How Ames, Iowa Housing Data Models Can Help You!

Julia Taussig

22 Mar 2019

Overview

- **Problem:** How do we use Ames, Iowa housing data and modeling techniques to predict property prices, and how can we use this knowledge to focus efforts to increase property value/price?
- **Data supplied by:** Dean De Cock, Truman State University
- **Explore data:** Cleaned data and inspected relationships between property features and sale price
- **Model with data:** Utilized Python and Scikitlearn and Matplotlib and other libraries to create Linear Regression model to predict sale price given certain features
- **Evaluate model:** Utilized linear regression metrics to evaluate model accuracy and precision
- **Answer problem:** I'll give you some recommendations!

Background

- Dataset: 81 variables and 2051 rows, compiled by Dean De Cock
    - Test dataset: 80 variables and 879 rows
- Data from Ames Assessor’s Office (used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010)
- Iowa State University located in Ames, Iowa
- Ames, Iowa population as of 2010 Census: 58,965
  (including students enrolled at ISU - over 36,000 students)

Sources: Dean De Cock, http://jse.amstat.org/v19n3/decock/DataDocumentation.txt
City of Ames, https://www.cityofames.org/about-ames/interesting-facts-about-ames

Data Cleaning

- Null values: inspected and filled with 0 or NA or other appropriate values
- Year garage was built column was removed due to issues with null values
- One additional feature was created:

  Total bathrooms =

  Basement Full Baths + Basement ½ Baths +

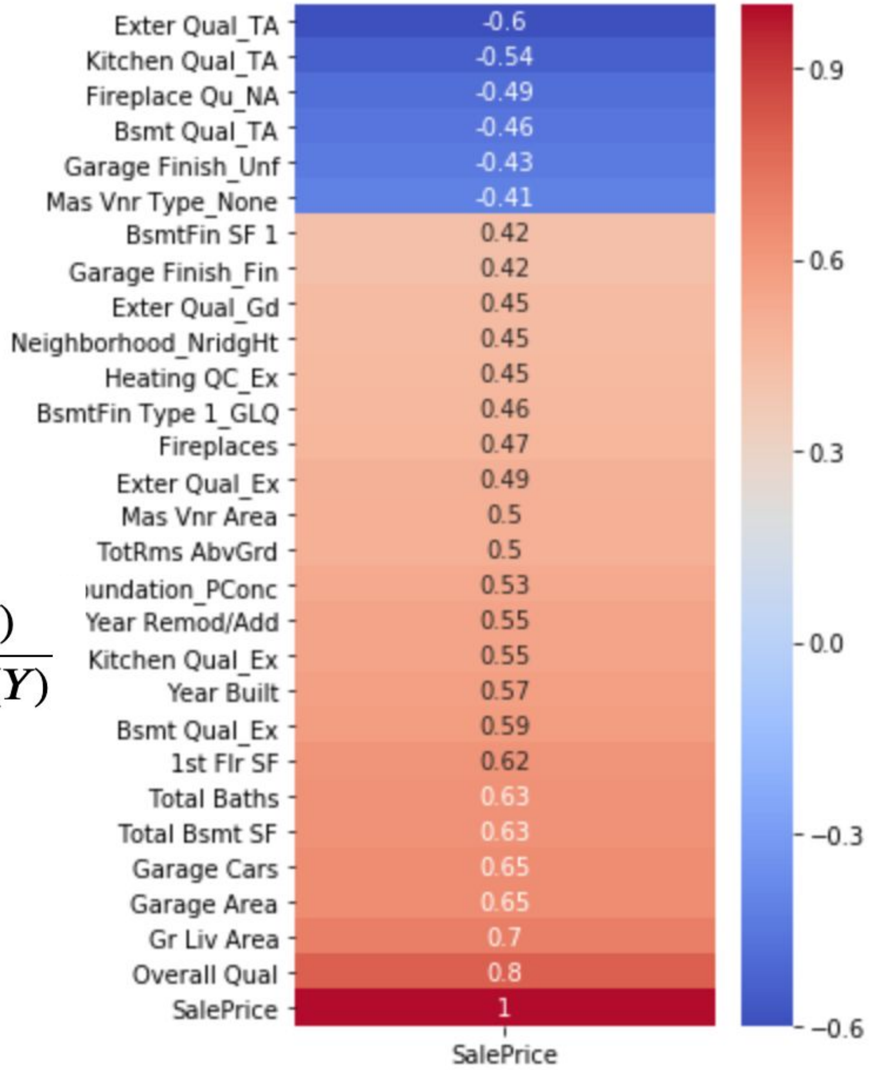
  Above Grade Full Baths + Above Grade ½ Baths

