BA305:
Predicting House Prices in Iowa

Team 20: Victor Floriano, Mahima Masetty, Aneri Patel, Jordan Teman

Disclaimer: The RMSE scores for our models have changed since the presentation. The issue we were running into at the time of the presentation with keeping the train/test split stable for all models has since been fixed which contributed to this change. Additionally, as suggested, we accounted for how new the houses are by creating a Years Since Remodel variable using Year Sold and Year Remodel Added variables instead of using the median of Year Remodel Added as a threshold to classify houses remodeled before that year as 'Old' and after that year as 'New'. All these changes are reflected in the Final Report and the Collaboration Notebook,

Meet the Team

Aneri Patel

Jordan Teman

Mahima Masetty

Victor Floriano

Project & Data Description

Predicting House Prices In Iowa with Business Analytics

According to Zillow, the nationwide median error rate for the Zestimate for on-market homes is 1.9%, while the Zestimate for off-market homes has a median error rate of 6.9%

Our Goal: Predict house prices in Ames, Iowa using various supervised learning methods taking inspiration from Zillow's Zestimate model. We also want to know the variables that highly effect price of a particular house.

Use Case: Our model could be used to check current house prices in Ames, Iowa to see if they are over/undervalued

Pre-Processing and Feature Selection Helped With Data Dimension Reduction

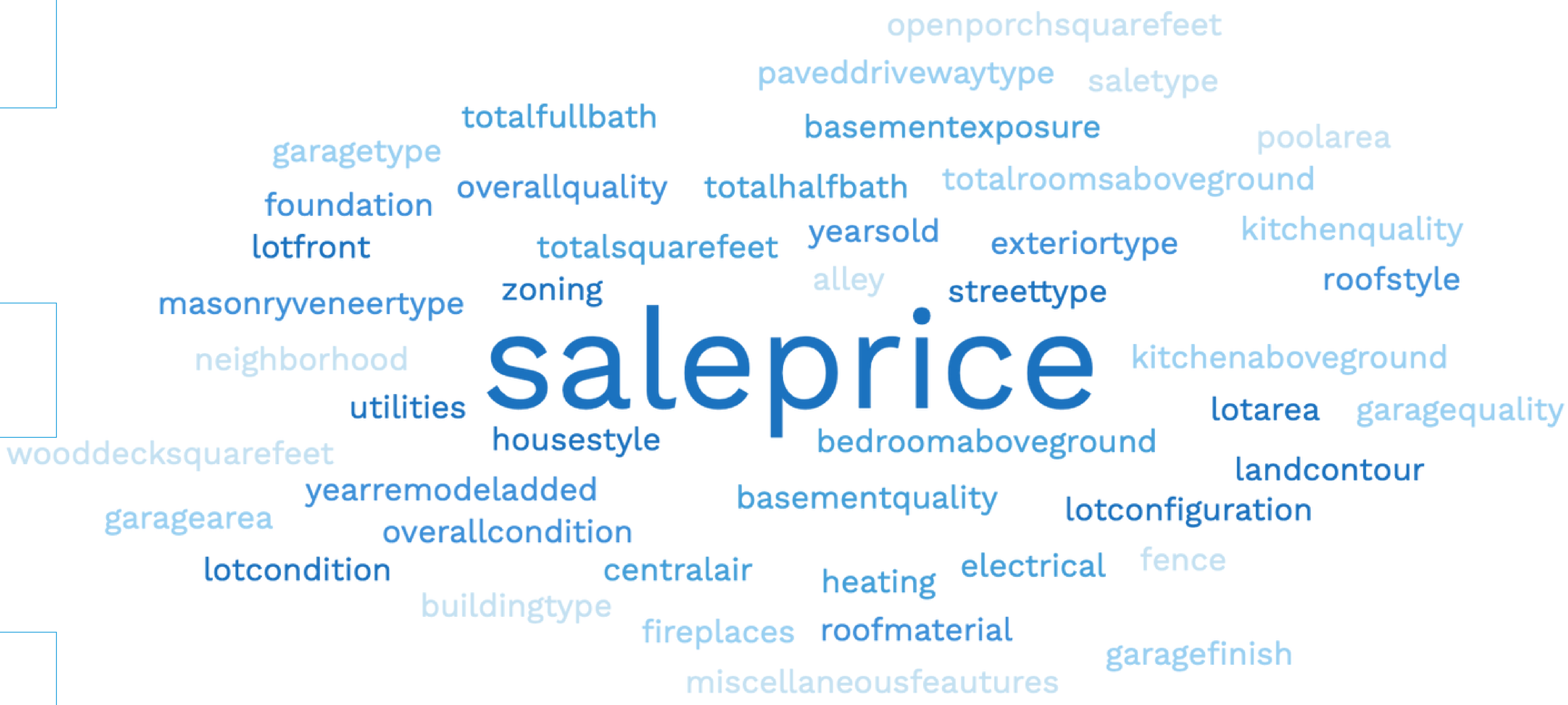
247 Predictors
37 Numerical | 210 Categorical

Pre-processing

115 Predictors
27 Numerical | 88 Categorical

Feature Selection

23 Predictors
11 Numerical | 12 Categorical



Data Pre-Processing

Dropping Predictors to Improve Data Usability

Based on NaN Values

Dropped predictors if more than 20% of their values had NaN

Some predictors dropped: Miscellaneous Features, Alley, Pool Quality, Fence

Based on Intuition

Looked through all the variable descriptions and manually dropped about 25 predictors because:

1. They did not intuitively seem to be relevant
2. They explained/captured the same information as some other predictor(s)
3. They were being transformed into a new variable that captured/explained more of our dataset (feature engineering)

