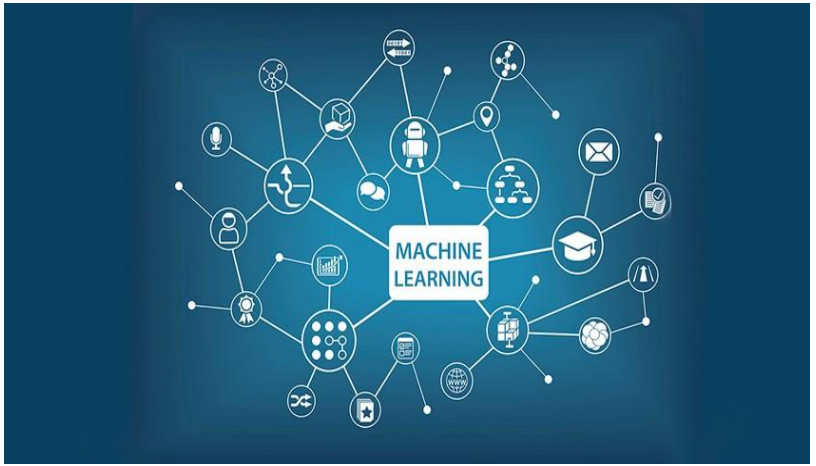
Ames Iowa Housing Prediction

Jason Washam

Define Problem/Purpose

Analyze the Ames Iowa housing data and apply machine learning concept to build a model that can predict the sale price of the houses.

Method



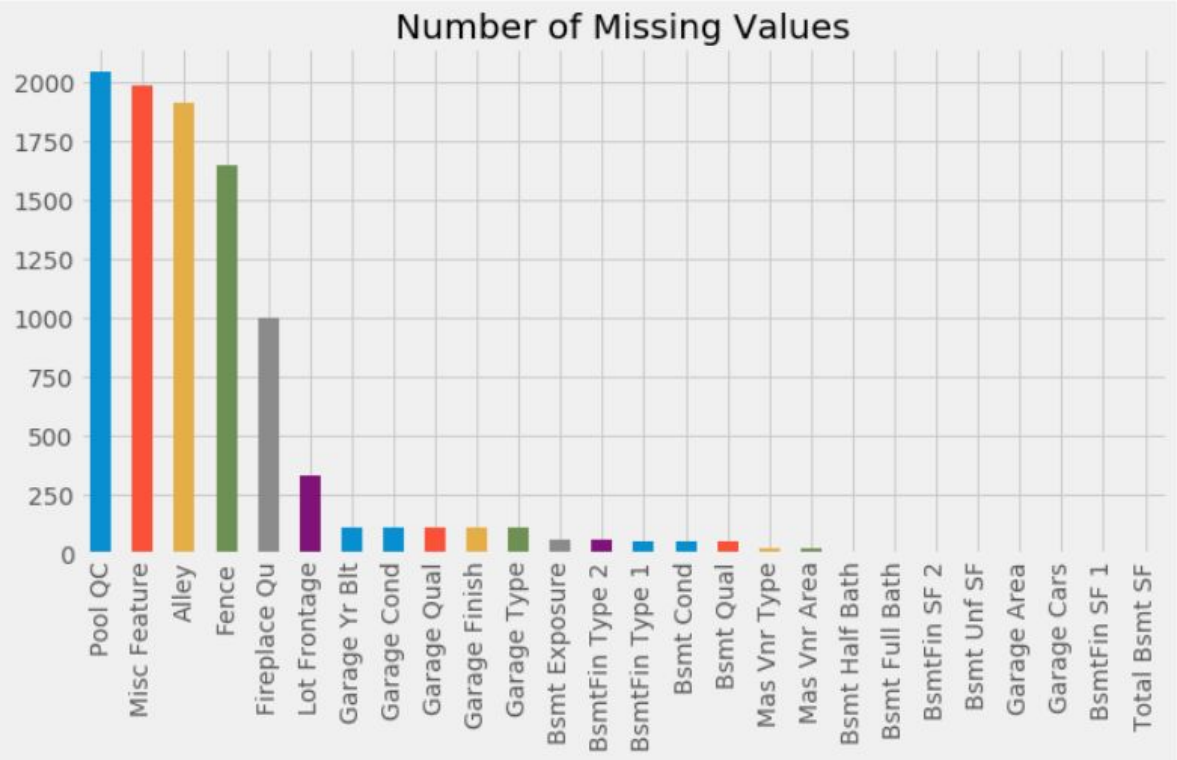
- Import and check the data
- Identify data types and missing values
- Handle missing values and clean the data
- Create charts and graphs to identify trends
- Analyze the relationship between sale price and predictors
- Evaluate the predictors and build a model
- Check the performance of the model
- Interpret the model

