Analyzing Climate Hazard Risk to a Loan Portfolio

by Hrolfur Sveinsson, Peijia Wu, Yixuan Yang

Business Analytics Practicum Final Report
George Washington University
School of Business

December 10, 2021

Table of Contents

1. Executive Summary

In this report, we have analyzed the feasibility of incorporating climate hazard data into traditional credit portfolio analysis to improve the accuracy of the credit portfolio analysis. The U.S. Small Business Administration's (SBA) 7(a) loan program portfolio is used as loan data, and the Federal Emergency Management Agency's (FEMA) National Risk Index for Natural Hazards is used as climate hazard data.

We have answered how much of the SBA's portfolio is currently exposed to the physical risks of climate hazards and shown the differences in risk across the different portfolio segments. We have quantified the impact of actual climate events on the portfolio, integrated climate hazard data with loan data, and utilized multiple machine learning techniques to determine the relationship between climate hazard risk and credit risk. Initial analysis suggests no evident relationship between climate hazard risk and credit risk in the form of interest rate. However, analysis shows that a significant part of the SBA's portfolio is exposed to climate hazard risk.

With better data becoming available each year, there is a chance for lenders to start incorporating this additional information of risk due to climate hazards into their risk management strategies. Higher climate hazard risk areas should, on average, be charged a higher interest rate.

2. Introduction

2.1 Problem Understanding

Business Problems

Based on the project's goal stated by the client, we have determined the business problems needed to be addressed as follows:

- Have historical climate hazards had a significant impact on loan risk? If yes, to what extent?
- What specific hazards have had a significant impact on loan risk?
- What is the loan portfolio's potential future exposure to climate hazards?

Business Objectives

We defined the objectives of our project based on the business problems defined above:

- Integrate climate hazard risk into loan risk assessment
- Measure climate hazards impact on loan risk
- Determine the portfolio's current exposure to climate hazards
- Project the portfolio's future exposure to climate hazards and potential climate scenarios, along with the impact of these scenarios on losses

Solutions & Steps

Since we already defined the business problems and objectives, clear solutions and action steps are needed to answer the problems and meet the objectives:

- Determine which datasets and what part of proposed datasets are needed
- Determine metrics to measure loan portfolio performance
- Integrate climate hazard dataset with loan risk dataset
- Determine the portfolio's exposure to climate hazards
- Explore the relationship between loan portfolio and climate risk

2.2 Background

ESG Research Field: Transition Risk & Physical Risk

Nowadays, banks and financial institutions are trying to assess the influence on financial risks from environmental, social, and governance (ESG) factors, especially how climate change and extreme weather conditions can affect their risk management processes, business strategies, and investment policies.

There are two major categories of financial risk derived from climate-related risk: Transition risk and physical risk. Transition risk is associated with the transition to a lower-carbon economy, which may entail extensive policy, legal, technology, and market changes. Physical risk is associated with the physical impacts on climate change, driven by extreme weather events such as hurricanes and floods, as well as chronic long-term shifts such as temperature increase and sea-level rise *(BofA-task-force-climate)*.

Why Focus on Physical Risk?

Physical risks are composed of two categories: acute hazards caused by extreme climate events such as droughts, floods, storms, and chronic hazards which arise from progressive shifts in climate patterns such as increasing temperatures, sea-level rise, and changes in precipitation *(Integrating Climate Risks into Credit Risk Assessment)*. Based on the climate data acquired from FEMA's National Risk Index on 18 types of natural hazards, the risk score for each hazard, and their expected annual loss, it can be said that the data is more representative of the acute hazard side of physical risk. There is no data on changes in temperatures, sea-level rise, precipitation, or carbon emissions. Therefore, this report is less focused on the chronic hazard side of physical and transition risk.

3. Data Understanding

3.1 Data Sources

Loan Data
SBA's 7(a) loan program:
https://data.sba.gov/dataset/7-a-504-foia

Climate Hazard Data
FEMA's National Risk Index for Natural Hazards:
https://hazards.fema.gov/nri/data-resources

3.2 Data Analyzed

Loan Data
There are two types of loan data available from the SBA: 7(a) & 504 Freedom of Information Act (FOIA) loan program data.

The FOIA data for the 7(a) and 504 programs are updated quarterly. The data is typically available one month after the quarter has ended *(7(a) & 504 FOIA)*.

Definition of 7(a) & 504 Loan Programs
7(a): The 7(a) loan program is the SBA's most common loan program. It includes financial help for small businesses with special requirements. The 7(a) loan program is the best option when real estate is part of a business purchase, but it can also be used for:

- Short and long-term working capital
- Refinance current business debt
- Purchase furniture, fixtures, and supplies

(7(a) Loans).

504: The 504 Loan Program provides long-term, fixed-rate financing of up to $5 million for major fixed assets that promote business growth and job creation. A 504 loan can be used for a range of assets that promote business growth and job creation. These include the purchase or construction of:

- Existing buildings or land
- New facilities
- Long-term machinery and equipment

(504 Loans).

Why 7(a)?

According to the definition and explanation of the two types of loans, we decided to study and research the 7(a) loan datasets instead of the 504 loan dataset. The 504 loan dataset is related to companies' fixed assets, and we do not have specific enough information about those fixed assets. Also, the 7(a) datasets have 1.6 million records, while 504 only includes 194,181 records. This makes the 7(a) loan datasets more suitable for the project's needs as we need to ensure we have enough data for modeling and visualization.

Information About the 7(a) Datasets

There are three files for the 7(a) loan program that segment the data by decade (FY1991-FY1999, FY2000-FY2009, FY2010-Present).

There are 32 fields in total for all three files. The fields relevant for our study are:

- GrossApproval: Total loan amount
- GrossChargeOffAmount: Total loan balance charged off (includes the guaranteed and the non-guaranteed portion of the loan)
- InitialInterestRate: Initial interest rate - total interest rate (base rate plus spread) at time loan was approved
- BorrState: Borrower state
- ProjectCounty: County where project occurs. The state and county information are used for the latter data integration with climate hazard data. There are no borrower counties in the loan data, but by sampling we found project counties match with borrower address shown in the loan data, so project county can be used as borrower county.
- NaicsCode: North American Industry Classification System (NAICS) code. The leading two-digits of NAICS code can be identified as each business company's industry title. We introduced NAICS code list in the industry-level study *(SIC Identification Tools)*.

(7(a) & 504 FOIA).

Information About the Climate Hazard Dataset

There are two levels of granularity for the National Risk Index data for all 50 states. Those granularities are county and Census Tract level:

- County Level: 365 fields in total and 3,142 records or counties.
- Census Tract Level: 375 fields in total, and 72,739 records or Census Tracts.

We decided to use the county-level data as a starting point for our study and research. The dataset contains 18 types of climate hazards. We limit the data to the 10 most common climate hazards in the U.S. The data includes risk scores, expected annual loss and the number of events for each climate hazard *(National Risk Index)*.

4. Methodology

We use the following methodology to analyze the impact of climate hazard risk on the SBA's 7(a) loan program portfolio:

1. **Find the relevant metrics to measure or quantify climate hazard risk.**

In general, climate hazards will result in physical damage to houses, buildings, factories, and equipment, disrupted supply chains, decreased revenue and cash flow, and in the worst-case scenario, physical injuries to staff. Therefore, relevant variables in the data are needed as metrics to measure to what extent climate hazards are causing damage. For more accurate analysis, the data should also have a time frame for when a specific climate hazard happens. Given a particular time frame, we can check the data to see if a relationship can be determined between when a climate hazard happens and when a company's loan defaults or a company takes another loan to cover the damages that the hazard has caused.

Looking at the provided climate hazard dataset, there exists variables such as the number of climate hazard events, risk scores of different hazards, expected annual loss on building value, which can be used to measure the climate hazard risks. However, there is something needed to be mentioned that the climate hazard dataset lacks the occurrence date of each climate hazard. Therefore, we cannot be sure that the borrowing actions of business companies exactly happen behind the occurrence of the climate hazard. For example, we have data about coastal flooding risk scores and expected loss in Los Angeles. But we don't have data to state when the hazard happened so we cannot be sure the company A in Los Angeles borrowed the 7(a) loan because of the coastal flooding in a specific year or month. Therefore, the variables in the provided dataset are limited so it will bring the bias on the third step. In order to avoid the bias, we extracted the loan data from 1991-2019 for study, which is basically consistent with the climate hazard data. Because the data of 10 common climate hazards we choose recorded as of Year 2019. Therefore, it's assumed that all the borrowing actions happened in the same year of the climate hazards. And it also means the occurrence of borrowing actions has the possibility of due to the climate hazards in the same year.

