Final Report: Coastal Vulnerability Risk Assessment through Data Scraping and Geospatial Visualization

Completed by:

Wila Mannella wmannell@usc.edu **USC ID: 749 785 2853**

Sofia Young sofiayou@usc.edu **USC ID: 746 883 1224**

Short Description

California contains more than three thousand four hundred miles of shoreline that support millions of residents and trillions of dollars in infrastructure. These areas face increasing climate-related threats, especially sea level rise. Scientific research on sea level rise is abundant, yet it is scattered across agencies and formats that are not always accessible to the public. A tool that consolidates these data in a clear and intuitive way can support more informed conversations about coastal risk and the need for resilient planning. This project addresses that need by scraping key datasets, calculating a normalized risk score, and creating three visual outputs: an interactive map of the California coastline displaying risk levels for each region, a bar graph that compares risk factors among stations, and a scatterplot that examines the relationship between mean housing values and predicted sea level rise. Together, these outputs offer an approachable resource that shows how vulnerable different coastal locations may become throughout the century.

Data

The project required information on both ocean conditions and coastal economic characteristics. To characterize ocean dynamics, we retrieved raw tide station data from NOAA, including station identification information, latitude and longitude, state indicators, and water level time series for the previous thirty days.

To examine property values, we used the California housing census dataset collected in 1990. More recent datasets were inaccessible due to paywalls, but the 1990 dataset still provided reliable spatial patterns that describe how housing values vary across the state. We extracted housing values using the fetch_california_housing method in the sklearn.datasets package and calculated average regional housing values for properties located near each tidal gauge station. Location matching was performed using latitude and longitude coordinates.

All NOAA data were accessed through URLs that returned information in JSON format. Rather than using tools such as Beautiful Soup or ElementTree, we created a function called safe_json that parsed the JSON directly, stored the results, and saved them to CSV files. This procedure was used for both the tidal gauge station metadata and the water level time series, which came from two separate NOAA CO-OPS API endpoints but were ultimately merged into a single CSV. From this procedure we collected 7200 data points from each NOAA tidal station, totaling over 120,000 data points.

For housing data, the fetch_california_housing function loaded the dataset directly into a pandas DataFrame. This provided housing values and corresponding geographic coordinates. From this data, we were able to extract 20,640 data points.

Data Cleaning, Analysis and Visualization

Cleaning and preprocessing were essential before any analysis could occur. The combined NOAA dataset was first loaded with the pandas readcsv function to create tide_df. Because the NOAA API initially returns all United States stations, we filtered the DataFrame so that only rows with the state indicator set to CA remained. Water level values were converted into numeric form with the pd.to_numeric method.

Time series refinement involved converting timestamp strings into datetime objects with pd.to_datetime and removing all null values with dropna. The calculation of sea level rise trends required transforming datetime values into seconds using dt.total_seconds. These values were used by a custom function called compute_trend, which applied np.polyfit to generate an estimate of sea level rise in meters per second. This value was later converted to meters per year to align with typical reporting conventions.

To prepare the housing dataset, we isolated the single column required for the risk score, median housing value, and created a reduced housing DataFrame called housing_df. This DataFrame was then merged with tide_df to create a combined dataset that contained both economic and oceanographic variables. Latitude and longitude values were maintained so that each station could be plotted accurately.

For the folium map, only latitude and longitude columns derived from the tide data were necessary. Therefore, the combined DataFrame was trimmed to retain only the fields needed to calculate and visualize risk scores.

The combined dataset supported three primary analyses. First, we calculated a normalized risk score for each station that incorporated both sea level rise trends and local housing values.

Risk score was calculated as: Risk score = sea level trend in m/year * housing value index (0-5 score indicating the cost of nearest housing). Sea-level trends were calculated using short-term water-level records to ensure data availability across stations. This score was recorded alongside supplementary metrics on an interactive folium map. The map provides an intuitive way for users to examine spatial patterns in risk across the state. Second, we created a bar graph that compared the risk scores among stations. This visualization allowed direct comparison of vulnerability levels and helped identify outliers or clusters of high risk. Third, we produced a scatterplot which plotted mean housing value against predicted sea level rise. This graph showed the relationship between economic exposure and environmental change and highlighted regions where valuable properties may experience significant future threats.

