PREDICTING HOUSE PRICES WITH MULTIPLE LINEAR REGRESSION

Quynh Nguyen - BDS24
May 2022

I. Introduction

a. Background

House prices are rising rapidly and hard to predict. Many people consider houses not only accommodations but also a profitable source of investment. However, the risk still exists, especially real-estate bubble. Therefore, to reduce this risk, it is crucial to have a data-driven approach to price a house.

House value is dependent on many factors: location, functions, area, neighborhoods, utility, etc. This project only takes into account the house features itself, such as number of rooms, garage area, floor area, and a few more, for the sense of simplicity and data availability. The method to model this prediction is multiple linear regression (MLR) - a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable. [1]

b. Dataset

Source: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

The dataset above contains many columns, including both categorical and numeric values. Because we are doing regression, only numeric values are considered. Besides, for simplicity, only 9 attributes are chosen to train the model based on the common sense about the crucial factors on house prices.

Independent Variables

- TotalBsmtSF: Total square feet of basement area (m2)
- 1stFlrSF: First Floor square feet (m2)
- 2ndFlrSF: Second floor square feet (m2)
- LowQualFinSF: Low quality finished square feet (all floors) (m2)
- GrLivArea: Above grade (ground) living area square feet (1stFlrSF + 2ndFlrSF) (m2)
- WoodDeckSF: Wood deck area in square feet (m2)

- GarageArea: Size of garage in square feet (m2)
- PoolArea: Pool area in square feet (USD)
- MiscVal: $Value of miscellaneous feature (USD)

Dependent Variable: SalePrice (USD)

The first 5 rows of this dataset:

	TotalBsmtSF	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	WoodDeckSF	GarageArea	PoolArea	MiscVal	SalePrice
0	856.0	856.0	854.0	0.0	1710.0	0.0	548.0	0.0	0.0	208500.0
1	1262.0	1262.0	0.0	0.0	1262.0	298.0	460.0	0.0	0.0	181500.0
2	920.0	920.0	866.0	0.0	1786.0	0.0	608.0	0.0	0.0	223500.0
3	756.0	961.0	756.0	0.0	1717.0	0.0	642.0	0.0	0.0	140000.0
4	1145.0	1145.0	1053.0	0.0	2198.0	192.0	836.0	0.0	0.0	250000.0