2. **Find the relevant metrics to measure or quantify the credit risk of SBA's 7(a) loan program portfolio.**

We need a metric for the credit risk so we tried two different metrics in our project. First, we used the initial interest rate to indicate loan loss risk. In general, a higher interest rate means higher credit risk. Second, we defined a new variable called loan loss ratio. This metric is calculated by the gross charged-off amount divided by the total approved loan amount. The higher the loan loss ratio, the more money the borrowers owed and could not repay to their lenders.

3. **Finding a relationship between the indicators of climate hazard risk and credit risk.**

Based on the previous two steps, we need to find a relationship between the variables of the two datasets that can indicate that when climate hazard risk increases then credit risk will increase. We fit different machine learning models and assess the models' performance to determine the relationship. The relationship results can show the impact of climate hazards risk on the SBA's 7(a) loan program portfolio. We will formulate different granularities of the datasets such as firm, industry, county level, and thereof to analyze the impact and differences in risk across the different portfolio segments.

5. Data Preparation

5.1 Data Cleaning

Loan Data

After collecting the climate dataset and loan datasets, we begin to do the data cleaning. The cleaning steps consist of data quality checking and data integration. First, we cleaned the loan datasets. One by one, we cleaned the three 7(a) loan program datasets from 1991 to 2021. First, we check the data completeness for each loan dataset and we got the results in the following:

- In the FY1991-FY1999 file, there were 469 missing values for ProjectCounty (0.001392%) and 17 missing values for BorrState (0.000050%).
- In the FY2000-FY2009 file, there were 186 missing values for ProjectCounty (0.000269%) and 11 missing values for BorrState (0.000016%).
- In the FY2010-Present file, there were 2 missing values for ProjectCounty (0.000003%) and 0 missing values for BorrState (0.000000%).

As we can see, the proportion of missing values for these two variables is very small and if we lack the county and state data we could not generate FIPS code and connect with the climate data. About dealing with the missing data, there are many ways like imputation or directly dropping. We also tried to impute the missing county and state using geocoding the borrower street address but the process cost a lot of time and the result is not accurate. In fact, the small proportion of missing values will not influence the final result, so we decided to drop the total missing values around 657 records of the total loan data.

Secondly for the data selection, in the climate data, most of the records for the climate hazards happened before 2020, so we also decided to choose the records in the loan data that happened before 2020 to be consistent with the climate data.

Thirdly, we then found there are 990,757 missing values in the variable initial interest rate (60%). Based on the client's requirement, we need to regard the initial interest rate as the metric to measure credit risk. We filled in the missing values for that variable so that we could continue our study and research. In the raw data, we found that the initial interest rate was acquired from the prime rate published in a daily national newspaper. In other words, the initial interest rate is not completely related to the industry categories so we can not use the industry interest rate to impute the missing initial interest rate. Therefore, we used historical bank loan prime loan rate data for each day from 1991 to 2019 from the Federal Reserve Economic Data *(Bank Prime Loan Rate)*. We found that the prime rate changed within a certain interval for each year and at the same time, the official data is not complete, so we used a uniform distribution to generate the initial interest rate for each year. This method can make sure that the numbers generated were inside the interval for that year. And because the interval we gave is not large, the values generated randomly will keep a steady mean. As for why we don't impute the missing values using the mean value of interest rate, it's because the interest rate changes for different dates or different months in the real world and the uniform distribution generation method can simulate the fluctuation. If we use the mean value method, the interest rate is fixed, which would deviate from the real world. In fact, we strongly advised that we had better use the dataset which doesn't have too many missing values in the future.

For the loan data, the raw data is complete and tidy already, so it wasn't necessary to do a lot of data cleaning. Lastly, we stacked up all three loan datasets to form a final loan dataset from 1991 to 2019.

Fourthly, for connecting the loan data with the climate data, we need to find a key to make the connection. We found that state-county FIPS codes are existent in the climate data and that they could be used to make the connection. The state-county FIPS code (5-digit) consists of state FIPS code (2-digit) and county FIPS code (3-digit). The state-county FIPS code is missing in the loan data. We found an external dataset that includes all of the county and state codes in the U.S. This dataset could be used to extract the state and county FIPS codes from the variables project county and borrower state in the final loan dataset. After having extracted the two different types of codes, we were able to combine them into one state-county FIPS code for each record. Then finally, we could use that 5 digit state-county FIPS code from the final loan dataset to connect with the 5 digit state-county FIPS in the climate dataset.

Climate Hazard Data

Next, we cleaned the climate data. In the raw climate data, there are a lot of missing values for hazards that simply cannot happen in some counties, for example, an avalanche in a county that has no mountains. We filled those missing values with 0. As mentioned above, we need to use state-county FIPS codes as the key to connecting the loan data with the climate data. Some

state-county FIPS codes in the climate data were incomplete and we filled those incomplete codes as 5-digit codes in the climate data. After reading the American Red Cross (Common Disasters Across the U.S.) report on the ten most common climate hazards across the U.S., we decided to focus on those 10 climate hazards. They are coastal flooding, drought, earthquake, hail, hurricane, ice storm, landslide, riverine flooding, tornado, and wildfire.

Loan Data and Climate Hazard Data

After having finished cleaning both datasets, we began to integrate the loan data with climate data using the state-county FIPS codes. After the integration, we also cleaned the final combined large dataset by dropping duplicate rows, and so forth.

Adding the Loan Loss Ratio Metric

Apart from the initial interest rate as the metric to measure credit risk, we also explore the loan loss ratio. We define the loan loss ratio as to how much proportional money the banks and SBA lose of their original investment, given that a loan defaults. The higher the ratio, the higher proportion of their investment the banks and SBA lose. The loan loss ratio is defined as the gross charged-off amount divided by the gross approval amount. In the final combined dataset, we added a new column called loanloss_ratio that contains the calculated value of this metric for each record. In fact, the raw loan data shows another problem, that is, the data show each company's loan status like paid in full, charged off or cancelled. But it didn't show when they paid in full to the bank, whether it happened after the climate hazards or before the hazards, therefore, we have to assume that these status happened after the occurrence of climate hazards. We reasonably think after the climate hazards some companies can't pay back to the bank so that the bank charges off the money. The amount of money that is charged off by the bank is the amount of loan loss.

Finding Outliers

We also explored part of variables' outliers in the combined data. As the plots Figure 5.1 including 6 plots below shown, the variables, initial interest rate, loan loss ratio, gross approval amount, gross charge off amount, risk score and total expected annual loss (EAL_VALT). And the percentage of outliers of risk score in the data is 3.9%, that of InitialInterestRate is 0.1322%, that of loanloss_ratio is 4.37%, that of gross approval amount is 2.17684%, that of gross charge off amount is 1.4675% and that of total Expected annual loss is 5.393978%. Figure 5.1 shows the outlier distribution of the variables we choose and the distribution of the variables in the whole dataset. As we can see, these variables are right skewed.

