Project 2
Ames Iowa
Housing Data Analysis

Purpose

Using the Ames Iowa Housing dataset to create a good model for predicting the price of homes at sale

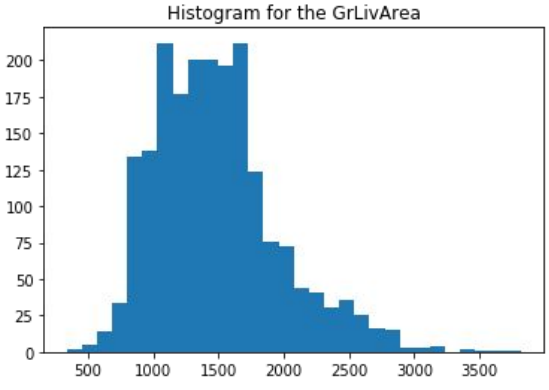
Feature Engineering

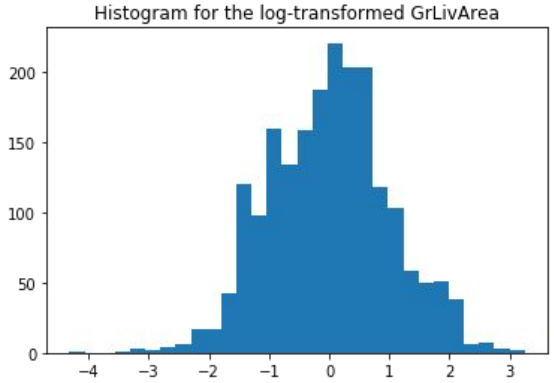
For trying various features, 3 datasets were created.

