Impact of Climate Change on House Mortgages in United States Coastal Areas

Junyi Li
Swati Sharma
Sneha Tirchy Shekar
Vinayak Saxena
Zhexuan Franklin Tang
Zheyan Chen

2023.12

Part I. Abstract

The study aims to find out whether variations in housing markets exist in correlation with increasing climate change vulnerabilities, and if they do, does this correlation between costs to acquire housing and climate change factors mirror the real life reports of reluctance in lenders or climate change pessimism in buyers. (spglobal.com, 2022) The study collects available federal government data and third party open data and visualizes the pertinent climate change trends and housing prices comparisons. In merging all relevant data into a manipulatable data set, all data points are unified into its correct zip codes across state boundaries. The merged data is then normalized and standardized to mirror normal distribution. The study tried both a simple linear regression and a more complex random forest prediction method. After studying the tested results, the study concludes that a random forest prediction method demonstrates a high effectiveness in predicting rise in housing prices and loan rejection as climate change vulnerabilities continue to worsen. Further recommendations include increased public awareness of climate change impacts on cost of house ownership, better public data stewardship of federal agencies, and further research using more advanced machine learning algorithms.

Part II. Literature Review

There are previous data science reports that focus on exploring whether any correlation exists between rising sea level, mortgage costs, and real estate prices. The New York times have reported that rising sea levels threaten the validity of residential mortgages, part of the economic foundation for the United States (Newyorktimes.com, 2020) . This report confirms some existing studies that have explored the relationship between rising sea levels exposure and changes in housing and mortgage markets from 2001 to 2020, that leverages data on home transactions, mortgage applications, flood insurance, and SLR forecasts at the census tract level to investigate how sea level rise exposure may have influenced housing and mortgage markets. (Dahl et al., 2017) This report aims to build upon these previous research and reports, hoping to determine whether the correlation has worsened or the datasets do not support the previous correlation claims.

Part III. Methodology

As initially outlined in our research proposal, the raw data we aim to research would mainly focus on three different directions: 1) housing prices, 2) mortgage data, and 3) environmental data. Initially hoping to find sufficient data to conduct research at one of the seaboards surrounding mainland USA (in anticipation of insufficient/incomplete available data source), we were pleased to find sufficient relevant, high quality data to resume our scope to mainland United States. After 2

weeks of research, we were able to collect a wide range of national level of data from many federal and non-profit agencies that allow us to maintain our scope of research to the mainland United States.

In our search, we have gathered rise in sea level and projected sea level data from National Oceanic and Atmospheric Administration (NOAA Tides and Currents) by individual monitoring stations, federal flood insurance claims (by Hurricanes or other flood-related natural disasters) OpenFEMA National Flood Insurance Program (NFIP) by US zip code, National Mortgage Database (NMDB) aggregate statistics by zip code, and Zillow Metropolitan Housing Data by zip code, and Home Mortgage Disclosure Act (HMDA) Data from 2010 and 2017.

We first visualized climate change data based on NOAA, and compiled the change in sea levels in the past 10 years in Figure 3A, while Figure 3B is a vulnerability index for US coastal locations by combining several factors including distance from coast by zip code, rise in sea level, and federal flood insurance claims. Figures 3C and 3D explored the vulnerability of US coastal regions separately, showing that at-risk regions in coastal Texas have seen the highest rise in sea level, and therefore having the highest rise of ocean flooding. Delaware, New Jersey, and Florida are of medium risk comparatively, while California has the lowest relative flood risk on the Pacific ocean coast. Figure 3E shows in the population confluence of US coastal area by state, where New York, California, and Florida have the top 3 highest populations that live near a coastal zone, with the state of New York's coastal population being disproportionately the largest.

Then we compiled and visualized housing related data. Figure 3F shows the housing value of coastal cities on the US mainland, with some Californian, New York, and Connecticut counties taking the top 5. Figure 3G and 3H displays the disparity between rent and house value index change in US coastal areas, in which the house value change is significantly recorded in California, Texas and Gulf coast, Floria, and the Northeast Corridor. Building upon previous climate change vulnerability data and existing property value data, we have also compiled a US East Coast property value loss from the sea based natural disasters, visualized in figures 3I and 3J. Figure 3J demonstrates that Florida and New Jersey suffer the most property loss from the east coast floods.

Based on the visualized data, we explored their relationship further with a correlation matrix in Figure 3K. The matrix demonstrates featurs all the available independent data variables from our collected data source. Based on our visualization results, we were able to identify the most significant correlation: Mortgage rejects are positively correlated with sea level rise, flood insurance claims and the higher loan amount requested. Housing value on the other hand is negatively correlated with mortgage rejects, and positively correlated with flood insurance claimed

In order to prepare the data for running in machine learning models, we need to also make sure all downloaded data are standardized with at least one standard identifier. We chose county level zip-codes, which are present on all datasets. For the seven years of HMDA data we gathered between 2010 and 2017, we converted the

HMDA's reporting of census tract into standardized zip codes using another third party translation dataset that detailed the relationships between state level census tracts and state zip codes, that would otherwise be duplicated across state boundaries. After removing confusion, we standardized the available dataset and normalized mortgage loan values, application and rejection numbers all by standardized zip code. The data sets are large enough to stress our computing resources, especially on system memory, but we were able to complete the zip code standardization process eventually.

