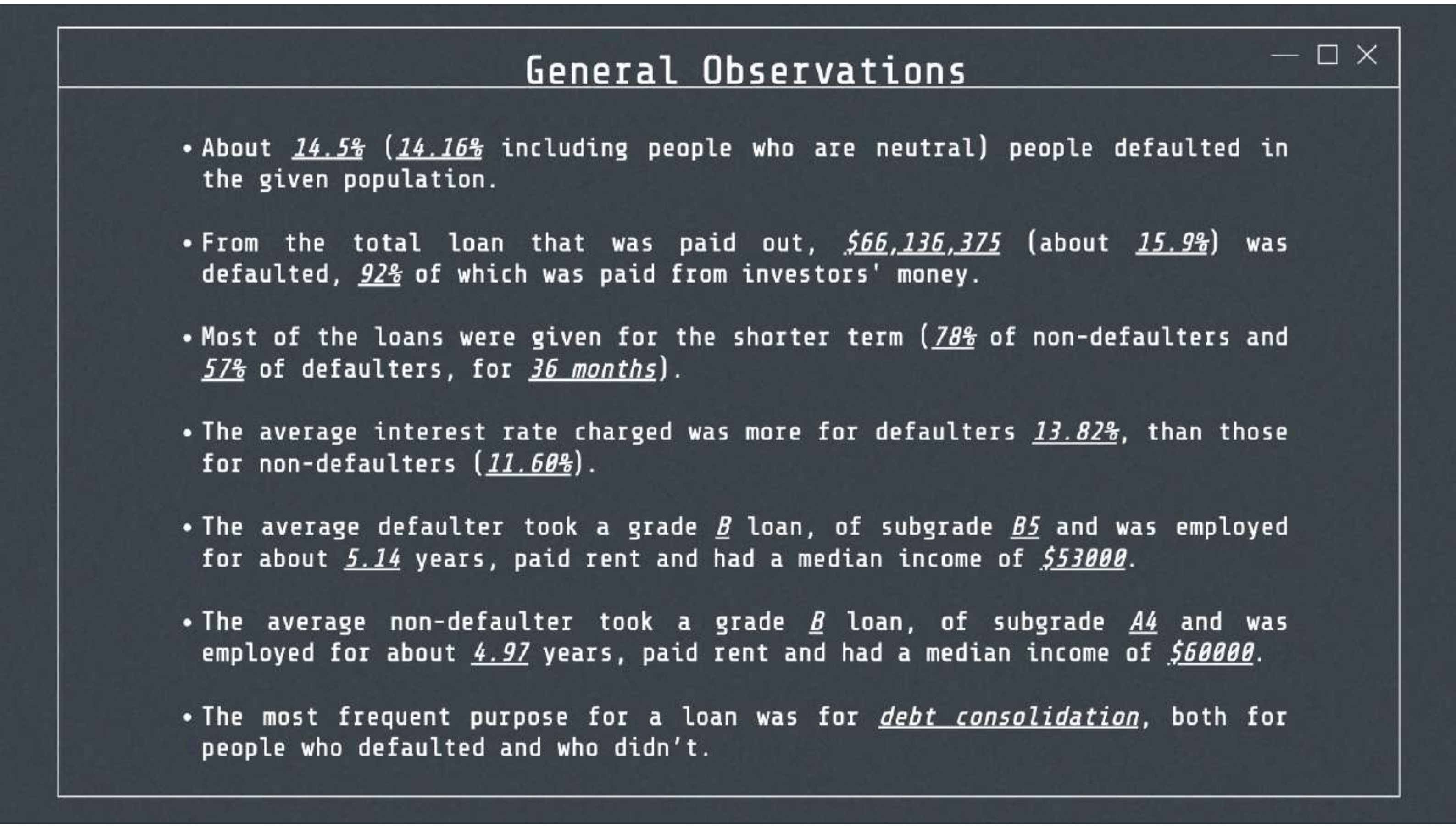
- About *8.4%* of defaulters had derogatory public records, while the same percentage is about *5.04%* for non-defaulters (despite there being *32950* non-defaulters and only *5627* defaulters).

- Following the trend, about *6.57%* of defaulters had a public record for bankruptcies, nearly twice of non-defaulters records for bankruptcy (at about *3.8%*).

