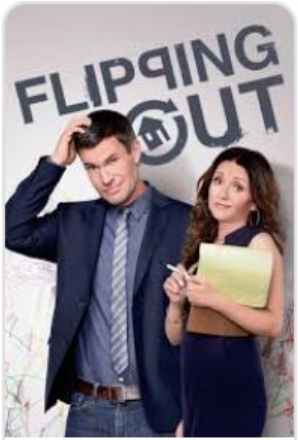
Pricing home renovations for real estate investment

Matt Hope, Aidan Hughes, Steven Lantigua, Garrett Sikes

NYC Data Science Academy, Fall 2020

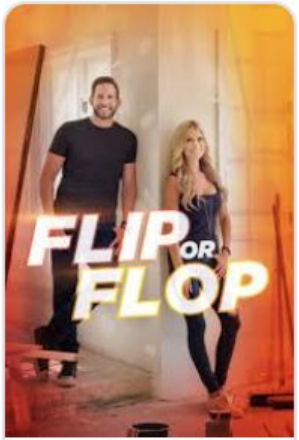
Real estate investment is an area of cultural fascination...



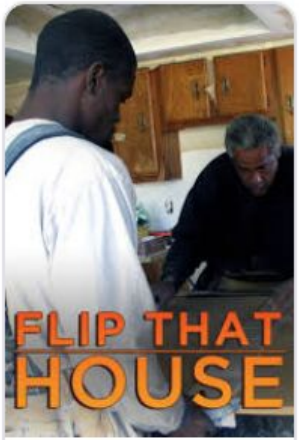
Flipping Out

Flip This House

Masters of Flip



Flip or Flop



Flip That House

Flipping Vegas

...that glosses over the financial reality of fixing and flipping a property.

Homes flipped in 2019 represent 8.6% of all sales, which is an **eight-year high**

However, median gross profit on flipped homes dropped 3.2% to $62,900, which is an **eight-year low**

(According to ATTOM Data Solutions)

Source: https://www.attomdata.com/news/market-trends/flipping/attom-data-solutions-year-end-2019-u-s-home-flipping-report/

Pain points for flipping properties:

- Technical expertise and capital are required

  Experience with the local market, construction and renovation know-how, and financing are all vital.

- Carrying costs for a renovation make flipping time-sensitive

  Utilities, insurance, interest on financing, and property taxes add up quickly.

What aspects of a property renovation should be prioritized?

1) The expected increase in sale price outweighs cost of renovation.

2) The amount of time needed to renovation should be minimal.

Approach: Use Ames Housing Dataset to estimate sale price increases from renovation

- Records of approximately 2,500 home sales from 2006-2010
- Extensive number of features ranging from square footage of the home to location and zoning information (59 in total)
- A classic dataset in the machine learning community to explore models.

Data Cleaning / Pre-processing

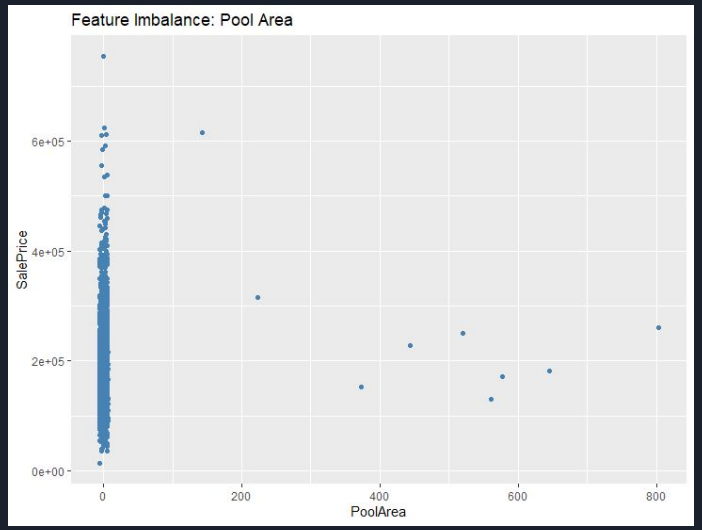
- Missing Values
- Ordinal Variables
- Categorical Variables
- Feature Imbalance