Exploring the Data

The variables shown in the heatmap have |correlation| with sale price ≥ 0.4

$$\text{pearson correlation } r = cor(X,Y) = \frac{cov(X,Y)}{std(X)std(Y)}$$

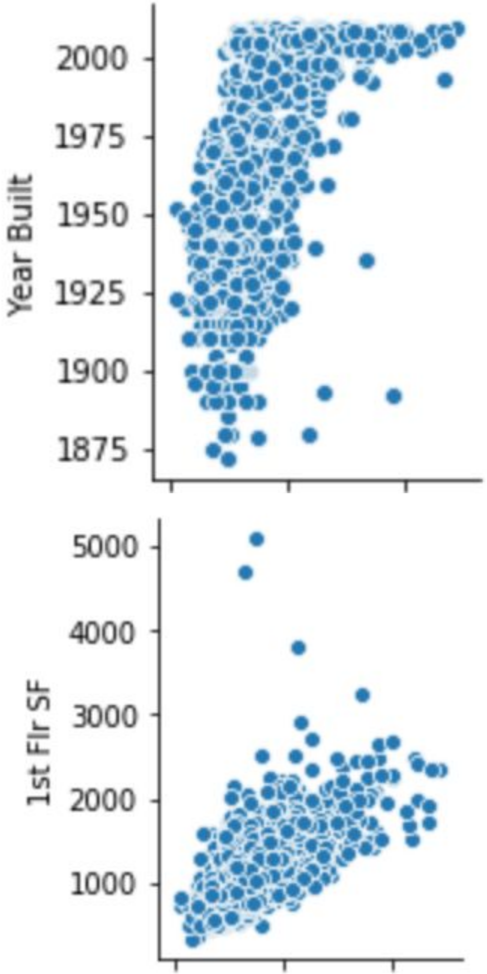
Source: General Assembly lesson by *Kiefer Katovich (SF), Minor updates by David Yerrington (SF):* [*http://localhost:8888/notebooks/Desktop/DSI-US-7/Lessons/2.04-lesson-eda/2_04-basic-eda-walkthrough.ipynb#cov_cor*](http://localhost:8888/notebooks/Desktop/DSI-US-7/Lessons/2.04-lesson-eda/2_04-basic-eda-walkthrough.ipynb#cov_cor)



Linear Regression Assumptions

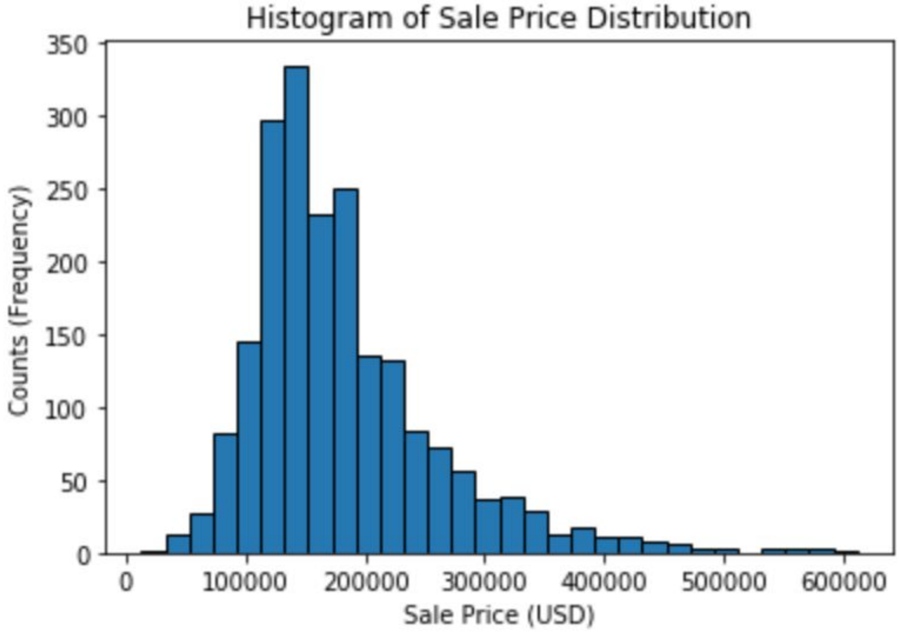
- Each feature linearly related to sale price (see plots of some features vs. sale price used in model to right)
- Independence of errors
- Normality of errors (mean of 0)
- Equality of variance (*e.g.,* errors don't increase as feature values increase)
- Independence of predictors (features)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_p X_{pi}$$