- The average defaulter has about *9.17* open credit lines, while the average non-defaulter has about *9.29*. The range being (*2, 44*) for non-defaulters and (*2, 38*) for defaulters.

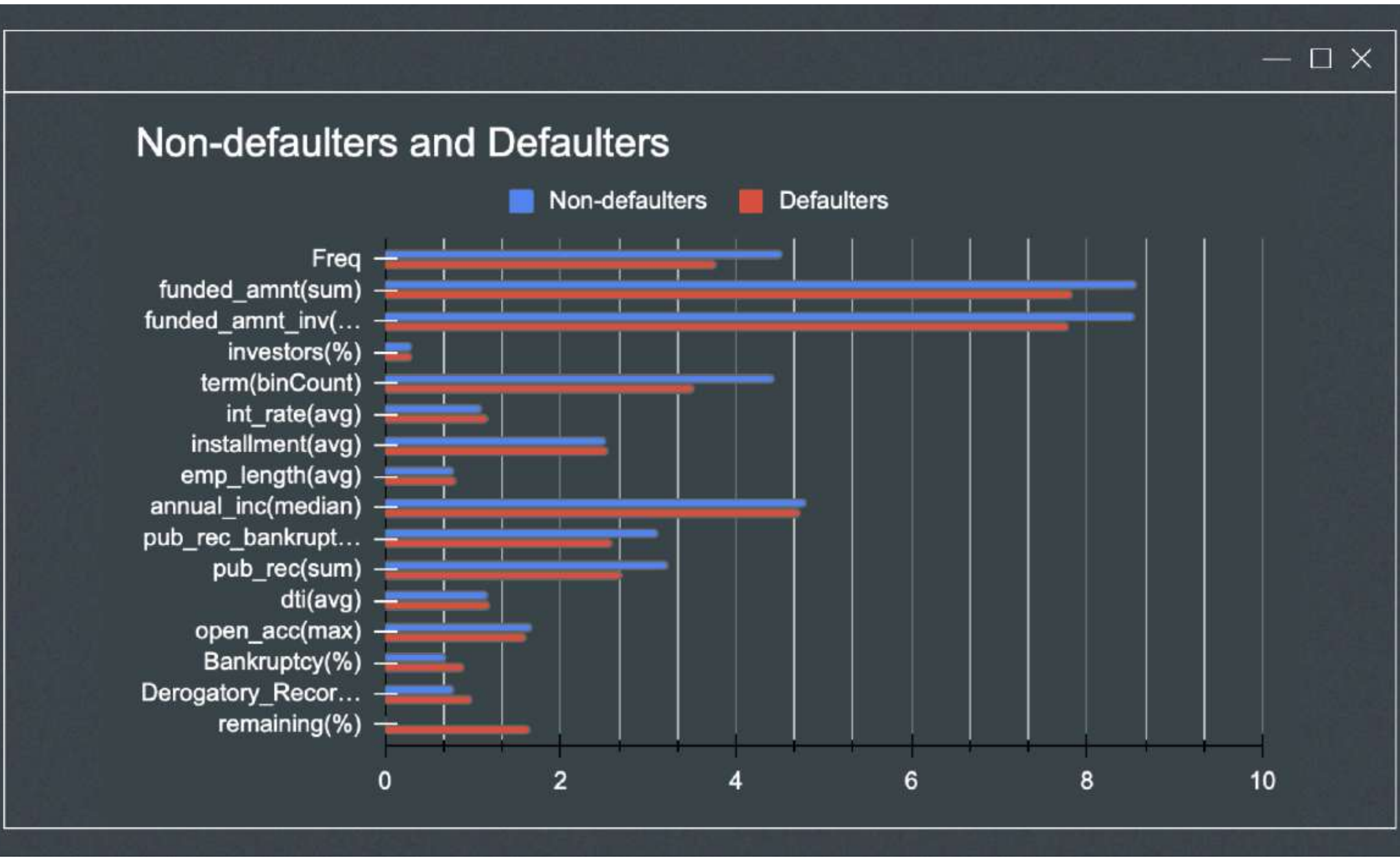
- The state with the most loans and thus most defaults is *California*.