Ames Iowa Housing Data

- **2050 rows and 81 columns**
- **39 numeric columns**
  - Year Built
  - 1st Floor area (square feet)
  - Garage area (square feet)
  - Pool area (square feet)
- **42 categorical columns**
  - Neighborhood
  - House style
  - Garage quality
  - Central air conditioning

Fireplace Qu	Garage Type	Garage Yr Blt	Garage Finish	Garage Cars	Garage Area	Garage Qual	Garage Cond	Paved Drive	Wood Deck SF	Open Porch SF
NaN	Attchd	1976.0	RFn	2.0	475.0	TA	TA	Y	0	44
TA	Attchd	1997.0	RFn	2.0	559.0	TA	TA	Y	0	74
NaN	Detchd	1953.0	Unf	1.0	246.0	TA	TA	Y	0	52
NaN	BuiltIn	2007.0	Fin	2.0	400.0	TA	TA	Y	100	0
NaN	Detchd	1957.0	Unf	2.0	484.0	TA	TA	N	0	59
Gd	Attchd	1966.0	Fin	2.0	578.0	TA	TA	Y	0	0
NaN	Basment	2005.0	Fin	2.0	525.0	TA	TA	Y	0	44
NaN	Attchd	1959.0	RFn	2.0	531.0	TA	TA	Y	0	0
NaN	Detchd	1952.0	Unf	1.0	420.0	TA	TA	Y	0	324
TA	Attchd	1969.0	Unf	2.0	504.0	TA	TA	Y	335	0

Missing Data



- 9822 total missing values
- About 6% of the data

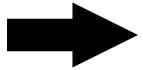
Handling Missing Data

"If you cannot fill the missing values, just drop them"

Categorical Columns

- House with no pool = Pool quality do not apply (NA)

Pool Area	Pool QC
0	NaN
0	NaN
0	NaN
0	NaN

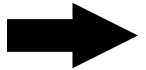


Pool Area	Pool QC
0	NA
0	NA
0	NA
0	NA

Numeric Columns

- House with no garage = No garage year built (0)

Garage Area	Garage Yr Blt
240.0	1937.0
0.0	NaN
576.0	2003.0
542.0	1981.0



Garage Yr Blt
1937.0
0.0
2003.0
1981.0

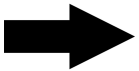
Data Cleaning

- **Ordinal data (19 total)**
  - Ex: Excellent = 5
  - Gd: Good = 4
  - TA: Typical = 3
  - Fa: Fair = 2
  - Po: Poor = 1
  - NA: None = 0

- **Binary data (2 total)**
  - N: No = 0
  - Y: Yes = 0

BsmtCond: General condition of the basement

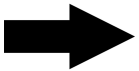
- Ex Excellent
- Gd Good
- TA Typical - slight dampness allowed
- Fa Fair - dampness or some cracking or settling
- Po Poor - Severe cracking, settling, or wetness
- NA No Basement



Bsmt Cond
3
3
3
3
4

Fence: Fence quality

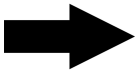
- GdPrv Good Privacy
- MnPrv Minimum Privacy
- GdWo Good Wood
- MnWw Minimum Wood/Wire
- NA No Fence