#1. Training data including some labeled or dummified categorical data

#2. Log-transformed Training data using #1 data

#3. Polynomial or interactive numeric training data using #1 data

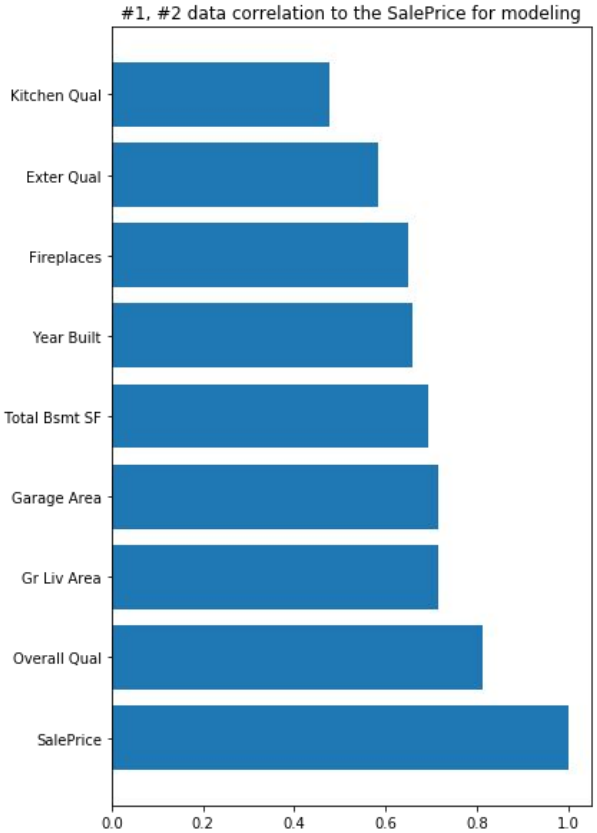


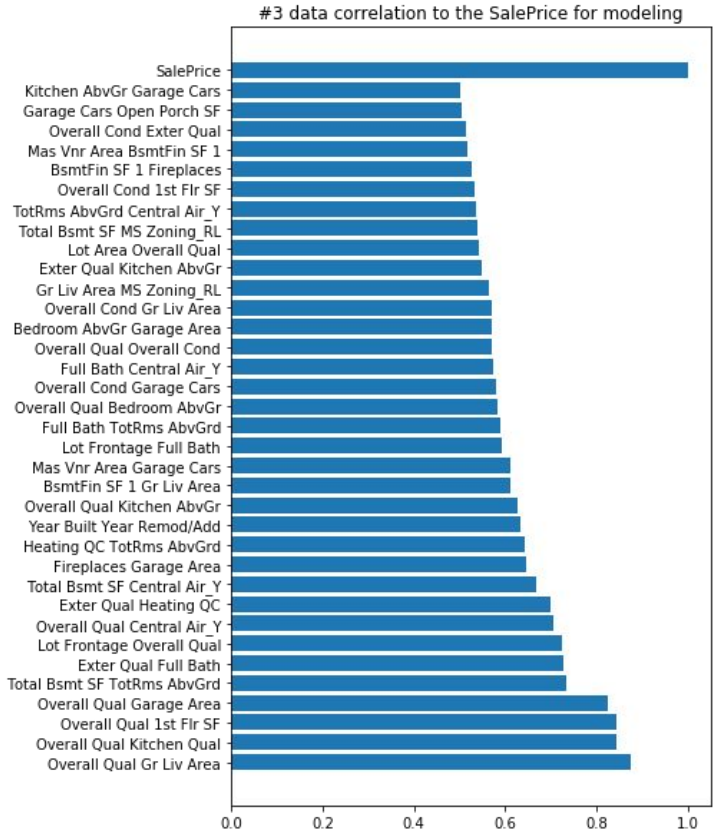


Features for modeling

- These features were used for modeling

- They have more than 0.5 correlation coefficients, but are not highly correlated (less than 0.8) each other





Modeling

Created 3 models (linear regression, ridge regression, and Lasso regression) for each training dataset

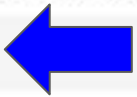
	Linear Regression	Ridge Regression	Lasso Regression
#1 traing data	0.8594	0.8593	0.8594
#2 traing data	0.8567	0.8572	Failed to make a model
#3 traing data	0.9119	0.9128	0.9130

Modeling

Created 3 models for each training dataset

Production model = the Lasso Regression using #3 data

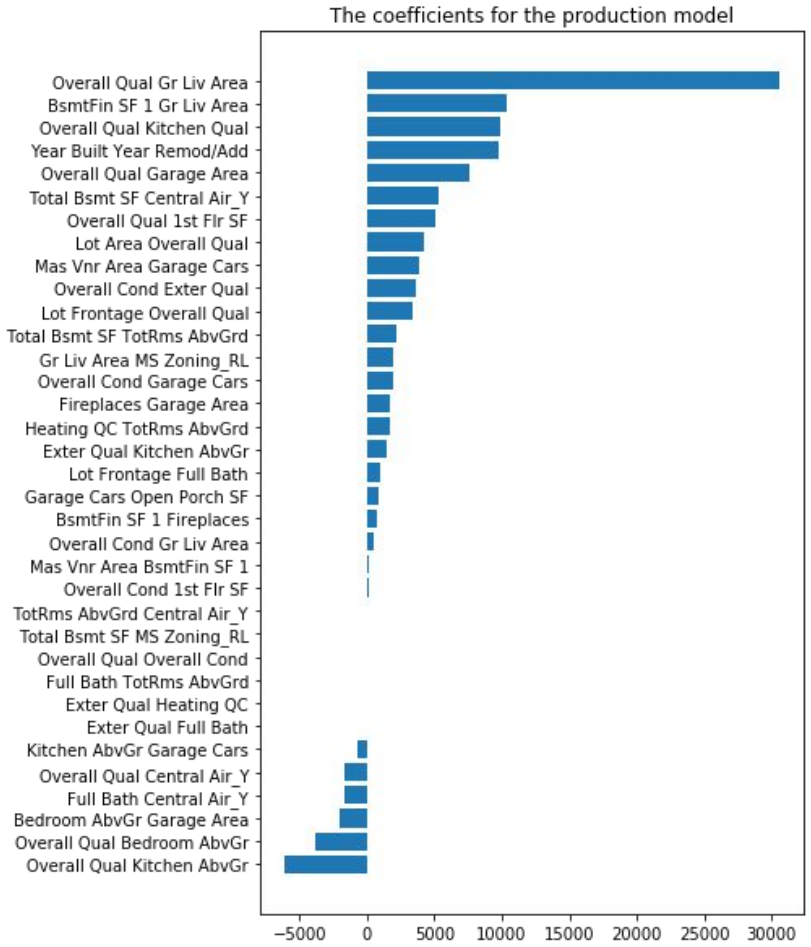
	Linear Regression	Ridge Regression	Lasso Regression
#1 traing data	0.8594	0.8593	0.8594
#2 traing data	0.8567	0.8572	Failed to make a model
#3 traing data	0.9119	0.9128	0.9130