- The total amount funded, on average, to defaulters was *$11753.34*, while to non-defaulters as *$10618.52* but the total amount recovered was about *$12725.78* from non-defaulters, but only *$6838.03* from defaulters, leaving over *41%* of the principal not being paid



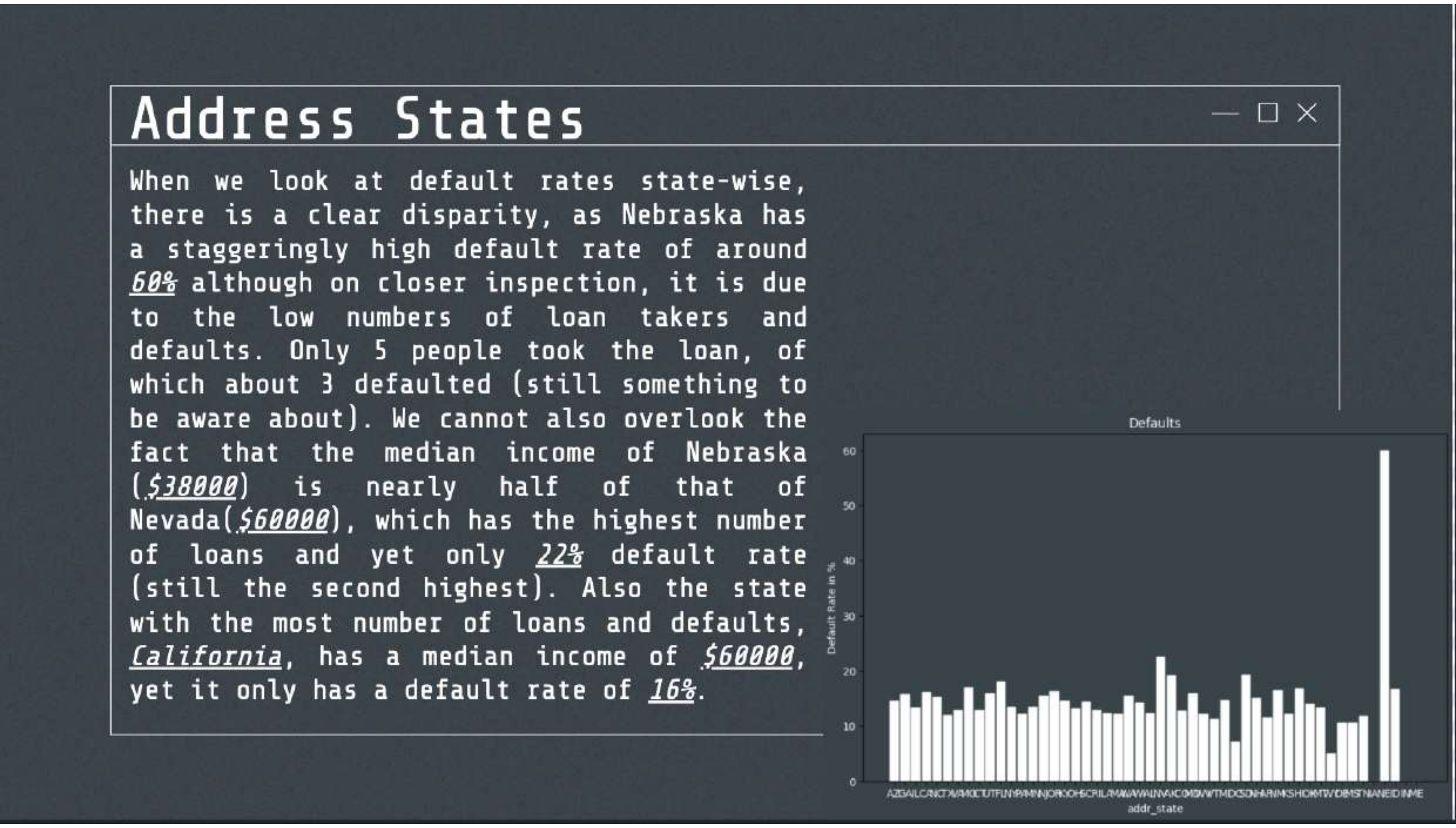
General Observations

- About *14.5%* (*14.16%* including people who are neutral) people defaulted in the given population.
- From the total loan that was paid out, *$66,136,375* (about *15.9%*) was defaulted, *92%* of which was paid from investors' money.
- Most of the loans were given for the shorter term (*78%* of non-defaulters and *57%* of defaulters, for *36 months*).
- The average interest rate charged was more for defaulters *13.82%*, than those for non-defaulters (*11.60%*).
- The average defaulter took a grade *B* loan, of subgrade *B5* and was employed for about *5.14* years, paid rent and had a median income of *$53000*.
- The average non-defaulter took a grade *B* loan, of subgrade *A4* and was employed for about *4.97* years, paid rent and had a median income of *$60000*.
- The most frequent purpose for a loan was for *debt consolidation*, both for people who defaulted and who didn't.

