# Final Report

# Housing Prices Analysis with Linear Regression

Yingshi Huang, Runjia She

Boston University, Metropolitan College

CS 521: Information Structures with Python

Alan Burstein

May 8th, 2022

**Introduction:**
In this project, we used python to analyze those factors that would influence house prices. We are given a big housing dataset with 81 columns, so we first need to download the dataset to the desktop, and then we use the Jupiter notebook to clean the dataset. We need to determine which factors would influence the house price and pick them out from the original dataset. In the first method, we used our general thinking and choose those factors that we care about when renting a house. On the other hand, we go to the internet, and observe the factors that the professional house rental and sales website, Zillow.com, shows on its homepage. Then we decide to choose "Lot area", "YearBuilt", "FullBath", "HalfBath", "BedroomAbvGr", and "GrLivArea" to find the relationship between these factors and sale price. In our first method, we do the data visualization to show the relationship between our choosing factors and the sale price. In the second method, Zillow.com choice, we first find the correlation between those factors we choose and then determine the coefficient for each of these factors, so that we can realize the relationship between them, and how they influence the sale price.

REDAME.md will provide instruction to use makefile.

1. Data range [IQR1 – IQR3] to explore how factors affect SalePrice
2. Data range reduce 2 standard deviations from each side of Edges
3. Train, and test data with different percentage
4. Compare our prediction with Zillow prediction

**Method data range [IQR1 – IQR3]: What factors we think would influence the sale price (Year Used, Sale price, GrLivArea). Extension (Linear, Polynomial Regression)**

In our general thinking, we think the used year of the house, and the total living area would influence the sale price. Then we pick these factors and build a new data frame. We first plot data and a linear regression model fit (fig 1. a). In this graph, we can observe the price increases when the living area increases. However, this graph is not clear to observe the relationship, since there are many extreme values in the graph, which would influence the trendline. Then we pick the medium 50% data to reduce outliers, and make the linear model fit our data better. From the second linear graph (fig 1. b), since the amount of data reduction, we can easily observe the relationship. The trendline follows an upward curve, when the living area is larger, which fits our thinking.

GrlivArea as independent value, SalePrice as dependent value:

**Fig 1.a**

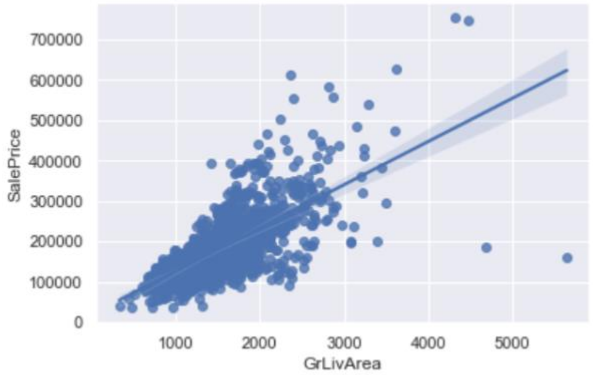
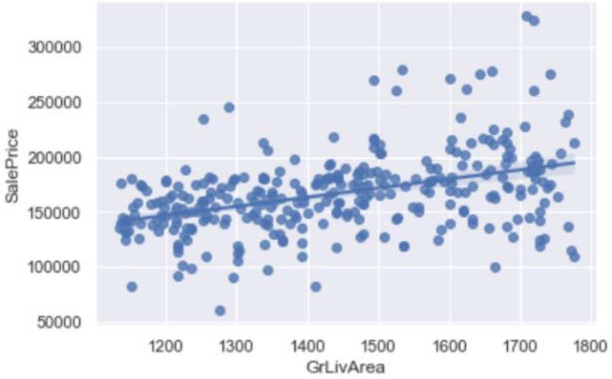


**Fig 1.b**



What is more, we analyze the relationship between house used years and sale price by using a polynomial curve graph to visualize. We build the polynomial curve graph with a degree of 4 (fig 2), which is the highest degree such that every polynomial equation can be solved by radicals, according to the Abel–Ruffini theorem. The goal we used polynomial regression here is to model the non-linear relationship between the year used and the sale price. In the graph, we can observe the curve going down when the year used increases, which also follows our general thinking, old houses are always cheaper than new houses.

YearUsed as independent value, SalePrice as dependent values:

**Fig 2**

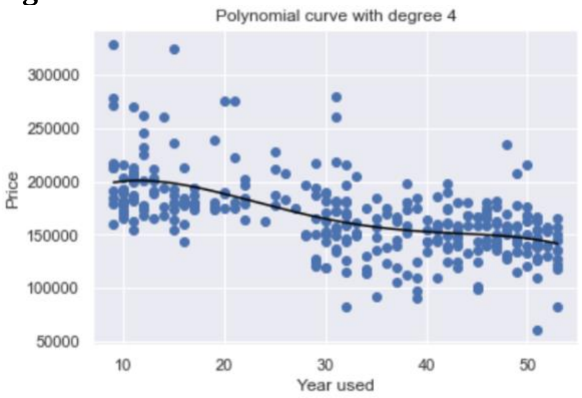
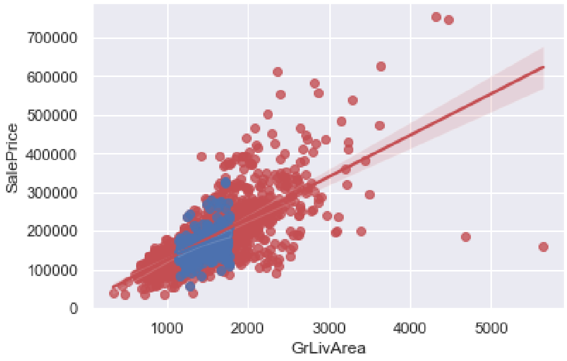


