

AMES House Price Prediction

Group TTP

Bee Kim

Drucila Lefevre

Ira Villar

Tomas Nivon

Background information

- Ames, Iowa
- 80 Variables
- 2006-2010
- Train Set
 - Sale Price
 - 1460 Data
- Test Set
 - 1459 Data
- Root-Mean-Squared-Error (RMSE) between predicted value and observed sales price



Source: GoogleMap

Objectives & Process

- Predict Sale Price using Train and Test dataset
- 1. Understanding the data
 - EDA
- 2. Cleaning the data
 - Drop Outliers
 - Drop Columns with Many Missing Values
 - Imputation on Missing Values
 - Skewed Numerical Variable Distribution -> Box-Cox Transformation
- 3. Feature Engineering
- 4. Machine Learning models

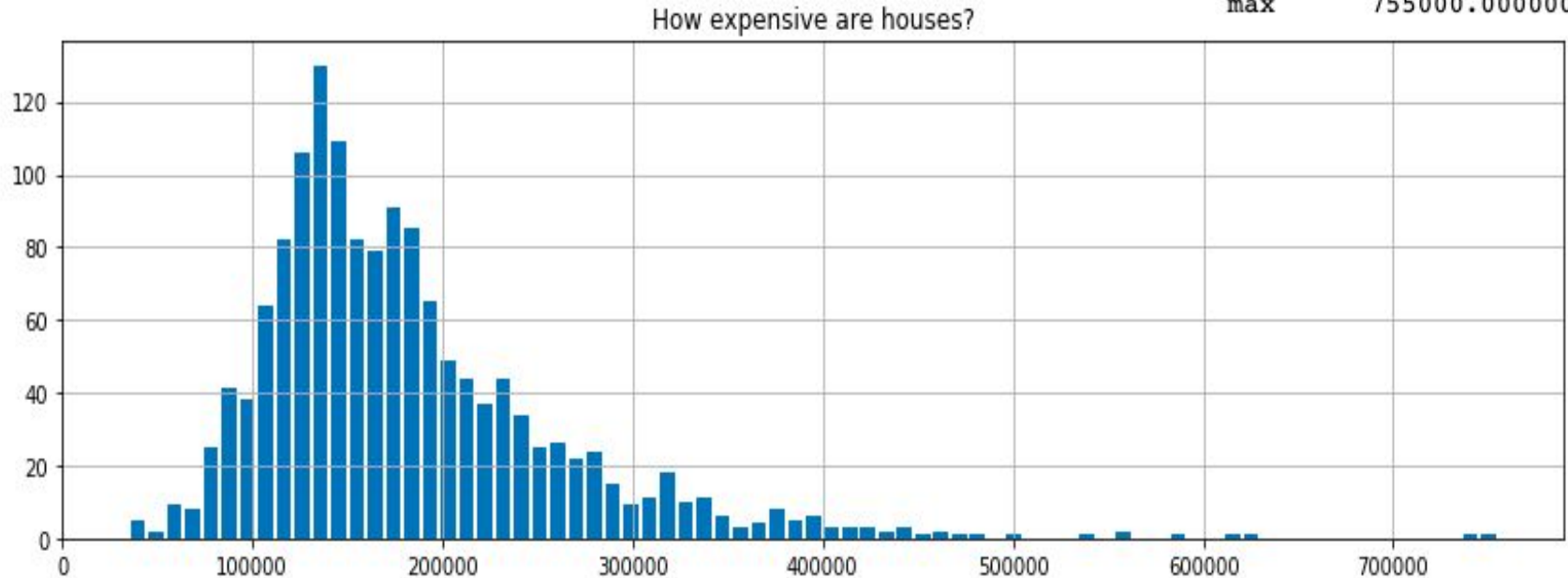
1

Understanding the Data

Objective: House Sale Price

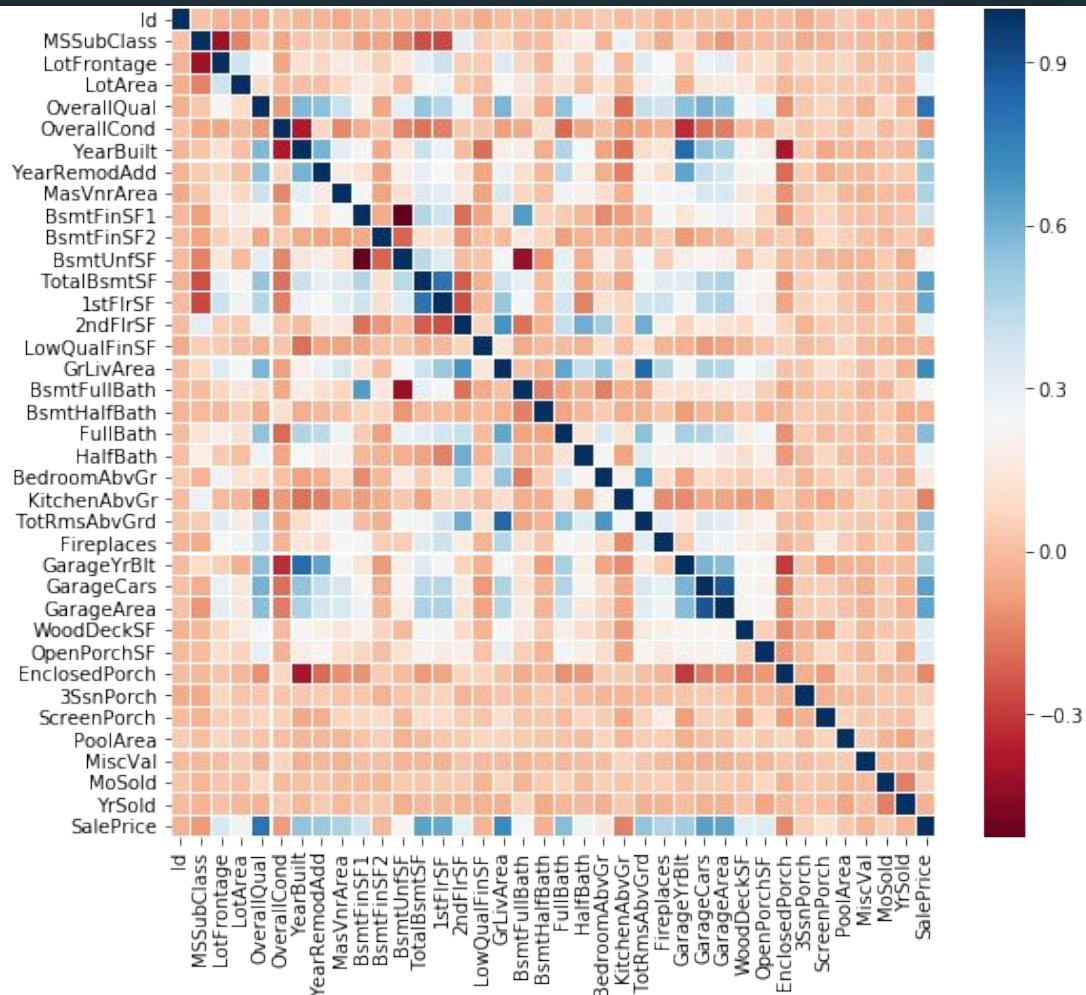
The cheapest house sold for \$34,900 and the most expensive for \$755,000
The average sales price is \$180,921, while median is \$163,000

count	1460.000000
mean	180921.195890
std	79442.502883
min	34900.000000
25%	129975.000000
50%	163000.000000
75%	214000.000000
max	755000.000000

