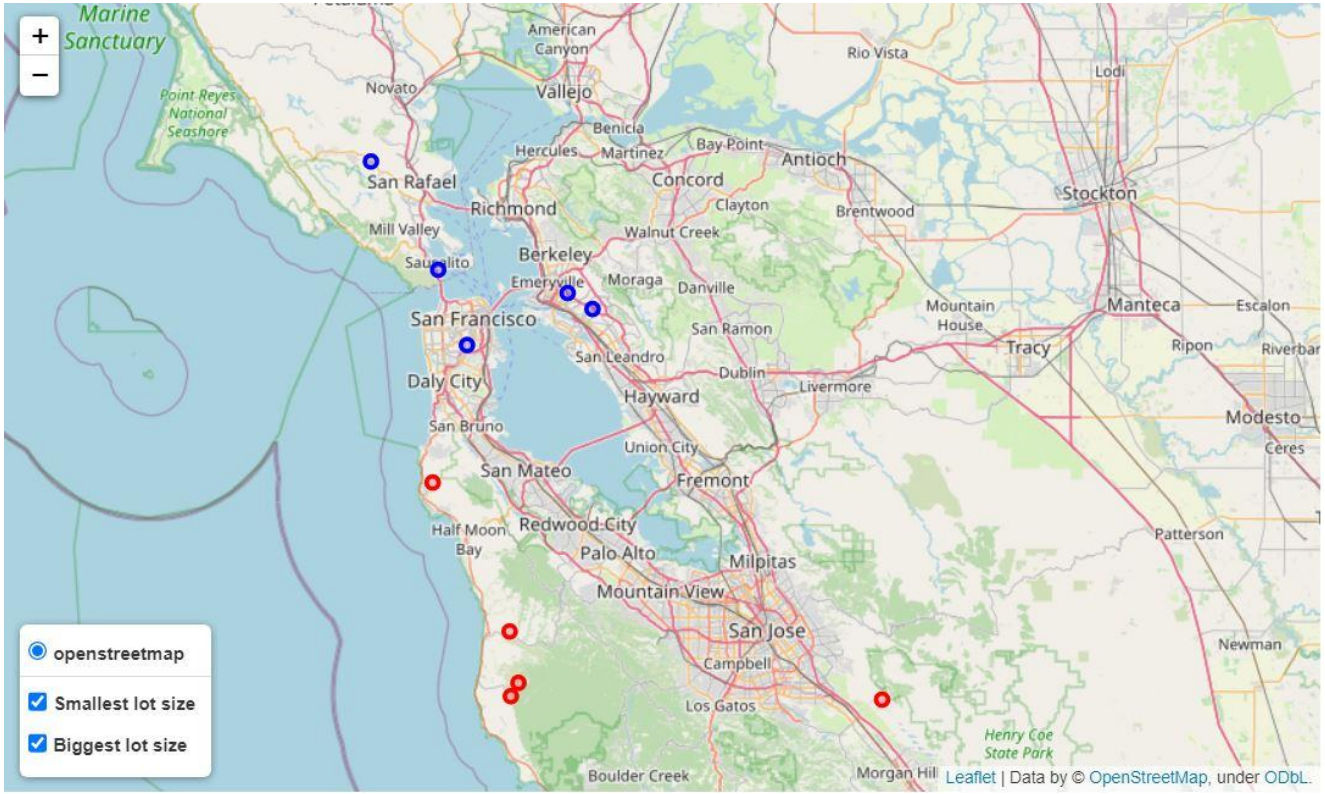
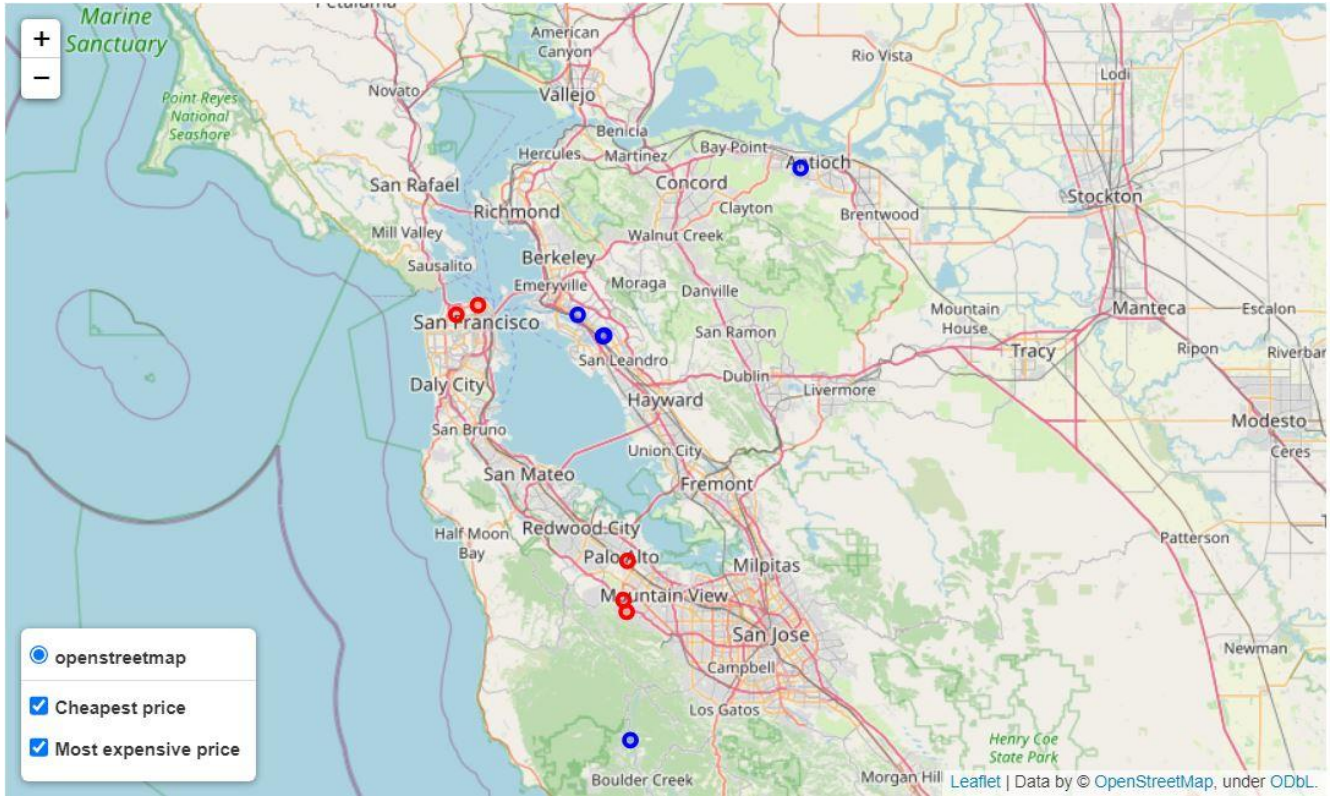
Predicting Bay Area House Prices

