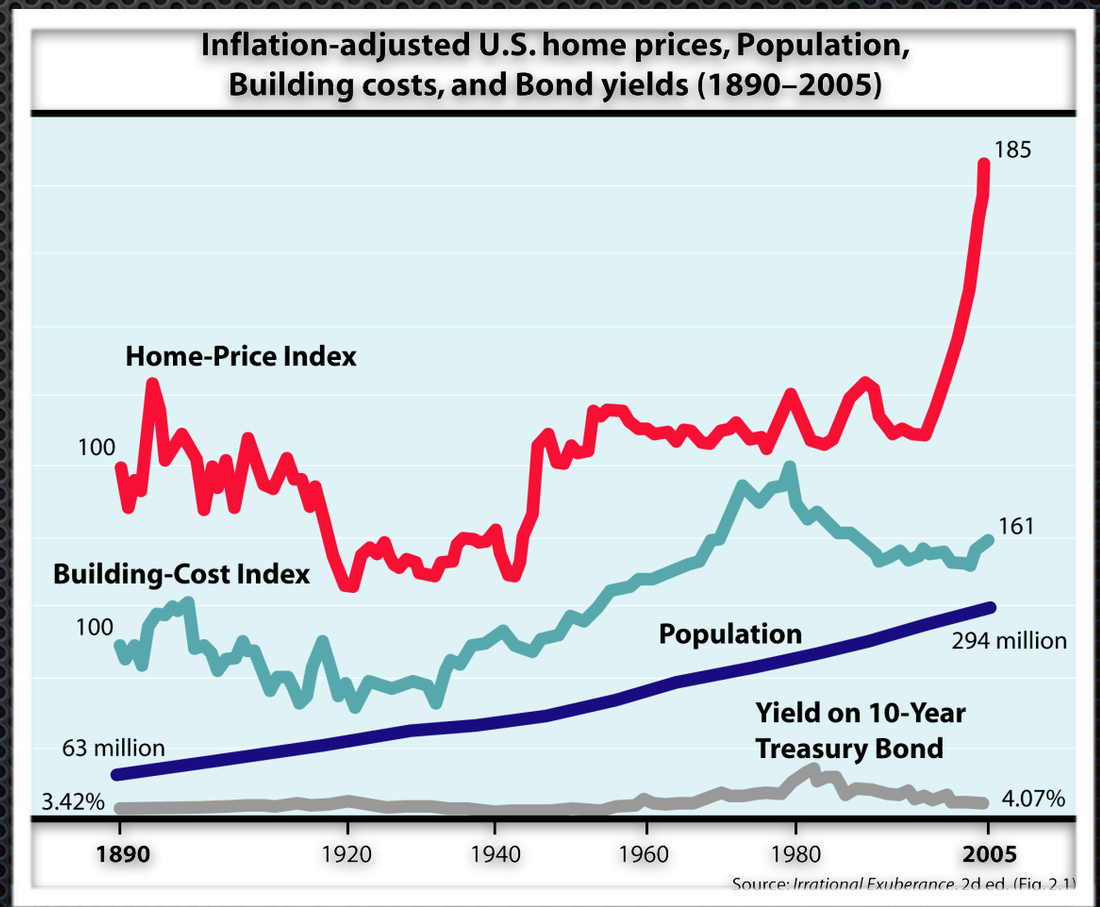
Predicting Sales Prices for the Kaggle Ames, Iowa Housing Dataset

Team Proxima Centauri - Wenchang, Kenny, James, Danny

Brief Overview of Data Set

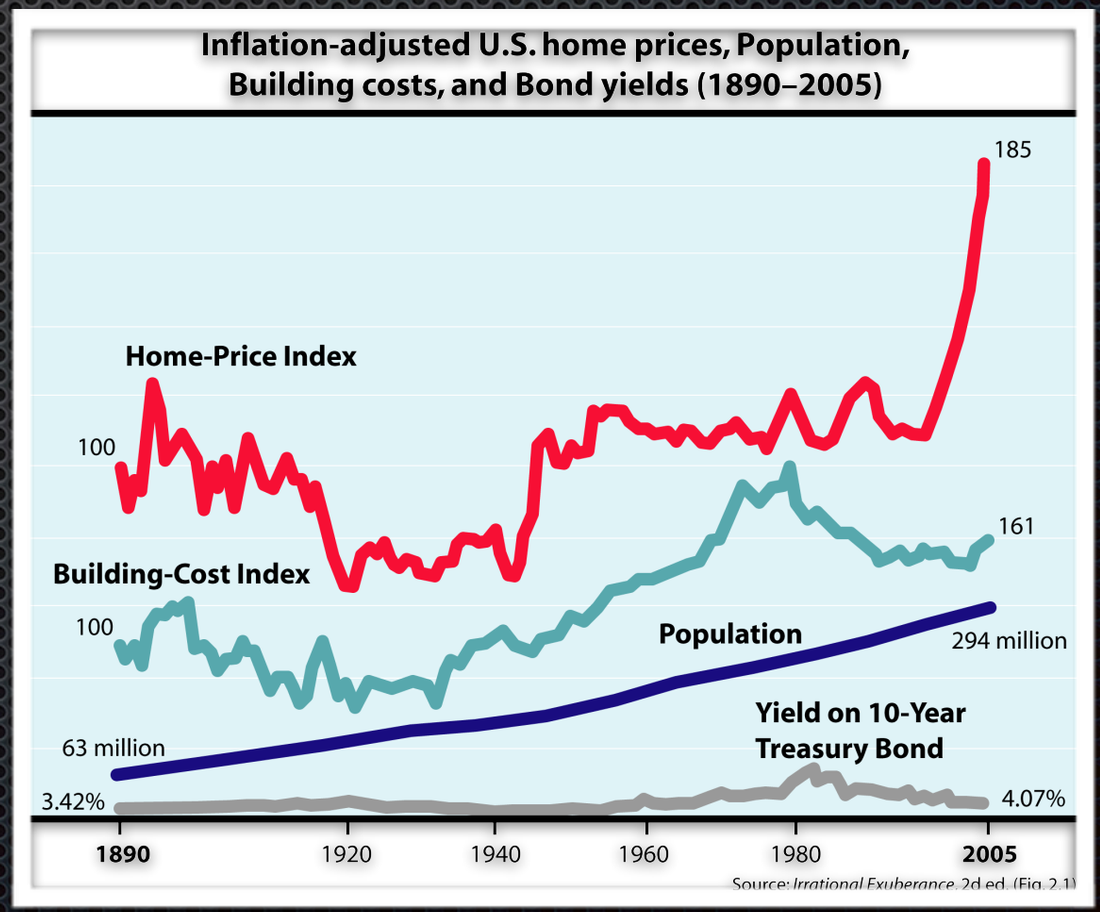
- 80 variables tied to price of home sales in Ames, Iowa between 2006-2010
- Initially obtained from Ames City Assessor's Office with 113 variables that described 3970 property sales -> cleaned and collated by Dean De Cock {Journal of Statistics Ed., **19**, (3) 2011}
- Reduced to 80 predictors:
  - 20 continuous variables {total dwelling sq. footage etc}
  - 14 discrete variables {number of bathrooms, kitchens}
  - 46 categorical variables - 23 nominal/23 ordinal {street pavement, neighborhood}

The US Housing Bubble



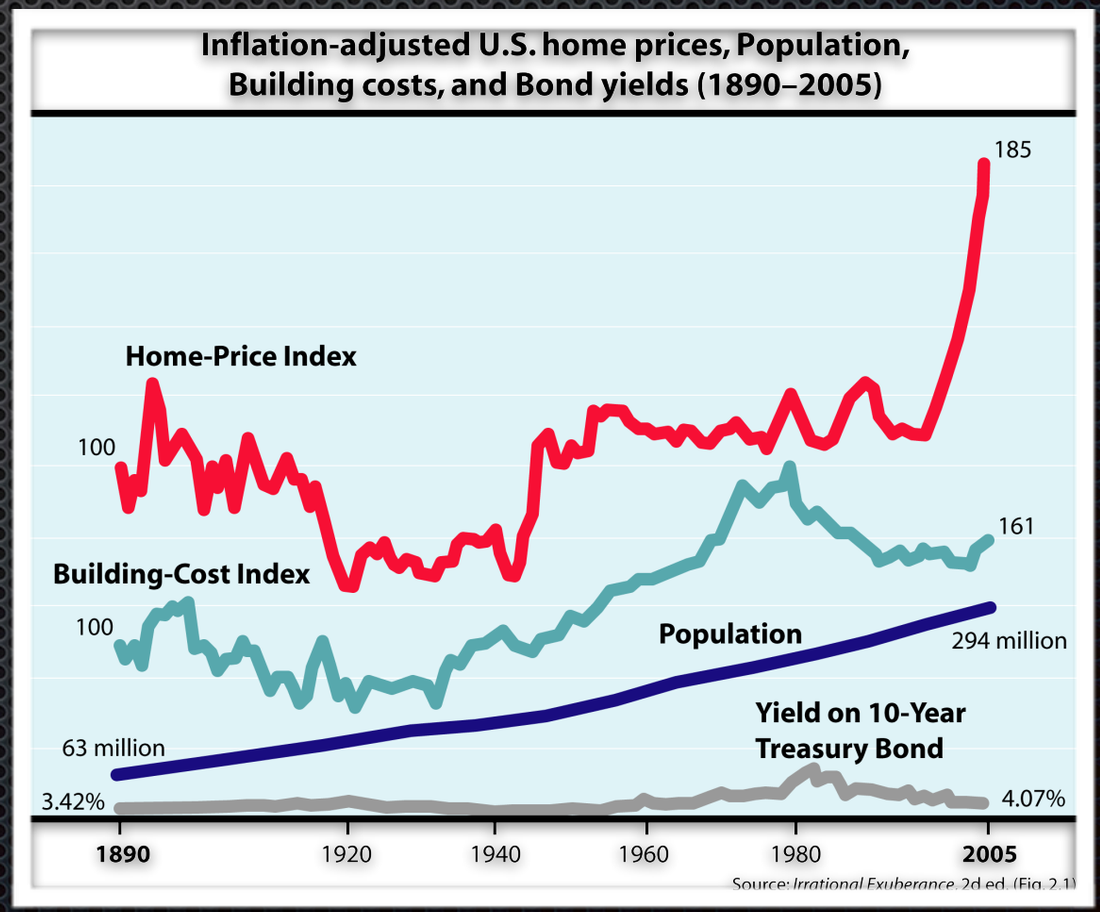
The US Housing Bubble

- What was going on between 2006-2010 in our country?
- Will there be a noticeable effect?