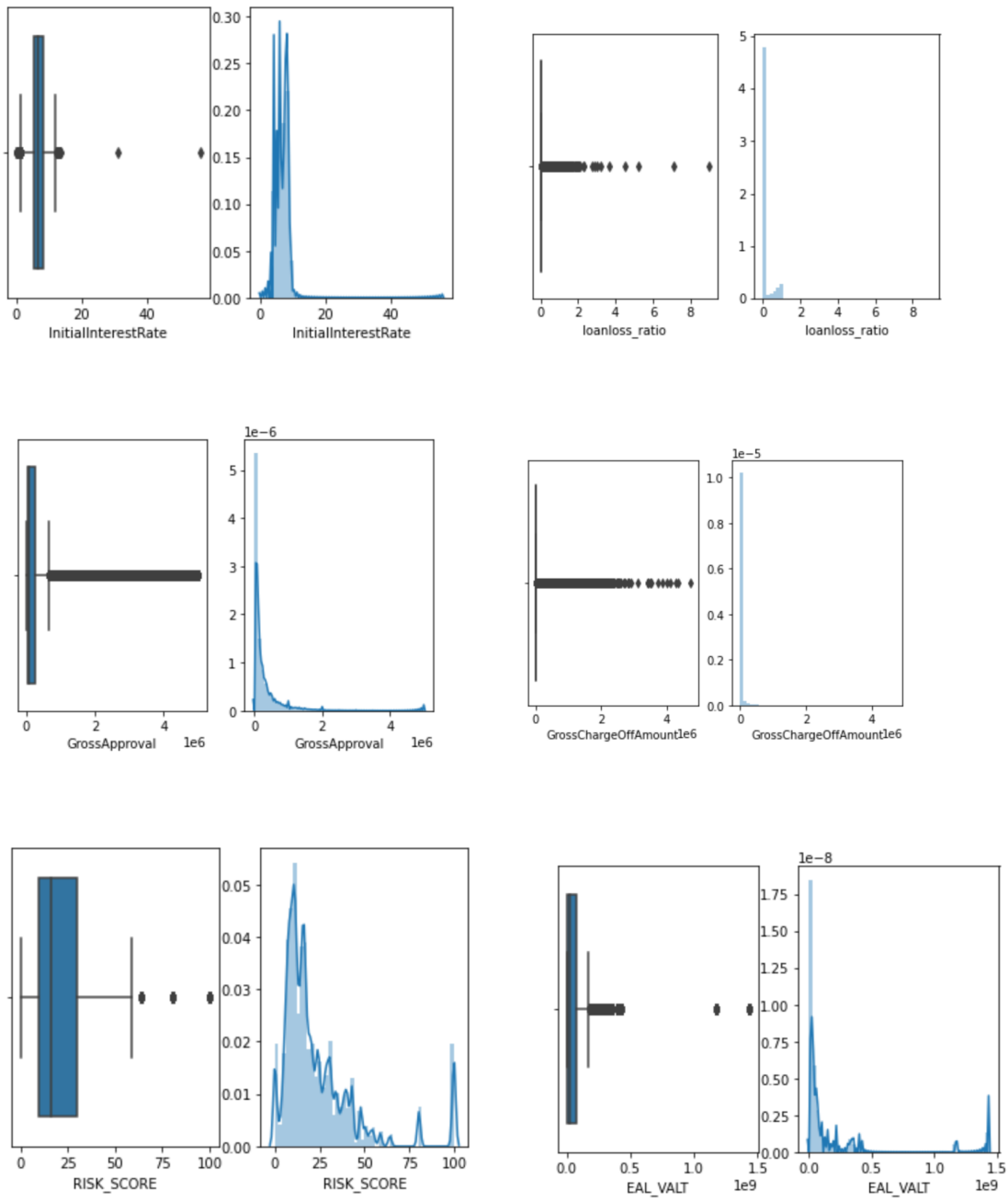

Figure 5.1: A Boxplot of a Part of the Variables' Outliers and a Density Plot of a Part of the Variables.

Dealing with Outliers

After finding out the outliers, we can see the proportion of outliers is small so we can delete it directly or delete the values below the lower bound value. We used the latter method to deal with the outliers because the original data show right skewed. The Z-score method can help to transform the data into a normal distribution which will be more reliable for the analysis. We set the threshold is 3 or -3 so if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers and then removed. Because there exist many variables, we have to choose two dependent variables to remove their outliers. They are loan loss ratio and initial interest rate. As we can see from Figure 5.2, after removing the outliers, the boxplots show the less outliers and the distribution shows normal distribution. The Figure for loan loss ratio shows the values in loan loss ratio mainly are 0 and some are between 0.5 and 0.75.

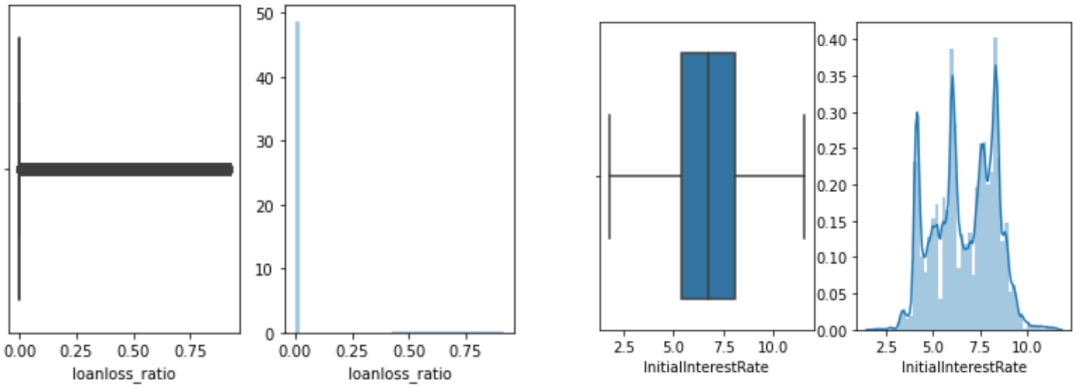

Figure 5.2: A Boxplot of a Part of the Variables' Outliers and a Density Plot of a Part of the Variables After Removing the Outliers.

Normalization

We decided to run normalization for all independent variables for more accurate relationship exploration. By doing so, we converted the data into values between 0 and 1. The independent variables are primarily from the climate data, and the climate data conforms to a normal distribution with few outliers involved. Figures 5.3 and 5.4 below show that the independent variables SOVI_SCORE (Social Vulnerability-Score) and RISK_SCORE conform to a normal distribution. As we can see from Figure 5.6, the risk score variable shows normal distribution in general but a little bit right-skewed in the data. We used a Min-Max feature to normalize the data. Figure 5.5 below shows the difference between original data and normalized data for the variable WFIR_RISKS.

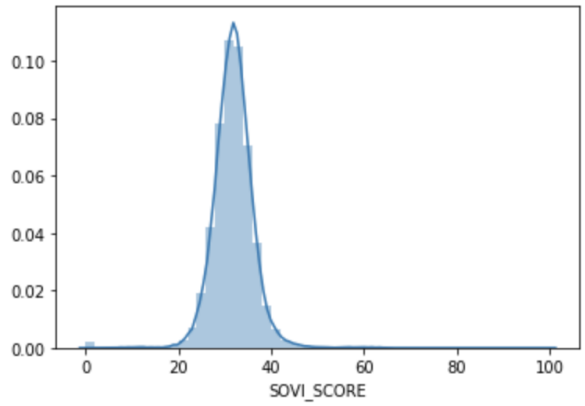

Figure 5.3: Density Plot of SOVI_SCORE.

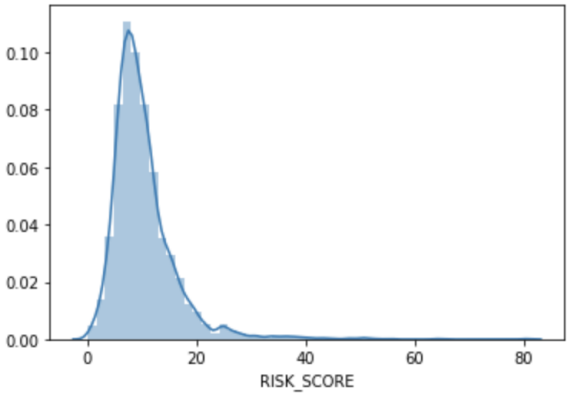

Figure 5.4: Density Plot of RISK_SCORE.

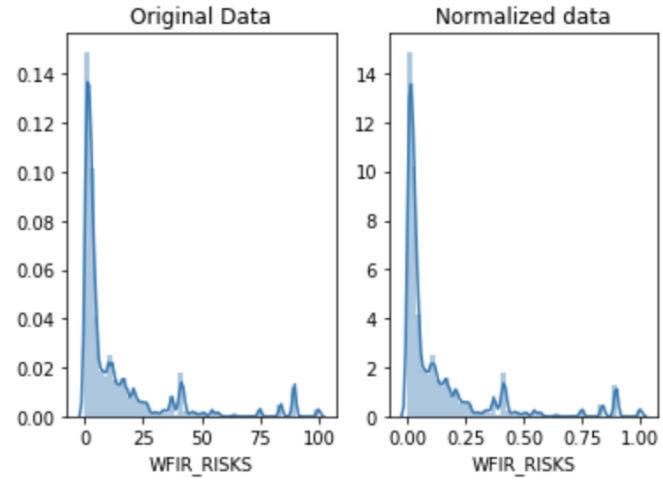

Figure 5.5: Comparison of the Density Plot between original WFIR_RISKS and normalized WFIR_RISKS.

5.2 Data Exploration

Adding the Industry Sector and Industry Title

We added a dimension to the analysis by analyzing the different industries of the companies receiving a loan. We found that we could use the NAICS code that already existed in the loan dataset to add this dimension to the analysis. The first two digits in the NAICS code stand for the different industries categories, and the different codes can be mapped to specific industry titles. Therefore, we manage to successfully map the industry for each company.

Adding Regions and Divisions

In order to analyze the credit risk for companies in the different regions and divisions across the U.S. and their exposure to climate hazards, we added new columns/variables that contained the region and division of each company's loan in the final dataset.

