Project Title: Modeling the Determinants of Housing Prices Using Kaggle Data

Team Member: Lihui Xiong (7345487924)

What problem are you trying to solve?

Housing prices are influenced by many factors such as location, size, and neighborhood characteristics, but it is often unclear which features are most important and how strongly they are associated with price. This project aims to investigate the relationship between residential house prices and a set of potential explanatory variables using the dataset.

How will you collect data and from where?

I plan to use an existing open-source housing dataset from Kaggle, the "House Prices – Advanced Regression Techniques" dataset. This dataset contains sale prices for a large number of residential homes along with many potential predictors, including: living area square footage, lot size, number of bedrooms and bathrooms, garage size, year built, overall quality scores, neighborhood, and other house condition ratings. The data will be downloaded directly from Kaggle as a CSV file and imported into Python.

Revised Method:

Based on the instructor's feedback, I revised the data collection approach to meet the course requirement of demonstrating programmatic web-based data acquisition. Instead of downloading a pre-packaged CSV dataset from Kaggle, I now collect housing transaction data directly through the NYC Open Data Socrata API using Python. This revised approach involves writing API requests to specify fields, apply filters, and retrieve raw JSON data, which is then cleaned and processed before analysis.

What analysis will you do and what visualizations will you create?

Correlation analysis: Compute and visualize the correlation matrix between numeric variables to identify strong linear relationships with sale price.

Feature selection and modeling: Build regression-based models.

Planned visualizations include:

- Histograms and boxplots of sale prices and key predictors.
- Scatter plots of sale price vs. continuous variables with trend lines.
- Boxplots of sale price across categorical variables such as neighborhood or overall quality rating.