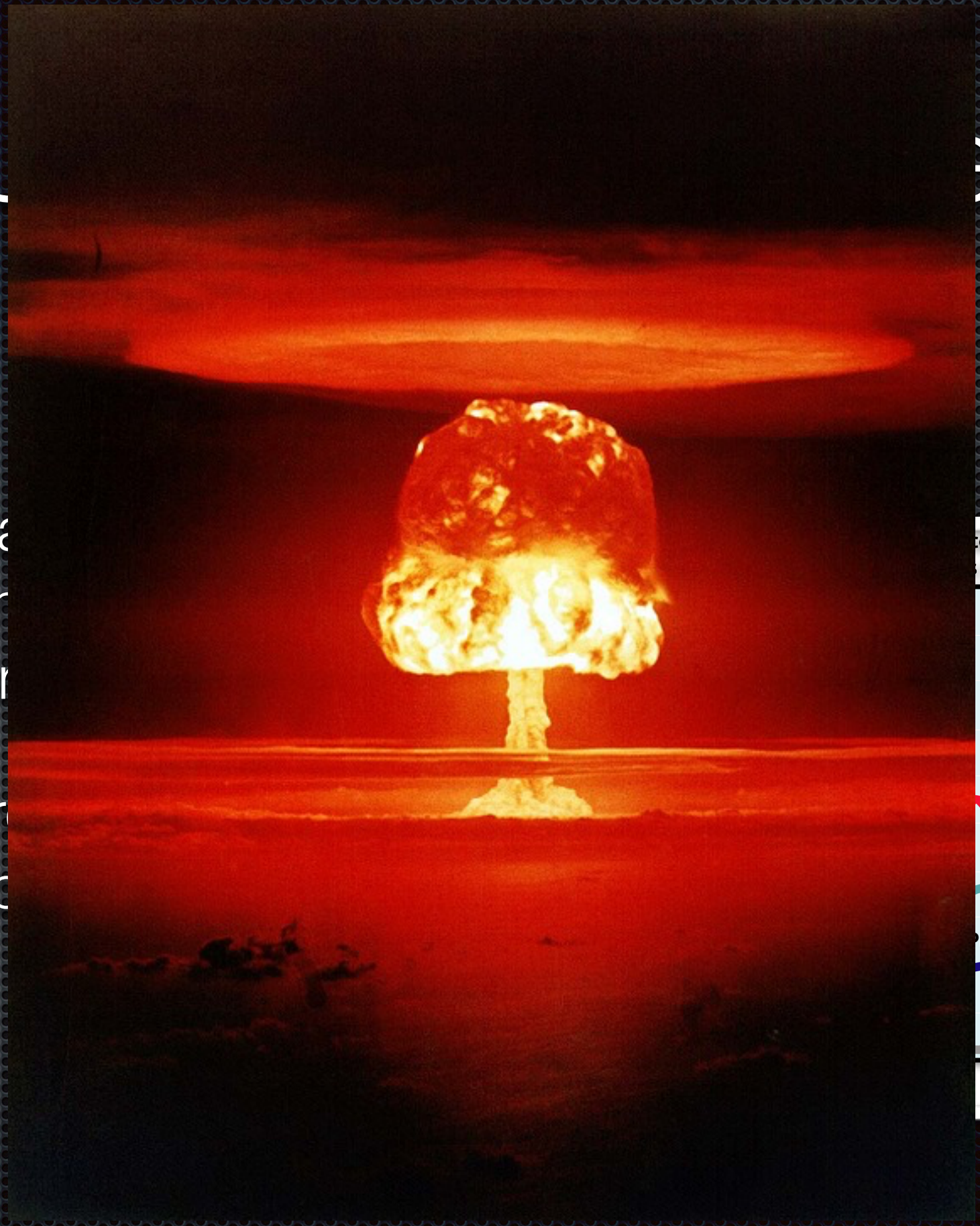
The US Housing Bubble

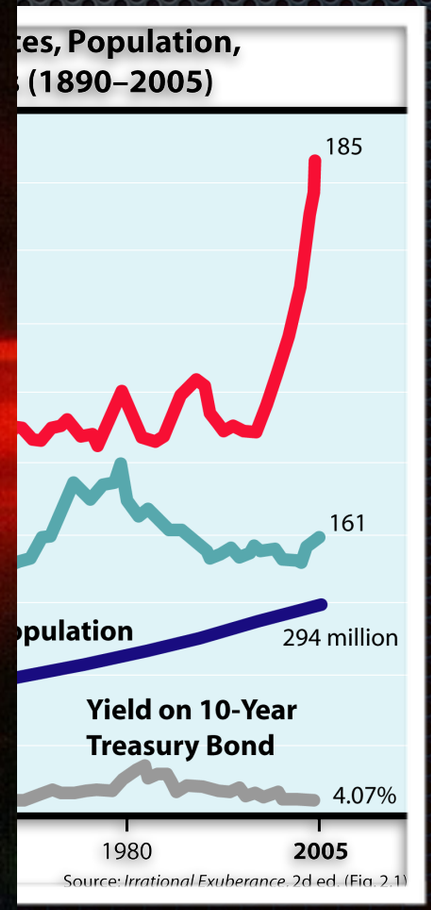
- What was going on between 2006-2010 in our country?
- Will there be a noticeable effect?



The U

- What wa
  betweer
  our cour
- Will ther
  noticeab





Considerations Involving Home Prices

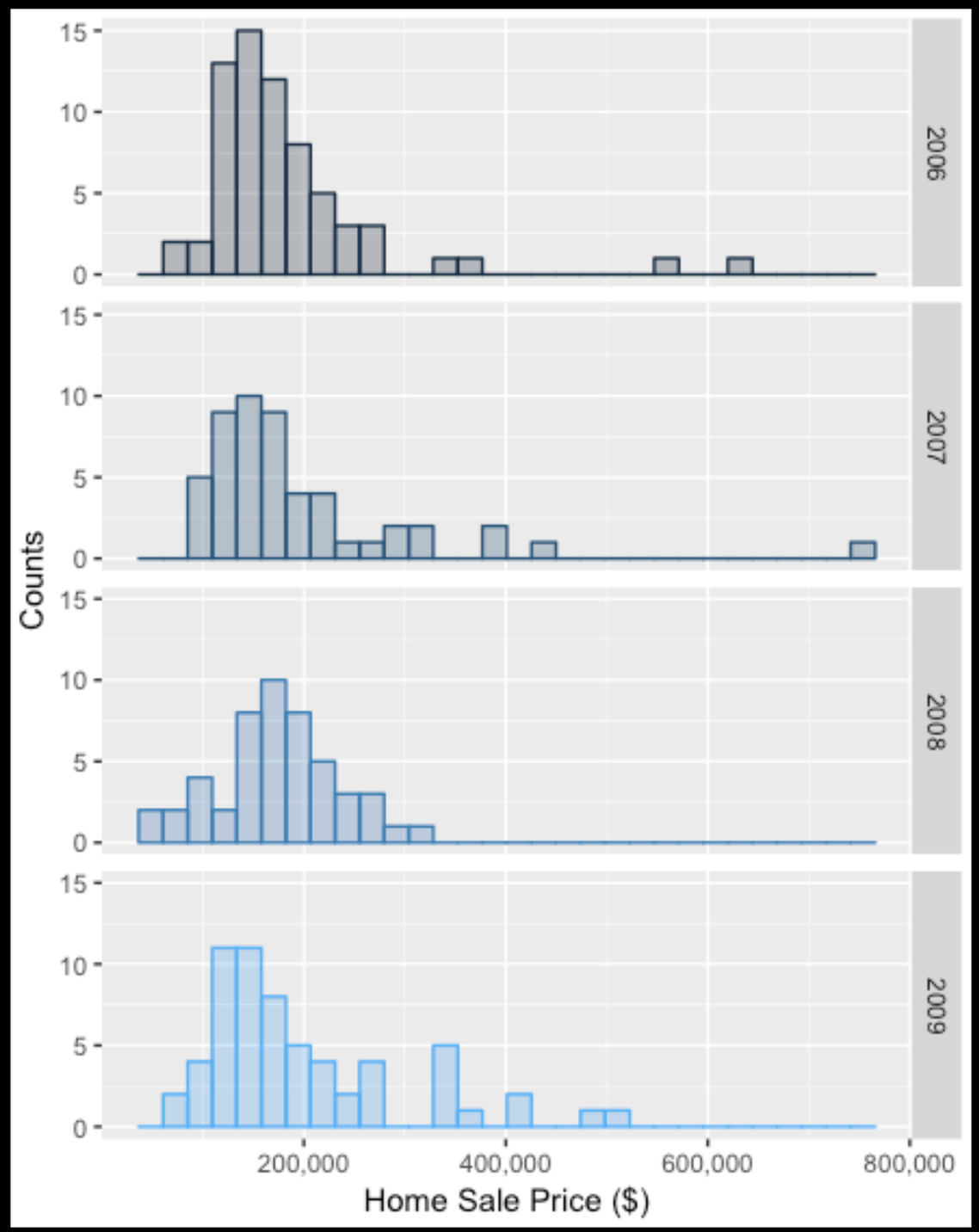
How did monthly home sales vary during the timespan represented in dataset?

https://plot.ly/~dbubb/13/

Buyer's market in Spring, Seller's in Fall

Kruskall-Wallis (~Anova): p=0.97, df=3, χ^2 = 0.24345

Levene(~Barlett's): p=0.97, df=3, F = 1.325



Considerations Involving Home Prices

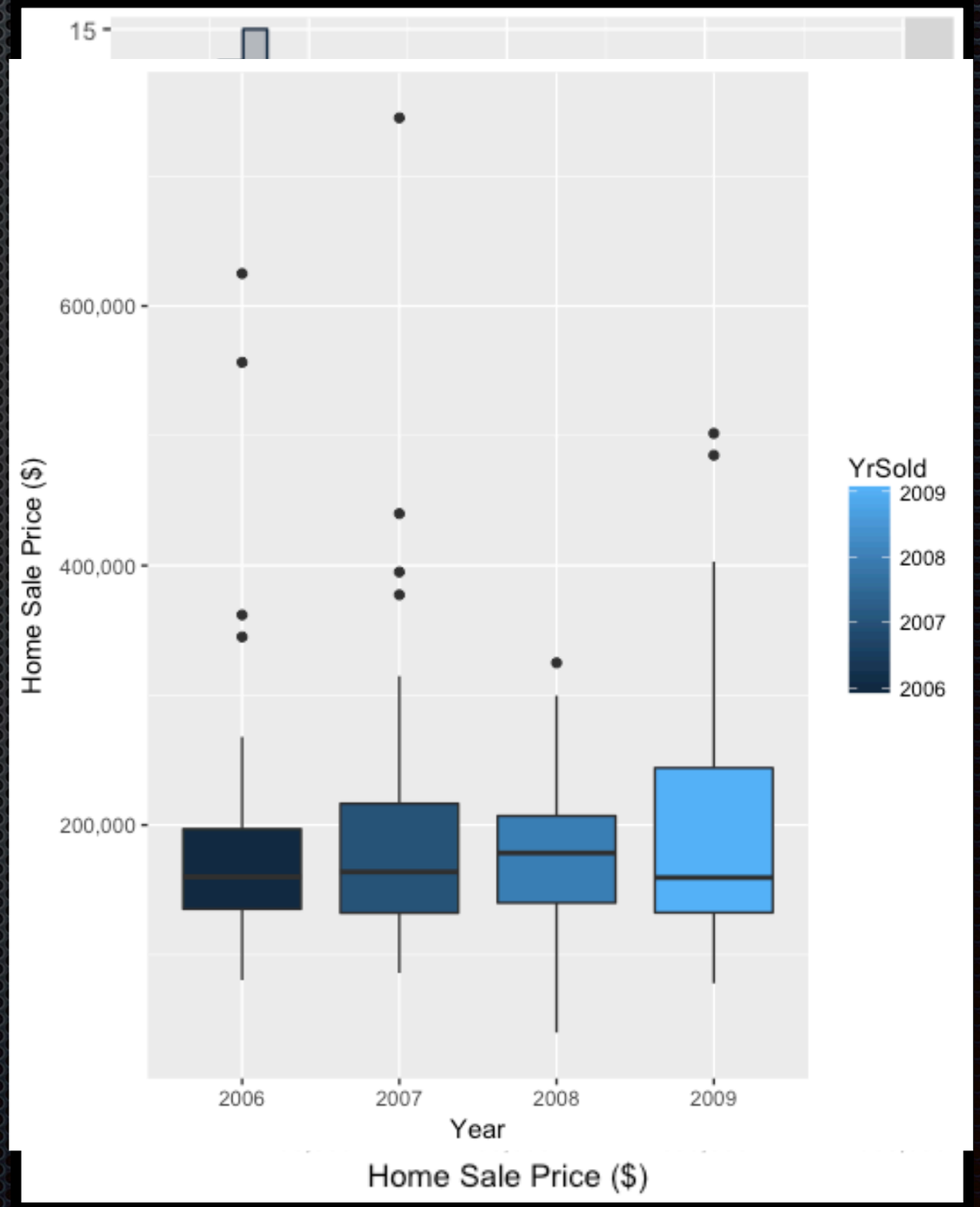
How did monthly home sales vary during the timespan represented in dataset?

