

It's all about location, right?



PREDICTING HOUSING PRICES, MACHINE LEARNING

Prepared by Katie Kwan, Data Scientist, [LinkedIn](#)
July 2024, [Github](#)

A G E N D A

- Project Overview
- Project Objectives
- Processing Methodology
- Predictive Modelling
- Descriptive Modelling
- Conclusion
- Appendix

PROJECT OVERVIEW

This project uses the AMES Iowa housing dataset to study the housing market, develop predictive models for pricing, and produces an easy to use online quiz, for realtors and homeowners to quickly calculate house prices.

This dataset was initially hosted on [Kaggle](#).

Full codework for this presentation can be found on Katie Kwan's [Github](#).



A G E N D A

- Background
- **Project Objectives**
- Processing Methodology
- Predictive Modelling
- Descriptive Modelling
- Conclusion
- Appendix

PROJECT OBJECTIVES

develop **Predictive Modelling**

Develop a clear methodology and process for modeling housing prices.

Test various models, linear regression, penalized models, and tree based models and boosting.

Highlight pros and cons.

produce **User friendly model**

Develop a V.1 user friendly model with under 15 features. Features should be easy to measure -- a ruler should be used minimally!

Model should be concise, with good accuracy.

User can be real estate agents or residents of AMES looking for a low lift way to find results.

explain **Descriptive Modelling**

Examine features , relevance, narratives, and storylines that explain SalesPrice behavior.

Crowd source hypotheses to test and leverage model to find answers.

enrich **User model**

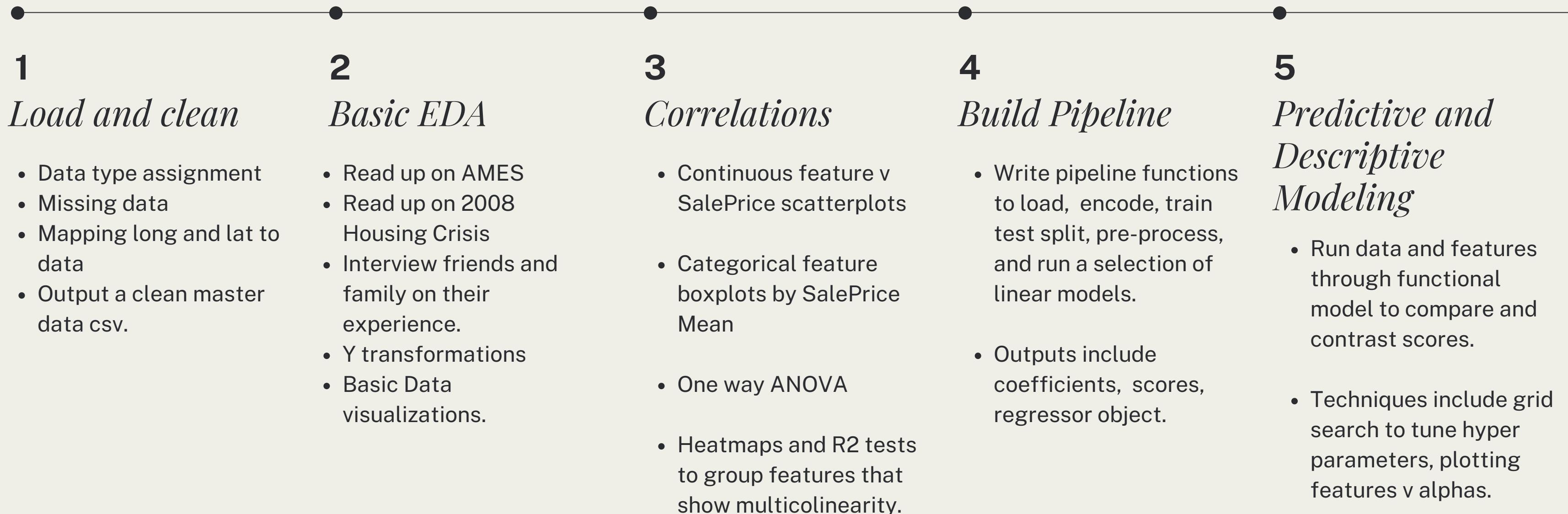
Apply learnings to V.1 model to extract more meaning and relevance from the model.

Ensure balance between Bias and Variance with a thoughtful selection of features.

A G E N D A

- Background
- Project Objectives
- **Processing Methodology**
- Predictive Modelling
- Descriptive Modelling
- Conclusion
- Appendix
 - Tree Based Models

PROCESSING METHODOLOGY



DATA CARD



2580 Records

80 Features

**Jan 2006
June 2010**

FEATURE DICTIONARY

Hosted on [Kaggle](#).

Cleaned Data Set

2559 records.

5 added Features.

FEATURES

Land: 'PID', 'MSZoning', 'LotFrontage', 'LotArea', 'Street', 'Alley', 'LotShape', 'LandContour',
'Utilities', 'LotConfig', 'LandSlope', 'Neighborhood', 'PavedDrive',

Type: 'MSSubClass', 'BldgType', 'HouseStyle', 'YearBuilt', 'YearRemodAdd', 'GarageYrBlt',

Finishes: 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'ExterCond',
'Foundation', 'BsmtExposure', 'BsmtFinType1', 'GarageFinish',

Quality: 'Condition1', 'Condition2', 'OverallQual', 'OverallCond', 'BsmtQual', 'BsmtCond',
'KitchenQual', 'FireplaceQu', 'Functional', 'GarageQual', 'GarageCond', 'ExterQual',

FEATURES, GROUPED

Size: 'GrLivArea', 'MasVnrArea', 'BsmtFinSF1', 'BsmtFinType2', 'BsmtFinSF2', 'BsmtUnfSF',
'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath',
'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'TotRmsAbvGrd', 'GarageArea', 'WoodDeckSF',
'OpenPorchSF', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea',

Amenities: 'Fireplaces', 'GarageType', 'GarageCars', 'PoolQC', 'Fence', 'MiscFeature', 'MiscVal',
'HeatingQC', 'CentralAir', 'Electrical', 'Heating',

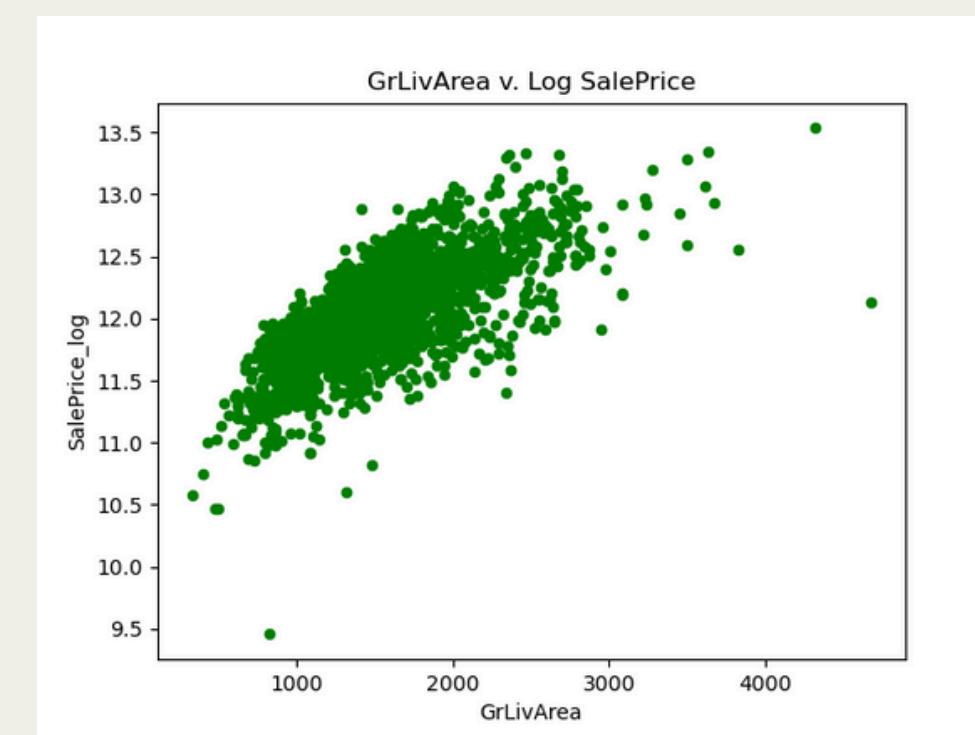
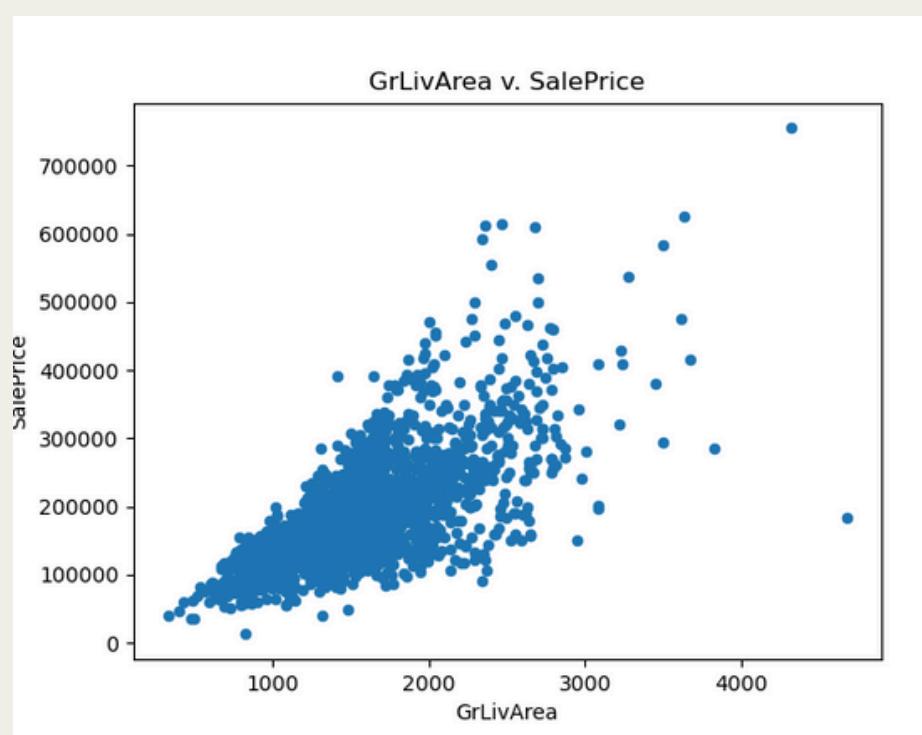
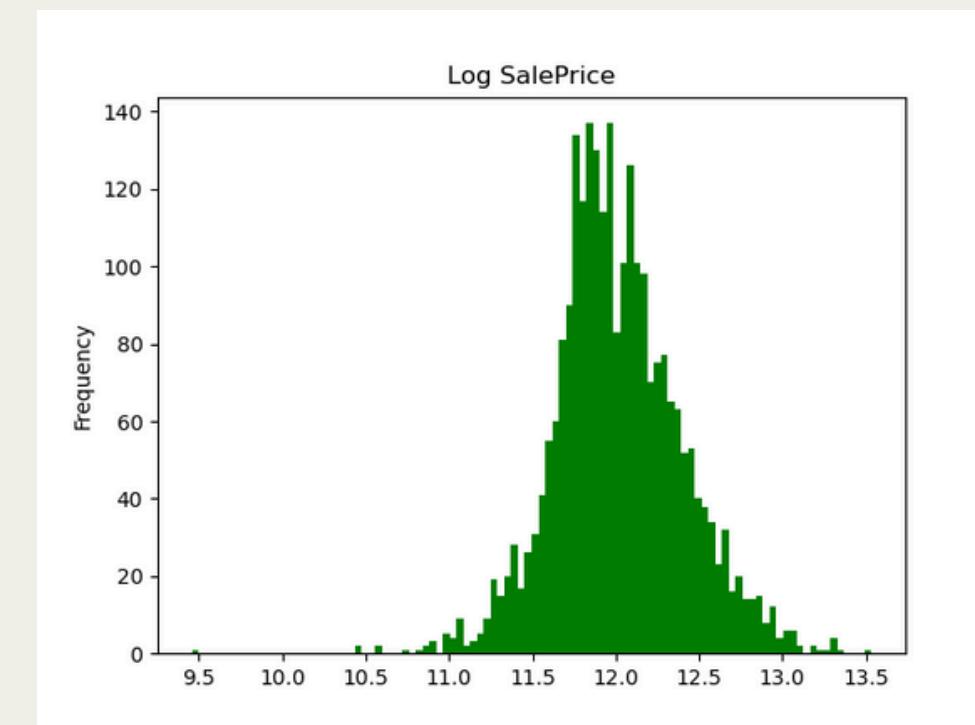
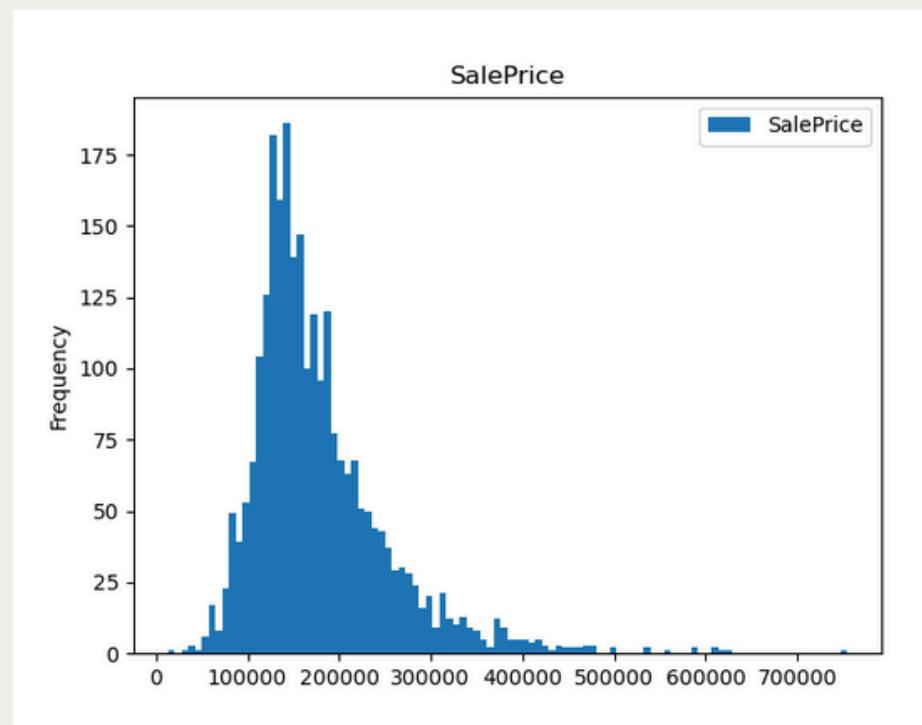
Sales Terms: 'MoSold', 'YrSold', 'SaleType', 'SaleCondition'

KEY DECISIONS DURING DATA PREP

Y Transformation

SalePrice distribution did not follow a Gaussian curve. Although this does not directly defy the assumptions of linear regression, normalizing can help the residuals meet variance assumption.

Scatterplots of continuous variables v SalePrice show a fanning behavior, which is not as linear. By transforming to log SalePrice, the scatterplot is more linear, with a more consistent variance.



KEY DECISIONS DURING DATA PREP

Features within groups need to be eliminated

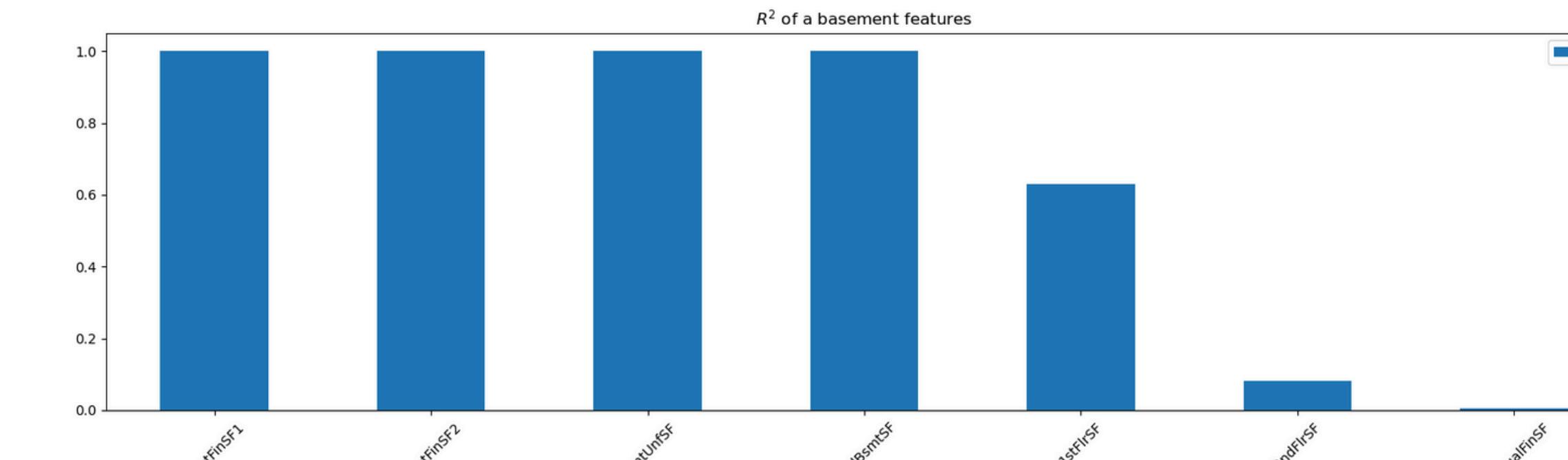
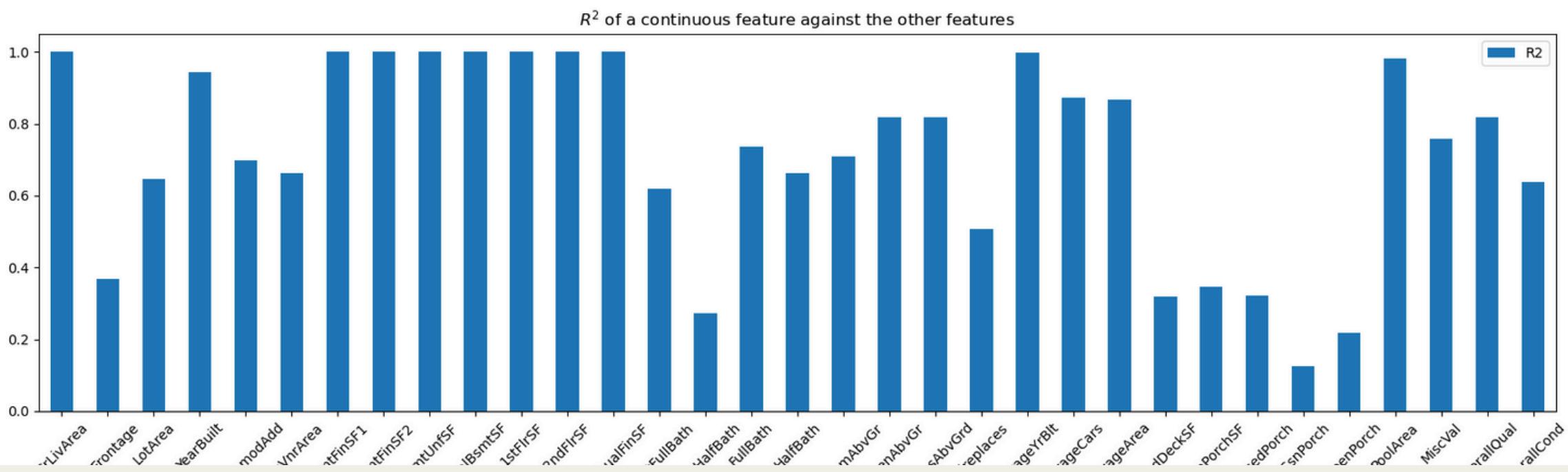
Plotting R Squared values of linear regression models where:

- Each feature is the dependent variable
- All other features are independents

R² = 0, exhibits no signs of multicollinearity.
R² = 1, is multicolinear with features.

