DSI Project 2

PREDICTION OF HOUSING PRICE IN AMES,IOWA

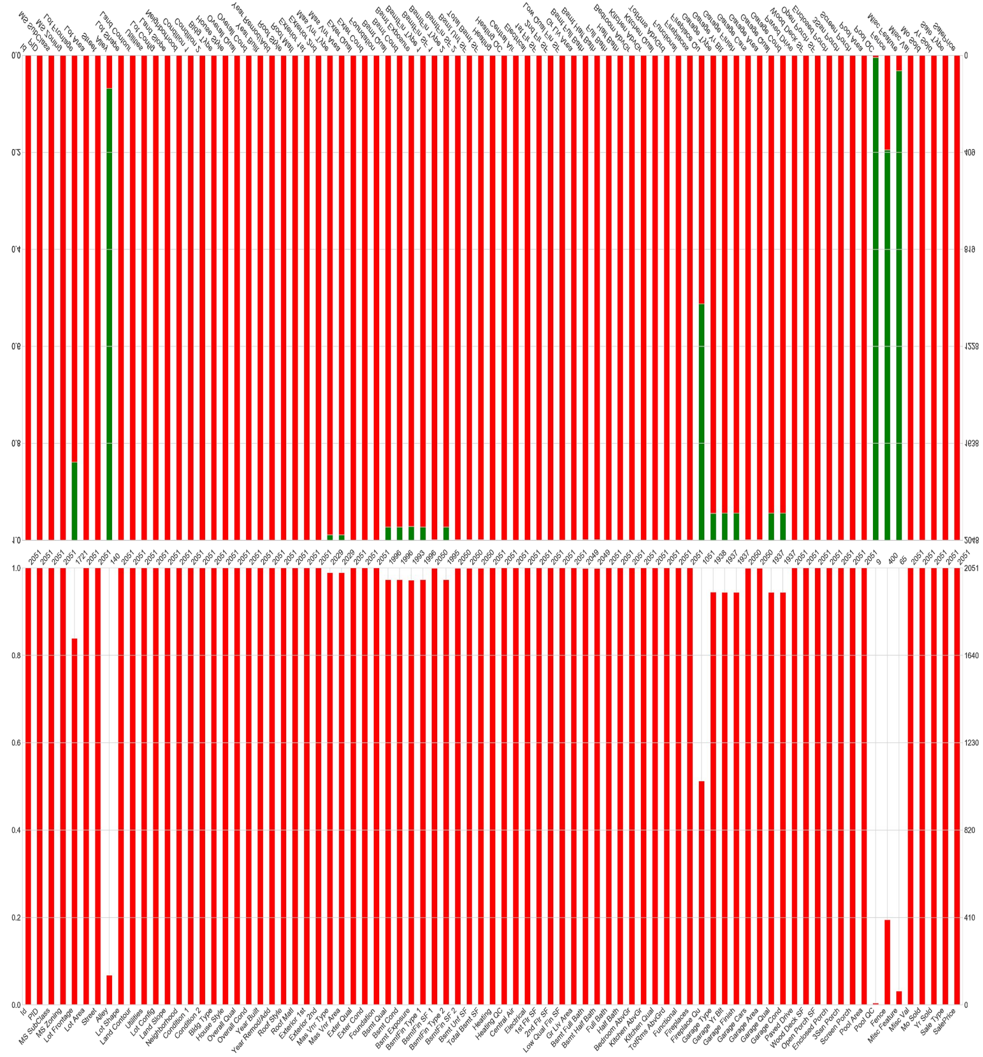
Presentation summary

1. **Problem statement**
2. **Data Cleaning & Exploratory Data Analysis (EDA)**
3. **Pre-processing**
4. **Modelling**
5. **Model Evaluation**
6. **Conclusions & Recommendations**

Problem Statement

Predicting the price of a house on sale at Ames,Iowa based on the features of the home and determine the key drivers of the home price

- Developing a regression model to study how the various features contribute to the pricing.
- Influence of the features on the pricing
- Maximise profits by building homes with appropriate features