By: Rohan Athalye and Paul Chon

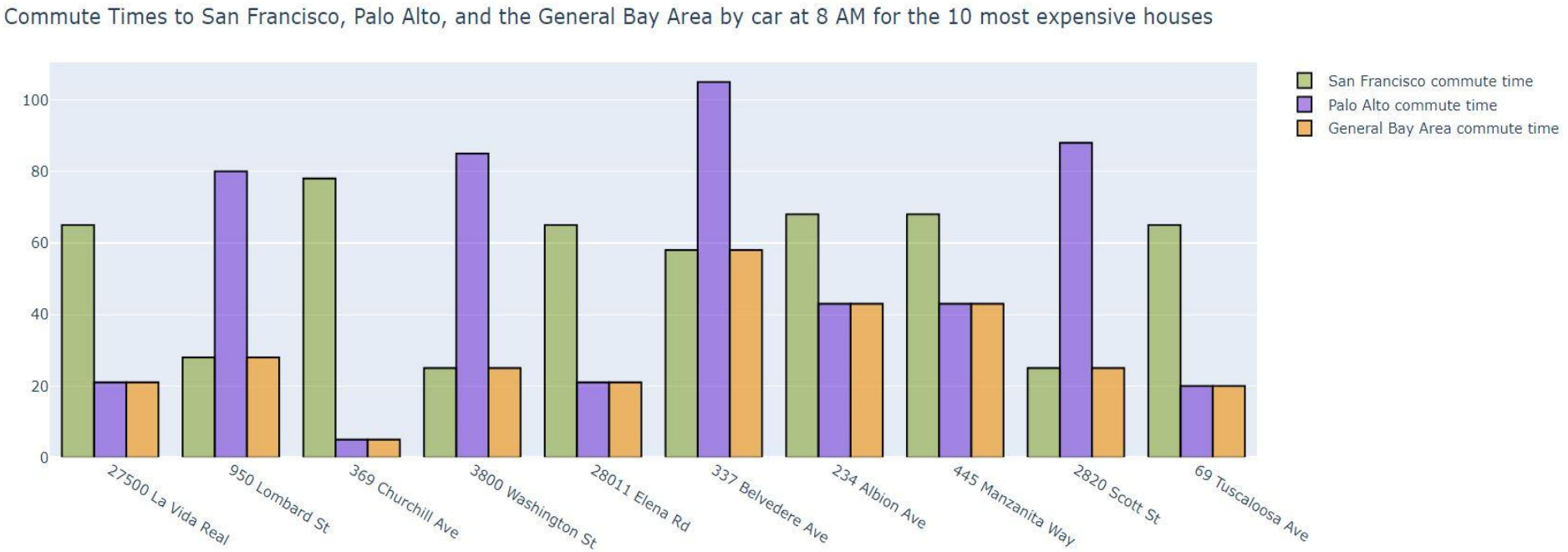
Where are the 5 houses with the biggest lot sizes and 5 houses with the smallest lot sizes located?