The Basement features alone, show high R Squared values and should be consolidated to one.

No features were cut before running the model, however the end model needs to limit multicollinearity for stability.

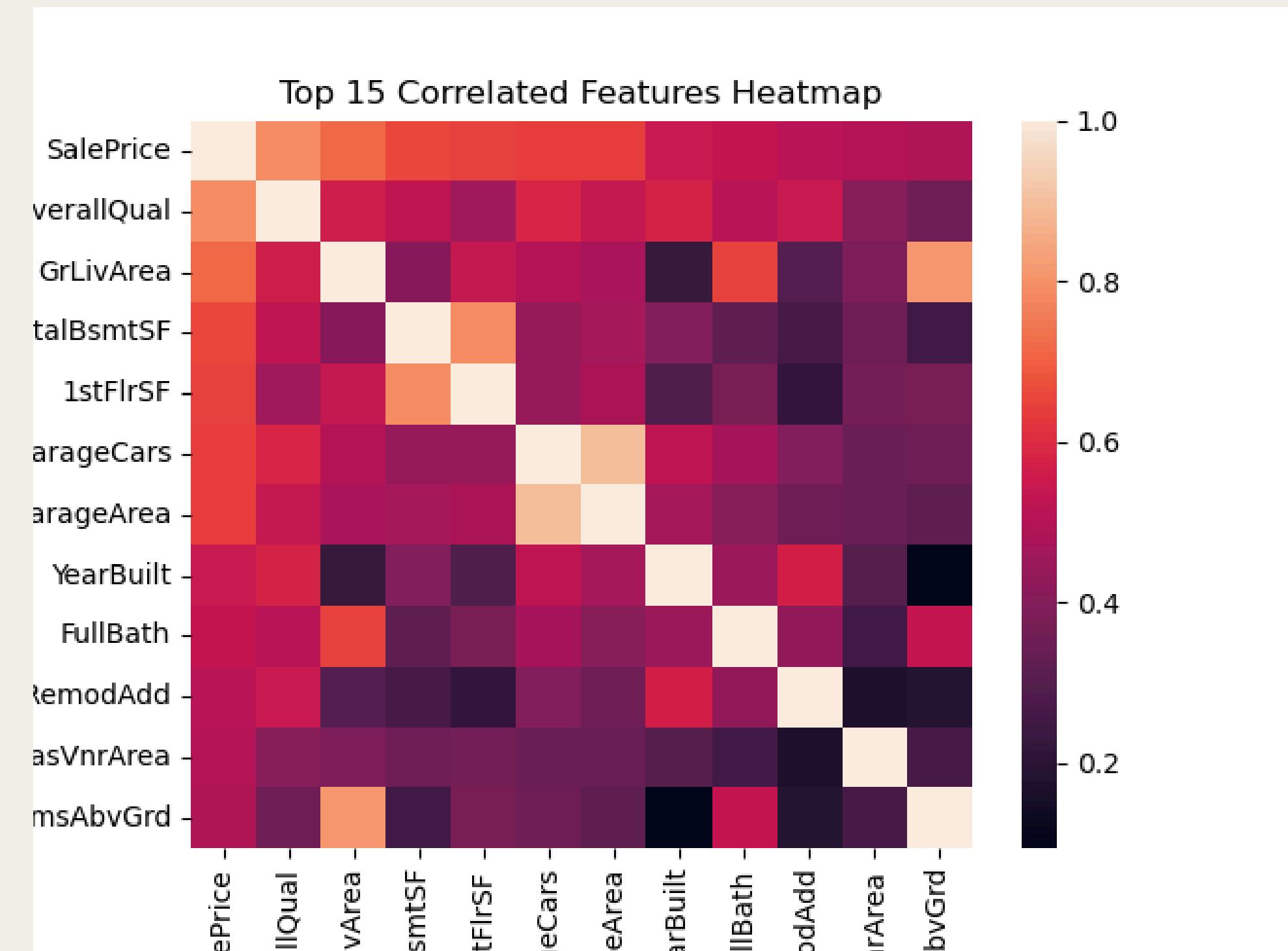


KEY DECISIONS DURING DATA PREP

Heatmaps should be used to plot direct feature -feature colinearity.

Ensure that final model does not have multiple features listed in this Heatmap.

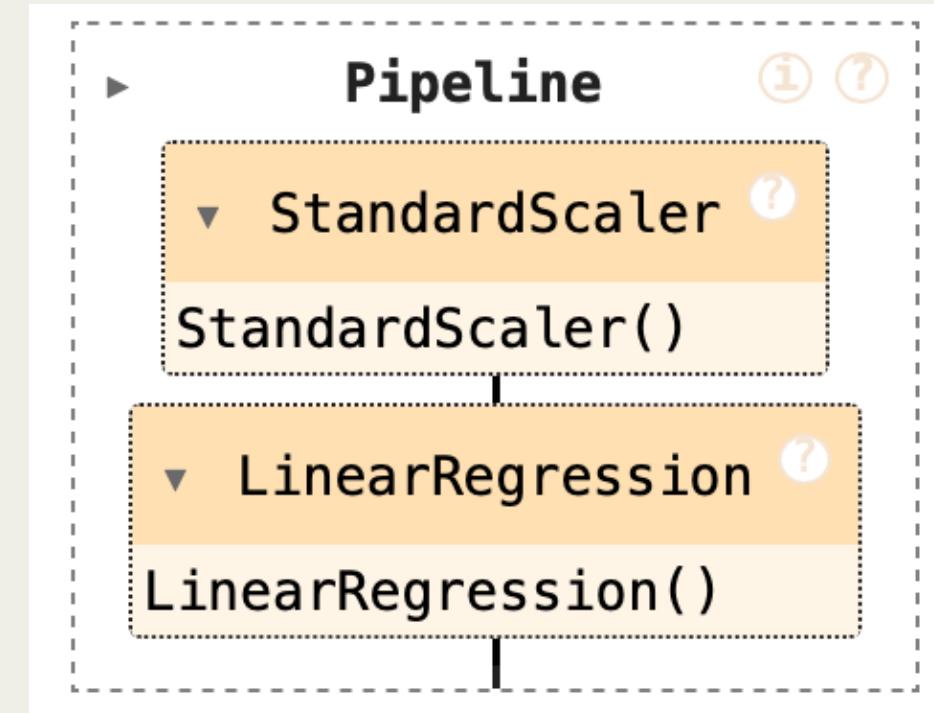
This heatmap lists the top 15 features that are most correlated with SalePrice (high likelihood they will be the most impact features in the model).



KEY DECISIONS DURING DATA PREP

Preprocessing Process:

1. One Hot Encoder to dummify all categorical features
2. Test SalesPrice and log of SalePrice as dependent variable for each model to determine best outcome.
3. Train Test Split (80/20)
4. Transform continuous features
 - a. Use Normalizer to transform when dealing with the saturated model. All features scaled from 0 to 1. Enable feature selection based on coefficient values.
 - b. Use StandardScaler on all features for models with fewer features.
5. Fit training data to regressor.
6. Use test data to produce and R squared and Adjusted R squared.
 - a. Use Adjusted R Squared to compare models.



LOG MODEL INTERPRETATION

Every Feature Produces is a Multiplier

1. By using the log, we replace Y with $\log(\text{SalePrice})$
2. We can re-write this equation to isolate SalePrice .
3. SalePrice is now the product of a set of multiples each ($e^{\beta_i \cdot x_i}$), where β_i is the coefficient returned by the model and x_i is the features.
 - a. $e^{\beta_0} = e^{\text{intercept}}$
 - i. Base Price for a House/Land
 - b. For a dummmified feature, if $x_i = 0$, $e^{0 \cdot \beta_i} = 1$.
 - i. Multiplier = 1, no change on outcome.
 - c. For a dummmified feature, if $x_i = 1$, $e^{\beta_i \cdot 1} =$
 - i. Multiplier = e^{β_i} .
 - d. For a continuous feature,
 - i. Multiplier = $e^{\beta_i \cdot x_i}$, this can range from [0,infinity].
 1. If <1 , reduces house price
 2. If >1 , increase house price.

$$Y = \beta_0 + \beta_1 \cdot \text{features}$$

$$Y = \log(\text{price})$$

$$\log(\text{price}) = \beta_0 + \beta_1 \cdot \text{features}$$

$$\text{price} = e^{\beta_0} \cdot e^{\beta_1 \cdot x_1} \cdot e^{\beta_2 \cdot x_2} \cdot \dots \cdot e^{\beta_n \cdot x_n}.$$

$$\text{price} = e^{\text{intercept}} \cdot e^{\beta_1 \cdot x_1} \cdot e^{\beta_2 \cdot x_2} \cdot \dots \cdot e^{\beta_n \cdot x_n}.$$

$$\text{price} = \prod_{i=0}^n e^{\beta_i \cdot x_i}$$

* $e \approx 2.72$.

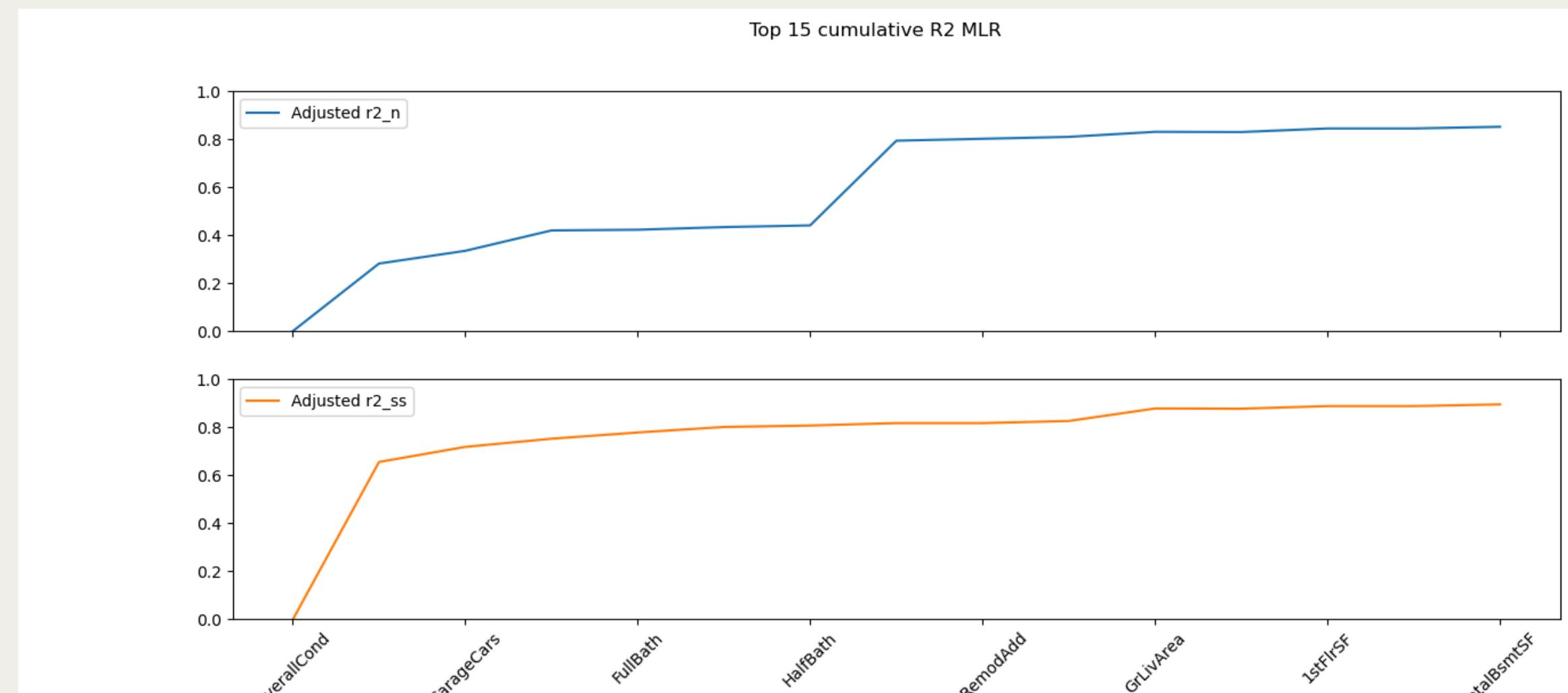
A G E N D A

- Background
- Project Objectives
- Processing Methodology
- **Predictive Modelling**
 - **Linear Models**
 - Tree Based Models
 - Easy Feature Model V1
- Descriptive Modelling
- Conclusion
- Appendix

PREDICTIVE MODELLING

Feature Selection Process:

1. Run saturated models.
2. Extract top 15 features.
3. Run through other models to compare Adjusted R2.
4. Find best top 15 and best model.



The graph above shows adjusted R squared values for models, each with increasing number of the Linear Regression Top 15 features.

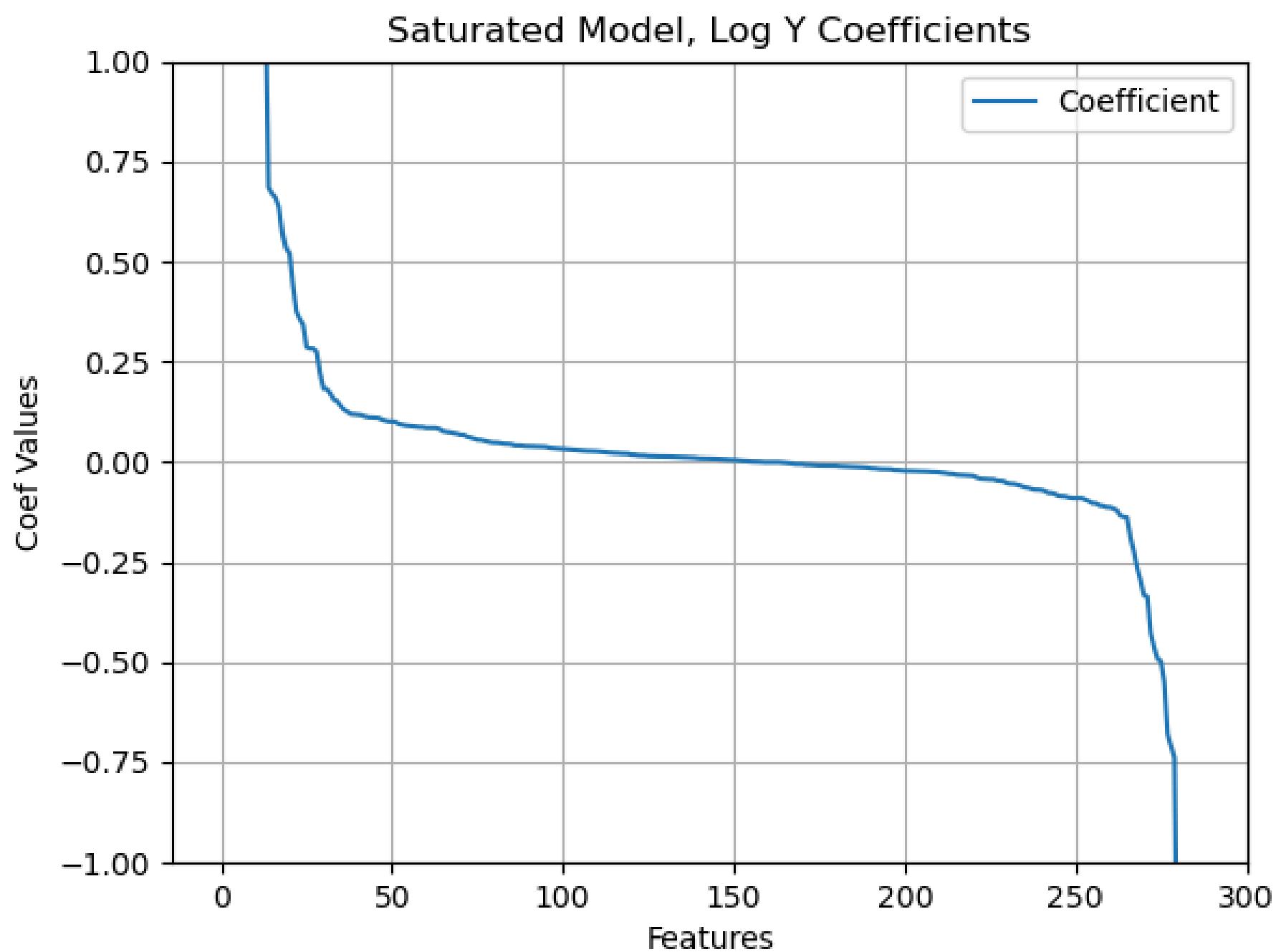
- With StandardScaler in preprocessing, the R squared goes to ~.9 quicker than Normalizer.
- Beyond 15 features, there is marginal improvement to scoring.

PREDICTIVE: COEFS AT A GLANCE

Coefficient Significance

- At $\text{coef} = .1$,
 - $e^{.1 \cdot x}$ ranges between [1.1,1] as a multiplier. The most it can shift the model is 10%.
- At $\text{coef} = 0$,
 - $e^{.1 \cdot 0} = 1$ and has no impact on the model
- At $\text{coef} = -1$,
 - $e^{-1 \cdot x}$ ranges between [.36,1] as a multiplier. The most it can shift the model is 10%.

To get an understanding of most impactful coefficients, we need to take the absolute value and sort. Take the top 15-20 and explore.



PREDICTIVE: LINEAR RESULTS (SATURATED)

Saturated Model Results

All models show similar behavior when compared testing to training. Does not display obvious signs of overfitting.

However, the number of parameters (almost 300 when taking the saturated model with dummmified features) is very large and the adjusted R Squared is much lower.