Some predictors dropped: Id, Land Slope, Masonry Veneer Area, No. of Cars In Garage (Garage Sq. Ft. gives similar information), Year Remodel Added (variable to show if a house was new or old was created instead)

Changing String Rankings to Numerical Ranking To Reduce Data

Predictors About Ranking

- These predictors provided ranking in strings
- They would have to be turned into 5/10 categorical variables each
- To reduce our data, we turned the string rankings into numerical rankings

	BsmtQual	BsmtCond	KitchenQual	HeatingQC	GarageQual	GarageCond
0	Gd	TA	Gd	Ex	TA	TA
1	Gd	TA	TA	Ex	TA	TA
2	Gd	TA	Gd	Ex	TA	TA
3	TA	Gd	Gd	Gd	TA	TA
4	Gd	TA	Gd	Ex	TA	TA

	BsmtQual	BsmtCond	KitchenQual	HeatingQC	GarageQual	GarageCond
0	4	3	4	5	3	3
1	4	3	3	5	3	3
2	4	3	4	5	3	3
3	3	4	4	4	3	3
4	4	3	4	5	3	3

Combining Variables with Similar Information About House Size to Reduce Data

Predictors About House Size

- These predictors divided up the house into basement and above ground levels
- To reduce data, we combined different pairs to capture information about the entire house

	BsmtFullBath	FullBath	BsmtHalfBath	HalfBath	TotalBsmtSF	GrLivArea
0	1	2	0	1	856	1710
1	0	2	1	0	1262	1262
2	1	2	0	1	920	1786
3	1	1	0	0	756	1717
4	1	2	0	1	1145	2198

	Total_Full_Bath	Total_Half_Bath	Total_SF
0	3	1	2566
1	2	1	2524
2	3	1	2706
3	2	0	2473
4	3	1	3343

Combining Variables with Similar Information to Reduce Data

Predictors About Iterations of the Same Feature

Here we look at two sets of predictors:

1. What kind of exterior the house has (the house can have 2 exteriors) - 14 different types
2. What is the condition of the lot area (the house can have 2 lot areas) - 8 types of conditions

To reduce data, we combined these predictors.

	Condition1_Feedr	Condition2_Feedr	Condition1_Norm	Condition2_Norm	Condition1_PosA	Condition2_PosA
0	0	0	1	1	0	0
1	1	0	0	1	0	0
2	0	0	1	1	0	0
3	0	0	1	1	0	0
4	0	0	1	1	0	0

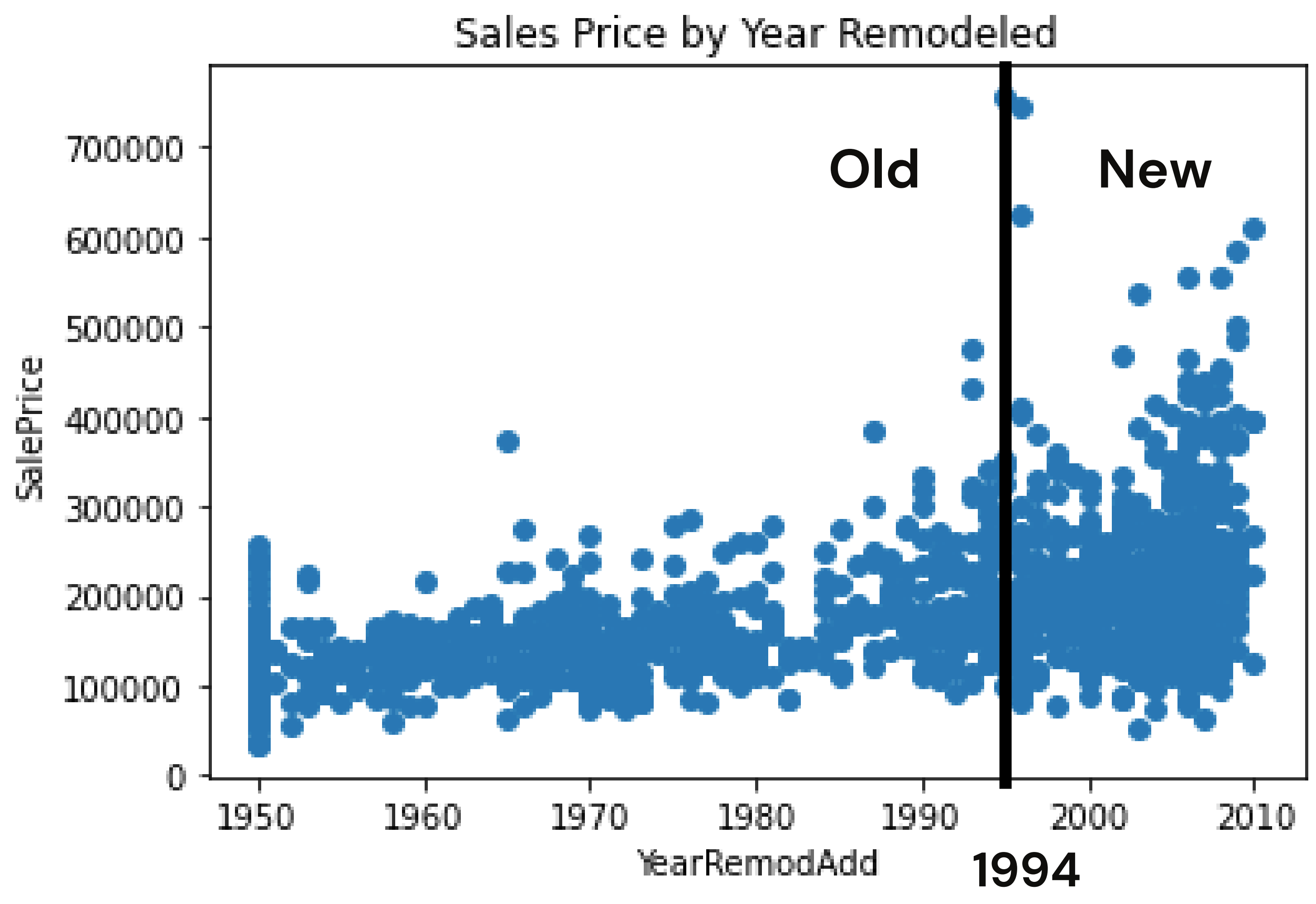
↓

	Condition_Feedr	Condition_Norm	Condition_PosA
0	0	2	0
1	1	1	0
2	0	2	0
3	0	2	0
4	0	2	0