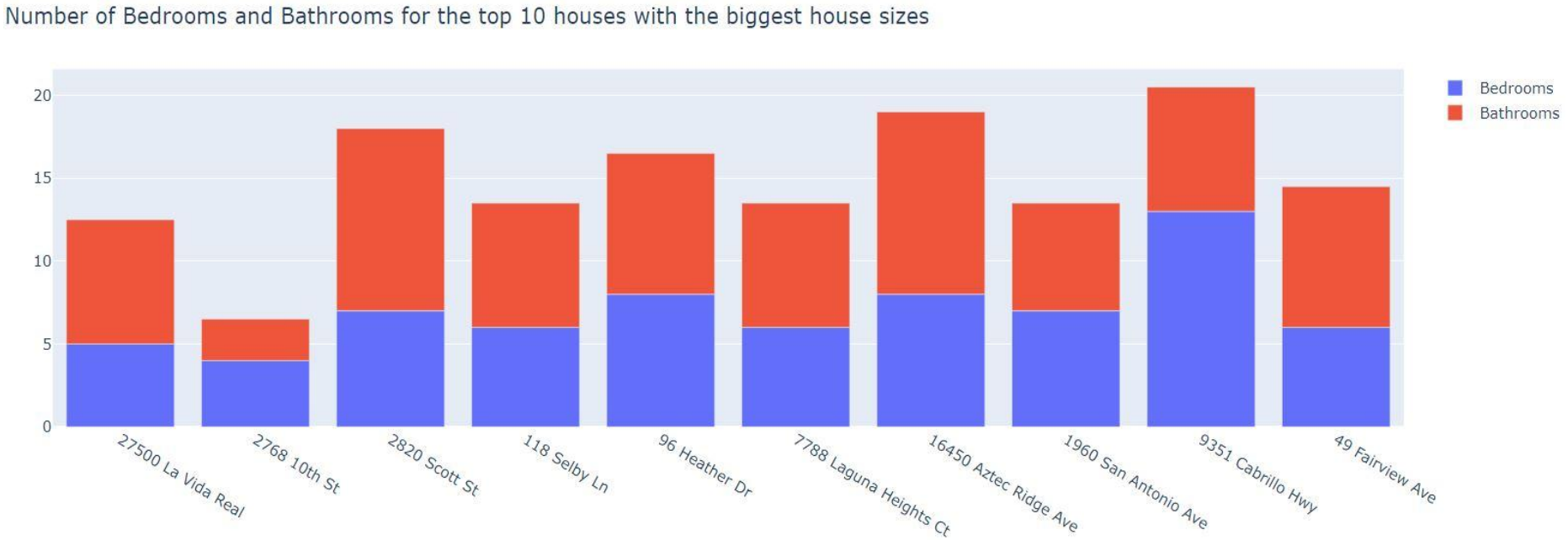
Where are the 5 most expensive and 5 cheapest houses located?