Table 1: The first 5 rows of the dataset

II. Exploratory Data Analysis and Data Preparation

a. Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   TotalBsmtSF   1460 non-null   float64
 1   1stFlrSF      1460 non-null   float64
 2   2ndFlrSF      1460 non-null   float64
 3   LowQualFinSF  1460 non-null   float64
 4   GrLivArea     1460 non-null   float64
 5   WoodDeckSF    1460 non-null   float64
 6   GarageArea    1460 non-null   float64
 7   PoolArea      1460 non-null   float64
 8   MiscVal       1460 non-null   float64
 9   SalePrice     1460 non-null   float64
dtypes: float64(10)
memory usage: 114.2 KB
```

Figure 1: The information of the dataset

	TotalBsmtSF	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	WoodDeckSF	GarageArea	PoolArea	MiscVal	SalePrice
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	1057.429452	1162.626712	346.992466	5.844521	1515.463699	94.244521	472.980137	2.758904	43.489041	180921.195890
std	438.705324	386.587738	436.528436	48.623081	525.480383	125.338794	213.804841	40.177307	496.123024	79442.502883
min	0.000000	334.000000	0.000000	0.000000	334.000000	0.000000	0.000000	0.000000	0.000000	34900.000000
25%	795.750000	882.000000	0.000000	0.000000	1129.500000	0.000000	334.500000	0.000000	0.000000	129975.000000
50%	991.500000	1087.000000	0.000000	0.000000	1464.000000	0.000000	480.000000	0.000000	0.000000	163000.000000
75%	1298.250000	1391.250000	728.000000	0.000000	1776.750000	168.000000	576.000000	0.000000	0.000000	214000.000000
