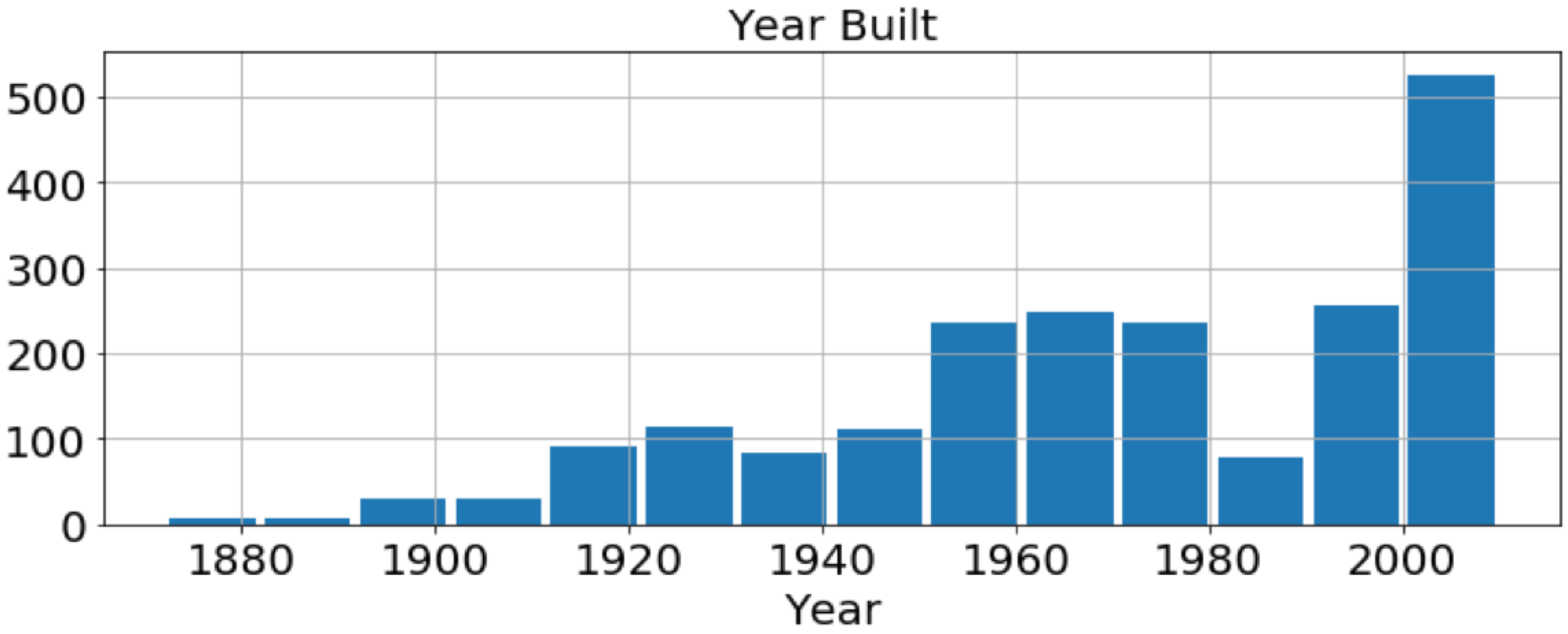
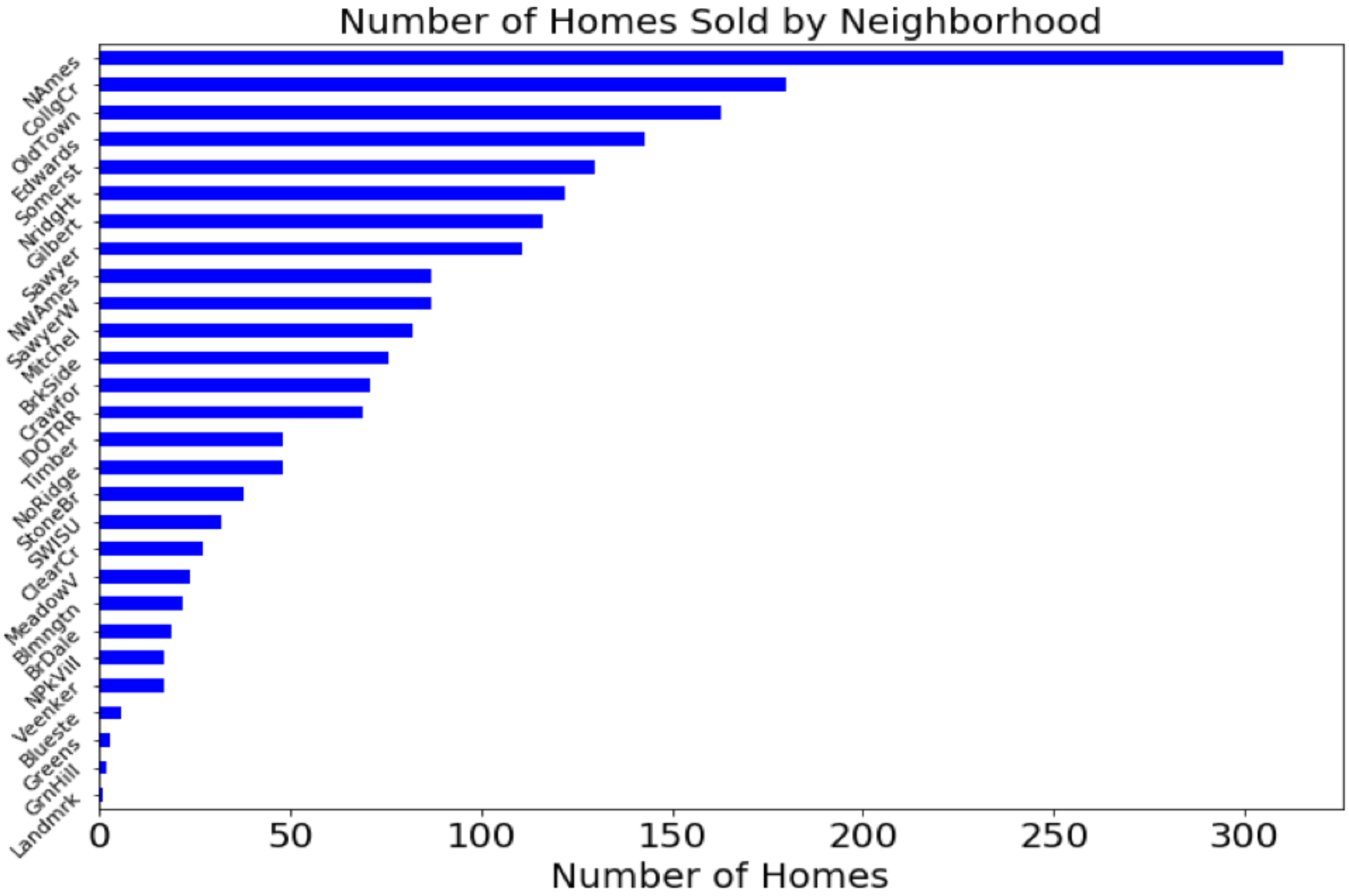
A Model to Estimate Home Values in Ames, Iowa

