Houses Price Prediction Regression Analysis

The median home price in the United States is $374,900 as of the second quarter of 2021. Home prices increased by 16.2% from 2020 to 2021. The median home price increased by 416% from 1980 to 2020. The Zillow Home Value Index puts the typical home price in the United States at $293,349. The increasing house price catches my attention, therefore I create a dataset and find out what variables affect the house price most. And I want to create a regression model which can predict the house price so that I could find out the undervalue house in the community.

1. Data

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

- Kaggle Dataset

2. Method

There are five main types of regression model used in practice today:

1. **Linear Regression:** A linear regression is a method to observe the relationship between a dependent variable, denoted as y, and one or more independent variables, denoted as x.

2. **Decision Tree:** The decision criteria is different for classification and regression trees. Decision trees regression normally use **mean squared error (MSE) to decide to split a node in two or more sub-nodes**. Suppose we are doing a binary tree the algorithm first will pick a value, and split the data into two subset.

3. **Random Forest:** Random Forest is a tree-based machine learning algorithm that leverages the power of multiple decision trees for making decisions. As the name suggests, it is a "forest" of trees! But why do we call it a "random" forest? That's because it is a forest of **randomly created decision trees**. Each node in the decision tree works on a random subset of features to calculate the output. The random forest then combines the output of individual decision trees to generate the final output. Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees.

4. **K-nearest neighbors:** KNN regression is a **non-parametric method** that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

5. **SVM: Support vector machines** (SVMs) are a set of supervised learning methods used for classification, regression and outliers' detection. The advantages of support vector machines are: Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples.

I choose the above five regression model to work with and generate the most appropriate predictive model. By comparing these model's mean absolute error, mean square error and accuracy, we can pick the better model and do a further evaluation.

3. Data Cleaning

In the dataset there are 1460 rows and 79 variables which related to the house sale price. I also had to clean & normalize the sale price.

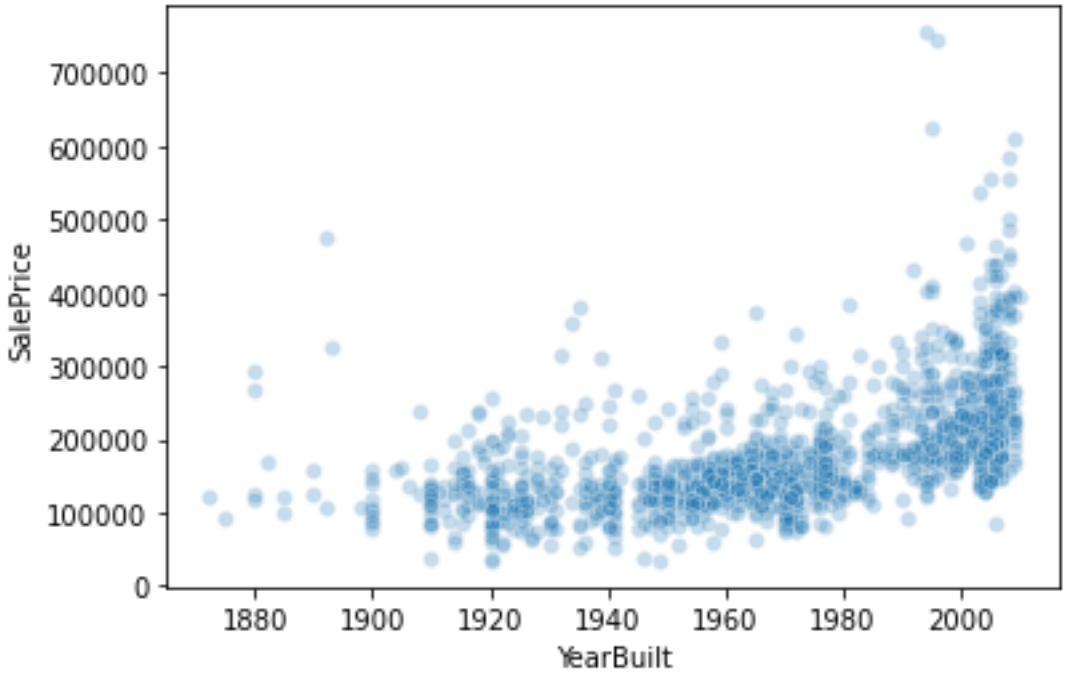
- **Problem 1:** This dataset has 79 variables related to the house sale price, too many variables will makes the model overfitting. **Solution:** Figure out the 20 highest correlation variables and drop other variables to avoid overfitting

- **Problem 2:** Multicollinearity. Multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. The variables are highly correlated will affect the

