Capstone Project 1: Milestone Report

1. Problem statement

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.

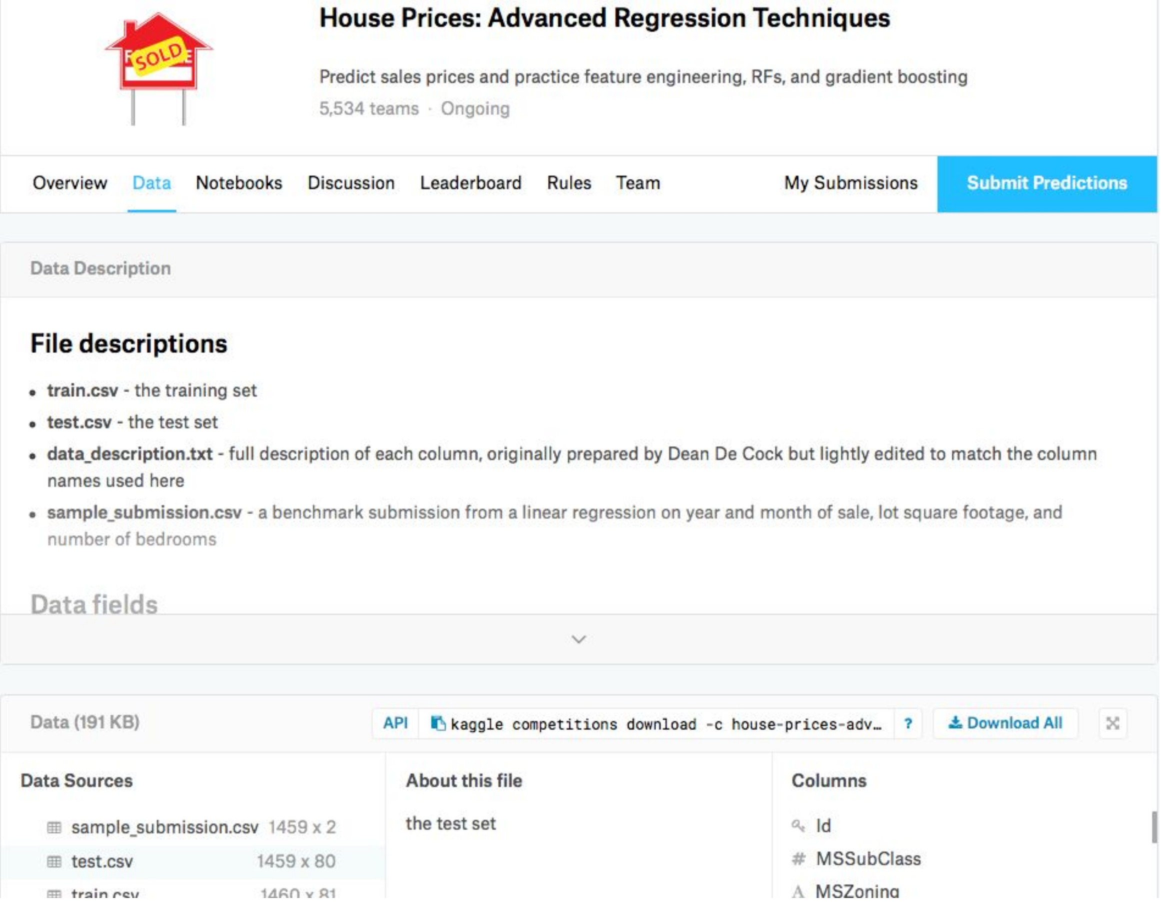
With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa (central Iowa in America) , I propose to predict the sales price for each house. For each Id in the test set, I must predict the value of the SalePrice variable.

Client

My clients are home buyers and they will decide to buy a dream house at a reasonable price.

2. Data Set

- The listing data of House Prices can be found on Kaggle: [House Prices: Advanced Regression Techniques]



House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

5,534 teams · Ongoing

File descriptions

- **train.csv** - the training set
- **test.csv** - the test set
- **data_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
- **sample_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data fields

Data (191 KB)

```
kaggle competitions download -c house-prices-adv…
```

Data Sources

sample_submission.csv	1459 x 2
test.csv	1459 x 80
train.csv	1460 x 81

About this file

the test set

Columns

Id
MSSubClass
MSZoning

3. Data Wrangling

First, I input data & do data wrangling. I conduct the following steps to clean up data and pick useful features.

Select columns that might be useful

- Visually inspect data & select columns that might be useful.

Analyze further & use only most relevant columns

- Originally there are 80 features/columns in the dataset, I further use only 9 most relevant variables with SalePrice that have more than 0.5 correlation with SalePrice.These can prove to be important features to predict SalePrice.