Production model

- R^2 score = 0.91, alpha = 117
- The coefficients show that the ‘Overall Qual’, the ‘Gr Liv Area’, the ‘Garage Area’, etc. have high weights for the sale price

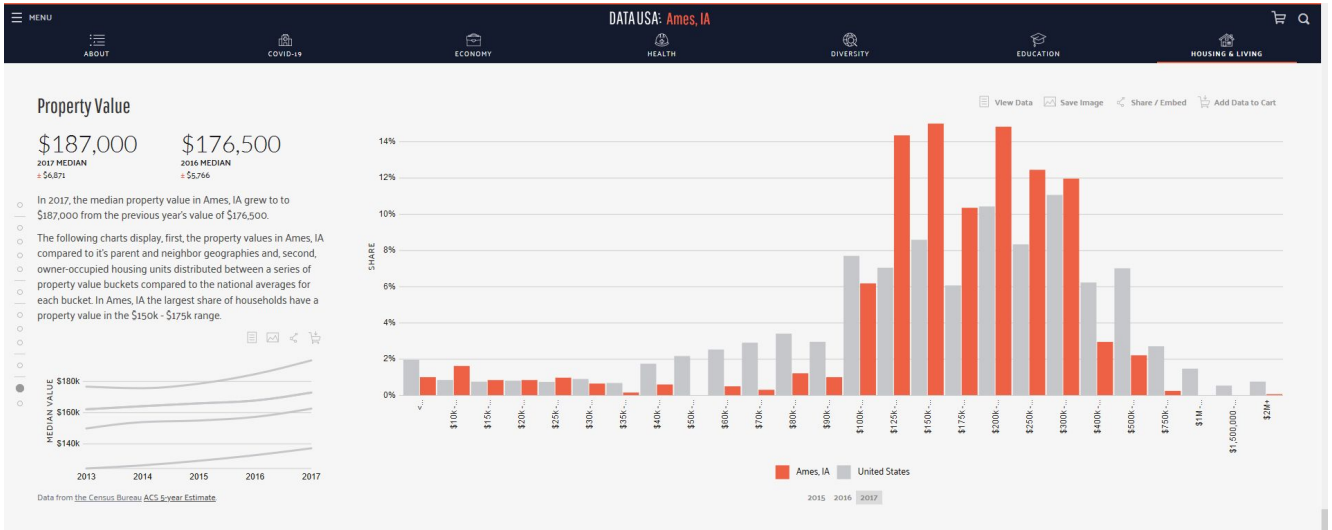
⇒ The quality and size are important to determine the house sale price



Model constraint

- The production model would be able to be applied to other city as a similar sity as Ames
- If the model would be applied to other city like New York, the prediction would be lower