https://plot.ly/~dbubb/13/

Buyer's market in Spring, Seller's in Fall

Kruskall-Wallis (~Anova): p=0.97, df=3, $\chi^2 = 0.24345$

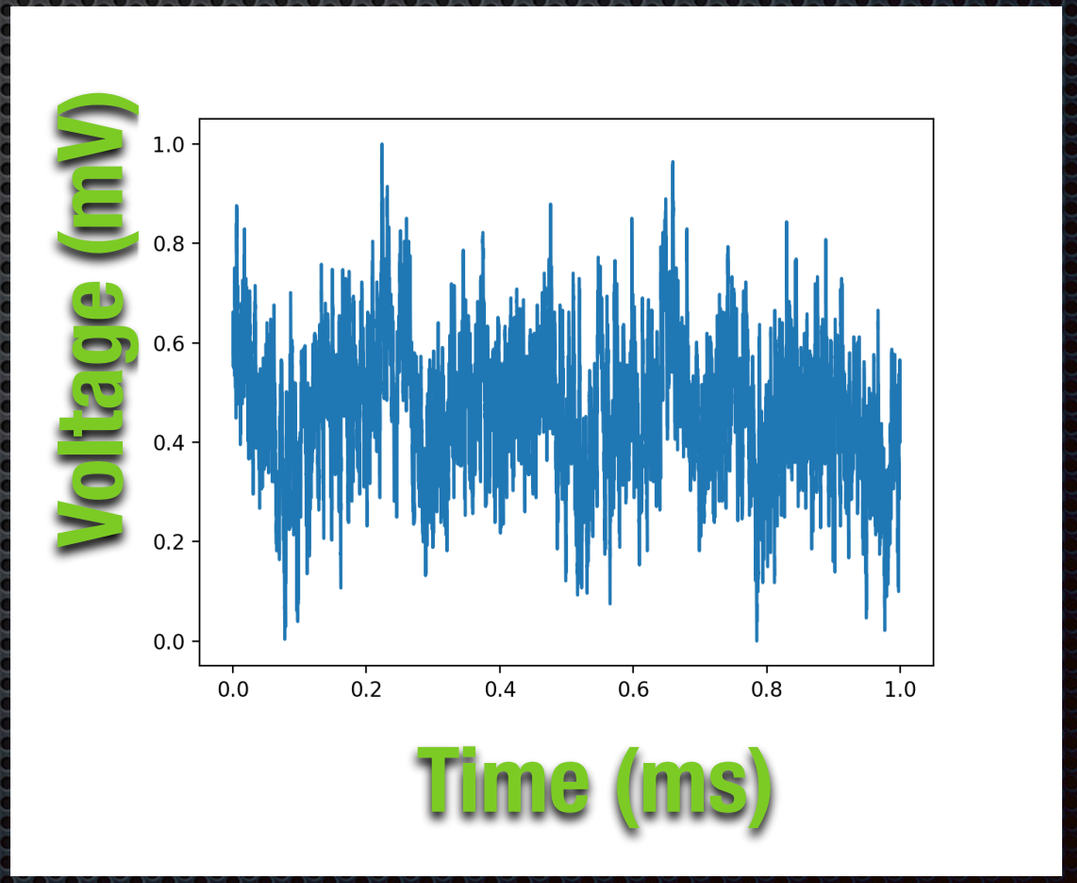
Levene(~Barlett's): p=0.97, df=3, F = 1.325



Hurst exponent

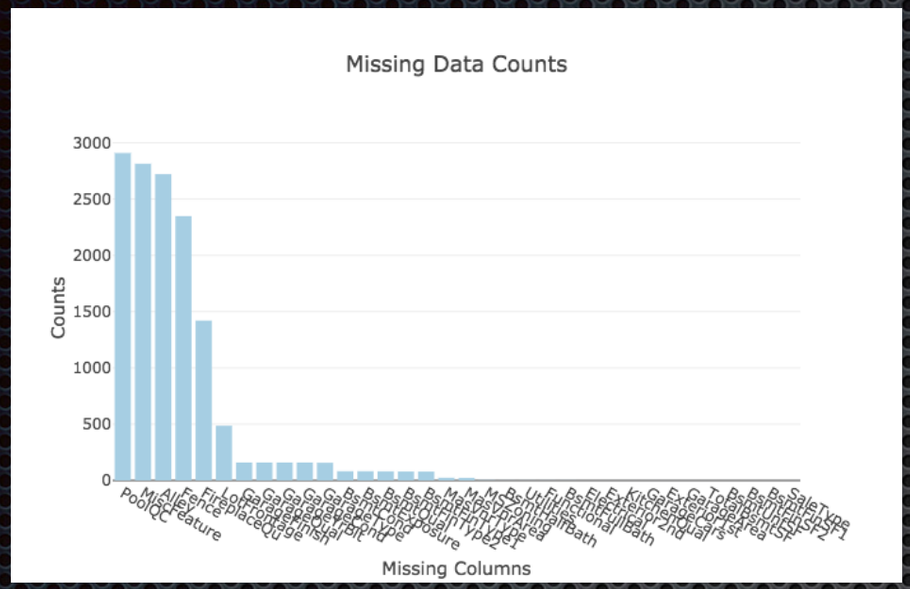
$$D_f = -\lim_{\epsilon \to 0} \frac{ln(N(\epsilon))}{ln(\epsilon)}$$

- H = 2 - D_f = 0.536 - ~ Brownian Motion
  - 0 < H < 0.5 - mean-reverting, anti-persistent
  - 0.5 < H < 1- persistent $\Longrightarrow$ increases tend to follow increases, same for decreases…

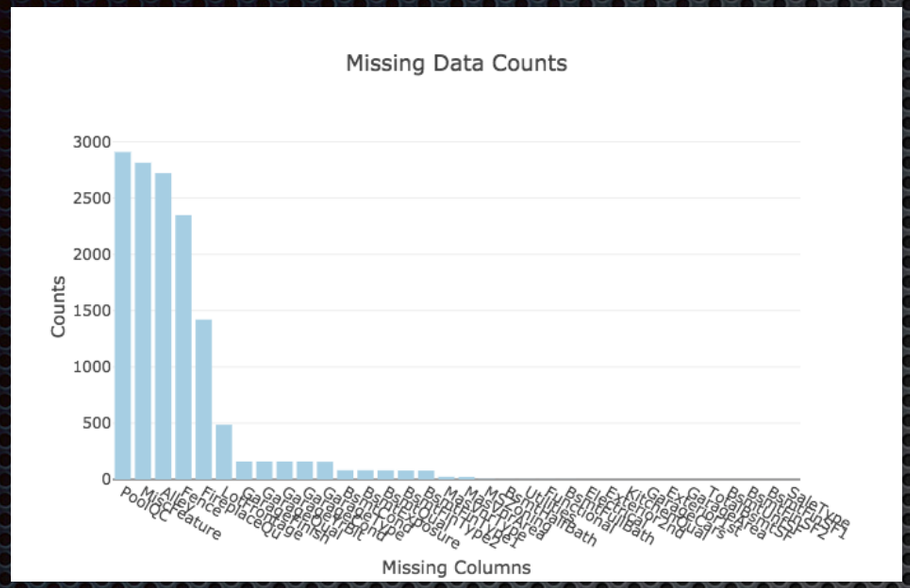


Good discussion here re: S&P 500

Data Cleaning, Feature Eng.

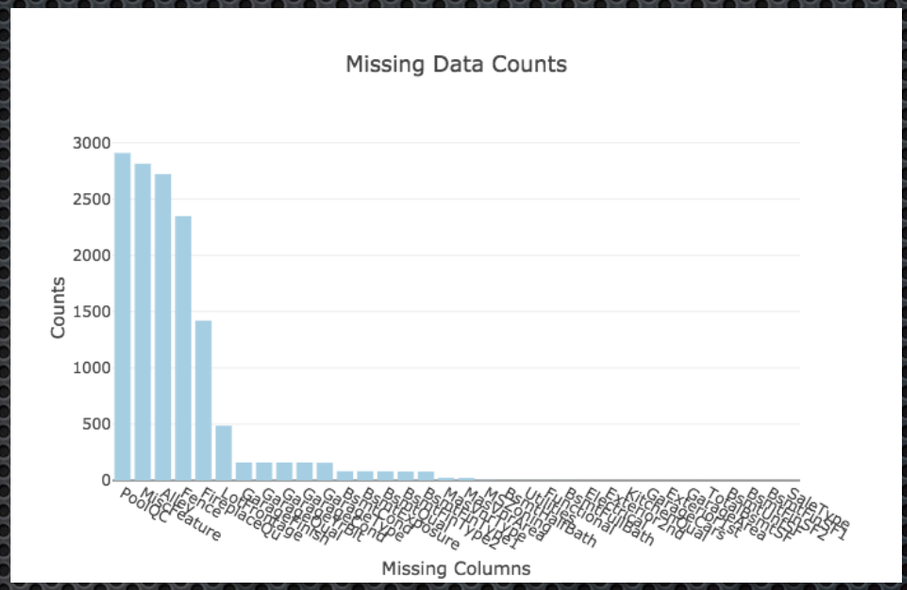


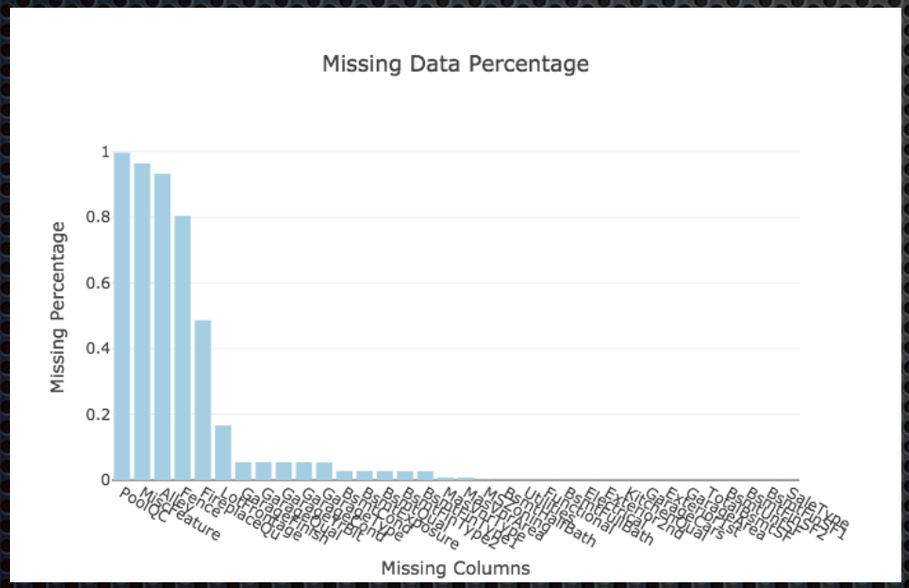
Data Cleaning, Feature Eng.



- Missingess
- Imputation
- Added features
- Normalization

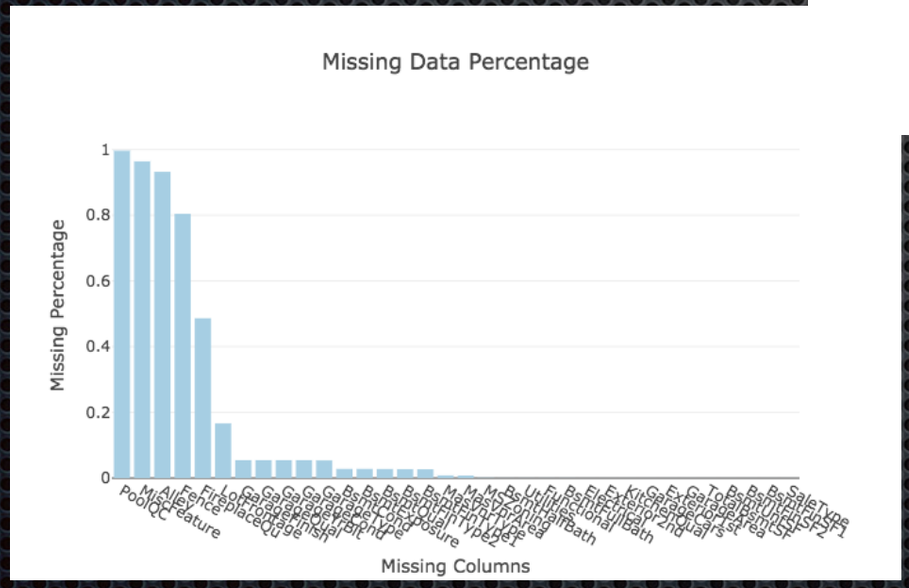
Data Cleaning, Feature Eng.