Creating a New Variable to Capture the Age of the House

Predictors About Age of House

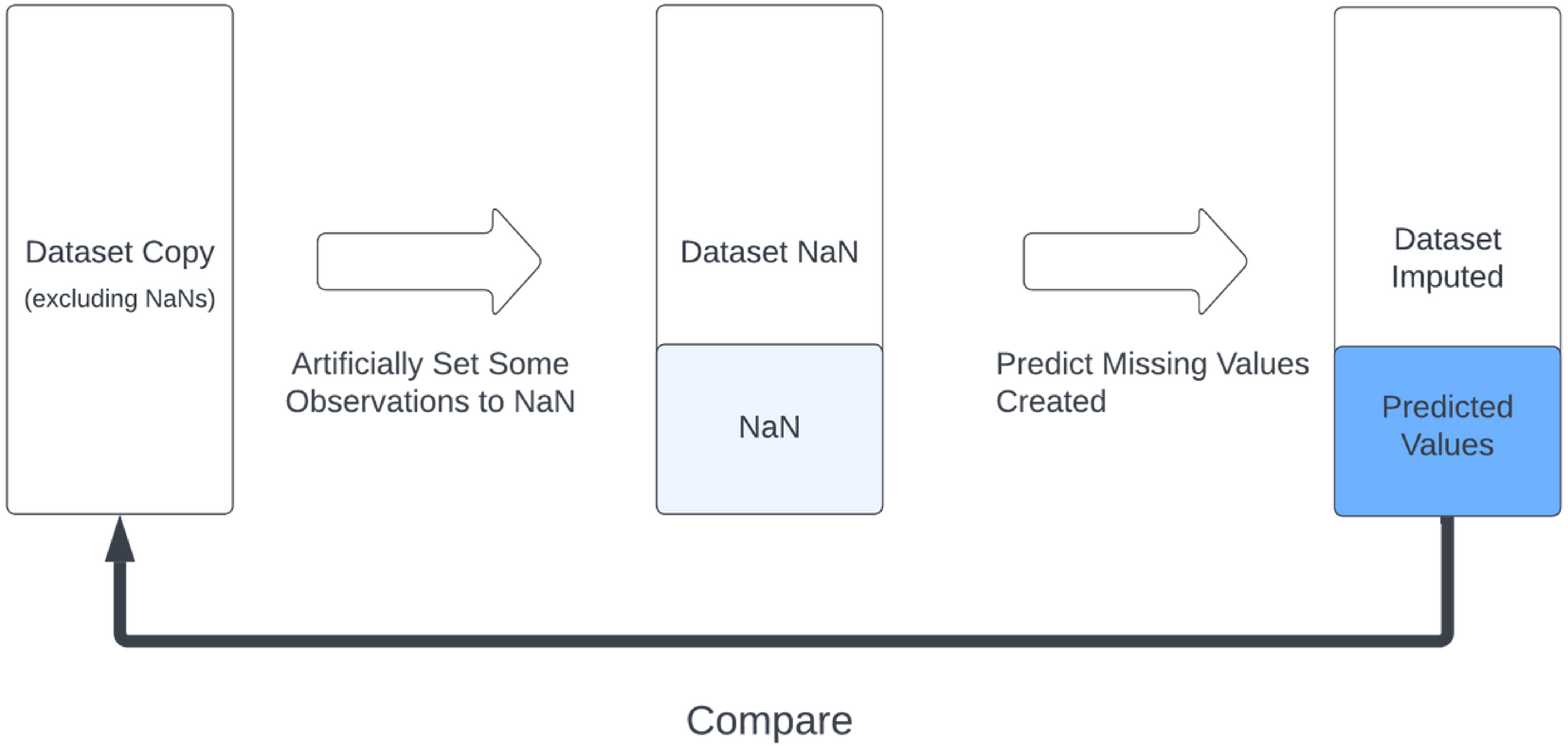
- Our dataset includes a predictor that provides information about when the house when was remodeled
- Since there are values of about 60 years, from 1950 to 2010, we would have had 120 more categorical predictors
- To reduce data, we used the median year of Year Remodel Added predictor to create a new variable that simply showed if the house was new or old