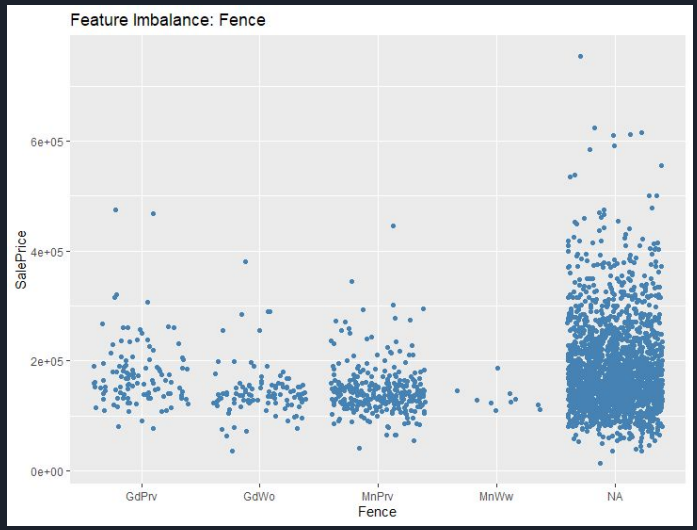
Missing Values

- Many NA's were coded intentionally in the original dataset when a rated feature was missing from a household
  - E.g. Quality
  - Straightforward fix: encode as zeros on the ordinal scale

- Other features had NA's that were genuinely just missing values
  - Very small # relative to overall number of rows
  - Replace with mode (categorical) or mean (numerical) of feature

Feature Imbalance





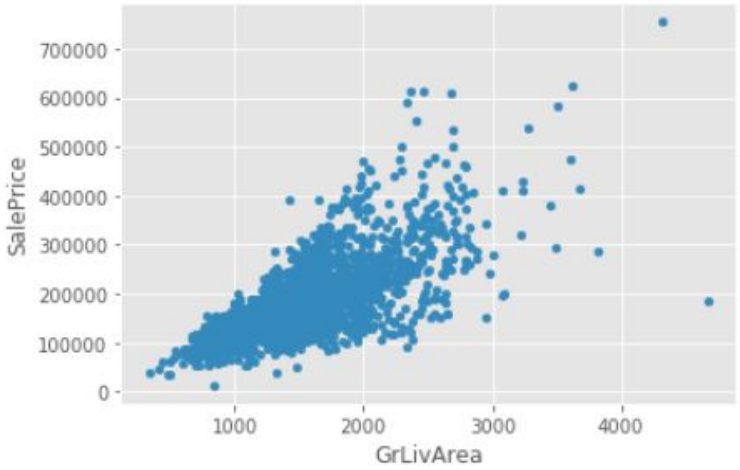
- Sometimes a feature was missing from so many homes that utilizing it within any model would be fruitless.
  - Challenge: some features weren't as clear-cut re: including or excluding based on feature imbalance... What % should we consider usable/unusable?

Outliers

- Additional fine tuning included testing models with exclusion of outliers
  - SalePrice
  - GrLivingArea

```
1 housing.plot.scatter(x='GrLivArea', y='SalePrice')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19963df06c8>
```



Ordinal and Categorical Variables

```
#assign values to ORDINAL categorical variables
#doing this manually bc label_encoder doesn't seem to preserve correct order of values

ordinal_features_1 = ['ExterQual', 'ExterCond', 'BsmtQual', 'BsmtCond', 'HeatingQC', 'KitchenQual']

ordinal_values_1 = {'Po':1,'Fa':2,'TA':3,'Gd':4,'Ex':5}
cleanup_dict = {}

for i in ordinal_features_1:
    cleanup_dict[i] = ordinal_values_1


ordinal_values_2 = {'No':1, 'Mn':2, 'Av':3, 'Gd':4}
cleanup_dict['BsmtExposure'] = ordinal_values_2

ordinal_values_3 = {'Unf':1, 'LwQ':2, 'Rec':3, 'BLQ':4, 'ALQ':5, 'GLQ':6}
cleanup_dict['BsmtFinType1'] = ordinal_values_3
cleanup_dict['BsmtFinType2'] = ordinal_values_3

ordinal_values_4 = {'Unf':1, 'RFn':2, 'Fin':3}
cleanup_dict['GarageFinish'] = ordinal_values_4
```