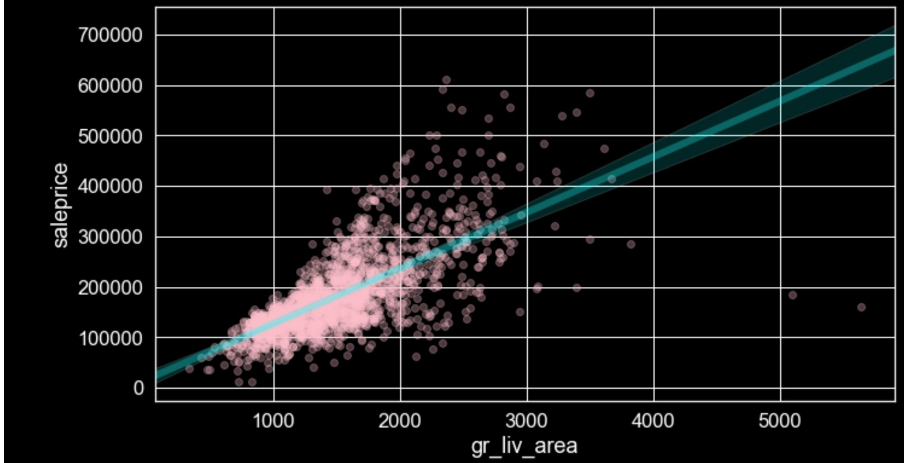


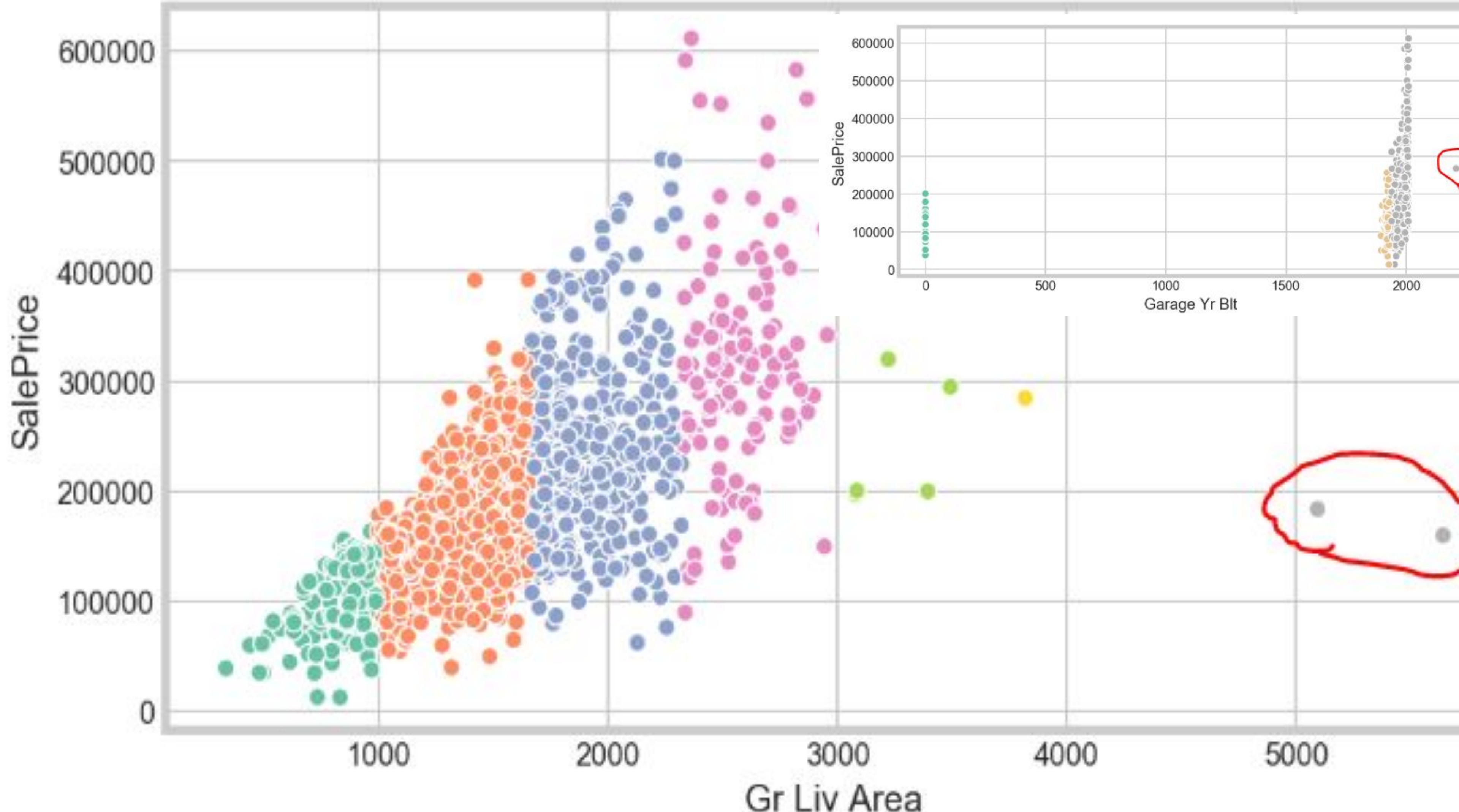
data

CLEANING

Challenging set of data

- Large Amount of Missing Data
- Imputation methodoloy
- Outliers

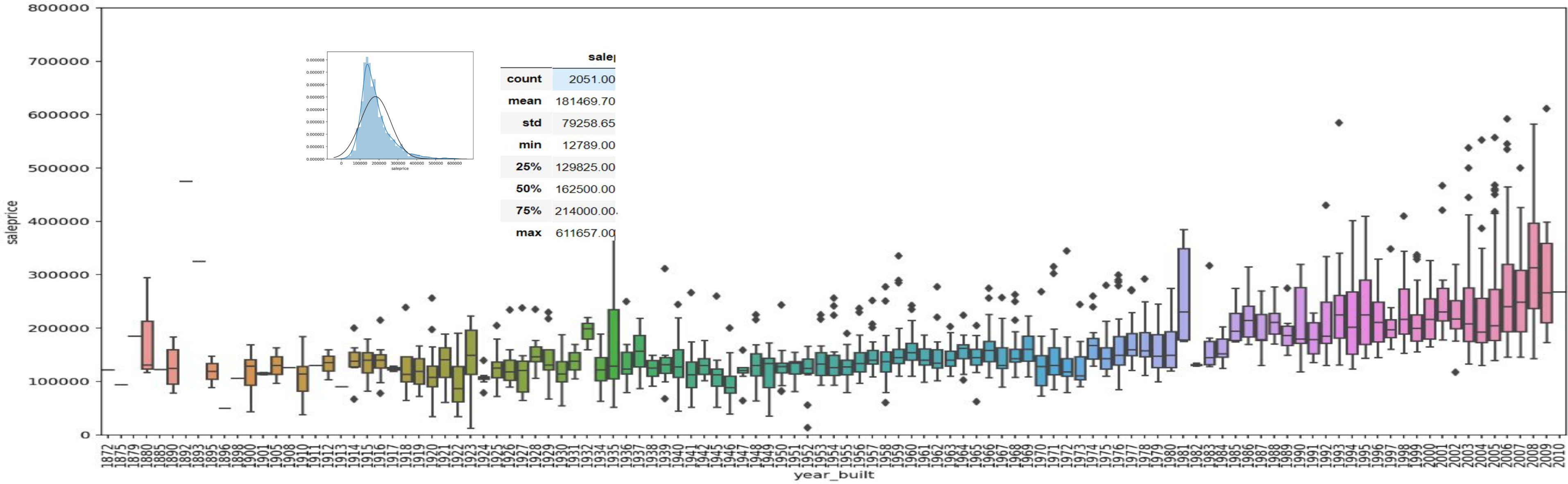




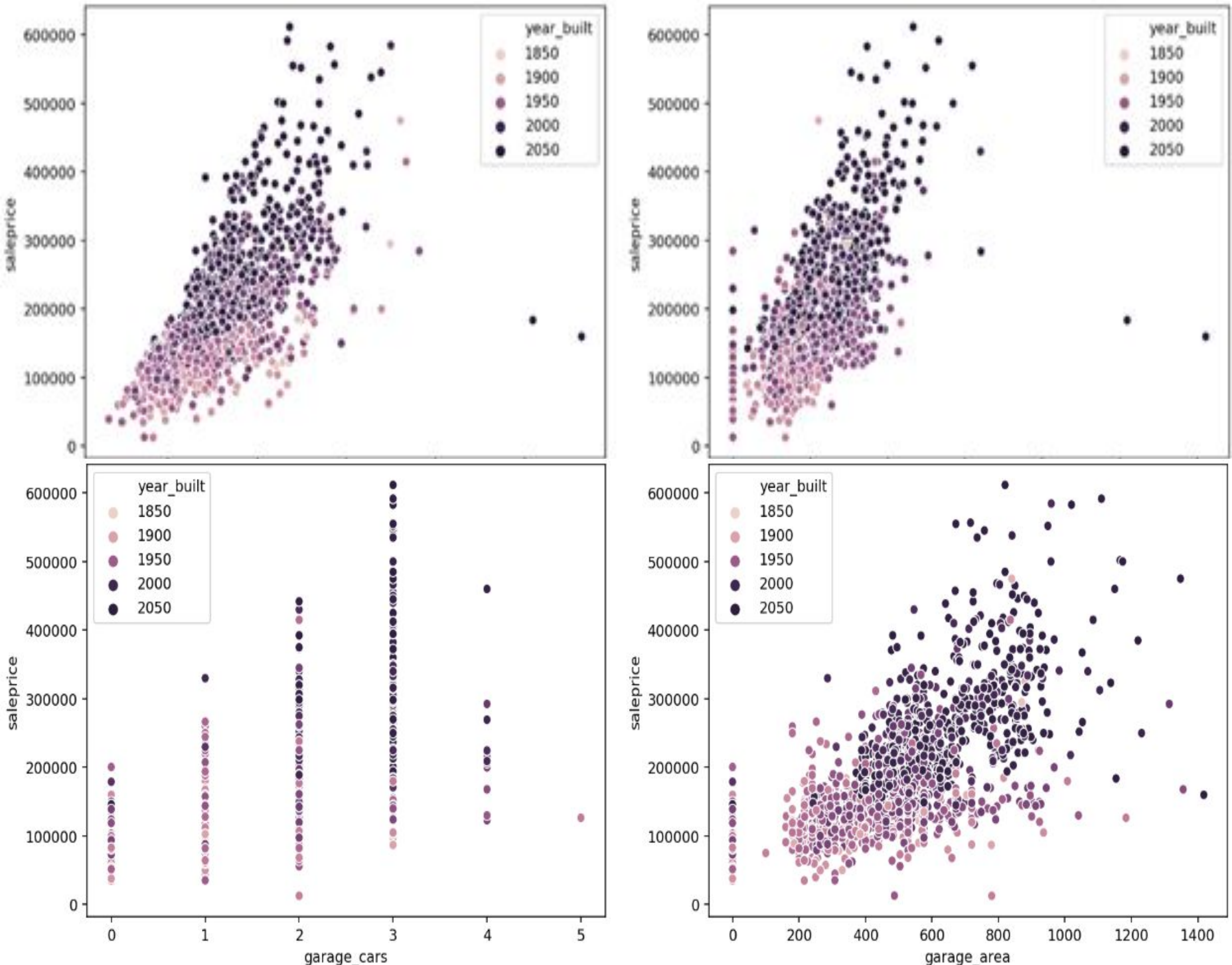
SalePrice
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
Gr Liv Area
SalePrice
600000
500000
400000
300000
200000
100000
0
0
500
1000
1500
2000
Garage Yr Blt

EDA, Data Munching & Data Engineering

- ➢ Sales price is positively skewed and kurtosis show peakedness (ie. there are outliers)
- ➢ Majority of the transactions were transacted between $130k to $250
- ➢ Sales has been increased steadily for more than century
- ➢ There are many features positively correlated to sales
- ➢ Some features exhibits unique characteristics