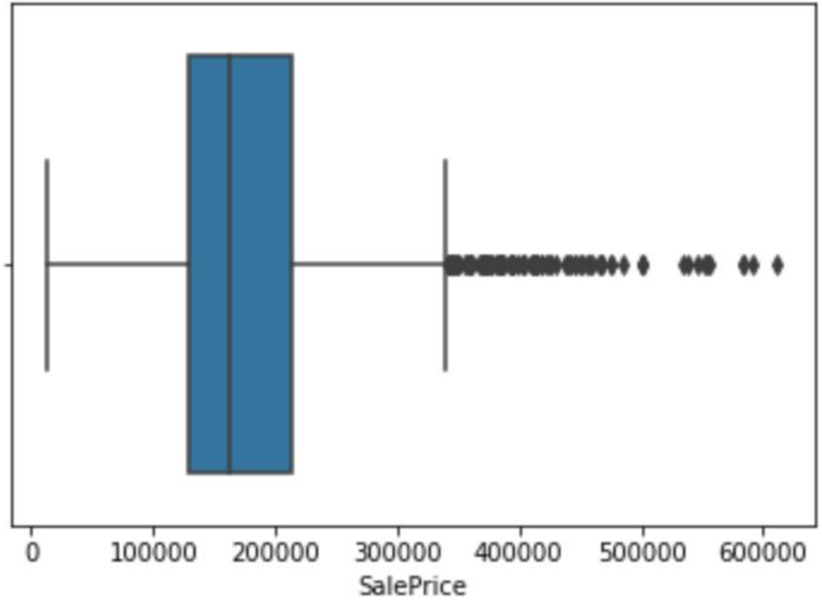


Source: General Assembly lesson by *Matt Brems (DC), Marc Harper (LA):*
http://localhost:8888/notebooks/Desktop/DSI-US-7/Lessons/3.01-lesson-linear_regression/sta
pynb