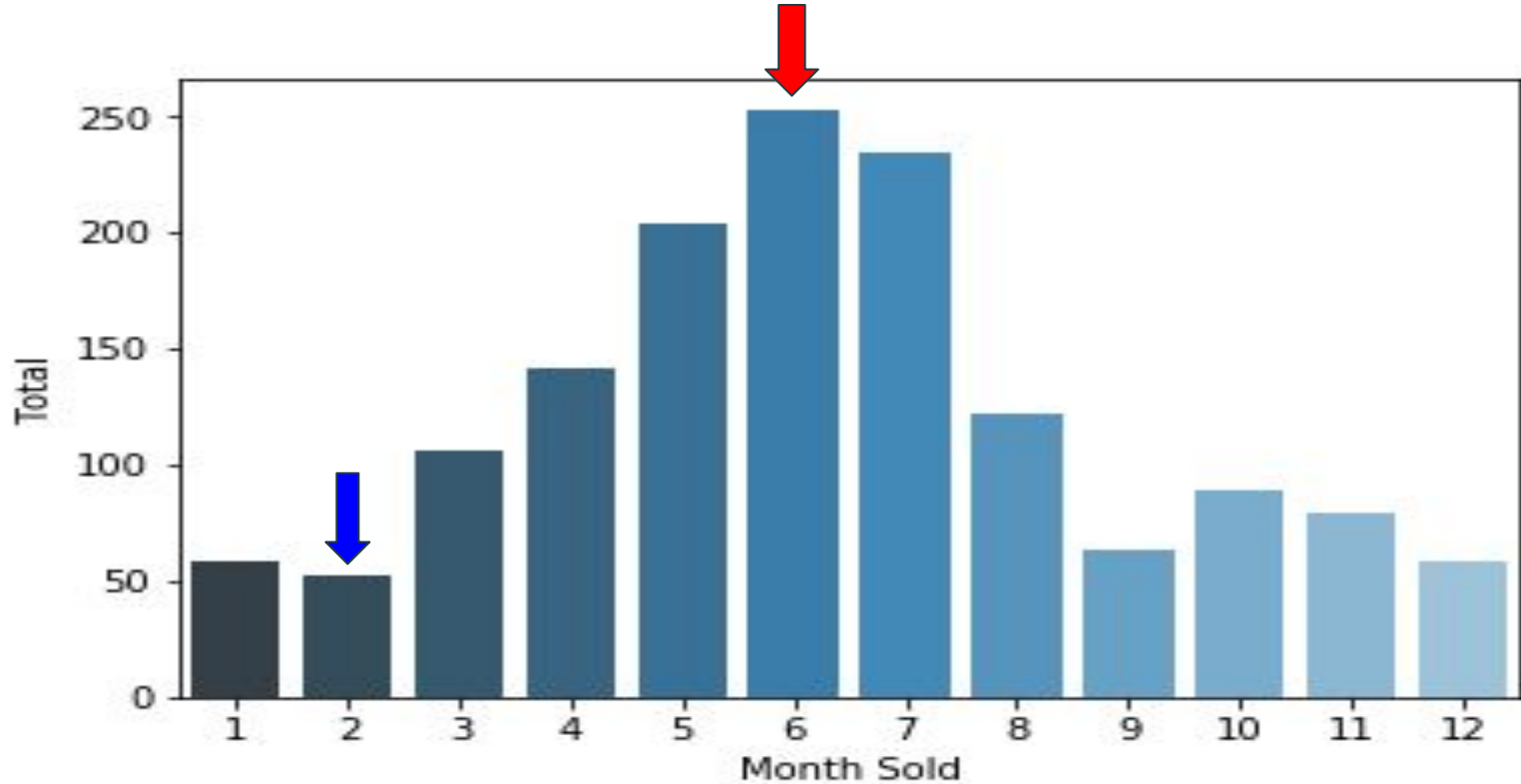


Numeric Variables Heatmap

SalePrice	OverallQual	0.800858
	TotalBsmtSF	0.646584
	1stFlrSF	0.625235
	GrLivArea	0.720516
	GarageCars	0.649256
	GarageArea	0.636964