The premise of this project was to develop an accessible, data driven tool that enables stakeholders to better understand the risks posed by sea level rise to current and future housing development along the California coastline. California’s coast supports dense population centers

and high value infrastructure, yet information describing sea level rise trends, exposure, and vulnerability is often fragmented across technical sources. The underlying hypothesis was that consolidating publicly available oceanographic and housing data into a single analytical framework and presenting the results through intuitive visualizations would reveal meaningful spatial patterns of risk and support clearer interpretation by planners, policymakers, and the public.

The analysis supports this hypothesis. By integrating NOAA tidal station data with regional housing values and translating these inputs into a normalized risk score, the project demonstrates that risk from sea level rise is not evenly distributed along the coast. The interactive map shows clear geographic variability, with certain regions consistently exhibiting higher relative risk due to the combined influence of rising sea levels and concentrated housing value. This spatial visualization allows users to quickly identify areas where exposure may be particularly consequential, offering insight that is difficult to obtain from raw data tables alone.

The bar graph further reinforces these findings by enabling direct comparison of risk among stations. Differences in risk scores highlight how local sea level rise trends and economic exposure interact, underscoring that areas with moderate physical change can still face substantial overall risk when housing value is high. Conversely, some locations with higher rates of sea level rise may present comparatively lower immediate economic exposure. This comparative perspective emphasizes the importance of considering both environmental and socioeconomic variables when assessing coastal vulnerability.

The scatterplot examining mean housing value in relation to predicted sea level rise provides additional context for future planning. While the relationship is not perfectly linear, the visualization illustrates that many high value coastal areas are already experiencing measurable sea level rise. This reinforces the conclusion that continued development in these regions carries increasing long term risk, particularly if adaptation or retreat strategies are not incorporated into planning decisions.

Overall, the project demonstrates that transforming complex datasets into a unified, visual narrative meaningfully improves interpretability and relevance for stakeholders. The conclusions drawn from the analysis suggest that tools like this can support more informed conversations about coastal development, risk management, and adaptation planning. By making sea level rise risk more tangible and location specific, this project contributes to a broader effort to align scientific understanding with practical decision making along California’s coast.

Deviations from Original Proposal

Our proposal went through several rounds of back-and-forth revisions and iterations with the TA, who advised us to firstly include more detail in our data assertion methodology and then to secondarily add more types of visualizations into our final product. For our second project proposal after the first round of revision, we more specifically defined exactly what API we would use to scrape the NOAA data, what databases it was coming from, and how we planned to extract it. We also went into more detail regarding the housing census data we collected, but

that was a less complex database. For our third proposal we added more types of visualizations in addition to the interactive California coastal risk map, which displayed the calculated risk for various locations across the coastline. We added in a bar graph to compare the risk scores between stations for easy comparative analysis. We also added in a scatterplot of mean housing value against predicted sea level rise, making it easy to identify properties of high risk in the future. Both of these changes have resulted in our final project not having any deviations from this 3rd project proposal.

Mention of Future Work

The calculations of risk score, the water-level time series, and the housing data were all very rudimentary. To improve the accuracy of these calculations, the water level timeseries can include more data scraped from other sources that would allow for a calculation using more than the past month of water level trends. Alternatively, we could pull the trend as calculated by NOAA from another source and skip the manual calculation of the trend. Regarding the housing data, the data comes from the 1990 census, which was the most robust and recent dataset containing the housing data we needed. To improve this project we could scrape housing data from websites like Zillow or Redfin to represent more recent housing trends in the state. For risk score, if we made the improvements to the water-level trend calculation so that it was representative of longer-term trends, the risk score could be calculated directionally so that sea-level rise would increase risk while sea-level declines would lower risk. Currently, our risk score is normalized by magnitude, which means the higher the trend in sea-level change in any direction, the higher the risk score. With long-term trends, the calculation of risk score could instead consider sea-level rise as increases in risk and declines as reductions in risk. This would likely be a more conducive understanding of risk.

Conclusion

This project demonstrates the value of synthesizing public data into accessible and meaningful formats. By scraping, cleaning, and analyzing information from NOAA and the California housing census, we created visual tools that illustrate both physical and economic dimensions of coastal vulnerability. These tools can inform public understanding, support community conversations, and contribute to broader efforts to manage risk from climate driven sea level rise.