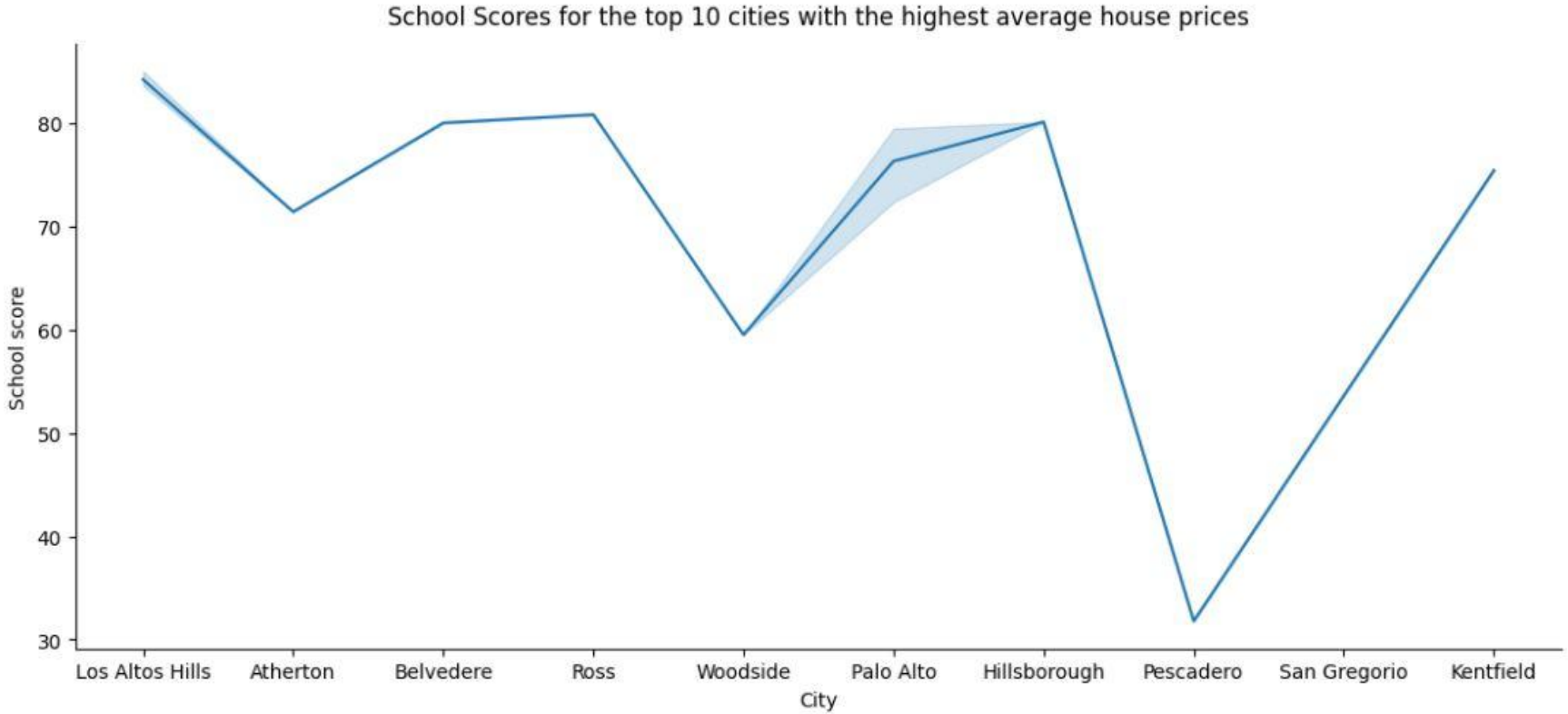
How do commute times by car at 8 AM to San Francisco, Palo Alto, and the general Bay Area compare for the 10 most expensive houses?



