# Housing prices: Regression Techniques

Hey fellow Data enthusiasts, today I would like to give you a glimpse at my take on solving a regression problem to predict housing prices. This is yet another of those classic datasets that Data Scientists of all caliber work on with varying levels of modeling approaches.However what I find interesting is the level of detailed effort you could put on this data to experiment what works and what doesn’t.

## Preview: The data

The dataset consists of about 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, and the variable to predict is the price of each house.I used Dataiku for data cleaning and feature engineering which reduced the time spent on each of the aspects by atleast half of what I would have spent through traditional python programming.

So without further delay let’s jump right into the pool of data!

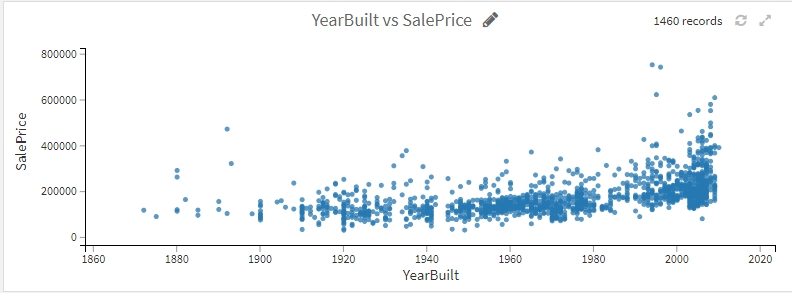
## Step 1: Explore the data

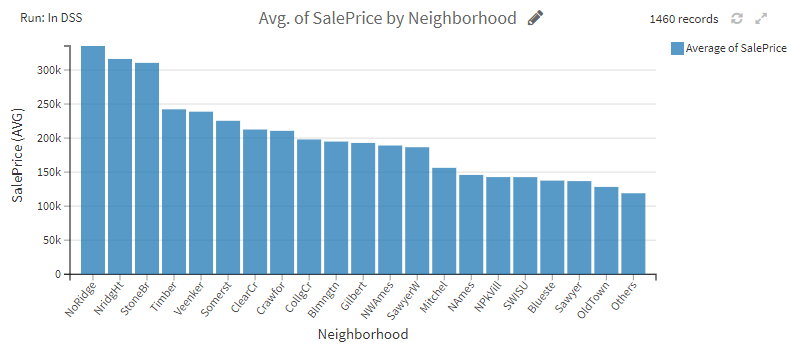
What I mean by explore is just spend some time understanding what the variables mean rather than getting right into statistics or graphs.

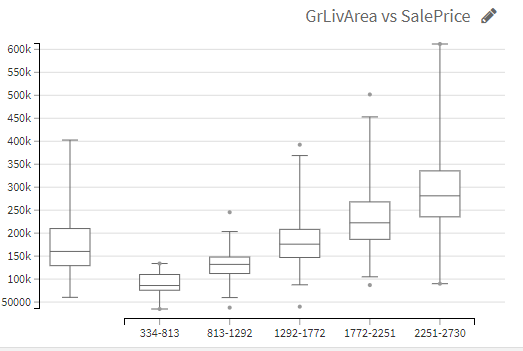
So what did I observe:

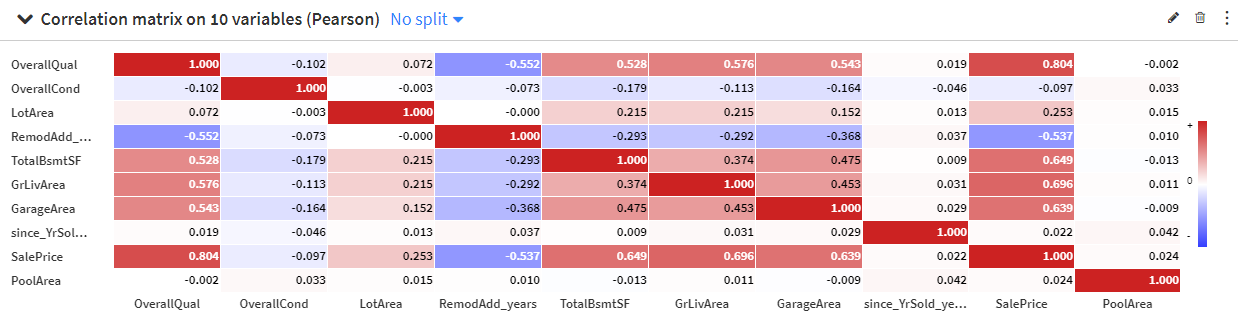
* There are a couple of similar features describing Lot such as LotFrontage, LotArea, LotShape, LotContour
* OverallQual and OverallCond are ratings for each of the houses on a scale of 1 to 10
* And then there are dates of “YearBuilt” and “YearRemodelled”
* Couple of variables are categorical and describe say the type of foundation or the basement type etc.
* Then comes utilities like CentralAir, Electrical, Heating etc.
* And finally lots of columns describing the number of full/half bedrooms and bathrooms, garage, kitchen and living area etc.