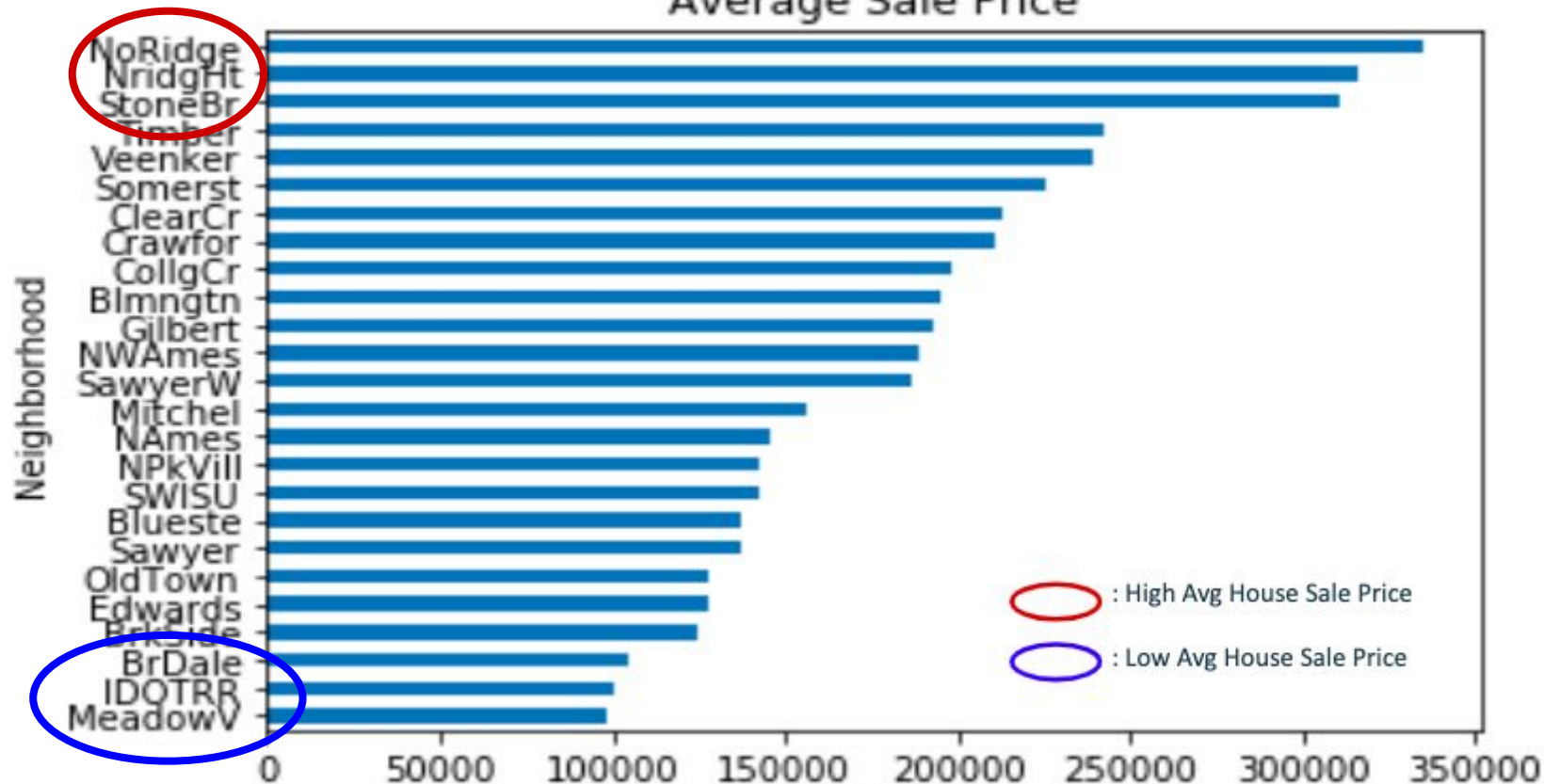
Number of House Sold by Month



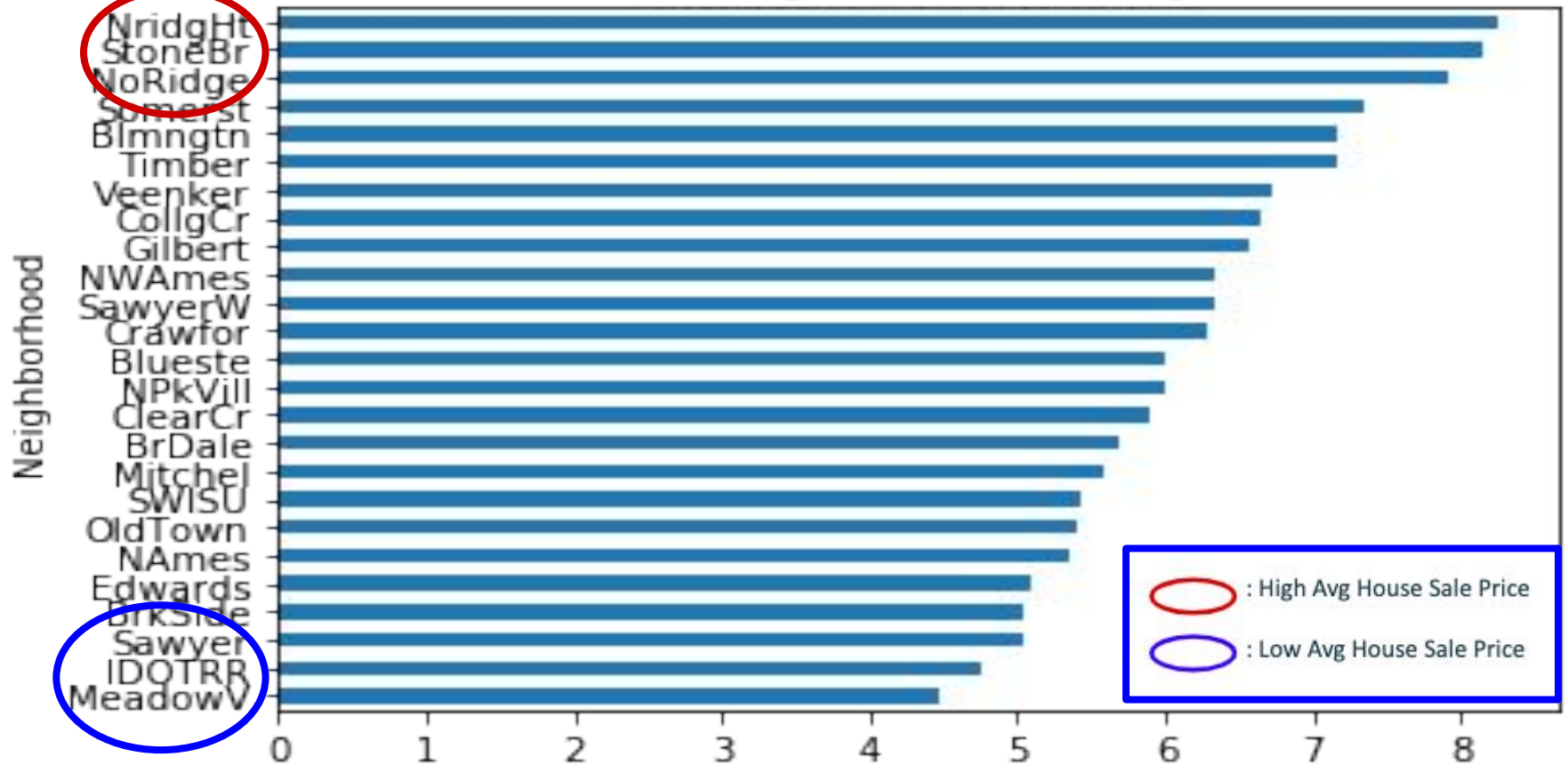
Average Sale Price by Month



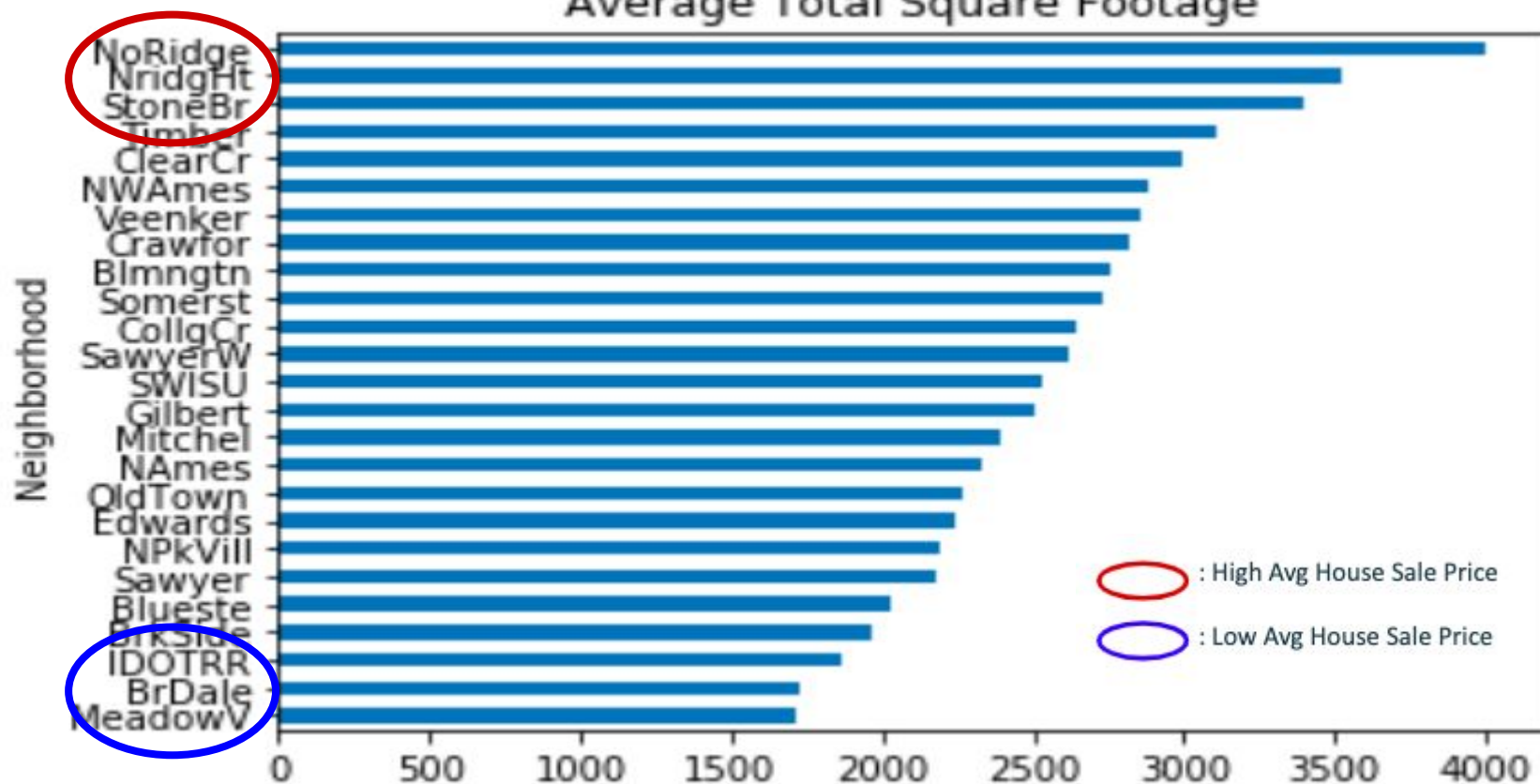