Fix data types

- I look through all the data & fixed data types. For example, I identify categorical data & use sklearn LabelEncoder to encode then for further usage.

Fix missing values

- Plot scatter plot for columns with missing values and inspect the trend. Confirmed that most of the features are either linear or random to target.
- Fill missing values:
    - With NA because it means the house does not have PoolQC, MiscFeature, Alley, and Fence.
    - With most frequent value because LotFrontage column is categorical columns but it has numerical values.
    - With median because MasVnrArea Masonry in square feet

Find & fix outliers

- Remove Faroutlier 3(IQR rule) above 75th percentile of the data.

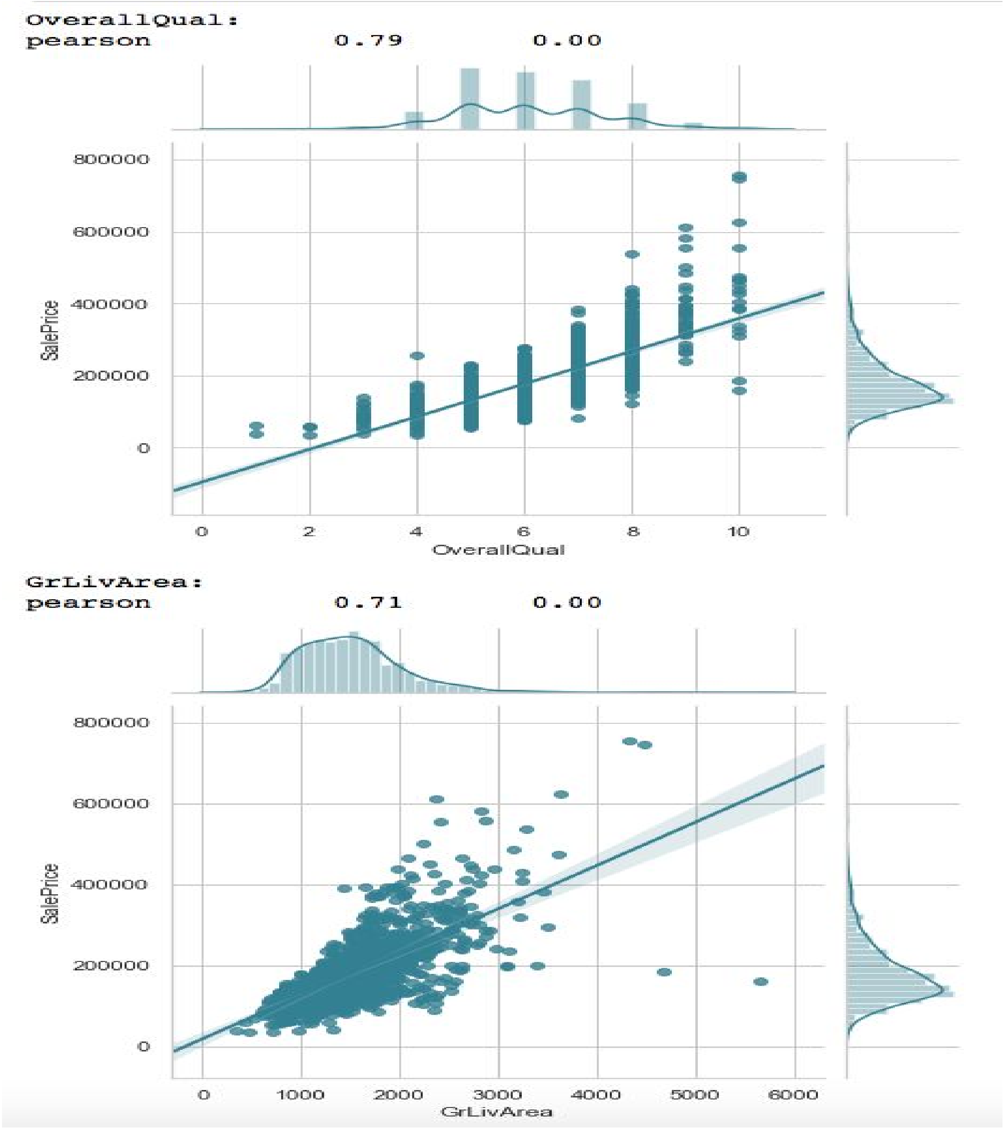
4. Initial Findings

During exploratory data analysis, I ask the following question:

- What might be the most important features that affect House price.