max	6110.000000	4692.000000	2065.000000	572.000000	5642.000000	857.000000	1418.000000	738.000000	15500.000000	755000.000000

Table 2: The table of descriptive statistics of the dataset

Key Observations:

- The dataset contains 1460 entries and 10 columns
- All columns in the dataset do not contain null values
- The datatype of all values in the dataset is float

- About Descriptive statistics, we will focus on the house prices in the "SalePrice" column:
  - The value is in the range from 349,000 USD to 755,000 USD
  - The mean is: 180,921 USD; The median is: 163,000 USD, this dataset is right-skewed.

b. Distribution

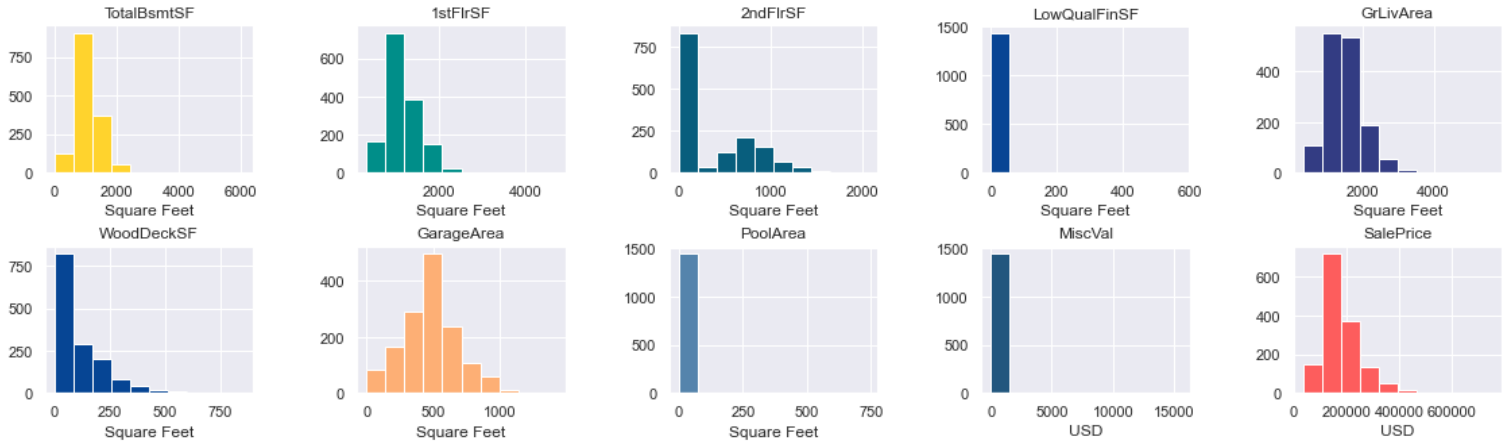

Figure 2: The distribution of the dataset attributes

Key Obsevations

- TotalBsmlSF, 1stFlrSf, GrLivArea, GarageArea, SalePrice: Right-skewed distributed
- The majority of values in 2ndFlrSF, LowQualFinSF, PoolArea, MiscVal,WoodDeckSF is zero or close to zero, which imply many houses do not or have an inconsiderable value of the second floor, low-quality area, pool, valuable miscellaneous, and wood deck area respectively.

c. Boxplots

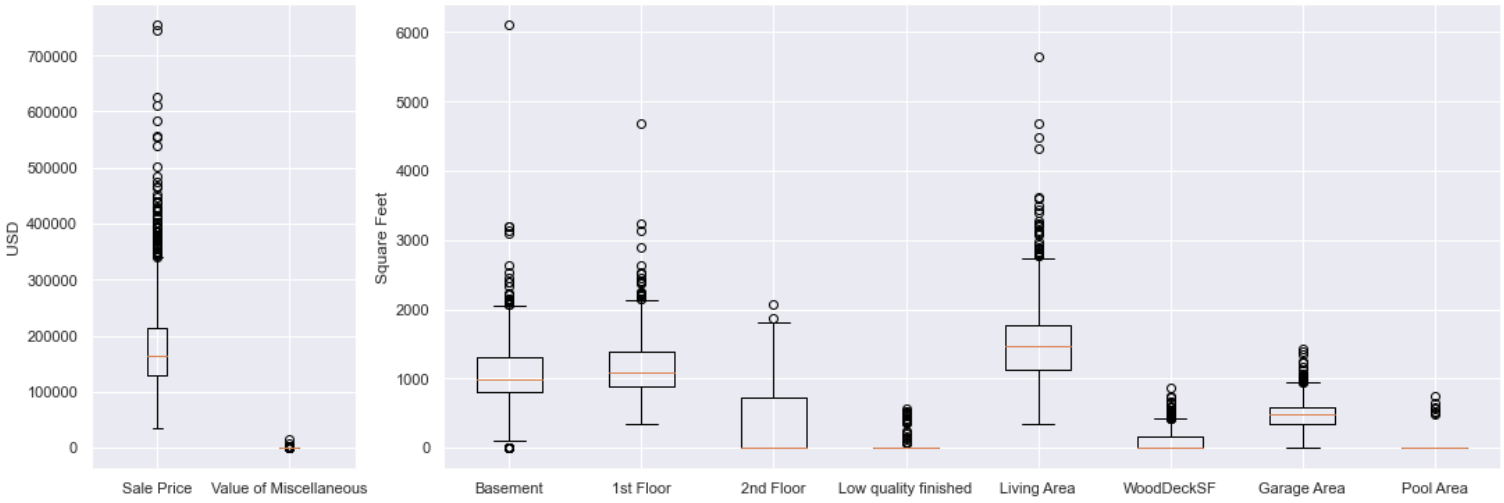

Figure 3: The boxplots of the dataset attributes (before removing outliers)

Key Obsevations:

All attributes of this dataset contain outliers. Therefore, to make sure that outliers doesn't affect the analysis, we will eliminate every records that contain outliers in any attribute. The rule to define an outlier is:

X is an outlier when x satisfies:

$$x > Q3 + 1.5IQR \quad or \quad x < Q1 - 1.5QR$$

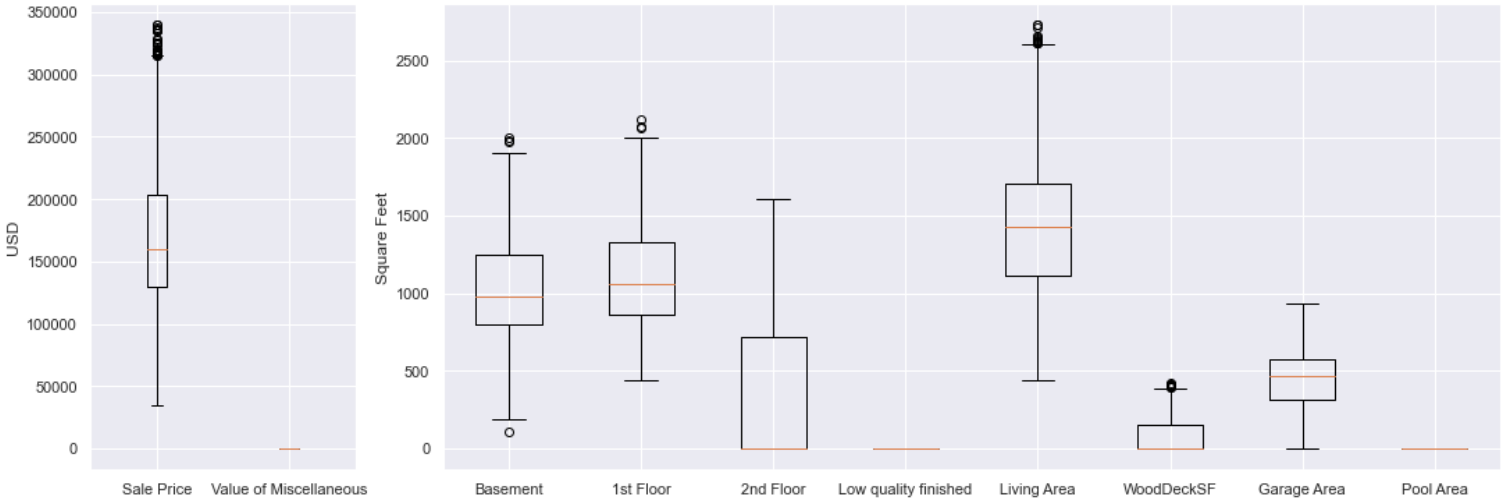