This is not suitable since:

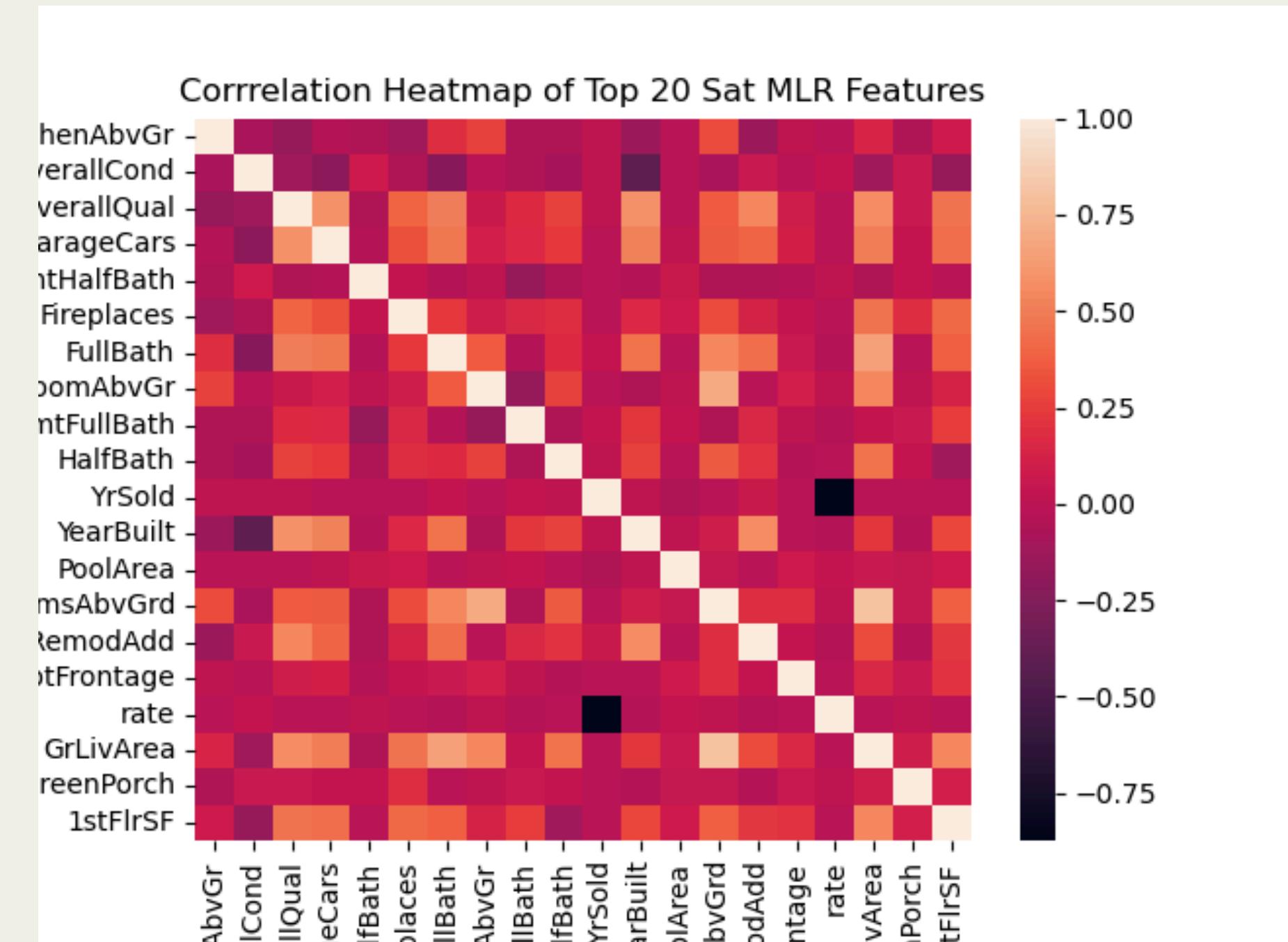
- All features would need to be collected to run data.
- Stability of model is compromised with having colinear features.

Model	Test R Squared	Test Adj R Squared	Train R Squared	MSE	RMSE
Linear Regression Normalized, Log Y	.910	.796	.934	.014	.118
Lasso Standard Scaler, Y (alpha = 4500)	.916	.809	.94	$5.33 * 10^8$	23.090
Ridge Standard Scaler, Log Y alpha = undetermined	.930	.841	.964	.011	.105
Elastic Net Standard Scaler Y alpha = .028, l1-ratio = .11	.901	.755	.912	$6.32 * 10^8$	25,156
LassoCV Standard Scaler, Log Y alpha = .0029	.925	.833	.934	.012	.11

PREDICTIVE: LINEAR TOP 15

Saturated Model Results

1. Plot absolute value of coef.
 - a. Extract top 20.
 2. Plot heatmap to indentify highly correlated features. Strip out one or all of features that show multicollinearity and add them back in through iterative loops.
 3. Find Feature set that optimizes either accuracy or accuracy and simplicity for adjusted r2.



PREDICTIVE: LINEAR TOP 15

Accuracy Based Feature Set

MLR 15 = OverallCond, OverallQual, GarageCars, Fireplaces, FullBath, BsmtFullBath, HalfBath, YearBuilt, YearRemodAdd, LotFrontage, GrLivArea, ScreenPorch, 1stFlrSF, 3SsnPorch, TotalBsmtSF

adjusted r2 = .896

Easy Measure Feature Set

Includes Features that are easy to measure (no measuring tape, except for total square footage).

MLR Easy Features = GarageCars ,Fireplaces, FullBath, BsmtFullBath, HalfBath, YearBuilt, OverallCond, OverallQual, GrLivArea, remodeled (bool, added feature)

adjusted r2 = .874

	r2	Adjusted r2	train r2	MSE	Abs Error	RMSE	0	Dropped Features
	0.897000	0.894000	0.885000	0.016000	skip	0.126000	30	[]
	0.897000	0.894000	0.885000	0.016000	skip	0.126000	4	['HalfBath']
	0.896000	0.893000	0.885000	0.016000	skip	0.126000	8	['1stFlrSF', 'HalfBath']
	0.896000	0.893000	0.885000	0.016000	skip	0.126000	0	['1stFlrSF']
	0.890000	0.887000	0.880000	0.017000	skip	0.130000	3	['Fireplaces']
	0.890000	0.887000	0.880000	0.017000	skip	0.130000	14	['Fireplaces', 'HalfBath']
	0.890000	0.887000	0.879000	0.017000	skip	0.130000	11	['TotalBsmtSF', 'HalfBath']
	0.890000	0.887000	0.879000	0.017000	skip	0.130000	1	['TotalBsmtSF']
	0.888000	0.886000	0.879000	0.017000	skip	0.130000	20	['1stFlrSF', 'Fireplaces', 'HalfBath']
	0.888000	0.885000	0.879000	0.017000	skip	0.130000	7	['1stFlrSF', 'Fireplaces']
	0.884000	0.881000	0.874000	0.018000	skip	0.134000	23	['TotalBsmtSF', 'Fireplaces', 'HalfBath']
	0.884000	0.881000	0.874000	0.018000	skip	0.134000	10	['TotalBsmtSF', 'Fireplaces']
	0.880000	0.877000	0.871000	0.019000	skip	0.138000	5	['1stFlrSF', 'TotalBsmtSF']

Table of linear model scores.

Top 15 Features = []. Features Dropped shows th model[] without specified feature.

Goal is to drop features and maintain accuracy.

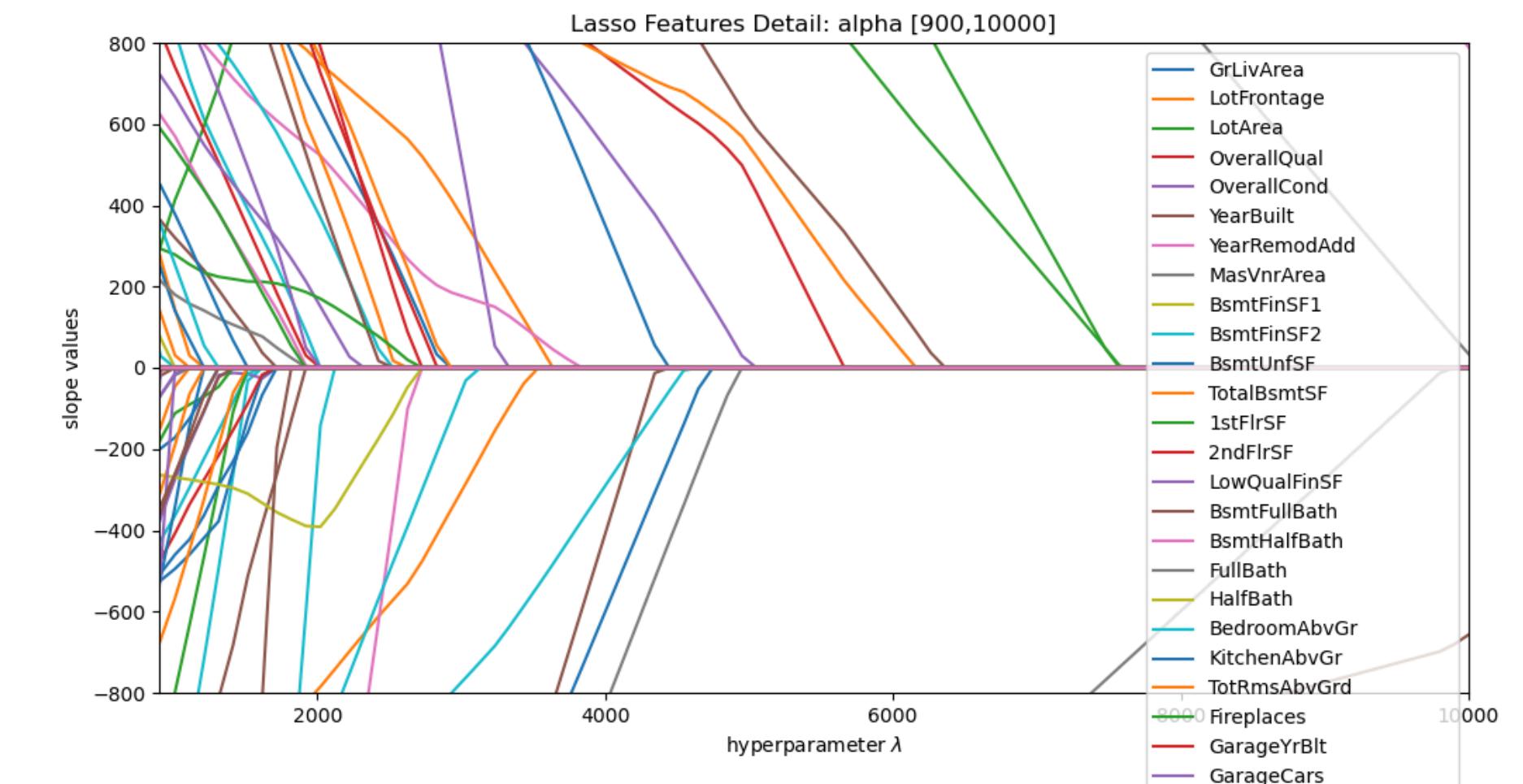
Top MLR
Features

PREDICTIVE: LASSO TOP 15

Feature Selection

1. Run a loop of alpha values for the saturated model.
2. Take all features that are not equal to zero at the max alpha.
3. Run through other models to find alpha where Adjusted R squared and train R Square converge.
 - a. MLR Top 15 in Lasso: 85.9%
 - b. Lasso Features: 87.1%
 - i. **Lasso 15 = OverallQual, GrLivArea, TotalBsmtSF, BsmtFinSF1, GarageArea, 1stFlrSF, YearBuilt, GarageCars, BsmtExposure_Gd, YearRemodAdd, MasVnrArea, ExterQual_TA**

Top Lasso Features



As alpha (or lambda here) goes to 10000, features go to zero. At 10,000, there are only a select few features that are non-zero. These are the top features. [Alpha tuning visualization](#).

PREDICTIVE: LASSOCV TOP 15

Feature Selection

1. LassoCV provides cross validation and auto alpha tuning.
2. LassoCV Top includes features beyond just the physical aspects of the house.
 - a. Lasso CV Top 15 = GrLivArea*, OverallQual*, YearBuilt*, OverallCond*, TotalBsmtSF*, BsmtFinSF1**, GarageCars*, Neighborhood_Somerst, Neighborhood_Crawfor, YearRemodAdd**, Fireplaces*, LotArea, SaleCondition_Normal, Neighborhood_NridgHt, MSZoning_RL

	r2	Adjusted r2	train r2	MSE	Abs Error	RMSE
Lassocv Top 15	0.910000	0.907000	0.900000	0.014000	skip	0.118000
MLR 15	0.897000	0.894000	0.885000	0.016000	skip	0.126000
Lasso Top 15 coef	0.881000	0.878000	0.871000	0.019000	skip	0.138000
MLR Easy Features	0.876000	0.874000	0.867000	0.019000	skip	0.138000

LassoCV best score is usign its own returned features. It is worth exploring if these features - neighborhood and zoning - can be explored in a more orgnaized fashion in descriptive analysis.

*seen in MLR and Lasso Top Features

**seen Lasso Top Features

A G E N D A

- Background
- Project Objectives
- Processing Methodology
- **Predictive Modelling**
 - Linear Models
 - **Tree Based Models**
 - Easy Feature Model V1
- Descriptive Modelling
- Conclusion
- Appendix

PREDICTIVE: TREE FAMILY RESULTS

Saturated Model Results

1. Run saturated models with default settings.
2. Run saturated models with some tuning techniques:
 - a. Grid Search
 - b. Custom tuning function.
3. Record best r2.

SalePrice used instead of log SalePrice.

With tuning, smaller difference between test and train r squared.

	train r2	test r2	hyperparameters
Decision Tree <i>Default Params</i>	.666	.666	criterion = mse min_samples_split= 325
Decision Tree <i>through Grid Search</i>	.892	.813	criterion = abs error min_samples_leaf = 8
Random Forest <i>Default params</i>	.981	.876	criterion = mse min_samples_split = 2 ooh score = .86
Random Forest <i>through Grid Search</i>	.907	.885	criterion = friedman_mse min_samples_leaf = 2 n_estimators = 100
Boosting	.981	.933	learning rate = .001 max_depth = 4

PREDICTIVE: TREE SCORES FOR FEATURE SETS

	MLR train r2	MLR test r2	Lasso train r2	Lasso test r2	hyperparameters
Decision Tree	.678 <u>tree in appendix</u>	.678	.661 <u>tree in appendix</u>	.669	criterion = mse, min_samples_split= 325
Decision Tree Grid Search	skipped	skipped	skipped	skipped	skipped
Random Forest	.984	.904	.982	.865	criterion = mse, min_samples_split = 2, ooh score = .86
Random Forest Grid Search	.907	.885	.965	.906	criterion = friedman_mse min_samples_leaf = 2 n_estimators = 100
Boosting w tuning	.963	.910	.939	.86	learning rate = .001, max_depth = 4

MLR Features = OverallCond, OverallQual, GarageCars, Fireplaces, FullBath, BsmtFullBath, HalfBath, YearBuilt, YearRemodAdd, LotFrontage, GrLivArea, ScreenPorch, 1stFlrSF, 3SsnPorch, TotalBsmtSF

Lasso Features = OverallQual, GrLivArea, TotalBsmtSF, BsmtFinSF1, GarageArea, 1stFlrSF, YearBuilt, GarageCars, BsmtExposure_Gd, YearRemodAdd, MasVnrArea, ExterQual_TA

EASY FEATURE MODEL V1

Objective: Build a simple model with few inputs and high impact. Users are real estate agents, and buyers and sellers that want a quick answer on house prices.

From MLR, we have developed the Easy Features feature set:

Easy Features = GarageCars ,Fireplaces, FullBath, BsmtFullBath, YearBuilt, OverallCond, OverallQual,
GrLivArea, Remodeled (bool, added feature)

Scores:

- Linear Regression, we get an adjusted $r^2 = .871$
- LassoCV, we get an adjusted $r^2 = .871$
- Random Forest an adjusted $r^2 = .866$
- When run with Boosting and tuning we get an adjusted $r^2 = .892$

In the next section we will perform some descriptive analysis. From our descriptive analysis, we will revisit this model and see how we can trade in features for better results.

A G E N D A

- Background
- Project Objectives
- Processing Methodology
- Predictive Modelling
- **Descriptive Modelling**
 - **Assumptions and Insights**
 - Easy Feature Model V2
- Conclusion
- Appendix

APPROACH

This half of the study focuses on the macro features impacting SalePrice -- those features beyond the four walls of the house.

To kickstart this inquiry, I asked people in San Francisco, what they thought drove pricing in the housing market. I list the following assumptions and analyses to strengthen our understanding of the market and hopefully the model.