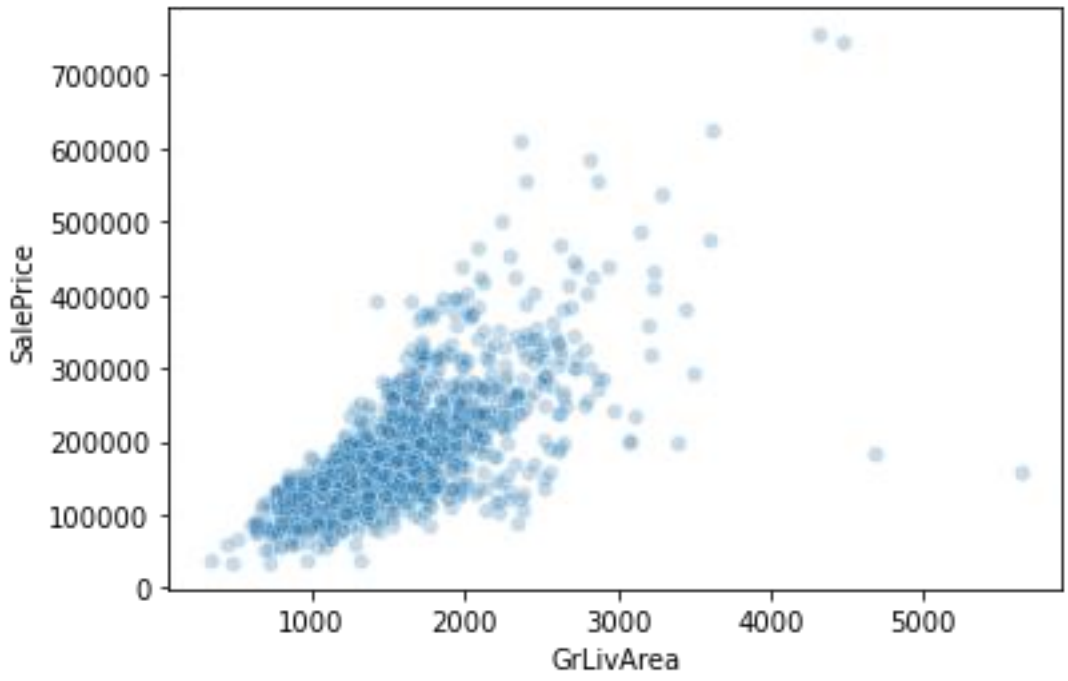
performance of the model. **Solution:** For the two highly correlated variables, remove either one variable and keep the other variable.

4. EDA

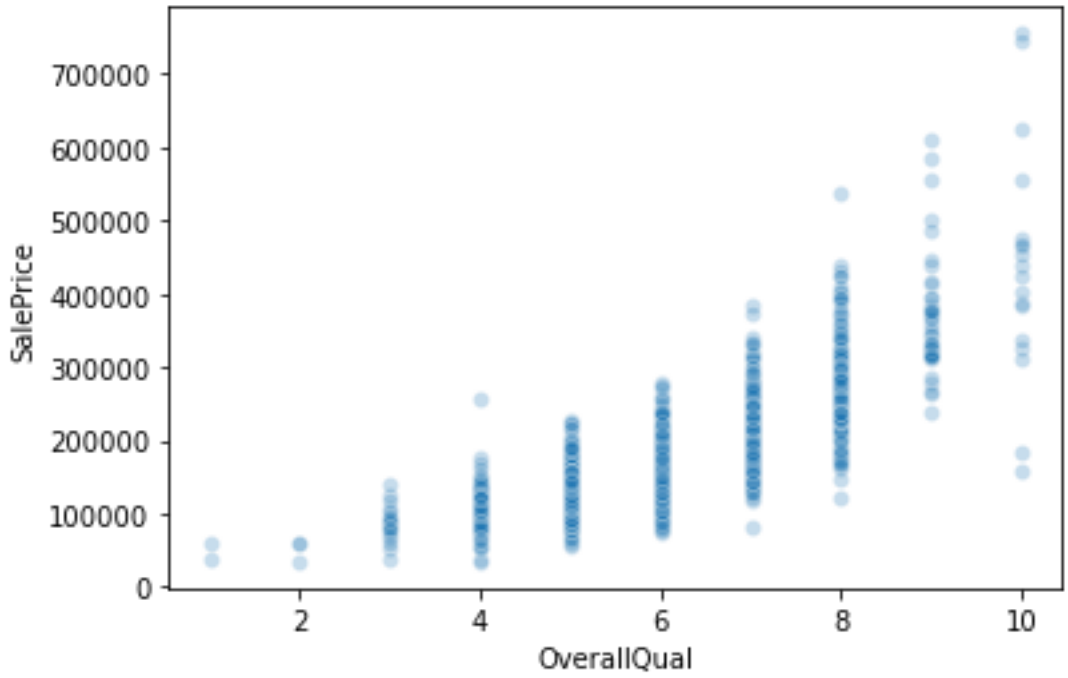
From the below graph, we can see that the newer the house, the price has a upper trend.