Fence
0
0
3
3

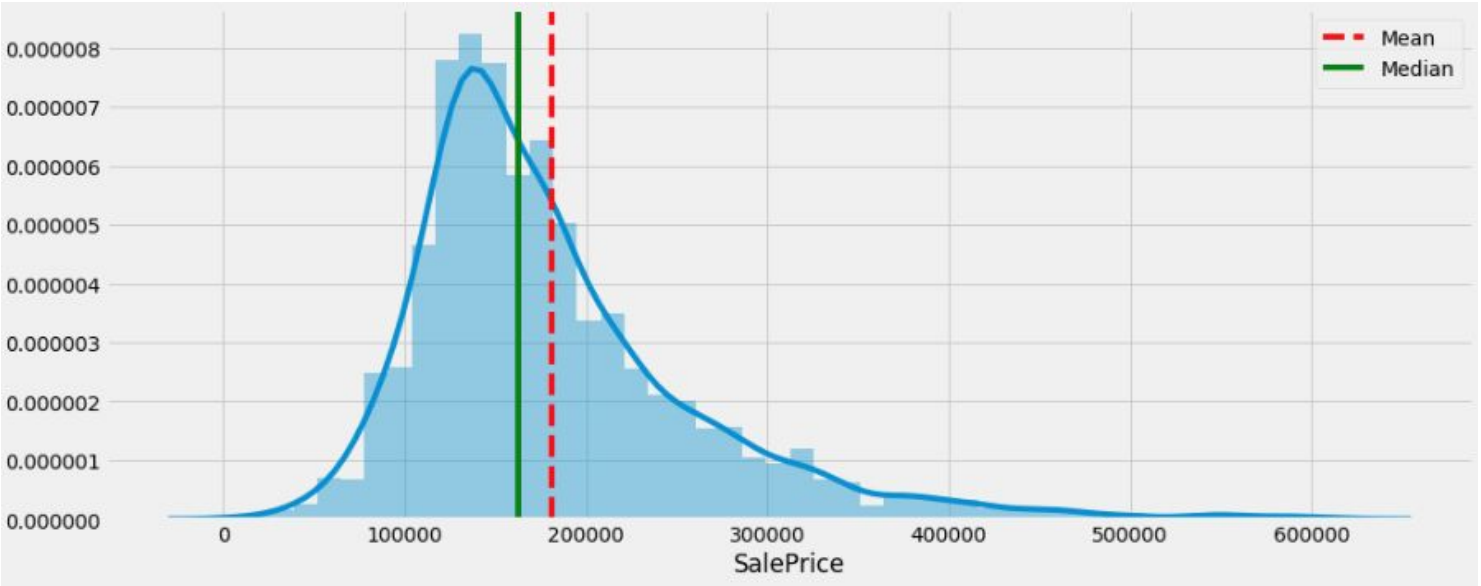
CentralAir: Central air conditioning

- N No
- Y Yes



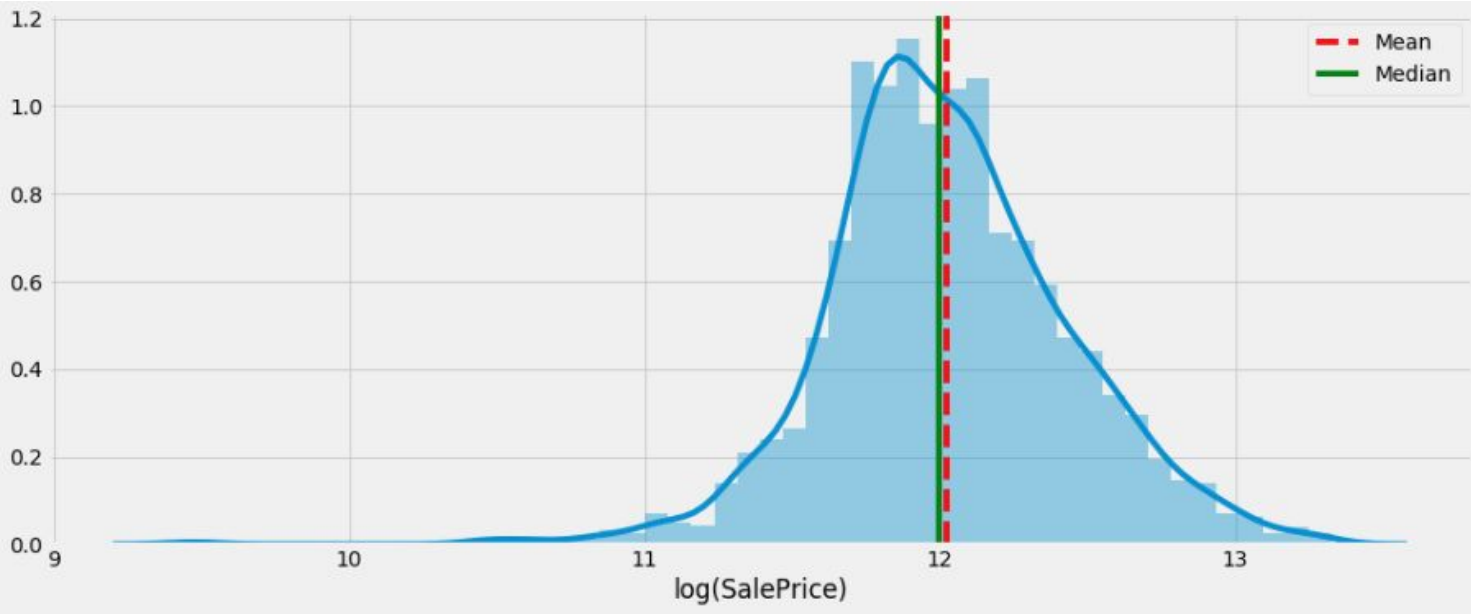
Central Air
1
0
1

Distribution of Sale Price



- Average (mean) of sale price > Midpoint (median) of sale price
  - Increase in error in model
  - Bad prediction