- Needed to encode ordinal/categorical variables so ML models could take them into consideration
  - Challenge: properly encoding/preserving priority of ordinal values

Quick note:

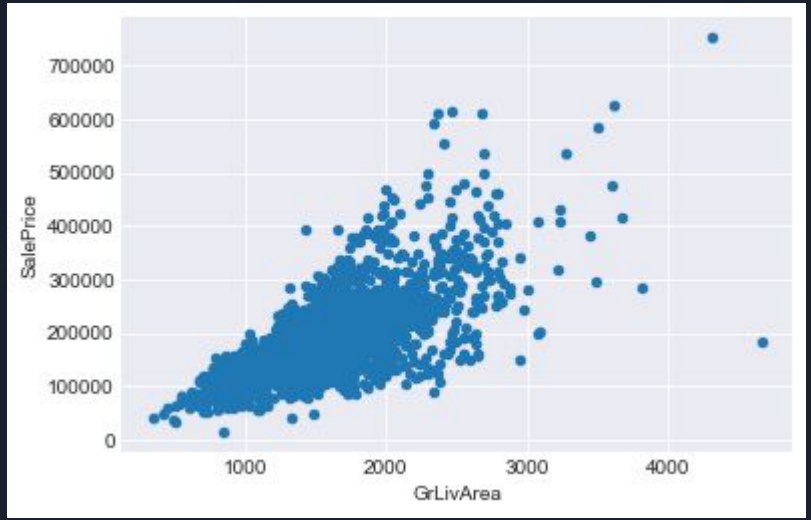
normalize=True vs normalize=False when working with encoded features

GarageType_Attchd	1116.033409
GarageType_Basment	-5935.613663
GarageType_BuiltIn	3295.598803
GarageType_CarPort	321.989470
GarageType_Detchd	-2066.599946

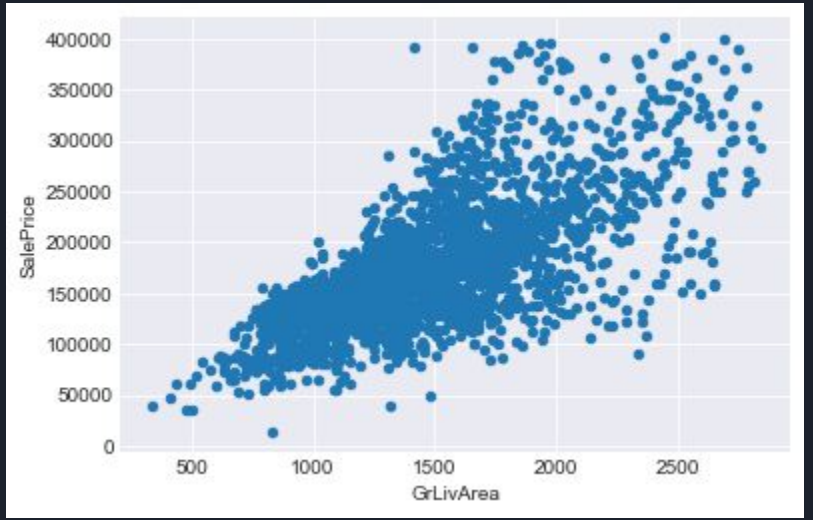
GarageType_Attchd	-1213.223054
GarageType_Basment	-3166.221179
GarageType_BuiltIn	-3587.902138
GarageType_CarPort	5661.592653
GarageType_Detchd	1748.368945

(we'll be showing you normalized data from here on out)

Outliers: Z-score > 3



Before



After

56 of 2559 rows removed

Problem: Feature selection greatly informs model predictions, but what's important?