For all the other non-HMDA sourced data, we first removed all empty rows in required columns, aggregated by US zip code and merged the various datasets on common state, city, county, zip code and census tracts, and merged them with HMDA data. After merging the data, we took sum, count and mean for fields depending on the nature of the relevant fields of data. For the loan amount, we extracted the sum based on their zip code. For the sea level rise in data, housing price, we took the mean per zip code, and for mortgage applications and rejection data we found out the count per zip code.

However, before moving the data into our machine learning model, we first identified the distribution of the merged data to see if the shape of the data is suitable for serving as the training model for the machine learning algorithm. The distribution histogram of the training shows that the data is skewed to the right, which is unsuitable for serving as the base for machine learning models. The closer they are to a normal distribution, the better suited the data are for machine learning models. To solve this issue, we first applied an automatic standardization algorithm where we aim to bring the data close to a normal distribution. After examining the intermediate output of the data, we further standardized the distribution by removing outliers beyond the 99th percentile. After multiple manual cleaning runs, we aim to arrive at an end result of distribution that looks much closer to a normal distribution instead of skewing heavily towards the right. We can finally move to creating a suitable machine learning model.

Part IV. Analysis

Based on the information gathered after cleaning and visualizing the national datasets, we can start regression modeling between the relationships between climate change vulnerability factors (such as rise in sea level, distance to coast, and previous federal flood insurance claims) and cost of housing ownership (housing cost and mortgage rejection numbers). Between these pairs we aim to examine which machine learning methods are the most effective, and have the least overfitting and extraneous errors. After dividing our dataset into training and test models, we first applied a simple linear regression method between loan rejection numbers, and vulnerability index (set by a combination of rise in sea level, distance to coast, and flood insurance claims), and housing prices and the vulnerability index and examined its R-squared value. The R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. As the R-squared value for both correlation pairs are very low (close to 0), this linear regression model does not explain much of the variance in the dependent variable, and none of the independent variables

('seaLevelChange', 'distance', 'floodProneDegree') seem to have a statistically significant effect on the dependent variable, as indicated by their p-values (which are both above 0.05). Hence, we realized that the linear regression method is not sufficient in determining whether climate change vulnerabilities would have an impact on loan rejection numbers or housing value changes. Figure 4A-4C demonstrates the scatterplot and associated linear regression model between housing value and the vulnerability indexes. Figures 4D and 4E exhibit the lack of relationship between loan rejection numbers and climate change vulnerabilities.

After trying simple linear regression methods, we have also attempted random forest methods for the same correlation sets. In a basic random forest method, the Mean Square Error (MSE) measures the average squared differences between predicted values and actual values and the resulting MSE of 0.101 and 0.485 for loan rejection and housing value respectively indicates that the model's predictions are close to the actual values on average.

At the same time, Feature importances reveal the significance of each feature in making accurate predictions. In our model, distance has the highest importance, followed by floodInsuranceClaimed, while seaLevelChange has relatively lower importance. This insight helps understand how the model utilizes different features for prediction. Figures 4F to 4H show the correlation between housing value and climate change vulnerabilities, with its feature importance being 0.452 0.345 0.201 respectively, while having a low MSE value of 0.101 in the model. the feature importance of each variable shows that the sea level rise contributes the most to housing value while the distance is the main factor of number of loan rejection.Both sets of illustrations demonstrate that this random forest model is more accurate as a regression model than linear regression, and is able to produce housing value and loan rejection count in the face of mounting climate change vulnerabilities more accurately than simple linear regression.

Part V. Conclusion and Further Recommendations

The Analysis result concludes that our intuition where increasingly austere climate change events have had an upward pressure in mortgage rejection and overall housing cost in coastal metropolitan areas. It is regrettable that our national data set on rent and federal insurance claims lacks further granularity. Rental index data was less detailed compared to our mortgage data, and federal insurance claims contained multiple rows of "NaN" reasons for the floods, as compared to some non-null values dating to the exact hurricane names. There are further directions of research, where we could potentially filter our mortgage/rental price movement normalized by inflation, or we could have separated coastal areas by population density, and we could have combined population changes in the timeframe we considered. Either of these directions would

yield a more accurate picture between climate change and cost of housing, adjusted for inflation and population congregation and density.

References

Dahl, Kristina A., et al. “Effective inundation of continental United States communities with 21st century sea level rise.” *Elementa: Science of the Anthropocene*, vol. 5, 2017, https://doi.org/10.1525/elementa.234.