Average Sale Price



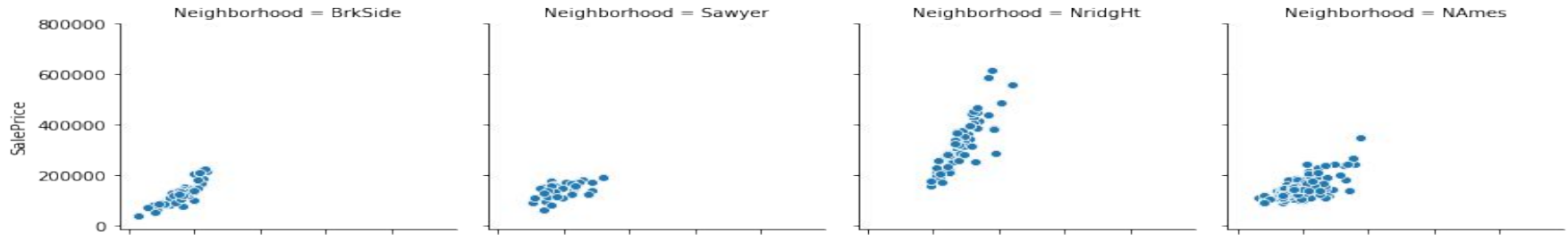
Average Overall Quality



Average Total Square Footage



Impact of Total Square Feet on Different Neighborhoods

