# data-memo

Haocheng Zhang

2023-08-20

```r
# Read the CSV file into a dataframe
real_estate_data = read.csv('RealEstate_California.csv')

# View the first few rows of the dataframe
head(real_estate_data)
```

```
##   X                id stateId countyId cityId country datePostedString
## 1 0 95717-2087851113       9       77  24895     USA       2021-01-13
## 2 1   94564-18496265       9      189  36958     USA       2021-07-12
## 3 2   94564-18484475       9      190  36958     USA       2021-07-08
## 4 3   94564-18494835       9      191  36958     USA       2021-07-07
## 5 4 94564-2069722747       9      192  36958     USA       2021-07-07
## 6 5   94564-18484390       9      193  36958     USA       2021-07-06
##   is_bankOwned is_forAuction           event         time  price
## 1            0             0 Listed for sale 1.610496e+12 145000
## 2            0             0 Listed for sale 1.626048e+12 675000
## 3            0             0 Listed for sale 1.625702e+12 649000
## 4            0             0 Listed for sale 1.625616e+12 599000
## 5            0             0 Listed for sale 1.625616e+12 299000
## 6            0             0 Listed for sale 1.625530e+12 575000
##   pricePerSquareFoot     city state yearBuilt    streetAddress zipcode
## 1                  0 Gold Run    CA         0 0 Moody Ridge Rd   95717
## 2                404   Pinole    CA      1958   1476 Belden Ct   94564
## 3                459   Pinole    CA      1959  3540 Savage Ave   94564
## 4                448   Pinole    CA      1908     2391 Plum St   94564
## 5                  0   Pinole    CA         0  2693 Appian Way   94564
## 6                407   Pinole    CA      1958  2611 Doidge Ave   94564
##   longitude latitude hasBadGeocode
## 1 -120.8345 39.16787             0
## 2 -122.3006 38.00121             0
## 3 -122.2714 37.98204             0
## 4 -122.2892 38.00439             0
## 5 -122.2984 37.98631             0
## 6 -122.2573 37.98238             0
##
## 1
## 2
## 3 "Light-filled mid-century 3 BR 2 BA home in the heart of Pinole Valley. Beautiful refinished hardw
## 4
## 5
## 6
##   currency livingArea livingAreaValue lotAreaUnits bathrooms bedrooms
## 1      USD          0               0        Acres         0        0
## 2      USD       1671            1671         sqft         2        3
## 3      USD       1414            1414        Acres         2        3
## 4      USD       1336            1336         sqft         2        3
## 5      USD          0               0        Acres         0        0
## 6      USD       1413            1413         sqft         2        3
##   buildingArea parking garageSpaces hasGarage    levels pool spa
## 1            0       0            0         0         0    0   0
## 2         1671       1            2         1 One Story    0   0
## 3         1414       1            2         1 One Story    0   0
## 4         1336       1            1         1 Two Story    0   1
## 5            0       0            0         0         0    0   0
## 6         1413       1            2         1 One Story    0   0
##   isNewConstruction hasPetsAllowed      homeType               county
## 1                 0              0           LOT        Placer County
## 2                 0              0 SINGLE_FAMILY Contra Costa County
## 3                 0              0 SINGLE_FAMILY Contra Costa County
## 4                 0              0 SINGLE_FAMILY Contra Costa County
## 5                 0              0           LOT Contra Costa County
## 6                 0              0 SINGLE_FAMILY Contra Costa County
```

## Overview of the Dataset

### What does it include?

- The dataset includes information about real estate listings in California, such as property details, price, area, number of bedrooms, type of property, and more.

### Where and how will you be obtaining it?

- Kaggle: Real Estate California

### About how many observations? How many predictors?

- Observations (rows): 35,389
- Predictors (columns): 39

### What types of variables will you be working with?

- Numerical variables: price, living area, bathrooms
- Categorical variables: city, state, home type
- Binary variables: hasGarage, isNewConstruction

### Is there any missing data? About how much? Do you have an idea for how to handle it?

- datePostedString: 3 missing values
- time: 289 missing values
- zip-code: 25 missing values (planning to delete these)
- description: 279 missing values
- And i am planning to delete those without zip-code, because it is essential. Also, deleting those 25 data would not impact too much on assurance.

## Overview of Your Research Questions

### What variable(s) are you interested in predicting? What question(s) are you interested in answering?

- I could focus on predicting property prices, depending on different areas and property types.

### Will these questions be best answered with a classification or regression approach?

- Price prediction would typically be a regression problem, while classifying property types would be a classification problem.

### Which predictors do you think will be especially useful?

- Features like living area, bedrooms, bathrooms, location, and property type may be particularly useful.

### Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

- The primary goal of my model is predictive since I aim to forecast property prices using various features like location (area) and property type.

## Proposed Project Timeline

- Week2: data loading
- Week3: exploratory data analysis
- week4-5: model building, evaluation,
- week6: finalization.

## Any Questions or Concerns

- I am worried that if I should find another dataset for real estate sales in a specific time range, such as 2020-2021, so I can see the trend for price change, which should be helpful to predict future prices. However, the current dataset already includes useful information, such as price, property types, zipcode, which should be already enough for my research.