Distribution of Sale Price Data



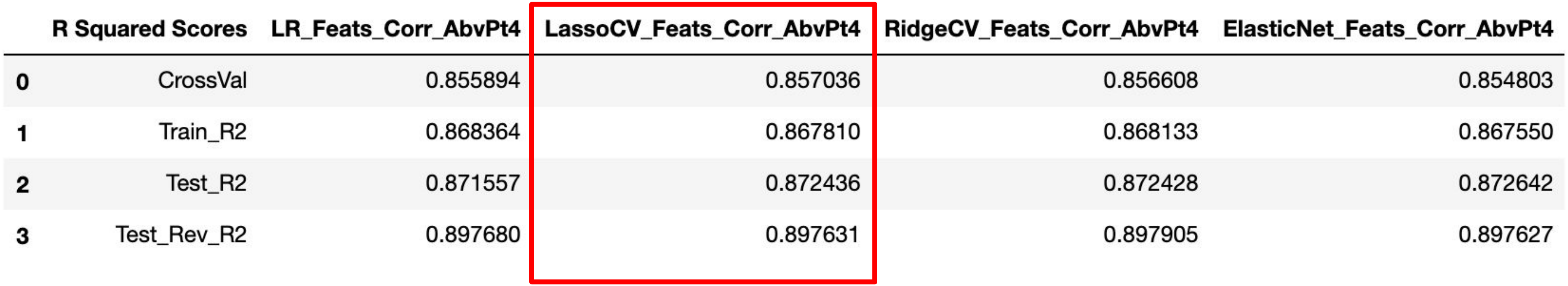
Boxplot Showing Sale Price Outliers



There is a right-skew (positive skew) of the data -> did PowerTransform

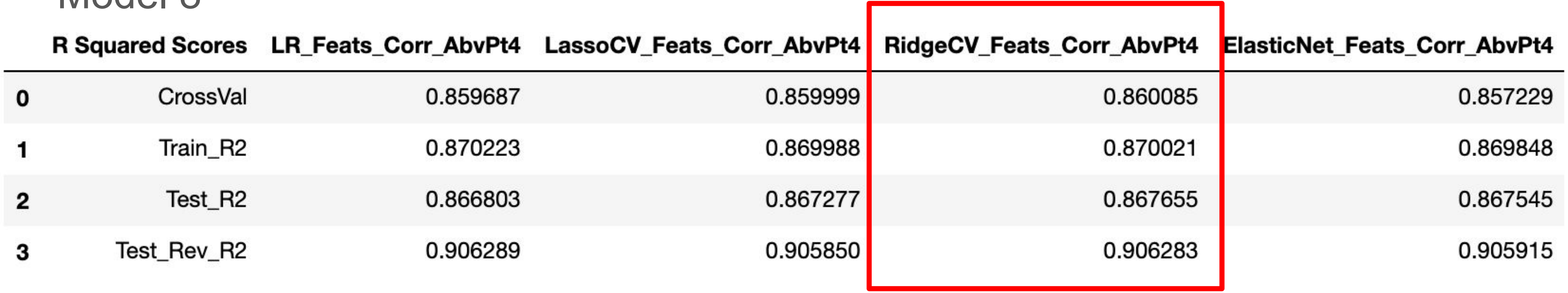
Model R2 Score data - showing how chose model

Model 6



	R Squared Scores	LR_Feats_Corr_AbvPt4	LassoCV_Feats_Corr_AbvPt4	RidgeCV_Feats_Corr_AbvPt4	ElasticNet_Feats_Corr_AbvPt4
0	CrossVal	0.855894	0.857036	0.856608	0.854803
1	Train_R2	0.868364	0.867810	0.868133	0.867550
2	Test_R2	0.871557	0.872436	0.872428	0.872642
3	Test_Rev_R2	0.897680	0.897631	0.897905	0.897627