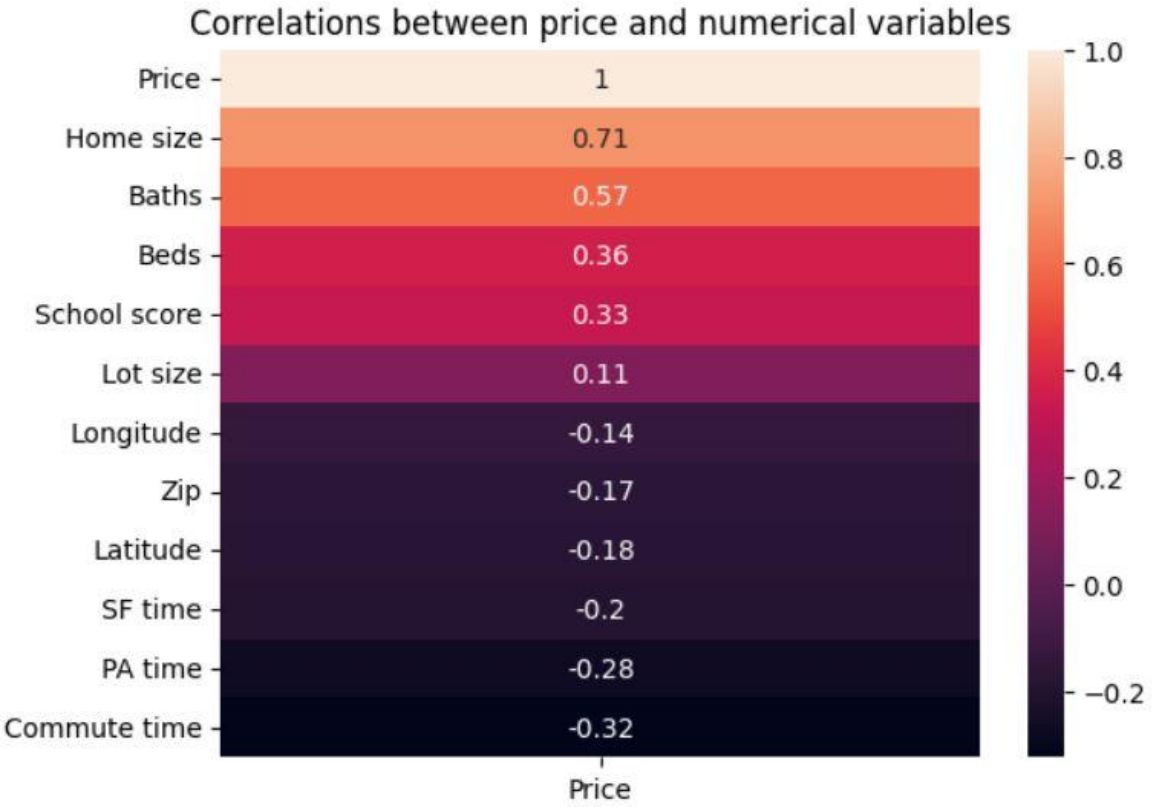
How do the number of bedrooms and bathrooms compare for the top 10 houses with the biggest house sizes?



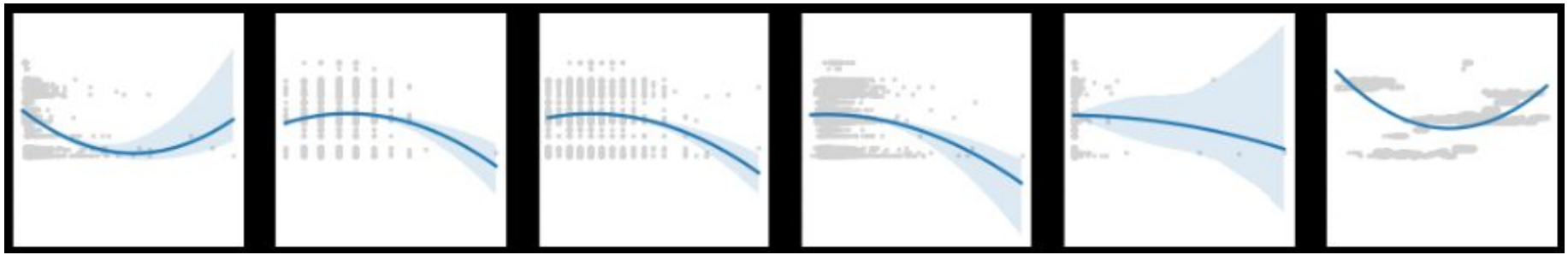
Do cities with expensive house prices also tend to have consistently high school scores?



What are the correlations between price and the numerical variables?



What are the pairwise relationships?