- **Goal**: to build a model that will make the best possible prediction of home values in Ames, Iowa.

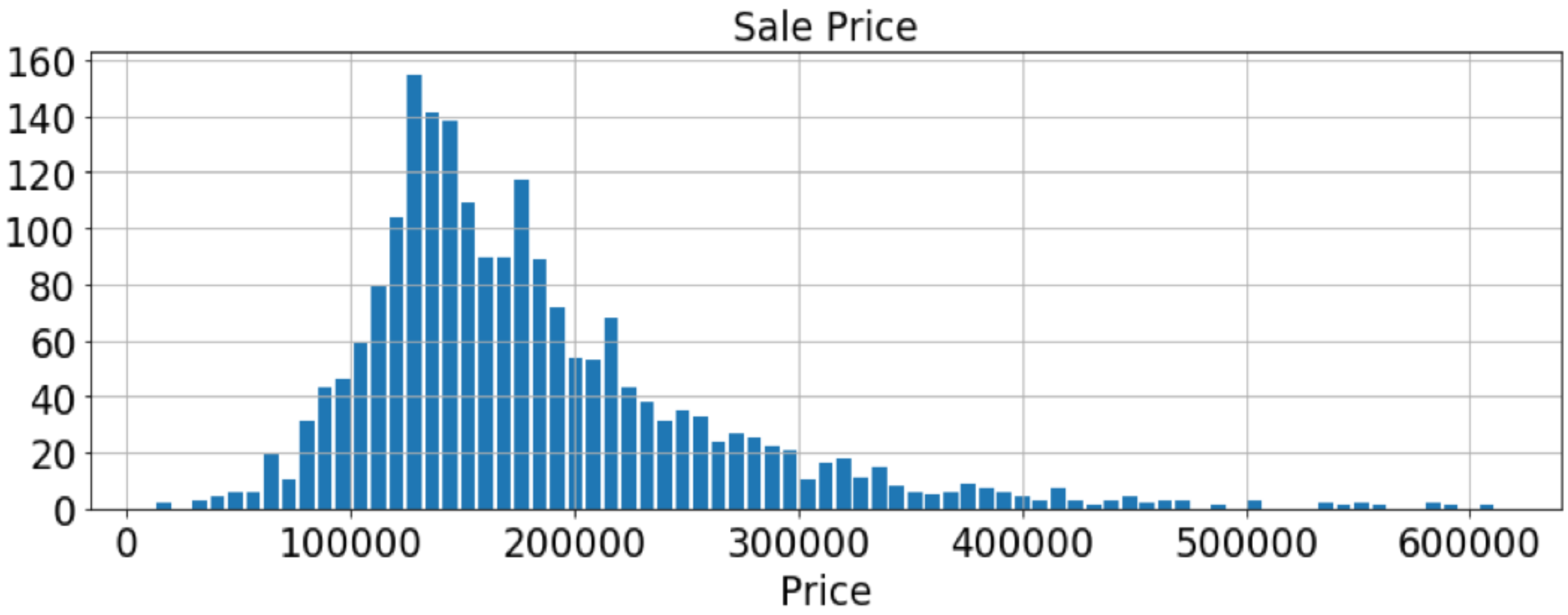
- **Materials/Data**:
    - A train data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010.
        - 79 variables related to the quality and quantity of many physical attributes of the property.
        - 51 columns were categorical and 28 were continuous.
    - A test data with which to feed data into the regression model.



