**Predicting Housing Prices using Advanced Regression Techniques**

**Introduction.**

Ask a home buyer to describe their dream house, and they probably won’t begin with the height of the basement ceiling or the proximity to an east-west railroad. But this dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. The goal of this project is to use EDA, visualization, data cleaning, preprocessing, and linear models to predict home prices given the 79 features, describing (almost) every aspect of residential homes in Ames, Iowa, and interpret linear models to find out what features add value to a home. The data was originally taken from Kaggle. (<https://www.kaggle.com/c/house-prices-advanced-regression-techniques#description>)

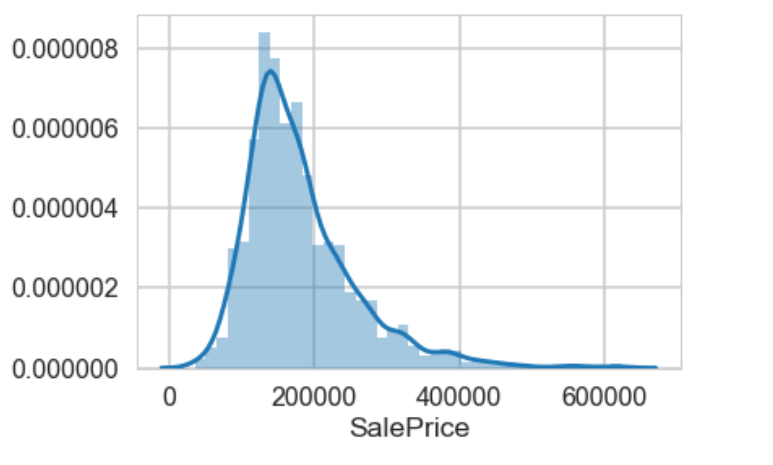
**Dataset description.**

As stated on the Kaggle competition description page, the data for this project was compiled by Dean De Cock of Truman State University for educational purposes, and it includes 79 predictor variables (house attributes) and one target variable (sale price). Data set describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values.

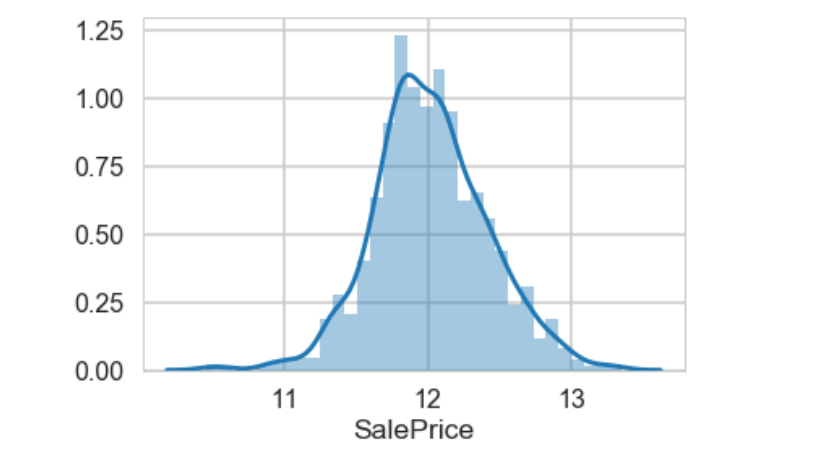
The variables included are basic characteristics that anyone wanting to buy a house would be interested in. For the most part, the different variables may be split up into specific groups. In general, the 20 continuous variables relate to measurements of area dimensions for each observation. These include, among others, the sizes of lots, rooms, porches, and garages. The 14 discrete variables mostly have to do with the number of bedrooms, bathrooms, kitchens, etc. that a given property has. There are several geographic categorical variables. The rest of the nominal variables identify characteristics of the property and dwelling type/structure. Most of the ordinal variables are rankings of the quality/condition of rooms and lot characteristics.