Descriptive Statistics

Figure 5.6 shows part of average gross approval amount, initial interest rate, gross charge off amount and loan loss ratio by 3116 FIPS codes. And they are shown in ascending order. As Figure 5.6 shows, in Michigan ST Joseph, there exists the initial interest rate 7.64 % on average and in this county, its total average gross approval amount is 20,000. Figure 5.7 shows the summary statistics of the combined data. The median and highest initial interest rate in the data is 6.73% and 56% respectively.

	sab	cname	fips	GrossApproval	InitialInterestRate	GrossChargeOffAmount	RISK_SCORE
1291	MI	ST JOSEPH	26149.0	20000.000000	7.643616	13949.0	8.814694
2717	TX	TERRELL	48443.0	39250.000000	7.323193	26218.0	5.206978
690	IL	ALEXANDER	17003.0	167283.333333	8.325676	110876.0	18.946870
2191	OR	GILLIAM	41021.0	150000.000000	5.203495	92944.0	4.738763
1490	MO	WORTH	29227.0	20000.000000	7.500000	12059.0	6.624264
...	...	...	...	...	...	...	...
1623	MT	TOOLE	30101.0	289942.125000	7.099809	0.0	4.617867
1627	MT	WIBAUX	30109.0	52062.500000	8.183429	0.0	3.707758
2613	TX	IRION	48235.0	67999.500000	6.601902	0.0	5.792034
2601	TX	HEMPHILL	48211.0	136777.777778	7.455452	0.0	13.756451
0	AK	ALEUTIANS EAST	2013.0	40000.000000	6.000000	0.0	4.508506

3116 rows × 300 columns

Figure 5.6: Average Gross Approval Amount and Other Variables by FIPS Code.

	InitialInterestRate	GrossApproval	GrossChargeOffAmount	loanloss_ratio	RISK_SCORE
count	1.566529e+06	1.566529e+06	1.566529e+06	1.566529e+06	1.511728e+06
mean	6.692763e+00	2.561719e+05	1.310443e+04	9.429175e-02	2.459024e+01
std	1.634211e+00	4.694302e+05	7.313488e+04	2.603867e-01	2.182042e+01
min	0.000000e+00	7.500000e+01	0.000000e+00	0.000000e+00	0.000000e+00
25%	5.500000e+00	3.500000e+04	0.000000e+00	0.000000e+00	1.058310e+01
50%	6.734834e+00	1.000000e+05	0.000000e+00	0.000000e+00	1.667308e+01
75%	8.050100e+00	2.606000e+05	0.000000e+00	0.000000e+00	3.025643e+01
max	5.600000e+01	5.000000e+06	4.706180e+06	8.979200e+00	1.000000e+02

Figure 5.7: Descriptive Statistics of the Combined Data Set.

Relationship Exploration

After normalization, we did some exploratory analysis on the updated dataset. We want to know if there exists a linear relationship between the independent variables and the dependent variables initial interest rate and loan loss ratio. As we can see from Figure 5.8 and 5.9 below, we found no significant linear relationship between the independent and dependent variables (Note: We tried two dependent variables, the initial interest rate, and the loan loss ratio). And even we cannot fit regression lines on the plots. So we have reasons to believe that there is no relationship between them, neither linear nor non-linear. Our next modeling section will focus on the exploration of the non-linear relationship. Figure 5.10 shows the average Initial Interest Rate by each FIPS Code as dependent variable and risk score as the independent variable, and then we can see all the data points show a very subtle linear relationship (nearly horizontal line) so we can get the primary hypothesis as well, there is no (linear) relationship between the independent variable and dependent variable. (Notice: We also added Figure 5.10 here for relationship exploration. It is based on the mean value of initial interest rate and we kept this for better readability but for the plot, the mean value is easy to be influenced by the extreme values.)

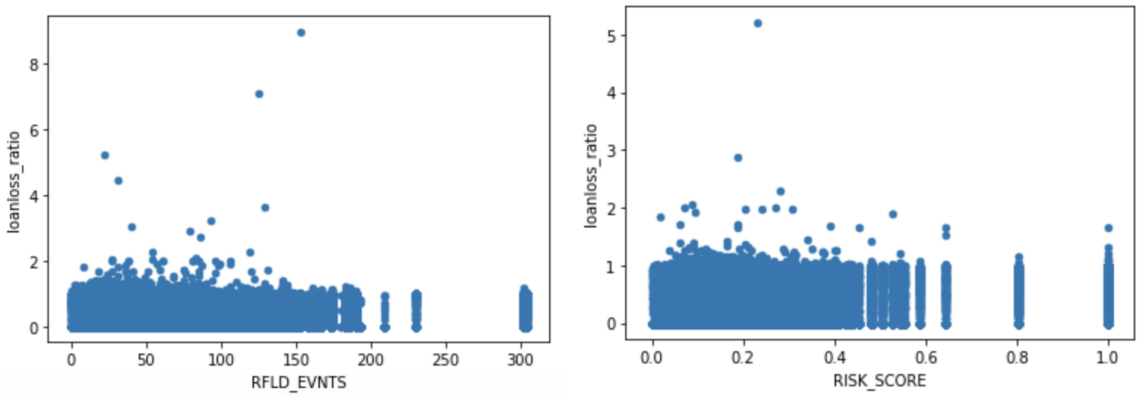

Figure 5.8: A Scatter plot of the Independent Variables and The Dependent Variable Loan Loss Ratio (RFLD_EVENTS: Riverine Flooding - Number of Events).

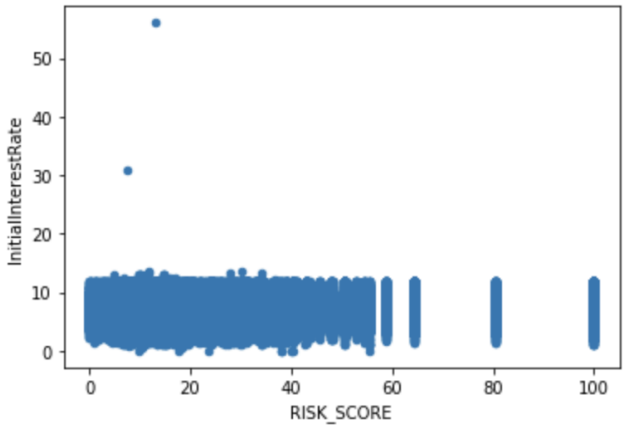

Figure 5.9: A Scatter Plot of an Independent Variable and the Dependent Variable Initial Interest Rate.

We also did the correlation exploration, shown in Figure 5.11. On the left-hand side plot, it shows the correlation between initial interest rate and all the independent variables (we just extract 15 independent variables here). On the right-hand side plot, it shows the correlation between loan loss ratio and all the independent variables (we just extract 15 independent variables here). They all show very low correlation, just around 0.01 and 0.03 respectively. So it validated our primary hypothesis, there is no (linear) relationship between the independent variable and dependent variable.

	Pearson Correlation Coefficient
DRGT_RISKS	0.017505
ISTM_EALP	0.016083
RFLD_RISKS	0.015303
TRND_EALT	0.013108
CFLD_RISKS	0.013099
TRND_RISKS	0.012869
SOVI_SCORE	0.012861
TRND_EALP	0.012573
HRCN_EXPP	0.012448
HRCN_EXPT	0.012442
DRGT_EALS	0.012162
HRCN_RISKS	0.011916
HRCN_EXPB	0.011902
TRND_EALB	0.011770
ISTM_RISKS	0.010358

	Pearson Correlation Coefficient
GrossChargeOffAmount	0.488601
TRND_RISKS	0.042794
RISK_SCORE	0.037665
TRND_EALS	0.035065
RFLD_RISKS	0.033456
TRND_EXPB	0.033265
ERQK_EXPB	0.033265
HAIL_EXPB	0.033264
ERQK_EXPT	0.032212
TRND_EXPT	0.032212
HAIL_EXPT	0.032212
ERQK_EXPP	0.032195
TRND_EXPP	0.032195
HAIL_EXPP	0.032195
HRCN_EXPP	0.032045

Figure 5.11: Correlation between Initial Interest Rate and All Independent Variables (left); Correlation between Loan loss ratio and All Independent Variables (Right).

6. Modeling & Evaluation

6.1 Modeling

The measures used to test the accuracy of the final models are R-squared and Root Mean Square Error (RSME). We randomly split the dataset into train and test subsets to mitigate bias due to sample dependency. The training dataset accounts for 70%, and the test part accounts for 30% of the full dataset. We ran Ridge, ElasticNet, and Gradient Boosting Machine regressions, and XG Boosting, and Random Forest models, as can be seen in Table 1 below. The ElasticNet Regression provides us with the benefits of both Lasso and Ridge regression. It can perform similarly to Lasso's feature selection and Ridge's feature-group selection.