From the below graph, we can see that the bigger the house, the price has a upper trend.



From the below graph, we can see that the nicer the house, the price has a upper trend.



5. Algorithms & Machine Learning

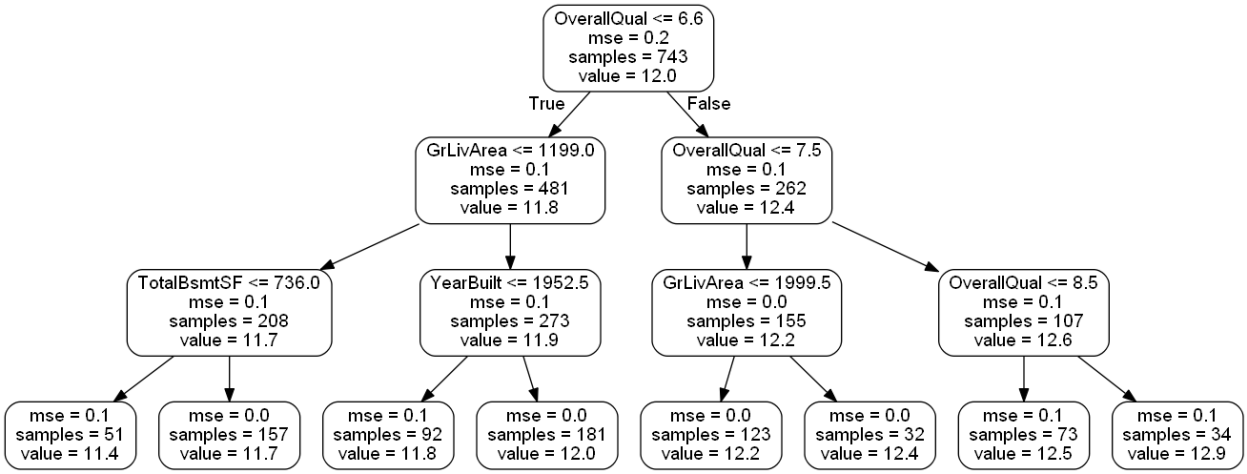
I chose to work with the Python surprise library scikit for training my predictive regression model. I tested the clean datasets on the 5 different algorithms, and decision tree regression and

random forest regression algorithm performed the best. It should be noted that this algorithm, although the most accurate is also the most computationally expensive, and that should be taken into account if this were to go into production.

```
LinearRegression
    MAE 0.10291878189465724
    RMSE 0.14888391131312223
    R2 0.8591102566487682
DecisionTreeRegressor
    MAE 0.00022486882573705542
    RMSE 0.0028646652757296854
    R2 0.9999478407177724
RandomForestRegressor
    MAE 0.04040169378708334
    RMSE 0.06066021857231781
    R2 0.9766120702582923
KNeighborsRegressor
    MAE 0.09713292928397774
    RMSE 0.13875533104064355
    R2 0.8776276797320525
SVR
    MAE 0.07664718635747168
    RMSE 0.1036357823880349
    R2 0.9317341818229748
```

NOTE: I choose RMSE as the accuracy metric, mean absolute error(MAE) and R square because they can give us a better picture which algorithm performs better. The RMSE is useful when large errors are undesirable. The smaller the RMSE, the more accurate the prediction because the RMSE takes the square root of the residual errors of the line of best fit. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The r-squared value provides an estimate of the strength of the relationship between the independent variables in the model and the dependent variable. The model performs better, the r square is closer to 1.

