HouseX: Price Prediction Startup

1. Problem Statement

The client, HouseX, wants to create a tool for individual home buyers/sellers of residential real-estate to help predict the value of their property or prospective property. The small startup has a team of market researchers that have compiled the provided dataset but are looking for more in-depth analysis before moving forward. Specifically, the goals set out for us were as follows:

a. Produce a model with the best predictive accuracy in estimating the value of a property based on numerous key features.
b. Identify which key features positively or negatively impact house price and to what degree.
c. An added goal would be an additional service for homeowners to evaluate potential renovations and estimate the added value they might expect.
    i. I.e., adding on additional square footage to specific rooms or converting a study to a bedroom by adding a closet, etc.

2. Dataset – Cleaning & Wrangling

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=data_description.txt

Our dataset started with 79 independent variables with 1 dependent variables. Our dataset was part of a competition for housing price prediction and came with a train and test dataset, one with labels and one without. For the purpose of this capstone, we strictly worked with the train dataset, with labels, and later split into train and test sets so that we could quantify the accuracy of our models.

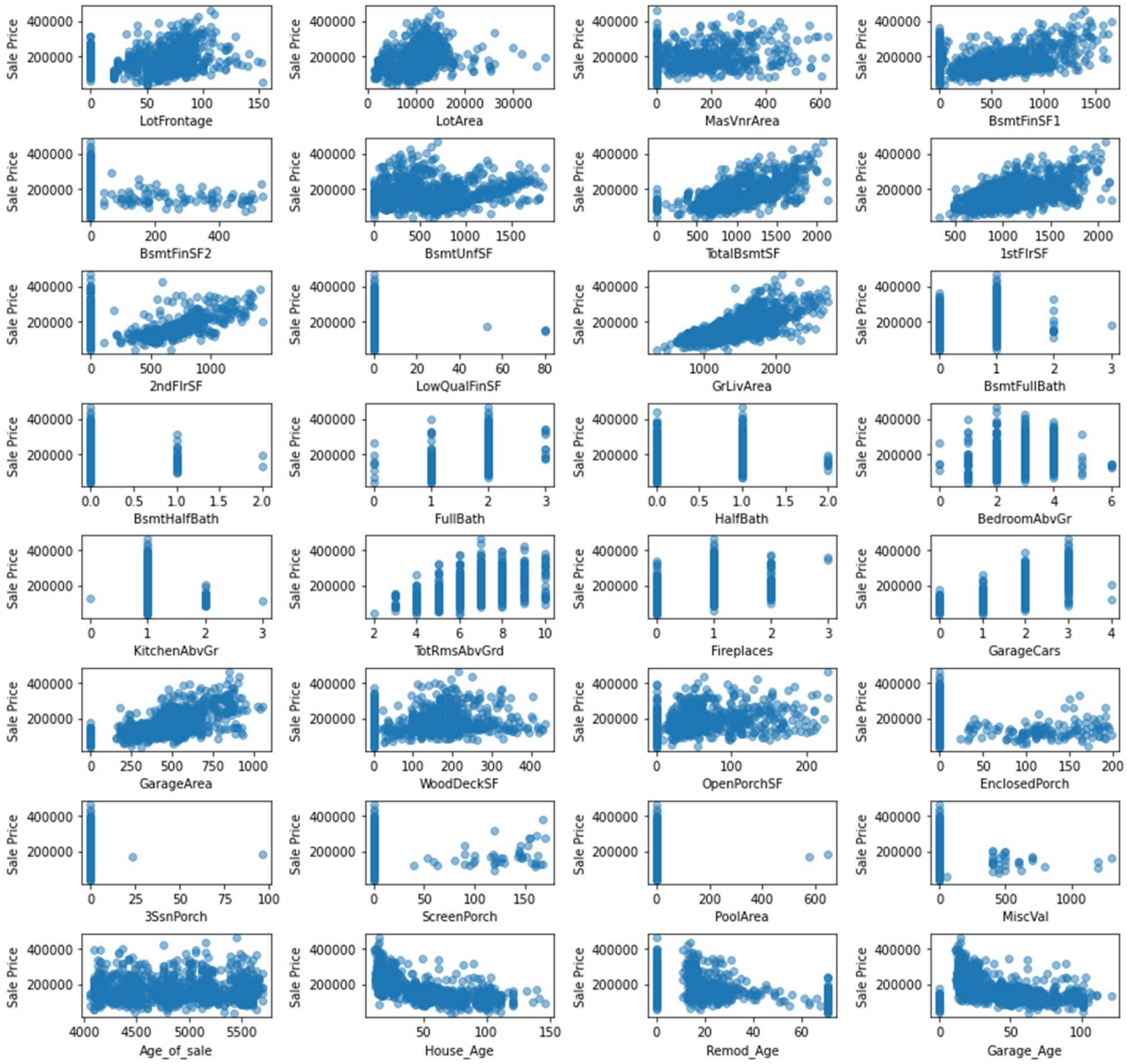
There were a number of outliers in our data which we decided to handle by removing any that were greater than 3 standard deviations from the mean via cutoff boundaries.