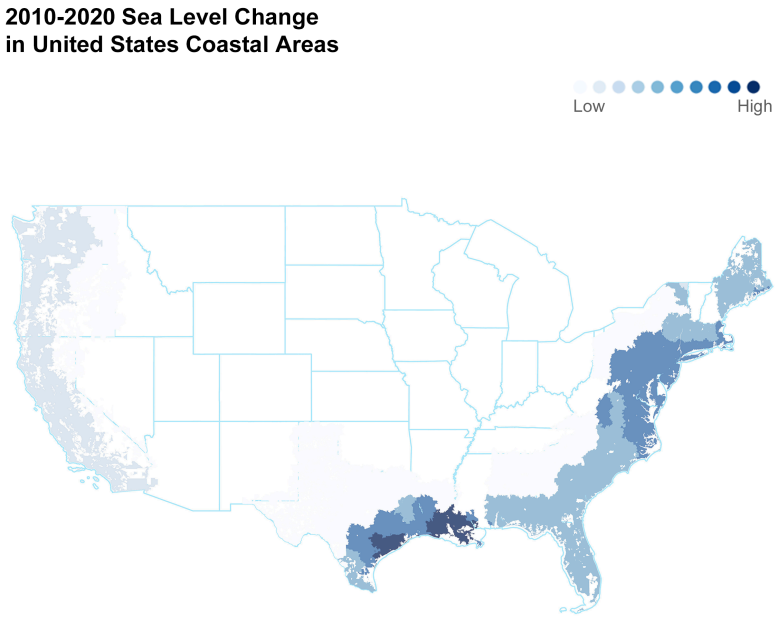
Flavelle, Christopher. “Rising Seas Threaten an American Institution: The 30-Year Mortgage.” *The New York Times*, The New York Times, 19 June 2020, www.nytimes.com/2020/06/19/climate/climate-seas-30-year-mortgage.html?smid=em-share.

S & P Case Studies. “A Bank Evaluates the Impact of Physical Climate Risk to Its Mortgages.” *A Bank Evaluates the Impact of Physical Climate Risk to Its Mortgages | S&P Global*, S & P Global, 5 May 2022, www.spglobal.com/esg/case-studies/a-bank-evaluates-the-impact-of-physical-climate-risk-to-its-mortgages.

Contribution Table

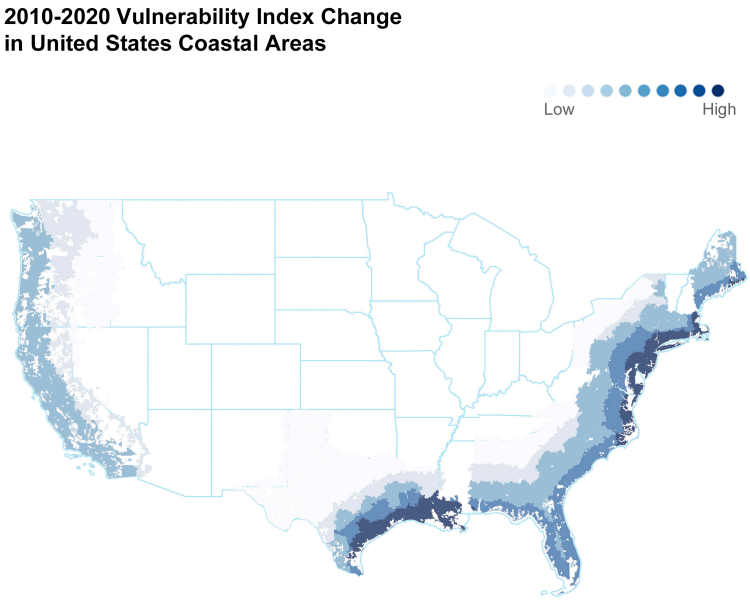
Team Member	Responsibilities
Junyi Li	Data Cleaning Distributions and Analysis of ML Dataset Data Visualization Machine Learning Modelling Video Presentation
Swati Sharma	Data Collection Data Cleaning Correlations and Analysis of Mortgage Data Data Visualization Machine Learning Modelling Video Presentation
Sneha Tirchy Shekar	Data Collection Geospatial Analysis of Housing & Pop. Data Data Visualization Video Presentation
Vinayak Saxena	Data Collection Geospatial Analysis of Rising Sea Level Vulnerability Index Calculation Data Visualization Video Presentation
Zhexuan Franklin Tang	Data Collection Data Aggregation Report Writing/Editing Presentation Editing Video Editing
Zheyan Chen	Data Cleaning Analysis of Housing Data Data Visualization Report Writing/Editing Presentation Editing