Using Imputation Method to Fill In Missing Values

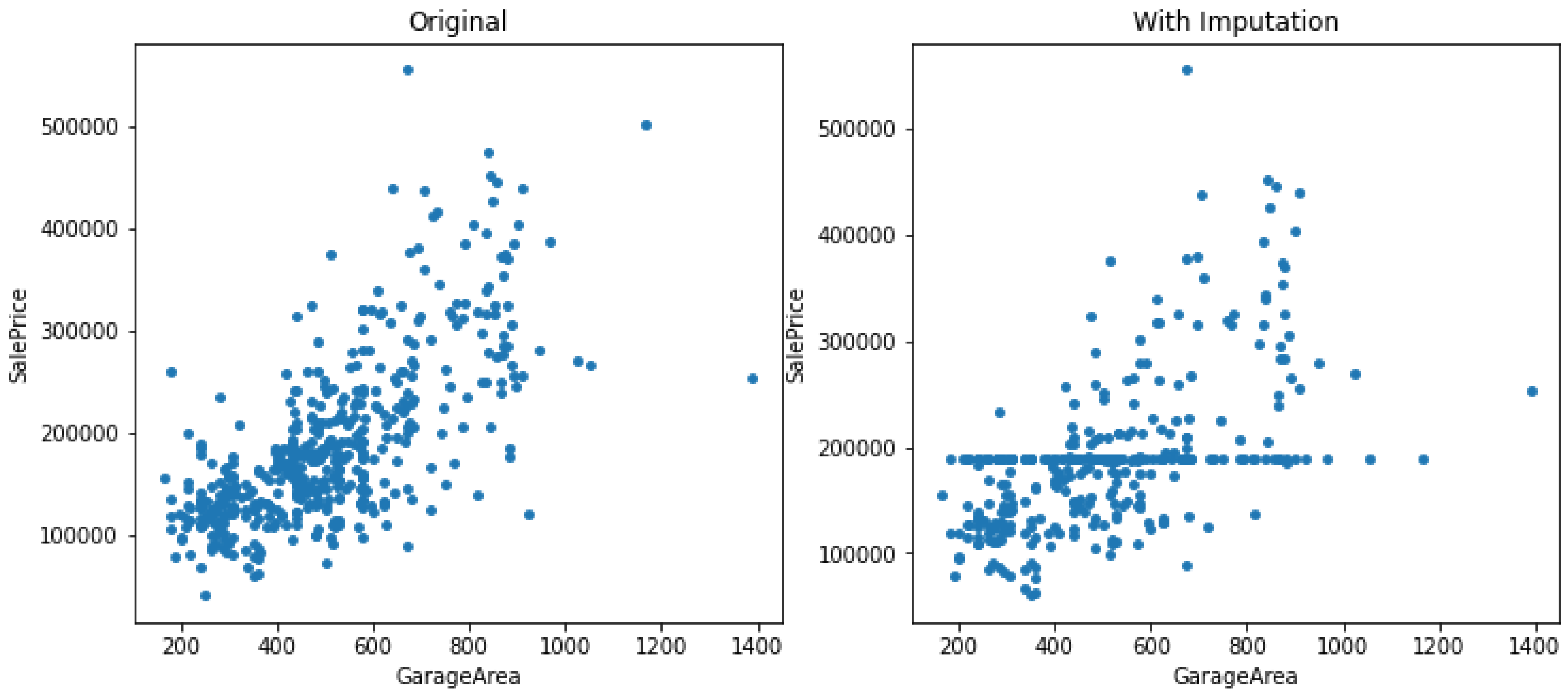
Motivation: To avoid removing instances that contained missing values, we decided to use an imputation method.

Process: Tested two types of imputation methods — Univariate Imputation using the Mean and Multivariate Imputation by Chained Equations (MICE).

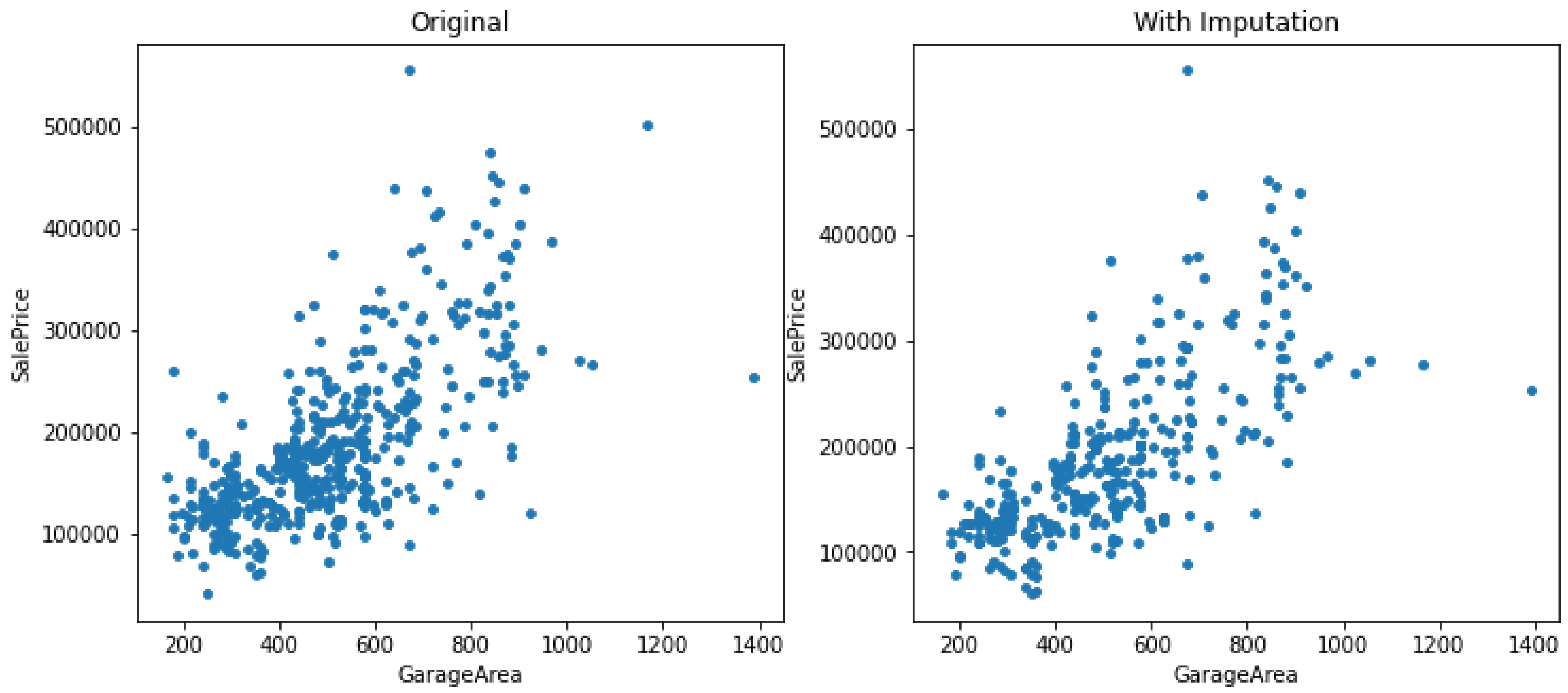


Imputating With MICE Results in a Similar Data Structure to the Original Dataset

Imputing with Mean



Imputing with Multivariate Imputation by Chained Equations (MICE)



