# Methodological Approach for Flood Risk Regression Analysis

This document outlines the methodological approach employed over the last week to perform regression analysis on flood risk across New York City (NYC) and the Bronx. The aim was to evaluate the relationship between flood risk and demographic/industrial characteristics.

## 1. Data Preparation

- \*\*Data Collection:\*\* Acquired zoning, demographic, and flood risk data for NYC neighborhoods using public datasets and existing project repositories.

- \*\*Data Cleaning:\*\* Loaded data into Python using libraries like `pandas` and `numpy`. Checked for missing values, outliers, and inconsistencies, particularly in zoning and demographic variables. Normalized the data for consistency, ensuring all numeric variables were scaled appropriately and categorical variables were encoded if necessary.

- \*\*Filtering:\*\* Extracted data for NYC as a whole and filtered for Bronx neighborhoods using Neighborhood Tabulation Area (NTA) codes. Segregated flood risk and demographic data by geographic regions for focused analysis.

## 2. Variable Selection

- \*\*Dependent Variable:\*\* `mean\_flood\_risk` (average flood risk for each neighborhood).

- \*\*Independent Variables:\*\*  
 - `industrial\_proportion` (percentage of industrial land use in the area).  
 - Demographic proportions:  
 - `White alone\_proportion`  
 - `Black or African American alone\_proportion`  
 - `Asian alone\_proportion`

Variables were chosen based on their theoretical relevance to flood risk and environmental justice considerations.

## 3. Data Aggregation

- Grouped data by NTAs to compute neighborhood-level means and proportions for selected variables.

- Aggregated flood risk data using statistical summaries (e.g., mean values per neighborhood).

## 4. Regression Analysis

- \*\*Model Specification:\*\* Used Ordinary Least Squares (OLS) regression to model the relationship between `mean\_flood\_risk` and independent variables.

- \*\*Steps:\*\*  
 - Implemented the regression using the `statsmodels` library in Python.  
 - Fit the model separately for NYC (citywide analysis) and Bronx (localized analysis).  
 - Extracted key statistics, including R-squared, adjusted R-squared, F-statistic, and p-values.  
 - Assessed the significance and direction of coefficients for each variable.  
 - Generated warnings for small sample sizes in the Bronx regression and addressed these as potential limitations.

## 5. Results Interpretation

- \*\*For NYC:\*\* Identified significant relationships between demographic variables (e.g., `White alone\_proportion`) and flood risk. Found industrial land proportion to be statistically insignificant.

- \*\*For Bronx:\*\* Observed a significant positive relationship between `Asian alone\_proportion` and flood risk. Noted reduced statistical power due to the limited number of observations (n=8).

- Highlighted discrepancies in results across scales (citywide vs. local) and explored potential drivers of these differences.

## 6. Documentation and Reporting

- Logged all data processing and analysis steps in a Python script (`regression.py`) for reproducibility.

- Compiled regression outputs, observations, and limitations into a structured format for review.

- Addressed methodological caveats, including sample size concerns and potential multicollinearity.