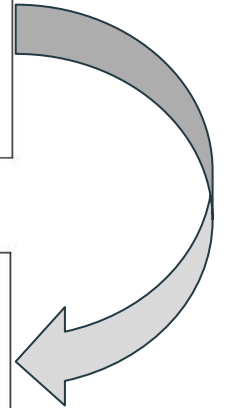
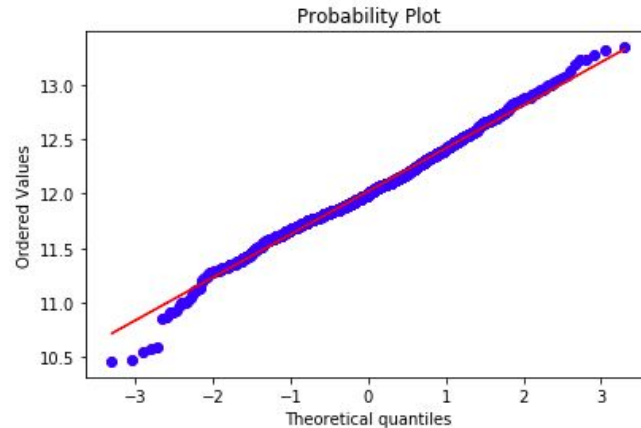
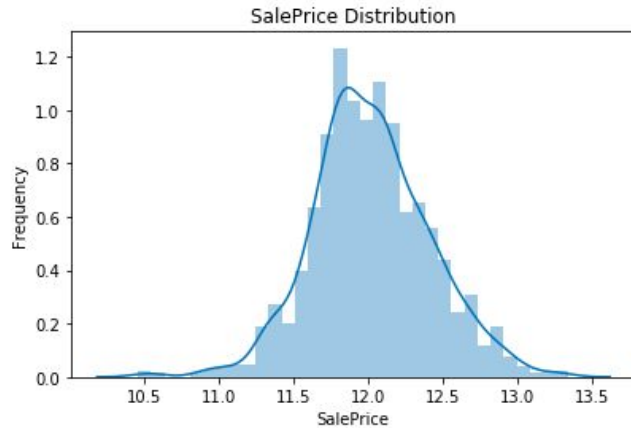
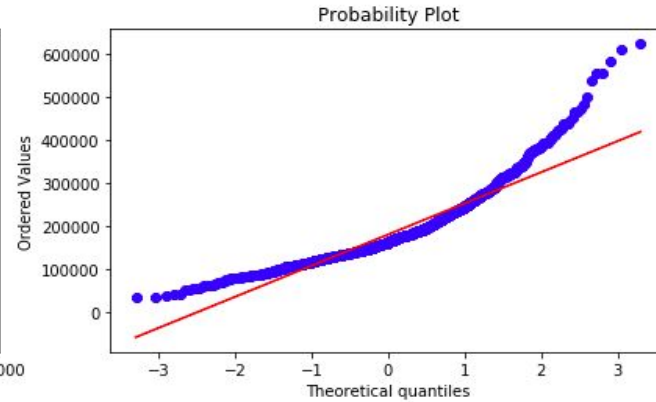
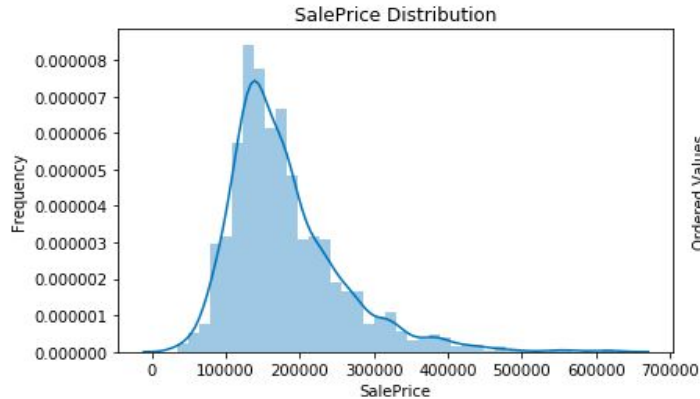


