Codebook for Real Estate California Project
General Description
Dataset Source: Kaggle: Real Estate California
Time Period: First 6 months of 2021
Observations: 35,389
Variables: 39

Variable Descriptions
Target Variable:
price: The listing price of the property in USD (numerical)

Predictor Variables:
livingArea: Total area of the property in square feet. (Numerical)
bathrooms: Number of bathrooms in the property. (Numerical)
bedrooms: Number of bedrooms in the property. (Numerical)
yearBuilt: The year when the property was built. (Numerical)
garageSpaces: Number of spaces available in the property's garage. (Numerical)
bed_bath_ratio: Ratio of the number of bedrooms to bathrooms. (Numerical)
price_per_area: Price of the property per square foot. (Numerical)
zipcode: Numeric representation of the property's zip-code. (Numerical)

Categorical Variables:
homeType: Type of the home (e.g., Single Family, Condo, Townhome). Converted to numerical representation. (Numerical) (1=Condo, 2=Lot, 3=Multi_Family, 4=single_family, 5=townhouse)
isNewConstruction: Whether the property is newly constructed. (Converted to numerical representation. (Numerical)) (1=yes 0=no)

Special Codes:
NA: Missing or unavailable information

Data Transformations:
Handling of Missing Values: Rows with missing 'zipcode' values were removed.
Observations with 'livingArea' or 'price' as 0 were also removed.
Observations with both ‘Bedrooms’ and ‘Bathrooms’ as 0 were removed.
Feature Engineering: 'bedrooms' and 'bathrooms' were combined into ‘bed_bath_ratio' as a new feature for analysis.
'Area' and 'Price' were combined into ‘price_per_area’ as a new feature for analysis.

Data Source Citation
This dataset is a co-production, with special thanks to the contributor and their colleague Jordan for compiling this

comprehensive dataset. Using this dataset requires citing the contributor as per the terms listed on the Kaggle page.