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

Introduction

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, particularly from sea-level change. Scientific research on sea-level dynamics 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, relative 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 observed sea-level trends. Together, these outputs offer an approachable resource that illustrates how vulnerability varies along the California coast based on both environmental change and economic exposure. Ultimately, this was produced more as a practice of the data acquisition and analysis techniques covered in the DSCI-510 course than as a finalized tool that accomplishes what we set out to accomplish. Improvements that could be made and further directions for this work are included in this final report that explores the steps needed to move this project into a more finalized state.

Data Collection

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 describing how housing values vary across the state. Housing values were accessed using the fetch_california_housing method from the sklearn.datasets package. Median housing values were associated with each tidal gauge station by matching locations using latitude and longitude coordinates.

All NOAA data were accessed through API endpoints that returned information in JSON format. Rather than using HTML parsing tools such as Beautiful Soup or ElementTree, we created a custom function to safely retrieve and parse JSON responses, store the results, and save them to CSV files. This approach was used for both tidal gauge station metadata and water-level time series, which originated from separate NOAA CO-OPS API endpoints but were ultimately integrated into a unified dataset.

From this process, we collected several thousand water-level observations per NOAA tidal station, resulting in over one hundred thousand total data points across all stations. Housing data were loaded directly into a pandas DataFrame, yielding over twenty thousand spatially referenced housing observations.

Data Cleaning, Analysis, and Visualization

Cleaning and preprocessing were essential before analysis could occur. The NOAA station metadata and water-level datasets were loaded into pandas DataFrames. Because the NOAA API returns stations across the entire United States, the data were filtered so that only rows associated with California remained. Water-level values were converted to numeric form using pd.to_numeric, and invalid or missing values were removed.

Time series refinement involved converting timestamp strings into timezone-aware datetime objects using pd.to_datetime, followed by the removal of null values. To calculate sea-level trends, datetime values were transformed into elapsed time in seconds. These values were passed to a custom trend calculation function that applied numpy.polyfit to estimate linear sea-level change over the thirty-day observation window. The resulting slope values were converted from meters per second to meters per year to align with standard reporting conventions.

For the housing dataset, only the variables necessary for the analysis were retained, specifically median housing value and geographic coordinates. A reduced housing DataFrame was created and spatially linked to the tide station dataset using a nearest-neighbor search implemented with a KDTree. This ensured that each tide station was paired with the closest available housing data point.

The combined dataset supported three primary analyses. First, a normalized risk score was calculated for each station using the calculated sea-level rise trend and the normalized median housing value of the nearest housing location. Sea-level trends were calculated using short-term water-level records to ensure consistent data availability across all stations. This presented a significant limitation of the project that is discussed in a later section. This risk score is intended as a relative and conceptual measure of vulnerability rather than a predictive hazard metric.

The risk score and associated metrics were displayed on an interactive Folium map, allowing users to explore spatial patterns of relative risk along the California coastline. Second, a bar graph was created to compare risk scores across stations, enabling direct visual comparison of relative vulnerability. Third, a scatterplot examined the relationship between median housing value and observed sea-level trend, highlighting areas where economic exposure coincides with measurable environmental change.

Interpretation of the Results

The premise of this project was to develop an accessible, data-driven tool that enables stakeholders to better understand relative coastal vulnerability along the California coastline. The underlying hypothesis was that consolidating publicly available oceanographic and housing data into a single analytical framework and presenting results through intuitive visualizations would reveal meaningful spatial patterns of risk. 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 relative vulnerability is not evenly distributed along the coast. The interactive map reveals clear geographic variability, with certain regions exhibiting higher relative risk due to the combined influence of short-term sea-level variability and concentrated housing value.

The bar graph further reinforces these findings by enabling direct comparison of risk scores among stations. Differences in risk scores illustrate how environmental variability and economic exposure interact, emphasizing that areas with moderate physical change may still face elevated relative risk when housing value is high. Conversely, locations with larger short-term sea-level fluctuations may exhibit lower relative economic exposure.

The scatterplot provides additional context by illustrating the relationship between housing value and observed sea-level trends. While the relationship is not strictly linear, the visualization highlights that many high-value coastal areas are already experiencing measurable sea-level variability, underscoring the importance of considering both physical and economic dimensions in coastal planning.

Deviations from the Original Proposal

The project underwent several iterations following feedback from the teaching assistant. Early revisions focused on clarifying data acquisition methodology, including the specific NOAA APIs used and how housing data would be incorporated. Later revisions expanded the scope of visualization beyond an interactive map to include comparative bar charts and scatterplots. These revisions were fully incorporated into the final implementation, and the completed project does not deviate from the final approved proposal.

Future Directions

The calculations of risk score, water-level trends, and housing exposure in this project are intentionally simplified. Future work could improve the accuracy of sea-level trend estimates by incorporating longer time series or by using NOAA-provided long-term trend products instead of short-term linear approximations. Similarly, housing data could be updated using contemporary sources such as Zillow or Redfin to better reflect current market conditions.

With longer-term trend data, the water_level trends would be more reflective of the actual long-term positive trends expected for sea level rise. In its current state, the trends used to calculate risk score do not represent the actual sea-level rise trends in m/year, and this resulted in numerous negative sea-level rise trends and thus negative risk scores. In the future, an alternative source could be used to get the data used to calculate the timeseries, one that allows for a longer period than 30 days of water-level timeseries data to be downloaded.

Limitations

This project has several limitations that should be considered when interpreting the results. Sea-level trends were calculated using approximately thirty days of water-level data, which may reflect short-term variability rather than long-term sea-level rise. As a result, some calculated trends are negative or exaggerated in magnitude. To address this, the analysis treats the risk score as a relative and conceptual metric rather than a predictive measure. Even so, the risk score is likely an inaccurate representation of risk in its current form and can be addressed if this project moves forward.

Housing data were obtained from the 1990 California housing census due to limited access to more recent datasets. While this dataset provides reliable spatial patterns, it does not represent current market conditions. Housing values were therefore used as a proxy for relative economic exposure.

Additionally, the spatial association between tide stations and housing data relies on a nearest-neighbor approach using geographic coordinates, which does not account for local coastal features or infrastructure. Finally, the risk score does not incorporate other important factors such as population density, coastal defenses, or long-term climate projections.

These limitations were acknowledged throughout the analysis, and results were interpreted within the scope of the available data and methods.

Conclusion

This project demonstrates the value of synthesizing publicly available data into accessible and interpretable formats. By scraping, cleaning, and analyzing information from NOAA and the California housing census, we developed visual tools that illustrate both environmental variability and economic exposure along the California coast. These tools support clearer interpretation of relative vulnerability and provide a foundation for more informed discussions about coastal development, adaptation, and risk management in the context of climate-driven sea-level change.