ASSUMPTION #1

*“Location, location,
location!”*

NEIGHBORHOODS

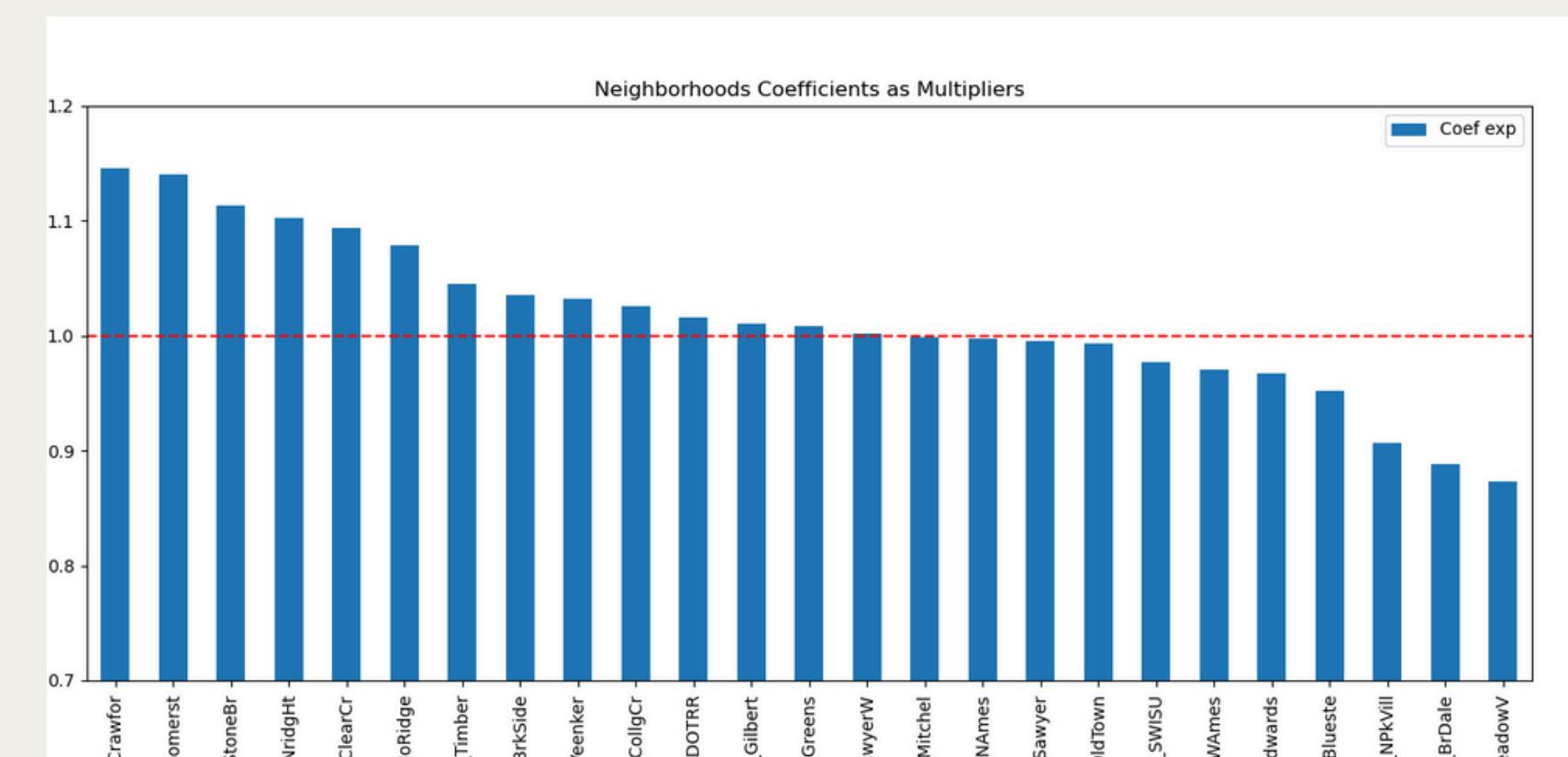
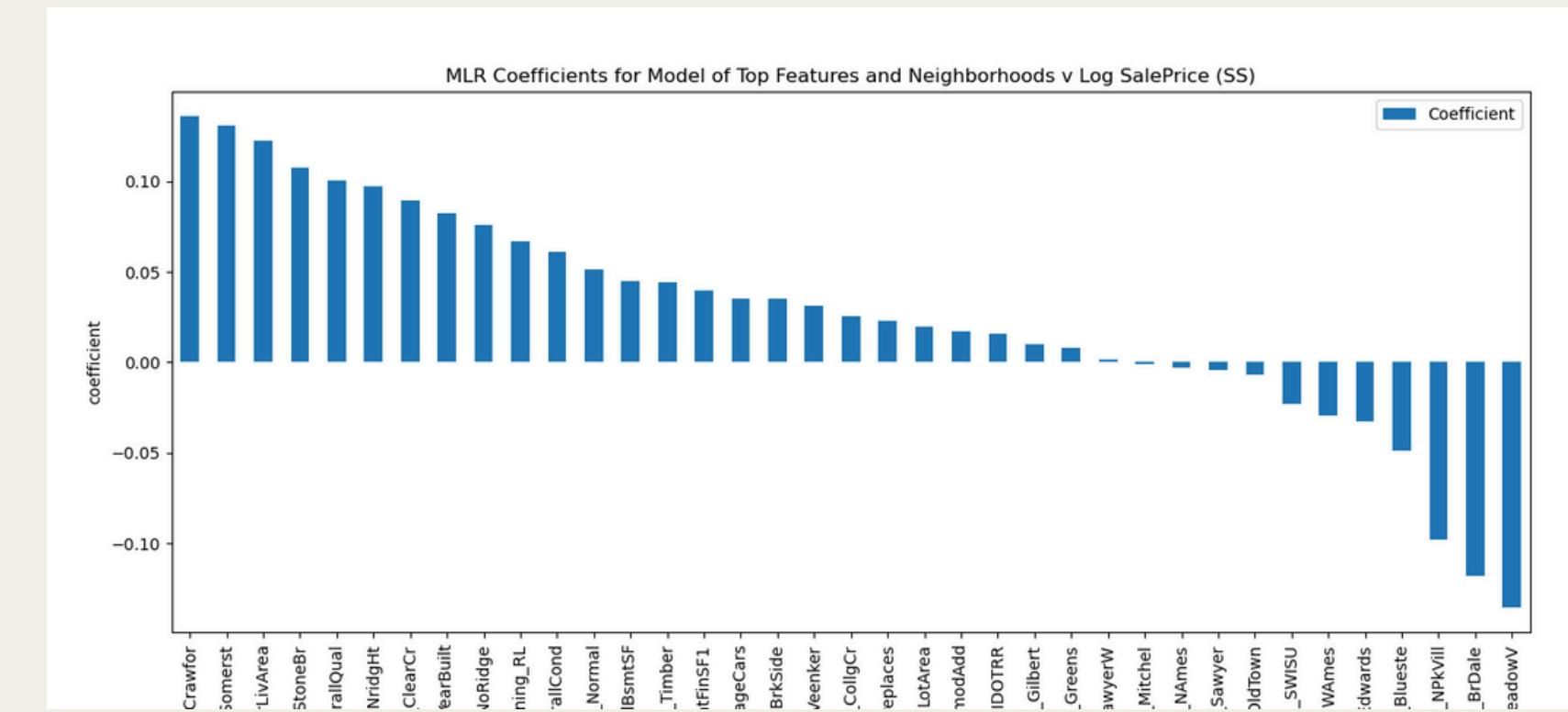
Using MLR to assign values to Neighborhoods

1. Use MLR with LassoCV Features + dummified Neighborhoods.
2. Plot to visualize impact of neighborhoods on model.
3. Sort and divide by quartile.
4. Assign a new feature: Neighborhood quartile (aka New_quant): Low, LowMed, MedHigh, High.

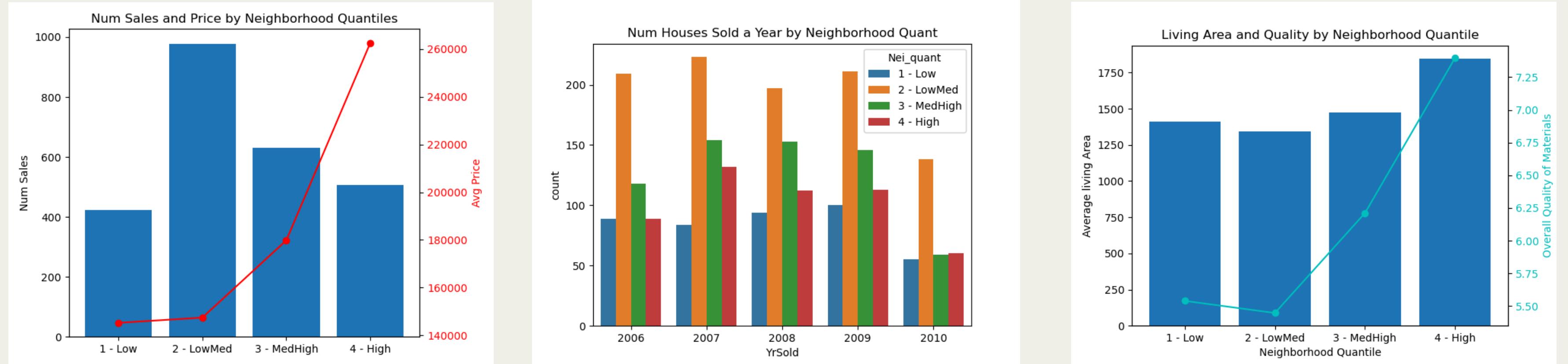
Y axis aka 'Multipliers' on second graph = $e^{\beta \cdot x}$.
 Remember,

$$price = e^{inter} \cdot e^{\beta_1 * x_1} \cdot e^{\beta_2 * x_2} \cdot e^{\beta_n * x_n}.$$

Each neighborhood either increases or decrease overall value by a percentage. Crawford increases SalePrice by 1.14 aka 14%.



NEIGHBORHOODS



Observations

By assigning Neighborhood Quantiles, we can visualize these neighborhoods and how they look:

- High Neighborhoods are typified by larger square footage, and more expensive houses. On average, their SalePrice is 180% of Low Neighborhoods. They are larger and are made of higher quality materials.
- MedHigh Neighborhoods have second highest number of house sales. MedHigh is substantially lower in SalePrice compared to High. It has the second highest number of sales per year.
- LowMed Neighborhoods have the highest sales numbers per year. These neighborhoods show the most activity and turnover.
- Low Neighborhoods have fewer sales than all other Neighborhoods. They have the lowest avg SalePrice.

ASSUMPTION #2

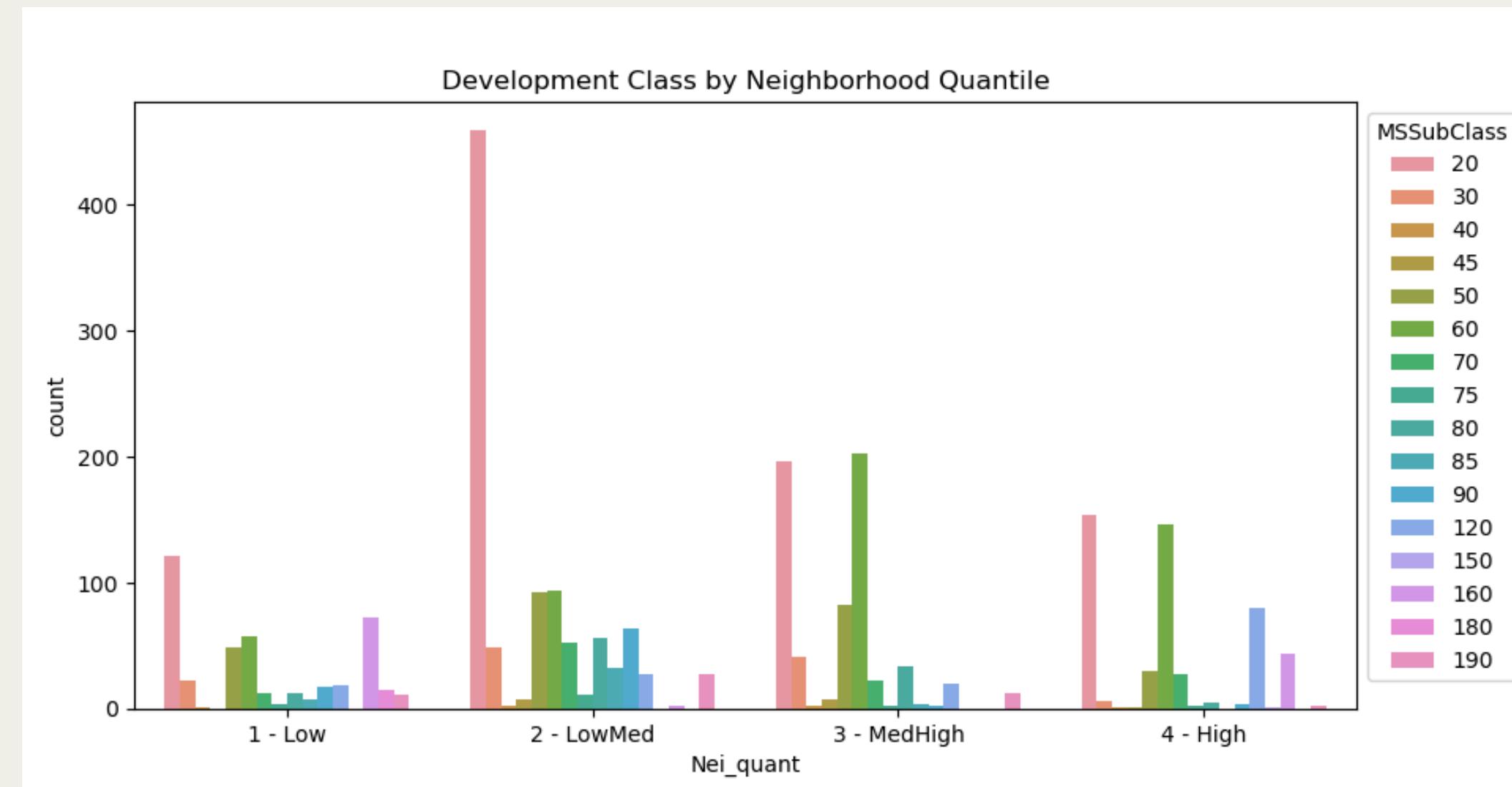
“How old is the city?”

People like old historic houses.”

DWELLING TYPE

MSSubClass: Type of dwelling

- 20* 1-STORY 1946 & NEWER ALL STYLES
- 30* 1-STORY 1945 & OLDER
- 40 1-STORY W/FINISHED ATTIC ALL AGES
- 45 1-1/2 STORY - UNFINISHED ALL AGES
- 50* 1-1/2 STORY FINISHED ALL AGES
- 60* 2-STORY 1946 & NEWER
- 70 2-STORY 1945 & OLDER
- 75 2-1/2 STORY ALL AGES
- 80 SPLIT OR MULTI-LEVEL
- 85 SPLIT FOYER
- 90 DUPLEX - ALL STYLES AND AGES
- 120* 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
- 150 1-1/2 STORY PUD - ALL AGES
- 160* 2-STORY PUD - 1946 & NEWER
- 180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
- 190 2 FAMILY CONVERSION - ALL STYLES AND AGES



Single Houses build after 1946 are most commonly sold. In fact, there are very few houses built before 1946 in this data.

MedHigh and High neighborhoods have more two story houses built after 1946.

High Neighborhood also show higher relative sales in PUD's, which we can think of as track homes/ new housing developments. With the majority of houses built after 1946, is being remodeled, relevant?

ASSUMPTION #3

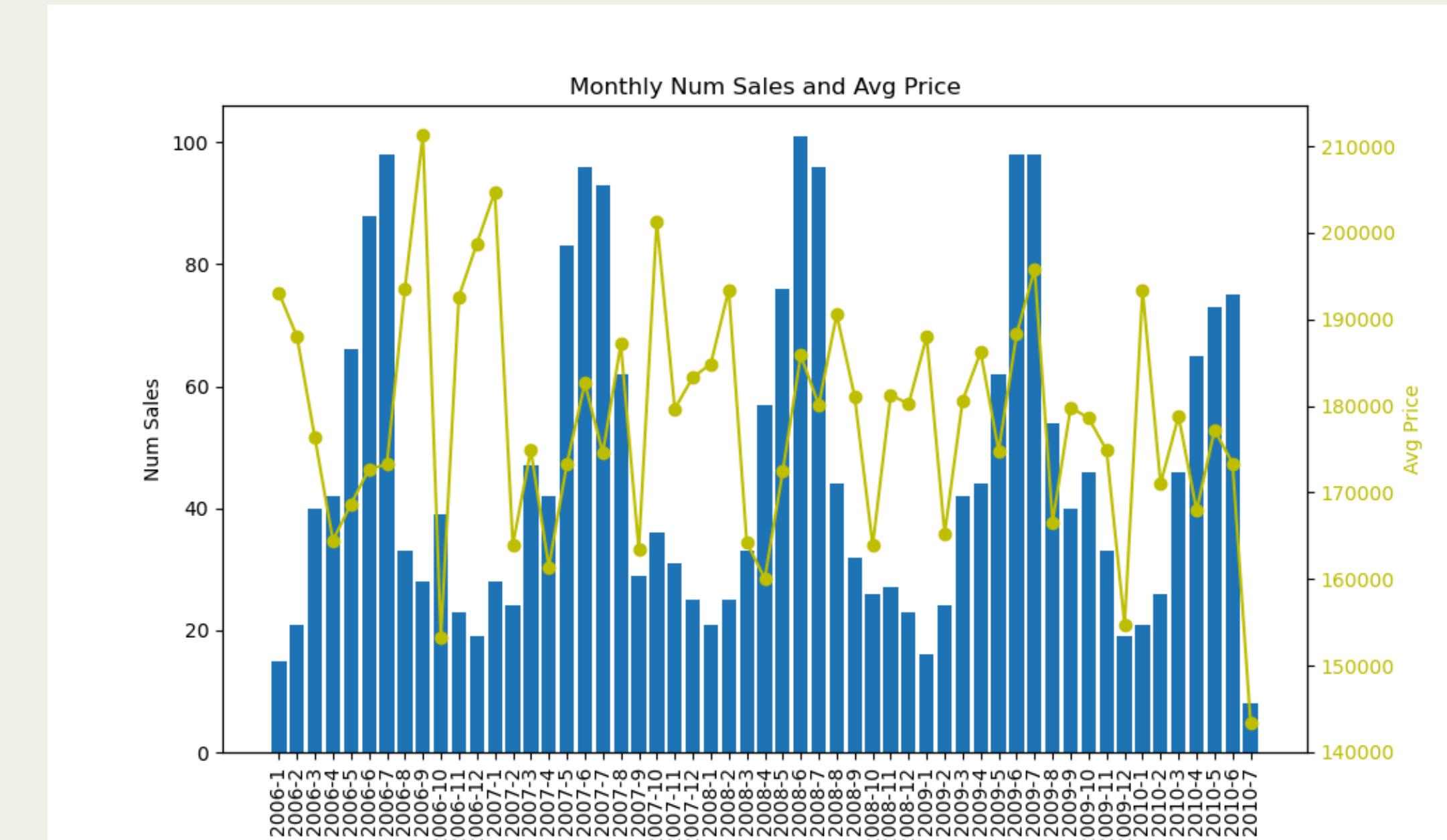
“It's all about Supply and Demand.

How many houses are for sale?”

MONTHLY NUM SALES V AVG PRICE

Supply and Demand and seasonality.

- Number of sales appear steady over the years. However, sales per month following a wave shape.
 - Sales per month peak in June with around 100 sales per month.
 - Second bump in September
 - Moving in the summertime or between school is a convenient time for life transitions.
- House prices are highest when sales are lowest, especially in 2006, 2007, 2008.
- In end of 2008, this pattern flips, with 2009 showing high avg price and high number of sales.
 - This may be a result of the 2008 housing crisis.