Figure 4: The boxplots of the dataset attributes (after removing outliers)

The number of outliers is reduced significantly so I assumed this is acceptable for further analysis.

d. Correlations

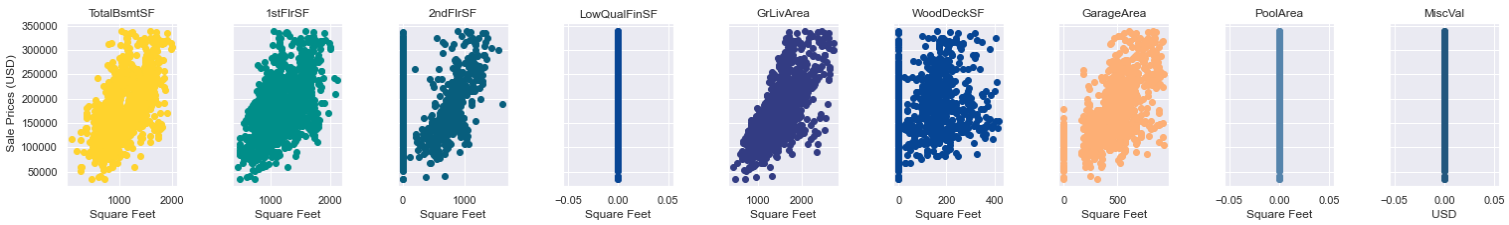

Figure 5: The scatter plots of the dataset attributes

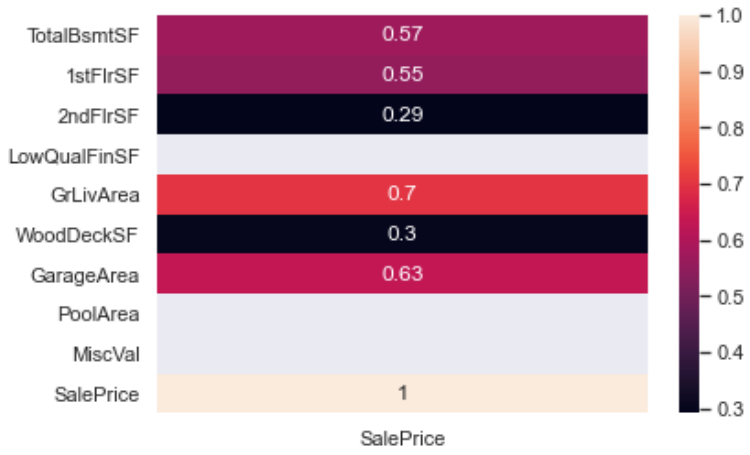

Figure 6: The correlations between the independent variables and the dependent variable.

Key Obsevations:

- The correlation between **LowQualFinSF, PoolArea, MiscVal** and the SalePrice in this dataset is approximately 0. Therefore, these attributes can be eliminated from the model because there is no relationship detected between these columns and house prices.
- **TotalBsmtSF, 1stFlrSF, GrLivArea,GarageArea** have fairly good correlations with the house prices. The attributes are expected to strongly influence the target variable.
- **2ndFlrSf** and **WoodDeckSF** have low correlations but still included in this analysis to increase the accuracy of the model

After the data preparation process, the number of rows is reduced to 1236 entries while the number of columns is reduced to 6.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1236 entries, 0 to 1458
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   TotalBsmtSF  1236 non-null   float64
 1   1stFlrSF     1236 non-null   float64
 2   2ndFlrSF     1236 non-null   float64
 3   GrLivArea    1236 non-null   float64
 4   WoodDeckSF   1236 non-null   float64
 5   GarageArea   1236 non-null   float64
dtypes: float64(6)
memory usage: 67.6 KB
```