## Step 2: Exploratory Data Analysis









So at this stage, I wanted to make a couple of assumptions regarding what could affect sale price and how. This made me check whether my assumptions had a pattern or not using graphs and other visual medium.

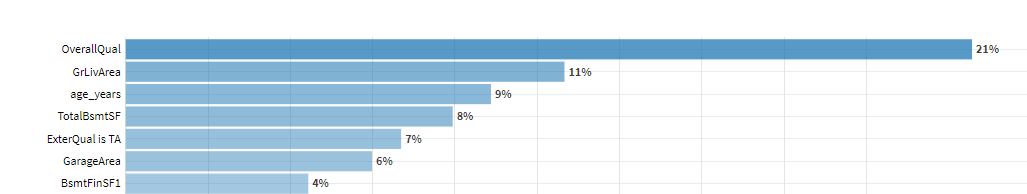
* It turns out year built did affect the price of the house as logically newer homes are costlier to build and hence sold at higher rates. (but note the exception of outliers here)
* My assumptions regarding if some neighbourhoods are costlier than other is kinda right but the chart gave me the idea to better group them since the differences between some of them are pretty small
* Box plots serve as a great way to identify outliers which might skew your results vastly. As an example the relationship between “greater living area” and “sale price” shows how some houses might be priced lesser than average even within the same area range.
* A correlation matrix also helped me to judge the strength of relation between variables with each other along with sale price.

## Step 3: Data Preprocessing

* Removing outliers: I removed the outliers from the range of 5\*IQR (classic Tukey method takes 1.5 \* IQR) since the original data had just 1460 rows and I wanted to preserve as much data as possible.
* Imputation of null values: I filled in missing numerical values such as LotArea with the mean value and most repeated values for categorical variables like Pavement.
* One hot encoding for categorical variables with less than 4 categories.
* Data conversions of years into age since it gave a better continuous flow and cleaning misspelled data types such as ‘Twnhs’ and ‘Twnhse’ for “Townhouse” in BldgType.

## Step 4: Modelling

* Considering it a regression problem I modeled the train data with Random Forest, Ridge (L2) regression and XGBoost techniques. While L2 produced similar results as Random Forest I found that with hyperparameter optimization Random Forest gave me a stronger accuracy and less error within a reasonable time limit (sorry XGBoost!)
* With a decent number of trees and max depth of 19 the model gave an R2 score of 0.895 and RMSLE of 0.136 which was pretty good.
* Further the variables importance showed that the Overall Quality of the house, Greater Living area size and how old the home was mattered more then rest of the factors.



## Step 5: Interpreting scoring and other factors

* Random forest model helped me avoid over-fitting of training data which paired well with hyperparameter optimization and resulted in confidence of my results.
* The effect of this is that the model introduces more bias , but becomes more statistically stable and invariant. In other words, it prevents us from overfitting and is better able to generalize to new data.

## Conclusion

Predicting housing prices is a great tool for many realtors and even seasoned home buyers since it will give that ball park price with reasonable judgement. Hence my approach involved around more explainable models than black box ones with higher accuracy. Something to note is how the pecularity of a certain house (antiquity or exquisite architecture) could result in outliers and hence drive the prices to skyrocket.

There are many more areas to explore and techniques to experiment with which could always result in accurate predictions. I continue to hope learning deeper insights with repeated problem solving even on the same data.