Log Transformation

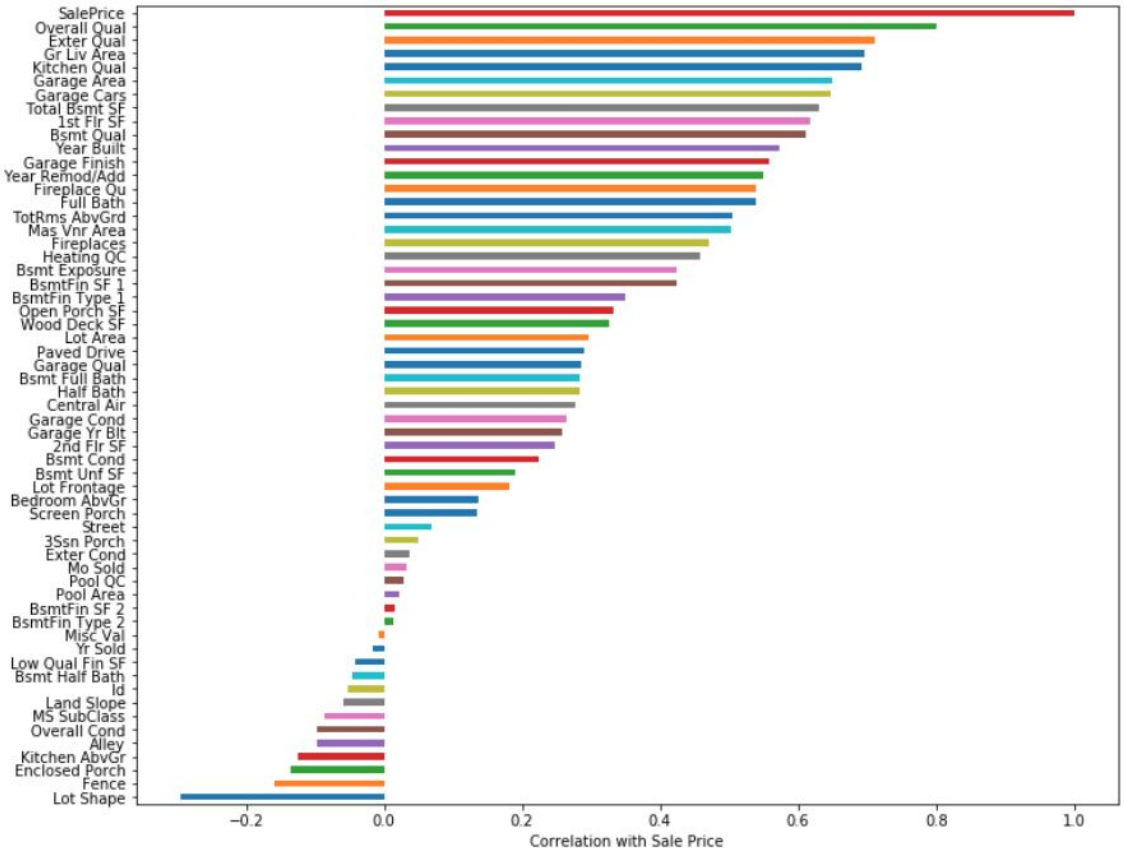


- Convert sale price to log(sale price)
  - Average (mean) of sale price ≈ Midpoint (median) of sale price
