House Sale Price: Modeling & Prediction

Suchanya (Mild) Trakarnsakdikul
Waya (Jumbo) Piyapanopas
Prem Honghimaphan (Prem)

You are looking to sell your house…

But, you don’t know what information you need to estimate the value of your house…

The problem

Identifying which features are most useful to homeowners that are looking to sell their homes, to help them get the best estimates of their house sale price.

The Data

TRAIN DATA SET

Data set consists of 78 features that help make up the sale price of a house.

Those features can be grouped into the following:

- ❏ Quality and Condition
- ❏ Size (Area)
- ❏ Age
- ❏ Style
- ❏ Location
- ❏ Raw materials used in building

TEST DATA SET

Data set consists of 78 features that help make up the sale price of a house.

Those features can be grouped into the following:

- ❏ Quality and Condition
- ❏ Size (Area)
- ❏ Age
- ❏ Style
- ❏ Location
- ❏ Raw materials used in building

The test data does not come with sale price

* Limited to housing data from Ames, Iowa.

About the Data

Taking a quick look at the City of Ames, Iowa:

- ❏ Has a population of around 65,000
- ❏ Home to Iowa State University (around 36,000 students in total)
- ❏ Since the weather can be really cold in winter, heaters and heating condition will be a necessity for the people living in Ames.

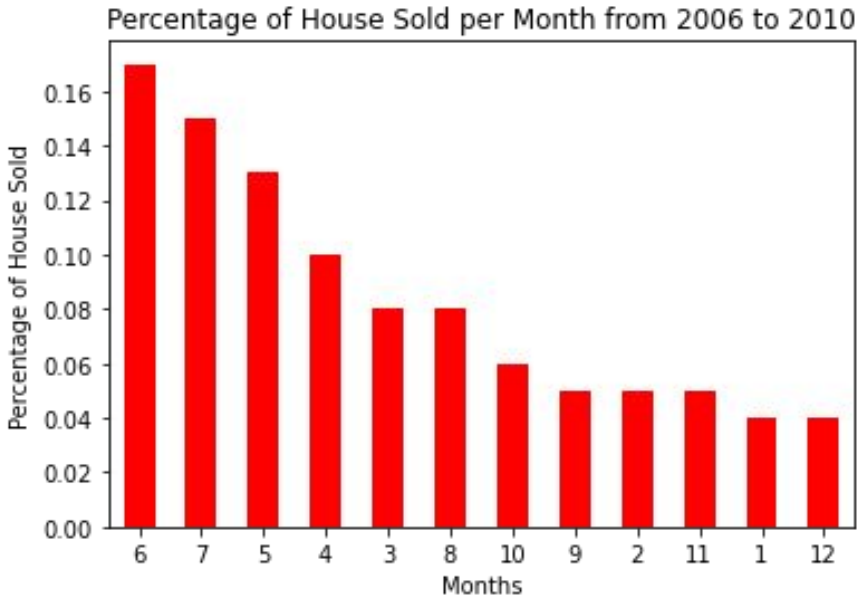
Looking into the data:

- ❏ Housing data is from the year 2006 to 2010.
- ❏ House with minimum price: $12,789
- ❏ House with maximum price: $611,657
- ❏ Average price of all houses in the dataset: $181,469

https://www.cityofames.org/about-ames/about-ames
https://weatherspark.com/y/10339/Average-Weather-in-Ames-Iowa-United-States-Year-Round

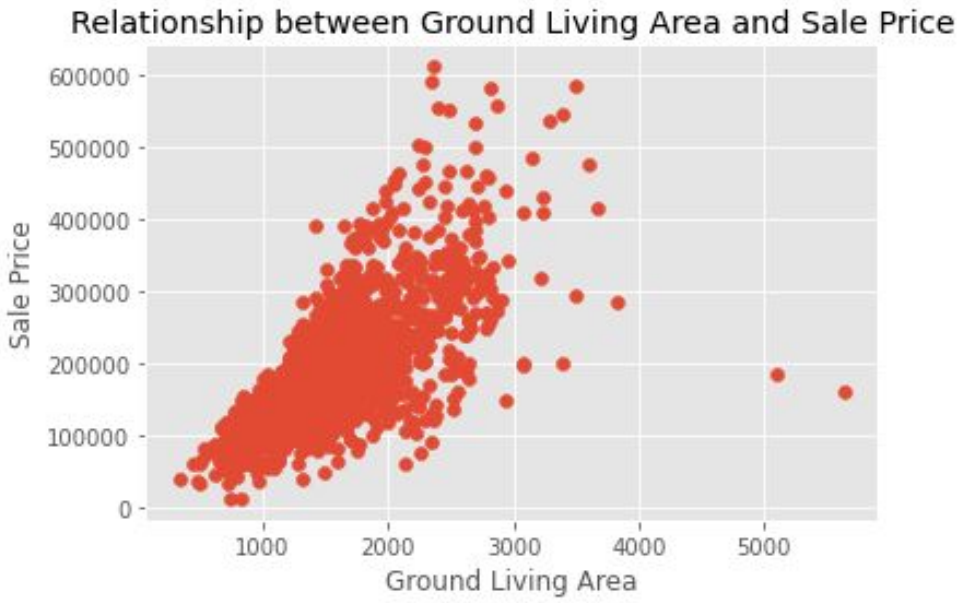
Key Insights

Houses are **mainly sold during summer** months (May to July) and least sold in winter months (November to January)