ASSUMPTION #4

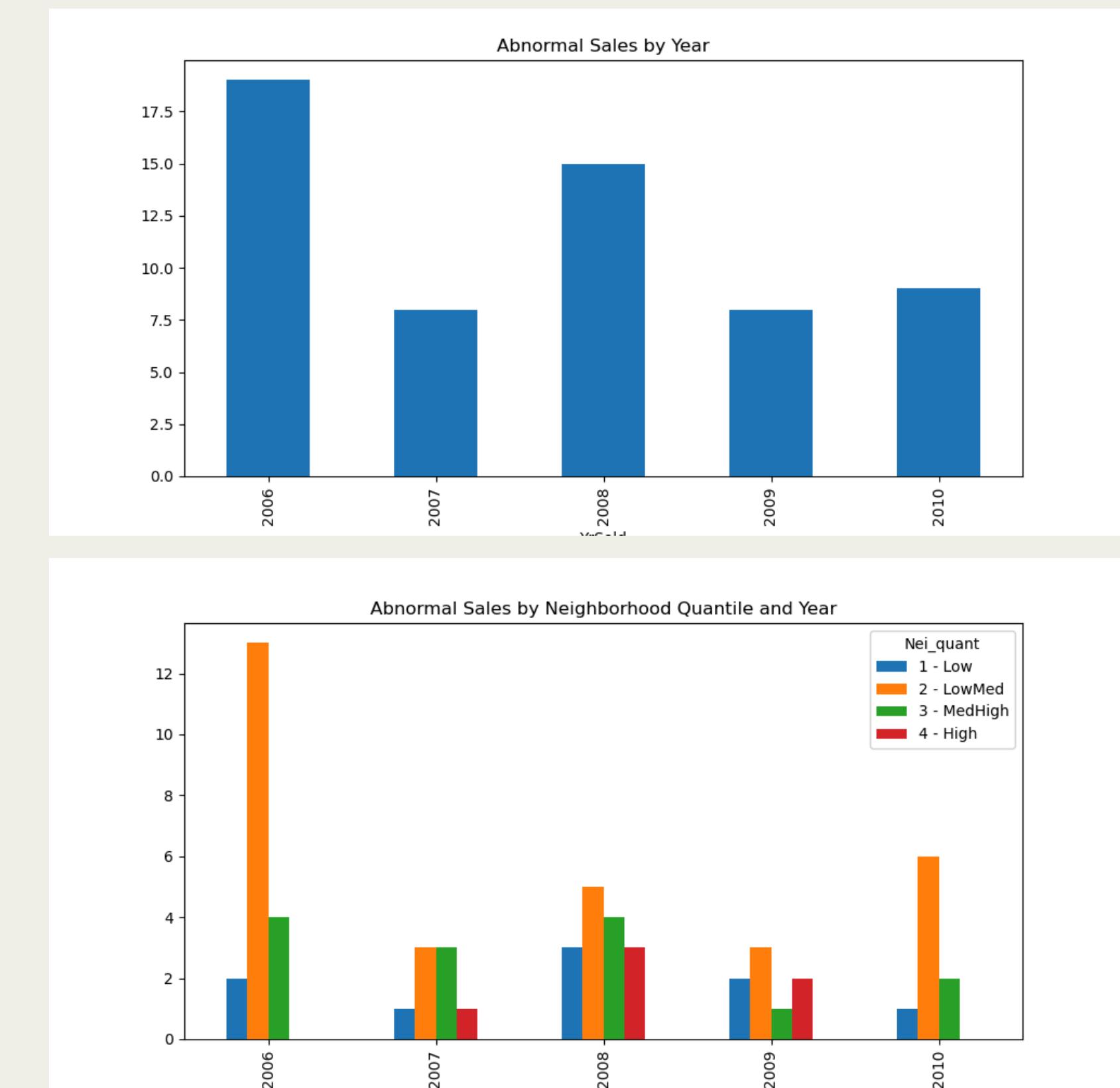
*“The market tanked in 2008,
not sure but it can’t be good.”*

ABNORMAL SALES BY YEAR

Features awareness

- Abnormal sales - trade, foreclosure, short sale - make up around 1-2% of all sales in the dataset.
- Number of Abnormal Sales are highest in 2006 and 2008, and most common in LowMed neighborhoods.
 - In 2006, there were 13 abnormal sales in the lowMed Neighborhoods.
 - In the years leading up to the housing crisis, banks were approving subprime mortgages to less qualified people resulting in higher volume of house sales, higher demand, and at times, foreclosure (when people were not able to afford their monthly premiums).
 - One would expect foreclosures to increase in 2008 and after, but rates stayed low. This may be a result of Congress encouraging banks to restructure mortgages instead of foreclose, increasing mortgage amount the Federal Housing Administration would insure. This was done in an effort to stabilize the housing market, and prevent housing prices from falling.
 - In fact, housing sales and prices remain constant throughout 2008-2010.

Source: federalreservehistory.org



ASSUMPTION #5

*“Mortgage rates are
down, you better
hurry up and buy!”*

KEY DECISIONS DURING DATA PREP

*Added
Feature*

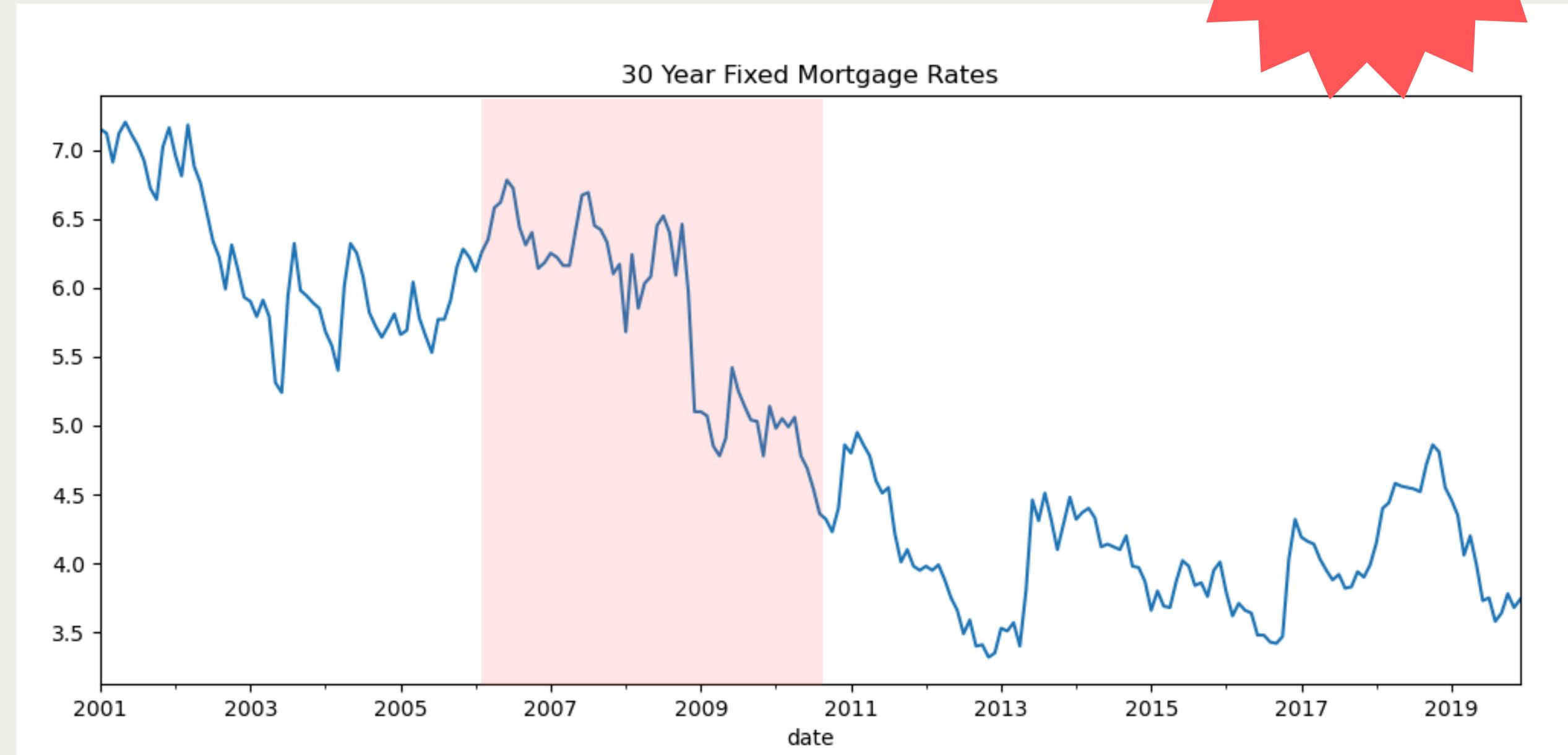
Added Feature

In order to investigate the relationship between mortgage rates and number of sales, a monthly average of the 30 year Fixed Mortgage Rate was added to the data set.

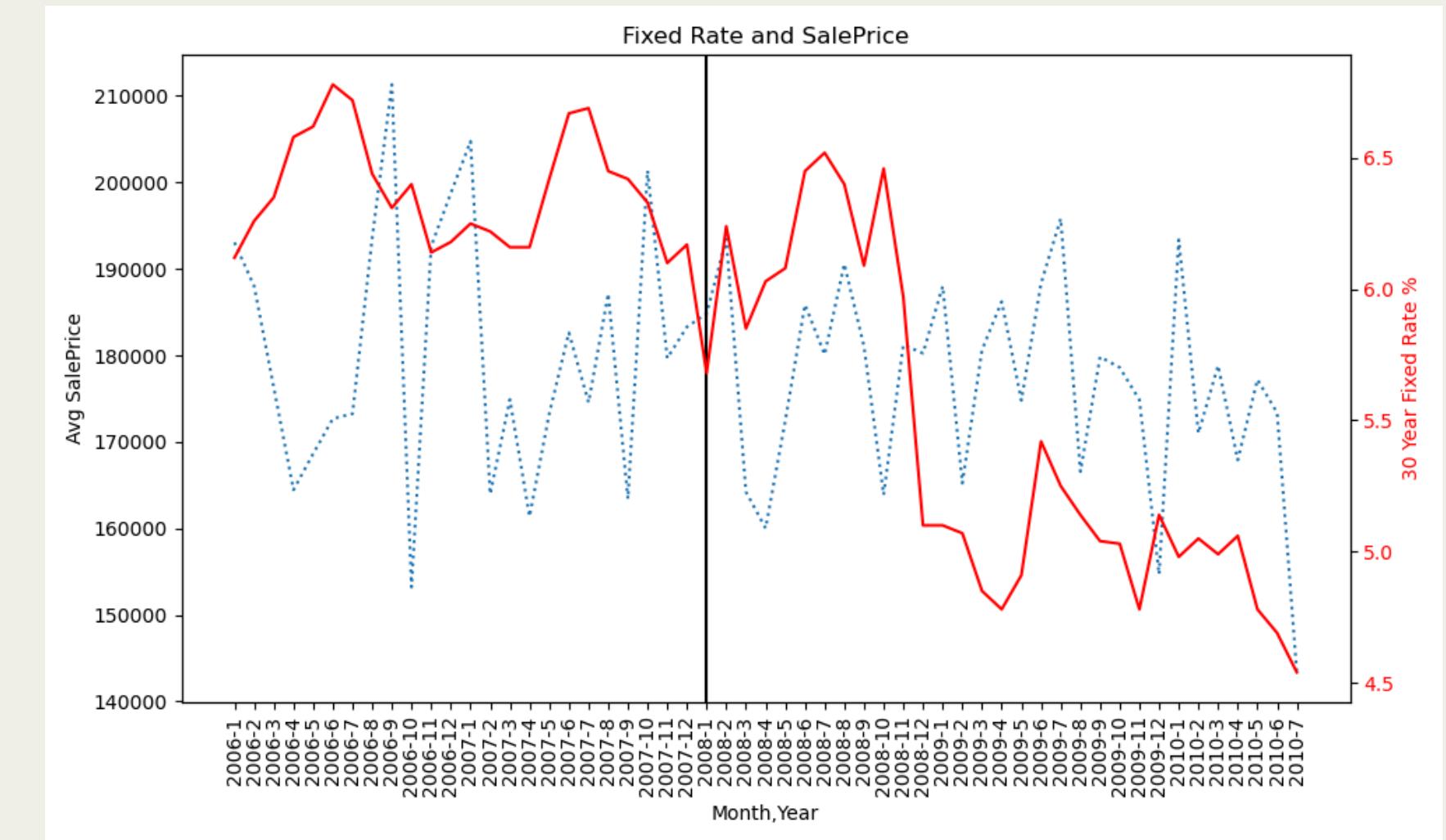
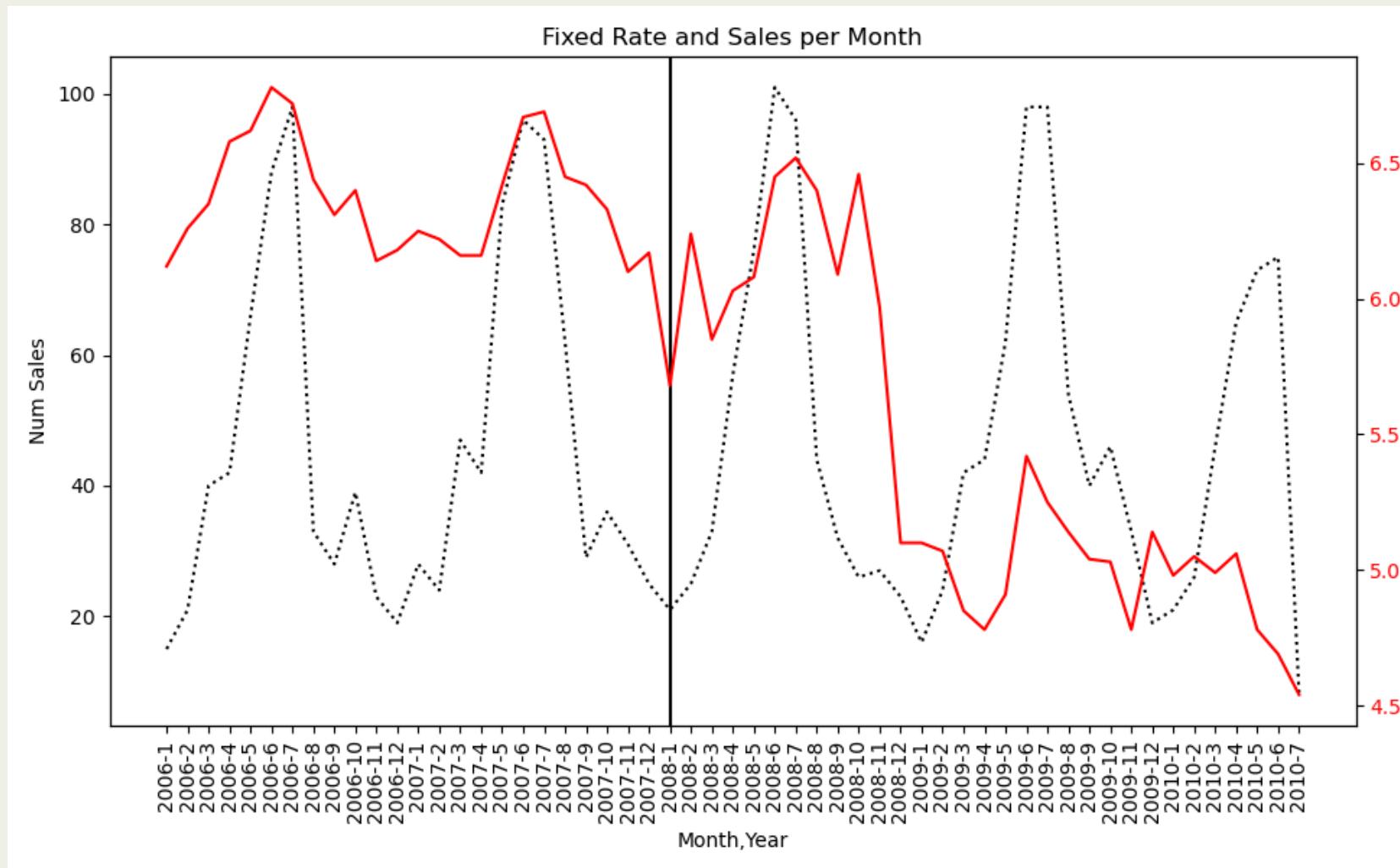
2006-2008 follow a similar pattern and rate, with peaks in the middle of the year.

In 2009, the Mortgage Rate drops by a point plus. This is because the Federal Reserve lowered rates to stimulate spending and stabilize the market to keep prices from falling.

At the beginning of the millennium, rates are at 7% and hit a low during 2013 at < 3.5%.



PREDICTIVE: PLOTTING ALPHA FOR LASSO



- Num Sales and Fixed Rate seem to follow the same shape and cadence with peaks in the middle of the year, especially up to 2008. In the 2009, Fixed Rate falls, but Num Sales does not and the peaks are less in line.
- Fixed Rate and SalePrice seem to have inverse behavior in 2006 and 2007. When mortgage rates are higher, one has less to spend on the premium, and house prices are lower, which is what the data up until 2008 reflects.
- In 2008, 2009, and 2010, this pattern is less consistent. Rates dropping muddies the trend. Nevertheless, lower rates coincide with higher SalePrice.

RATE, SALES, AND PRICE CORRELATIONS

Correlation Matrix by Year

	2006			2007			2008			2009			2010		
	Sales Month	Avg Price	Fixed Rate	Sales Month	Avg Price	Fixed Rate	Sales Month	Avg Price	Fixed Rate	Sales Month	Avg Price	Fixed Rate	Sales Month	Avg Price	Fixed Rate
Sales Month	1.00	-0.46	0.93	1.00	-0.06	0.77	1.00	0.03	0.58	1.00	0.54	0.56	1.00	0.71	0.18
Avg Price	-0.46	1.00	-0.51	-0.06	1.00	-0.05	0.03	1.00	0.12	0.54	1.00	0.05	0.71	1.00	0.59
Fixed Rate	0.93	-0.51	1.00	0.77	-0.05	1.00	0.58	0.12	1.00	0.56	0.05	1.00	0.18	0.59	1.00

A correlation matrix of Sales per Month, Avg Price, and the Mixed Rate show how correlation shows by year.

- In 2006, Sales per Month and Fixed Rate are highly correlated at .93. This value decreases by year and remains over .5 through 2009. In 2010, this value falls to .18
- In 2006, Fixed Rate shows some correlation to Avg Price, but this drops off in the subsequent years.

In non-recession and non-crisis years, trends between Sales per Month, Avg Price, and Fixed Rate are more correlated. However government intervention alters this relationship.

To see how YrSold, MoSold, and Rate compared in a model. The MLR Easy Features were plotted three times, each with one of the mentioned features. The results show that MoSold brought the most information, but also the most parameters. All features detracted from the accuracy of the model equally. Given all the outside forces on the market, these features did not have an impact on the model.