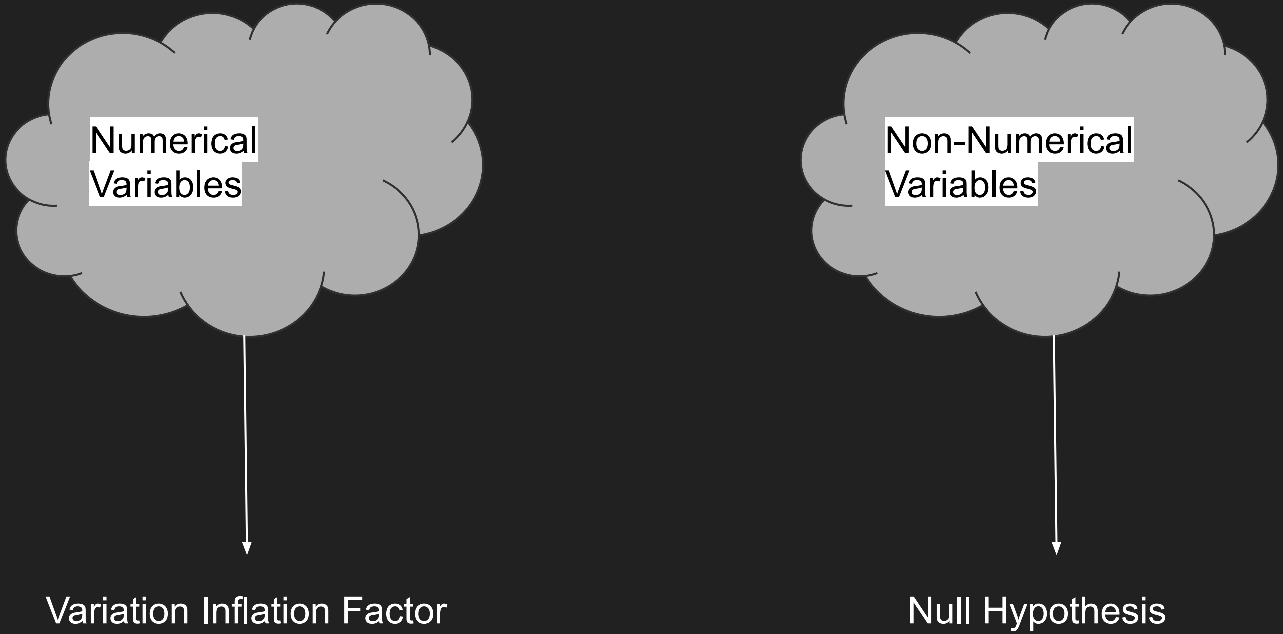
	salep
count	2051.00
mean	181469.70
std	79258.65
min	12789.00
25%	129825.00
50%	162500.00
75%	214000.00
max	611657.00



EDA, Data Munching & Data Engineering

Preprocessing

- Split training file into train/ validation sets to build an accurate model before scaling
- Apply one-hot encoding on selected categorical features
- Scale training and testing datasets excluding dummies ie. scaled numeric data only