	R^2		RMSE	
Selected Model	Initial Interest Rate	Loan Loss Ratio	Initial Interest Rate	Loan Loss Ratio
Ridge Regression	0.0198	0.2632	6.7998	0.1737
ElasticNet Regression	-2.908e-10	-2.908e-10	1.6322	0.2024
Gradient Boosting Machine Regression	0.0379	0.9932	1.60096	0.0167
XG Boost	0.0479 0.047552	0.9994	1.596936 1.592915	0.0051
Random Forest	-0.0895	0.99993	1.7037	0.0017

Table 1: Modeling Results.

6.2 Evaluation

How to Assess Credit Risk?

As mentioned in the Methodology section, we utilized two metrics to measure credit risk: Initial interest rate and loan loss ratio. The higher the value of those two metrics, the higher the credit risk.

How to Assess Model Performance?

We separated the datasets into training and testing data, which occupied 70% and 30% of the whole dataset, respectively. Then we used cross-validation to test the model performance. As shown in Table 1, we utilized R^2 and RMSE (Root Mean Squared Error) to assess model performance. R^2 measures the strength of the relationship between the climate hazard risk and credit risks. RMSE measures the differences between predicted values and observed values of the dependent variables.

7. Results, Conclusions, Visualizations, and Recommendations

7.1 Results and Conclusions

As is shown above in Table 1, five different machine learning models were compared based on two different metrics (initial interest rate and loan loss ratio). For Ridge regression, we chose an alpha value of 0.01. Results show that the independent and dependent variables have a weak relationship even though the RMSE is relatively low. For Gradient Boosting Machine regression, we can see that the R^2 for the initial interest rate is very low but very high for the loan loss ratio, and their RMSE shows the opposite results. We can see no significant relationship between our data's independent and dependent variables. This result is what we predicted before modeling. We will explain the reasons and how to improve our analysis in the Challenges & Improvements section.

7.2 Visualizations

Figure 12 below shows FEMA's National Risk Index score for each county combined with the SBA's 7(a) loan program gross approval amount per county. The size of the dots represents the gross approval amount, while the color of the dots represents the climate hazard risk. The National Risk Index score is a summarized and aggregated score based on 18 natural hazards, the expected annual losses from those natural hazards, social vulnerability, and community resilience. The score is calculated for each county. FEMA's website has good documentation on how each score for the composite risk score of the National Risk Index is derived, and the link can be found in the References section.

At first glance, states near the coastline seem to be at most risk, with California, Texas, and Florida standing noticeably out. California's wildfires and Texas's cold wave have been in the spotlight for 2021, and Florida's hurricanes regularly hit the spotlight. The size of the dots shows that these states are also receiving a significantly large amount of the loans provided by the 7(a) program, with Los Angeles being the county with the highest number of loans nationally, 66,516 or around 4.3% of all national loans.

If we switch the measure to the gross charge-off amount, we can see that the map barely changes; the gross charge-off amount is strongly positively correlated with the gross approval amount.

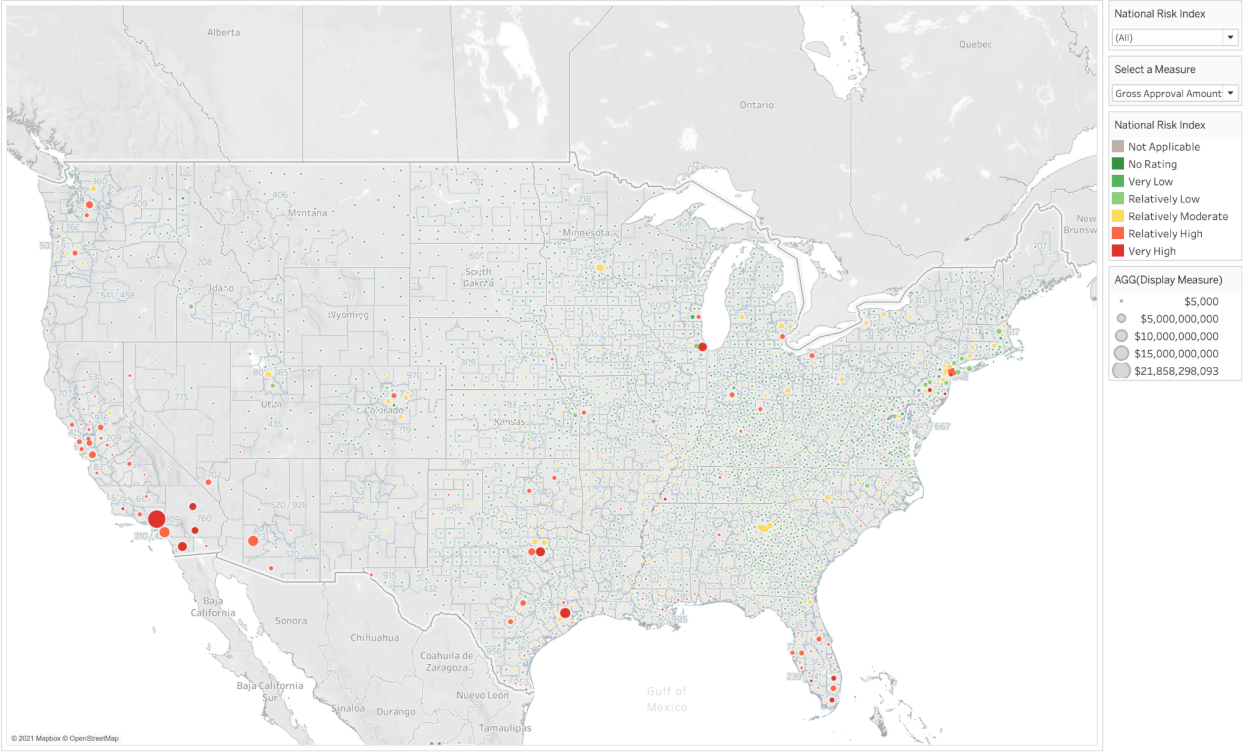

Figure 12: Gross Approval Amount for All Risk Levels of the National Risk Index.

Figure 13 below shows that counties with a relatively moderate risk of climate hazards are more densely distributed along the Midwest, Northeast, and South regions. However, there is still a significant portion of the loans at moderate risk in the West region. Again we see that the three noticeable states of California, Texas, and Florida stand out.

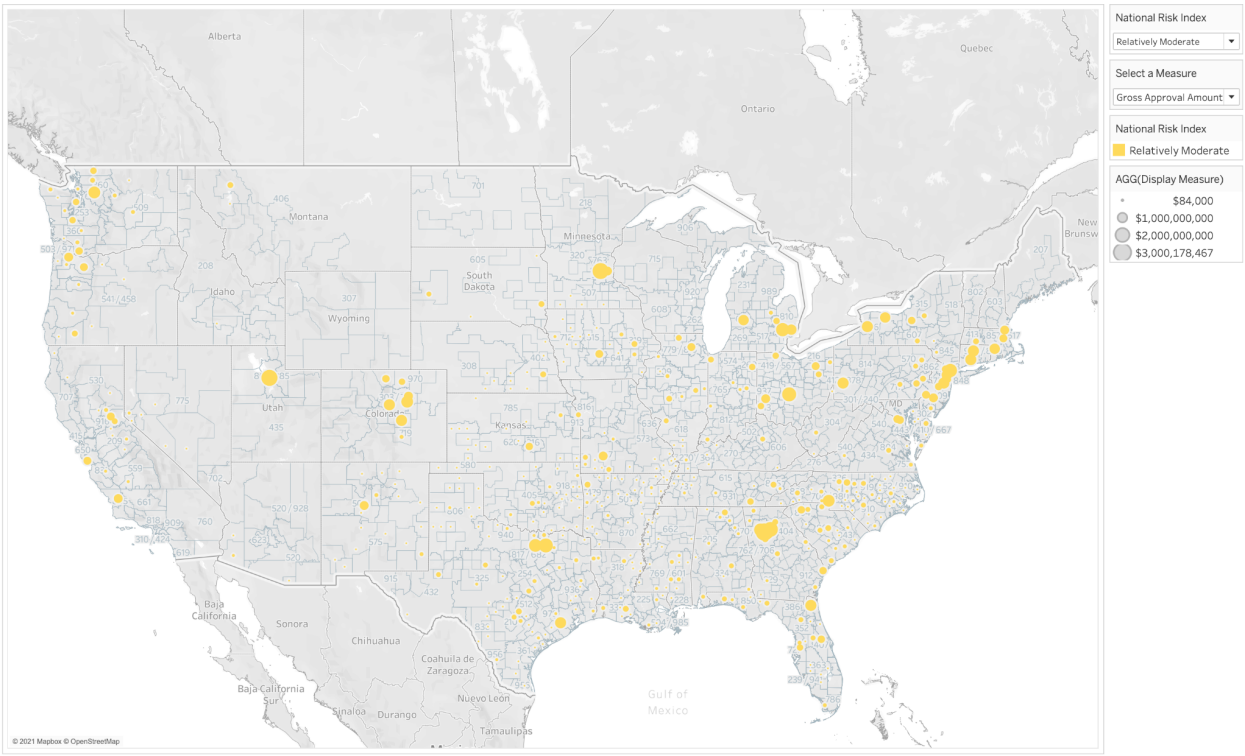

Figure 13: Gross Approval Amount for a Relatively Moderate Risk Level of the National Risk Index.