Appendix of Images



Source: National Oceanic and Atmospheric Administration

Figure 3A Sea Level Change in the United States Coastal Area



Source: National Oceanic and Atmospheric Administration

Figure 3B Vulnerability Index

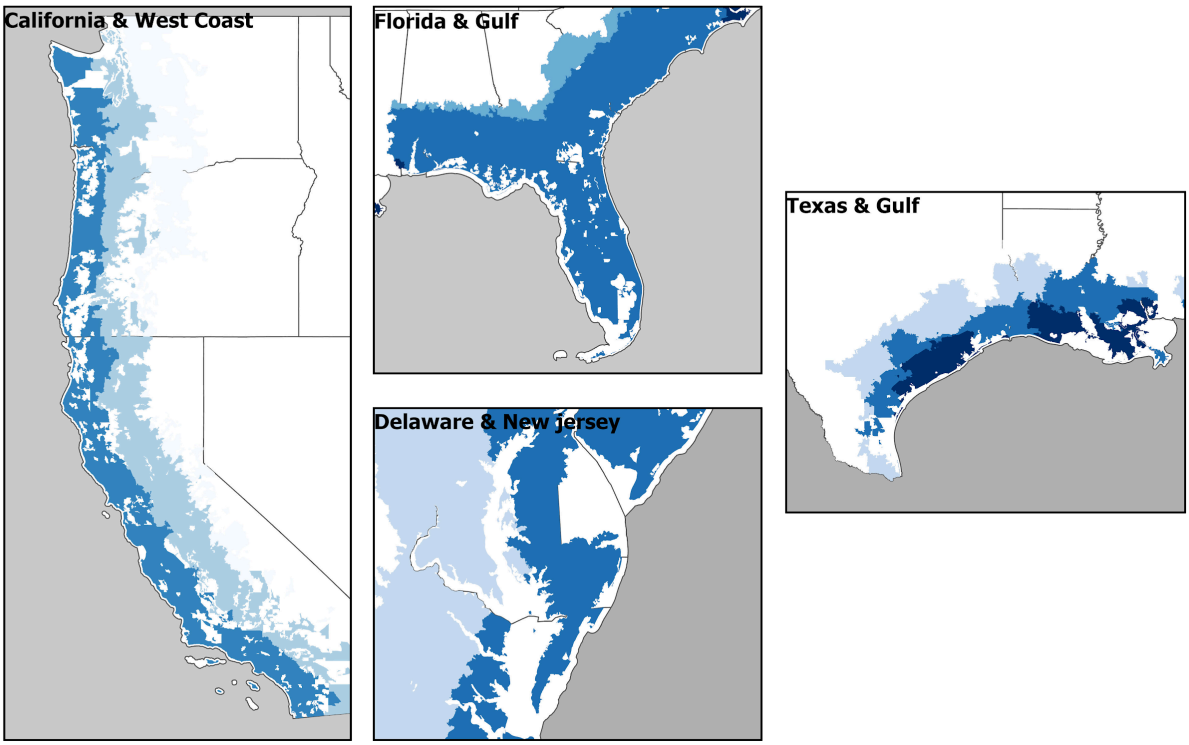