Data Visualization

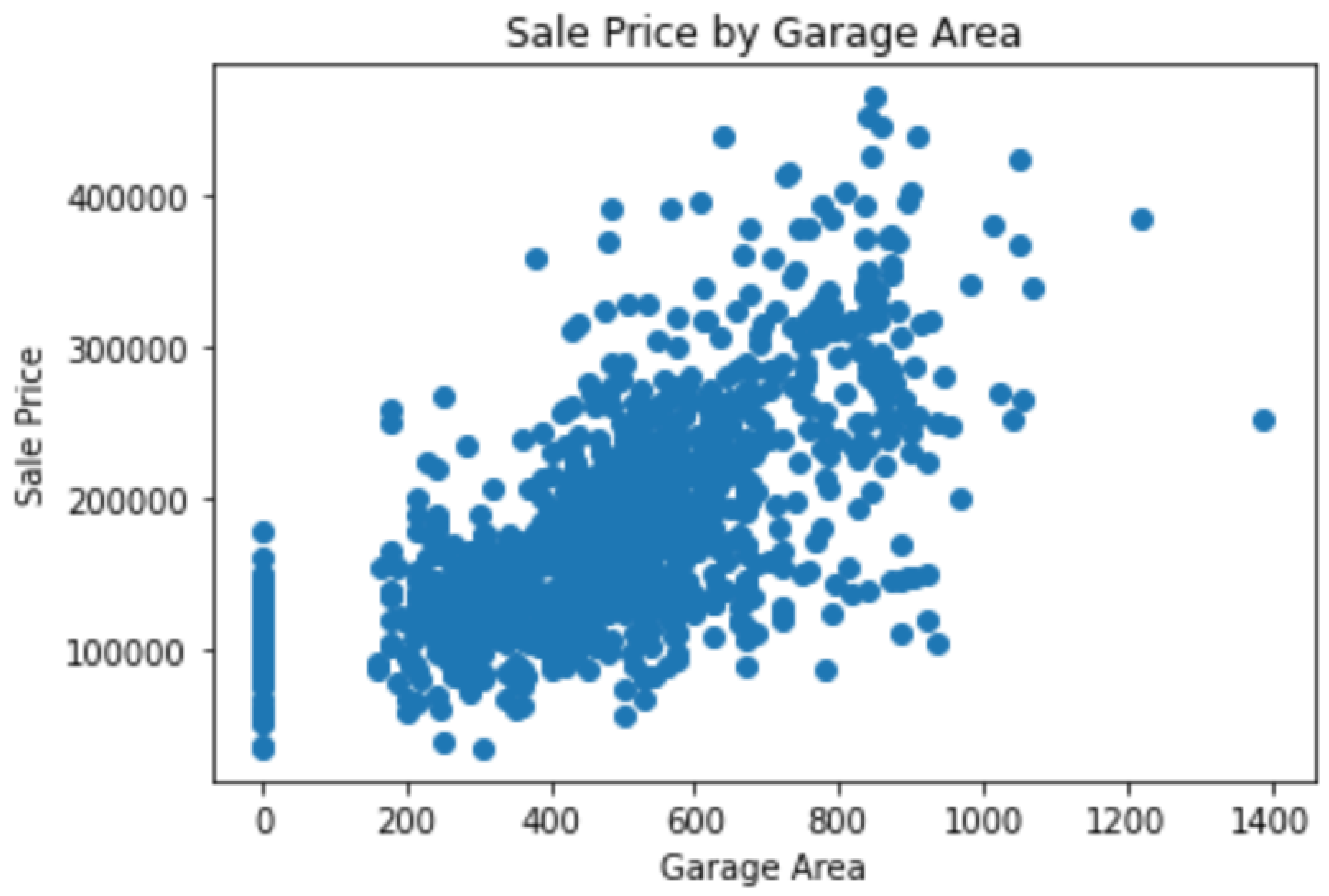
The Less Dense the Zone of the House, the Higher the Sale Price

Variable Descriptions:

C(all): mean of all the remaining zones
FV = Floating Village Residential
RH = Residential High Density
RL = Residential Low Density
RM = Residential Medium Density

- Grouped by different zones and captured the mean house price in each zone
- Floating Village Residential and Residential Low Density Zones have the highest priced homes
- The homes in less populated areas have higher prices

Sale Price Increases as Garage Area Increases



- Having a garage clearly impacts the sale price of the house
- The houses with no garage are priced lower on average than those with a garage
- The larger the garage the higher the sale price on average because there is space for more cars

Total Square Foot and Sale Price Have a Strong Correlation

- Expected total square footage to be one of the variables impacting sale price the most
- Removed outliers to better identify trend using interquartile ranges
- Decided to eliminate observations above the 75th quartile plus 1.5*IQR and below the 25th quartile minus 1.5*IQR

Lot Area is Less Correlated to Sales Price Than We Hypothesized Once the Outliers Are Removed