  - Better prediction

Correlation with Sale Price

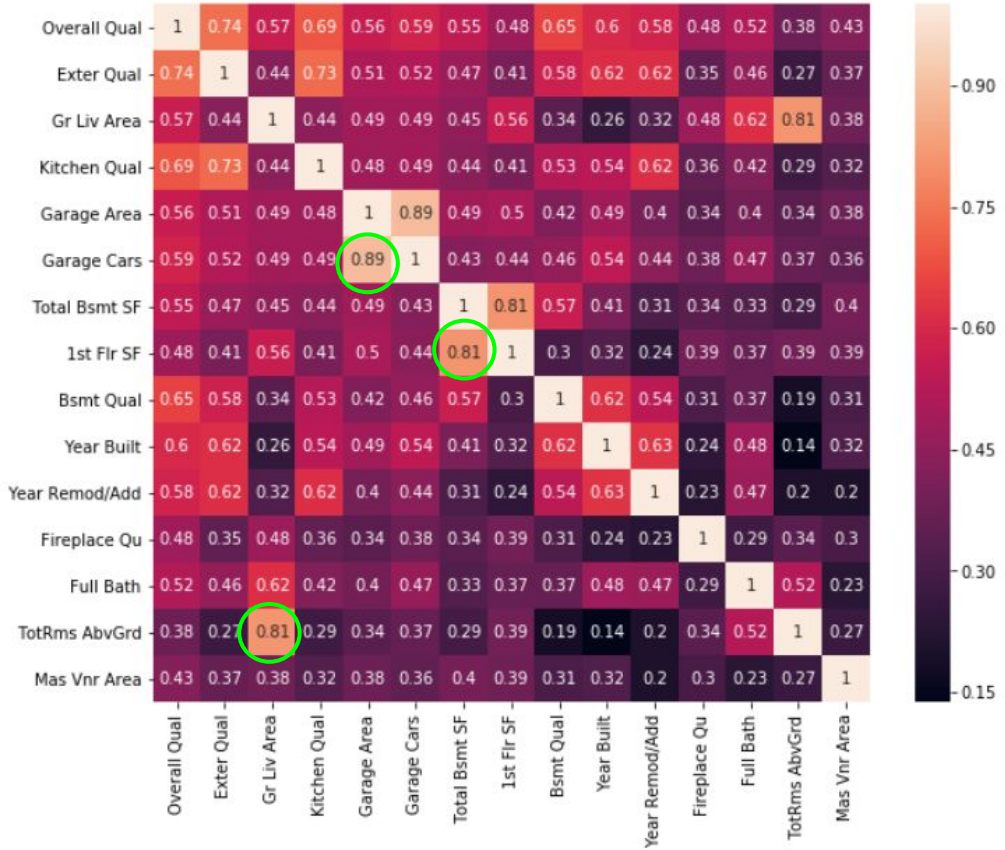
- **Positive correlation**
  - Overall Quality
  - External Quality
  - Living Area