Original Data:
81 features

Targeted Cleaned Data:
59 features

Full Cleaned Data:
280 features

Multiple Approaches

1. Selection by multiple linear regression
2. Tree-based feature importance

Standard errors on the features of multiple linear indicate important features

Multiple Linear Regression

↓

Five-fold cross validation

↓

Rank p-values of coefficients

	0	0	0	0	0	Mean
GrLivArea	2.872641e-22	6.443073e-31	6.106472e-21	5.268445e-23	2.370502e-25	1.289331e-21
OverallCond	4.871631e-21	3.162150e-23	1.059943e-22	1.637106e-19	8.450763e-20	5.064550e-20
OverallQual	3.877112e-19	1.774581e-21	2.846298e-19	4.312880e-20	5.257292e-19	2.485947e-19
TotalBsmtSF	2.852021e-20	3.422339e-23	5.064019e-15	2.098095e-17	2.919845e-19	1.017064e-15
BsmtFinSF1	1.266732e-14	4.811976e-19	2.005698e-16	1.327495e-15	3.276270e-21	2.839175e-15
ExterQual_Gd	1.407028e-15	1.992728e-17	5.154114e-14	7.206497e-13	1.100196e-17	1.547258e-13
BsmtQual_Gd	1.191265e-14	7.930602e-13	6.012071e-16	3.869013e-13	1.889088e-12	6.163127e-13

Both the targeted and full dataset converge on similar, important features

Targeted Cleaned Data:
59 features

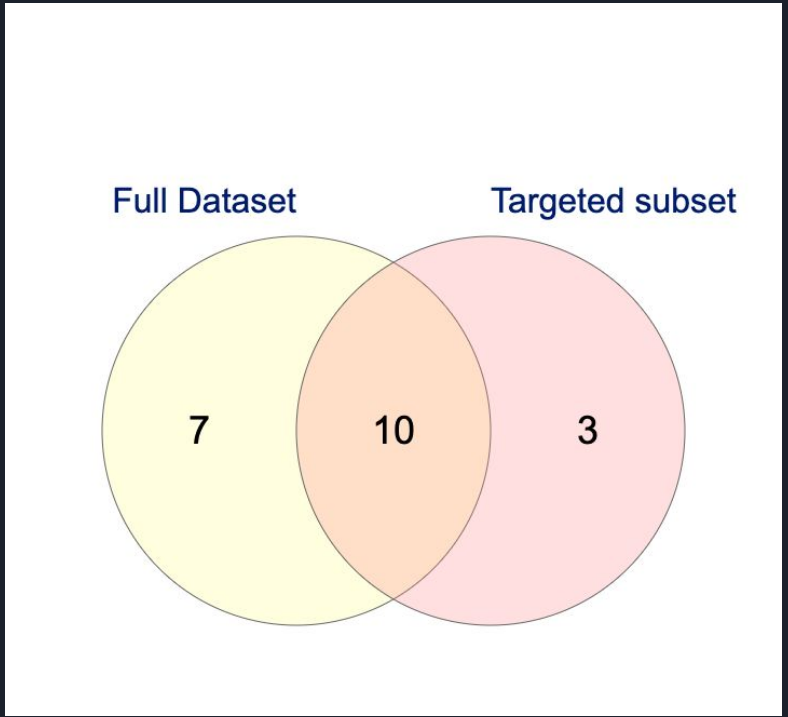
↓

13 features
(22%)