SYNTHESIS

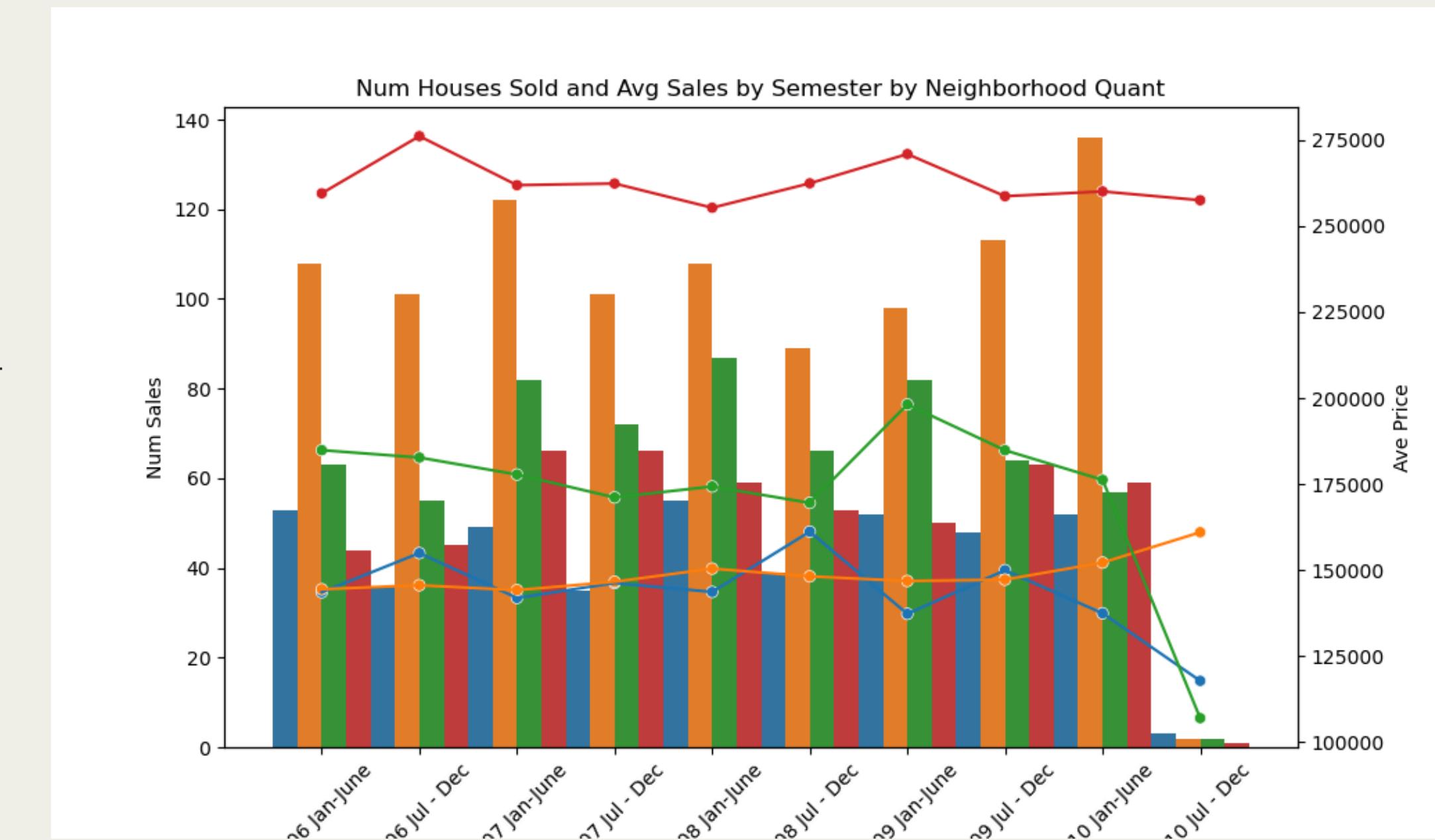
“So, what happened?”

LOW MED SALES TREND UPWARD

Prices remain stable

If we revisit the Number of Sales and Price chart , but now by half year (to include the first half og 2010), we can see the following:

- Low, MedHigh, and High Sales remain relatively constant, with a dip in number of sales in Jan-Juen 2010. Low and High Sales follow a Sales per Month and SalePrice inverse relationship.
- LowMed Sales show a different sales pattern. In July-Dec 2009, number of sales is expected to be less than its Jan-June 2009 numbers. However, it increases, and continues to increase through Jan-June of 2010.
 - Throughout this period, price remains constant.
 - For those selling, selling may be motivated by an inability to pay the mortgage or a desire to recoup investment.
 - For those buying, mortgages rates had just been lowered by the fed to an all time low. Although prices did not change, the overall amount owned monthly does which considering a lower mortgage rate.



A G E N D A

- Background
- Project Objectives
- Processing Methodology
- Predictive Modelling
- Descriptive Modelling
 - **Easy Feature Model V2**
- Conclusion
- Appendix

EASY FEATURE FINAL MODEL

Objective: Build a simple model with few inputs and high impact. Users are real estate agents, and buyers and sellers that want a quick answer on house prices.

1. Recall Easy Features = GarageCars ,Fireplaces, FullBath, BsmtFullBath, YearBuilt, OverallCond, OverallQual, GrLivArea, Remodeled (bool, added feature) from initial Linear Regression Exploration.
2. Iterate through, adding and subtracting features to find optimal balance between adjusted R squared and train R squared.
 - a. To add in: 'MSSubClass', 'Nei_quant', 'rate' (fixed rate mortgage), 'YrSold', 'MoSold'
 - b. To subtract: Full Bath, 'remodeled', 'YearBuilt'.
3. Run through LassoCV, Random Forest.

Final Features: GarageCars, Fireplaces, BsmtFullBath, YearBuilt, OverallCond, OverallQual, GrLivArea, MSSubClass, Nei_quant

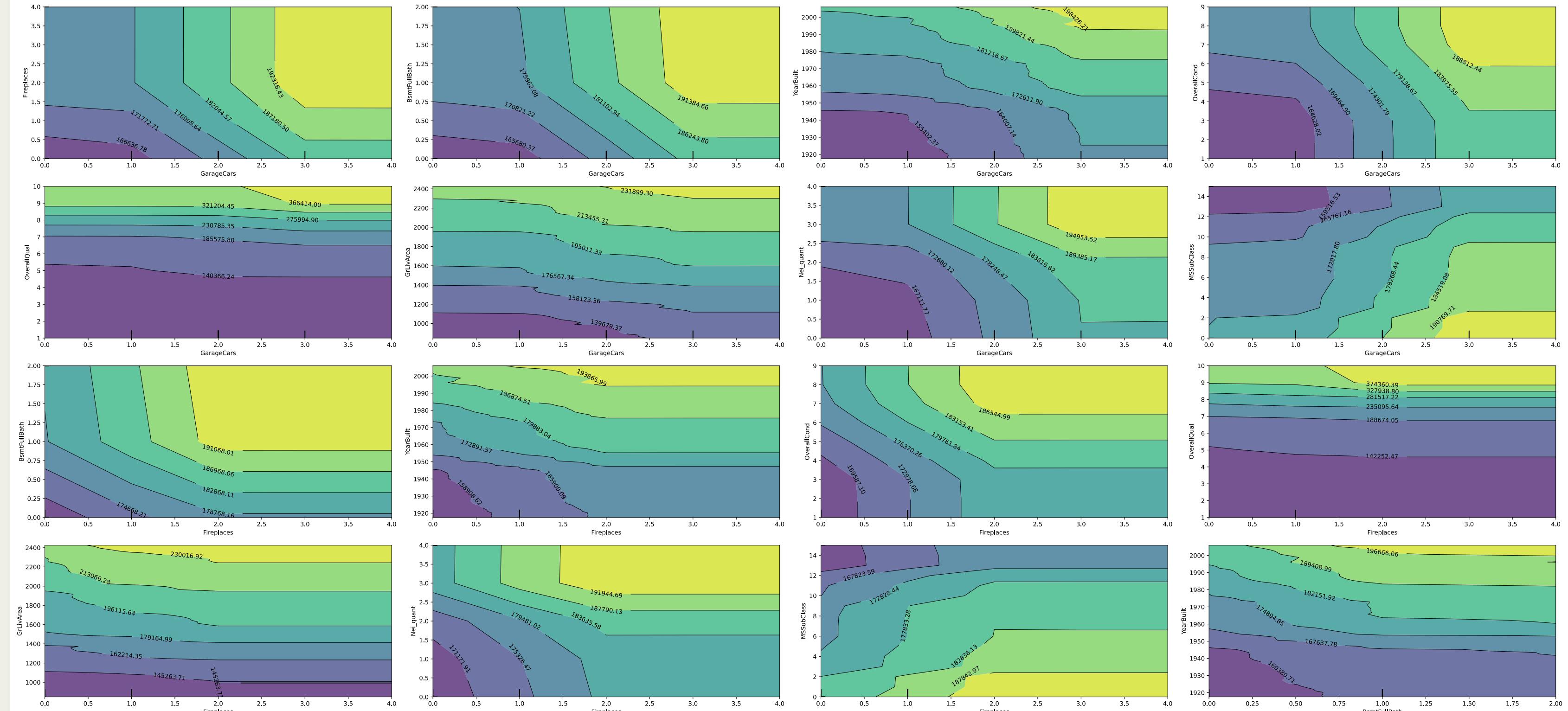
Model	Adjusted R2	Train R2	Hyperparameters
MLR (log saleprice, SS)	.902	.897	(log saleprice, SS)
LassoCV	.901	.897	(log saleprice, SS, cv = 5, alpha = 0.00035)
Random Forest w Grid Search tree visual	.899	.966	{'criterion': 'squared_error', 'min_samples_leaf': 2, 'n_estimators': 100}