- Expected Lot Area to be another strong predictor of sale price
- Unable to accurately see the trend due to outliers. Therefore, removed the outliers by using the same interquartile range method as total square foot
- Not as directly correlated as total square foot of the house, but a correlation is evident

Modeling

Creating A Baseline Model With Our Hypothesized Most Impactful Variables to Get a Benchmark RMSE

Variable Selected

- Selected 31 variables (8 numerical and 23 categorical) expected to have a strong correlation with the house sale price
- Selected variables based off of intuition that they are highly correlated and confirmed it by visualizing correlations

Outcome

- RMSE = $31,309
- Not very large of an error considering the average house price is $175,000

Data Reduction with PCA Was Unsuccessful Due to Too Many Categorical Variables

Component	% of Variance	Cumulative %
0	0.093	0.093
1	0.037	0.129
2	0.033	0.162
3	0.027	0.189
4	0.023	0.213
5...	0.023	0.235
51	0.008	0.769
52	0.008	0.777
53	0.008	0.785
54	0.008	0.792
55	0.008	0.800

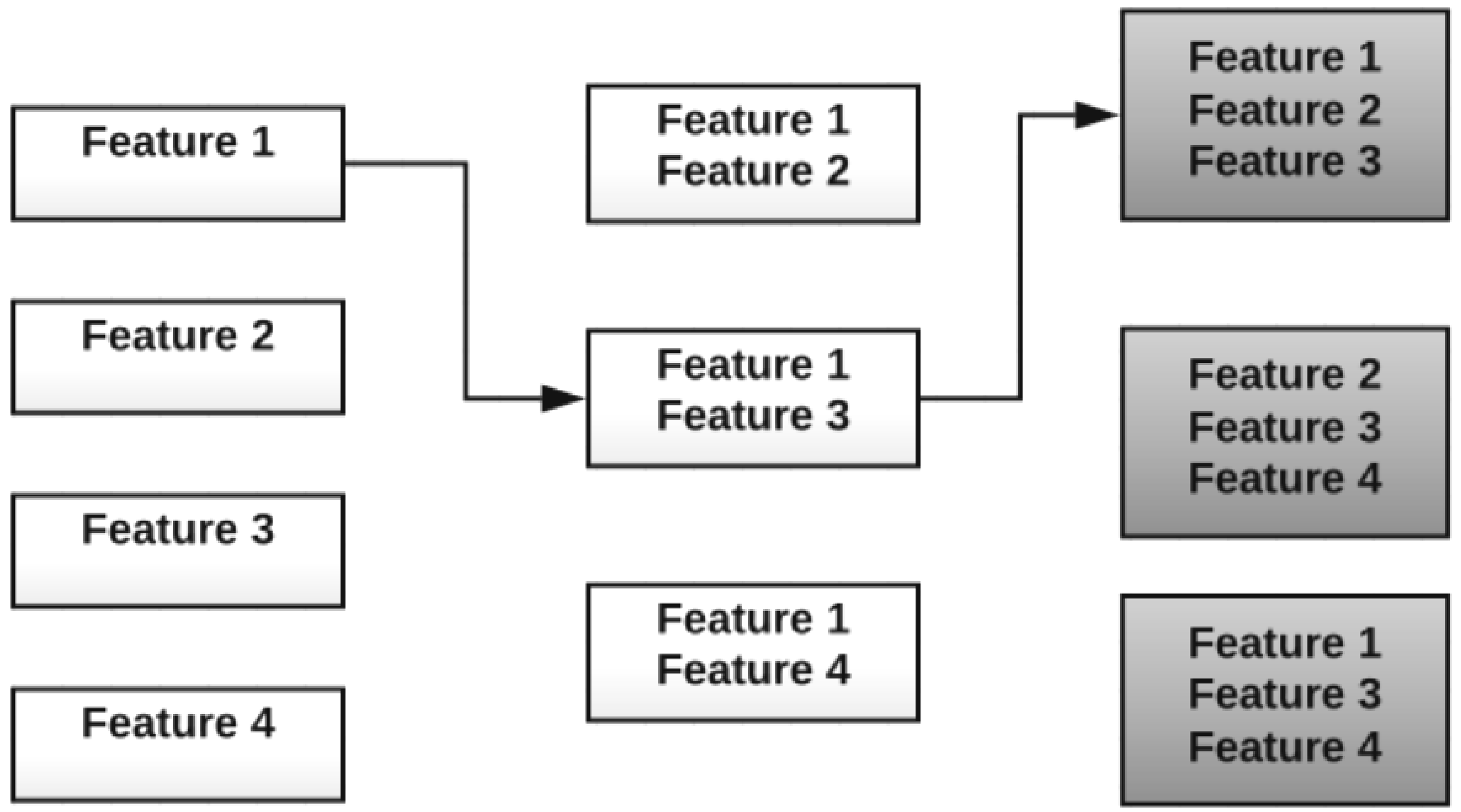
Ran a PCA on **115** variables and found that according to the cumulative proportions of these variables, we need **55** components to just explain **80%** of the variation.

This is likely because we have **88** categorical predictors after running a one-hot encoder.

Using Forward Selection to Select the Best Combination of Predictors

Problem: Current Dataset has 115 variables

Solution: To choose what variables to use in our models, we decided to use **Forward Selection** and setting it up to retain 20% of our original features.



Drawbacks of forward selection: It will not go through all possible combinations of predictors and it can take a lot of processing power with large datasets.

Some of the Feature Selected Overlapped With Our Baseline Model's Predictors

Scikit Learn method: SequentialFeatureSelection

Features Selected: 23 (20%) | 11 Numerical and 12 Categorical

Exterior - Cement Board

Garage Type - Built In

Land Contour - Hill Side

Zoning - Residential Medium Density

Fire Places

House Style - 2 Story

Lot Area

Kitchen Quality

Lot Frontage

Garage Finish - Rough Finish

Overall House Condition

No. of Kitchens

Overall House Quality

Total Sq. Ft.