Numerical Variables

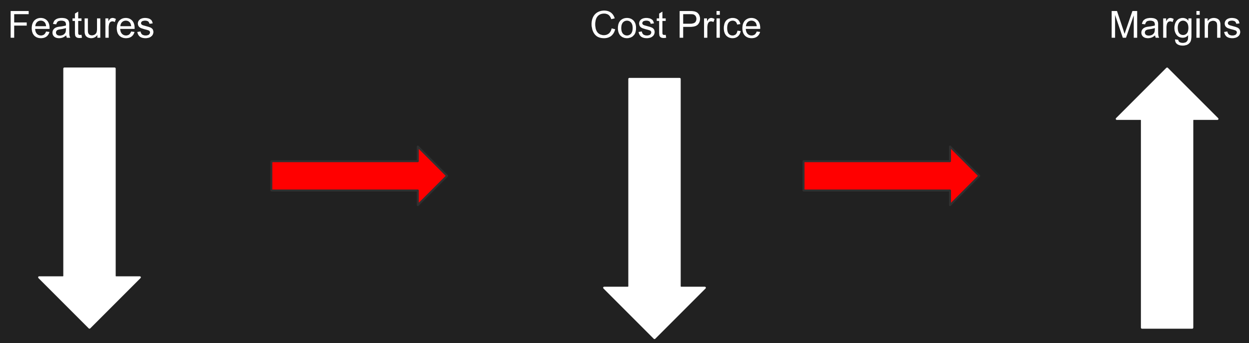
Variation Inflation Factor

Non-Numerical Variables

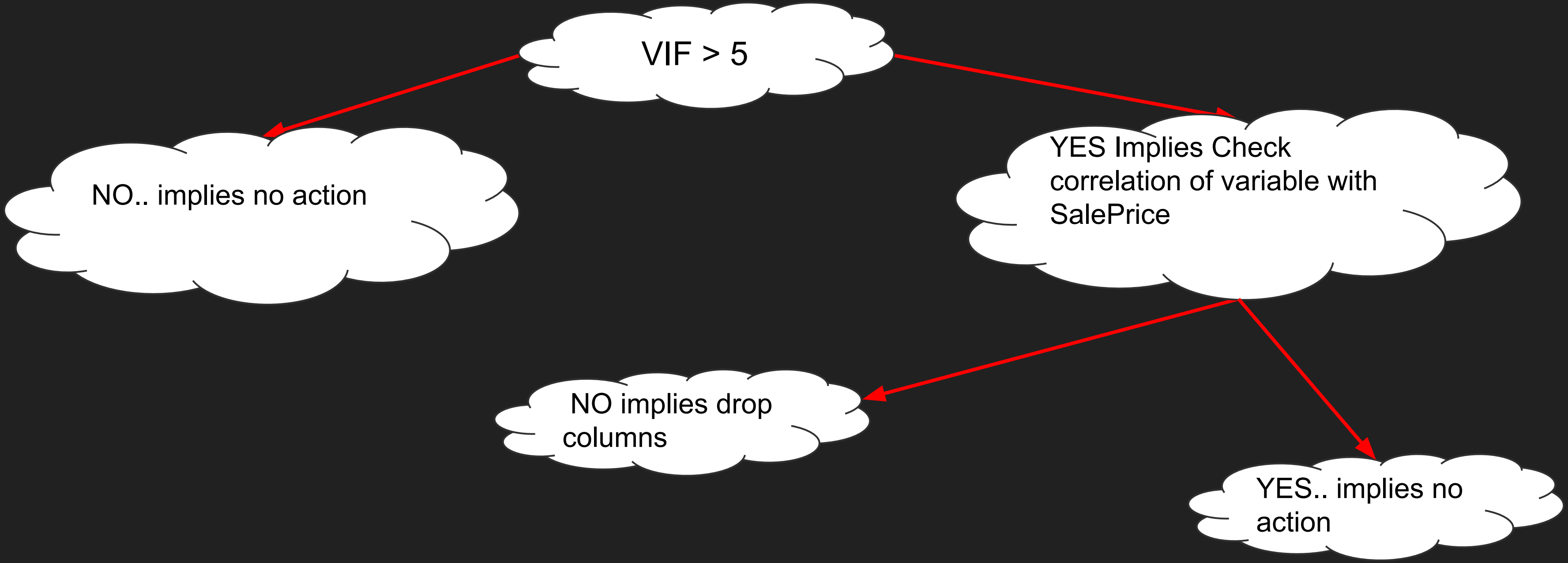
Null Hypothesis

Variation Inflation Factor

Goal: We want to drop features that are multicollinear



Variation Inflation Factor



VIF > 5

NO.. implies no action

YES Implies Check correlation of variable with SalePrice

NO implies drop columns

YES.. implies no action

Chi Square Test

Null Hypothesis:

➢ **The selected feature has high multicollinearity with other variables and has to be discarded**

Alternate Hypothesis:

➢ **The selected feature has a low multicollinearity with other variables and it can be used for for our modelling**

  ○ **If p-value < 0.05 we reject the null hypothesis and include the feature in our modelling.**

20% of the effort

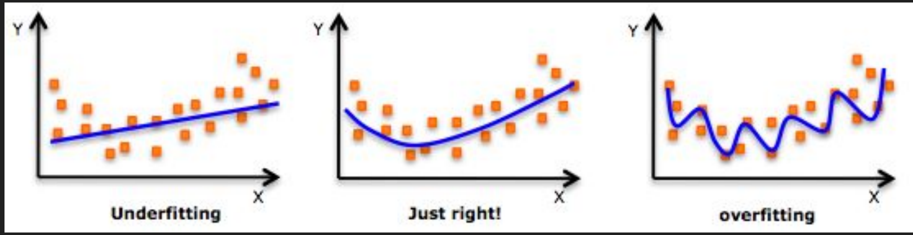
80% of the results

Feature selection

1. **Variation Inflation Factor(VIF)**
2. **Chi2 Test**
3. **Recursive Feature Elimination (RFE)**
4. **Built-in regularization from regression models**

Modelling

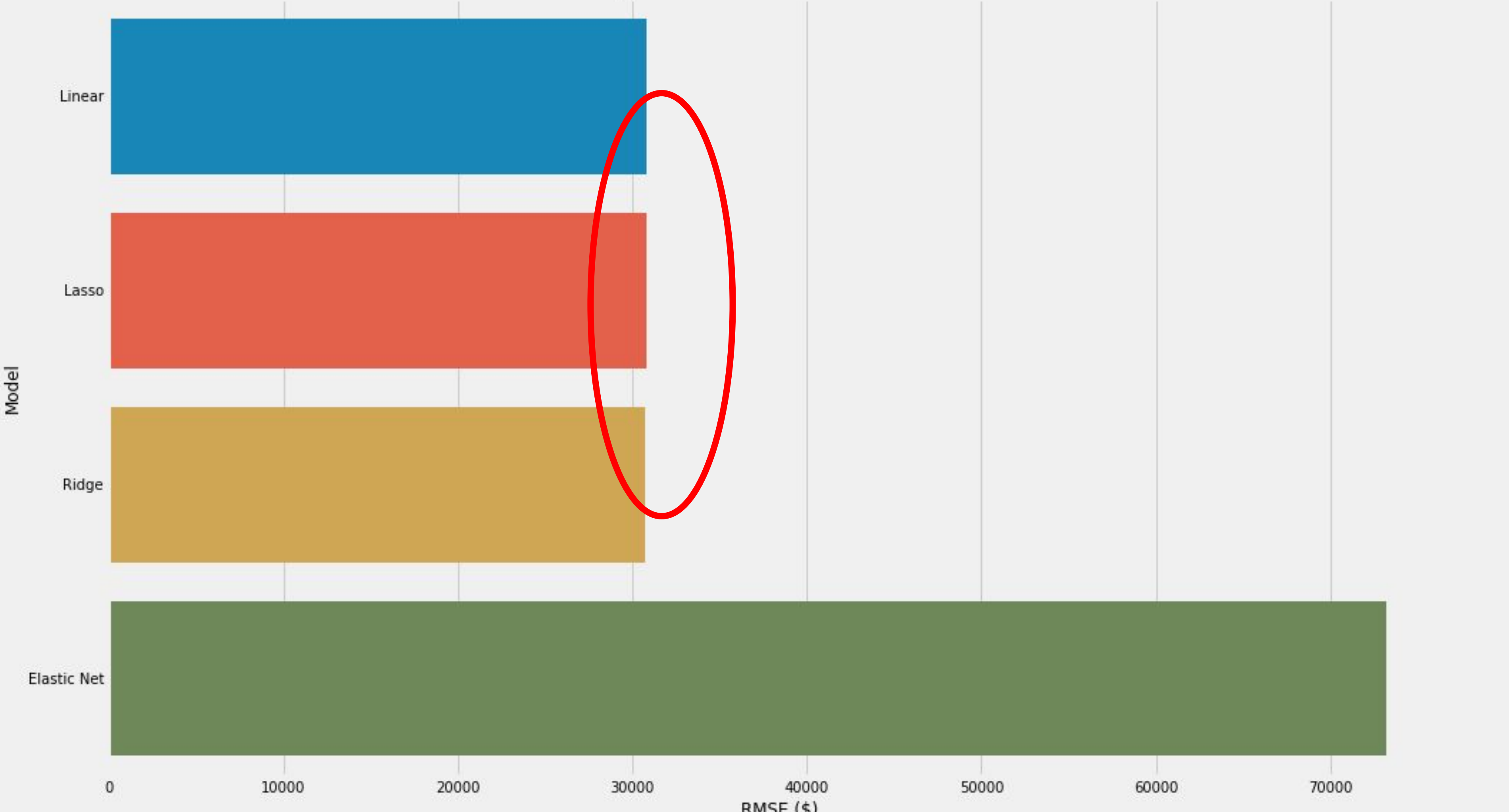
- Linear Regression
- Lasso
- Ridge
- Elastic Net