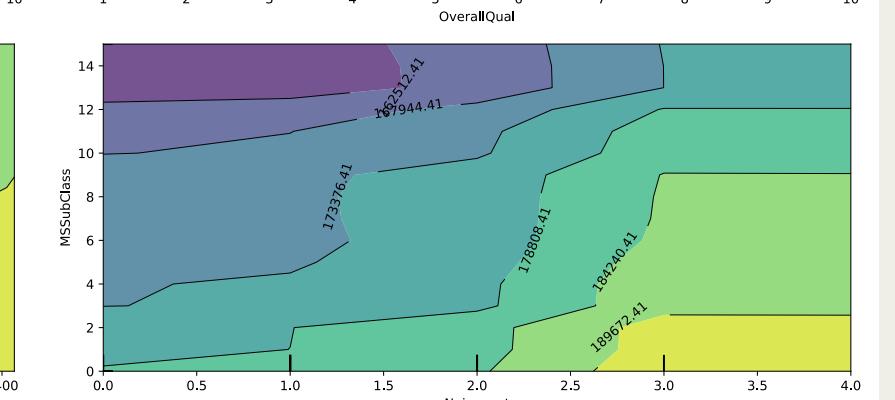
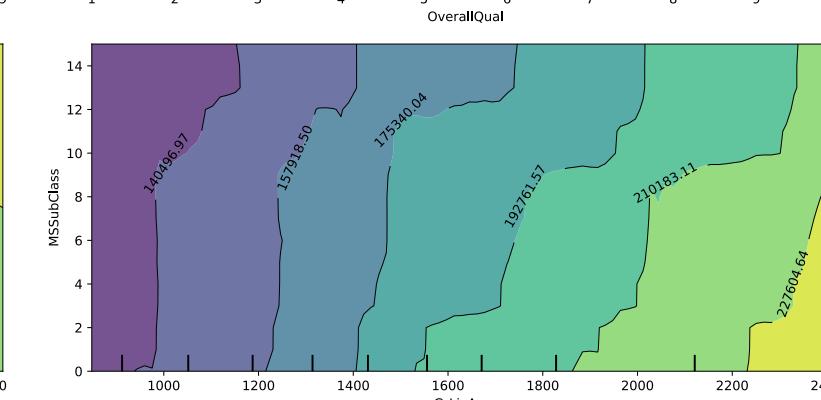
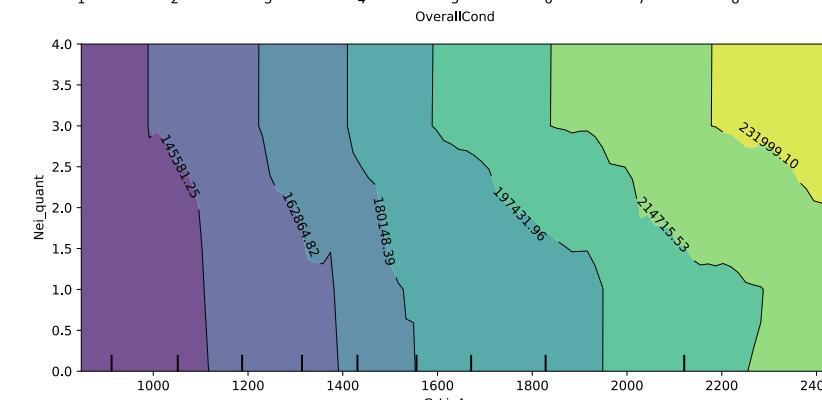
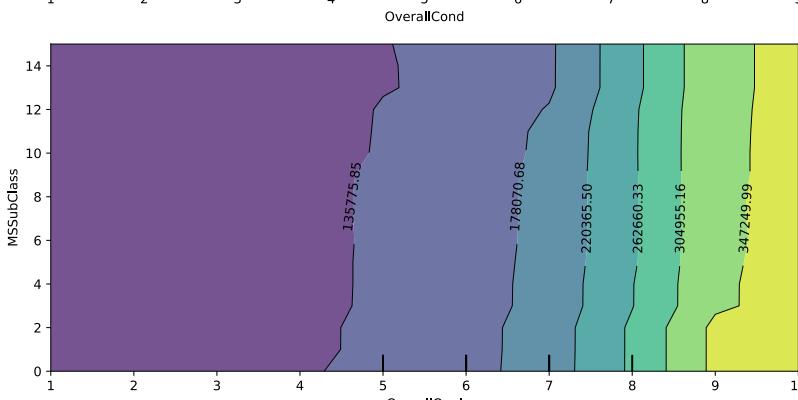
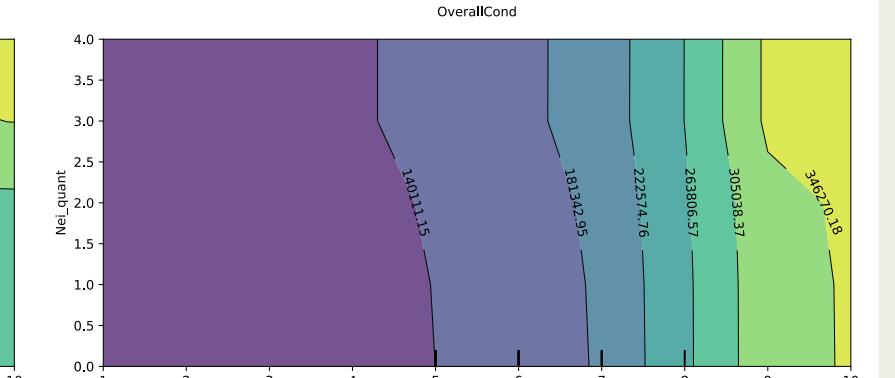
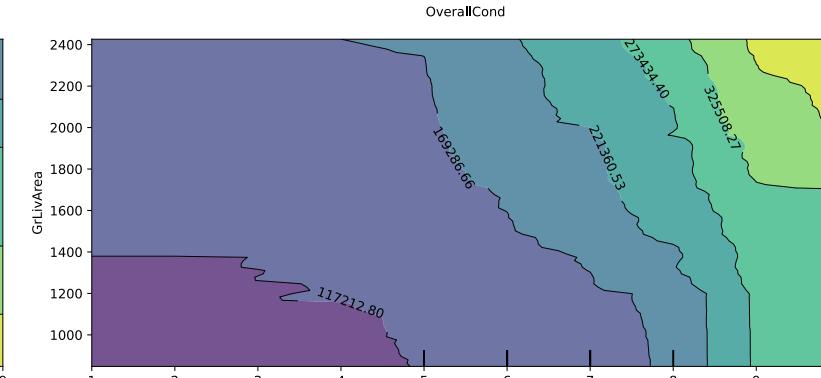
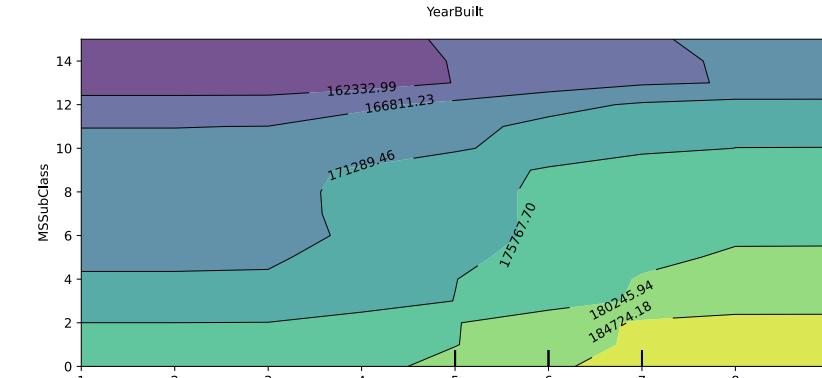
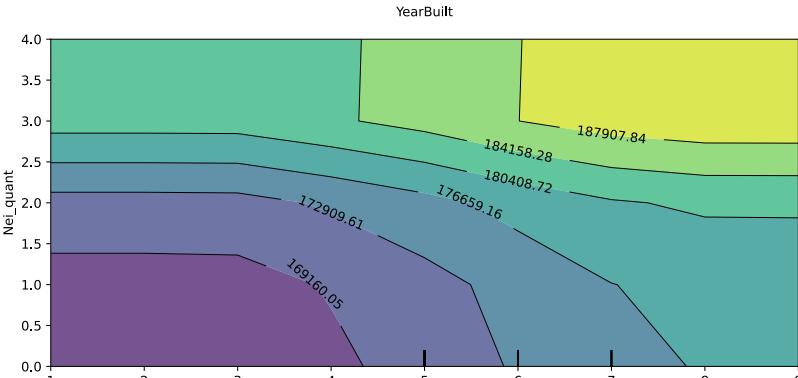
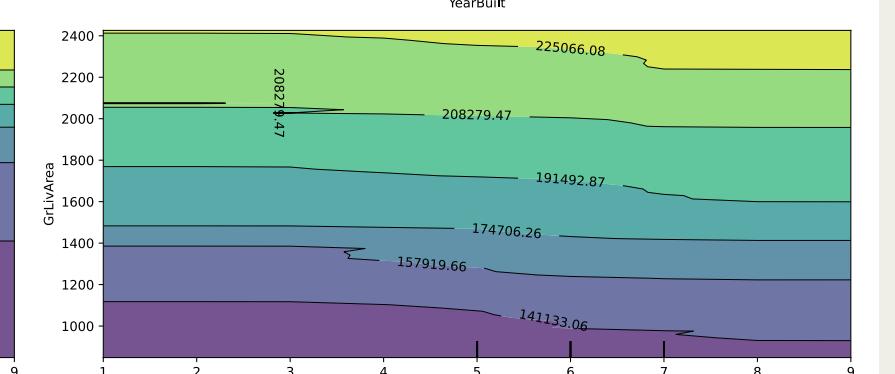
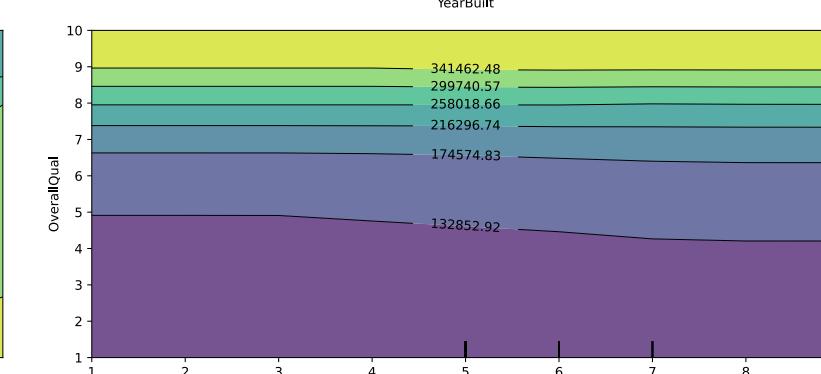
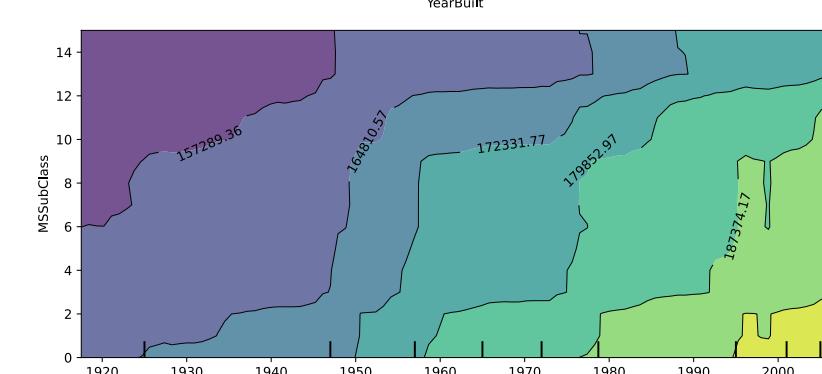
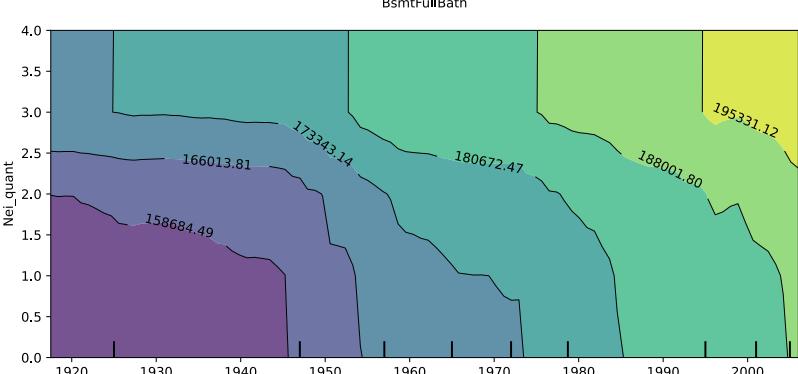
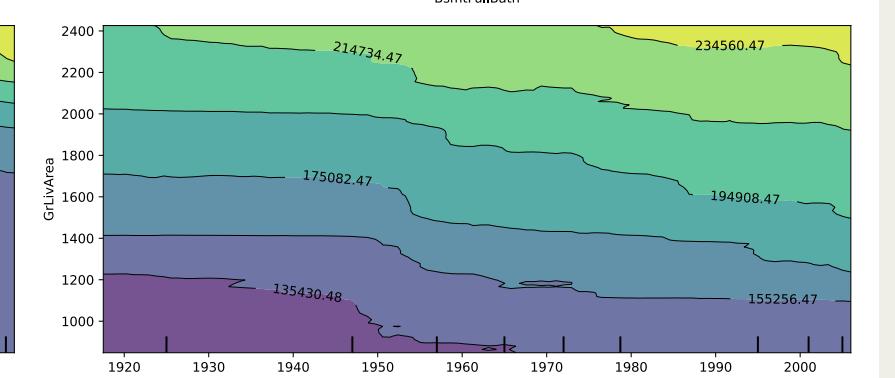
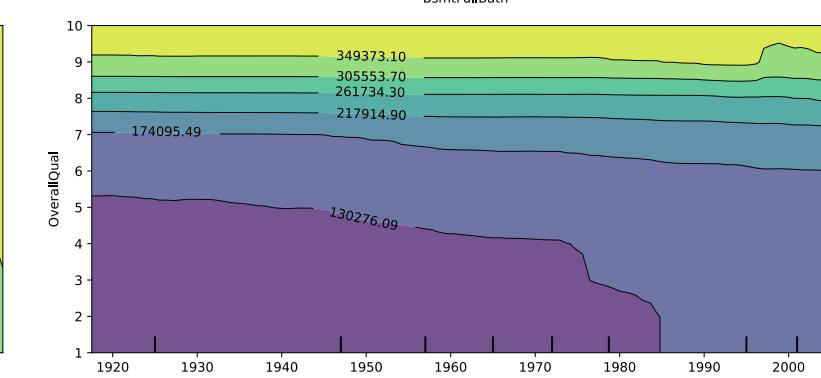
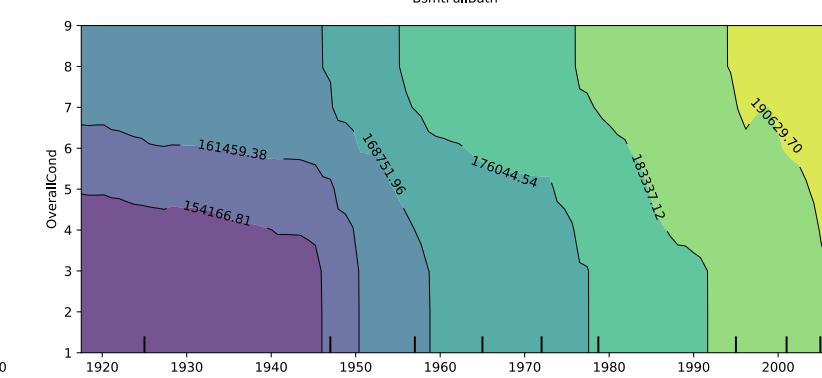
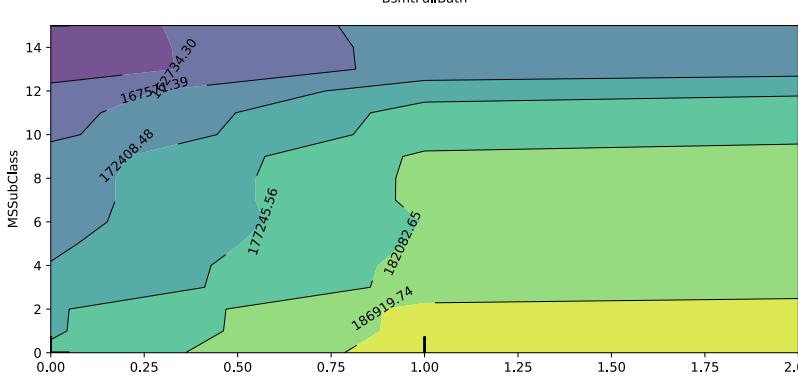
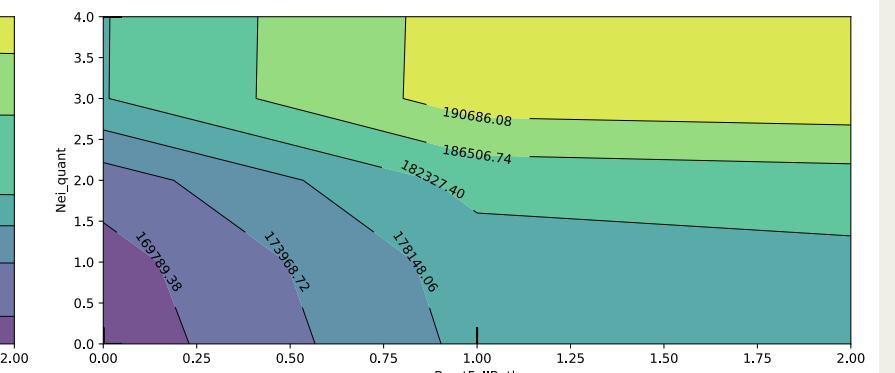
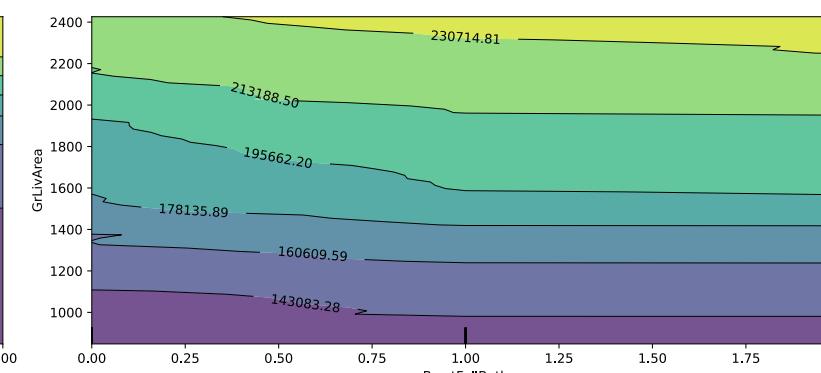
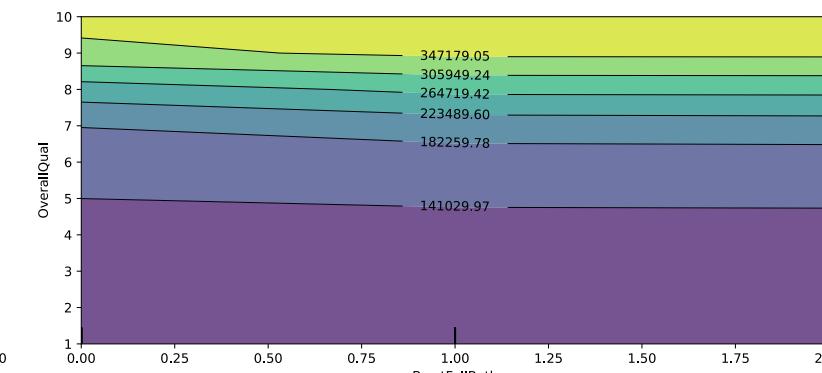
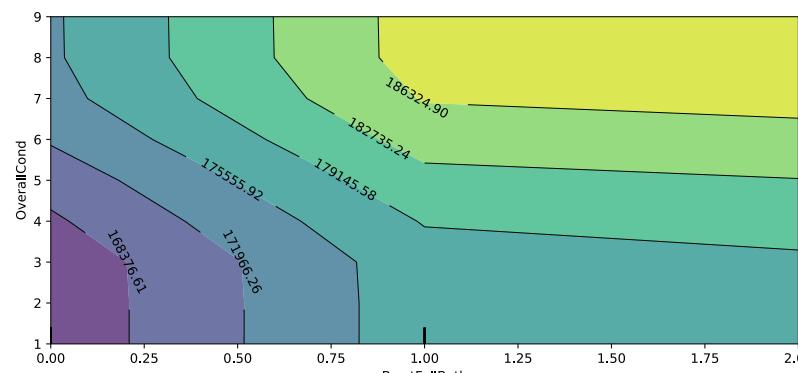
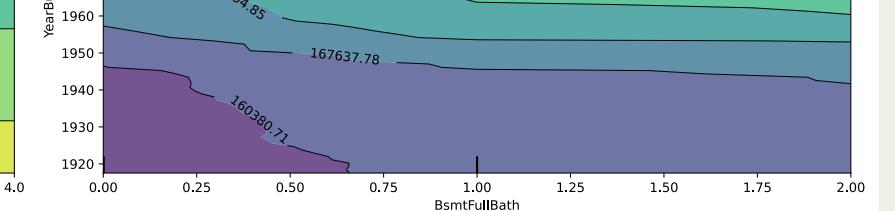
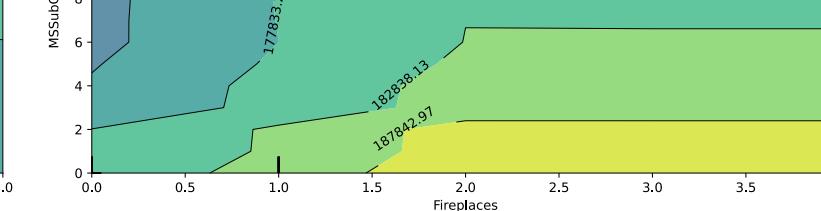
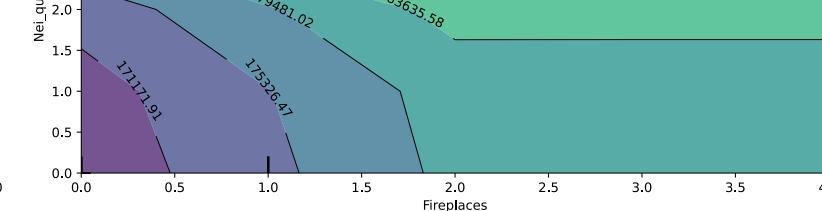
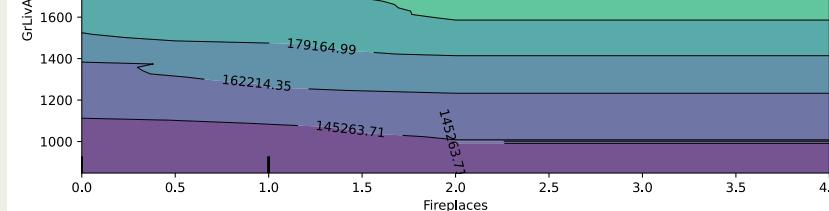
RANDOM FOREST FEATURE IMPORTANCE
[(0.593, 'OverallQual'),
(0.217, 'GrLivArea'),
(0.054, 'YearBuilt'),
(0.047, 'GarageCars'),
(0.021, 'Fireplaces'),
(0.021, 'MSSubClass'),
(0.019, 'BsmtFullBath')
(0.018, 'Nei_quant'),
(0.009, 'OverallCond')]



EASY FEATURE DEPENDENCE PLOTS

Partial Dependence Plots of Easy Features





A G E N D A

- Background
- Project Objectives
- Processing Methodology
- Predictive Modelling
 - Linear Models
 - Tree Based Models
- Descriptive Modelling
- Conclusion
- Appendix

LEARNINGS

**Predictive Model Selection is
an iterative process
complemented by research
and Descriptive Analysis.**

**Be willing to be disproven.
Anecdotal vs quantitative
evidence may differ.**

**This model may not work for
all locations. Models are
specific.**

**Define business object and set
attainable goalposts.**

**There is always more to do.
Know when to stop.**

FINDINGS & FUTURE WORK

Biggest surprise: SalePrice was impervious to the 2008 Housing Crisis.

Many things happened under the hood. Within this dataset and outside.

However, the SalePrice was not impacted. What degree of government intervention resulted in no change in the marketplace?

Quality, Quality, Quality!

More than Location, housing value is tied closely to Overall Quality, the overall material and finish of the house.

For those who want to remodel, quality drives prices more than square footage. Good news for those who don't want to increase the envelope of their house.

Investing in good quality materials is closely tied to increasing the value of a house.

Future Work

- Investigate how SalePrice Increase when improving the quality of certain aspects of the home.
 - Used to inform how to remodel in order to increase the value of the home.
 - Ex. Kitchen Quality vs Exterior Masonry, which increases SalePrice the most?
- If Low and LowMed neighborhoods are similar in size and OverallQuality, what is it about Low Neighborhoods that decrease SalePrice?
- Use Recursive Feature Elimination to improve feature set.
- Run Feature set through more models, XGBoost, and SVG to get to a better outcome.

Thank you!

For more, see Appendix



Katie Kwan, Data Scientist

A G E N D A

- Background
- Project Objectives
- Processing Methodology
- Predictive Modelling
 - Linear Models
 - Tree Based Models
- Descriptive Modelling
- Conclusion
- Appendix
 - **Lasso, Ridge, Elastic Net and Visuals**
 - **Tree Models, Ice Plots and Visuals**

PREDICTIVE: RIDGE AND ELASTIC NET RESULTS

Reduced Feature Set Results

1. Ridge

a. A loop on a range of alphas was run (similar to Lasso). Alpha was determined where training R squared cross testing R squared