Non-defaulters and Defaulters
🟦 Non-defaulters
🟥 Defaulters
Freq
funded_amnt(sum)
funded_amnt_inv(…
investors(%)
term(binCount)
int_rate(avg)
installment(avg)
emp_length(avg)
annual_inc(median)
pub_rec_bankrupt…
pub_rec(sum)
dti(avg)
open_acc(max)
Bankruptcy(%)
Derogatory_Recor…
remaining(%)
0
2
4
6
8
10

Default Rates

The next step is selecting around six variables that could potentially be indicators of whether an individual will default or not, and scrutinizing their default rates, i.e, what percentage of loan takers in those variables have defaulted compared to others. The variables are: Grade, Subgrade, Home Ownership, Annual Income, Purpose for taking the loan and the State they live in.



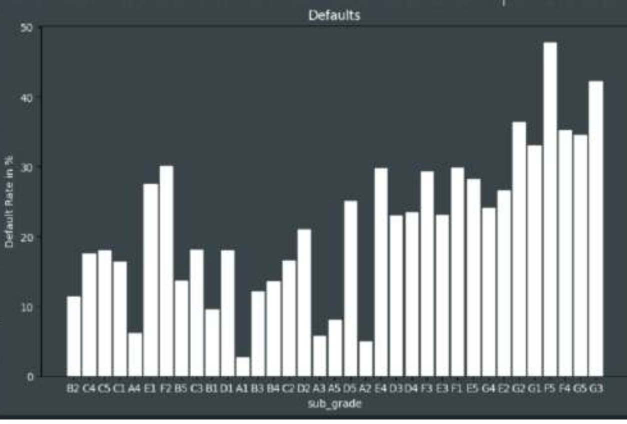
Address States

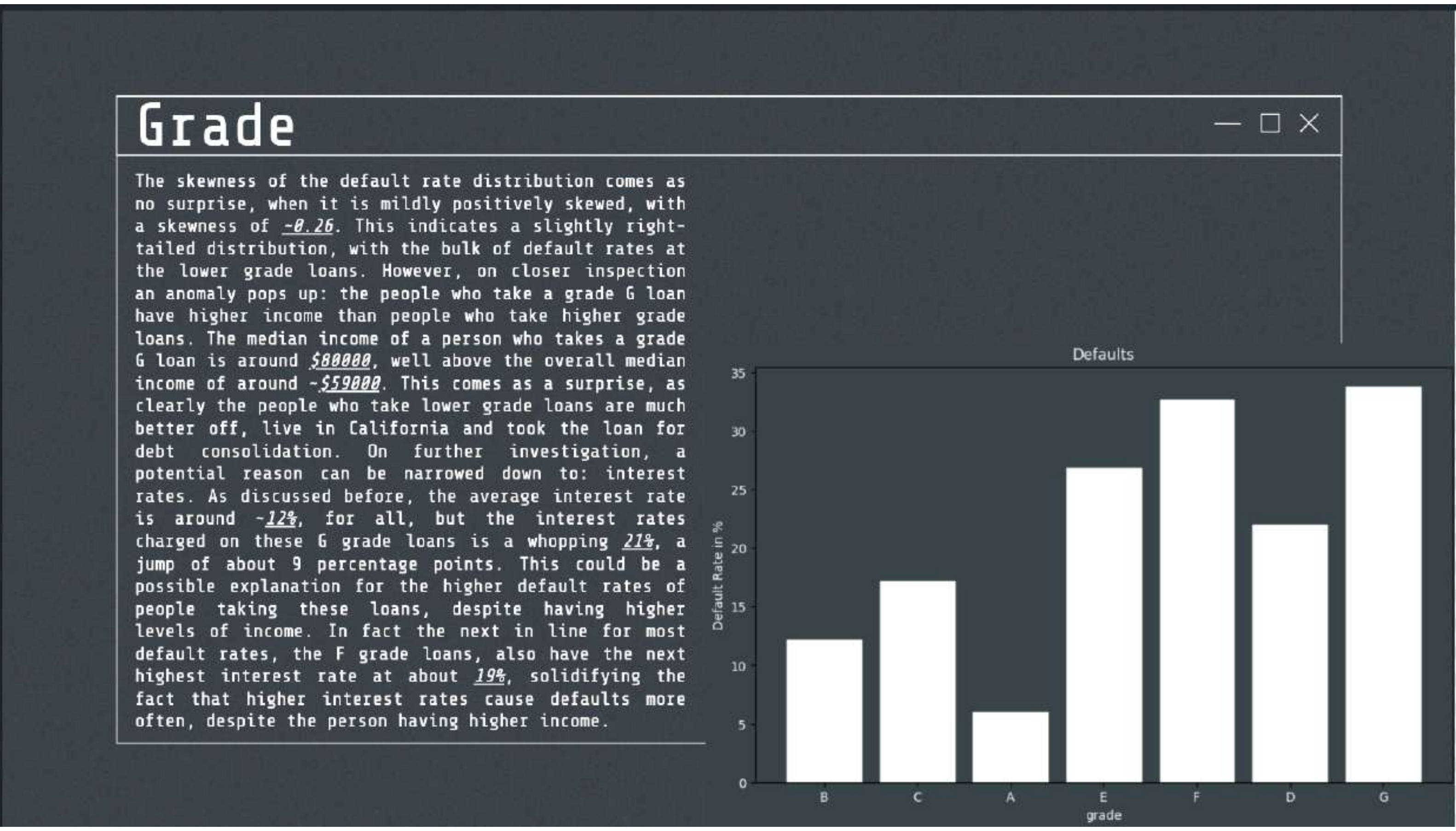
When we look at default rates state-wise, there is a clear disparity, as Nebraska has a staggeringly high default rate of around *60%* although on closer inspection, it is due to the low numbers of loan takers and defaults. Only 5 people took the loan, of which about 3 defaulted (still something to be aware about). We cannot also overlook the fact that the median income of Nebraska (*$38000*) is nearly half of that of Nevada(*$60000*), which has the highest number of loans and yet only *22%* default rate (still the second highest). Also the state with the most number of loans and defaults, *California*, has a median income of *$60000*, yet it only has a default rate of *16%*.

Defaults
Default Rate in %
60
50
40
30
20
10
0
AZGAILCANCTXVAMOCTUTFLNYPAMNNJORKYOHSCRILAMAWAWALNVAKCOMDWVWTMDCSDNHARNMKSHIOKMTWYDEMSTNIANEIDINME
addr_state

Subgrade

The same trend continues for the subgrade loans as well. The lower category loans have higher median or average annual income, while the interest rates are higher as well. This generally causes higher default rates in the lower grade loans, though not due to lower credit rating of the borrower, but maybe due to higher interest rates charged. The skewness is much like the ones for grades column: slightly positively skewed at *0.23*. The highest default rate is at *47%* for the F5 grade loans, followed by *45%* for the G3 loans.



