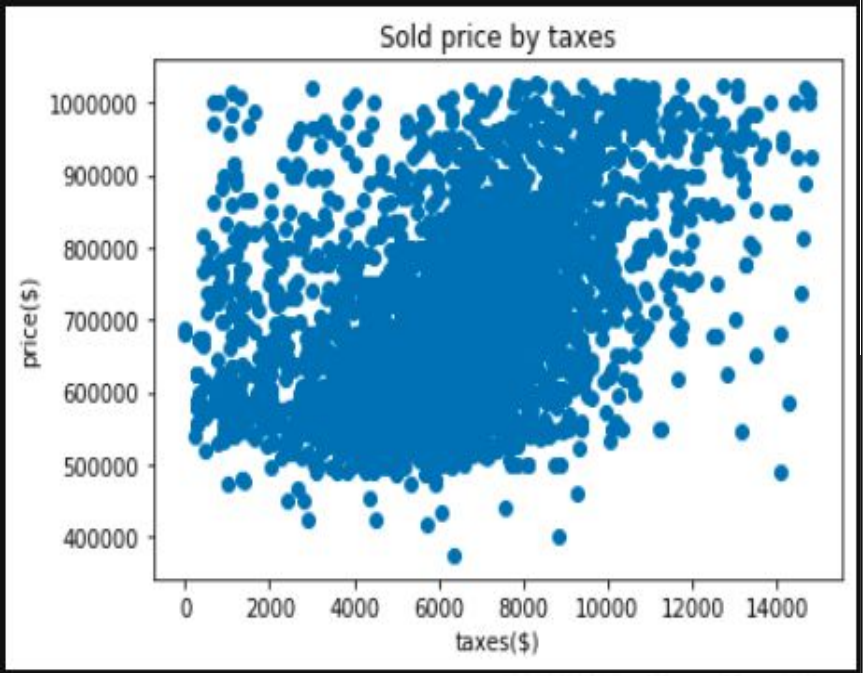
Linear Regression and Bayes Classification: Evaluating housing prices