Figure 14 below shows that California has a noticeably large amount of counties and loans at relatively high risk. The West and South regions stand out, with Texas and Florida also contributing to many counties and gross approval amounts at relatively high risk.

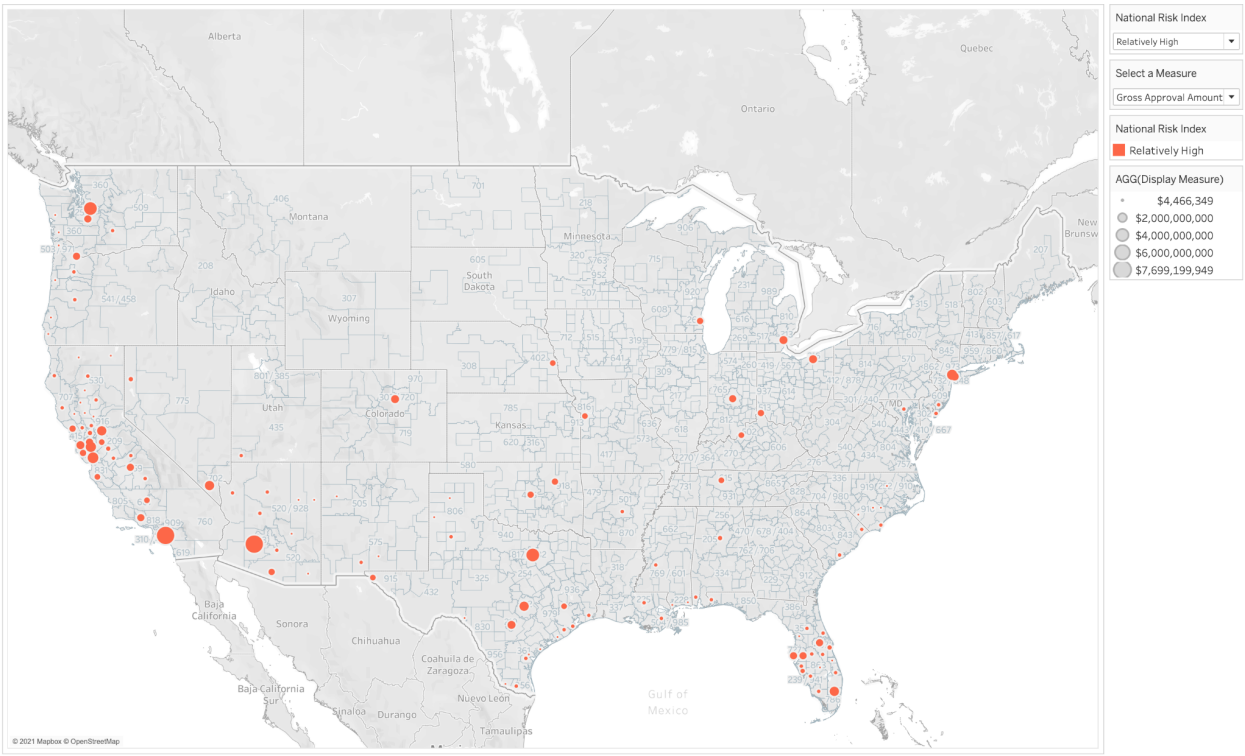

Figure 14: Gross Approval Amount for a Relatively High Risk Level of the National Risk Index.

Finally, Figure 15 shows that California and Texas have the highest gross approval amounts at very high risk due to climate hazards. Los Angeles receives a risk score of 100, any loans made in that county should be handled with extra care, and research should be done on the exact locations of these loans. It is recommended to consider a higher interest rate for loans in this county. We also notice that Florida is less noticeable this time but still has three counties at very high risk.

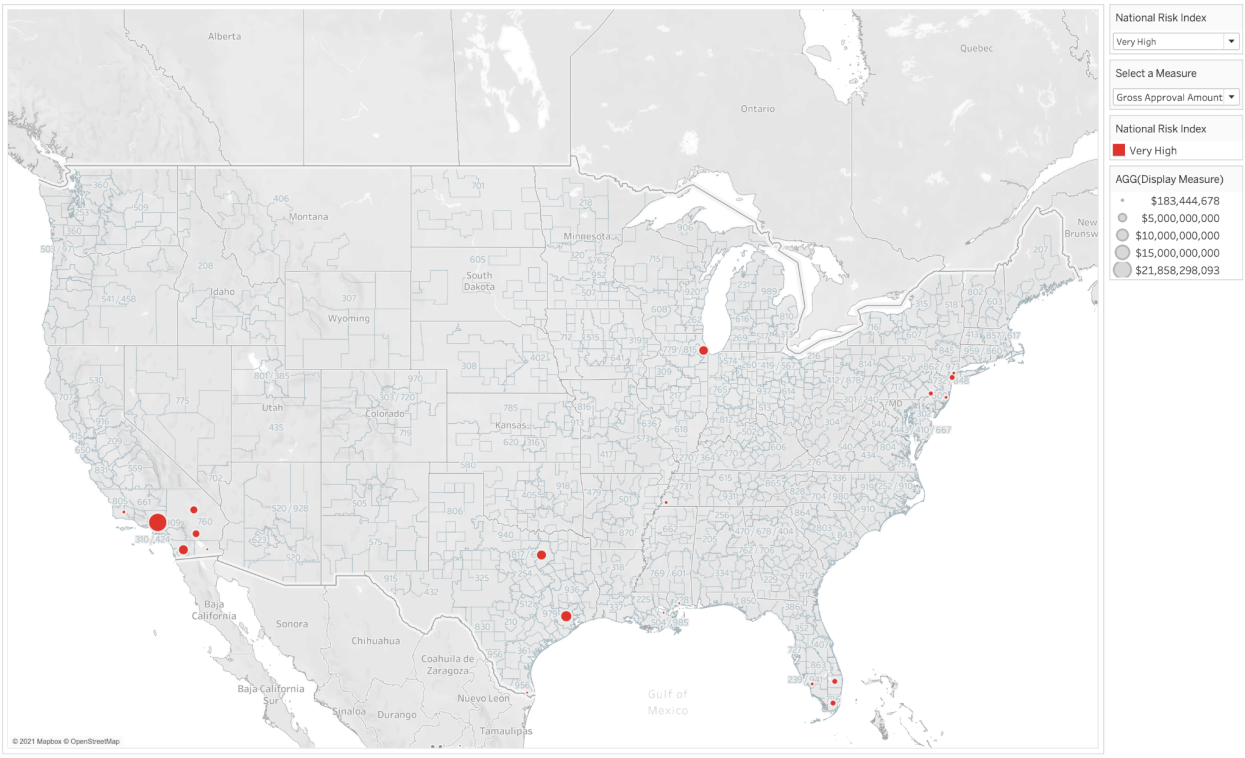

Figure 15: Gross Approval Amount for a High Risk Level of the National Risk Index.

Figure 16 below shows that more of the portfolio is exposed to relatively moderate to very high risk than relatively moderate to very low risk. That is a big concern and might indicate that integrating the loan data with the climate data risk scores on a county level is possibly making too many strong assumptions about each loan.