Figure 6: The information of the dataset after the preparation stage

III. Regression Analysis

a. Prepare Testing and Training Data

The dataset is divided into 2 parts:

- *The training data (989 rows)*
- *The testing data (247 rows)*

The ratio of number of records between these parts are 8:2.

b. Fit the model

In Linear Regression, coefficients are estimated using the least squares criterion, in which we try to minimize the sum of squared residuals. Multiple Linear Regression simply includes multiple features. It takes the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Each x represents a different feature, and each feature has its own coefficient. In this case:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

x_1, x_2, x_3, x_4, x_5 and x_6 represent for TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'WoodDeckSF', 'GarageArea' respectively.

After fitting the data, the coefficients for each attribute are:

```
const          -14271.48
TotalBsmtSF       63.56
1stFlrSF           7.63
2ndFlrSF          29.64
GrLivArea         37.28
WoodDeckSF        67.40
GarageArea        91.12
dtype: float64
```

Figure 6: The coefficients for each attributes of the multiple linear regression

The form of the model is:

$$y = -14271.48 + 63.56\,x_1 + 7.63\,x_2 + 29.64\,x_3 + 37.28\,x_4 + 67.40\,x_5 + 91.12\,x_6$$

More information is shown below:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.427e+04	4135.905	-3.451	0.001	-2.24e+04	-6155.255
TotalBsmtSF	63.5584	6.270	10.137	0.000	51.254	75.863
1stFlrSF	7.6341	4.153	1.838	0.066	-0.517	15.785
2ndFlrSF	29.6419	2.597	11.413	0.000	24.545	34.739
GrLivArea	37.2760	2.385	15.632	0.000	32.596	41.955
WoodDeckSF	67.3952	9.804	6.875	0.000	48.157	86.633
GarageArea	91.1223	6.135	14.854	0.000	79.084	103.161

Table 3: Detailed information about the coefficients of the attributes

c. Hypothesis Testing

- Null Hypothesis:

$$\beta_1 = \beta_2 = \beta_3 + \beta_4 = \beta_5 = \beta_6 = 0$$

Imply: There is no relationship between "TotalBsmtSF", "1stFlrSF", "2ndFlrSF", "GrLivArea", "WoodDeckSF", "GarageArea" and "SalePrice".

- Alternative Hypothesis:

$$\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \neq \beta_6 \neq 0$$

Imply: There is relationships between "TotalBsmtSF", "1stFlrSF", "2ndFlrSF", "GrLivArea", "WoodDeckSF", "GarageArea" and "SalePrice".

To test the hypothesis, the p-value -a representation the probability that the coefficient is actually zero - is used:

```
const          5.830809e-04
TotalBsmtSF    4.930425e-23
1stFlrSF       6.636268e-02
2ndFlrSF       2.039362e-28
GrLivArea      2.347743e-49
WoodDeckSF     1.104155e-11
GarageArea     3.570173e-45
dtype: float64
```

Figure 6: The p-value of the coefficients of the dataset attributes

The coefficients of "TotalBsmtSF", "1stFlrSF", "2ndFlrSF", "GrLivArea", "WoodDeckSF", "GarageArea" are all less than 0.05. Therefore, we are reject the null hypothesis and conclude that:

There are relationships between the area of basement, the area of first floor, the area of second floor, the area of ground living area, the area of wood deck, the area of garage and the price of house

d. Model Evaluation

To evaluate a model, the overall fit of that model is considered. We use **R-squared value** - the proportion of variance in the observed data that is explained by the model. We expected to have high **R-squared value** because it means that more variance is explained by the model.

The R-Squared Value of this model is 0.7032. This is typically a good value for house prices prediction, due to the fact that we only take into account a few attributes of the original dataset.

Besides the R-squared value, we can use **Root Mean Squared Error (RMSE)** to evaluate the model. Other options are Mean Absolute Error (MAE) and Mean Squared Error (MSE). However, in this project, RMSE is used because it not only punishes large errors but also interprets in "y" units.

The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

-

n is the testing sample size. In this case, n is equal to *247.*

The RMSE of the model is: 32424.27

However, to understand the RMSE value in respect with the value range of the house price, this RMSE is normalized by this formula [2]:

$$RMSE = \frac{RMSE}{House\ Price_{Max} - House\ Price_{Min}}$$

The normalized RMSE is: 0.10733

This value is 0.11, which implies the differences between values predicted by a model and the actual value is only about 11% difference from the range of house prices. This is an appropriate number to use to predict the house price.

IV. Conclusion

This project demonstrates a project of analyzing data and building a statistical model using multiple linear regression to predict house prices.

We have chosen 9 attributes from the dataset of 1460 records, including: total square feet of basement area, first floor square feet, second floor square feet, low quality finished square feet, above grade living area square feet, wood deck area in square feet, size of garage in square feet, pool area in square feet and value of miscellaneous feature. After exploratory data analysis stage, we eliminated 3 attributes due to its no correlation to the target variable: low quality finished square feet, pool area in square feet, and value of miscellaneous feature.

In the regression analysis stage, we apply the multiple regression model to produce the formula below:

$$y = -14271.48 + 63.56\,x_1 + 7.63\,x_2 + 29.64\,x_3 + 37.28\,x_4 + 67.40\,x_5 + 91.12\,x_6$$

With:
- x_1: total square feet of basement area
- x_2: first Floor square feet
- x_3: second floor square feet
- x_4: above grade living area square feet
- x_5: wood deck area in square feet
- x_6: size of garage in square feet

Although the input attributes are simplified, the model produces a fairly good prediction with R-square is 0.7032 and normalized RMSE is 0.11. These measures demonstrate that we can use this model to predict the house price in practice.

v. References

[1] https://corporatefinanceinstitute.com/resources/knowledge/other/multiple-linear-regression/
[2] https://www.statology.org/what-is-a-good-rmse/
[*] https://www.ritchieng.com/machine-learning-evaluate-linear-regression-model/