COMPSCI 1090A / AC209A

Exploring Determinants of Housing Prices Using Regression and PCA

Group Members:

- InJoo Kim (ink595@g.harvard.edu)
- Yao Xiao (yaoxiao@g.harvard.edu)
- Rolando Desantiago Jr. (rod316@g.harvard.edu)
- Julia Duro (julia_jimenezduro@gsd.harvard.edu)

Background and Motivation:

Housing prices are influenced by a wide range of factors, from property-level features (size, year built, condition) to broader neighborhood and city-level dynamics (land use, accessibility, demographics, and transit).

Our motivation is to identify and quantify the most important drivers in a clear and interpretable way using regression and PCA, methods that we have studied in class. By focusing initially on a single city (e.g., San Francisco), we can take advantage of well-curated public data while keeping the analysis tractable.

This project is also timely: San Francisco faces a major housing affordability crisis and upcoming policy changes (rezoning, density allowances). Our work could provide insights into which factors most strongly shape housing values, potentially linking to real policy debates.

Data:

- Zillow Research Data – housing prices, property characteristics (e.g., square footage, year built, number of bedrooms).
    - https://www.zillow.com/research/data/

- San Francisco Open Data – neighborhood-level datasets such as zoning, land use, building permits, and transportation networks.
    - https://data.sfgov.org/

- American Community Survey (ACS) – socio-demographic data at the tract or block group level (income, education, household size).
    - https://www.census.gov/programs-surveys/acs

- Additional layers:
    - Amenities from OpenStreetMap (e.g., green spaces, schools, commercial areas).
    - Transit accessibility (distance/travel times to employment centers or schools).

Scope:

- Begin with exploratory data analysis (EDA) to understand distributions and identify candidate predictors.
- Use multiple linear regression to quantify the impact of property-level and neighborhood-level factors on housing prices.
- Apply PCA to reduce dimensionality when including many correlated variables (e.g., different measures of accessibility or socio-economic indicators).
- Initial focus will be on a few major drivers (property details, accessibility, demographics). We can expand to additional layers (e.g., zoning, transit, amenities) if time allows.
- Deliverables: interpretable model outputs, visualization of key relationships, and discussion of policy-relevant insights.