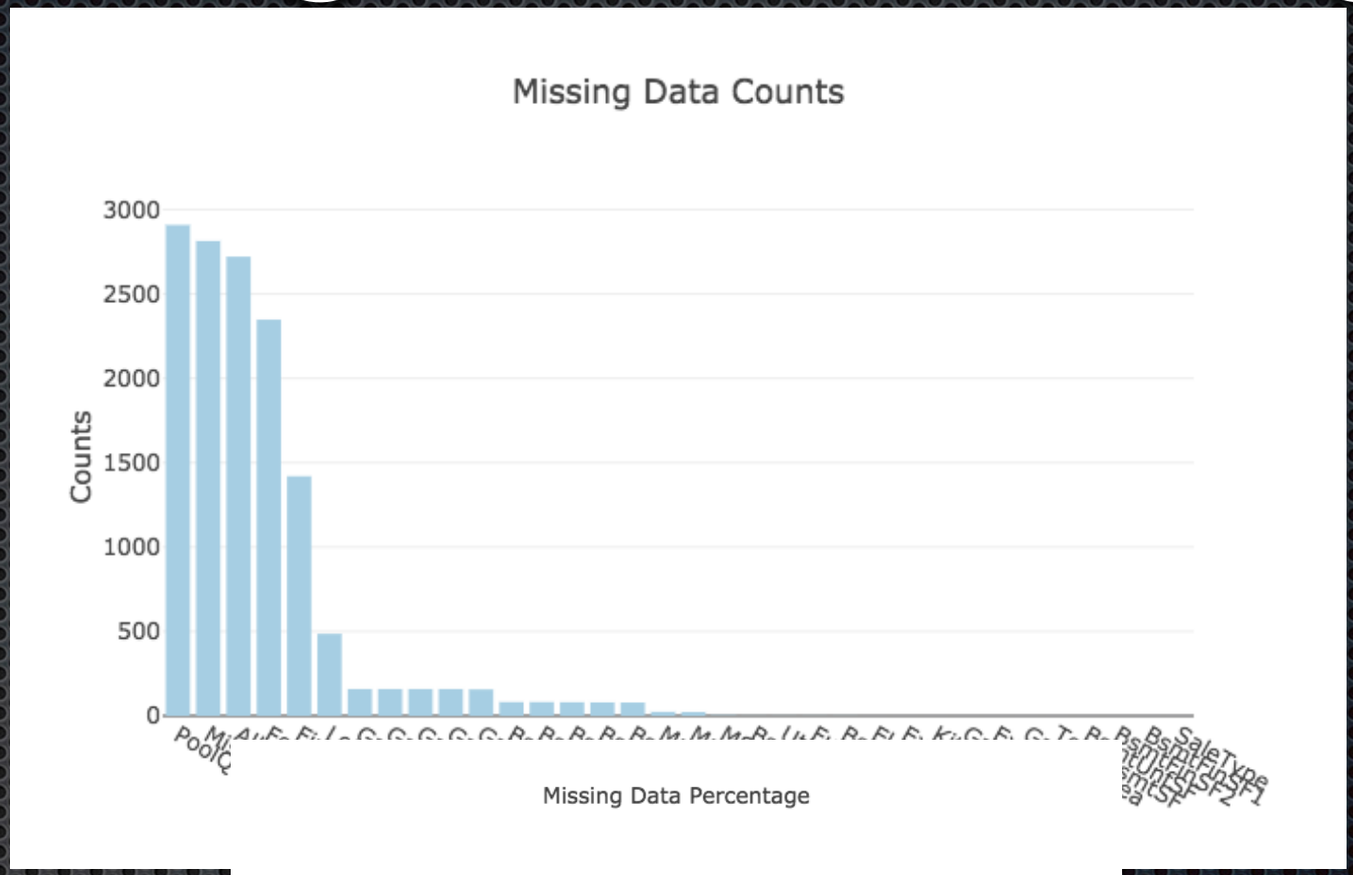
- Normalization

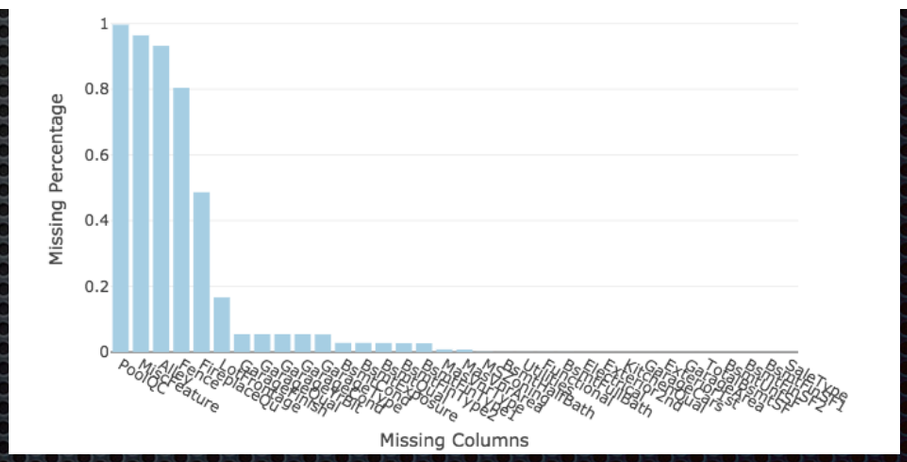
Data Cleaning, Feature Eng.



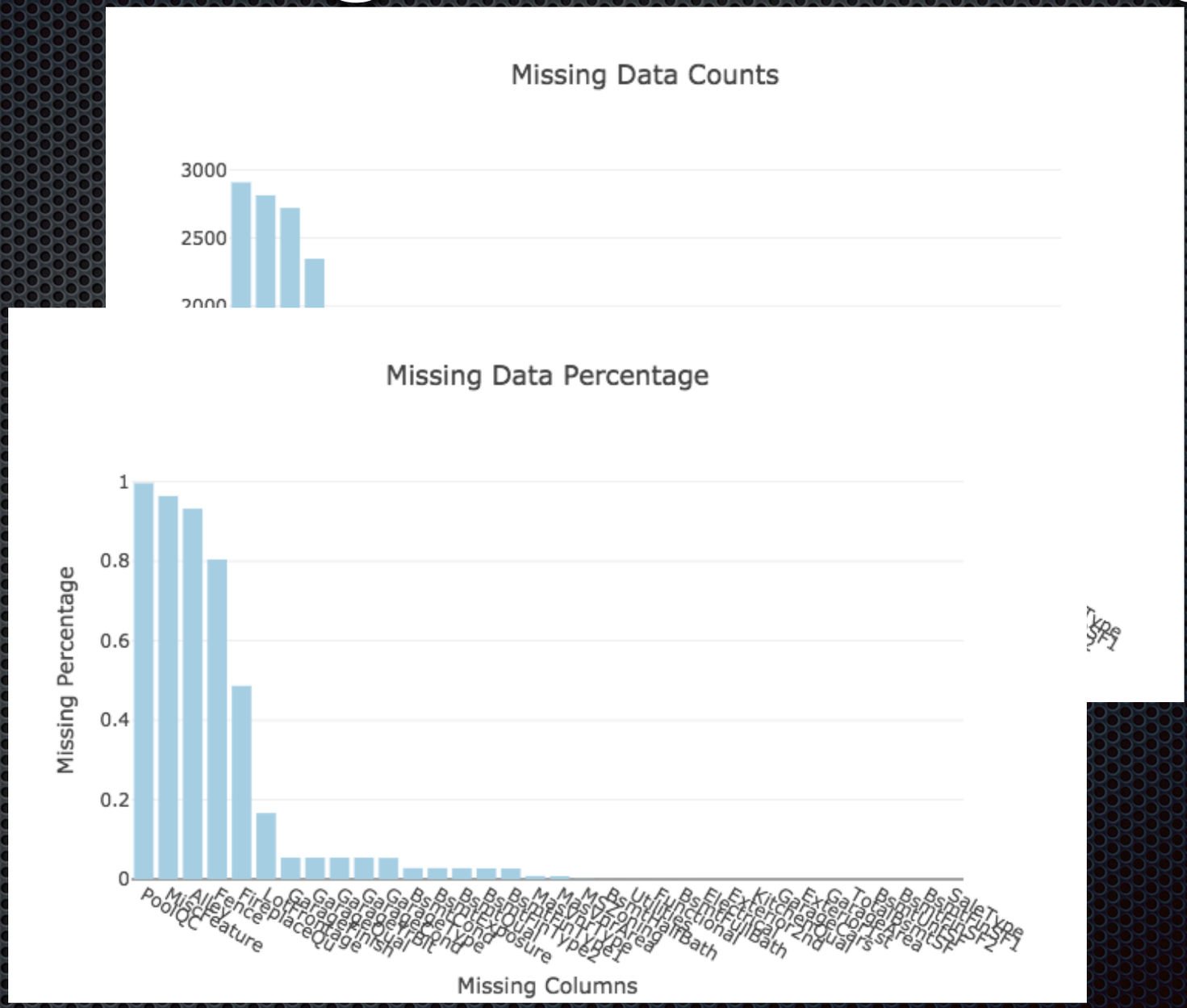
- Normalization

Data Cleaning, Feature Eng.