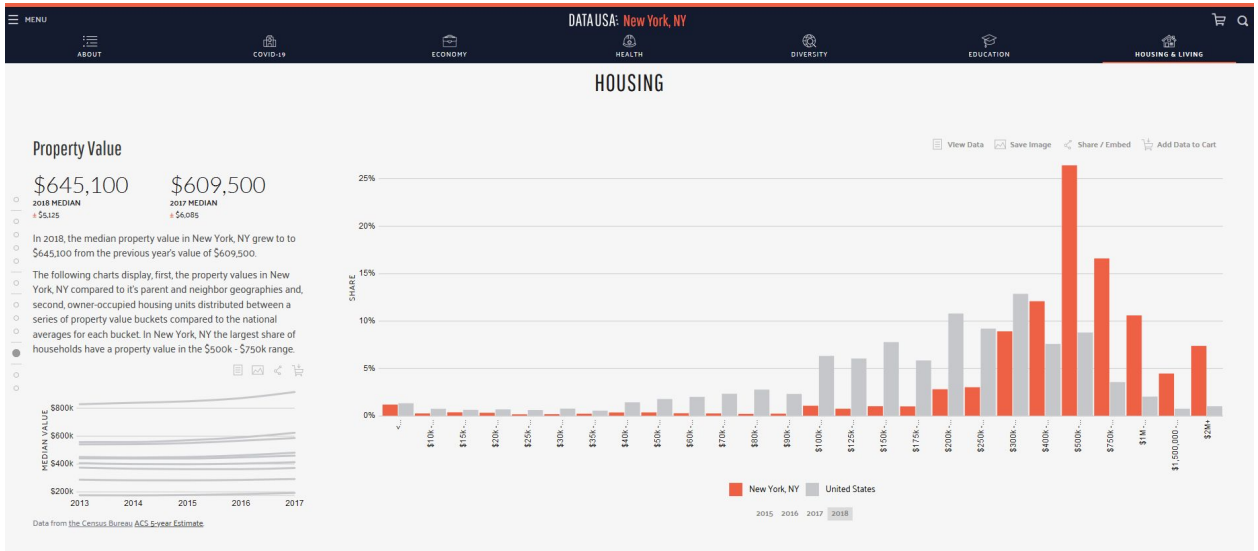
From DATAUSA(*1)



Model constraint

- The production model would be able to be applied to other city as a similar sity as Ames
- If the model would be applied to other city like New York, the prediction would be lower

From DATAUSA(*2)



Model constraint

- The production model would be able to be applied to other city similar to Ames
- If the model would be applied to other city like New York, the prediction would be lower

⇒ Models are basing on a training dataset

⇒ If we want to make a generalized model, we need data from other cities

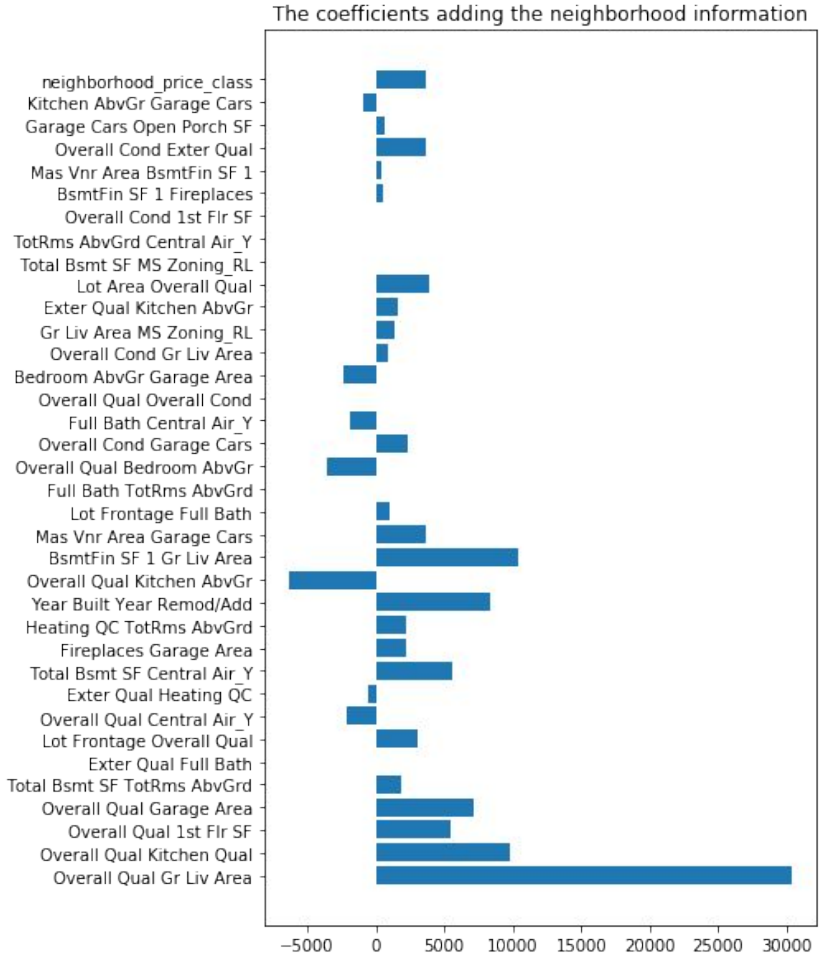
Improve the production model

- According to a paper (* 2), neighborhood information affects house sale price
- The model doesn’t have neighborhood information
- Trained the model again adding the neighborhood information labeled into 3 classes basing on the sale price

⇒ The R^2 score = 0.91 the score didn’t improve!