The dataset also had four date columns: year of sale, month of sale, year garage was added (if at all), and year of remodel. Using pd.date_time we changed all of these date columns to age, so that they could be handled more appropriately in our model.

3. Exploratory Data Analysis

Our features were categorical, ordinal as well as discrete. We split the features accordingly to get a better look at their relationships to one another and to our target variable, SalePrice.

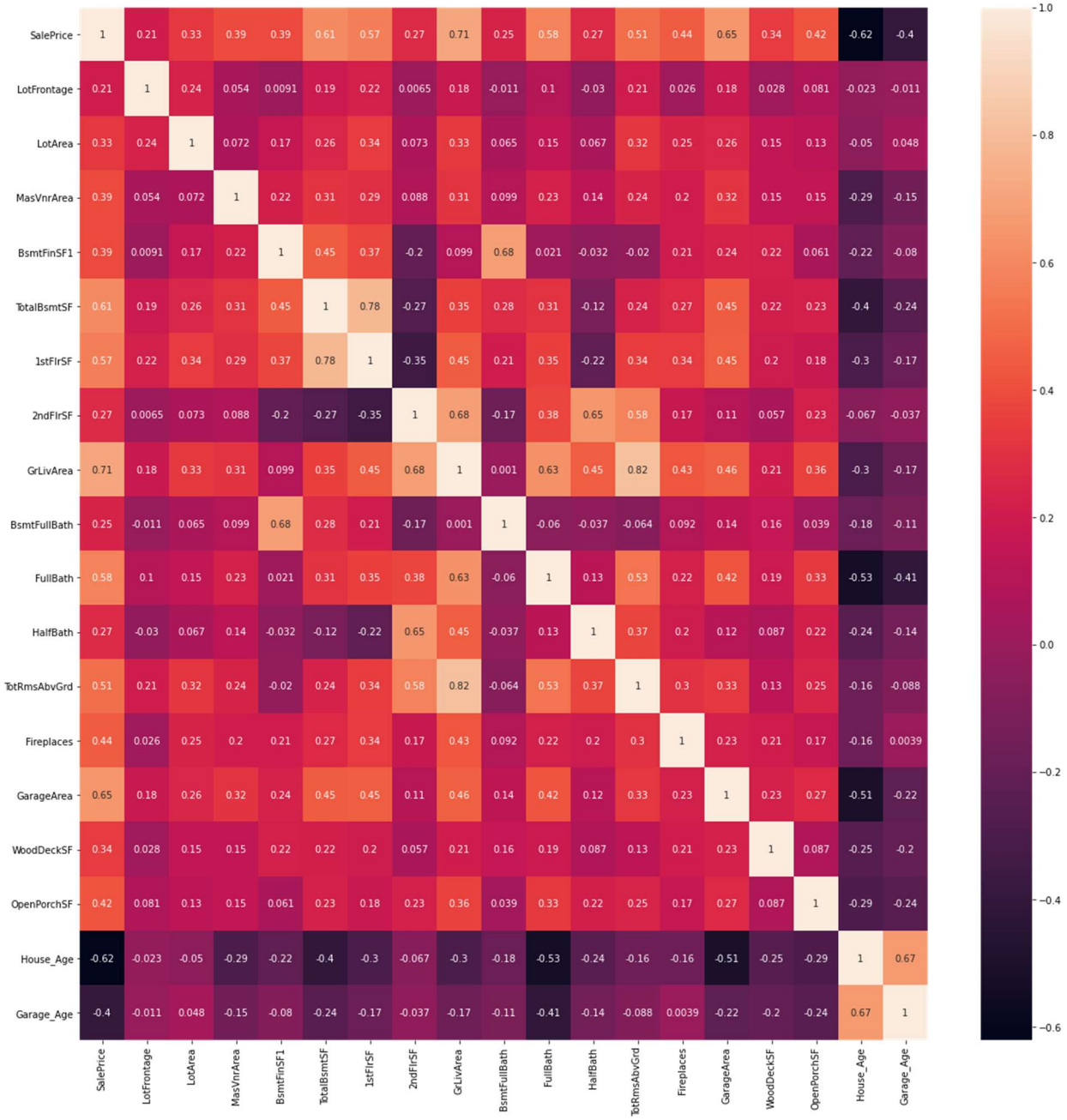
Clearly based on the sheer number of features our dataset had, it was clear we were going to have to narrow it down so that we didn't have a model that was too computationally expensive and also didn't overfit to our training data. We eliminated a number of features based on common sense, such as categorical ones that had only one of a certain category. We used strip plots and count plots to analyze our categorical variables. After narrowing it down some we produced their scatter plots to get a first look at their relationships to SalepPrice, seen below:



Seems DV's relationship to age-based variables is maybe logarithmic. Also worth noting is that extreme values of certain room attributes result in lower property values. Sales price declines with more than 1 half bathroom, 1 basement full bath, 4 above ground bedrooms, 1 above ground kitchen, 2 fireplaces, and 3 car garages. This relationship isn't entirely represented in TotRmsAbvGrd, however. Certain other intuitive relationships exist as well, such as finished versus unfished area. We would expect finished basements to generate more value, and the plots show that higher finished basement square footage lead to greater value than unfinished basements. A few interesting outliers seem to be present as well, such as the extreme value seen in BsmtFinSF1, 1stFlrSF as well as a house with 3 full baths in the basement and another (or the same) with 8 above ground bedrooms.

We dropped a number of features, like the age of sale and conditions of sale as we felt they weren’t relevant predictors for a home that has yet to be put on the market. We then plotted a

correlation heatmap to get a look at their relationship to saleprice as well as to one another, seen below:

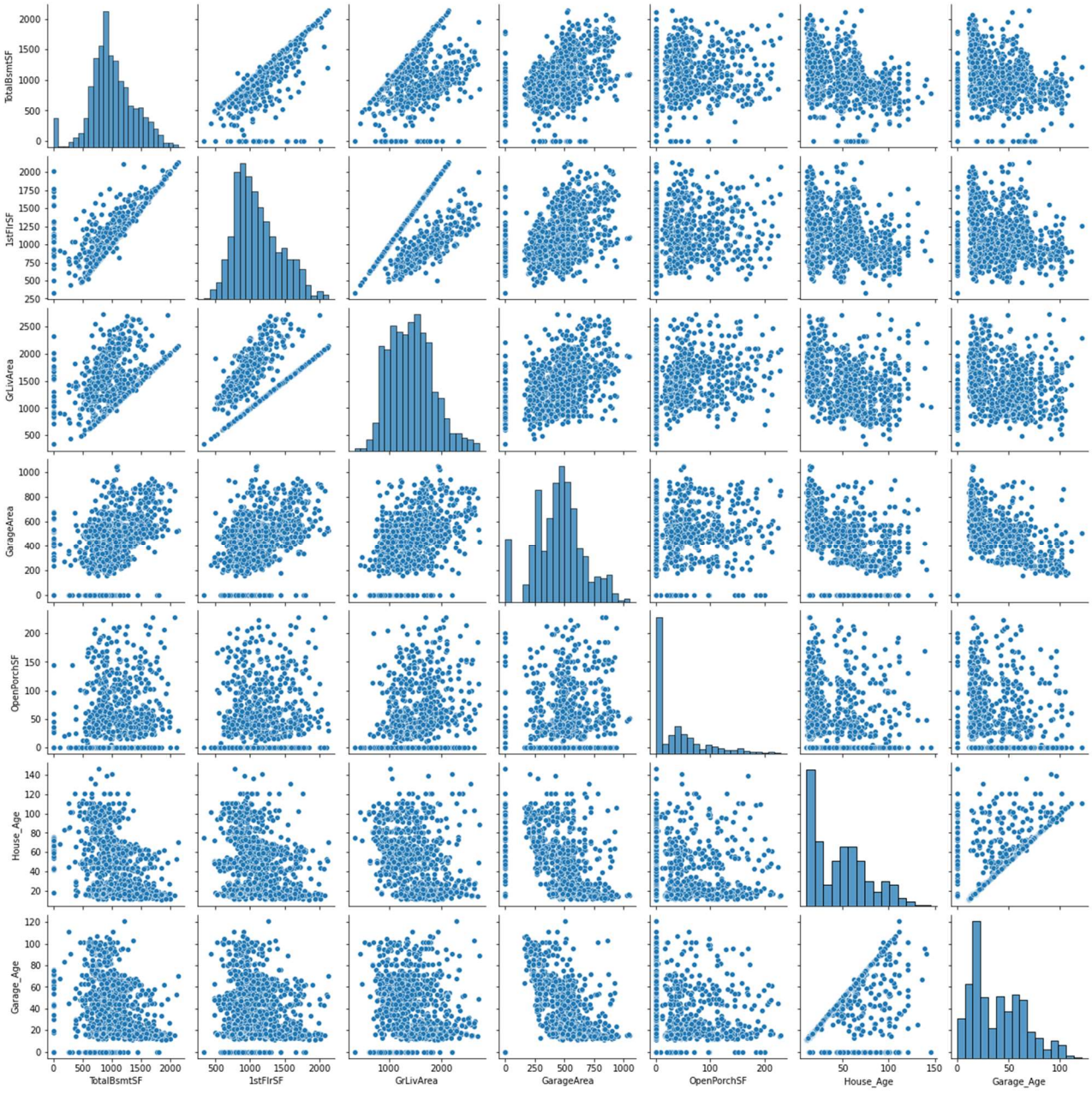


We have several correlated independent variables, so we are going to first make some observations regarding their relationships and then determine which we can drop before preprocessing. Things of note:

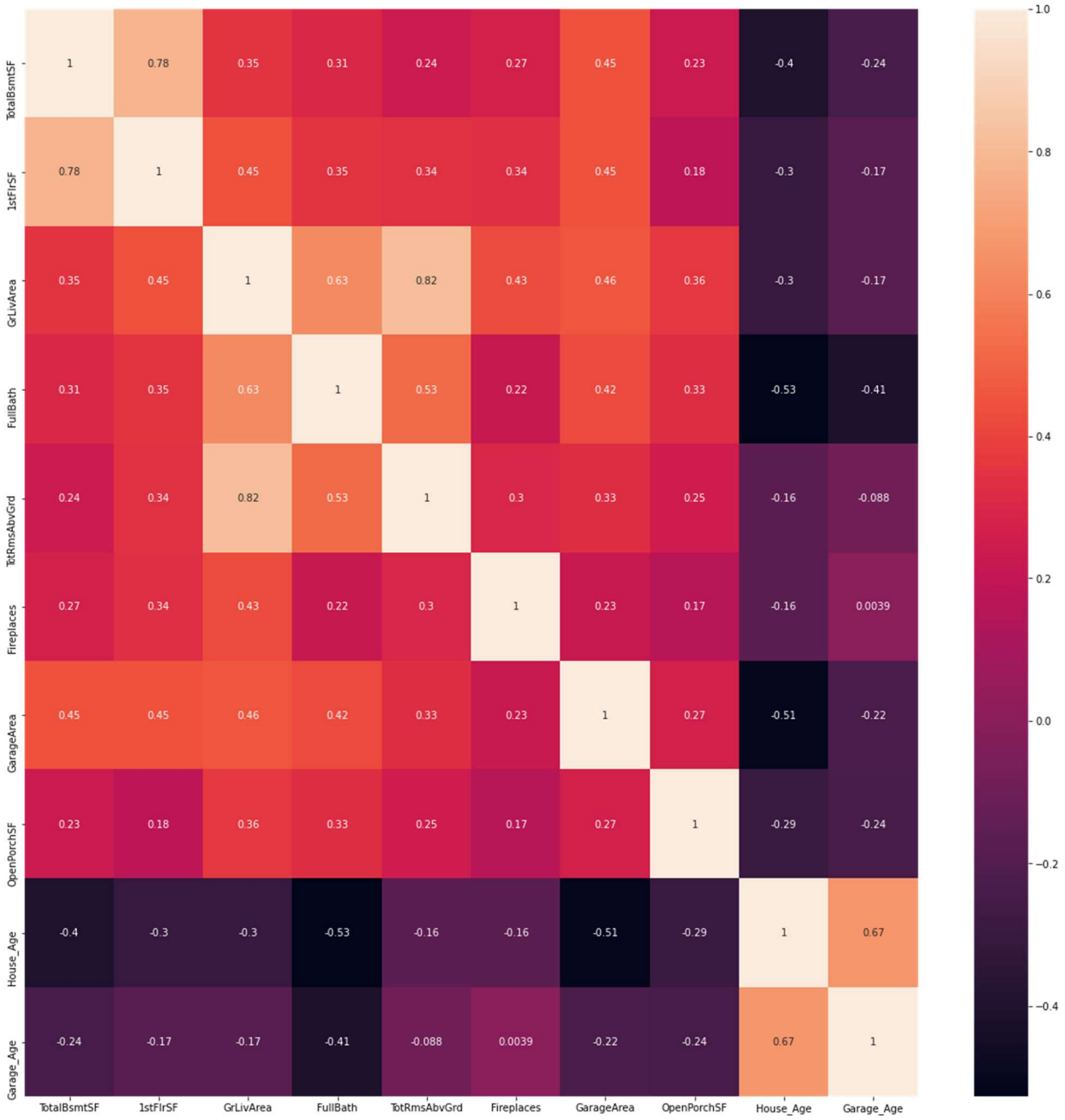
- Newer homes seem to have larger garages, or perhaps older homes are less likely to have a garage
- 1st floor square footage is highly correlated to total basement square footage, which makes sense as basements are generally only built under the foundation of the home
- Garage cars and garage area are clearly correlated because larger garages fit more cars.
- It seems enclosed porches are more common in older homes
- GrLivArea is highly correlated with basically all features related to rooms and above ground square footage

