# House Prices: Advanced Regression Techniques

- A Machine Learning project by Confidence Squared

# Kaggle Competition

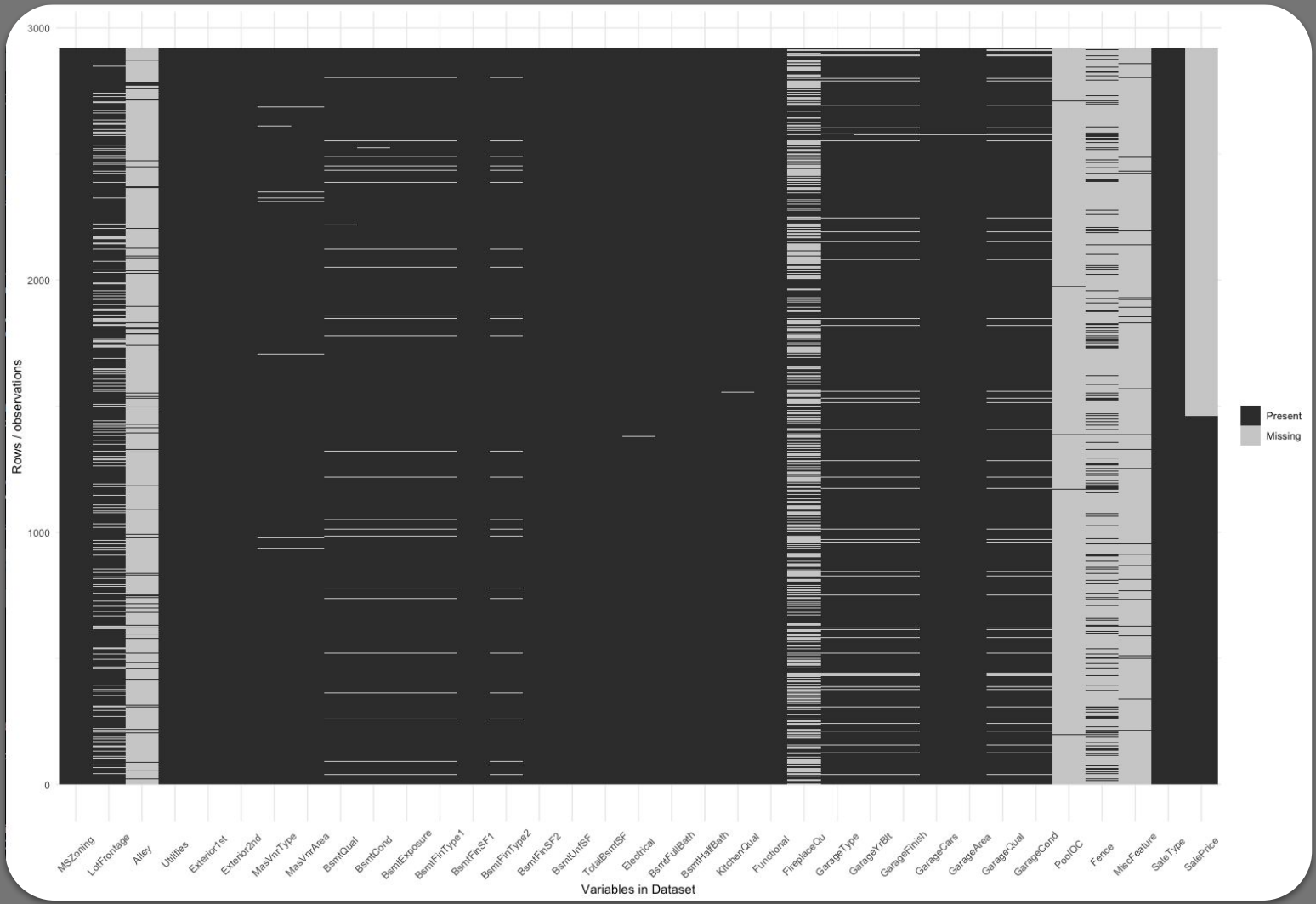
# Steps

- Preprocessing and EDA
- Feature Engineering
- Tree-Based Models
  - Multi-linear Regression
  - Ridge
  - Lasso
  - Elastic Net
- Tree-Based Models
  - Random Forests
  - Boosted
- Stacking

# Preprocessing and EDA

- ➔ Find missing values
- ➔ Systematically impute missing data according to each variable and data description
- ➔ Edit existing variables