Figure 3C/3D. At Risk Regions for Flooding by Sea Level Rise 2010-2020

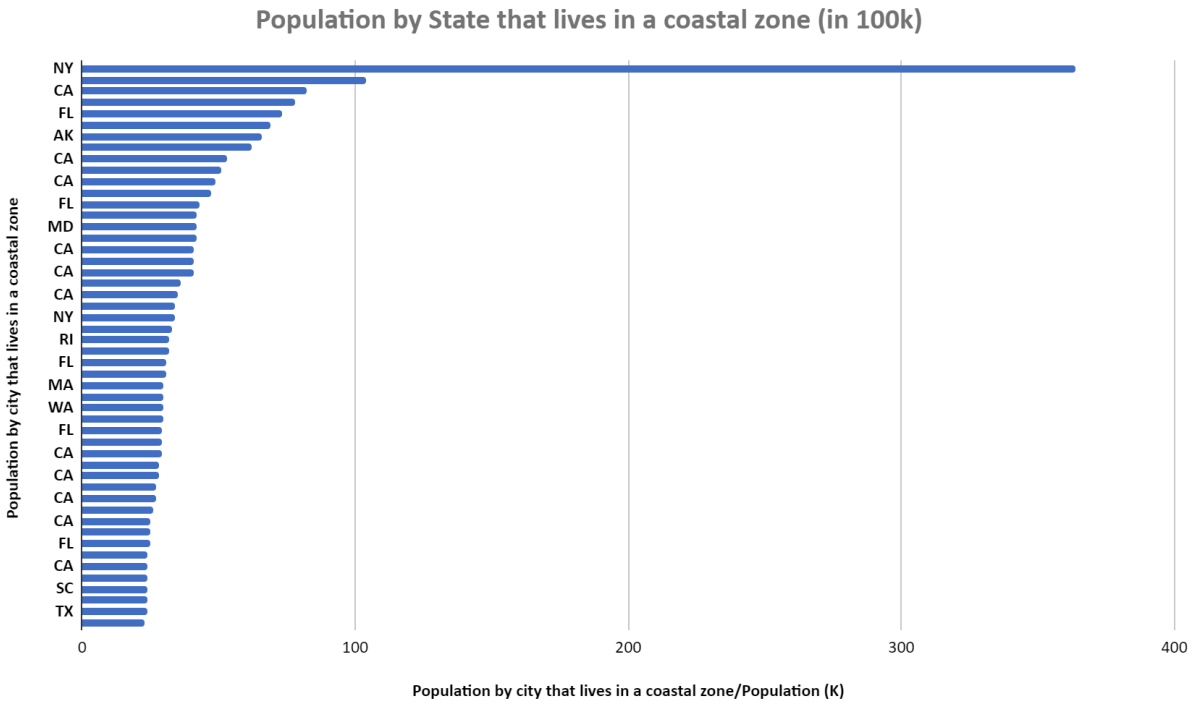


Figure 3E Population by State that lives in a coastal zone

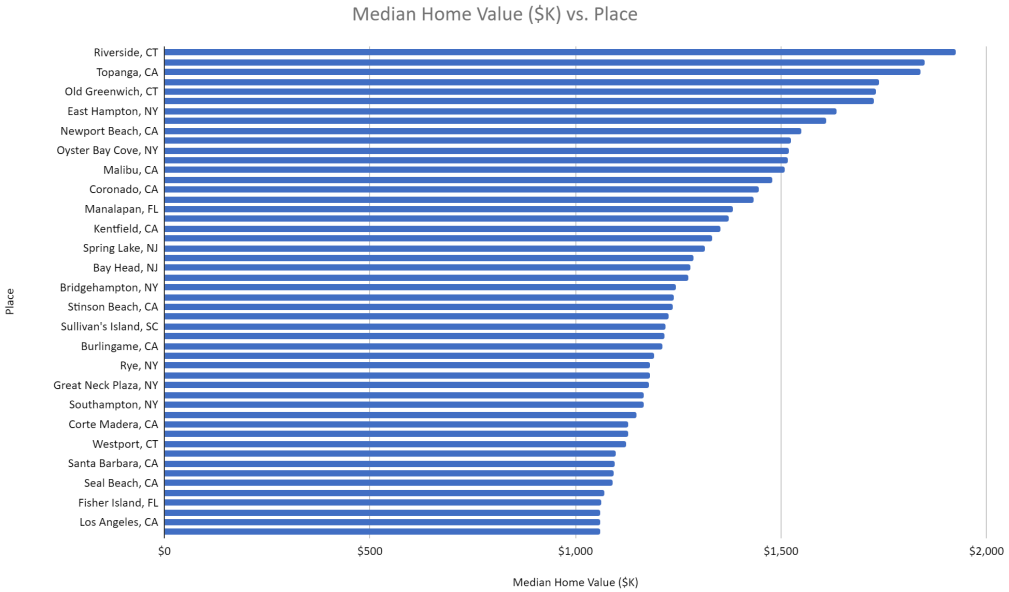