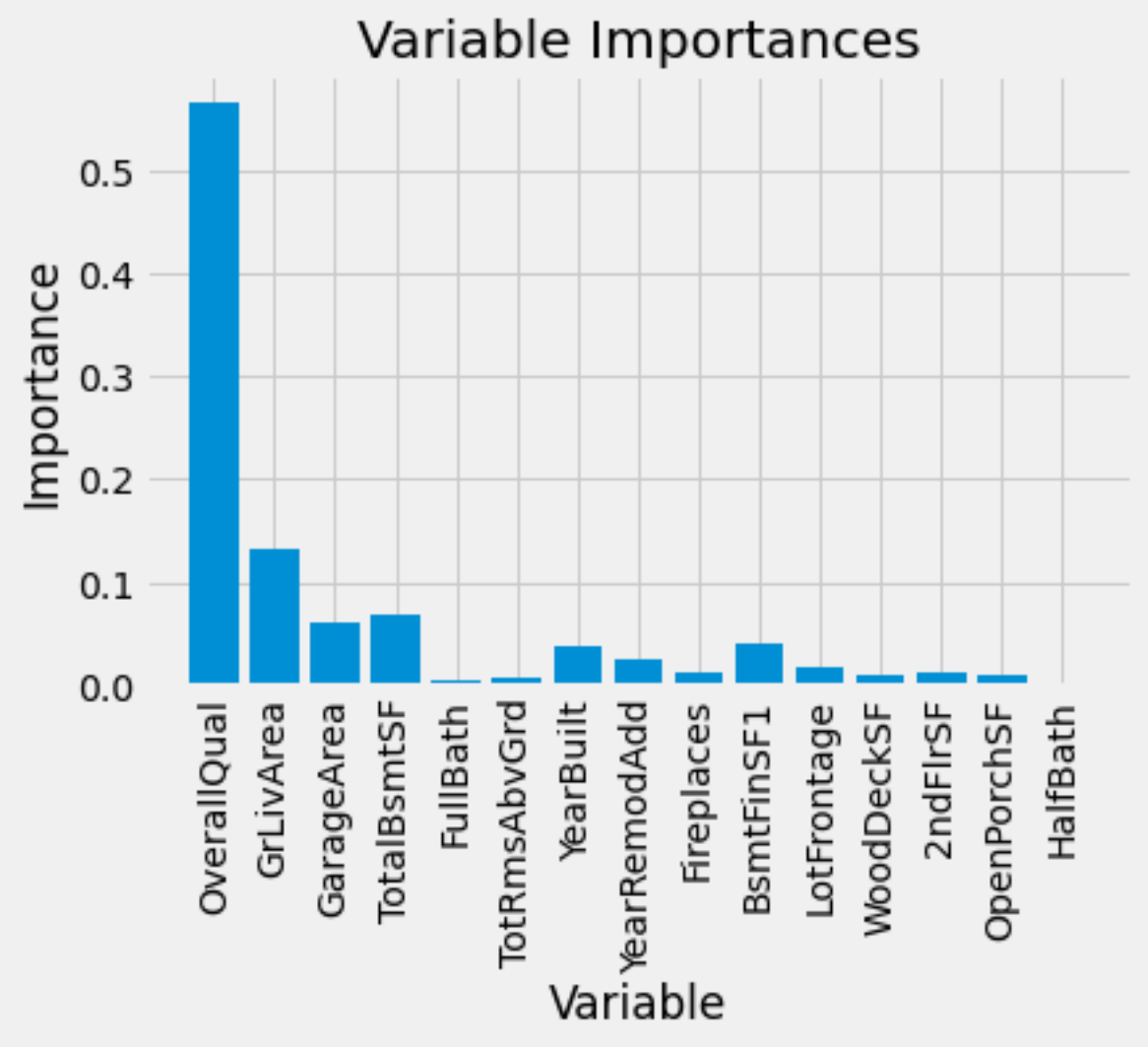
WINNER: Decision Tree and Random Forest



6. Feature Importance

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model, and the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on the problem.

```
Variable: OverallQual          Importance: 0.57
Variable: GrLivArea            Importance: 0.13
Variable: TotalBsmtSF          Importance: 0.07
Variable: GarageArea           Importance: 0.06
Variable: YearBuilt            Importance: 0.04
Variable: BsmtFinSF1           Importance: 0.04
Variable: YearRemodAdd         Importance: 0.02
Variable: LotFrontage          Importance: 0.02
Variable: TotRmsAbvGrd         Importance: 0.01
Variable: Fireplaces           Importance: 0.01
Variable: WoodDeckSF           Importance: 0.01
Variable: 2ndFlrSF             Importance: 0.01
Variable: OpenPorchSF          Importance: 0.01
Variable: FullBath             Importance: 0.0
Variable: HalfBath             Importance: 0.0
```



From the above graph, we can see that Overall Quality and Living Area have the 70% feature importance. They are the most important factors affect the house sale price.

7. Future Improvements

- In the future, I would love to spend more time on getting more up to date house price dataset to update the predictive model.
- The other way to improvement performance of the model is adjusting the hyperparameters. By adjusting the hyperparameters, we can get a better performance model.