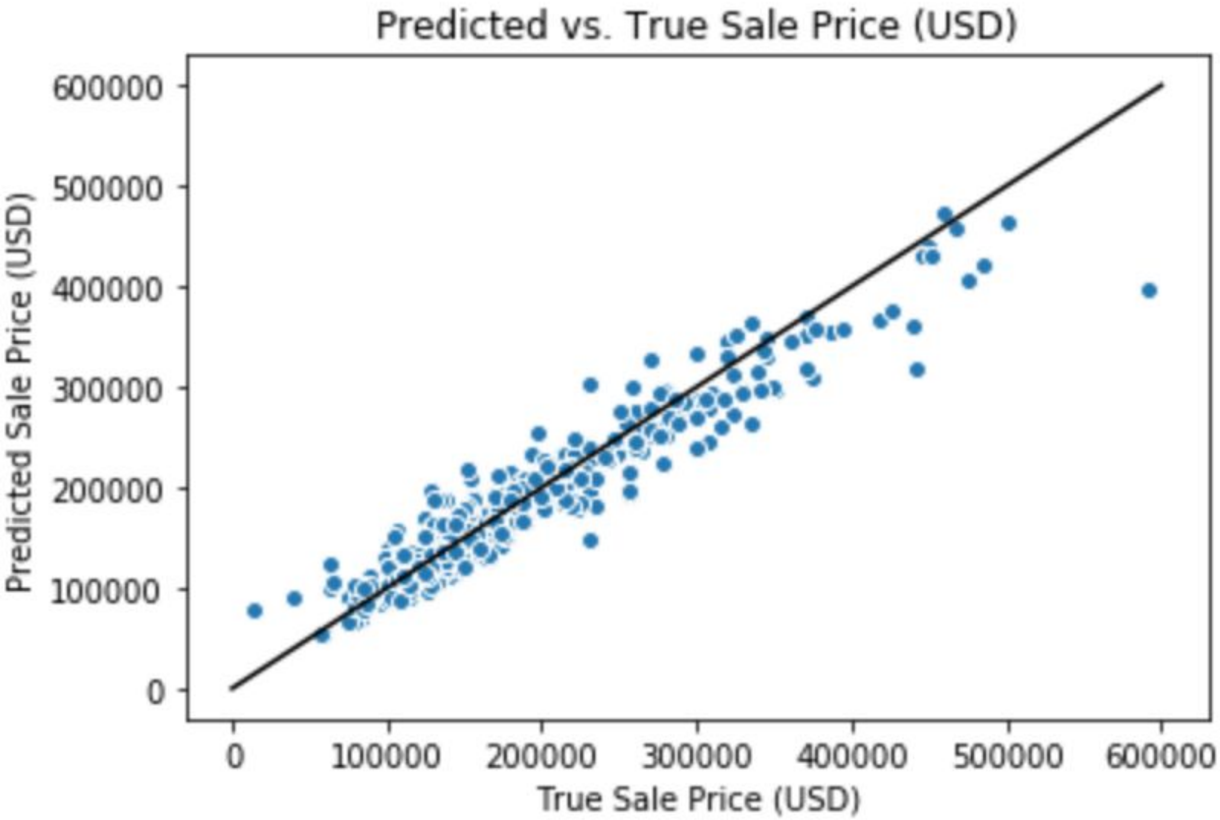
Model 8



	R Squared Scores	LR_Feats_Corr_AbvPt4	LassoCV_Feats_Corr_AbvPt4	RidgeCV_Feats_Corr_AbvPt4	ElasticNet_Feats_Corr_AbvPt4
0	CrossVal	0.859687	0.859999	0.860085	0.857229
1	Train_R2	0.870223	0.869988	0.870021	0.869848
2	Test_R2	0.866803	0.867277	0.867655	0.867545
3	Test_Rev_R2	0.906289	0.905850	0.906283	0.905915

Chose Model 8: More metrics for this model: MSE: approx. 603273361.69 $^2, RMSE: $24561.62,
Mean Absolute Error: $17106.22