# Missingness



# Missingness

```
[1] 2919   80
     PoolQC  MiscFeature    Alley       Fence    SalePrice  FireplaceQu  LotFrontage  GarageYrBlt
       2909         2814     2721        2348         1459         1420          486          159
 GarageFinish   GarageQual   GarageCond   GarageType     BsmtCond BsmtExposure     BsmtQual BsmtFinType2
        159          159          159          157           82           82           81           80
 BsmtFinType1   MasVnrType   MasVnrArea     MSZoning    Utilities BsmtFullBath BsmtHalfBath   Functional
           79           24           23            4            2            2            2            2
 Exterior1st  Exterior2nd   BsmtFinSF1   BsmtFinSF2   BsmtUnfSF  TotalBsmtSF   Electrical  KitchenQual
           1            1            1            1            1            1            1            1
 GarageCars   GarageArea     SaleType
           1            1            1
     MSZoning  LotFrontage       Alley    Utilities  Exterior1st  Exterior2nd   MasVnrType   MasVnrArea
     "factor"    "integer"    "factor"   "factor"   "factor"     "factor"     "factor"     "integer"
     BsmtQual     BsmtCond BsmtExposure BsmtFinType1   BsmtFinSF1 BsmtFinType2   BsmtFinSF2    BsmtUnfSF
     "factor"     "factor"     "factor"     "factor"     "integer"    "factor"     "integer"    "integer"
     TotalBsmtSF   Electrical BsmtFullBath BsmtHalfBath  KitchenQual   Functional  FireplaceQu   GarageType
     "integer"    "factor"    "integer"    "integer"    "factor"     "factor"     "factor"     "factor"
     GarageYrBlt GarageFinish   GarageCars   GarageArea   GarageQual   GarageCond       PoolQC       Fence
     "integer"    "factor"     "integer"    "integer"    "factor"     "factor"     "factor"     "factor"
     MiscFeature     SaleType    SalePrice
     "factor"     "factor"     "integer"
```

***There are 35 columns with missing values***

# 35 Missing Variables

Group 1:

- ✓ Alley
- ✓ BsmtQual
- ✓ BsmtCond
- ✓ BsmtExposure
- ✓ BsmtFinType1
- ✓ BsmtType2
- ✓ FireplaceQu
- ✓ GarageType
- ✓ GarageFinish
- ✓ GarageQual
- ✓ GarageCond
- ✓ PoolQC
- ✓ Fence
- ✓ MiscFeature

Not Group 2:

- ✓ Not Group 1

Individuals:

- ✓ MasVnrType
- ✓ MasVnrArea
- ✓ GarageYrBlt

# Editing Variables

Quality Levels:
Ex>Gd>TA>Fa>Po>None

- ✓ ExterQual
- ✓ ExterCond
- ✓ BsmtQual
- ✓ BsmtCond
- ✓ HeatingQC
- ✓ KitchenQual
- ✓ FireplaceQu
- ✓ GarageQual
- ✓ GarageCond
- ✓ PoolQC

Basement Finish:
GLQ>ALQ>BLQ>Rec>LwQ>Unf>None

- ✓ BsmtFinType1
- ✓ BsmtFinType2
- ✓ BsmtExposure
- ✓ MiscFeature
- ✓ Alley
- ✓ Fence
- ✓ GarageType
- ✓ GarageFinish
- ✓ Functional

Basement Exposure:
Gd>Av>Mn>No>None

- ✓ MasVnrType
- ✓ MSSubClass
- ✓ YrSold
- ✓ MoSold
- ✓ YearBuilt
- ✓ YearRemodAdd