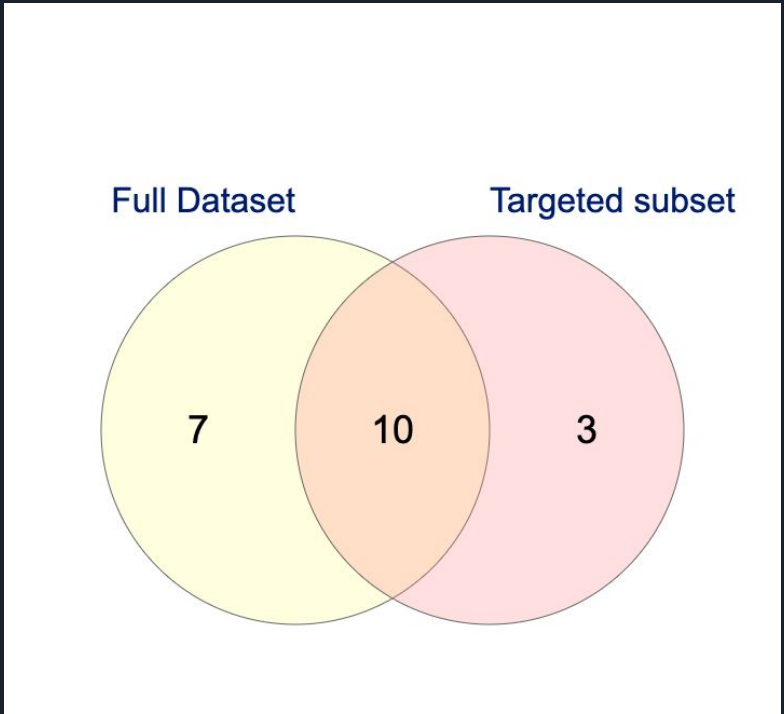
Full Cleaned Data:
280 features

↓

34 features
(12%)



Both the targeted and full dataset converge on similar, important features



BsmtFinSF1
ExterQual
GrLivArea
KitchenQual
LotArea
MasVnrArea
OverallCond
OverallQual
TotalBsmtSF
YearBuilt