Linear Regression Prediction Using RidgeCV Fit to True Values

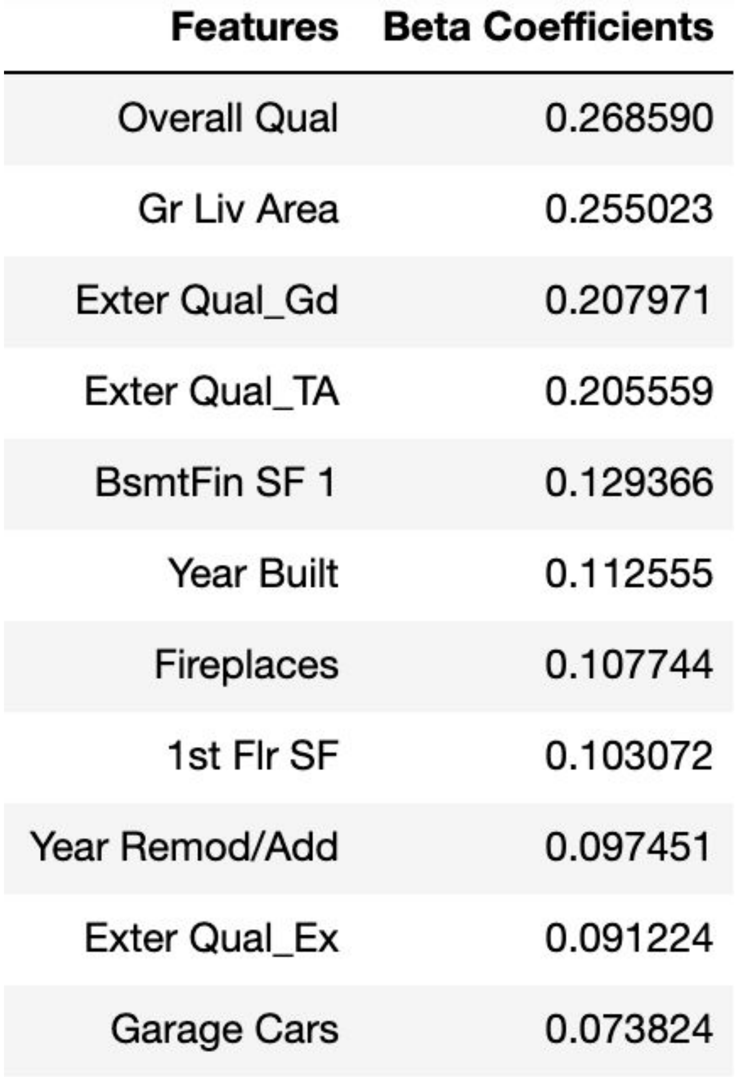


Top 10 features that add to value

These features appear to add the most value to a home

(larger β coefficients have larger affect on sale price prediction)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_p X_{pi}$$