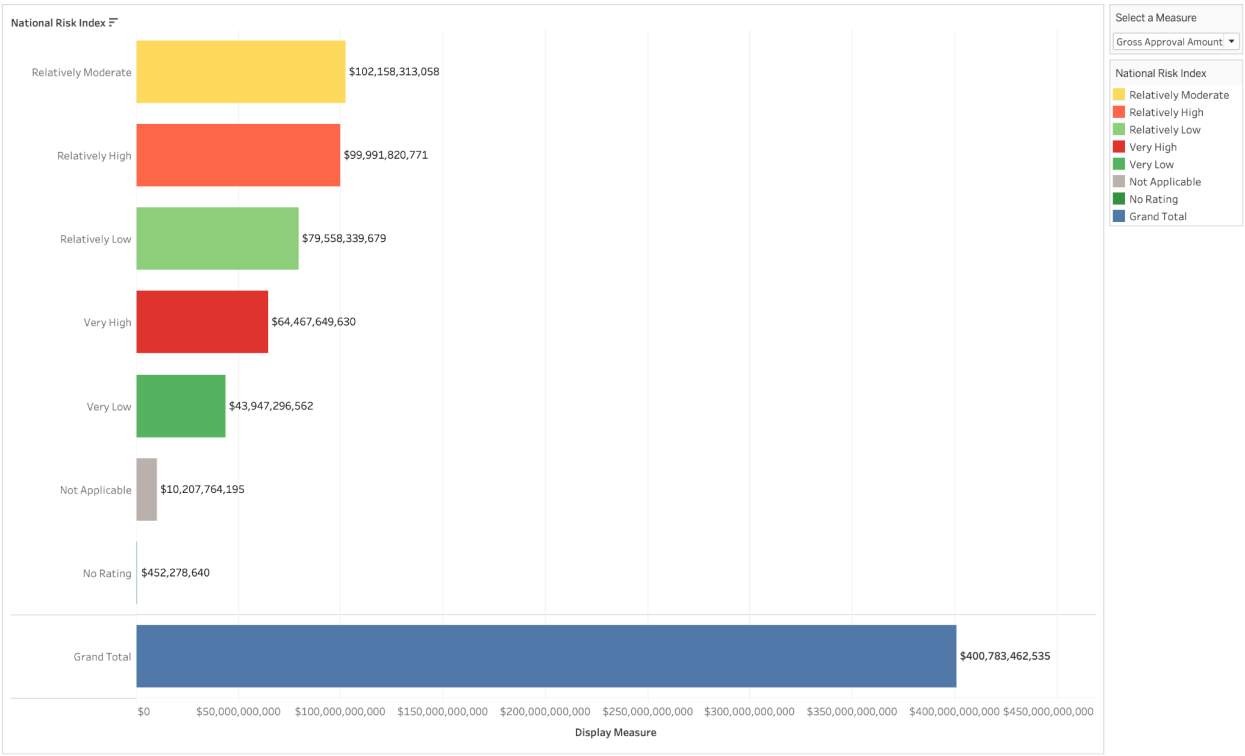

Figure 16: Distribution of Gross Approval Amount Across the Different Risk Levels of the National Risk Index.

Figure 17 below shows the exposure in percentages. 66.53% (25.49% + 24.95% + 16.09%) of the portfolio is exposed to a relatively moderate to very high risk of climate hazards.

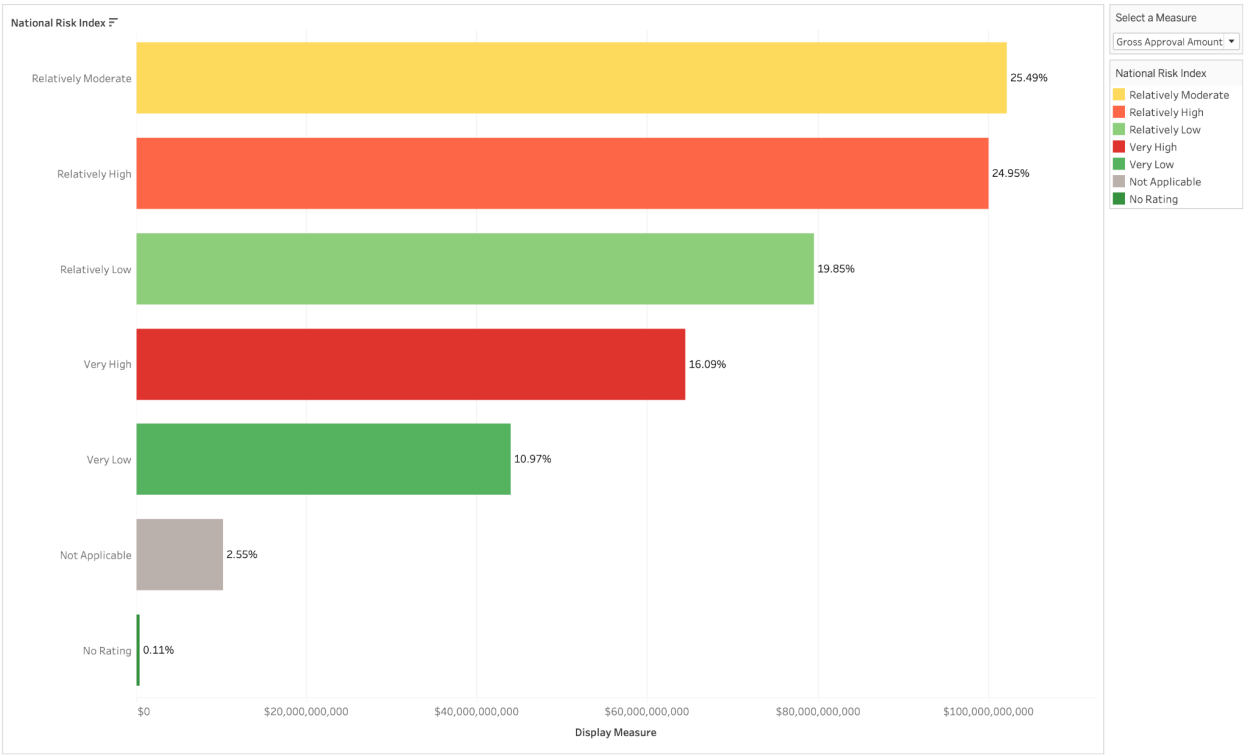

Figure 17: Distribution of Gross Approval Amount in Percentages Across the Different Risk Levels of the National Risk Index.

Figure 18 below shows that the South and West regions have the highest gross approval amounts. A significant part of these amounts is exposed to relatively moderate or very high climate hazard risk. This is aligned with the map chart visualizations and shows that these regions are more likely to be negatively impacted by climate hazards.

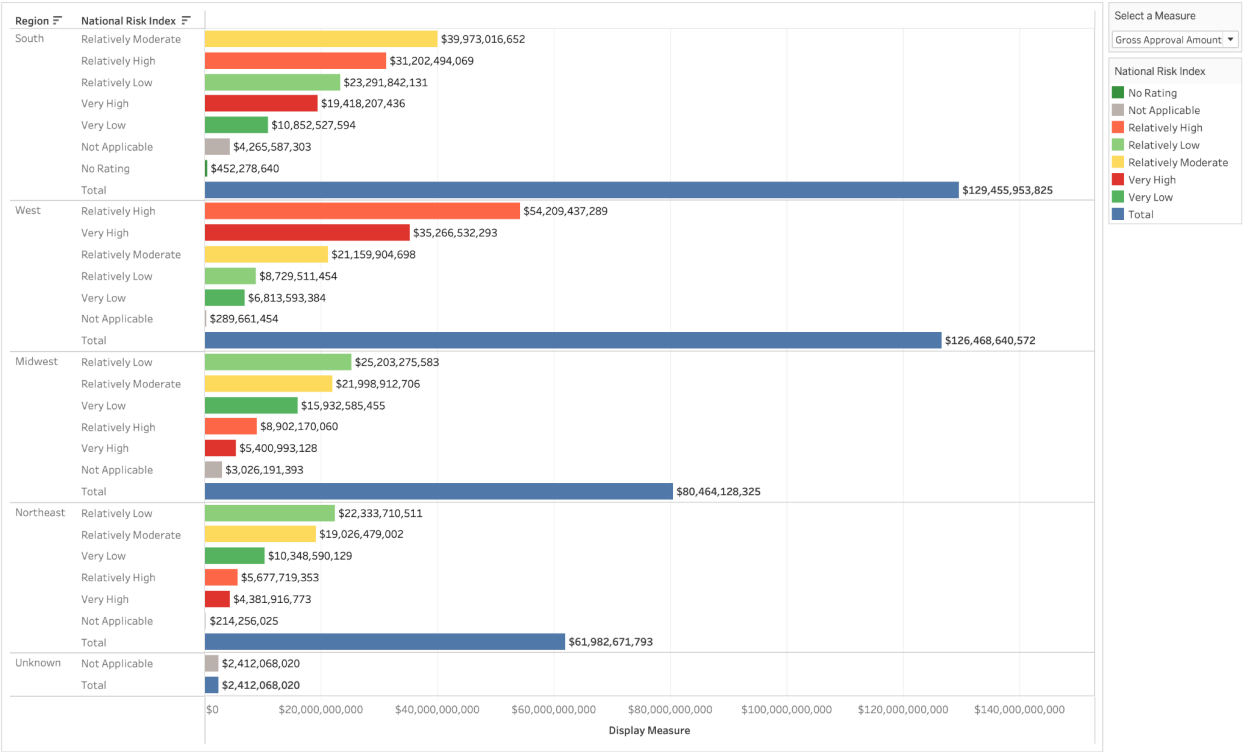

Figure 18: Distribution of Gross Approval Amount Across the Different Regions and Risk Levels of the National Risk Index.