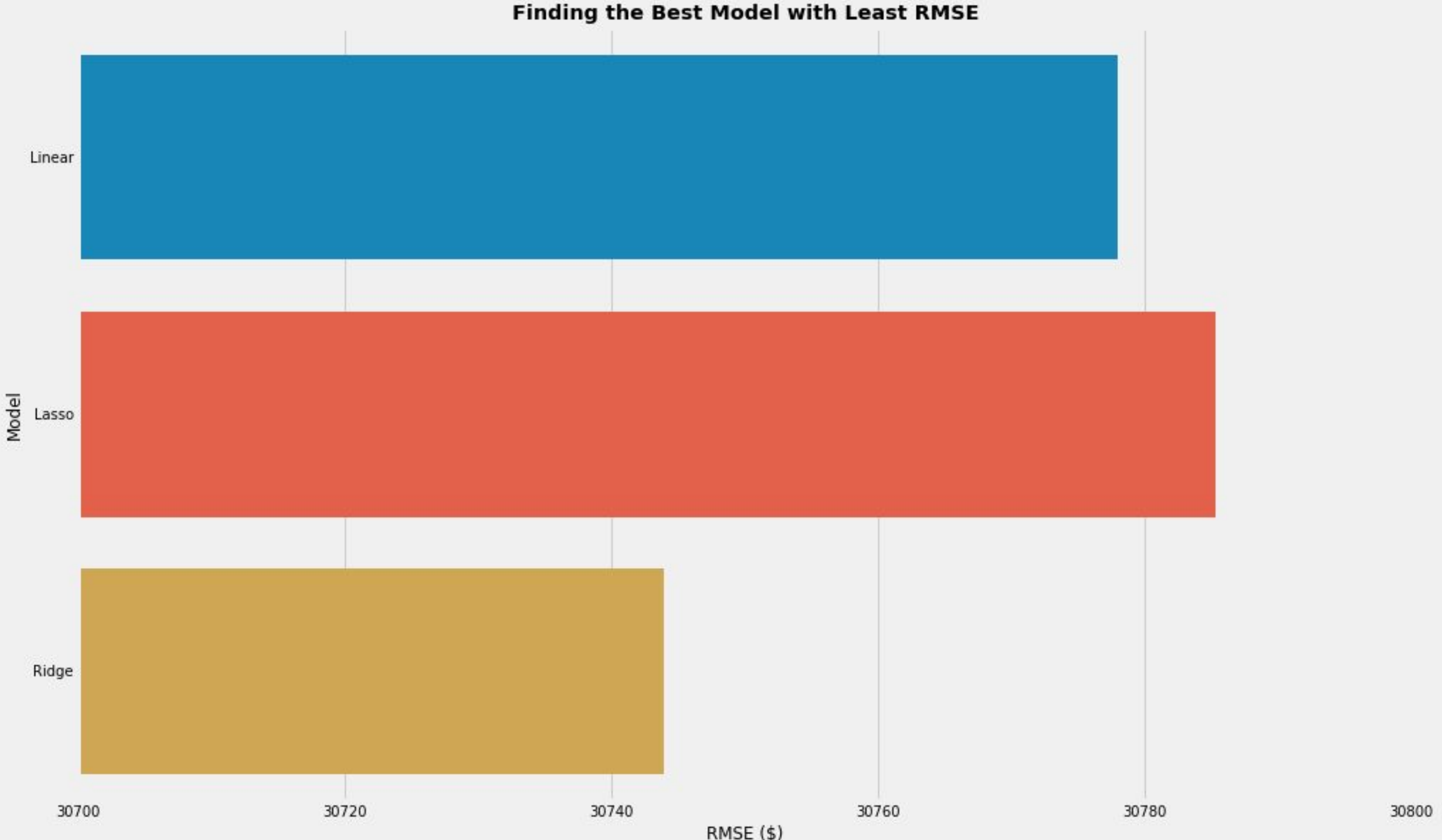


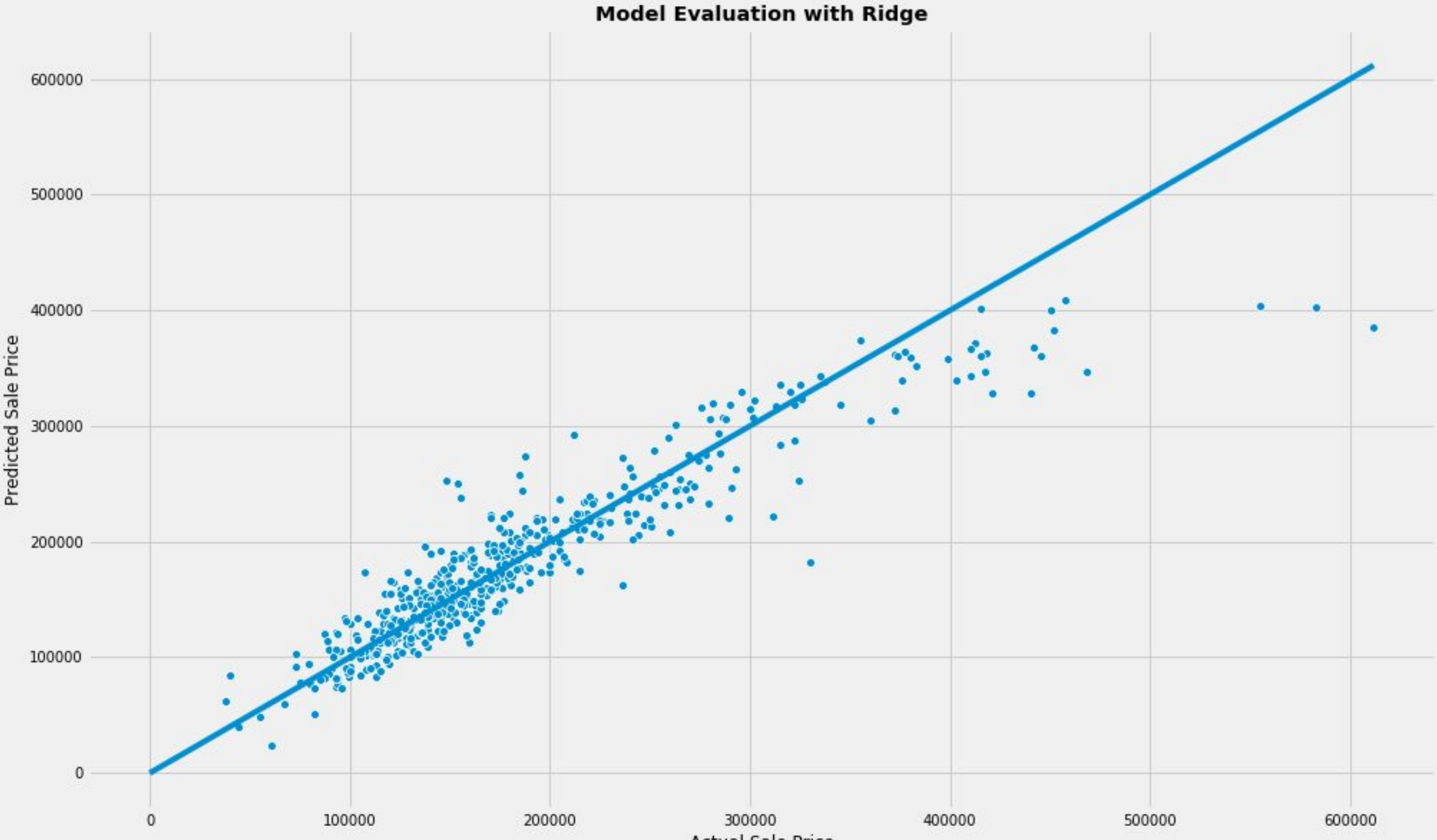
Baseline Score

$$RMSE(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\tfrac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

