



Predicting House Prices Using Image Detection Milestone 1

Kaitlyn Chen, Ryan Guo, Jonathan Wang
GenAI Gurus

Topic & Previous Solutions

- Develop a model for multi-step time series prediction of the sale price of houses based on image detection
- Example Solution: <https://www.sciencedirect.com/science/article/pii/S2667305322000217>
 - Study using Deep convolutional neural networks (CNNs)
 - Estimates house prices based on attributes such as interior, exterior, and satellite images
 - Created a model with the strength of being trained on visual information
- Example Solution: <https://github.com/tncy67/House-Price-Prediction-via-Computer-Vision>
 - CNNs for predicting prices based on images of the front of the house

Datasets

With prices:

<https://www.kaggle.com/datasets/ted8080/house-prices-and-images-socal>

- Exterior housing images and pricing dataset containing 8 variables & 15000+ rows in SoCal

<https://www.kaggle.com/competitions/house-price-estimation/overview>

- Images inside & outside house with prices of houses in Cali (Unable to access currently)

Without prices:

<https://www.kaggle.com/datasets/robinreni/house-rooms-image-dataset>

- Around 3000 collective images of Bathroom, Bedroom, Living Room, Dining, & Kitchen spaces

<https://www.kaggle.com/datasets/mikhailma/house-rooms-streets-image-dataset>

- Includes several image categories of the previous dataset including new ones for streets

<https://github.com/emanhamed/Houses-dataset>

- Around 2000 images of bedroom, bathroom, kitchen and frontal images of houses

Data exploration

Socal Dataset: street, ~15.5k images of exterior; columns: city, # bed, # bath, sqft, price

House Price Estimation Dataset: ~2.2k images of bathroom, bedroom, kitchen, exterior; columns: zip code, # bed, # bath, sqft, price

- Initialize dataset path, search in the directory to open the CSV files
- Display the first rows, summary statistics, column data types, amounts of rows & columns
- Plan to feed our neural network with the imaging datasets and train the model to make accurate predictions (ex. CNN)

```
Path to dataset files: /root/.cache/kagglehub/datasets/ted8080/house-prices-and-images-socal/versions/1
Files in dataset: ['socal2.csv', 'socal2']

image_id  street  citi  n_citi  bed  bath  sqft  \
0         0  1317 Van Buren Avenue  Salton City, CA  317  3  2.0  1560
1         1         124 C Street W  Brawley, CA  48  3  2.0  713
2         2      2304 Clark Road  Imperial, CA  152  3  1.0  800
3         3       755 Brawley Avenue  Brawley, CA  48  3  1.0  1082
4         4      2207 R Carrillo Court  Calexico, CA  55  4  3.0  2547

price
0  201900
1  228500
2  273950
3  350000
4  385100

image_id  n_citi  bed  bath  sqft  \
count  15474.000000  15474.000000  15474.000000  15474.000000  15474.000000
mean    7736.500000    216.597518    3.506398    2.453251    2173.913209
std     4467.103368    112.372985    1.034838    0.958742    1025.339617
min       0.000000     0.000000    1.000000    0.000000     280.000000
25%     3868.250000    119.000000    3.000000    2.000000    1426.000000
50%     7736.500000    222.500000    3.000000    2.100000    1951.000000
75%    11604.750000    315.000000    4.000000    3.000000    2737.750000
max    15473.000000    414.000000    12.000000   36.000000   17667.000000

price
count  1.547400e+04
mean    7.031209e+05
std     3.769762e+05
min     1.950000e+05
25%     4.450000e+05
50%     6.390000e+05
75%     8.349750e+05
max     2.000000e+06
image_id  int64
street    object
citi      object
n_citi    int64
bed       int64
bath      float64
sqft      int64
price     int64
dtype: object
(15474, 8)
```

Data cleansing

- Identify and fill/remove missing values in datasets (e.g., price, sqft, bed/bath count).
- Convert incorrect data types (e.g., numerical values stored as strings).
- Identify extreme values in price, sqft, or bed/bath count using statistical methods.
- Ensure all listings have corresponding images, remove unmatched records.
- Normalize categorical data (e.g., city names, zip codes) for consistency.

Data preparation

- Merge the datasets based on common attributes (e.g., zipcode, sqft, bed/bath count).
- Create new features (e.g., price per sqft, neighborhood ranking, image-based features).
- Resize, normalize, and extract features from house images
- Divide the dataset into training and validation sets to evaluate model performance.

Challenges

- Missing/Incomplete Data: Difficulty in handling missing property details and inconsistent records.
- Image-Text Matching Issues: Ensuring each house listing has a corresponding image for accurate predictions.
- Outlier Impact: Extreme values (luxury homes, distressed properties) affecting model performance.
- Data Imbalance: Uneven distribution of house prices across different zip codes and regions.

Questions?