Figure 3F Median Home Value of Coastal Area

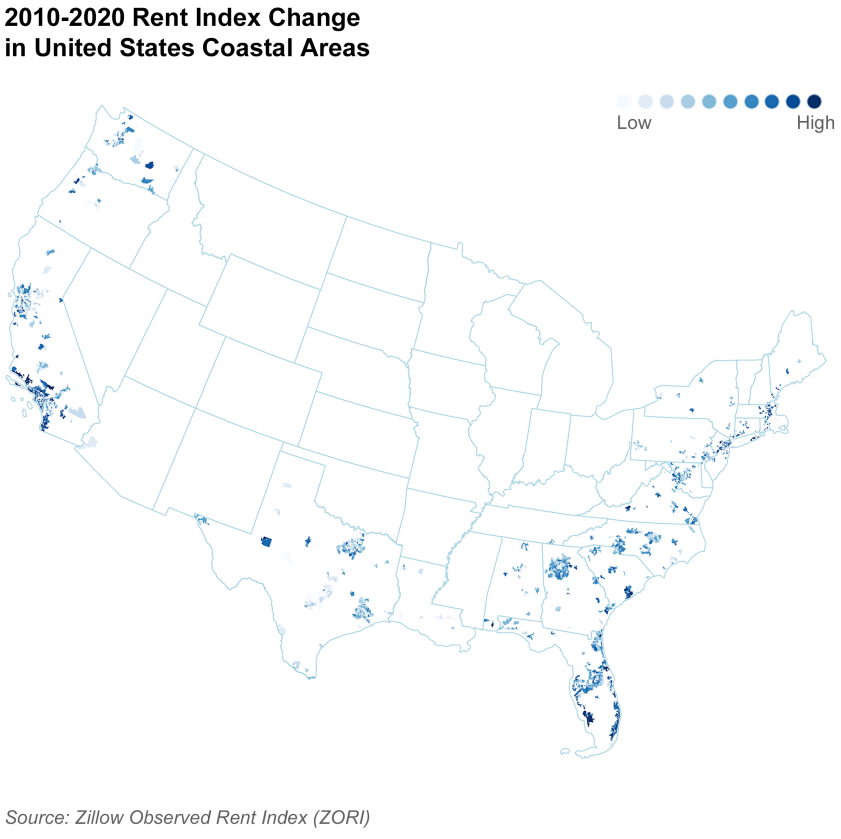


Figure 3G Rent Index Change in United States Coastal Area

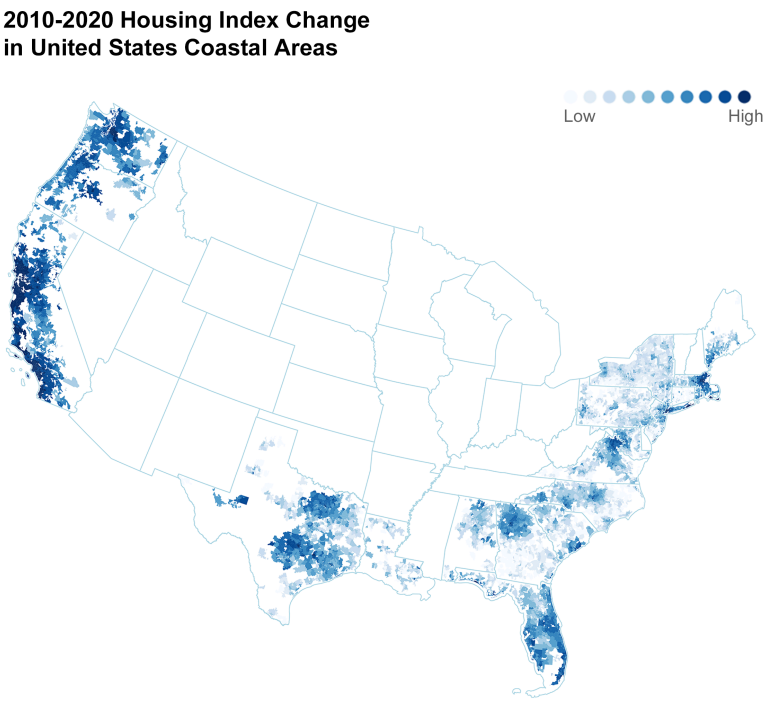
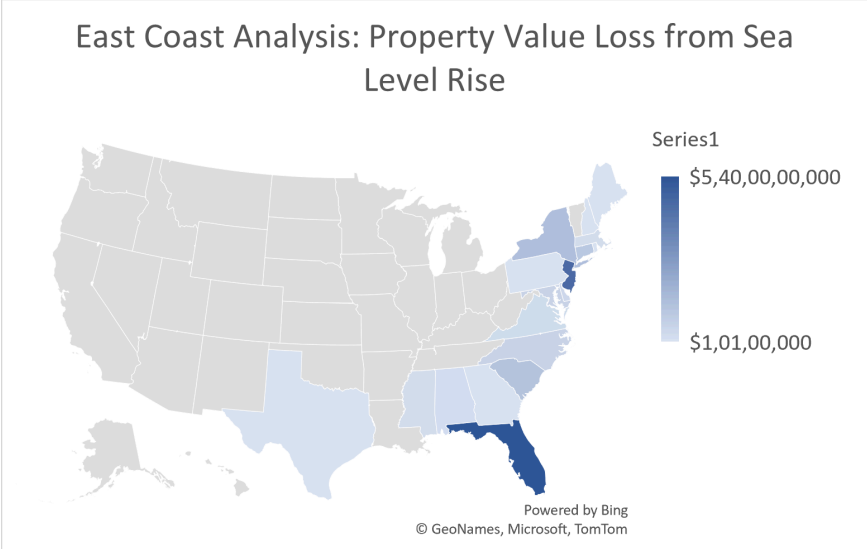
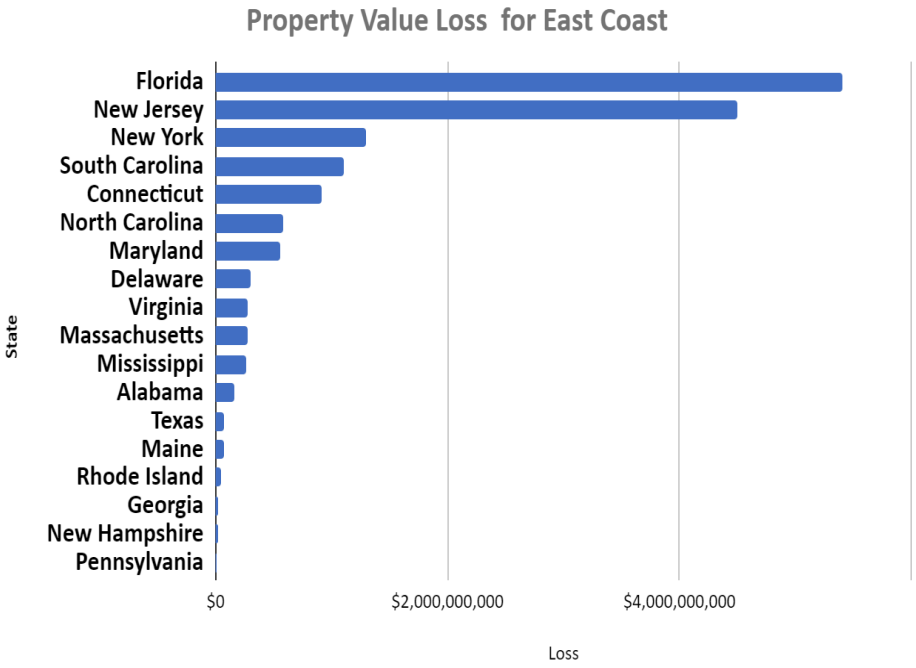