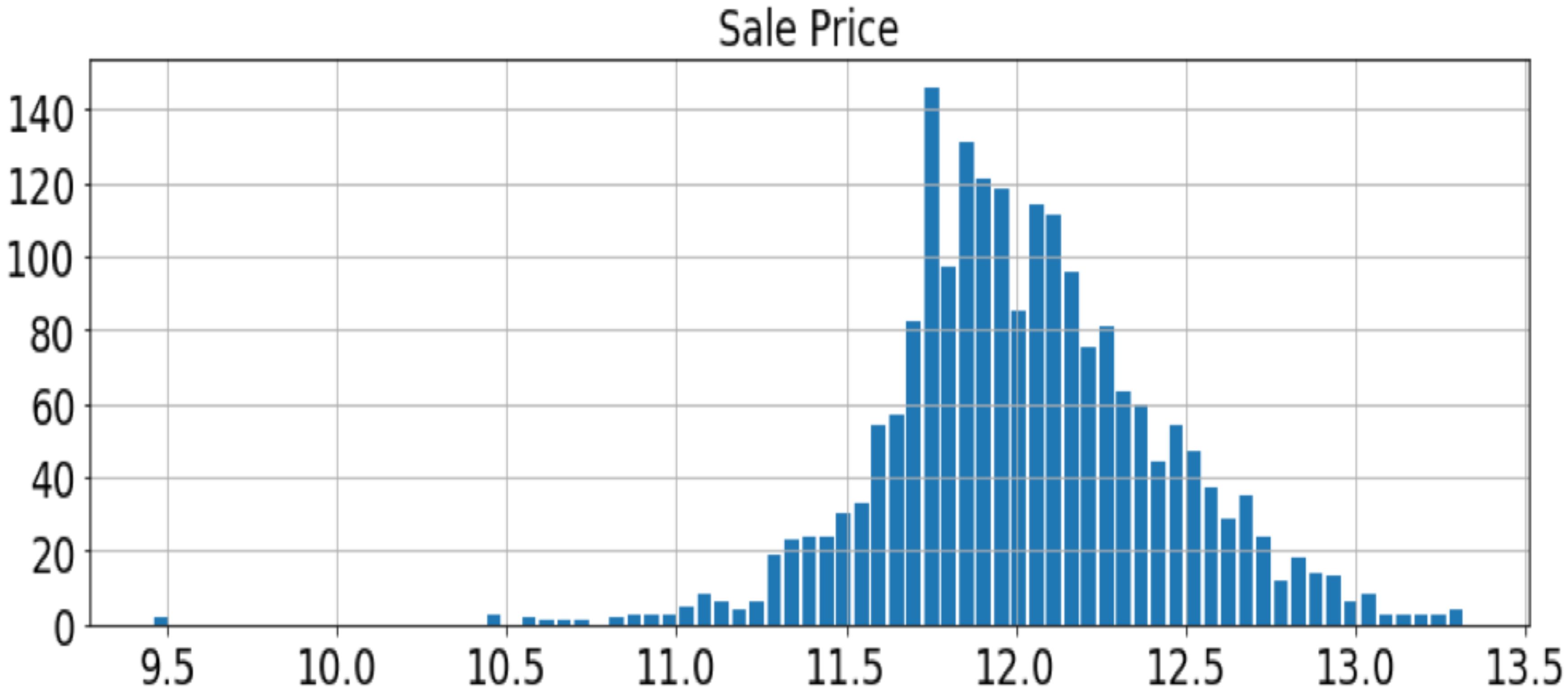
- The oldest house was built in 1872.
- The newest house was built in 2010.