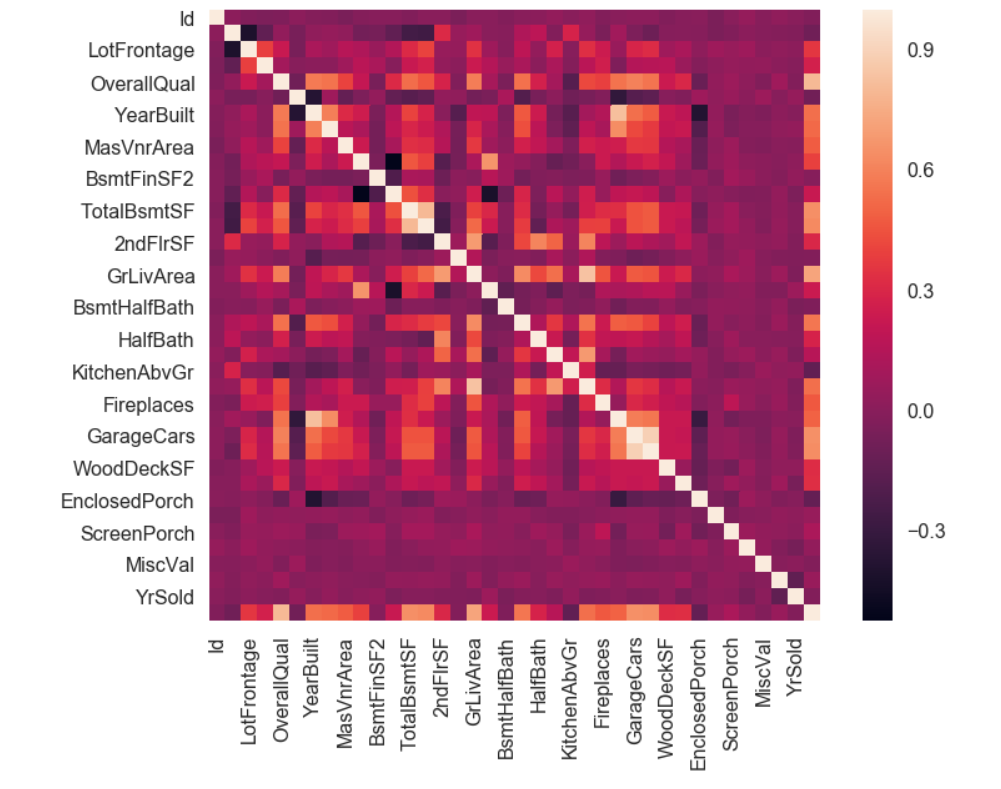
**Processing the Data, EDA.**

### As our response variable, Sale Price, is continuous, regression models will be used. One assumption of linear regression models is that the error between the observed and expected values (the residuals) should be normally distributed. Violations of this assumption often stem from a skewed response variable. Sale Price has a right skew, so we log + 1 transform it to normalize its distribution.



Log+1 transformation does pretty good job.



To explore further we’ll use visualisation methods to analyze the data better. The heatmap is the best way to get a quick overview of correlated features.

The most correlated features with sale price are: OverallQual, GrLivArea, GarageCars, TotalBsmtSF, GarageArea, 1stFlrSF, FullBath, TotRmsAbvGrd, YearBuilt, YearRemodAdd.

Pair plot between 'SalePrice' and correlated variables for better understanding.



The following pre-processing steps were taken:

1. **Removing outliers:** Removing the outliers may have improved predictions for average-priced homes, since outliers would have the most impact on the fit of linear-based models. Viewing the relationship between Above Ground Living Area and Sale Price, we noticed some very large areas for very low prices. The solution was to remove these observations, because they might increase our models' errors. Values with Above Ground Living Area > 4000 were removed as outliers.
2. **Dealing with missing values:**

Machine learning algorithms do not handle missing values very well, so we need to take care of them. Below I describe examples of some of the ways I treated these missing data.

Linear feet of street connected to property: Linear feet of street connected to property. It has quiet low correlation with the price so we can remove it or put mean instead of missing values.

Alley: Type of alley access to property. Many missing values, maybe these properties just don't have an alley access. Was changed to None.

Masonry veneer type/Masonry veneer area in square feet: both have 8 values missing, probably they are the same ones. The most frequent one was set.

Basement Variables: A number of variables in connection with the basement. About the same number of missing values. However, there are two basement-related variables without missing values. After comparison we can see that there is no basement for all missing values.

Electrical: Just one missing value - here just impute most frequent one.

Fireplace quality: I assume the properties with missing values just don't have a fireplace. There's also the variable Fireplaces (without missing values) - same as with basement -no fireplace.

Garage Variables: 81 missing in these columns. However, there are some Garage-related variables without missing values: Size of garage in car capacity, Size of garage in square feet, as we can see - no garage for missing values.

Pool quality - probably no pool - checked against Pool area in square feet (which has no missing values), no pool for all missing values.

Fence: Too many missing values - probably no fence, insert None.

**3. Created dummy variables for the categorical variables.**

**4. Split the data into a training set and a test set.**

**5. Scaled the data.**

**Modeling.**

Now that we have prepared our data set, we can begin training our models and use them to predict Sale Price. Results table is shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **ElasticNetCV** | **RidgeCV** | **LassoCV** | **RandomForest** | **Xgboost** |
| **RMSE** | 0.10502604065106295 | 0.11988702432763035 | 0.10694717410861436 | 0.3025445104080258 | 0.133160366862 |

The information in the table represents best results for each model on test data.

Probably, various linear models perform better than RandomForest model. We suspect this had a lot to do with the data itself, Sale Price (or rather, log(Sale Price)) likely has a relatively linear relationship with the predictor variables.

For most models the following features have the strongest impact on Sale Price predictions: above grade (ground) living area in square feet, lot size in square feet, total square feet of basement area, overall material and finish quality, overall condition rating, size of garage in square feet, remodel date, original construction date, unfinished square feet of basement area, year garage was built.

### **Conclusions**

### The objective of this project was to build models to predict housing prices of different residences. Our best model resulted in an RMSE of 0.1050, which translates to an error of about $9000 for the average-priced house.

The variables seen as most important or as strongest predictors through our models were those related to square footage, the age and condition of the home, the neighborhood where the house was located and the year the house was sold.