We went through and calculated the pearson correlation coefficient and removed features with a coefficient below 0.2.

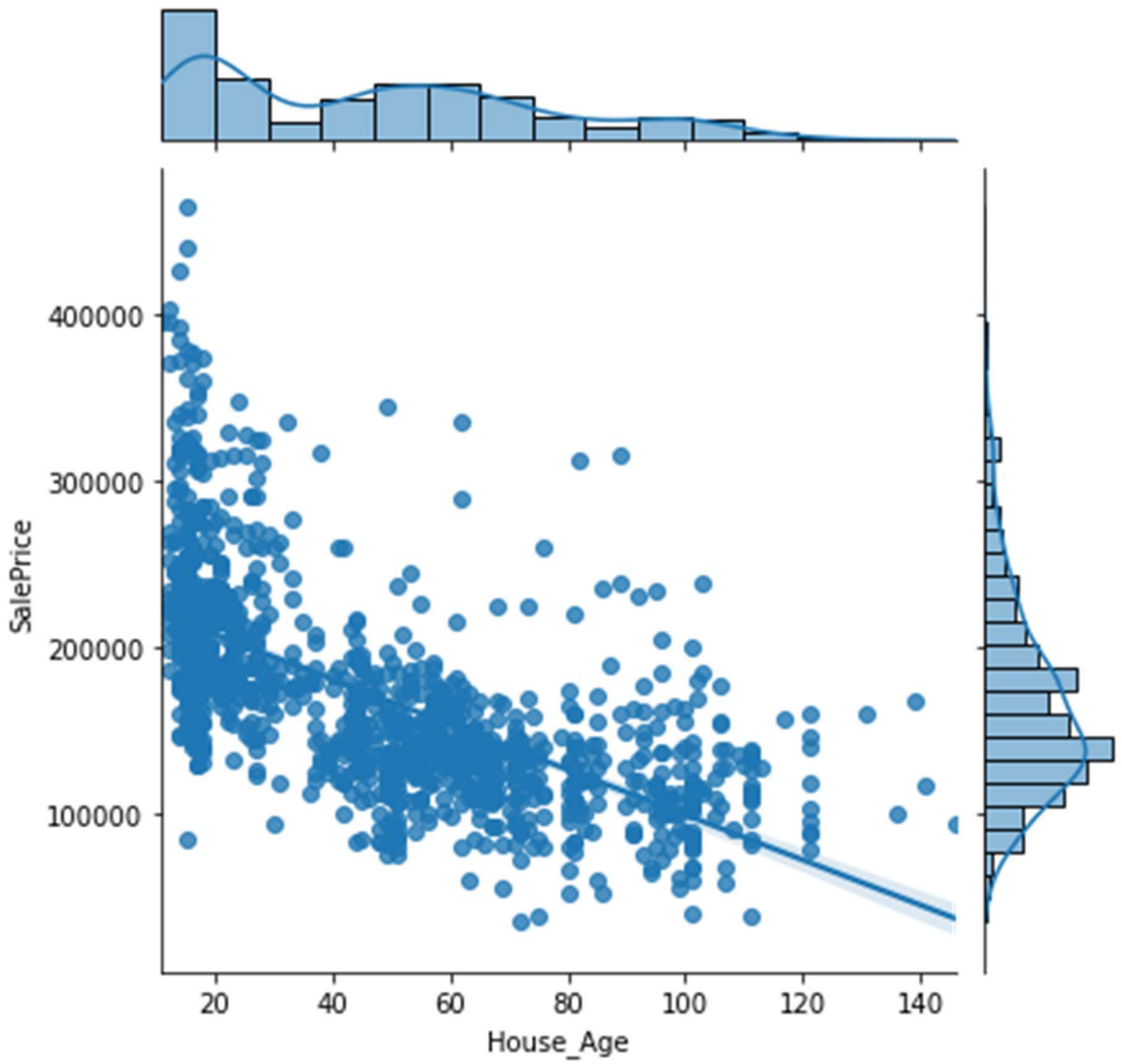
We then plotted the pairplots for the most correlated numerical features:



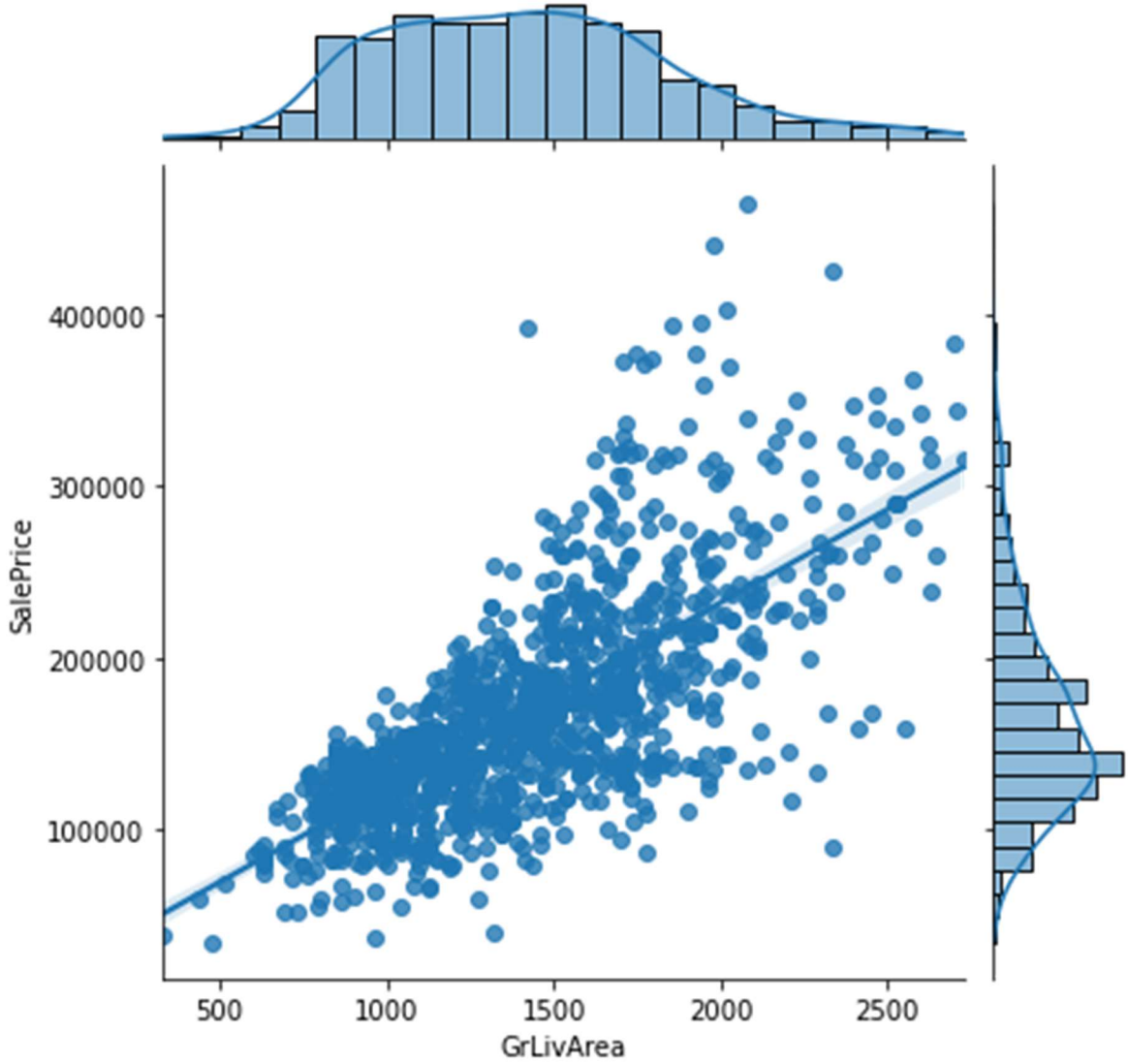
We also looked more closely at these highly correlated features once more in a heatmap:



There are some strong positive and negative relationships to SalePrice. The age of the home having a negative relationship makes sense as the newer the home, the higher the value:



Likewise, the greater the size of the home, the higher we would predict the value to be, and thus be positively correlated:

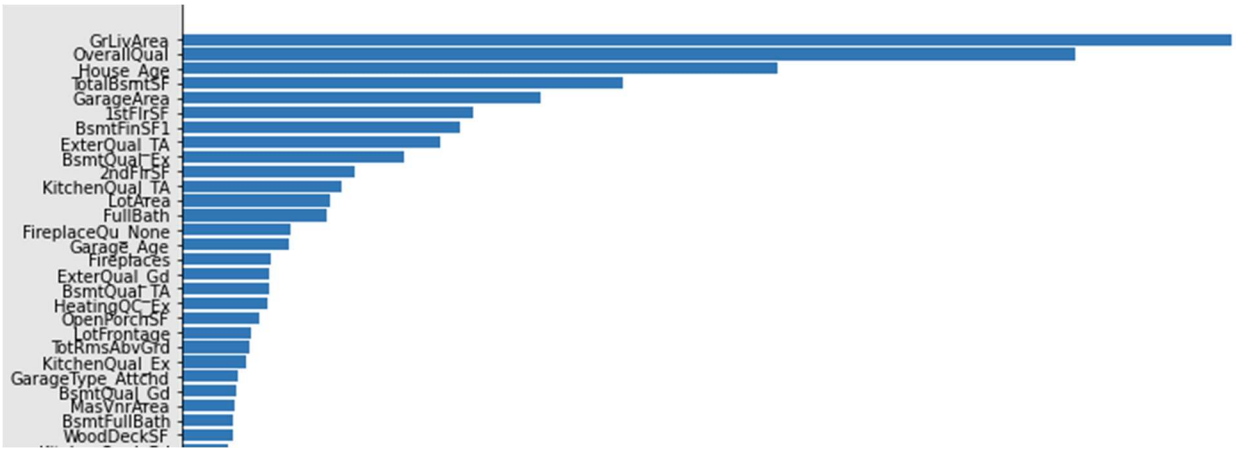


With our data relationships visualized and cleaned up a bit more, we were finally ready to move onto preprocessing and modeling.

4. Preprocessing and Feature Selection

Our preprocessing phase was simple, we separated our numerical and categorical features, scaled our numerical features using StandardScaler, and one-hot encoded our categorical features, and recombined them into our final preprocessed dataset.

Feature selection was going to be a huge portion of this project simply due to the large number of features we were dealing with. We looked at feature importance using RidgeCV, basic Random Forest and Random Forest with permutations. We did some hyperparameter tuning during these steps to get the best features we could before modeling. The plot below shows the feature importance ranking of our final RF permutation:



We selected the top 10 features that appeared in the top rankings across our feature importance models and ended up with the follow for our final model:

- House_Age
- GrLivArea
- Remod_Age_Avg
- TotalBsmtSF
- GarageArea
- 1stFlrSF
- 2ndFlrSF
- BsmtFinSF1
- LotArea
- FullBath

5. Final Model & Recommendations

After having narrowed down our features to the top 10 in terms of feature importance, we then ran four different regression models to determine which was best for predicting a home's sale price. We ran a basic linear regression model in addition to random forest, lasso and ridge regression models. After using these models to predict on our test set with 5-fold cross validation, we compared their performance based on their root-mean-squared-error, on the basis that the errors would have the same units and thus be comparable between models. The mean of the 5 fold cross validation RMSE for these models is as follows:

Linear Regression: 26,574.19

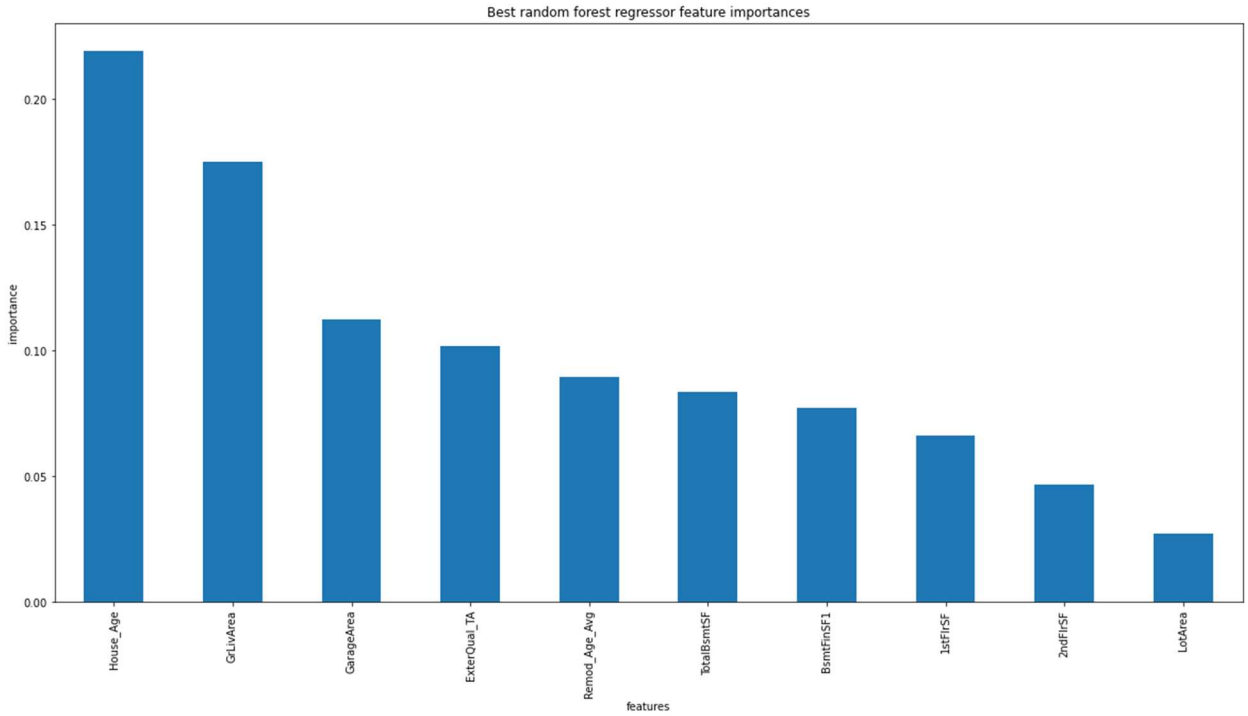
Random Forest: 23,516.36

Lasso: 26,549.76

Ridge: 29,872.18

Using RMSE as our performance metric, we have determined the Forest Regression model to be the preferred and best performing model. The linear regression and lasso regression models were very close, and the ridge model performed the worst of all four. Based on our chosen our model, the four greatest predictors for a home's value are the home’s age, the general living area square footage, the

exterior quality of the home, the size of the garage. The feature importance of our final model can be seen below:



Our model has a root mean squared error of roughly $27,345.54, so it can predict a home's value +/- 13,673 dollars.

6. Future Thoughts & Questions

We are quite confident in our final model's ability to predict a houses value but have a few questions for future models moving forward.

1. How much does highly correlated independent variables affect our model, and could we achieve a more accurate model by removing some of these? One example being 1st and 2nd floor square footage being correlated to GrLivArea square footage, as it is basically their sum. Also 1st floor square footage is highly correlated with basement square footage since most basements are only as large as the foundation of the home.
2. We noticed a strong correlation with house prices based on neighborhoods, but felt it wasn't general enough for the purpose of our model. However, in the future this could certainly be used in models targeting a more niche housing market, specific to a certain town or city.
3. Hyperparameter tuning of our random forest model was quite computationally expensive, but feel that with more time or computing power, parameters could potentially be tweaked a bit more.