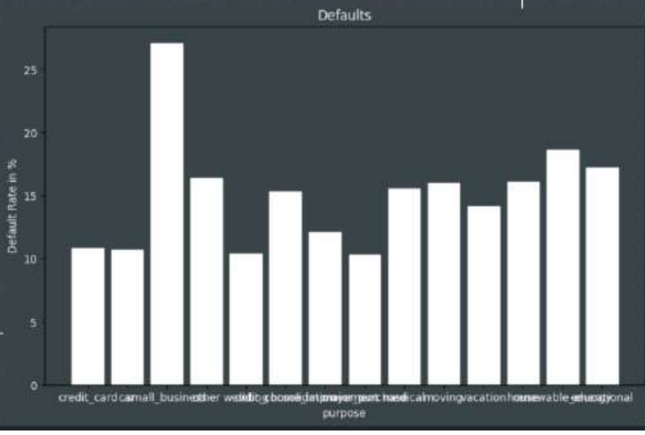


Grade

The skewness of the default rate distribution comes as no surprise, when it is mildly positively skewed, with a skewness of *~0.26*. This indicates a slightly right-tailed distribution, with the bulk of default rates at the lower grade loans. However, on closer inspection an anomaly pops up: the people who take a grade G loan have higher income than people who take higher grade loans. The median income of a person who takes a grade G loan is around *$80000*, well above the overall median income of around ~*$59000*. This comes as a surprise, as clearly the people who take lower grade loans are much better off, live in California and took the loan for debt consolidation. On further investigation, a potential reason can be narrowed down to: interest rates. As discussed before, the average interest rate is around ~*12%*, for all, but the interest rates charged on these G grade loans is a whopping *21%*, a jump of about 9 percentage points. This could be a possible explanation for the higher default rates of people taking these loans, despite having higher levels of income. In fact the next in line for most default rates, the F grade loans, also have the next highest interest rate at about *19%*, solidifying the fact that higher interest rates cause defaults more often, despite the person having higher income.

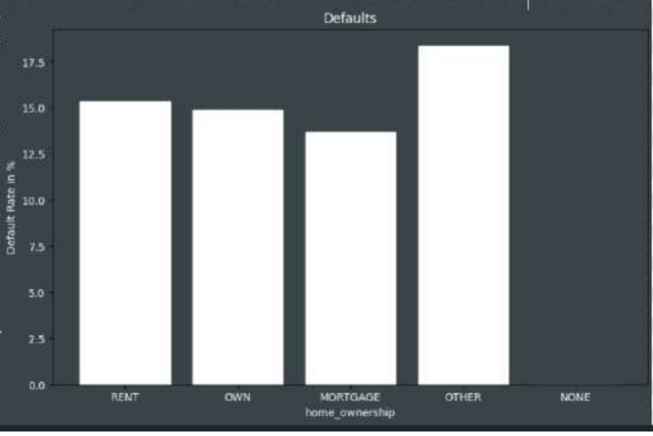
Purpose

The default rates according to the purpose for taking a loan are fairly equally distributed, until we see the staggering *27.08%* of default rates of people who took out loans for their small businesses. It follows that risks taken by the people in opening a small business may simply be transferred to the lending company, hence such high default rates among small business owners. The next highest rates are those of renewable energy, which are still 9 percentage points below the small business people