Outliers in the data:
- Ground Living Area > 4000
- Z Score from sale price

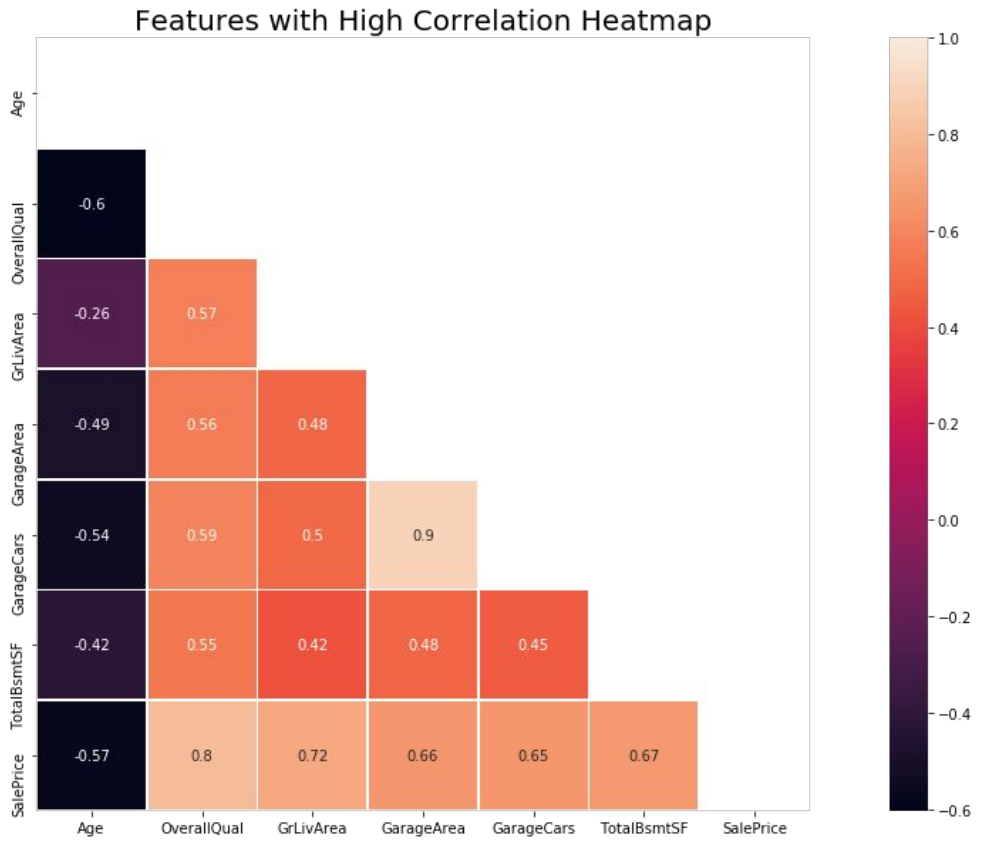
Removing them would help improve predictions



Features Engineered

Based on research and being curious, few features were created and used for analysis:

- **Age:** The difference year sold and year built
- **Total Interior:** Total interior space of 1st floor and 2nd floor combined
- **Has second floor:** Whether there is a 2nd floor or not
- **Remodel:** Whether there was remodeling done or not
- **Neighborhood:** Most people want to live in good neighborhoods, therefore it must have an effect.