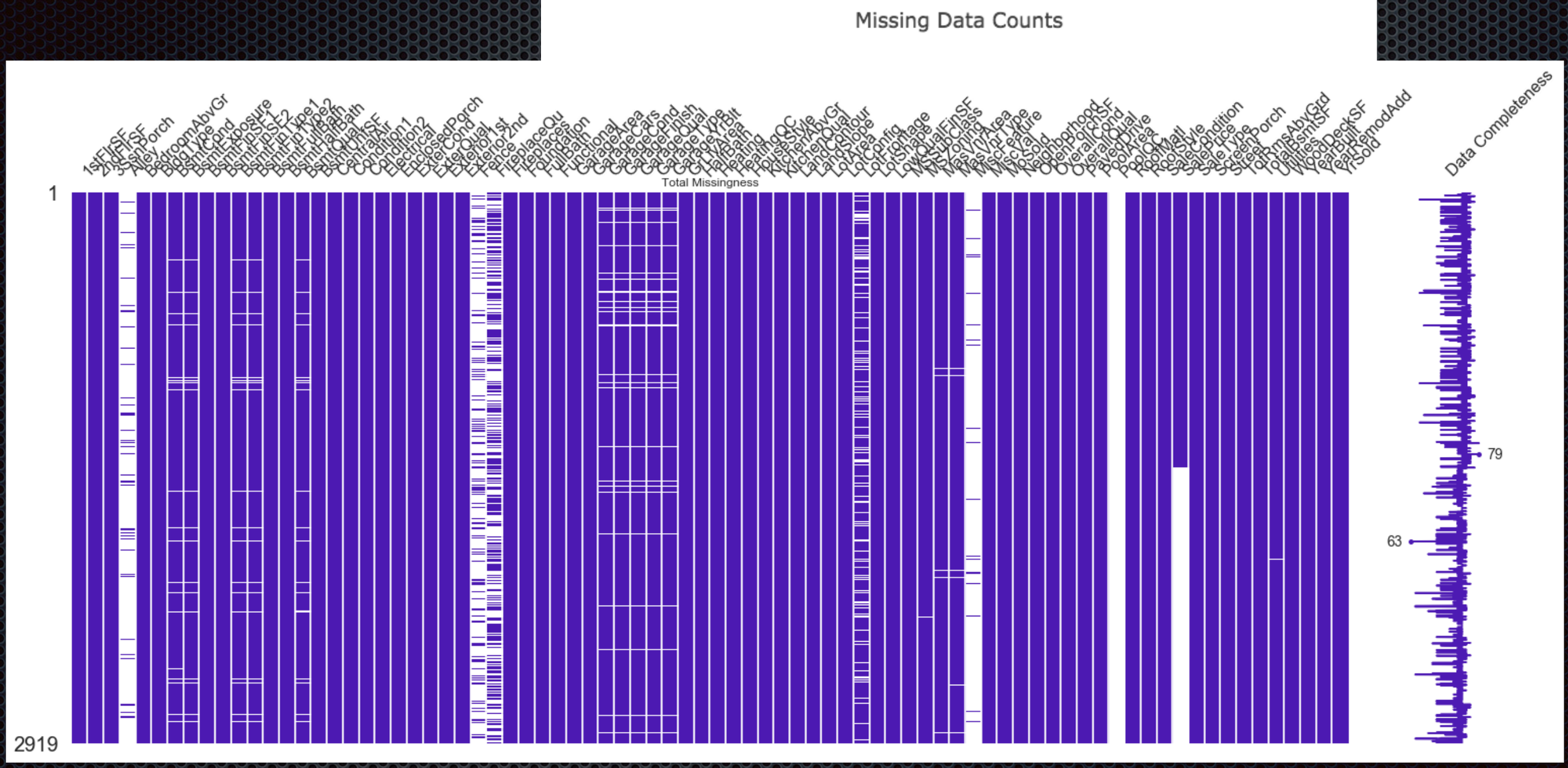


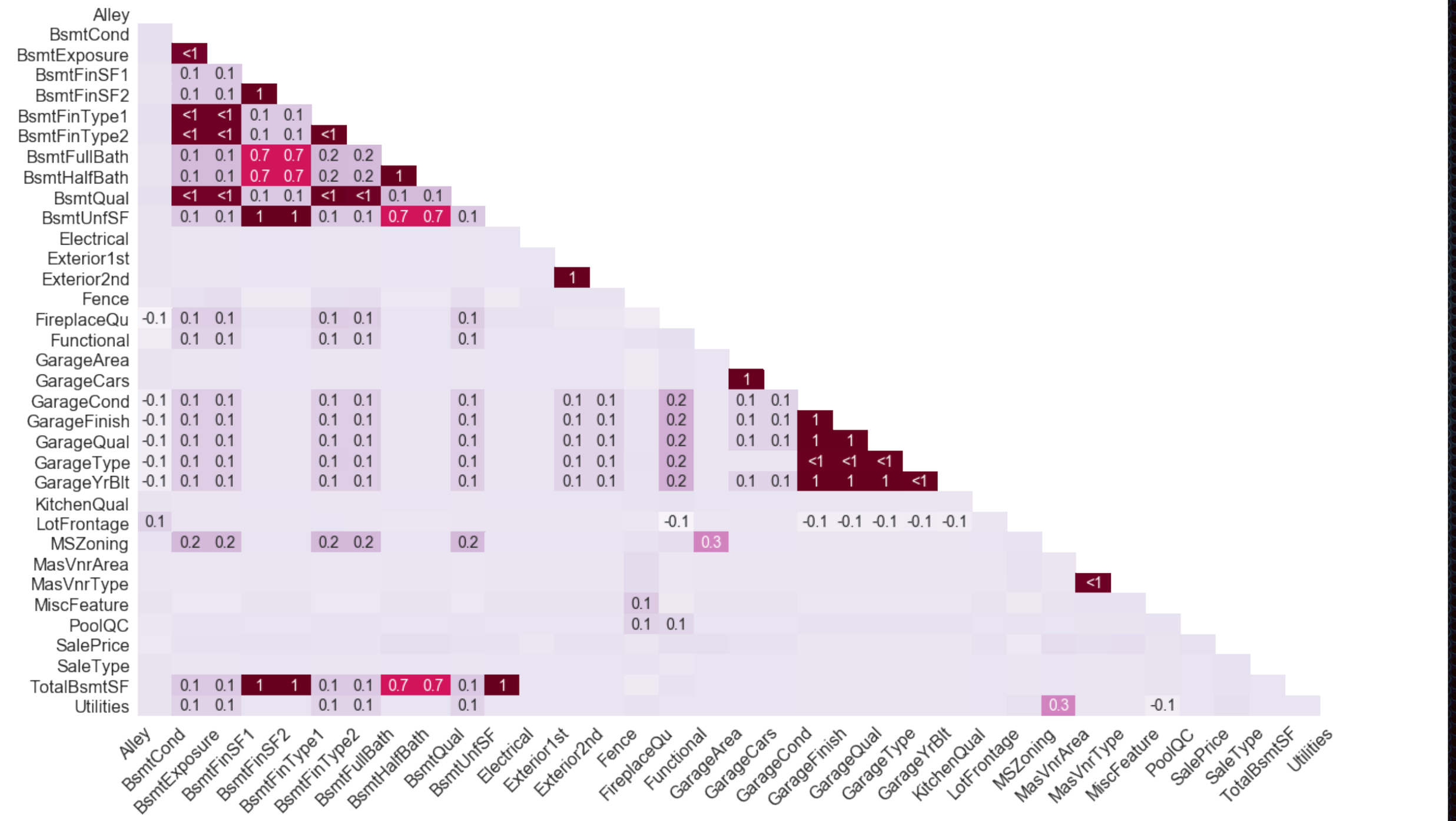


Data Cleaning, Feature Eng.



Data Cleaning, Feature Eng.

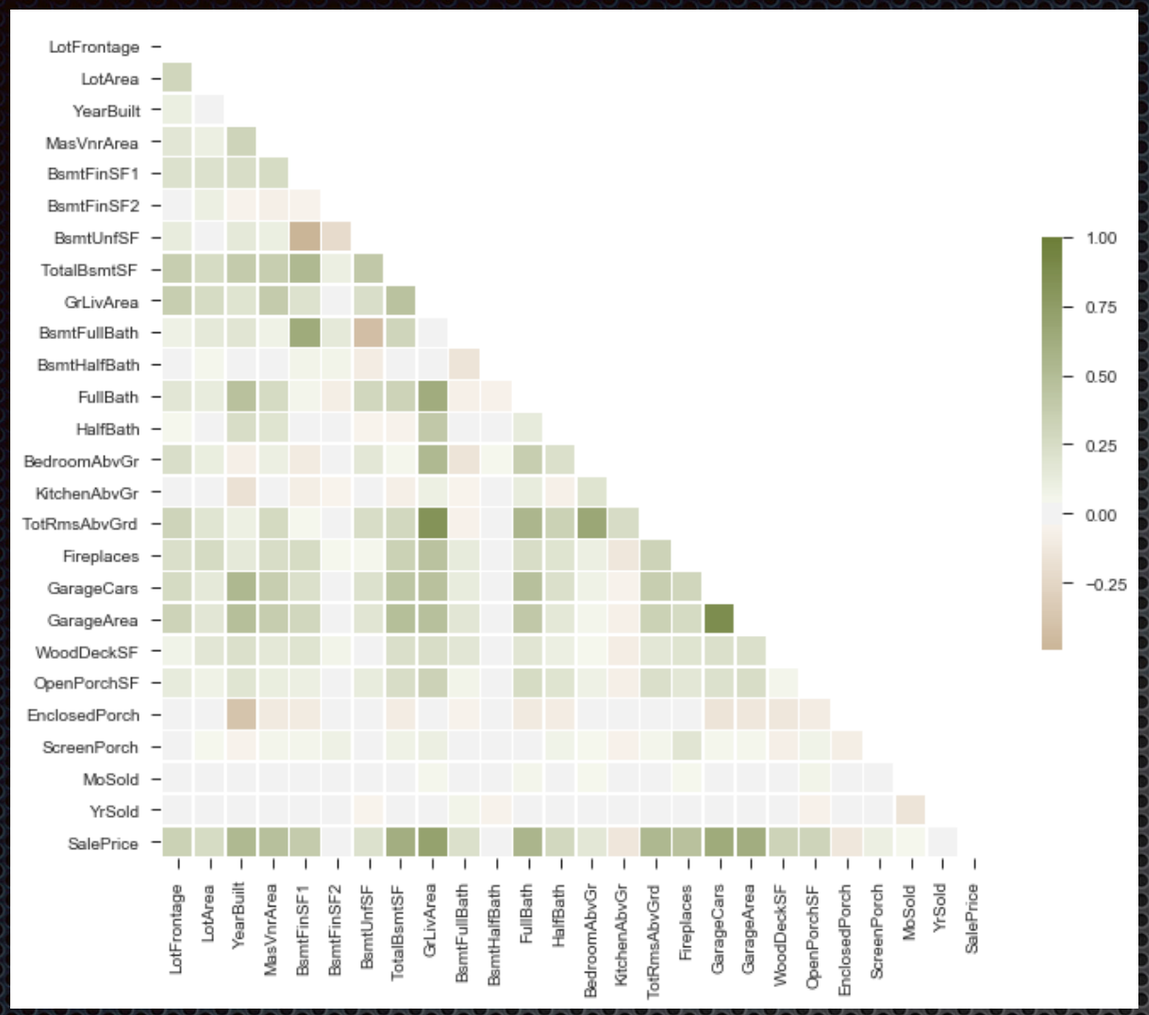


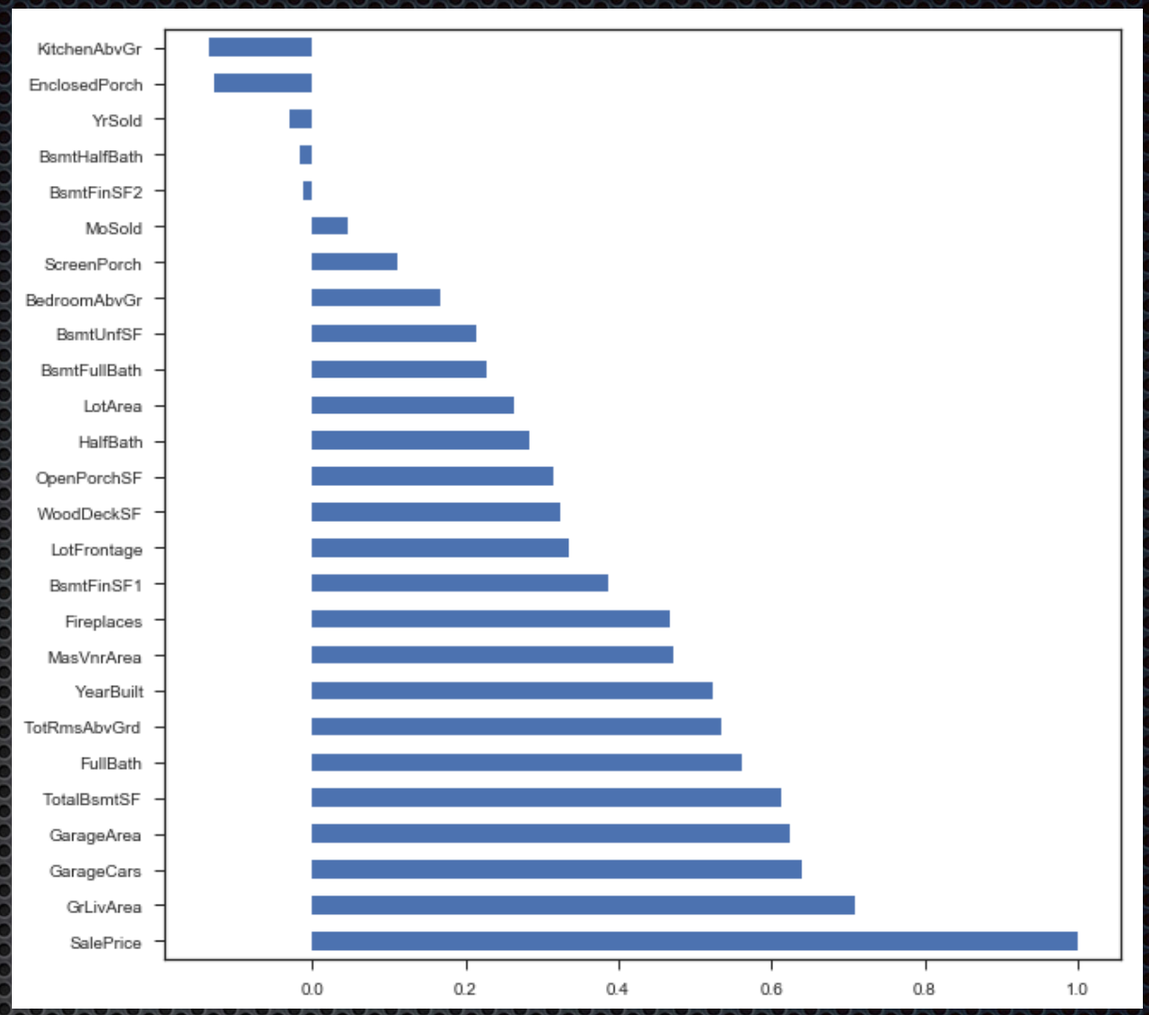


Correlated Missingness?