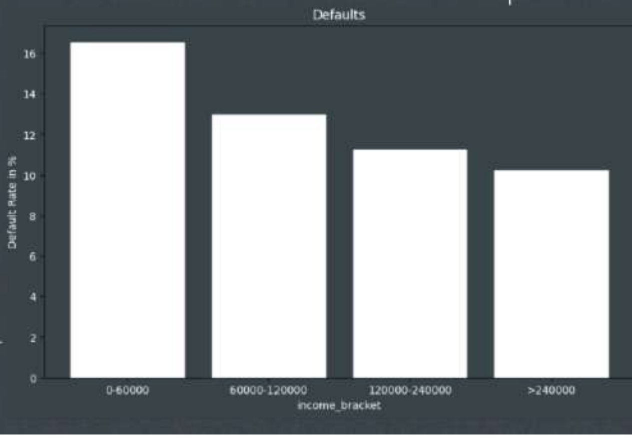
Home Ownership

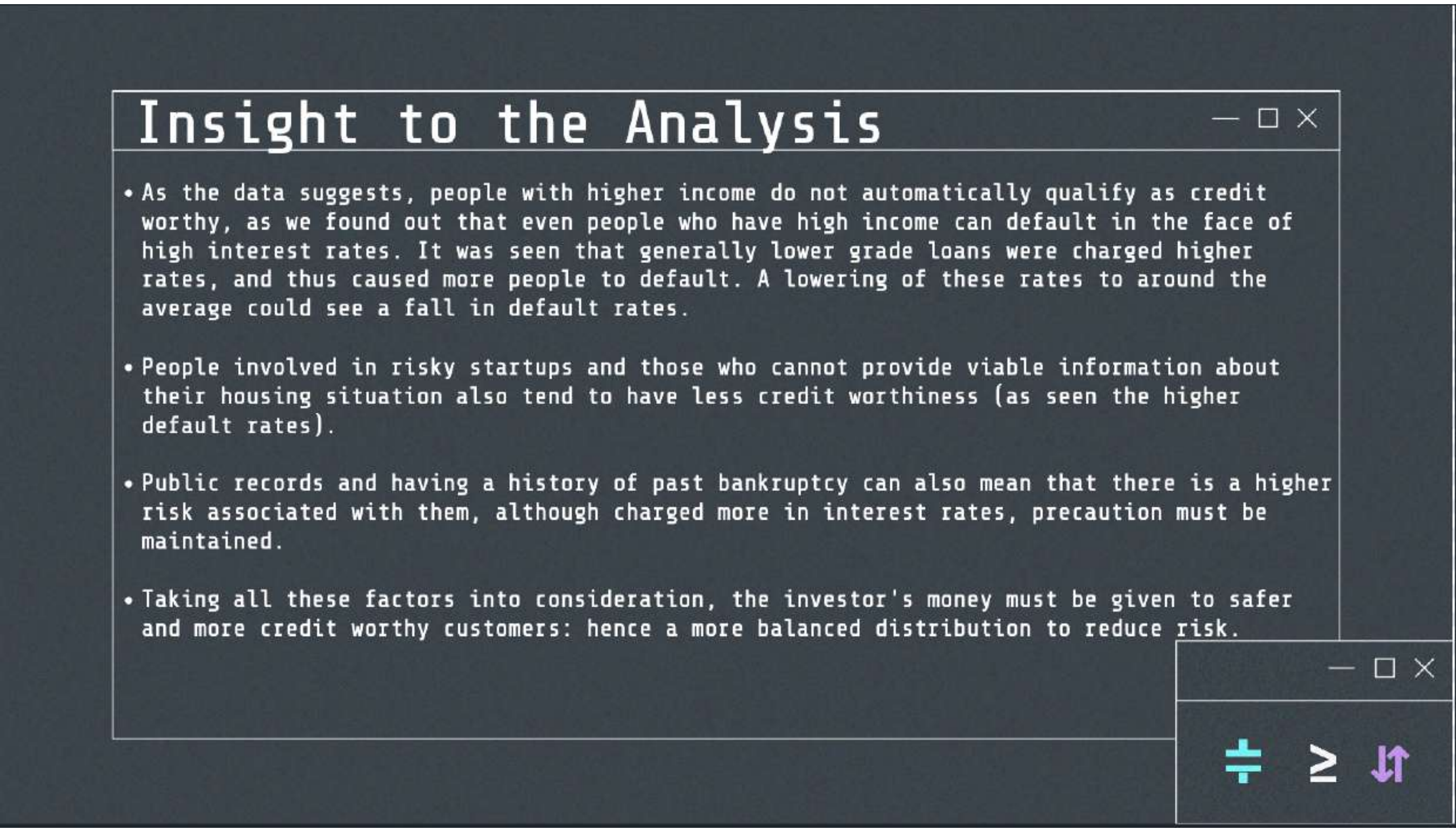
Although the total number of loans is significantly higher for those who have rents and mortgages, it is the 'other' category that has a higher default rate. This indicates that the category of 'other' home owners are generally credit risks, as they neither own, rent nor have a mortgage, implying some other means of staying with a roof over their head, which ultimately proves to be unreliable. The default rate is around 18%, which is around 3 percentage points higher than rent or mortgage, despite both the number of loan-takers renting a home or having a mortgage are significantly higher



Income Bracket

As expected, in the case of income brackets, we get a smooth distribution of default rates which is skewed to the right-tail (~*1.12*). This simply implies that the lower the income of an individual, the more credit risk they are. The default rate seems to drop around an average by *14.70%* as we go to higher income brackets. The positive skewness simply indicates that there are more higher default rates at lower levels of income.