Link to full plot: https://imgur.com/a/osJwMj5

Training/Testing Data Strategy

- 80/20 Stratified Shuffle split based on the Baths category

Pipeline:

1. Categorical columns
    a. Encode
    b. Impute
2. Numerical columns
    a. Impute
    b. Scale
3. PCA (Principal Component Analysis)

Chosen ML Models

1. Linear Regression
    a. Base test
2. Decision Tree
    a. Nonlinear
    b. Comparison
3. Random Forest
    a. Ensemble learning
4. Support Vector Machine
    a. Base test, but nonlinear

Performance of Trained Models

- Performance metric: root mean squared error

Model	Mean cross-validation score for training data with all columns	Mean cross-validation score for training data without Address and State columns
Linear Regression	1276886.45	1276827.61
Decision Tree	1535360.50	1503050.50
Random Forest	1181598.45	1144059.25
Support Vector Machine	2141957.29	2141957.08

Best ML Model Performance using Test Set

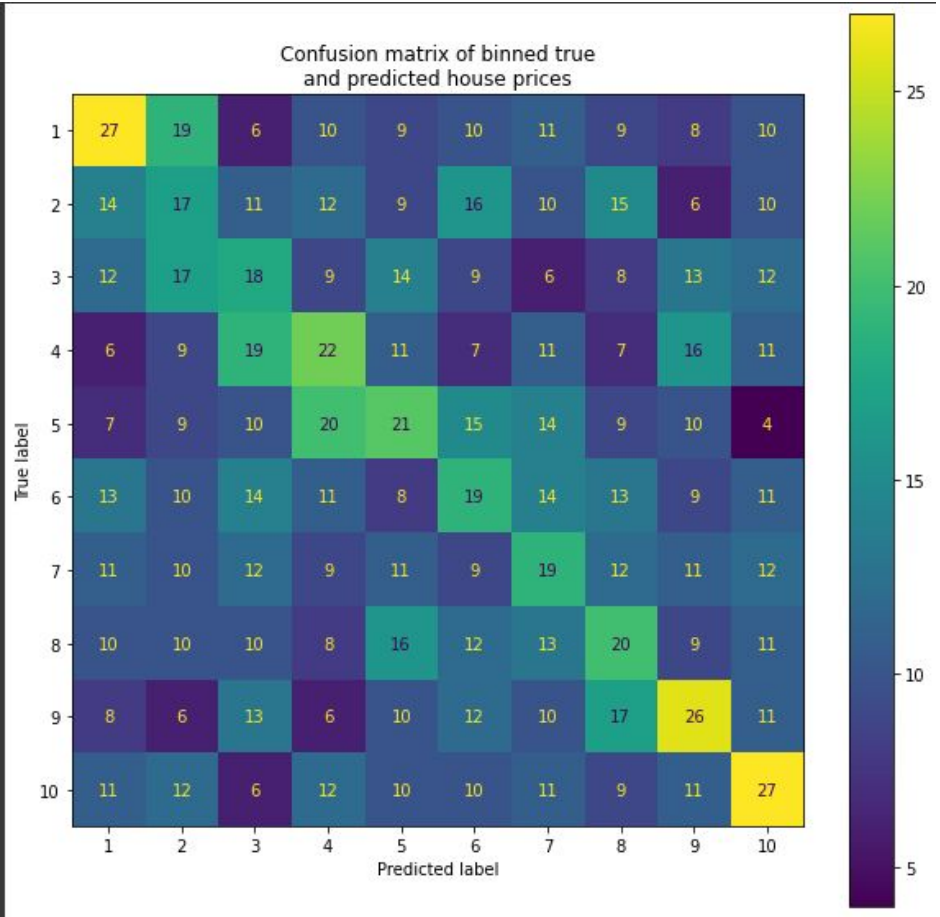
Model: Random Forest

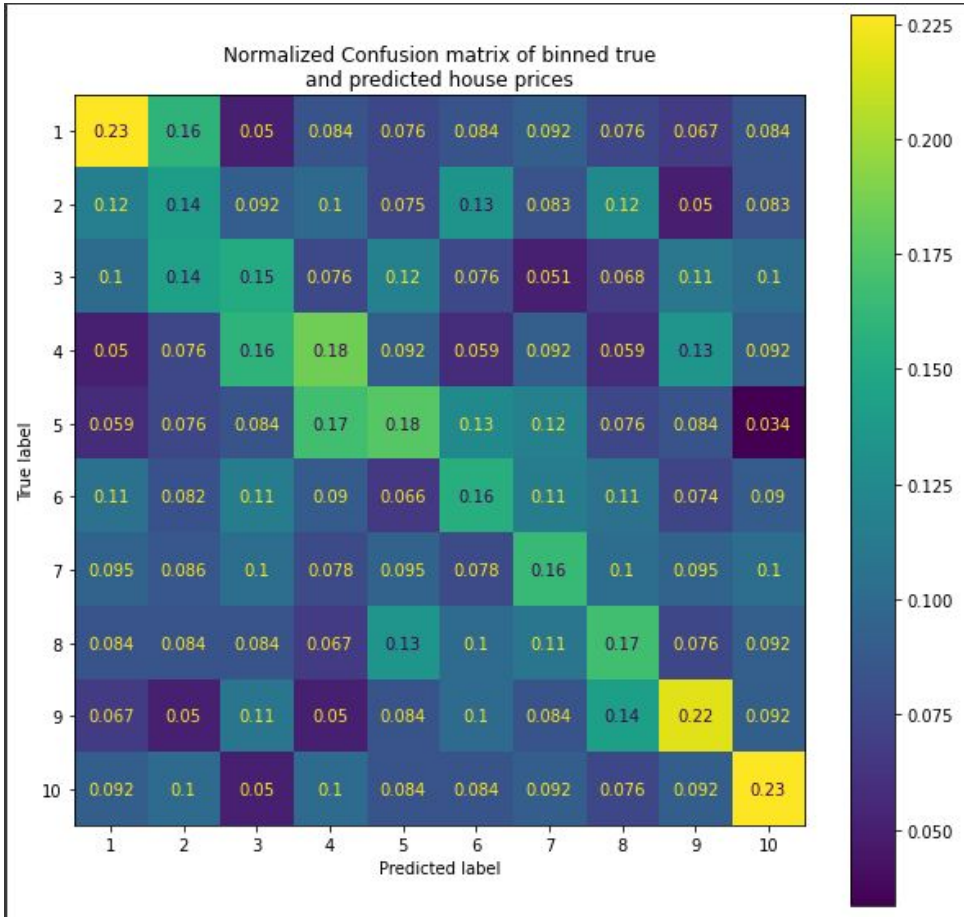
Features:

- max_features=30
- n_estimators=38

RMSE (train): 1161480

RMSE (test): 1237363


Confusion matrix of binned true
and predicted house prices
True label
Predicted label
1 2 3 4 5 6 7 8 9 10