Initial findings are:

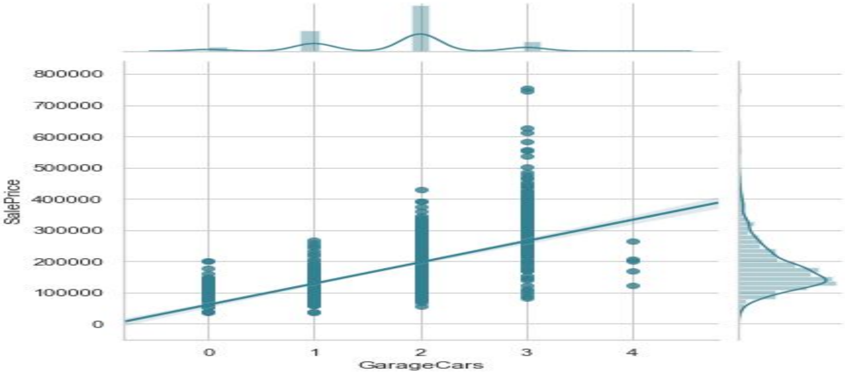
1. As we expected, independent variables, like OverallQual, GrLivArea, GarageCars, GarageArea,and TotalBsmtSF are correlated with the dependent variable, SalePrice.



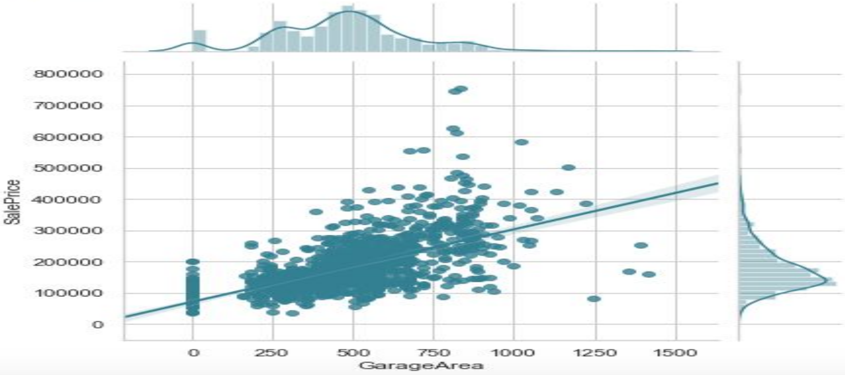
```
OverallQual:
pearson            0.79        0.00
```

```
GrLivArea:
pearson            0.71        0.00
```

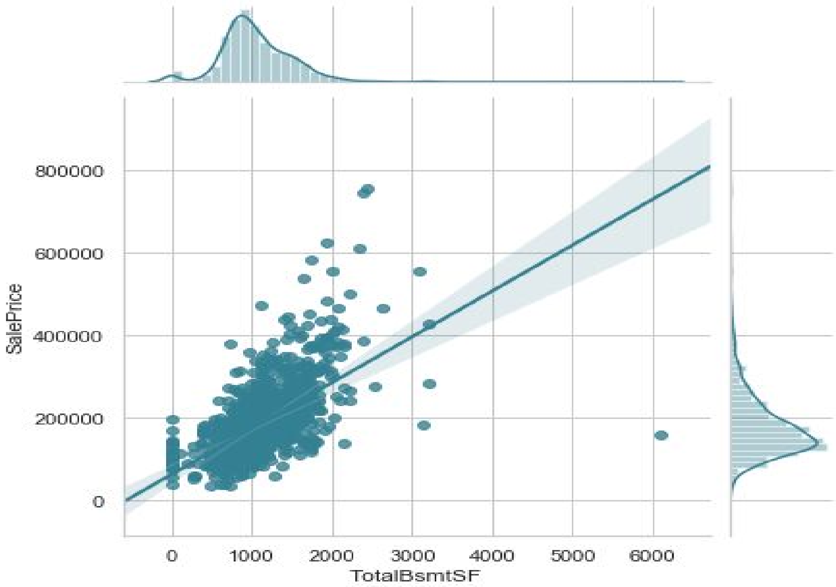
```
GarageCars:
pearson            0.64          0.00
```



```
GarageArea:
pearson            0.62          0.00
```



```
TotalBsmtSF:
pearson            0.61          0.00
```



2. However there might be some exceptions:
    a. There might be extremely high listing price for instance total living area is a little over 2,000 sqft and the listing price is more than $600,000.
    b. As overall quality increases, the listing price would not increase proportionally.
    c. As the number of garage cars increases, the listing price would not increase proportionally, either.

3. Living area wise, most of the houses have a living area of 1,000 to 2,000 sqft, and the listing price is around $200,000.

4. There are also strong correlations between pairs of independent variables such as OverallQual and GrLivArea, OverallQual and GarageCars, OverallQual and GarageArea, OverallQual and TotalBsmtSF. Finally, they all have pearson's correlation coefficient > 0.5, which correlate well with each other.