We have seen from the above figures that a large part of the 7(a) loan program portfolio is exposed to climate hazards. Regardless, figure 19 below shows no apparent macroeconomic relationship between climate hazard risk and credit risk in the form of average initial interest rate. However, we notice that as climate risk increases, the number of loans in each relatively high to very high climate risk county is increasing very noticeably. Climate hazards can have devastating effects on loan loss on the microeconomic level, but we do not yet know the effect of this increased risk on the macroeconomic level. The data shows a pattern of an increased number of loans in high-risk areas and should be a warning, encouraging mitigating risk measures.

Figure 19: Average Initial Interest Rate per FIPS Code (County) vs. Risk Score.

7.3 Recommendations

Lenders such as banks and the SBA have to start considering the fact of climate risk. The National Risk Index has shown that a significant part of the SBA's 7(a) program portfolio is exposed to climate hazard risk, but the analysis shows that these exposed loans are not on average charged a higher interest rate than loans less exposed to climate risk. However, the data makes strong assumptions and is not very granular. We recommend introducing new time-series data sets that; 1. have specific information on climate hazards in the areas and surrounding areas of the loans and 2. show cash flows of the companies that have received those loans. We also recommend introducing a more granular way of viewing the data using census tract instead of county-level data.

It would be desirable to see non-summarized and aggregated data on the time, date, frequency, and severity of climate hazards in these counties and census tracts. That way, it should be easier to see the relationship climate hazards have on companies' cash flows and loan default. We recommend starting to build the updated models using data from California, Texas, and Florida. Picking specific hazards that are frequent in those states can also be done. Starting modeling in Los Angeles County is desirable as the county is at very high risk due to climate hazards.

8. Challenges & Improvements:

8.1 Reflection on What Limitations Affect Our Project

Infeasible Input Data

In the current climate data analysis, feasible variables are insufficient and unspecific. Expected annual loss value on building/agriculture for each climate hazard, risk scores, number of occurrences in the provided data are feasible for this project. However, they are not detailed and not enough. The ideal input data should include the extent to which climate hazards have physical damage to houses, buildings, factories, and equipment, disrupted supply chains, and physical injuries to staff in the worst-case scenario. The data should also have a specific year, month, and duration about when a specific climate hazard happens to properly make sure that the loan transaction happened after the climate hazard happened. In other words, we may know for sure that the company has a loan from the bank precisely due to the climate hazards.

We lack companies' financial data in the loan data, such as cash flow information of borrowers, revenue loss, supply chain damage, and other loan programs. These data are hard to collect because they are too complicated, and sometimes the loss is hard to quantify, such as the physical injuries to staff. Some financial data are private for companies, so it is hard to acquire internal data. Apart from that, the loan data only focus on one loan program, 7(a). As we all know, to get through the hazards, companies may borrow different kinds of loan programs to apply in the rebuilding factories or compensation for staff. We do not know if they have loans precisely because of the hazards. That is why it is essential to know what the loan money is used for and when the loan transaction happens.

Lack of a Better Measurement for Credit Risk

We used the initial interest rate and loan loss ratio to measure credit risk. The banks determine the initial interest rate of each loan after a credit risk analysis. However, the credit risk analysis does not put great emphasis on climate risks. Therefore we cannot find an apparent relationship between credit risk in the form of initial interest rate and climate hazard risk. As mentioned in “Adding the Loan Loss Ratio Metric”, the loan loss ratio is calculated using the gross charge off amount. However, we cannot determine the loan status, “already charged off by the bank”, after a climate hazard has happened. If we blindly assume that the loan status changed after the climate hazards, that would lead to a biased result.

Difficult to Collect the Proper Datasets

Some financial data are private for companies, so it is hard to acquire internal data. Also, it is hard to find accurate climate data, including all needed variables such as time, damages amount, financial impact of physical risk, etc.

Data Granularity

We used the county level for our analysis which is not very granular. For example, riverine flooding may cause massive destruction on one side of a hill, while there is no destructive impact on the other side of the hill in the same county. For these situations, the county-level granularity is not accurate enough. We may wrongly think that riverine flooding severely impacts the whole hill. Therefore, this selection of data granularity will lead to inaccurate analysis results in many cases.

8.2 How Can We Improve on the Project in the Future? (Potential Next Steps)

Acquire More Detailed Data

As mentioned before, the input data is the biggest reason we cannot draw specific conclusions about our modeling results. In acquiring accurate input data, we advise focusing on a couple of companies and collecting their detailed financial data. The input data would be cash flow, damage caused to the company's assets by climate hazards, allocation of loan money, and so forth. We hope the small sample of companies can be located in states or regions where apparent and severe climate hazards exist each year. We could study the targeted climate hazards' effect on the small sample size loan portfolio. The loan data can include multiple loan programs for the same companies.

Find the Right Level of Data Granularity

As mentioned in the reflection above, we could use data sets with smaller granularity to improve the accuracy of modeling performance and have more specific map visualizations. FEMA's National Risk Index provides two granularities of the climate dataset: county-level and census tract level. With permission of enough time, we could acquire the census level of FIPS codes in the loan dataset by using Google's API or other third-party software to convert the borrowers' addresses into accurate longitude and latitude coordinates. Then, we can integrate the climate data with the loan data at a census-tract level, which will provide more accurate and rigorous modeling results. By doing so, we could also extract more information from the visualizations.

More Rigorous Data Processing and Models

We can utilize more rigorous methods once we acquire suitable detailed data. For example, for data cleaning, if we have complete data, we can avoid imputing the tons of missing values using the random generation method. When dealing with the outliers, we could use the z-score method to remove more variables' outliers or use other more rigorous methods to deal with outliers. These are all based on how the provided data sets look. It would be great to use advanced machine learning techniques on the new complete data.

Multiangle Analysis of the Impact of Physical Risk on a Company's Finances

There are three different angles to analyze the impact of physical risk on a company's finances: direct impact, indirect impact, and macroeconomic impact. Direct impact means the direct damages to a company's assets or disruptions to the company's supply chains and operations, which results in credit risk. Indirect impact means that physical risk may influence the company's supply chain, and then the company's suppliers are exposed to disruptions. Then, that results in price change which affects the company's customers, and thus the demand for the company's products and services diminishes. Consequently, revenue decreases, and credit risk increases. Macroeconomic impact means that physical risk affects the region's capital, social vulnerability, community resilience, labor productivity, and household demand. Macroeconomic impact results in the deduction of a company's loan repayment capacity, which results in higher credit risk. The macroeconomic angle is closely related to the macroeconomic impact on credit risk. However, this angle is abstract and more complicated. Therefore we advise beginning with the direct and indirect impact of physical risk on credit risk in the future.

Scenario Analysis

It is also a good idea to do rigorous scenario analysis after acquiring the complete raw input data. By doing scenario analysis, we can accurately know how loan loss will change dependent on if a specific climate hazard occurs more frequently or if there is an increase in on building value, and so forth. We believe that the results from scenario analysis would be very helpful and provide meaningful insight.

9. References

County FIPS Matching Tool & NAICS
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OSLU4G&version=1.0

SIC Identification Tools https://www.naics.com/search/

Bank Prime Loan Rate https://fred.stlouisfed.org/series/DPRIME

Common Climate Hazards
https://www.redcross.org/get-help/how-to-prepare-for-emergencies/common-natural-disasters-across-us.html#all

FIPS code
https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/OSLU4G/TJAKCV&version=1.0

SBA's 7(a) Loans
https://www.sba.gov/funding-programs/loans/7a-loans

SBA's 504 Loans
https://www.sba.gov/funding-programs/loans/504-loans

Bank of America TCFD report
https://www.fema.gov/sites/default/files/documents/fema_national-risk-index_technical-documentation.pdf

American Red Cross - Common Disasters Across the U.S.
https://www.redcross.org/get-help/how-to-prepare-for-emergencies/common-natural-disasters-across-us.html#all

Appendices

Python Jupyter Notebook

Tableau Public Dashboards and Worksheets