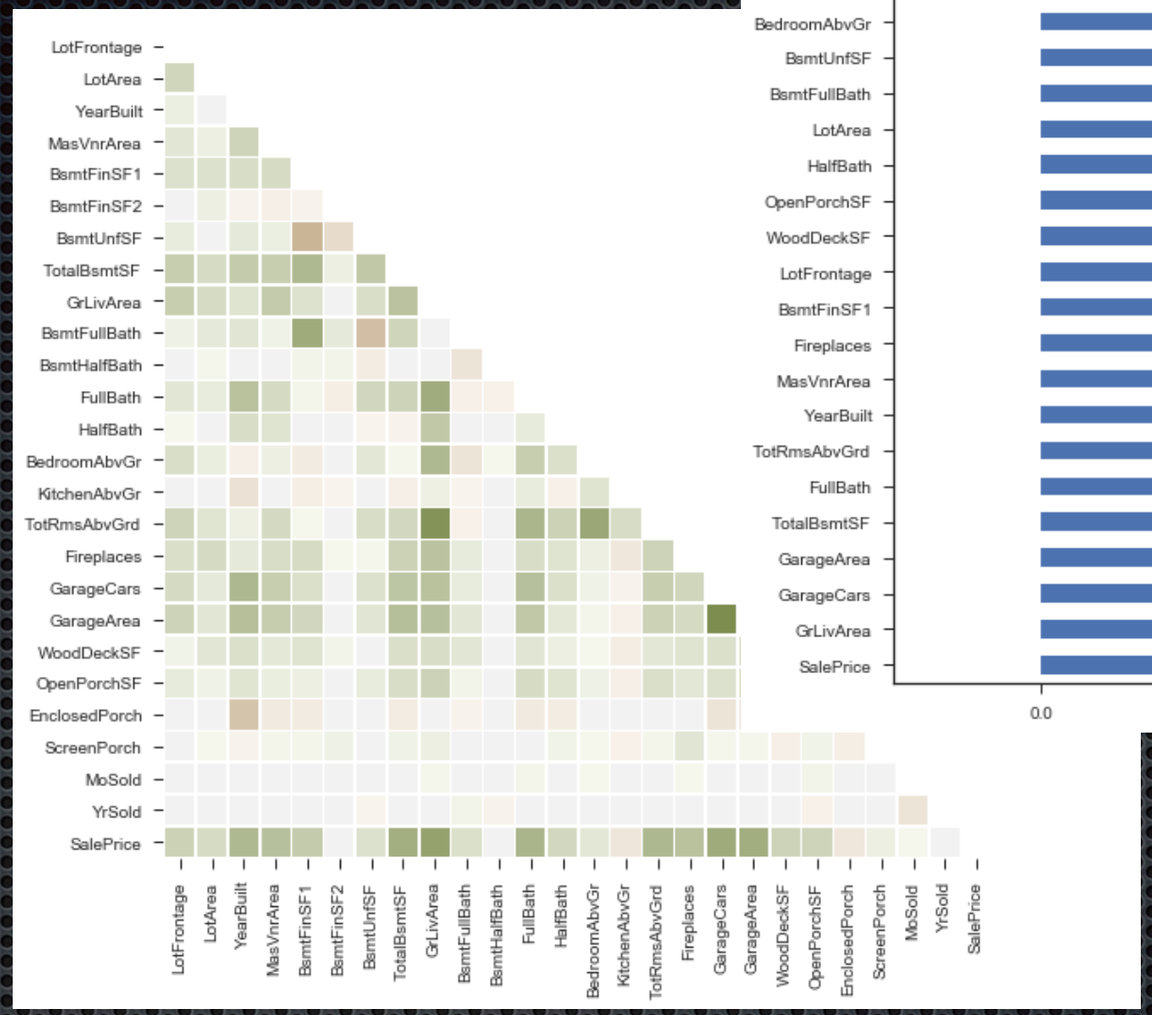
Skew and Dummifying

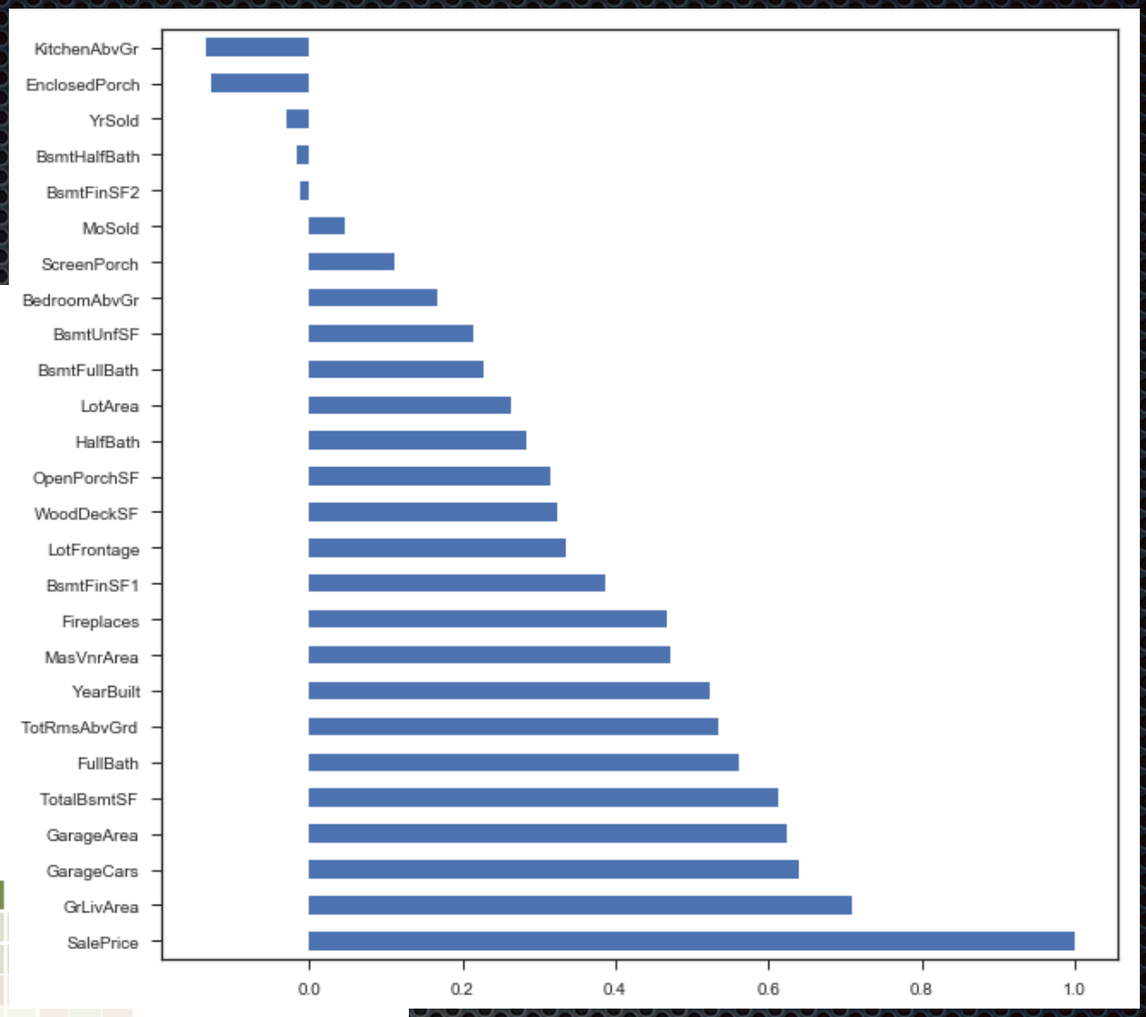
Sales Price info



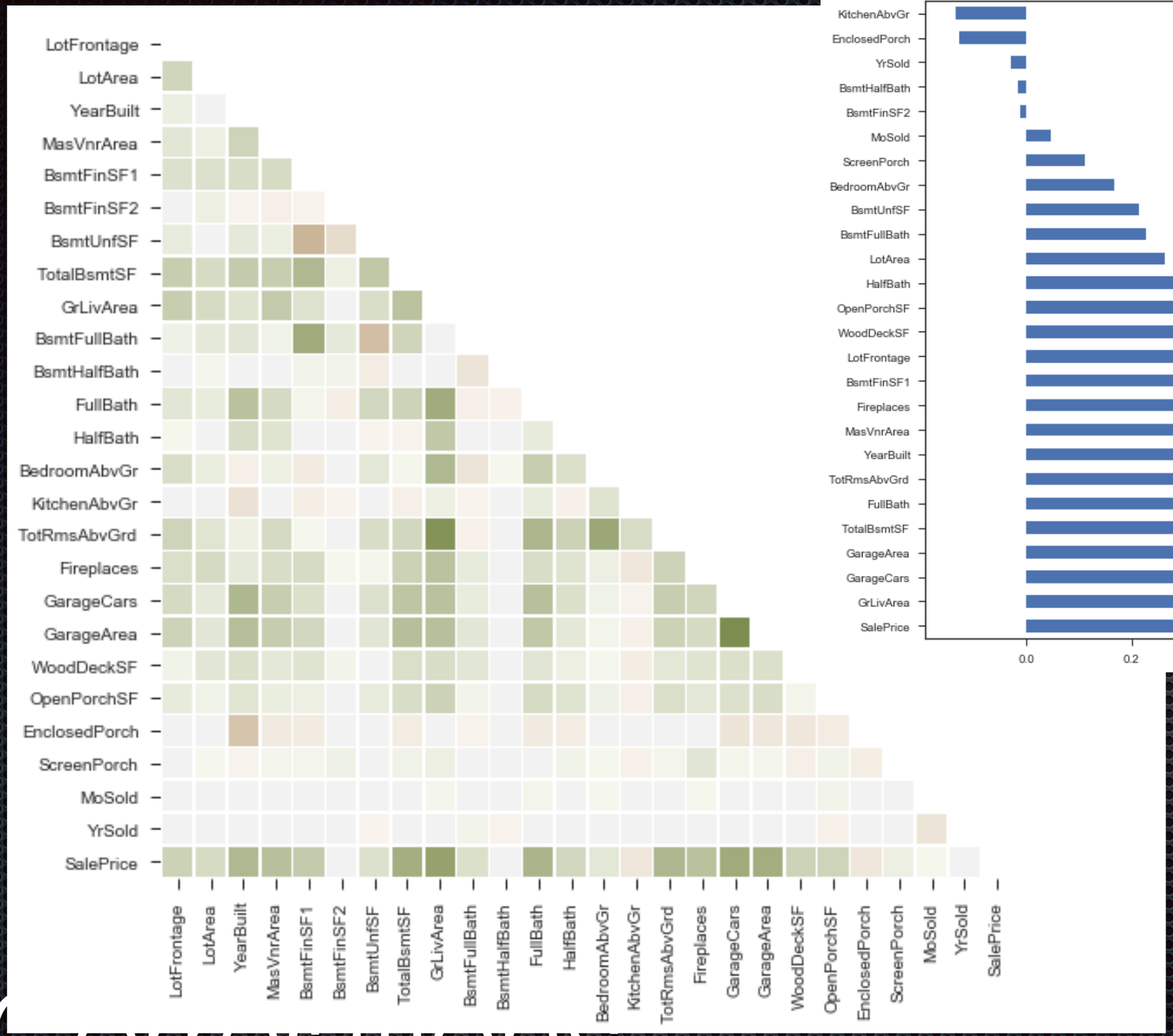


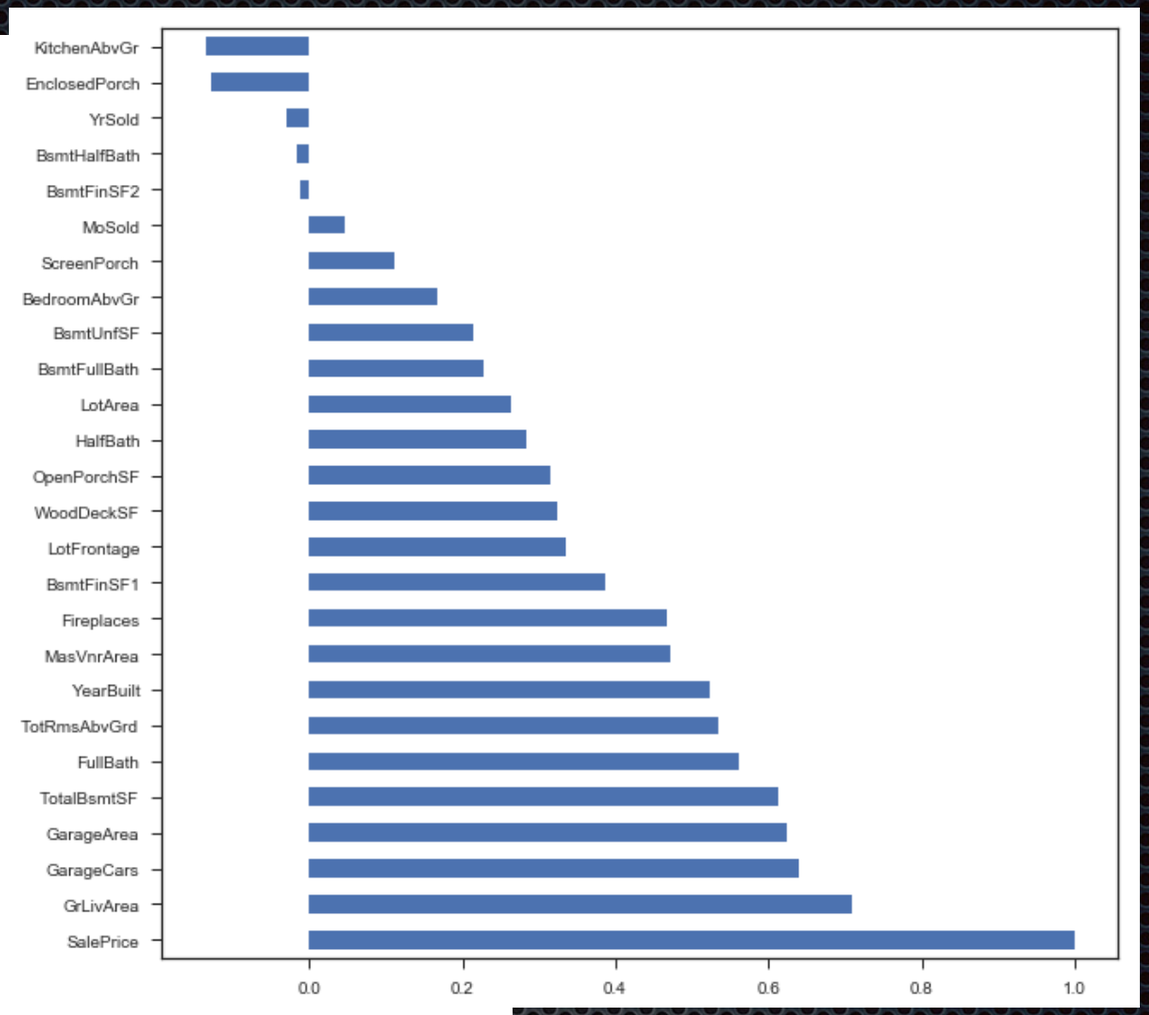
Correlations





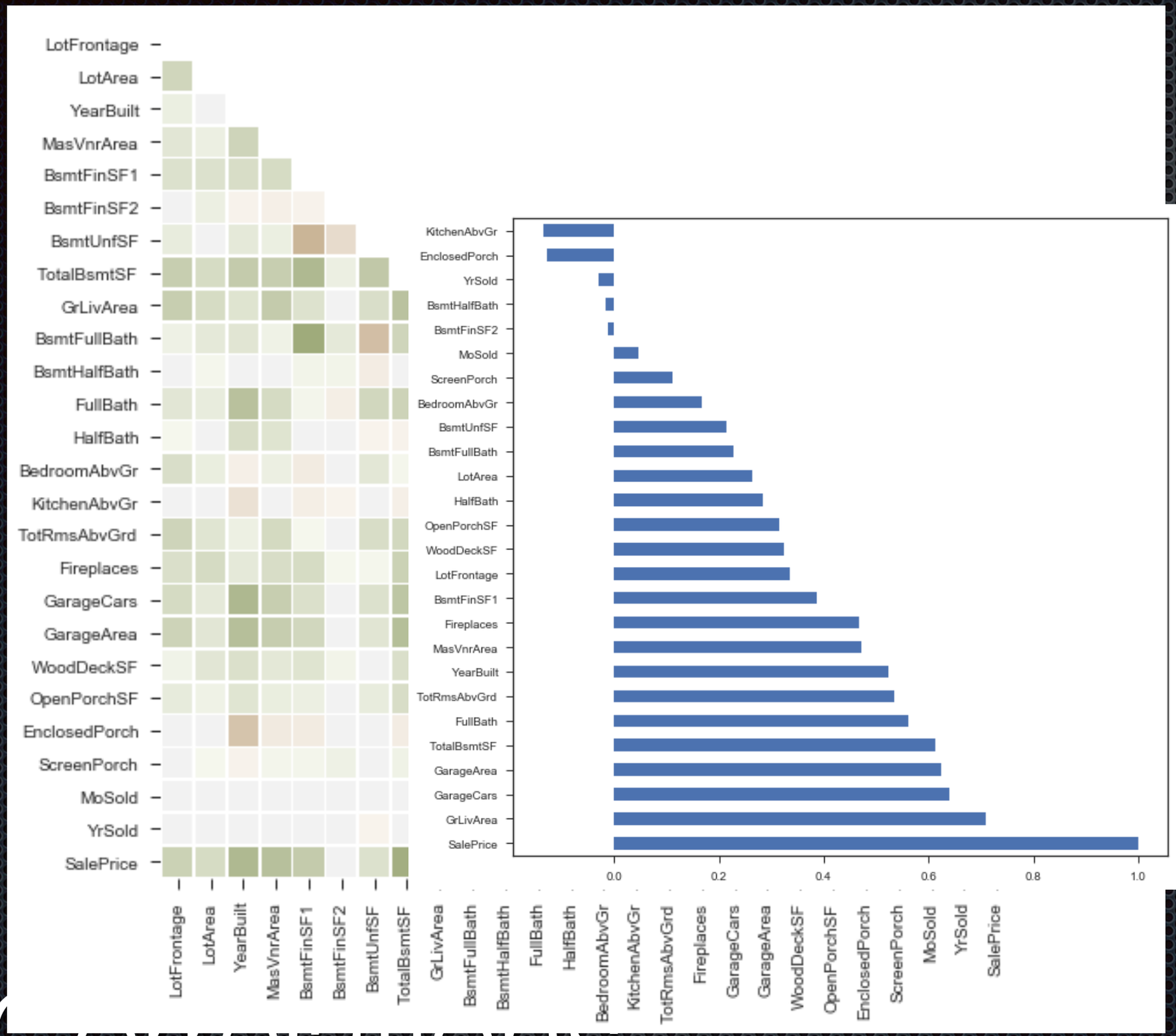
Correlations



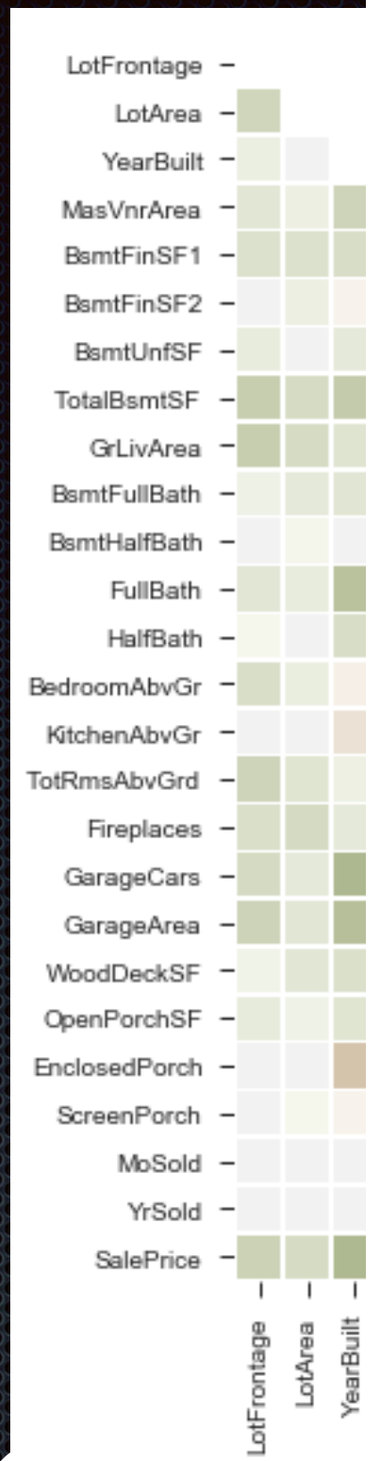


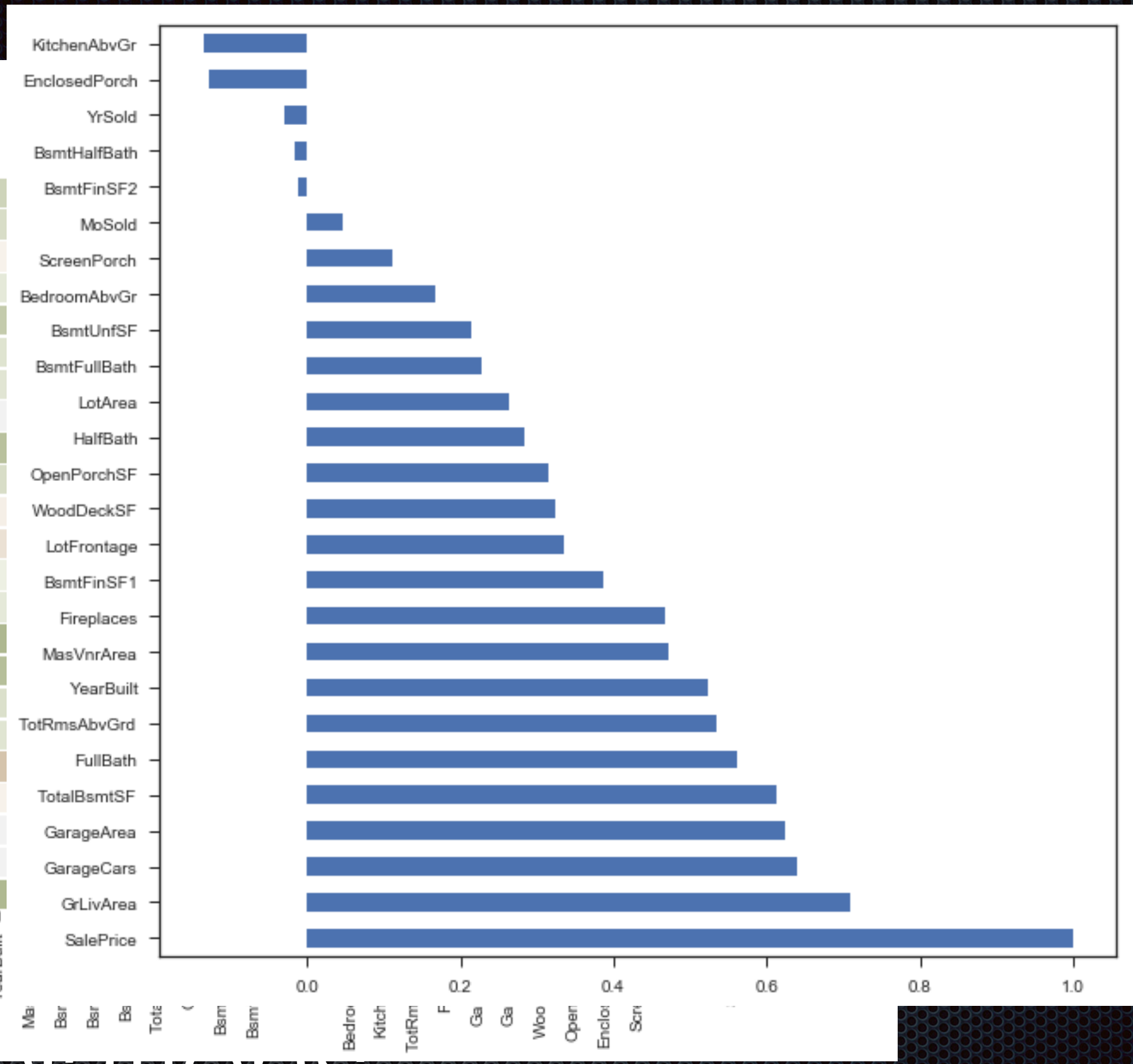
Correlations

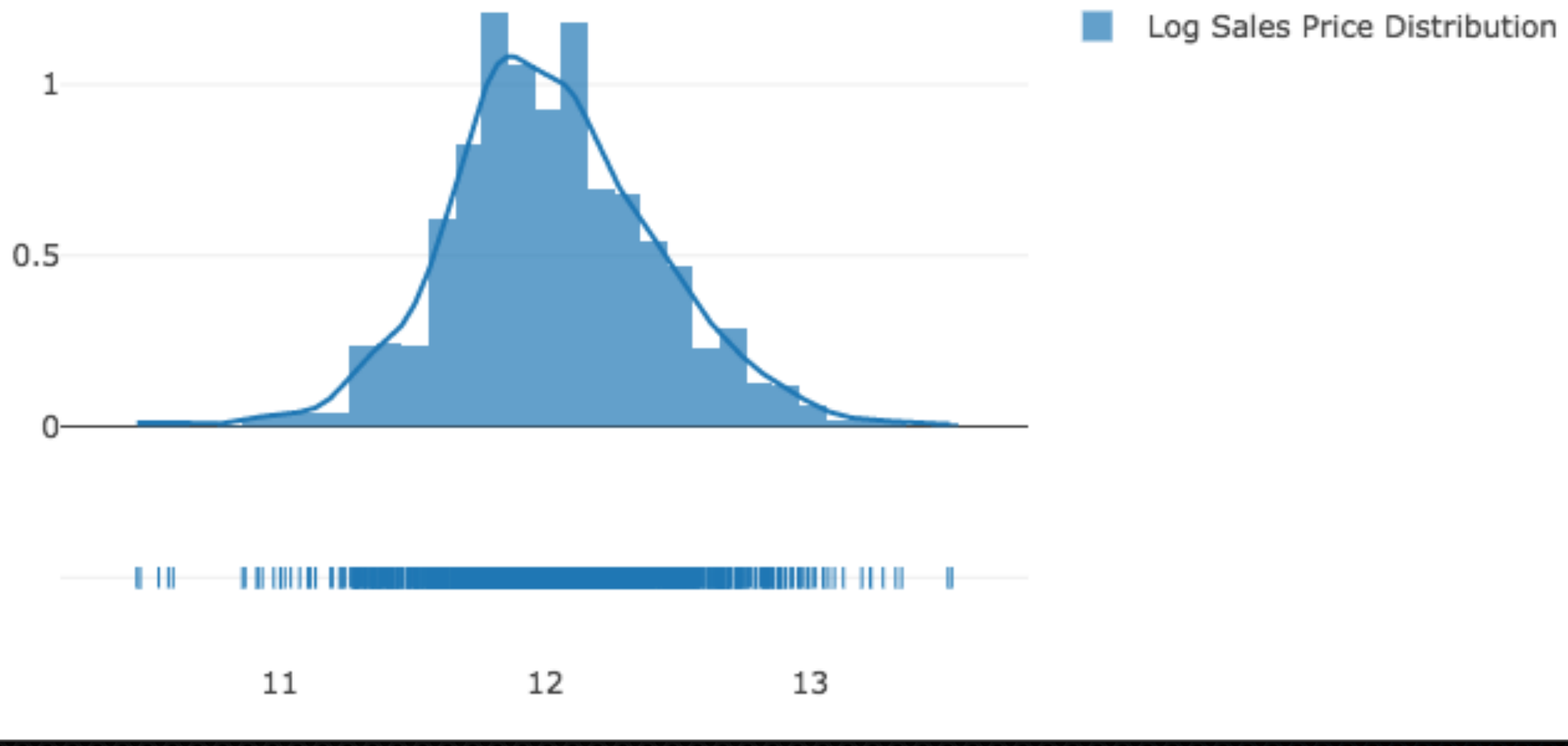
Correlations



Correlations



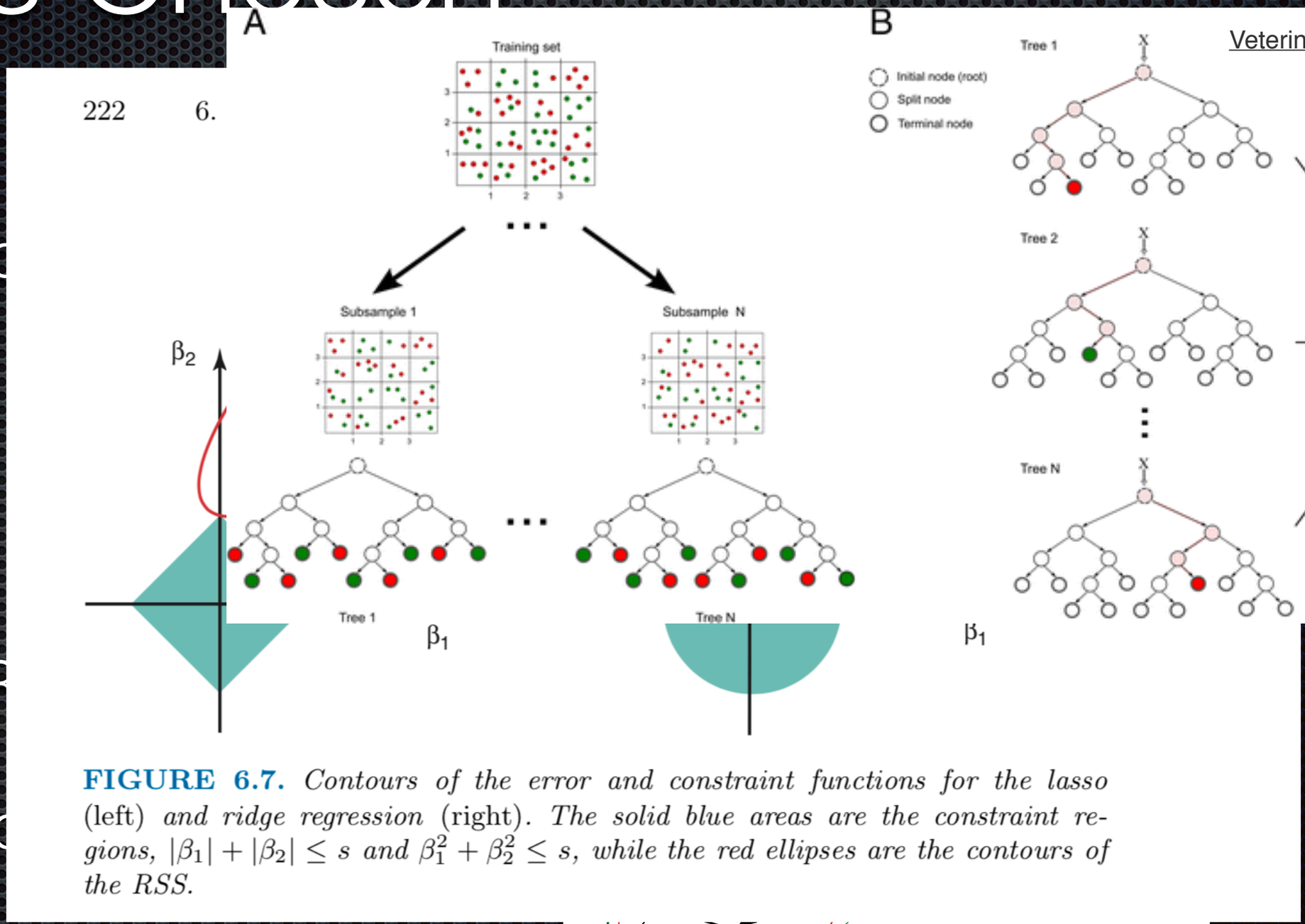




Normalizing Sales Price Distribution

Models Chosen

- Regularized
  - Ridge
  - Lasso
- PCA - MLR
- Random Fo
- XGBoost



222 6.

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Veterinary

Journey Through The Woods

- Decision Tree
- Tuned Decision Tree