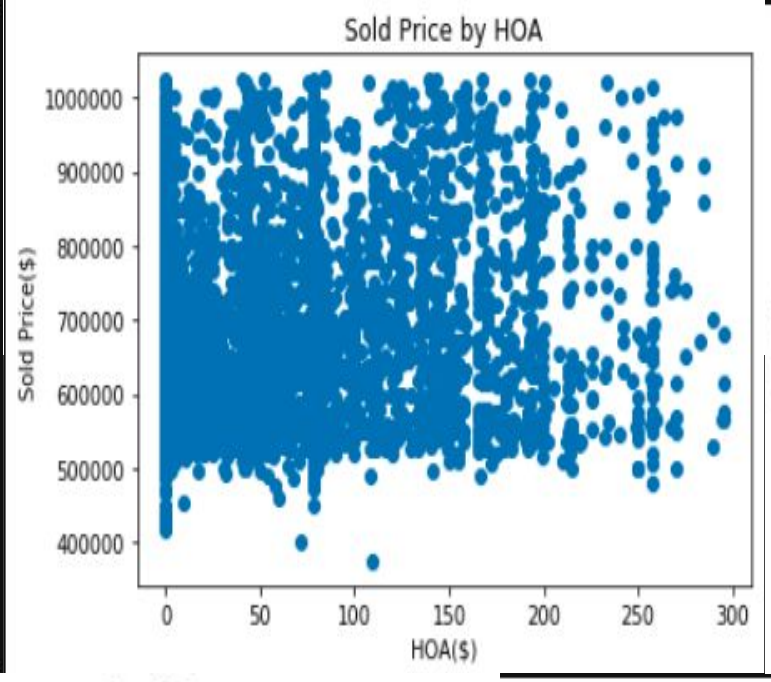
Enhance IT

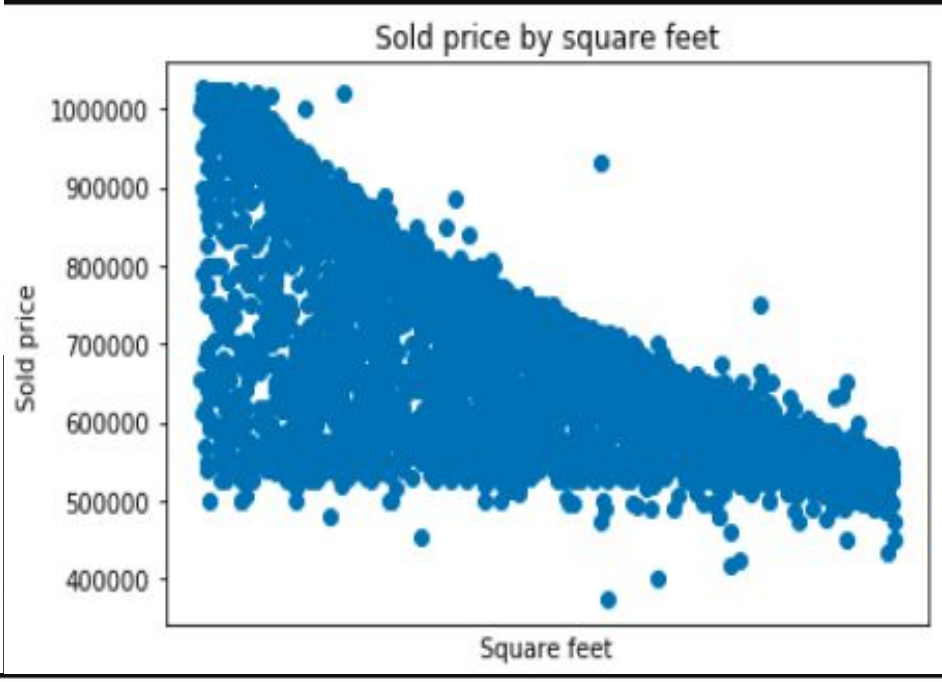
Outline

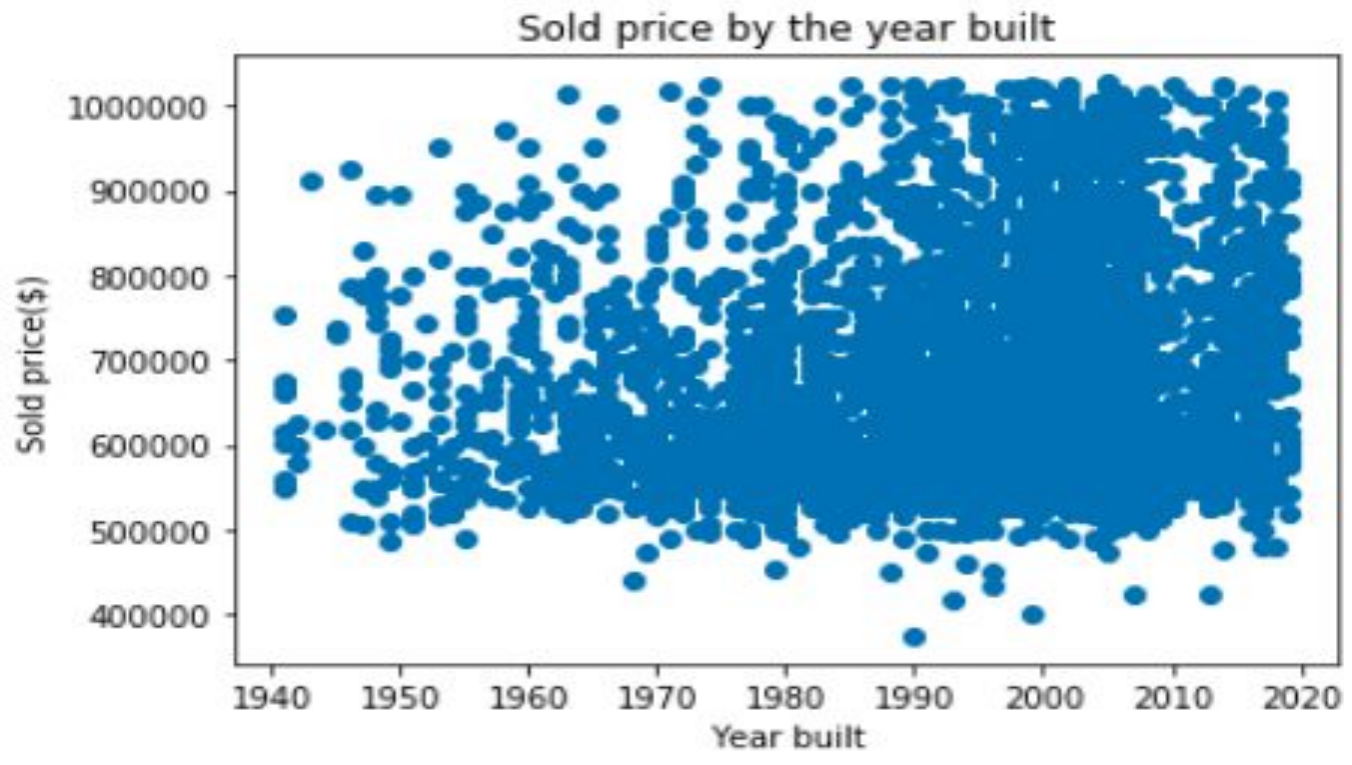
- Data exploration
- Linear Regression
- Bayes classification