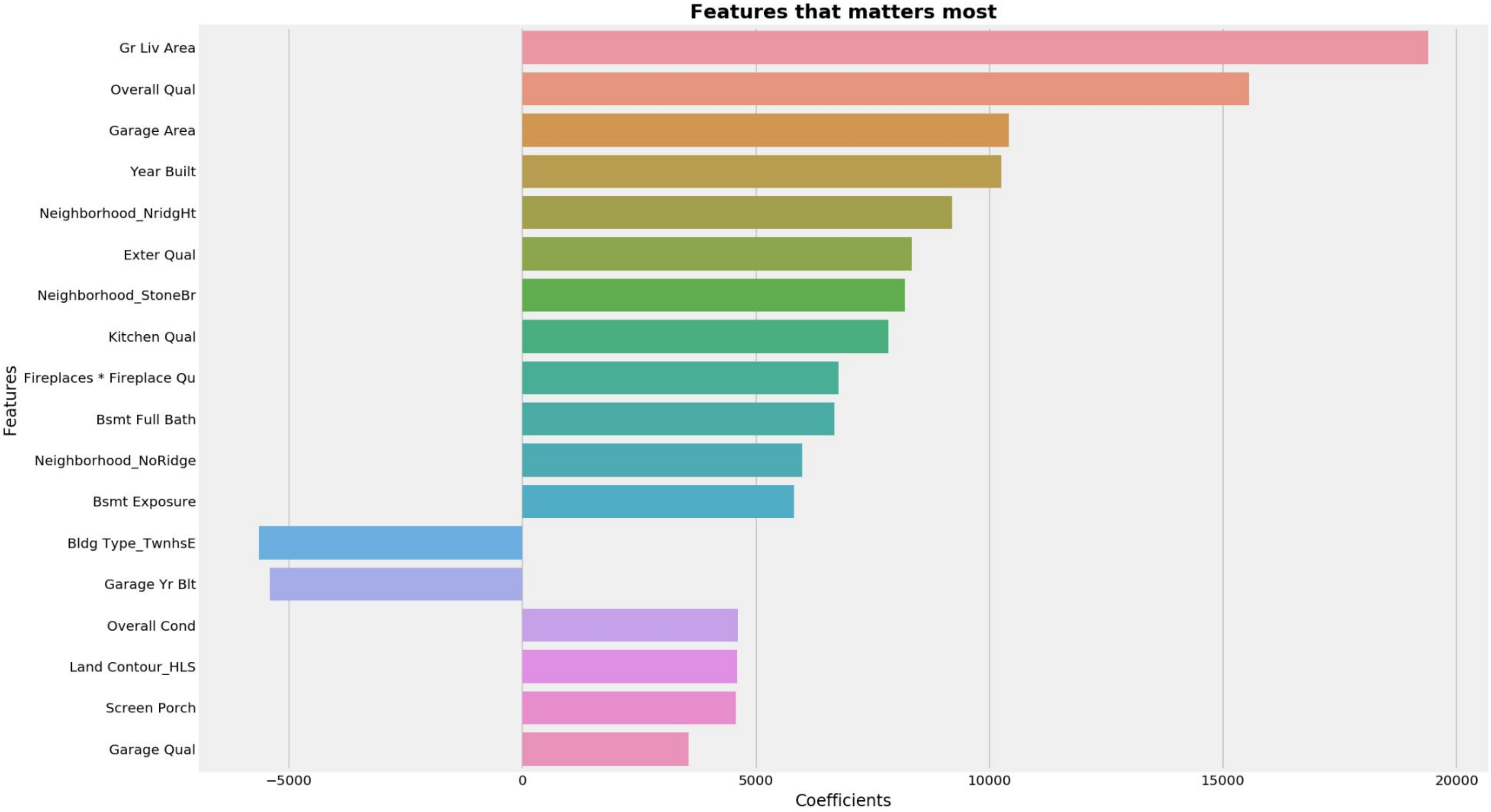
Finding the Best Model with Least RMSE



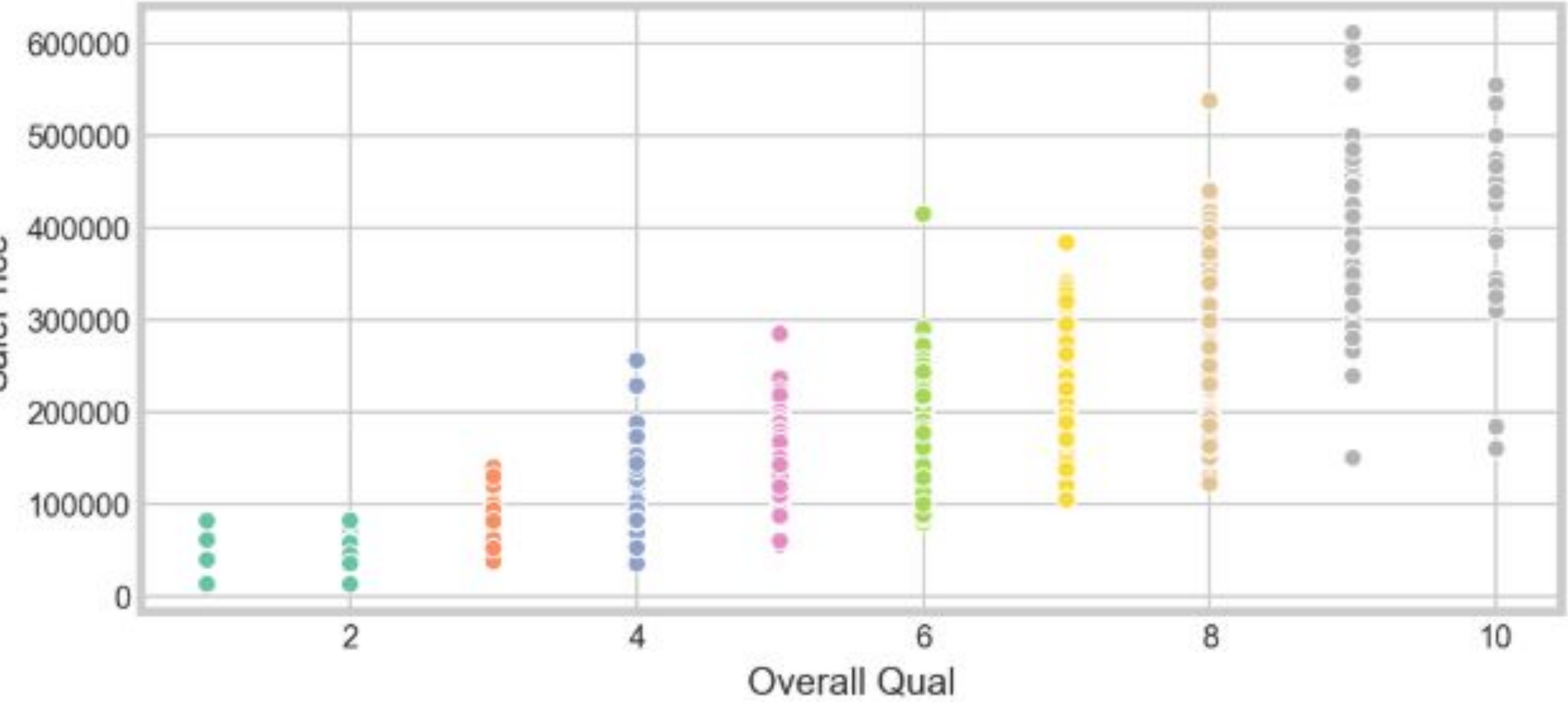

Finding the Best Model with Least RMSE
Model
Linear
Lasso
Ridge
30700
30720
30740
30760
30780
30800
RMSE ($)



Model Evaluation with Ridge


Features that matters most
Gr Liv Area
Overall Qual
Garage Area
Year Built
Neighborhood_NridgHt
Exter Qual
Neighborhood_StoneBr
Kitchen Qual
Fireplaces * Fireplace Qu
Bsmt Full Bath
Neighborhood_NoRidge
Bsmt Exposure
Bldg Type_TwnhsE
Garage Yr Blt
Overall Cond
Land Contour_HLS
Screen Porch
Garage Qual
Features
−5000
0
5000
10000
15000
20000
Coefficients

- Conclusions



Conclusions

- Never compromise on quality
- Lesser number of features does not translate to lower selling price
- A larger floor area → higher price
- Newer the house → higher the price

Lets build smartly, the data science way!!!!