⇒ But the coefficient has the power to explain the sale price

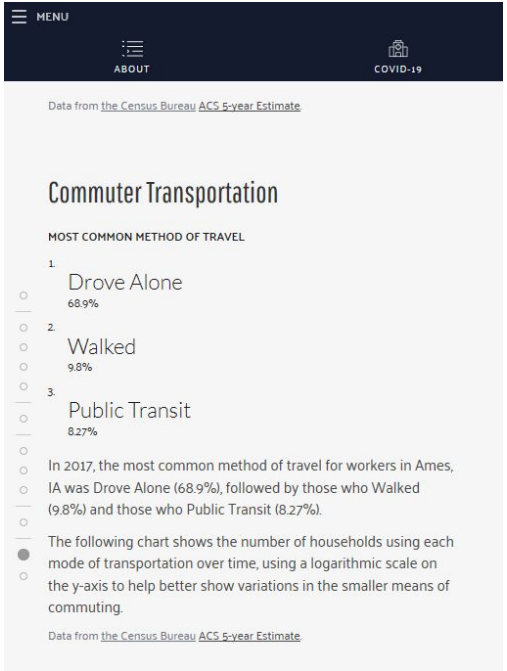
⇒ There might be a better feature for the neighborhood


The coefficients adding the neighborhood information
neighborhood_price_class
Kitchen AbvGr Garage Cars
Garage Cars Open Porch SF
Overall Cond Exter Qual
Mas Vnr Area BsmtFin SF 1
BsmtFin SF 1 Fireplaces
Overall Cond 1st Flr SF
TotRms AbvGrd Central Air_Y
Total Bsmt SF MS Zoning_RL
Lot Area Overall Qual
Exter Qual Kitchen AbvGr
Gr Liv Area MS Zoning_RL
Overall Cond Gr Liv Area
Bedroom AbvGr Garage Area
Overall Qual Overall Cond
Full Bath Central Air_Y
Overall Cond Garage Cars
Overall Qual Bedroom AbvGr
Full Bath TotRms AbvGrd
Lot Frontage Full Bath
Mas Vnr Area Garage Cars
BsmtFin SF 1 Gr Liv Area
Overall Qual Kitchen AbvGr
Year Built Year Remod/Add
Heating QC TotRms AbvGrd
Fireplaces Garage Area
Total Bsmt SF Central Air_Y
Exter Qual Heating QC
Overall Qual Central Air_Y
Lot Frontage Overall Qual
Exter Qual Full Bath
Total Bsmt SF TotRms AbvGrd
Overall Qual Garage Area
Overall Qual 1st Flr SF
Overall Qual Kitchen Qual
Overall Qual Gr Liv Area
−5000
0
5000
10000
15000
20000
25000
30000

Suggestion of other neighborhood features

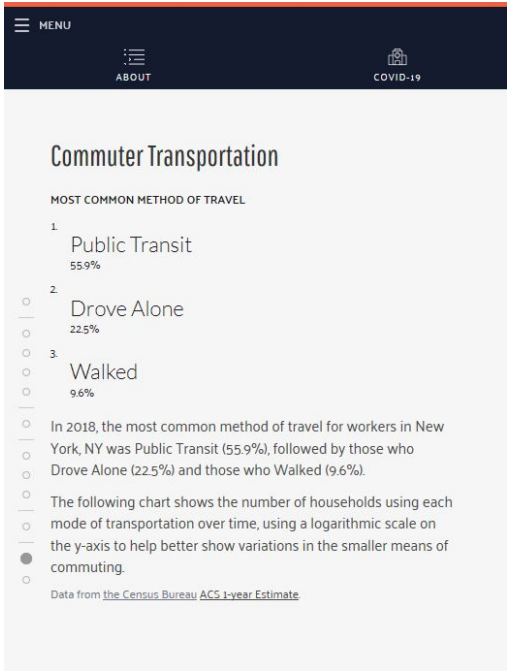
- The distance to a public transportation might affects the sale price