Different neighborhoods seem to have premiums/discounts for SF

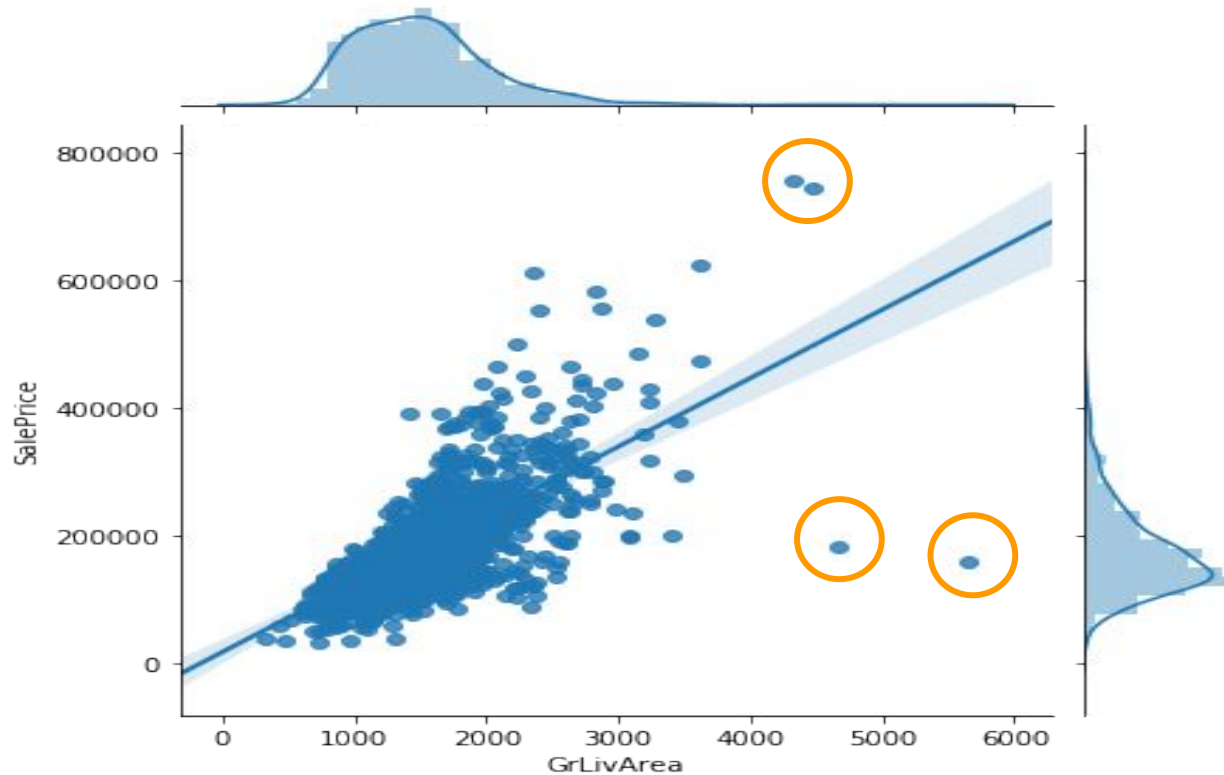
2

Cleaning the Data

Skewness in Sale Price



Data Preparation: Outlier Detection



Data Preparation: Missing Values

- PoolQC, MiscFeature, Alley, Fence
-> Removed the columns due to over 80% missingness.
- Some missing only in Test set
-> Combined Training and Test dataset to impute missing.

Training

	total	missing percent
PoolQC	1453.0	99.520548
MiscFeature	1406.0	96.301370
Alley	1369.0	93.767123
Fence	1179.0	80.753425
FireplaceQu	690.0	47.260274
LotFrontage	259.0	17.739726
GarageType	81.0	5.547945
GarageYrBlt	81.0	5.547945
GarageFinish	81.0	5.547945
GarageQual	81.0	5.547945
GarageCond	81.0	5.547945
BsmtExposure	38.0	2.602740
BsmtFinType2	38.0	2.602740
BsmtFinType1	37.0	2.534247
BsmtCond	37.0	2.534247
BsmtQual	37.0	2.534247
MasVnrArea	8.0	0.547945
MasVnrType	8.0	0.547945
Electrical	1.0	0.068493

Test

	total	missing percent
PoolQC	1456.0	99.794380
MiscFeature	1408.0	96.504455
Alley	1352.0	92.666210
Fence	1169.0	80.123372
FireplaceQu	730.0	50.034270
LotFrontage	227.0	15.558602
GarageCond	78.0	5.346127
GarageYrBlt	78.0	5.346127
GarageQual	78.0	5.346127
GarageFinish	78.0	5.346127
GarageType	76.0	5.209047
BsmtCond	45.0	3.084304
BsmtExposure	44.0	3.015764
BsmtQual	44.0	3.015764
BsmtFinType1	42.0	2.878684
BsmtFinType2	42.0	2.878684
MasVnrType	16.0	1.096642
MasVnrArea	15.0	1.028101
MSZoning	4.0	0.274160

Data Preparation: Imputation

- Fill with “0”
 - ScreenPorch -> If there is missing value, it might mean there is no screen porch area
- Fill with “No”
 - BsmtQual -> If there is missing value, it might means that there is no basement
- Fill with Mode
 - LotShape -> If there is missing value, most likely to follow the mode or “regular” lot shape

Data Preparation:

Skewed Numerical Variables -> Box Cox

- CONTINUOUS numerical variable skewness > 0.75
- LotFrontage, LotArea, MasVnrArea, BsmtFinSF1, BsmtUnfSF, LowQualFinSF, GrLivArea, GarageArea, PoolArea, MiscVal, TotalSF, TotalPorchSF

Dummy Variables

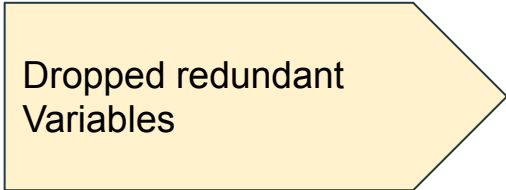
- Purpose: to run some machine learning algorithm (like linear regression) that can't handle label data

3

Feature Engineering

Added New Variables

- Total Square Feet of the House
- Total Number of Bathrooms
- Total Square Feet of the Porch
- Ratio of Above Ground Square Feet to Total Square Feet
- Percentage of Finished Basement over Total Basement
- Years between Year Remodeled and Year Built
- Years between Year Sold and Year Remodeled
- Neighborhood dummy variable * Total Square Foot
- Neighborhood dummy variable * Overall Quality