Number of Homes Sold by Neighborhood

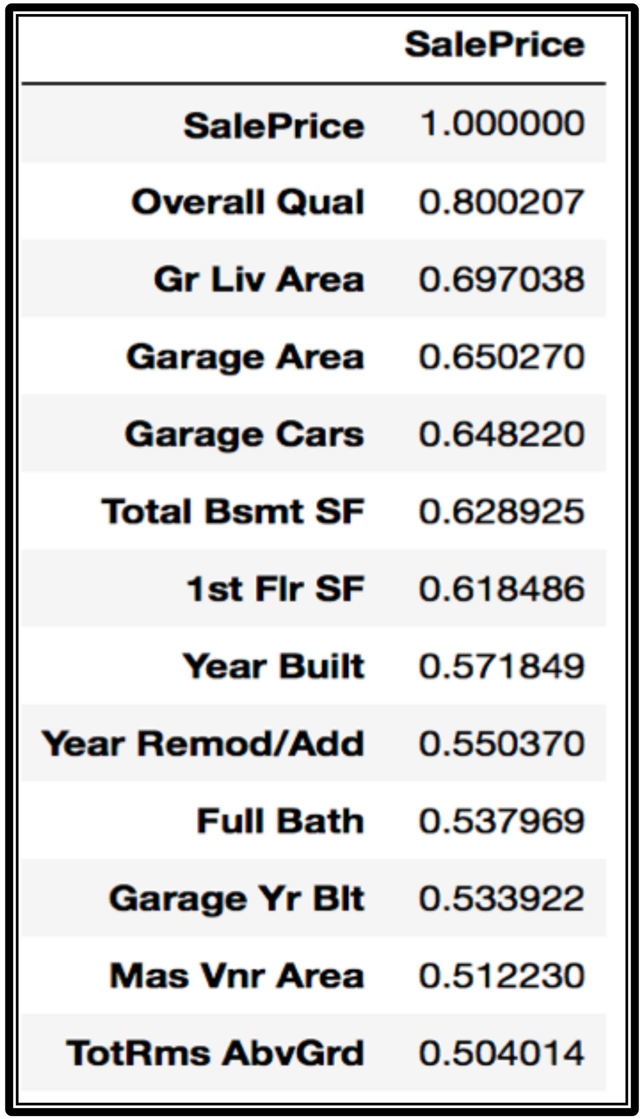


- The average sales price was $181,470.
- The median price was $162,500

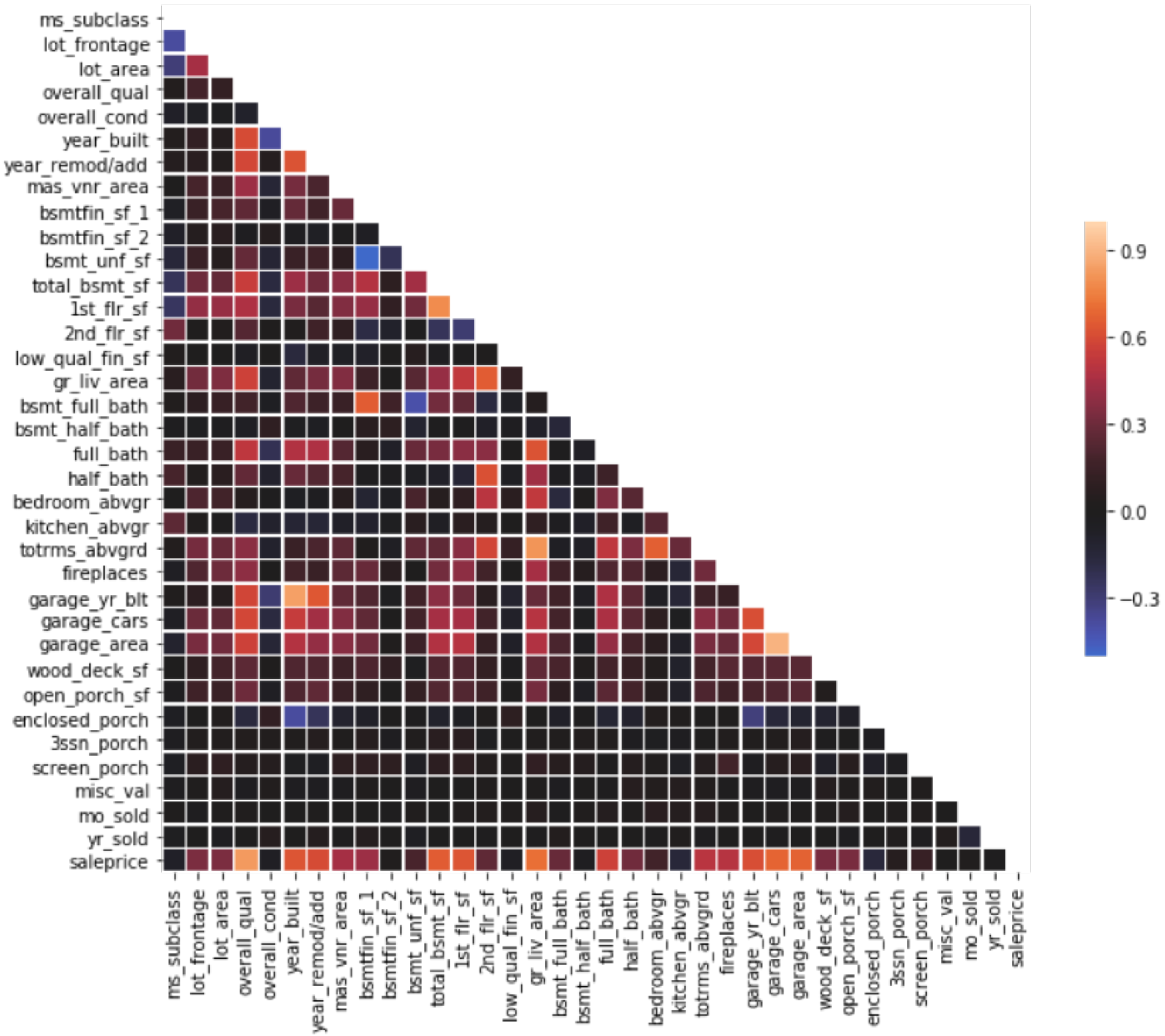


I applied a log transformation to the Sale Price