Both the targeted and full dataset converge on similar, important features

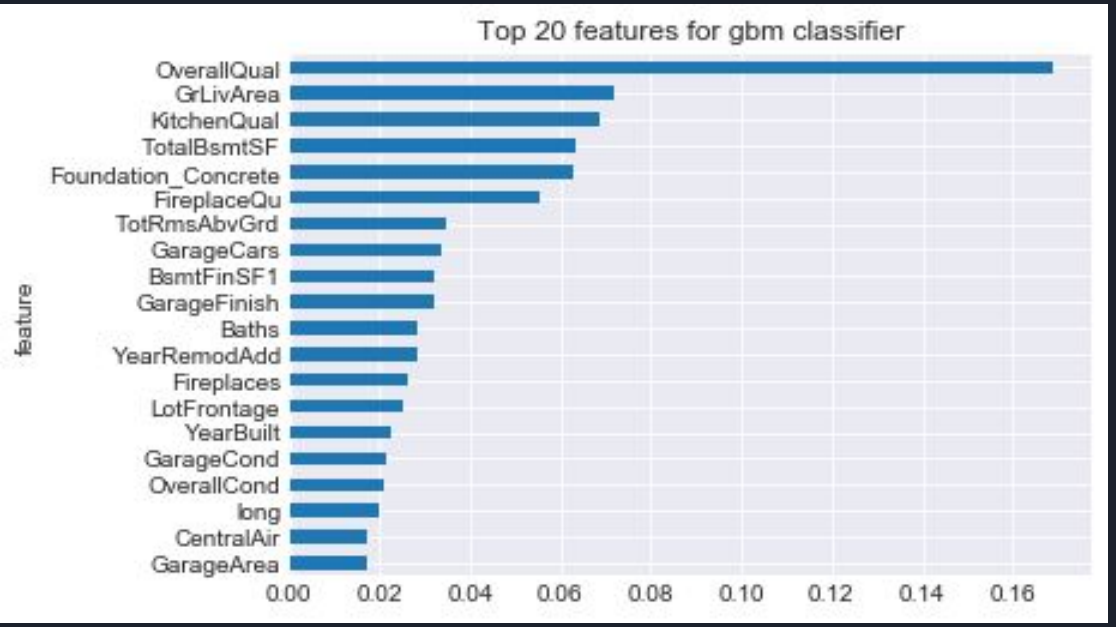
Full Dataset

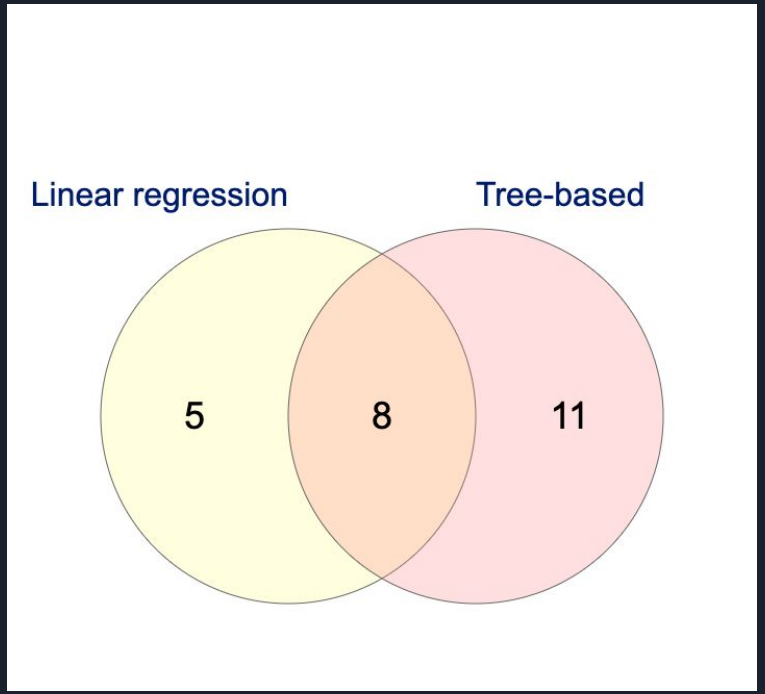
```
Train R2:  0.9234280355471544
Test R2:  0.9456976442289615
Train R2:  0.9356646431904241
Test R2:  0.92054041517471 39
Train R2:  0.9216717021130378
Test R2:  0.9496236984341639
Train R2:  0.9230707049005569
Test R2:  0.9432483444017128
Train R2:  0.9266163625091186
Test R2:  0.9336066060106832
```

Targeted Subset

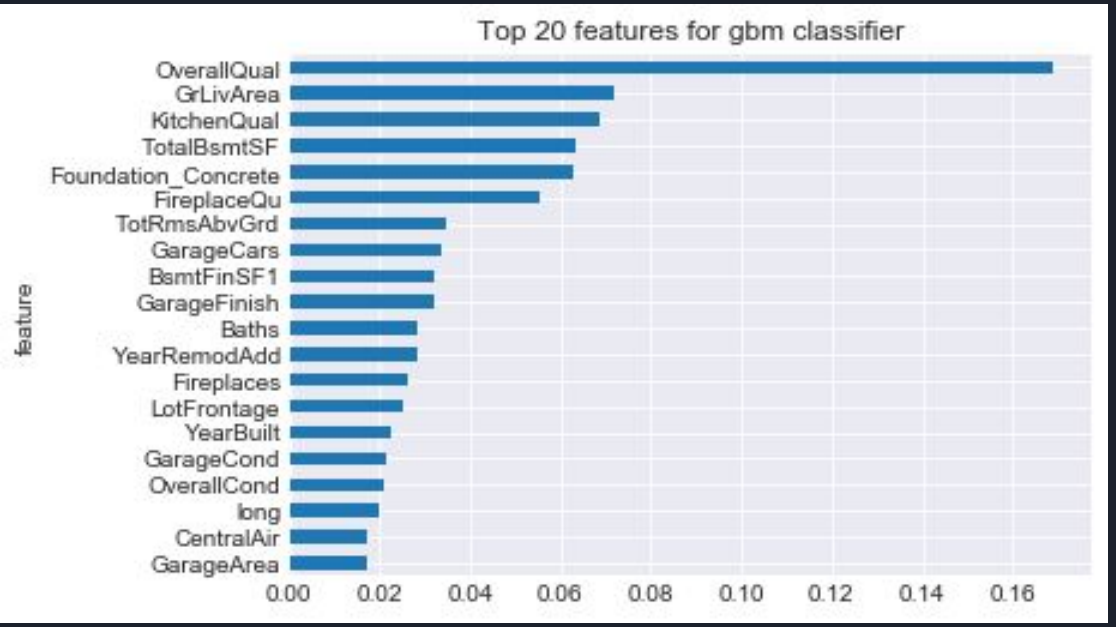
```
Train R2:  0.8671418976398184
Test R2:  0.8847112728871297
Train R2:  0.8795937946446714
Test R2:  0.8450861928516264
Train R2:  0.8651783180549923
Test R2:  0.8918588654588644
Train R2:  0.8679194711387048
Test R2:  0.8774979354281827
Train R2:  0.8696948175865222
Test R2:  0.873471672881484
```