Figure 3H Housing Index Change in United States Coastal Area





Figures 3I & 3J Property Value Loss from Sea Level Rise

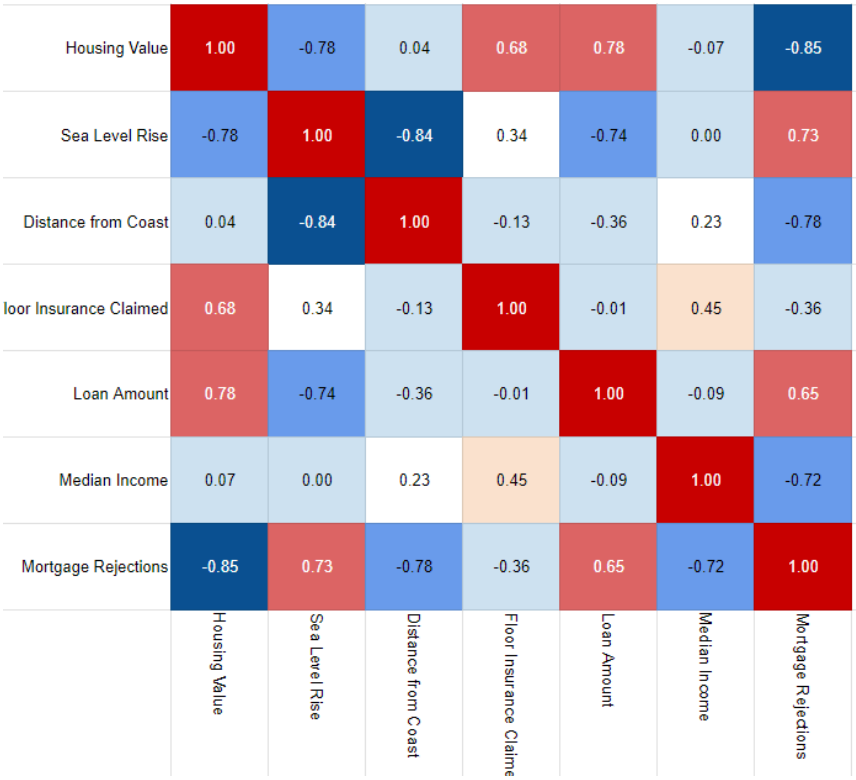
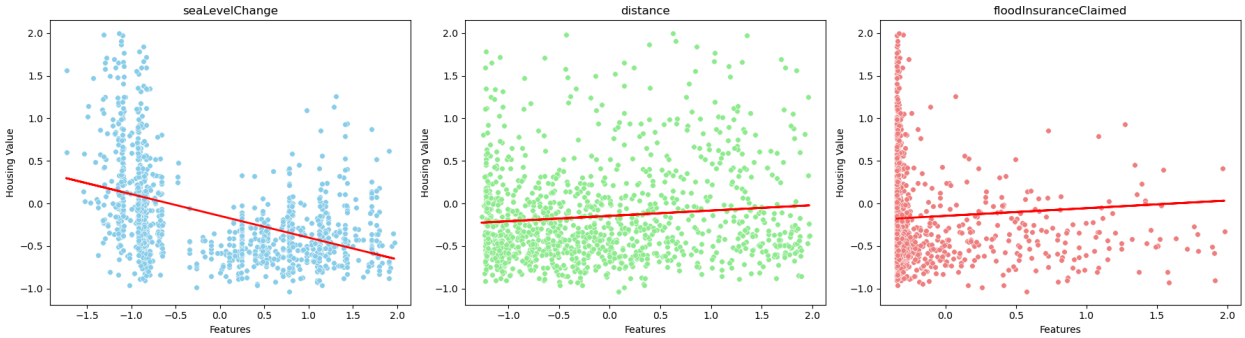
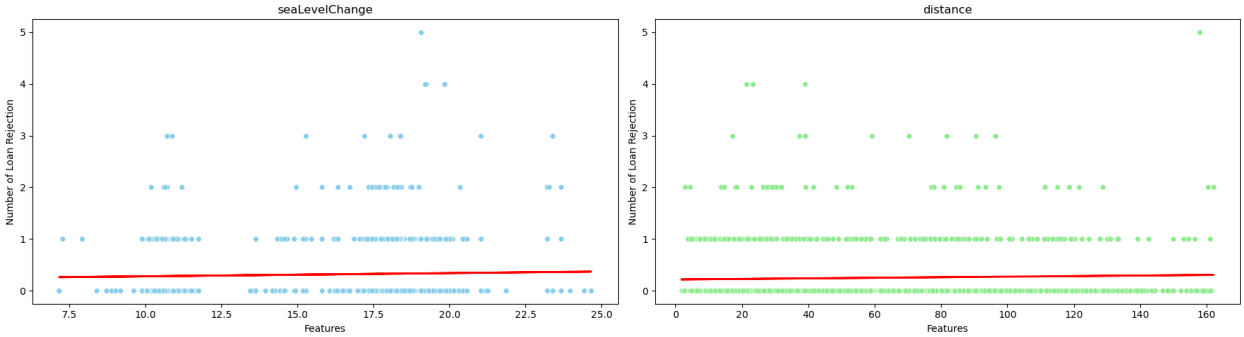