to see if it could make the skewed sale price take on a more normal distribution.

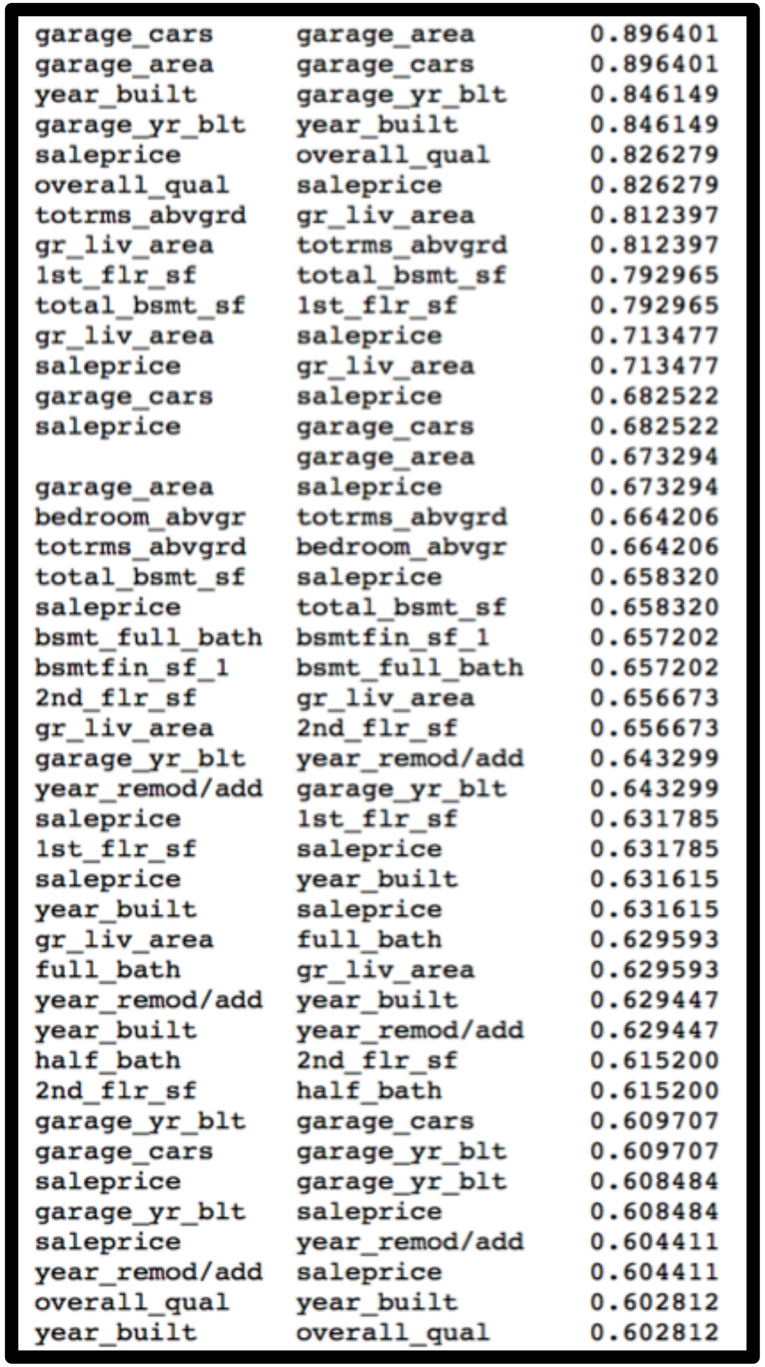
This table shows that features associated with square footage or quality have a strong correlation to the sale price.