- Bagged Tree
- Random Forest
- Tuned Random Forest
- Stochastic Gradient Boost

Bagged Tree

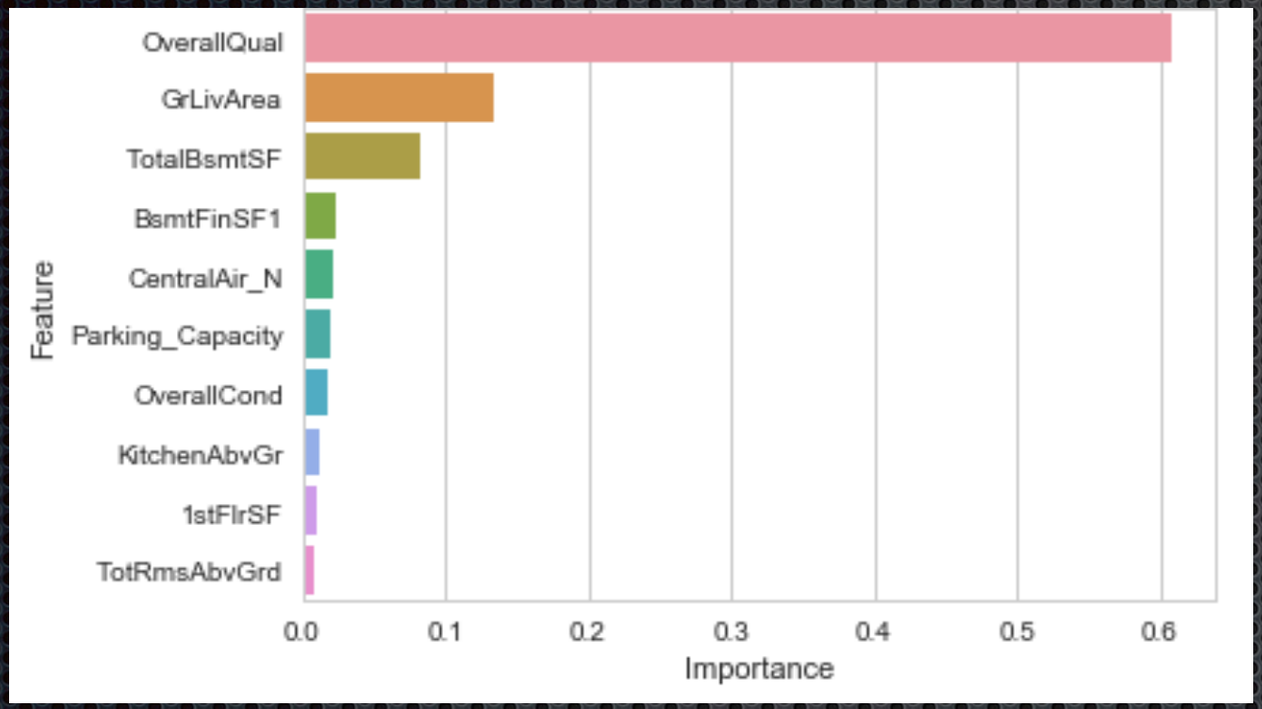
- **Parameters**
  - **Trees: 500**
  - **Max Samples: 783 (2/3rds of training set)**
- **R2**
  - **Training: 97%**
  - **Test: 87%**
- **RMSE**
  - **Training: .0745**
  - **Test: .1394**

First Attempt - Random Forest

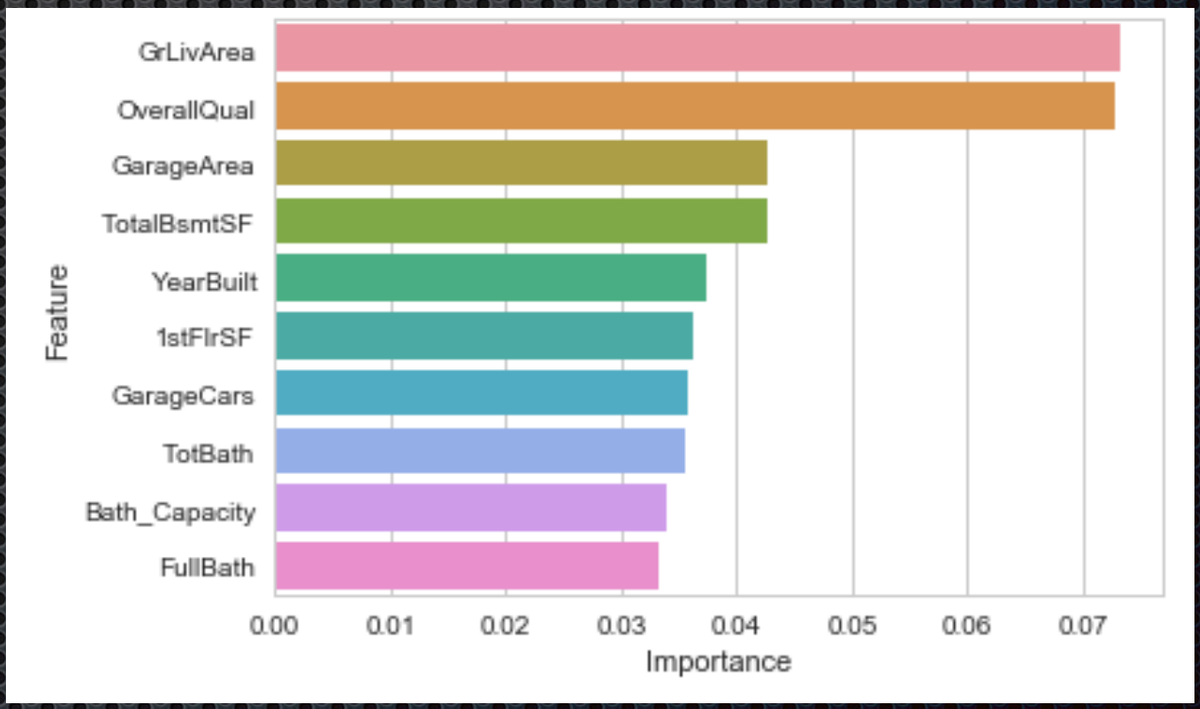
- **Parameters**
- **Trees: 500**
- **Max features: 17 (~ sqrt # of predictors)**
- **R2**
- **Training: 98%**
- **Test: 87%**
- **RMSE**
- **Training: .0538**
- **Test: .1397**

Feature Importance in Trees

Correlated (Bagged) Tree



Non-Correlated Random Forest



Next: On to Stochastic Gradient Boosting

Tuning with GridSearchCV

- **Best Parameters**
- **Max Features: 500**
- **Max Features: 36**
- **R2**
- **Training: 98%**
- **Test: 88%**
- **RMSE**
- **Training: .0526**
- **Test: .1361**

THOUGHTS: Again, not much changed for the Training and Test R^2 . What's going on?

Introducing Stochastic Gradient Boosting

- **Parameters**
- **Learning Rate: 0.1**
- **Subsample = 2/3**
- **R2**
- **Training: 96%**
- **Test: 92%**
- **RMSE**
- **Training: .0835**
- **Test: .1125**

Utilizing Tuned Stochastic Boosting

- **Best Parameters**
- Trees: 800
- Learning Rate: 0.04
- Subsample = 2/3
- **R2**
- Training: 99%
- Test: 93%
- **RMSE**
- Training: .0454
- Test: .1036

- Parameters
  - Learning Rate: 0.1
  - Subsample = 2/3
- R^2
  - Training: 96%
  - Test: 92%
  - RMSE
  - Training: .0835
  - Test: .1125