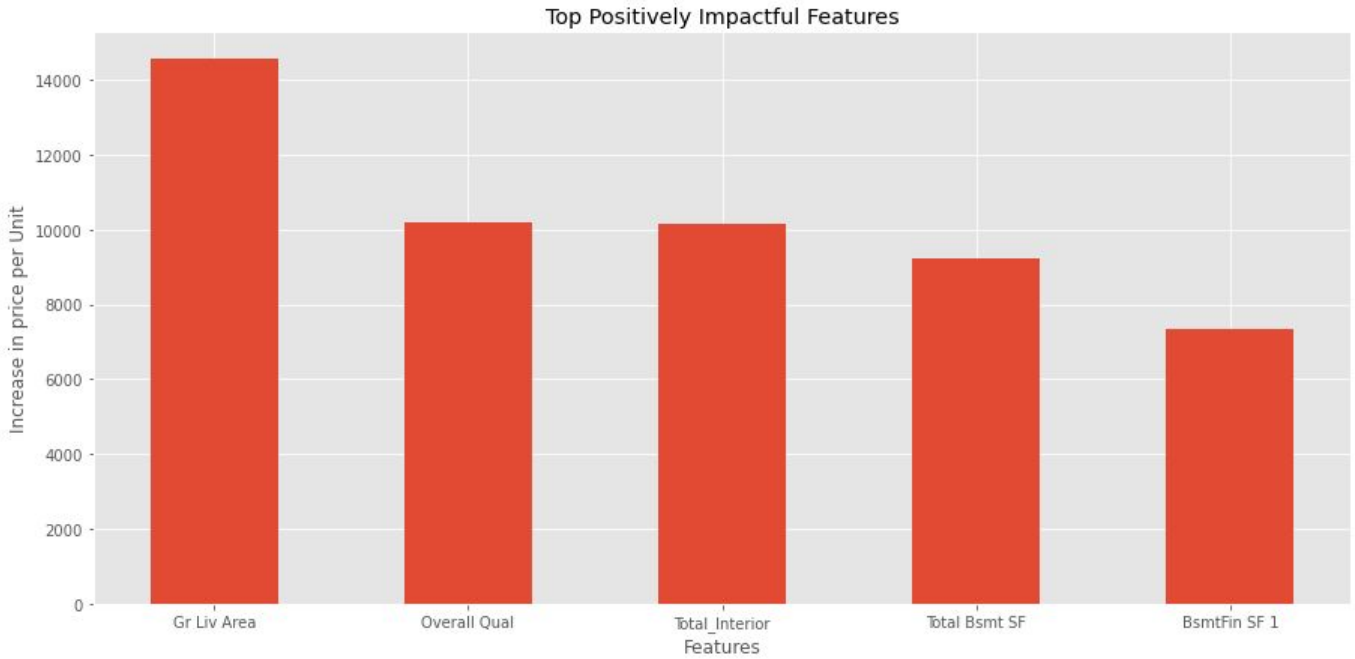


https://www.opendoor.com/w/blog/factors-that-influence-home-value

Modeling and Prediction

Using **Lasso** and **Elastic Net** Models:

- ❏ Determine the quality of each features
- ❏ Ones with negative impact will be zeroed out
- ❏ From the quality features an optimized model is created, where the variance is not too high or too low.



Product Model:

This **Ridge model** that was created, as the **optimized model,** had decent accuracy on unseen data.

The features selected to estimate the price of the house would include:

- Any features that deals with area, quality and condition, age, and location as
- They are having high correlation to the price
- They are also factors that helps create accurate price prediction.

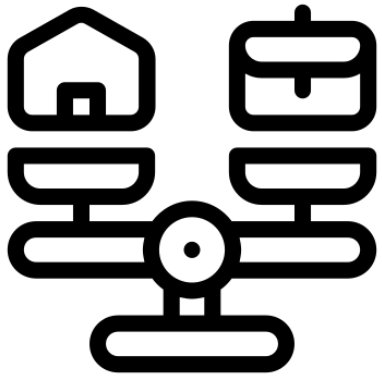
```
lr, R2: 0.9251819722280523
lr, RMSE: 21635.331152418134
Ridge, R2: 0.9253036357767206
Ridge, RMSE: 21617.565088617845
Elastic, R2: 0.9251613454145234
Elastic, RMSE 21640.421707590485
```

About the Features (Ordered)

- Above Ground Living Area
  - 14544 INCREASE PER UNIT
- Overall Quality of the house
  - 10194 INCREASE PER UNIT
- Total living space
  - 10159 INCREASE PER UNIT
- Basement Area
  - 9245 INCREASE PER UNIT
- Overall Condition of the house
  - 5962 INCREASE PER UNIT
- Kitchen Quality
  - 5656 INCREASE PER UNIT
- Exterior Quality
  - 5134 INCREASE PER UNIT
- Basement Quality
  - 4965 INCREASE PER UNIT
- Garage Area
  - 3511 INCREASE PER UNIT
- Age
  - 10437 DECREASE PER UNIT
- Neighborhood
  - RANGES FROM 5656 INCREASE TO 0.822 DECREASE DEPENDING ON LOCATION

Additional Data

To make this model universal we can consider:



Standard of Living Generalization, through GDP estimators

To generalize it more we can take data from places with similar standard of living using GDP per capita to predict places with similar levels of income

Location Generalization, segmented by area of living

Location in Neighborhoods contribute greatly on the quality of the model, the model is effective as a universal model.

Conclusion

The features that relates to **sizing** (size of the living space, lot, garage, etc.), **quality** and **condition** (in excellent quality), **location** (neighborhood), and **age** are most useful to identify the price of the house.

- Age being the most negatively impactful on the price. With an increase of only 1 year of age, the price of the house would roughly drop by USD 10,000

Since this model is built on Ames housing data, it cannot be applied to everywhere else. Therefore, for further investigation we need to

- Gathering information about different locations would help our model cater to a wider audience and generalized the model

Thank you

Appreciate your time.