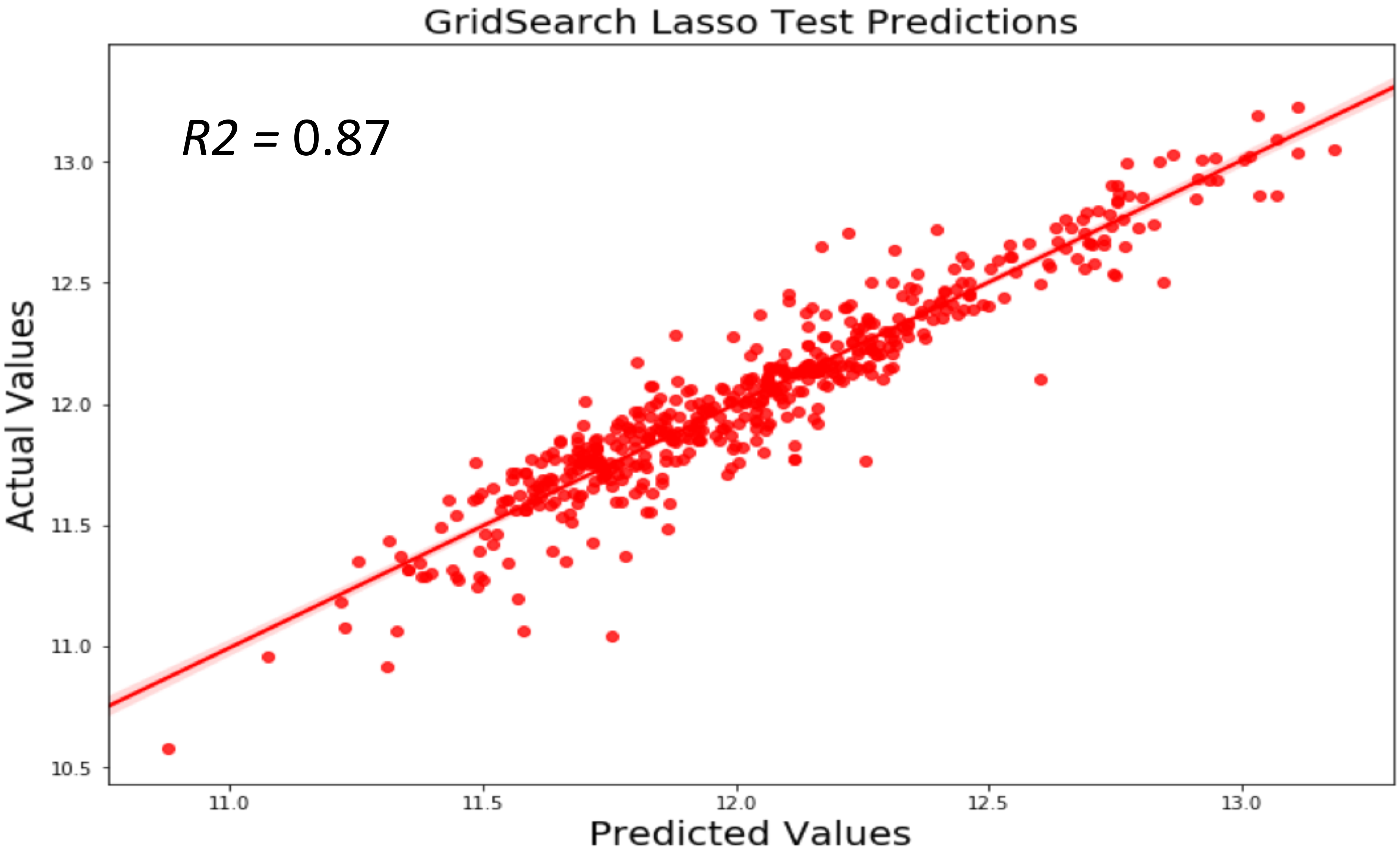
	SalePrice
SalePrice	1.000000
Overall Qual	0.800207
Gr Liv Area	0.697038
Garage Area	0.650270
Garage Cars	0.648220
Total Bsmt SF	0.628925
1st Flr SF	0.618486
Year Built	0.571849
Year Remod/Add	0.550370
Full Bath	0.537969
Garage Yr Blt	0.533922
Mas Vnr Area	0.512230
TotRms AbvGrd	0.504014