Ames



From DATAUSA(*1)

NY

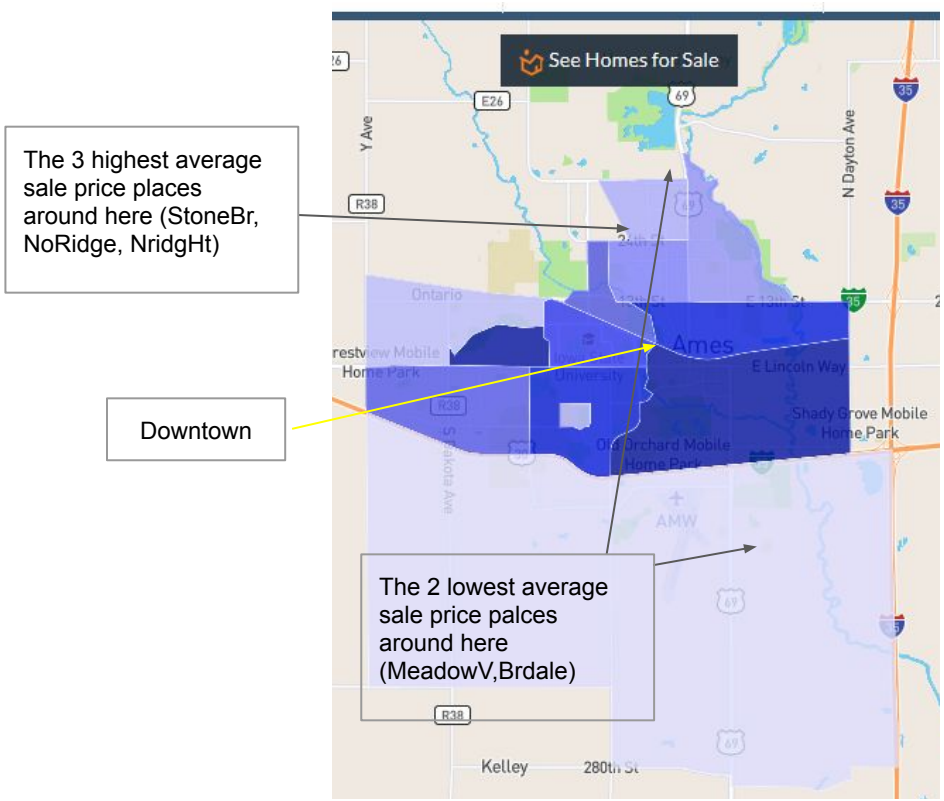


From DATAUSA(*2)

Suggestion of other neighborhood features

- The distance to a public transportation might affect the sale price
- Crime rate might affect the sale price
- The distance to downtown might affect the sale price
- Collecting these information might improve the model

Ames crime rate map (*4)



Sammary

- The model is good to predict the sale price in Ames since the R^2 score is about 0.91

- This model can be applied to other cities similar to Ames

- If this model would be applied to other places like New York, the model should be trained for the place's data since their city characteristics are different

- Neighborhood information should be important for a better model to predict house prices

- However neighborhood data this time didn’t improve the score but can explain the model

- It would be better to gather more neighborhood data such as distance to a station, crime rates, etc. to create a better model

Reference

- (*1) https://datausa.io/profile/geo/ames-ia/#housing
- (*2) https://datausa.io/profile/geo/new-york-ny#housing
- (*3) https://www.researchgate.net/publication/304597534_THE_IMPACT_OF_NEIGHBORHOOD_CHARACTERISTICS_ON_HOUSING_PRICESAN_APPLICATION_OF_HIERARCHICAL_LINEAR_MODELING
- (*4) https://www.neighborhoodscout.com/ia/ames/crime