- **Negative correlation**
  - Lot Shape
  - Fence
  - Enclosed Porch



Correlation

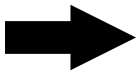
- **High correlation between predictors (Multicollinearity)**
  - Total Bsmt SF and 1st Flr SF
  - TotRms AbvGrd and Gr Liv Area
  - Fireplace and Fireplace Qu
  - Garage Yr Blt and Garage Qual
  - Garage Yr Blt and Garage Cond
  - Garage Area and Garage Cars
  - Garage Qual and Garage Cond
  - Pool Area and Pool QC
- **Removed**
  - Fireplace Qu
  - Garage Cond
  - Garage Cars
  - Pool QC
  - Garage Qual



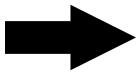
Feature Engineering

- **High correlation between predictors (Multicollinearity)**
  - Total Bsmt SF and 1st Flr SF
  - TotRms AbvGrd and Gr Liv Area
  - Total Bsmt SF and Gr Liv Area also have high correlation with sale price
- **Make new features instead of removing**
  - Total Area * Overall Qu = (Total Bsmt SF + 1st Flr SF + 2nd Flr SF) * Overall Qu
  - Gr Liv Area * Overall Qu

Total Bsmt SF	1st Flr SF	2nd Flr SF
725.0	725	754
913.0	913	1209
1057.0	1057	0

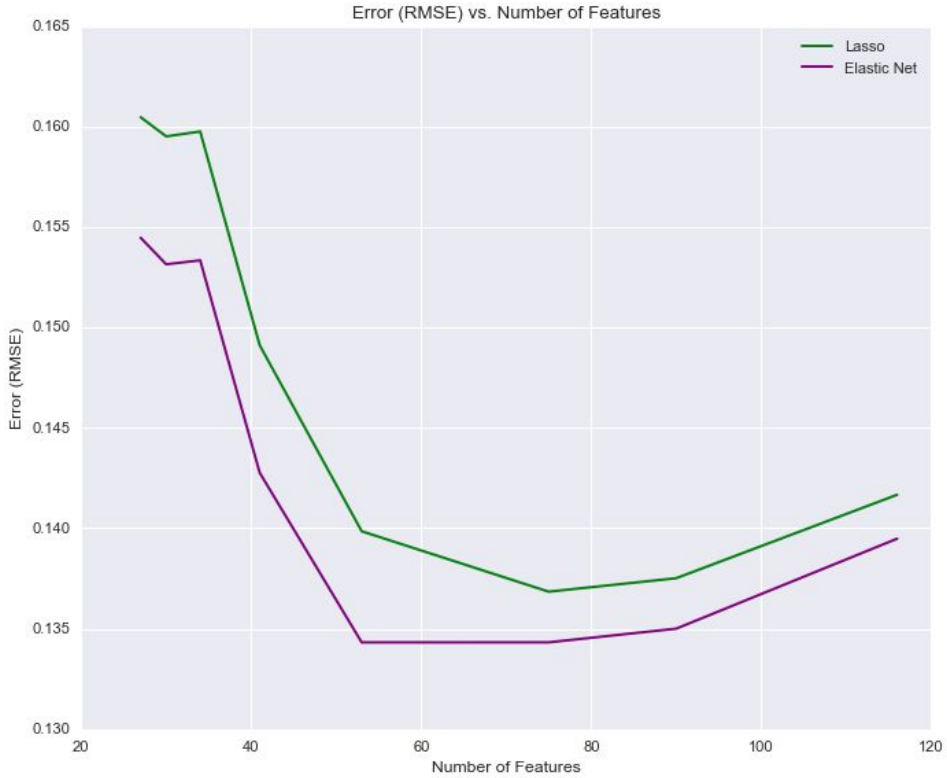


Total Area
2204.0
3035.0
2114.0



Total Area * Overall
13224.0
21245.0
10570.0

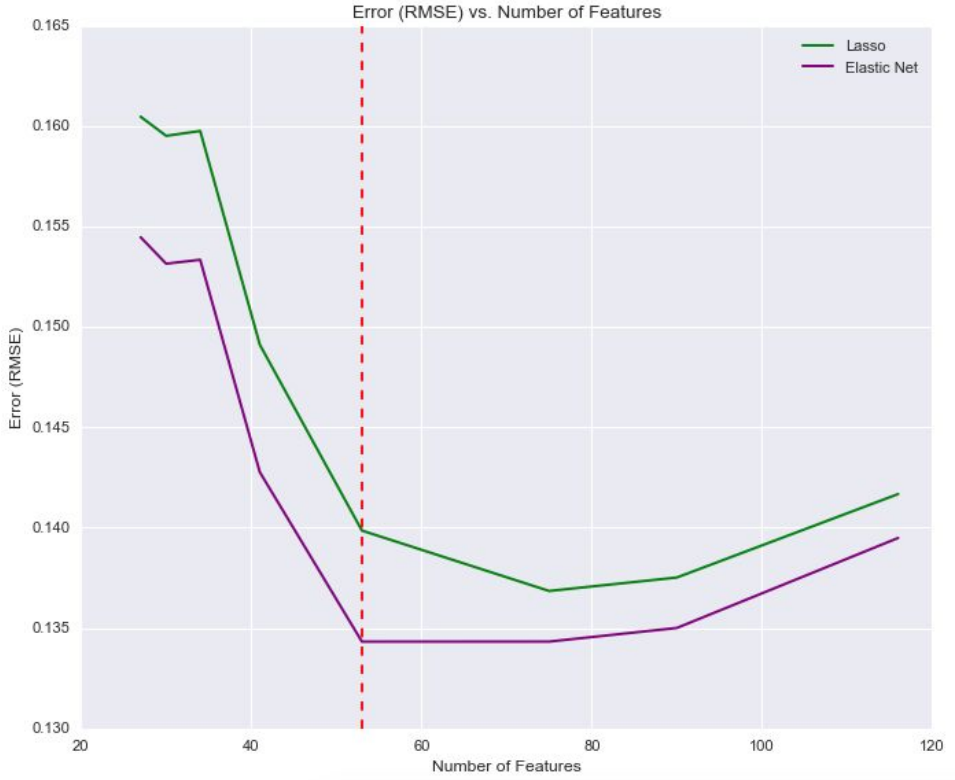
Feature Selection and Modeling



- **Selecting features**
  - Lasso method
  - Fit all features
  - Unnecessary features will have value (coefficient) of zero
  - Set different cutoffs for the value

- **Modeling**
  - Lasso and Elastic Net
  - Fit the model with selected features
  - Check the error (RMSE)

Final Model



Elastic Net Method

- With 53 features
  - Living Area
  - Overall Condition
  - Building Type
  - Neighborhood
  - Sale Type

Interpreting the model

Features that help the sale price to increase

- High positive coefficient
  - Gr Liv Area: above ground living area in square feet
  - Total Area: area of basement, 1st floor and 2nd floor in square feet
  - Neighborhood Crawfor: Crawford neighborhood in Ames, Iowa (Expensive neighborhood?)
  - Exterior 1st_BrkFace: Brick Face exterior covering on house

Features that help the sale price to decrease

- High negative coefficient
  - Roof Matl_ClyTile: Clay or tile roof material
  - Functional_Sal: Home functionality rating (salvage only)
  - Functional_Sev: Home functionality rating (severely damaged)
  - Neighborhood_Edwards: Edwards neighborhood in Ames, Iowa (Inexpensive neighborhood?)

Questions