ms_subclass
lot_frontage
lot_area
overall_qual
overall_cond
year_built
year_remod/add
mas_vnr_area
bsmtfin_sf_1
bsmtfin_sf_2
bsmt_unf_sf
total_bsmt_sf
1st_flr_sf
2nd_flr_sf
low_qual_fin_sf
gr_liv_area
bsmt_full_bath
bsmt_half_bath
full_bath
half_bath
bedroom_abvgr
kitchen_abvgr
totrms_abvgrd
fireplaces
garage_yr_blt
garage_cars
garage_area
wood_deck_sf
open_porch_sf
enclosed_porch
3ssn_porch
screen_porch
misc_val
mo_sold
yr_sold
saleprice
ms_subclass
lot_frontage
lot_area
overall_qual
overall_cond
year_built
year_remod/add
mas_vnr_area
bsmtfin_sf_1
bsmtfin_sf_2
bsmt_unf_sf
total_bsmt_sf
1st_flr_sf
2nd_flr_sf
low_qual_fin_sf
gr_liv_area
bsmt_full_bath
bsmt_half_bath
full_bath
half_bath
bedroom_abvgr
kitchen_abvgr
totrms_abvgrd
fireplaces
garage_yr_blt
garage_cars
garage_area
wood_deck_sf
open_porch_sf
enclosed_porch
3ssn_porch
screen_porch
misc_val
mo_sold
yr_sold
saleprice
0.9
0.6
0.3
0.0
−0.3