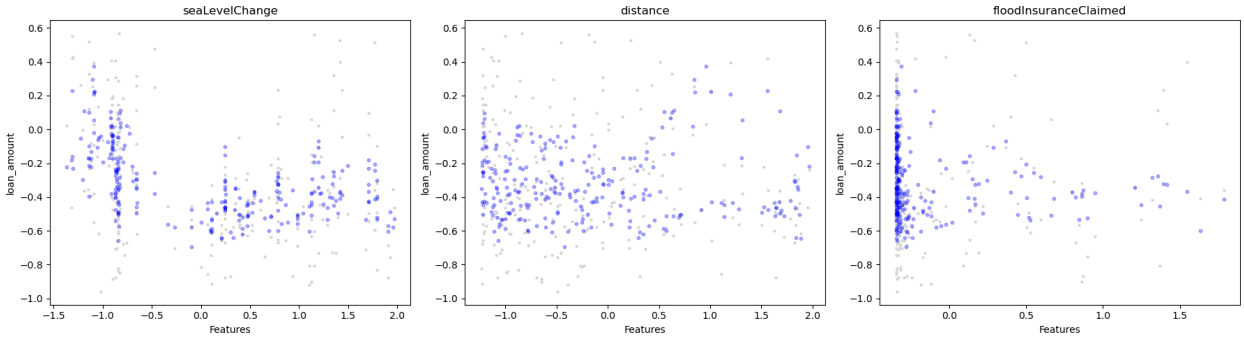
Figure 3K. Correlation Matrix of all main datasets.



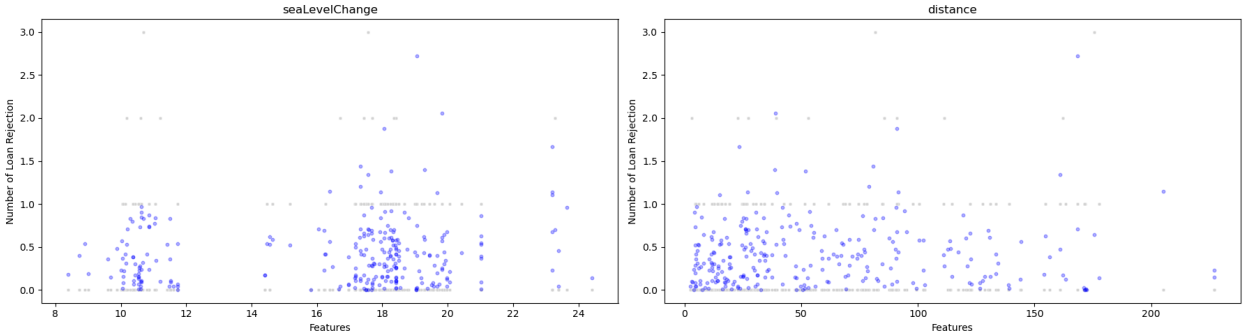
Figures 4A, 4B, 4C. Linear Regression results in housing value.



Figures 4D, 4E. Linear regression results on loan rejection numbers.



Figures 4F, 4G, 4I. Random forest Method in housing value.



Figures 4J, 4K, Random forest Method in loan rejection numbers.