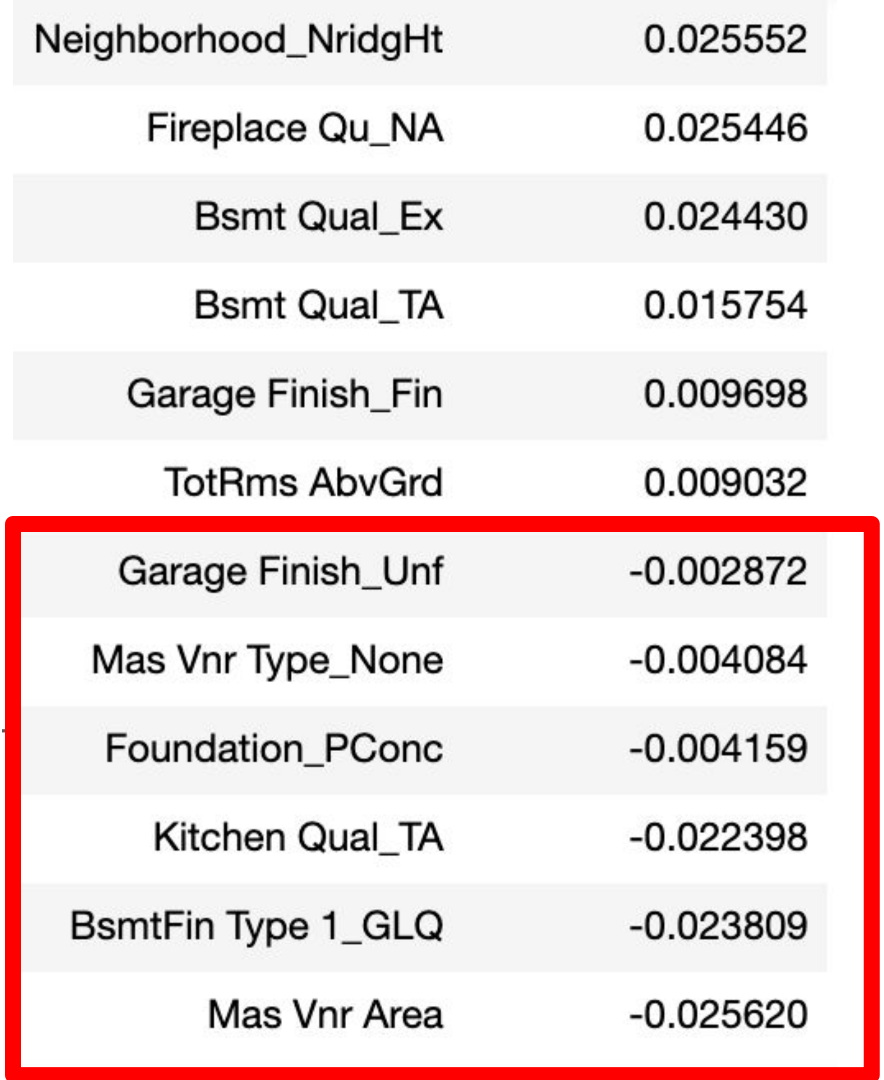


Features	Beta Coefficients
Overall Qual	0.268590
Gr Liv Area	0.255023
Exter Qual_Gd	0.207971
Exter Qual_TA	0.205559
BsmtFin SF 1	0.129366
Year Built	0.112555
Fireplaces	0.107744
1st Flr SF	0.103072
Year Remod/Add	0.097451
Exter Qual_Ex	0.091224
Garage Cars	0.073824

Features that hurt value

The following features hurt the value of a home the most:

- Unfinished garage
- No masonry vaneer type (e.g., if no brick, brick face, cinder block, or stone vaneer)
- Poured concrete foundation type (instead of cinder block, etc.)
- Typical/avg kitchen quality (instead o excellent or good)
- Rating of basement finish type: good living quarters - odd observation
- Masonry vareer area (sq ft)



Neighborhood_NridgHt	0.025552
Fireplace Qu_NA	0.025446
Bsmt Qual_Ex	0.024430
Bsmt Qual_TA	0.015754
Garage Finish_Fin	0.009698
TotRms AbvGrd	0.009032
Garage Finish_Unf	-0.002872
Mas Vnr Type_None	-0.004084
Foundation_PConc	-0.004159
Kitchen Qual_TA	-0.022398
BsmtFin Type 1_GLQ	-0.023809
Mas Vnr Area	-0.025620

Other findings / recommendations

- To increase value of home, homeowners should:
    - Increase overall quality of the home
    - Ensure good quality of exterior (including masonry vaneer)
    - Finish basement if it is unfinished
    - Remodel
    - Finish garage if it is unfinished
    - Increase kitchen quality (need to stand out!)
- Neighborhood that stands out as a good investment:
    - Northridge Heights (NridgHt)
    - Other good neighborhoods: Northridge, Stone Brook, Somerset, Timberland, Veenker, and College Creek (according to corr.)

Next Steps

- Model optimization
- This model can generalize to other city/cities if:
    - Demand and market information available (*e.g.*, general growth rates would help to scale the model)
    - Data similar to data used to build this model, especially variables on heatmap on Slide 4
- To make the model more universal (*e.g.*, to general U.S. regions):
    - Include data from various areas in U.S. (weighted equally for enough representation of each region)
    - Scaling factors for regions in U.S. with different priorities (*e.g.,* structural features needed in flood-prone areas)

Sources

City of Ames, https://www.cityofames.org/about-ames/interesting-facts-about-ames

Dean De Cock, http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

General Assembly lesson by *Kiefer Katovich (SF), Minor updates by David Yerrington (SF):*
http://localhost:8888/notebooks/Desktop/DSI-US-7/Lessons/2.04-lesson-eda/2_04-basic-eda-walkthrough.ipynb#cov_cor

General Assembly lesson by *Matt Brems (DC), Marc Harper (LA):*
http://localhost:8888/notebooks/Desktop/DSI-US-7/Lessons/3.01-lesson-linear_regression/starter-code.ipynb

https://www.google.com/search?biw=1280&bih=583&tbm=isch&sa=1&ei=pj2UXOzBPIO7jwS3_ZfgDQ&q=birds+eye+view+ames%2C+iowa+current&oq=birds+eye+view+ames%2C+iowa+current&gs_l=img.3...3318625.3321015..3321206...0.0..0.0.0.......13....1..gws-wiz-img.jSwJteC57P8#imgrc=Klagv1qes1xiKM:

Thank you!