Dropped redundant
Variables

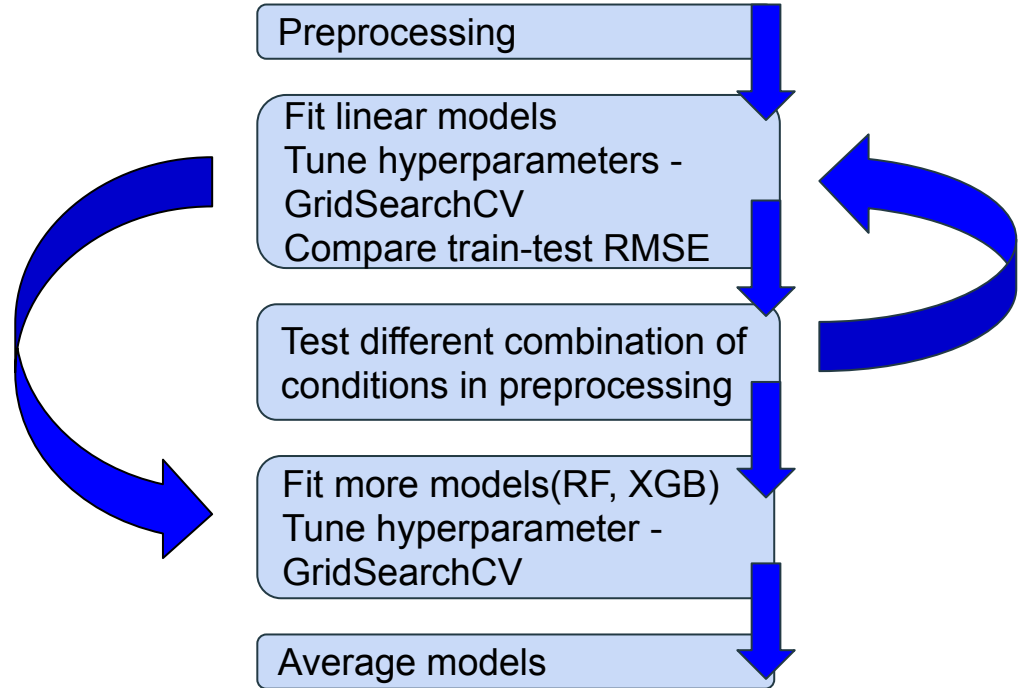
[Data Description](#)

4

Machine Learning Models

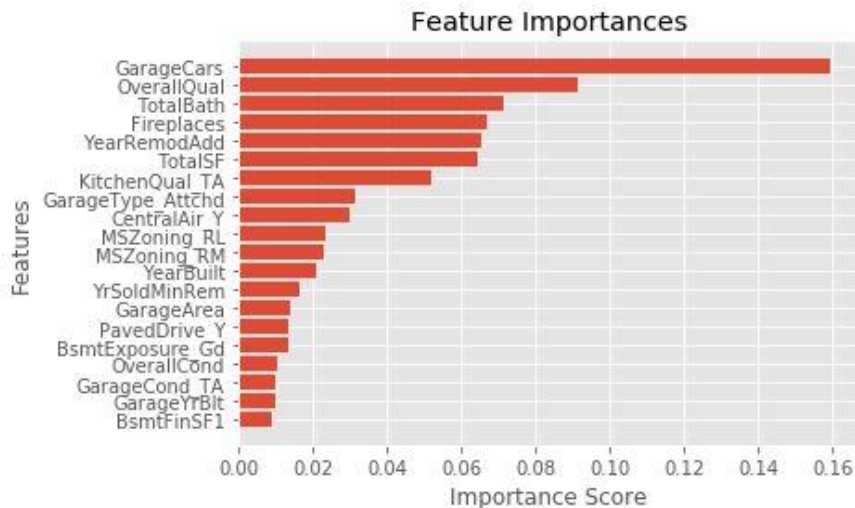
Machine Learning Algorithms

Model Type	Step 1	Step 2
Linear	OLS Lasso Elastic Net	Stacked
Nonlinear	Random Forest XGBoost	

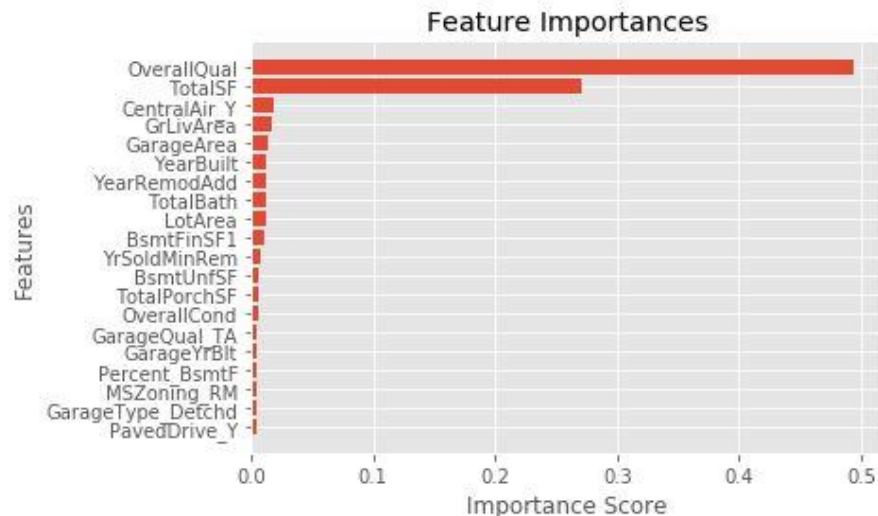


Feature Importance

XG Boost



Random Forest



Common Features across models:

: Overall Quality, Total SF, Year Built, Years since remodeled, Year Remodel, Percent of Basement Finished, SF of finished Basement

Conclusion

- Lasso produced the best result on Kaggle

- > More advanced models didn't outperform the linear regression models.

- Importance of Feature Engineering

Further Improvement

- Non-linear models without imputation and dummy variables.

- Seasonality -> time series.

- Combine similar Neighborhood and do imputation based on Neighborhood.

	Kaggle
OLS	0.12215
Lasso	0.12104
ElNet	0.12611
RF	0.1441
XGB	0.12787
Stacked	0.124

Thank you.