Feature importance from tree-based model also shows good overlap with selection by multiple linear regression





Feature importance from tree-based model also shows good overlap with selection by multiple linear regression



BsmtFinSF1
GarageArea
GrLivArea
KitchenQual
OverallCond
OverallQual
TotalBsmtSF
YearBuilt

LassoCV coefficient significance

Significant features

ExterQual	15115.150914
OverallQual	11278.667359
KitchenQual	8966.072474
Lot_CulDSac	8113.718520
Fireplaces	7040.967812
Foundation_Concrete	6495.444084
OverallCond	4684.012011
YearBuilt	246.304700
LotFrontage	106.933281
GrLivArea	49.919106
ScreenPorch	42.669911
GarageArea	29.914517
BsmtFinSF1	23.295081
TotalBsmtSF	21.144795
LotArea	0.532227

Real World - Exterior Quality

- Average Cost of Home Exterior Makeover: $7,700
  - Includes exterior painting, landscaping, door and window update, porch railing, and decoration

ExterQual	15115.150914

- Significant bang for your buck when it comes to renovating home
  - One of the cheaper parts of home to revamp, adds significant value

Real World - Kitchen Quality

- Average cost of renovating kitchen
  - $16,600 or $150 per square foot
  - Ranges from $12,800 to $21,200
- 80% of homebuyers list the kitchen in their top three most important spaces
- The national average ROI for renovating kitchens
  - Small project: 81%
  - Upscale: 54%
  - In Iowa it is around 68%
  -

KitchenQual	8966.072474

Real World - Fireplaces

- National cost of remodeling a fireplace
  - Ranges from $390 to just over $2000
- Can boost a home's value by as much as $15,000 in certain parts of the country

Fireplaces	7040.967812

A note on varying feature impact:

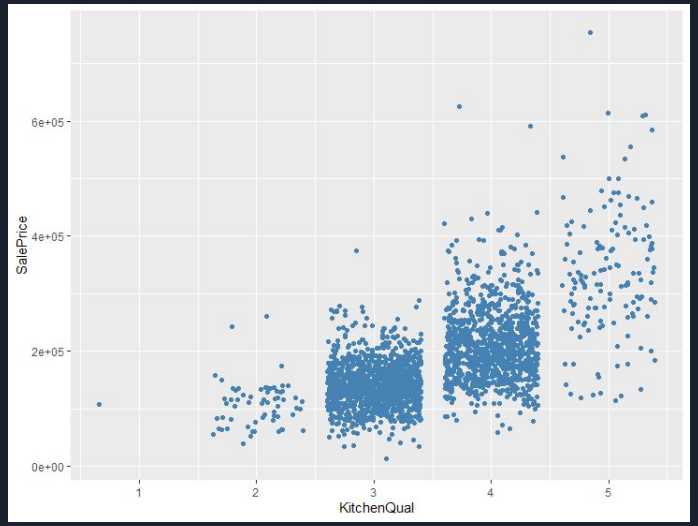
- Our model provides a great, generalizable tool to understand which types of renovations flippers should prioritize in a home

- However, high-quality renovations are not the only factors in a home's sale performance

- If other features / driving factors (that may or may not be controllable by a flipper) are lacking, upgrading features that can be easily improved isn't 100% guaranteed to have the impact the flipper might hope for

- The impact size of those upgrades might also have varying impacts based on other features...

..... such as Neighborhood



..... (the same holds true for some hard-to-change features too)