Garage Area

No. of Bedrooms

Foundation - Slab

Basement Exposure

Basement Unfinished Sq. Ft.

Sale Condition - Partially Finished House

House Style - 2.5 Story Finished

Condition - Normal

Foundation - Concrete

Decided Not to Use Naive Bayes Due To Our Mixed Datatypes

Models Considered

Gaussian Naive Bayes: For our numerical variables

Multinomial Naive Bayes: For our categorical variables

Issue: Dataset had both numerical and categorical predictors and each version of Naive Bayes does not work well with variables of the other type.

Possible Solution: Transform each continuous variable into bins and use the Multinomial Naive Bayes.

Downside:

- Information loss
- Hundreds of more predictors (new categories)

K Nearest Neighbors Model Already Showed an Improvement Over Our Baseline Model

Steps:

1. Standardize dataset with StandarScaler( )
2. Instantiate a KNeighborsRegressor
3. Fit the model to our training data
4. Tried multiple values for K

Baseline RMSE = $31,309

Untuned Model

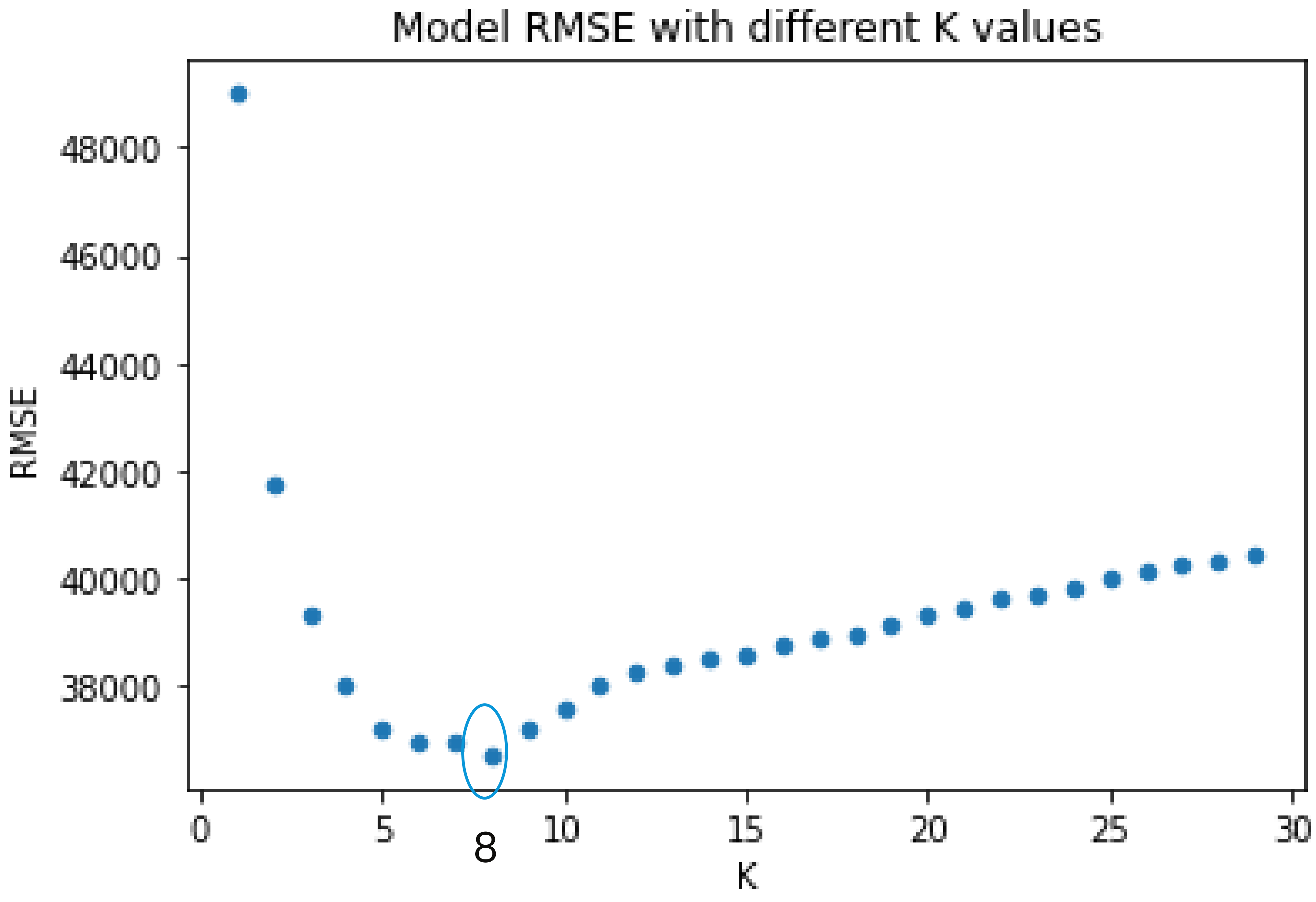
- K = 1
- RMSE = $31,901

→

Tuned Model

- K =8
- RMSE = $31,014

At K=8, Our Model Has the Smallest RMSE



Random Forest Model Showed a Significant Improvement Over the Baseline

Steps:

1. Create a new model with a Decision Tree Regressor
2. Loop through many values for the maximum number of leaves to prune the tree
3. Instantiate a Random Forest Model to try to improve over our Decision Tree