Enhance IT


Sold price by taxes
price($)
1000000
900000
800000
700000
600000
500000
400000
0
2000
4000
6000
8000
10000
12000
14000
taxes($)


Sold Price by HOA
Sold Price($)
1000000
900000
800000
700000
600000
500000
400000
0
50
100
150
200
250
300
HOA($)


Sold price by square feet
Sold price
1000000
900000
800000
700000
600000
500000
400000
Square feet


Sold price by the year built
Sold price($)
1000000
900000
800000
700000
600000
500000
400000
1940
1950
1960
1970
1980
1990
2000
2010
2020
Year built


Sold price by the lot of acres
Sold price
1000000
900000
800000
700000
600000
500000
400000
0
10
20
30
40
Lot of acres

Data exploration

Possible correlations with selling price:

- Taxes
- HOA
- Year built
- Lot of acres
- Square feet

Linear regression not applicable:

- Garage
- Fireplaces
- Bedrooms

Enhance IT

Linear regression

- Use standard deviation, mean and covariance to find correlation
- Use linear function to predict the sold price
- Evaluate the linear regression model with the mean squared error

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\hat{\alpha} = \overline{y} - \beta\overline{x}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Enhance IT

Linear regression - results

```
Taxes - a: 522380.5797296087 b: 25.512303510732085 MSE: 456522048615.9411
HOA - a: 657378.0709186892 b: 315.595839670267 94 MSE: 465019873713.07745
Acres - a: 679229.6197592629 b: 128.3723559922125 MSE: 465113054799.64996
Year - a: -1383608.0120514145 b: 1034.1410790330237 MSE: 462408560649.1306
```

Enhance IT

Bayes Classifier

- Using a Naive Bayes Classifier to predict the house value
- Categorizing sold prices in low, medium and high
- Using taxes, lot of acres, year built, HOA and square feet as features
- Changing features to binary (>mean or <mean)
- Using marginal and conditional probabilities of the features for training
- Result: **~47**% accuracy on testing set

$$\hat{y} = argmax \prod_{i=1}^{d} p(x_i|y)p(y)$$

Enhance IT

Conclusion

- Linear regression is not enough to predict the sold price of houses, when given taxes, HOA, lot of acres and built year with a good accuracy
- Naïve Bayes classifier with the same features of linear regression, but adding the square feet is not enough
- Explore with more features
- Explore with other methods and models

Enhance IT

Q&A

Enhance IT

Enhance IT

Thank You!