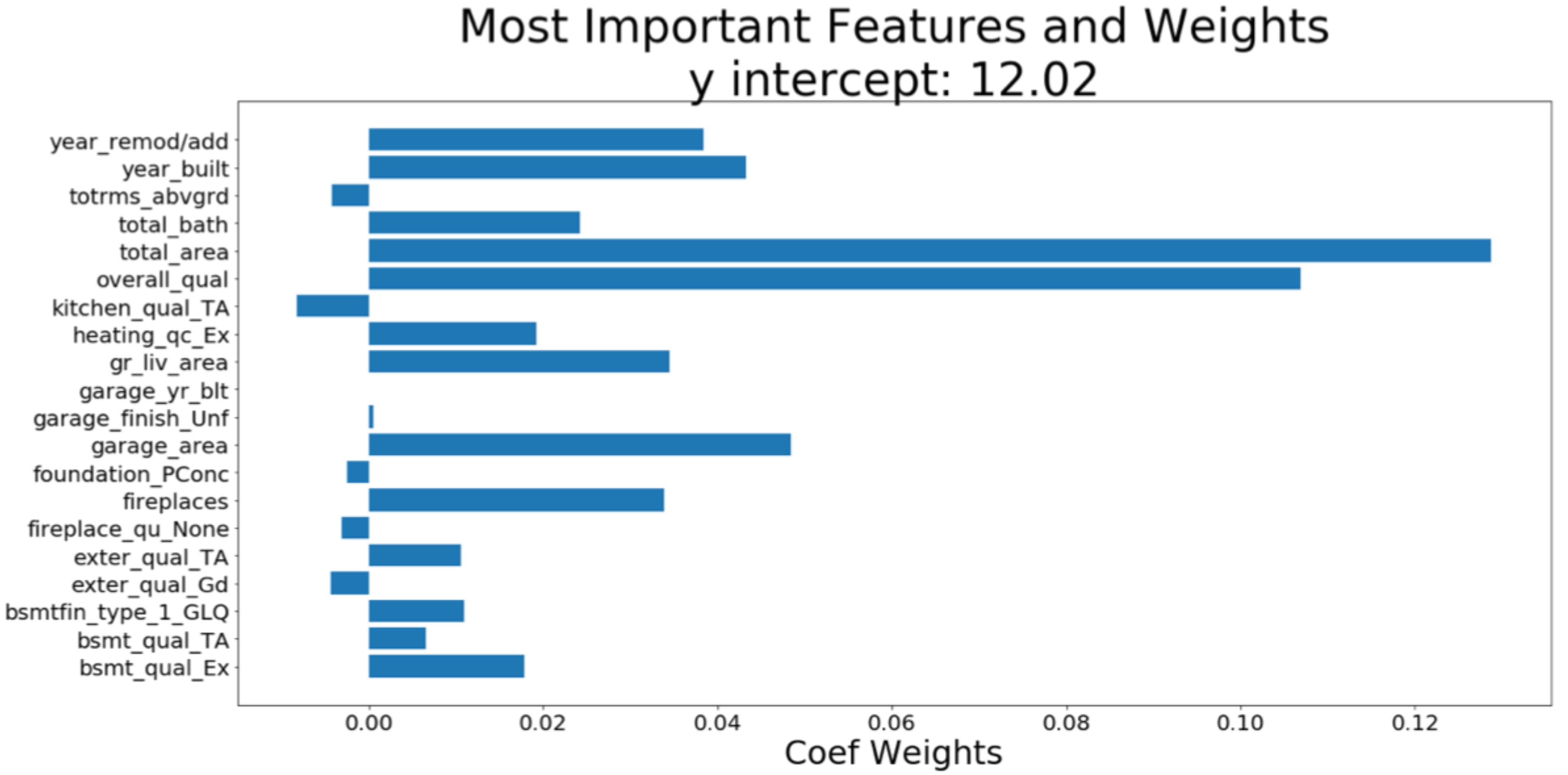
- To deal with multicollinearity I dropped variables that were highly correlated with others and grouped together multiple features that could be defined by one total feature.
- Example:
  - Dropped garage_cars and and kept garage_area.
  - Combined the half baths and full baths into one total bathroom feature



```
garage_cars     garage_area      0.896401
garage_area     garage_cars      0.896401
year_built      garage_yr_blt    0.846149
garage_yr_blt   year_built       0.846149
saleprice       overall_qual     0.826279
overall_qual    saleprice        0.826279
totrms_abvgrd   gr_liv_area      0.812397
gr_liv_area     totrms_abvgrd    0.812397
1st_flr_sf      total_bsmt_sf    0.792965
total_bsmt_sf   1st_flr_sf       0.792965
gr_liv_area     saleprice        0.713477
saleprice       gr_liv_area      0.713477
garage_cars     saleprice        0.682522
saleprice       garage_cars      0.682522
                garage_area      0.673294
garage_area     saleprice        0.673294
bedroom_abvgr   totrms_abvgrd    0.664206
totrms_abvgrd   bedroom_abvgr    0.664206
total_bsmt_sf   saleprice        0.658320
saleprice       total_bsmt_sf    0.658320
bsmt_full_bath  bsmtfin_sf_1     0.657202
bsmtfin_sf_1    bsmt_full_bath   0.657202
2nd_flr_sf      gr_liv_area      0.656673
gr_liv_area     2nd_flr_sf       0.656673
garage_yr_blt   year_remod/add   0.643299
year_remod/add  garage_yr_blt    0.643299
saleprice       1st_flr_sf       0.631785
1st_flr_sf      saleprice        0.631785
saleprice       year_built       0.631615
year_built      saleprice        0.631615
gr_liv_area     full_bath        0.629593
full_bath       gr_liv_area      0.629593
year_remod/add  year_built       0.629447
year_built      year_remod/add   0.629447
half_bath       2nd_flr_sf       0.615200
2nd_flr_sf      half_bath        0.615200
garage_yr_blt   garage_cars      0.609707
garage_cars     garage_yr_blt    0.609707
saleprice       garage_yr_blt    0.608484
garage_yr_blt   saleprice        0.608484
saleprice       year_remod/add   0.604411
year_remod/add  saleprice        0.604411
overall_qual    year_built       0.602812
year_built      overall_qual     0.602812
```



Variance Threshold: [0, .05, .1], Kbest: [10, 15, 20], Alpha: np.logspace(-3,3,7)


Most Important Features and Weights
y intercept: 12.02
year_remod/add
year_built
totrms_abvgrd
total_bath
total_area
overall_qual
kitchen_qual_TA
heating_qc_Ex
gr_liv_area
garage_yr_blt
garage_finish_Unf
garage_area
foundation_PConc
fireplaces
fireplace_qu_None
exter_qual_TA
exter_qual_Gd
bsmtfin_type_1_GLQ
bsmt_qual_TA
bsmt_qual_Ex
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Coef Weights

- Suggestions for improved modeling:
    - Better feature engineering and subset selection.
        - The most important features in all of my models were 'overall_quality', 'total_area', 'gr_liv_area', garage area', 'year_remodeled/add' and 'year_built'. I could run my models with these as the only features to see if it improves their accuracy.
        - I can also do more feature engineering to try and reduce the number of redundant variables, especially after I got dummies of all my categorical columns.
    - Finally, I might consider log transforming individual features that had above 0.7 or 0.8 in order to make their distributions closer to normal.
    - Adjust the Kbest and variance threshold parameters in my modeling.