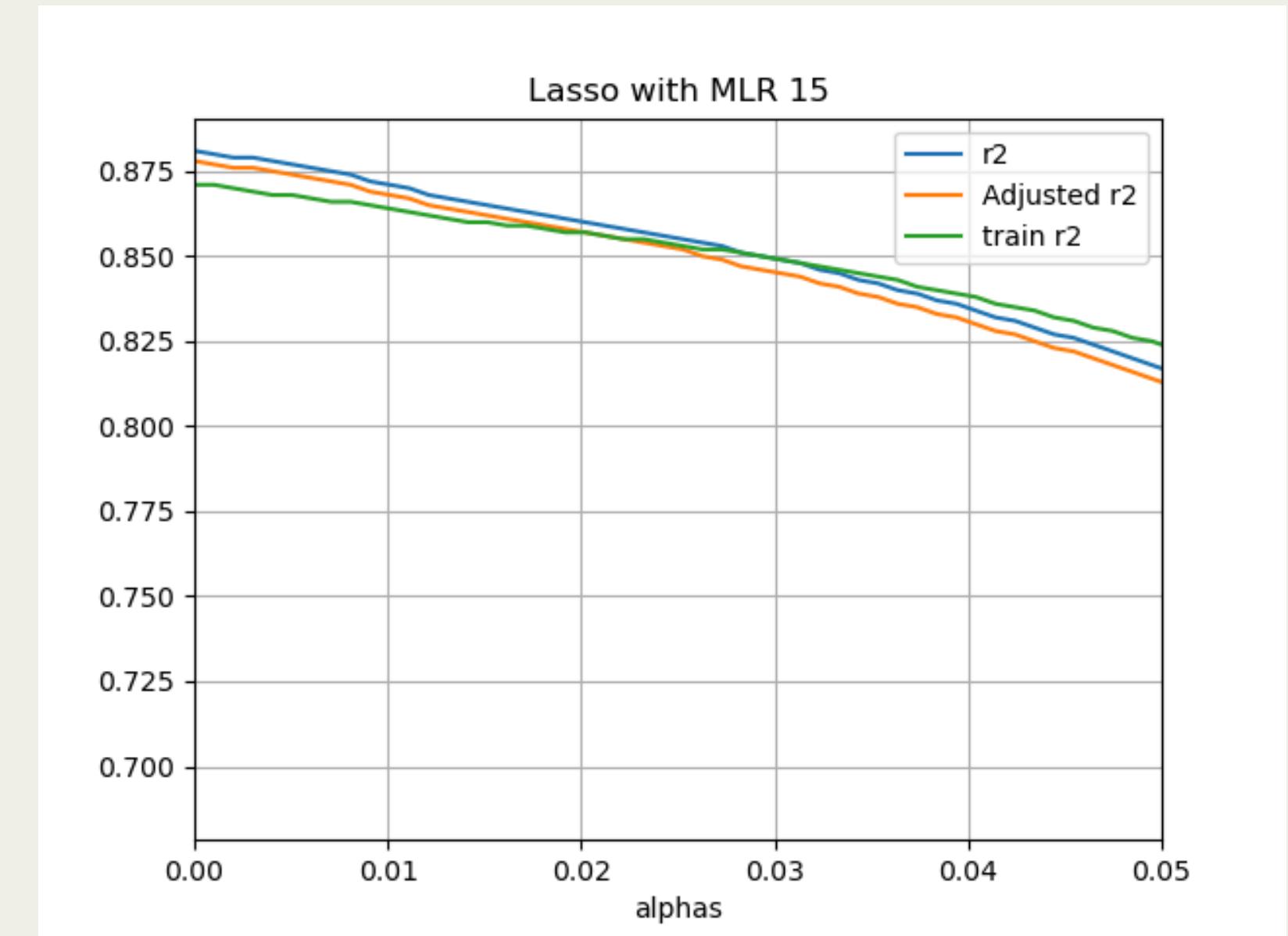
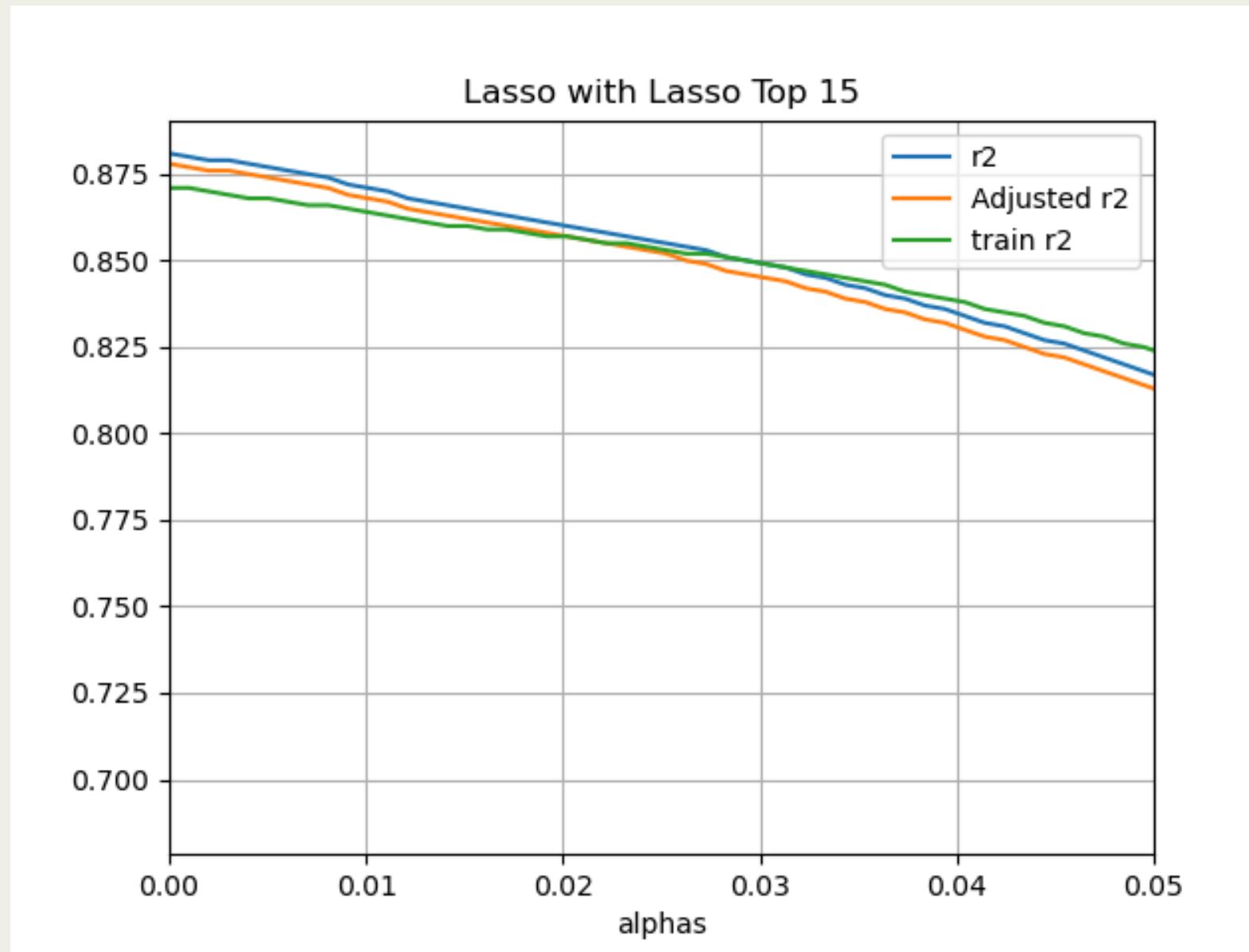
2. Elastic Net

a. A grid search was run on the feature options. Best alpha and l1_ratio were returned

	MLR Top 15	Lasso Top 15
Ridge	.846 alpha = 1666	.858 alpha = 707
Elastic Net	.872 alpha = .0006 l1_ratio = .666 cross validation = 5 scoring = r2	.885 alpha = .01 l1_ratio = 0.0 cross validation = 5 scoring = r2

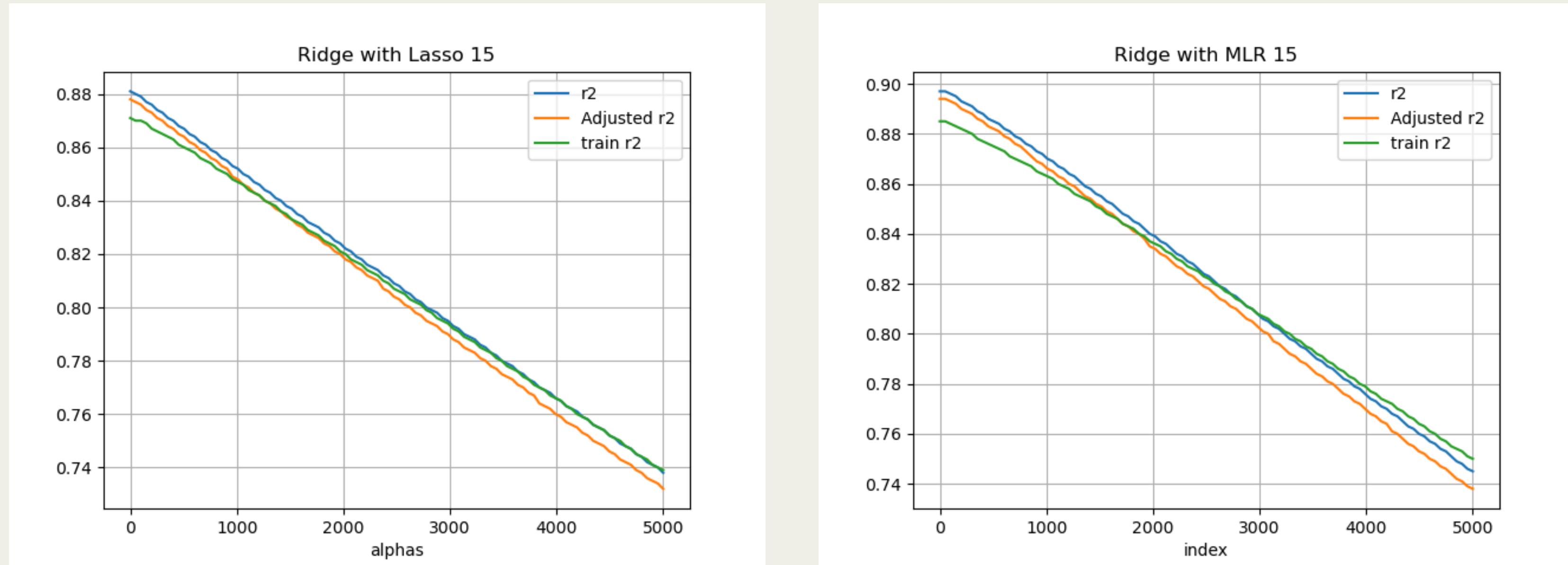
Elastic Net, a hybrid Ridge and Lasso model performs best better than Ridge.

PREDICTIVE: PLOTTING ALPHA FOR LASSO



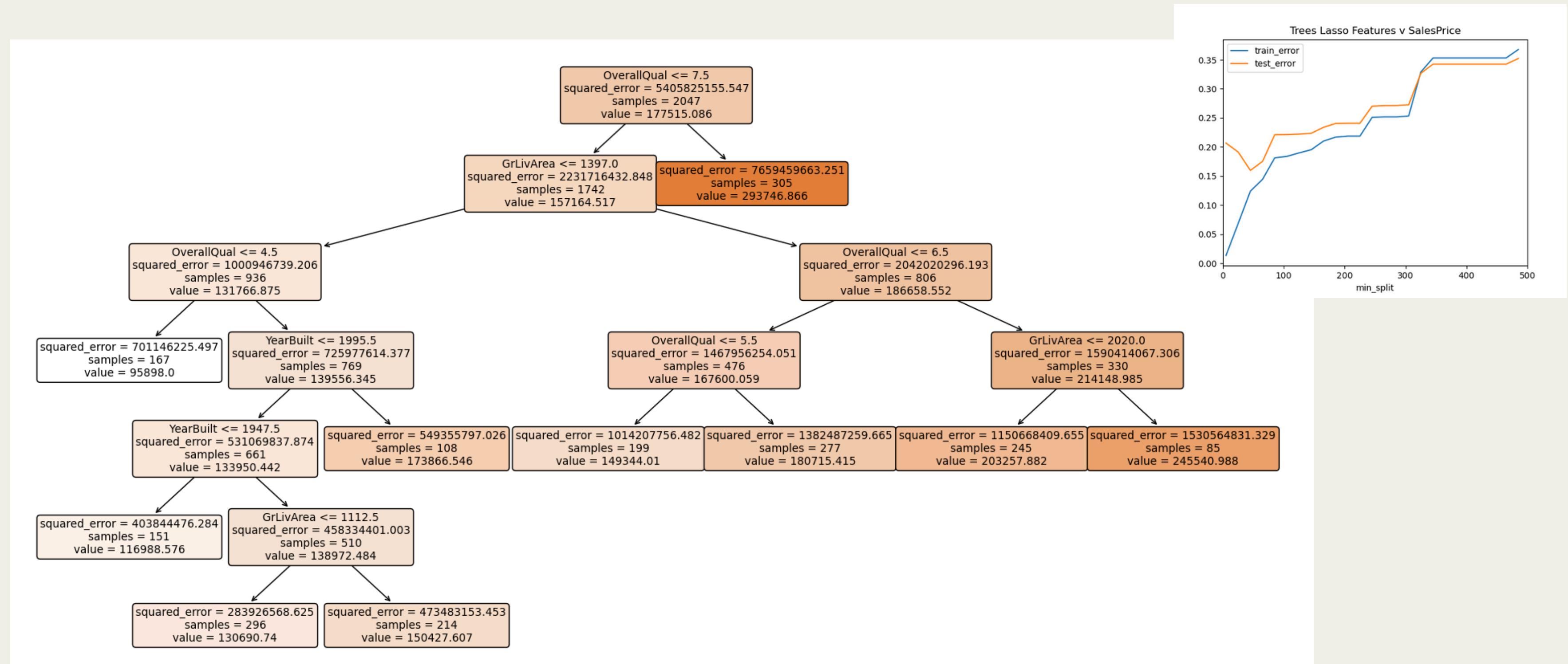
Adjusted R2 and train r2 cross at the optimal alpha.

PREDICTIVE: PLOTTING ALPHA FOR RIDGE



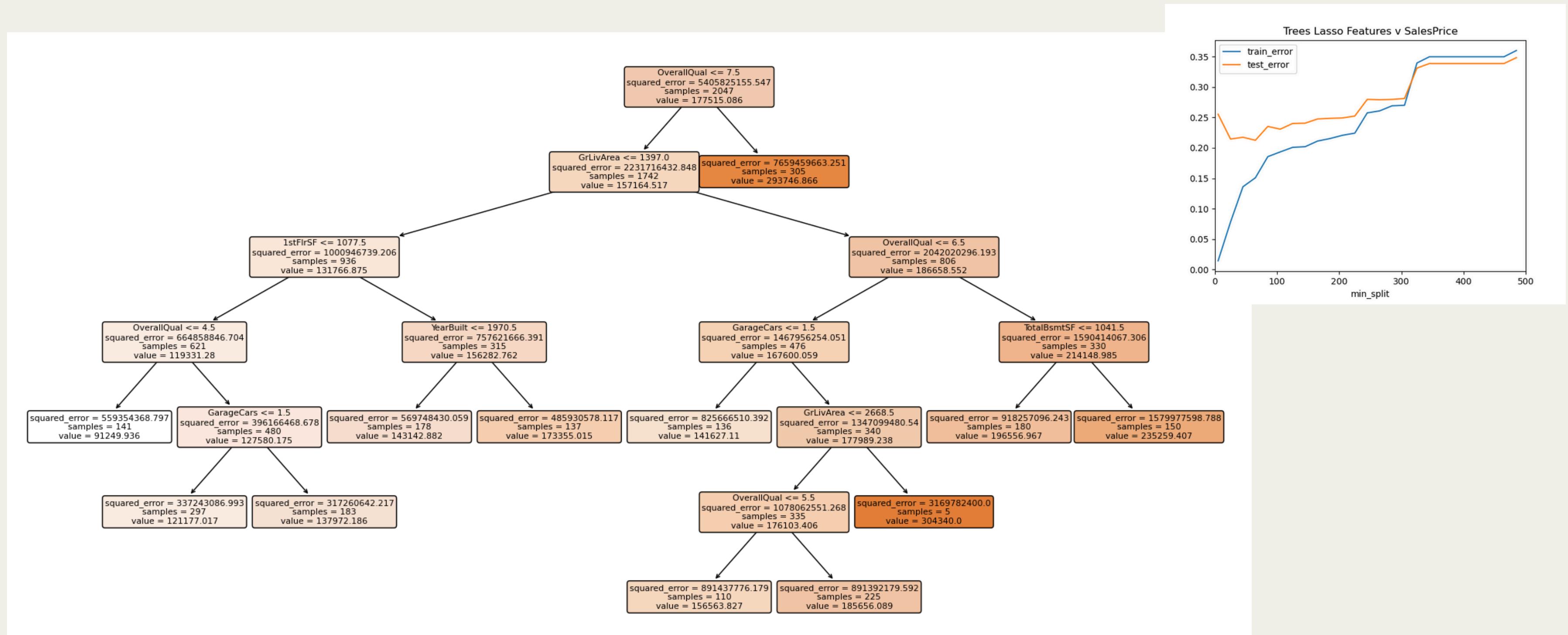
Adjusted R2 and train r2 cross at the optimal alpha. Accuracy v alpha appear linear.

PREDICTIVE: TREE OF LASSO FEATURES R2 .667



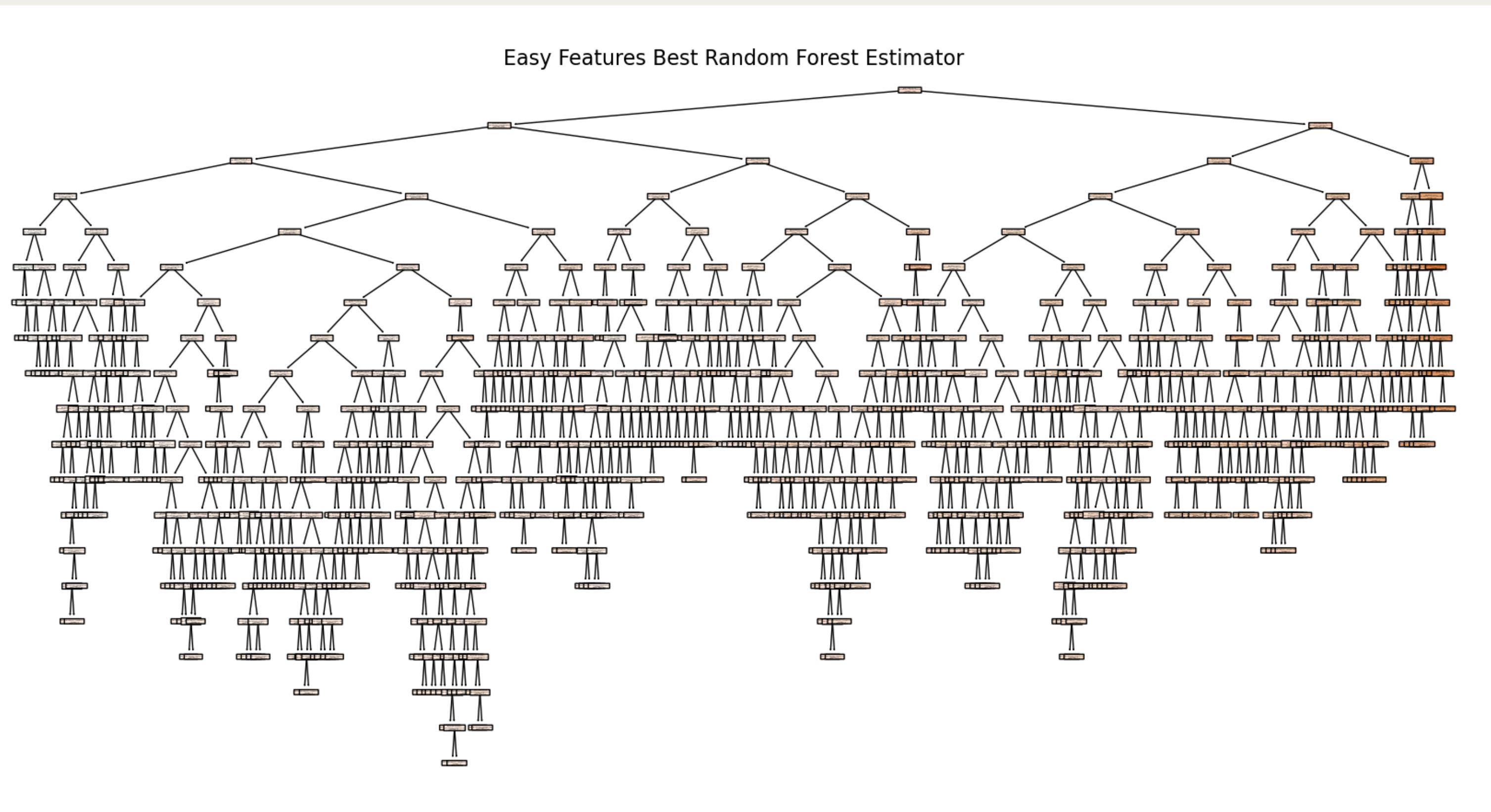
Only 3 features used: Overall Qual, GrLivArea, YearBuilt, 9 were fed to model

PREDICTIVE: TREE OF MLR FEATURES R2 .678



Only 3 features used: Overall Qual, GrLivArea, 1st FlrSF, GarageCars, TotalBsmtSF, YearBuilt. 15 were fed to the model

EASY FEATURE FOREST AND FEATURES



EASY FEATURE RANDOM FOREST ICE PLOTS