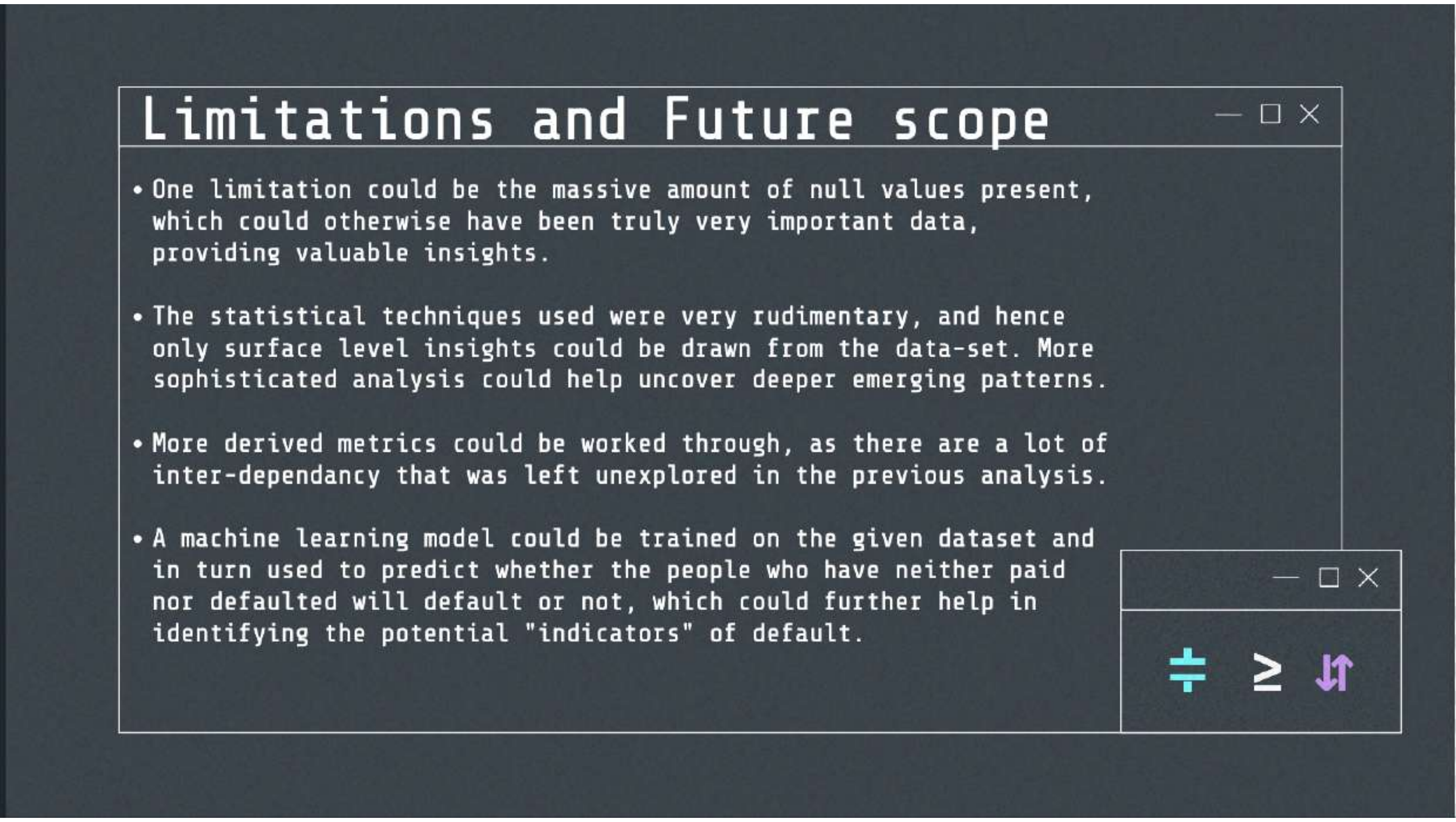
Insight to the Analysis

- As the data suggests, people with higher income do not automatically qualify as credit worthy, as we found out that even people who have high income can default in the face of high interest rates. It was seen that generally lower grade loans were charged higher rates, and thus caused more people to default. A lowering of these rates to around the average could see a fall in default rates.

- People involved in risky startups and those who cannot provide viable information about their housing situation also tend to have less credit worthiness (as seen the higher default rates).

- Public records and having a history of past bankruptcy can also mean that there is a higher risk associated with them, although charged more in interest rates, precaution must be maintained.

- Taking all these factors into consideration, the investor's money must be given to safer and more credit worthy customers: hence a more balanced distribution to reduce risk.



Limitations and Future scope

- One limitation could be the massive amount of null values present, which could otherwise have been truly very important data, providing valuable insights.
- The statistical techniques used were very rudimentary, and hence only surface level insights could be drawn from the data-set. More sophisticated analysis could help uncover deeper emerging patterns.
- More derived metrics could be worked through, as there are a lot of inter-dependancy that was left unexplored in the previous analysis.
- A machine learning model could be trained on the given dataset and in turn used to predict whether the people who have neither paid nor defaulted will default or not, which could further help in identifying the potential "indicators" of default.

Thanks!

by
Astle Dsa