1 27 19 6 10 9 10 11 9 8 10
2 14 17 11 12 9 16 10 15 6 10
3 12 17 18 9 14 9 6 8 13 12
4 6 9 19 22 11 7 11 7 16 11
5 7 9 10 20 21 15 14 9 10 4
6 13 10 14 11 8 19 14 13 9 11
7 11 10 12 9 11 9 19 12 11 12
8 10 10 10 8 16 12 13 20 9 11
9 8 6 13 6 10 12 10 17 26 11
10 11 12 6 12 10 10 11 9 11 27
25
20
15
10
5


Normalized Confusion matrix of binned true
and predicted house prices
True label
Predicted label
1 2 3 4 5 6 7 8 9 10
1 0.23 0.16 0.05 0.084 0.076 0.084 0.092 0.076 0.067 0.084
2 0.12 0.14 0.092 0.1 0.075 0.13 0.083 0.12 0.05 0.083
3 0.1 0.14 0.15 0.076 0.12 0.076 0.051 0.068 0.11 0.1
4 0.05 0.076 0.16 0.18 0.092 0.059 0.092 0.059 0.13 0.092
5 0.059 0.076 0.084 0.17 0.18 0.13 0.12 0.076 0.084 0.034
6 0.11 0.082 0.11 0.09 0.066 0.16 0.11 0.11 0.074 0.09
7 0.095 0.086 0.1 0.078 0.095 0.078 0.16 0.1 0.095 0.1
8 0.084 0.084 0.084 0.067 0.13 0.1 0.11 0.17 0.076 0.092
9 0.067 0.05 0.11 0.05 0.084 0.1 0.084 0.14 0.22 0.092
10 0.092 0.1 0.05 0.1 0.084 0.084 0.092 0.076 0.092 0.23
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050

Challenges and Thoughts

- Creating the confusion matrices
- Pre-processing
  - Deciding what to do
    - Feature selection
- Future
  - Try more ensemble learning
  - Try to better visualize the model performance