Functional:
Typ>Min1>Min2>Mod>Maj1>Maj2>Sev>Sal

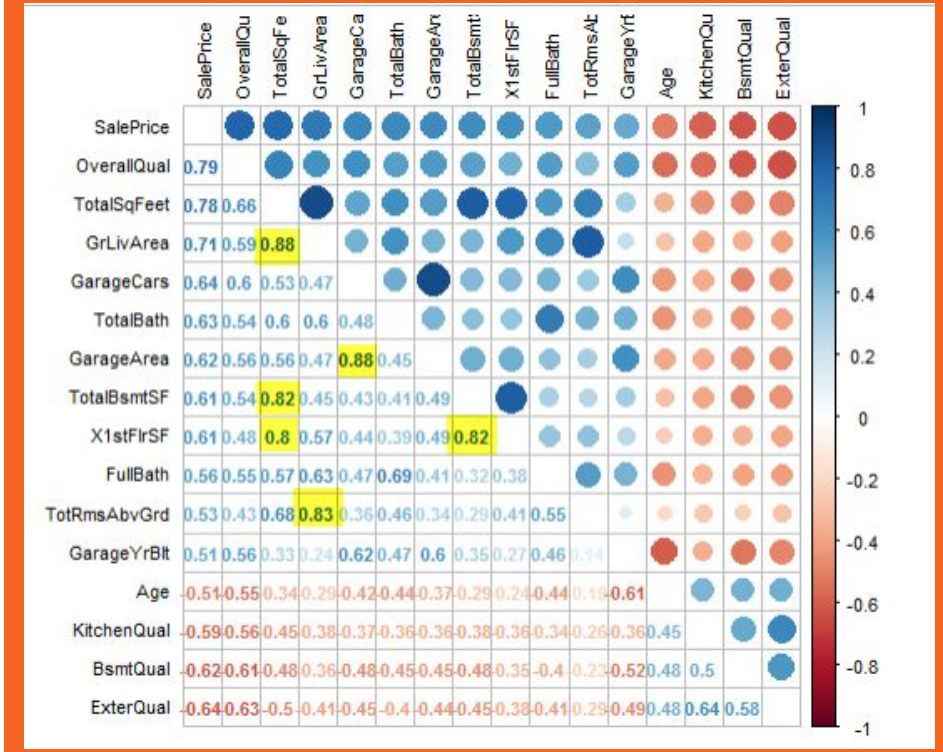
# Feature Engineering

Engineer new variables:

- Total square feet
- Total Porch square feet
- Total Bathroom #
- Whether house is new
- Whether house is remodeled
- Total age of house after build/remodeled

# Linear Models

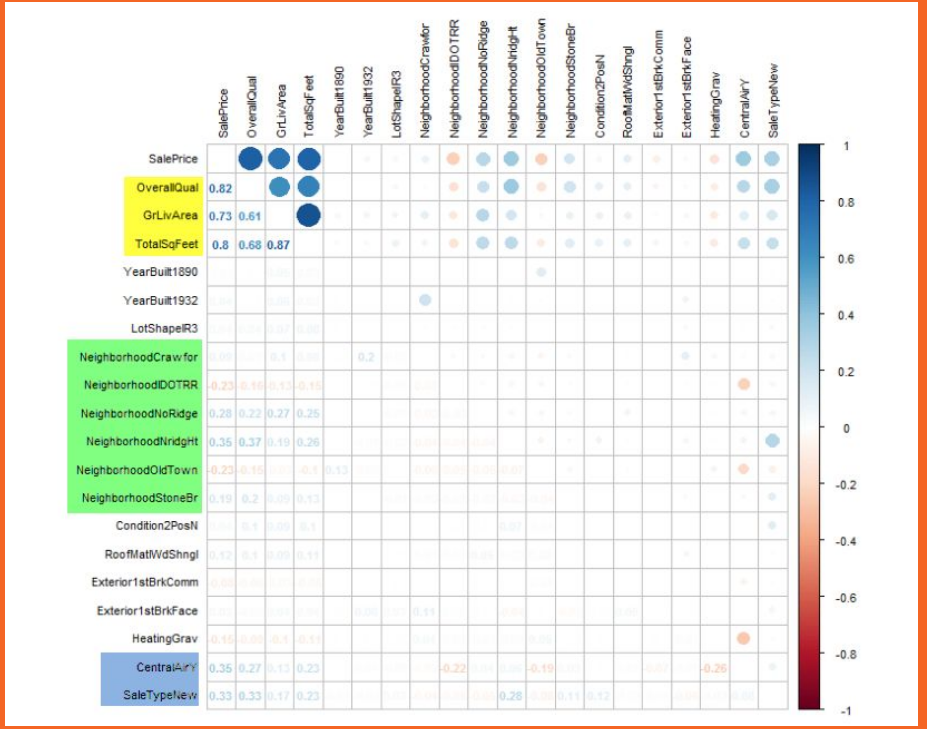
Check correlations with dependent variables, watch out multicollinearity



# Linear Model Results

| Method | RMSE | Challenge |
|---|---|---|
| MLR | 0.17 | Have to pick variables, cannot process 400+ variables, vector size 19G |
| Ridge | 0.1382 | Cannot reduce total numbers of variables, hard to interpret |
| Lasso | 0.1291 | Reduce variable # from 420 to 86. 2nd best RMSE |
| Elastic Net | 0.1290 | Similar as Lass, harder to interpret. Best RMSE |

# Linear Model Interpretations



## Trees: RF / Boosting

- While Decision Trees are not always an optimal approach, they are useful for their interpretability. By aggregating different approaches, we were able to improve our model's accuracy.
- RF approach averages over a collection of de-correlated Trees. With a big enough # of trees, we do not have to worry about overfitting. RMSE 0.1562
- Boosting
  - Trees are grown sequentially; each tree grown based on information from previously grown trees
  - Tuning Parameters
  - XGBoost (high accuracy, scalability, faster) RMSE 0.1294

# Conclusions

We decided to take a weighted average of our Lasso , Elastic and Boosted test predictions..

- Accuracy vs. Bias Tradeoff
- Hedge against risk of overfitting
- Kaggle Public Leaderboard results
    - Private Results