Baseline RMSE = $31,309

Decision Tree – Tuned

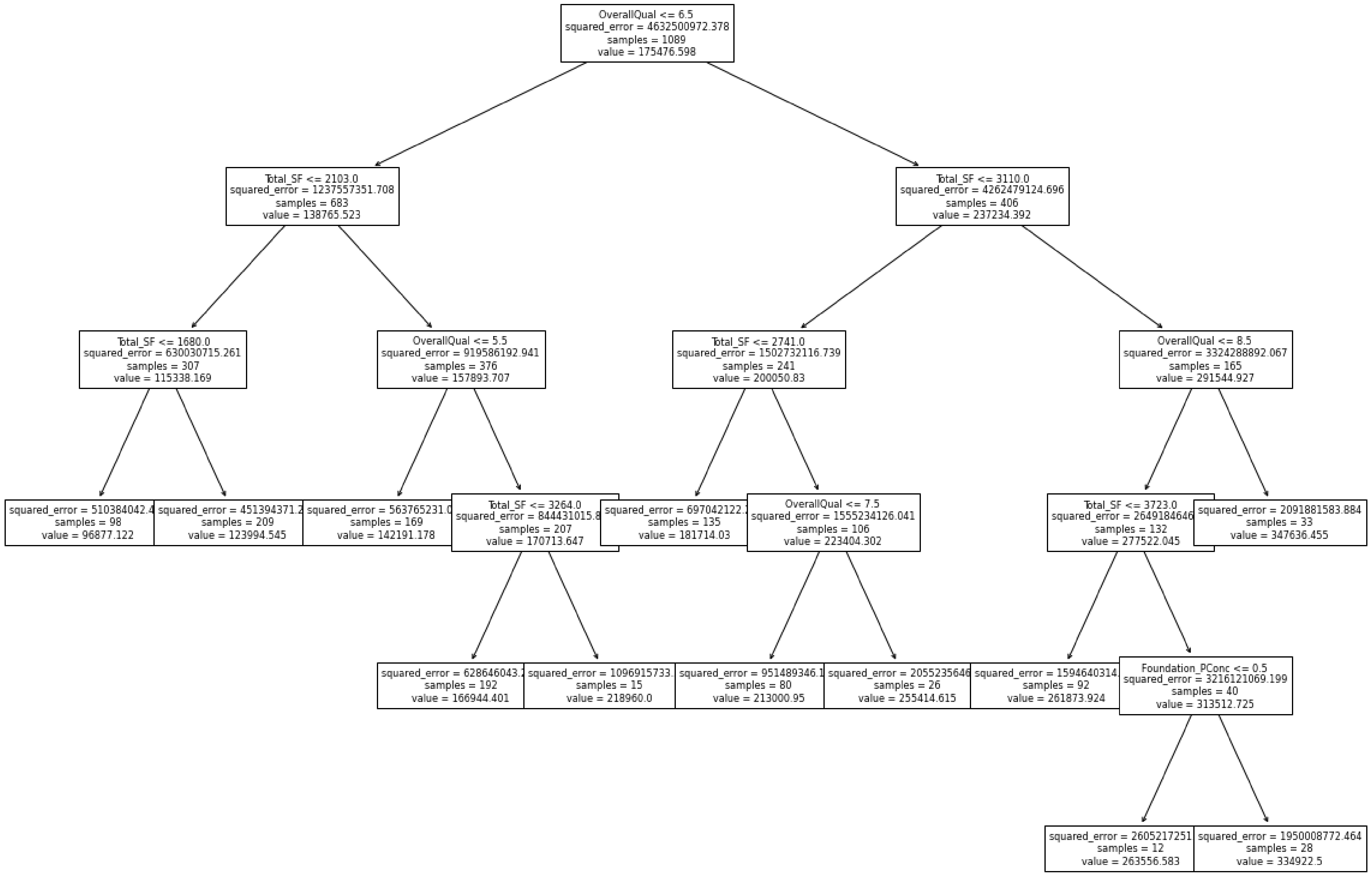
RMSE = $33,451
Max Leaves = 12

→

Random Forest

RMSE = $23,602

Pruning the Decision Tree Picked 12 as the Maximum Leaves



Explored Regularized Regressions to Lower Model Variance and Improve Predictive Power

Regularized Regressions:

1. Reduces overfitting by decreasing the flexibility of the model
2. Penalizes models with too many coefficients

Linear Regression Model

$$J = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y} - y)^2$$

Normal loss function

Lasso Model

$$L_{lasso}(\hat{\beta}) = \sum^{n} (y_i - x_i^T \hat{\beta})^2 + \lambda \sum^{m} |\hat{\beta}_j|$$

Sum of Square of Errors

Penalty Term

Using Lasso As Our Preferred Regularized Regression Method Led to the Best Model

Steps:

1. Create a baseline model — a Lasso model with alpha=0 (same as a Linear Regression)
2. Standardize data with StandardScaler( )
3. Instantiate a Lasso( ) model
4. Fit the model to our training data
5. Loop through many values for alpha

Baseline RMSE = $31,309

Linear Regression RMSE = $23,611

→

Lasso RMSE = $22,546

Tuned Lasso is the Best Prediction Model For Our Problem Due to its Low RMSE

Prediction Method/Model Used	RMSE
Baseline Model	$31,309
Decision Trees	$33,451
Random Forest	$22,699
KNN	$31,014
Lasso Tuned	$22,546

Key Takeaways

Key Takeaways

What Worked

- Forward selection validated some choices we made for predictors and allowed us to shrink our number of predictors
- Models based on Forward Selection features reached a low RMSE, indicating that the predictors we used had high predictive power
- Tuning our models with the best parameters led to lower RMSEs in every instance

What Could Be Better

- Larger Dataset - many of the predictors in our data had few observations
- Up to date data - only contained information up to the year 2010
- Better categorical variable descriptions
- Data from other states to scale the model to get accurate price predictions for other regions

Thank you!