**Fig 3**



**Method increases data size, only remove outliers. Extension (Correlation)**

From IQR method Fig 3, plot the filter scatter over original scatter, the data have reduced large amount of the data, compare the two lines, we found the affection could create bias if apply to real life.

Therefore, we decide to use second method to reduce errors. To use standard deviation to reduce outliers. Maintain most of the housing data and apply both linear regression and polynomial regression. Compare both regressions to see the different. Check the relationship between factors and SalePrice by exanimating correlation value.

**Fig 4**

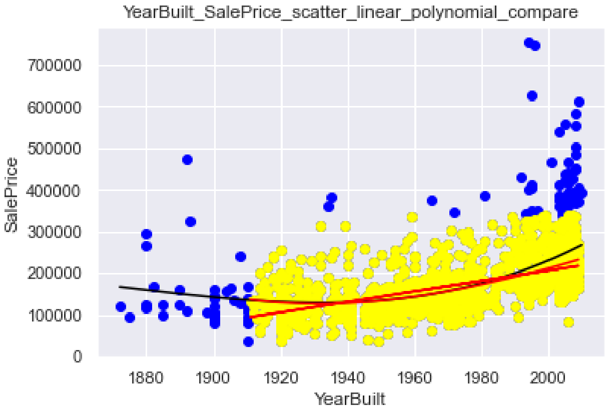


**Fig 5**

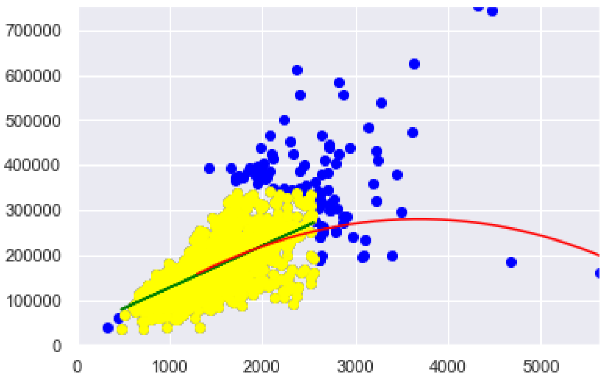


**Fig 4**, it is presenting how YearBuilt can affect SalePrice. The blue scatter present all the data from csv file, yellow scatter present the data which has removed outliers. Black line is representing the polynomial regression over entire data set. The red lines are polynomial and linear regression after filter outliers. In graph the lines have overlap each other, which means the data is concentric. We also notice that the minimum and maximum of SalePrice is slowly steadily increase if the YearBuilt increase.

**Fig 5**, use the same method to show entire data set by blue and yellow for remove outliers. The green line is linear regression, and the red line is polynomial regression. The two lines is overlap each other from minimum to 2500.

Linear Regression makes the estimation procedure simpler, like in this case we can just use calculate to predict SalePrice, not necessary a computer. Polynomial Regression is able to produce flexible and higher accuracy SalePrice in most case. However, in the housing data set we notice most of the time, Linear Regression and Polynomial Regression are overlap, so that it is better to use Linear Regression to predict price.

**Extension (Correlation)**

Organize housing data (clean invalid data – N/A data). We elevate all the factors that from guessing. Categorize housing data, and then pick 20 factors that might affect SalePrice. In practical only absolute value larger than 0.4 is meaningful in correlation. Therefore, we only collect factors have the absolute correlation value is higher than 0.4.

The graph only applies the large influence factors as X value, and the correlation values as Y value.

**Fig 6**

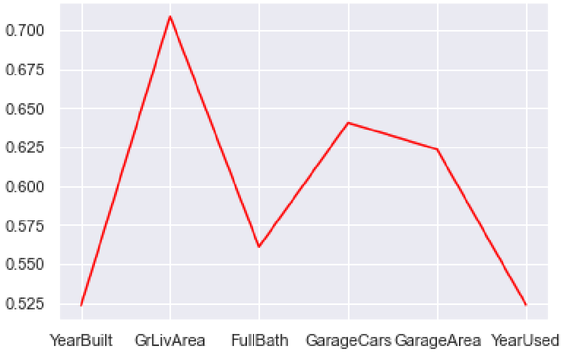


**Fig 6,** there are only 6 factors out of 20 able to pass 0.4 level. All the correlations are between 0.525 to 0.75. The highest is actual living area of the house on sale 0.75. We guess there are many factors can affect the SalePrice but people believe living area is the most important part for buying a house. Additionally, when the customers buy houses, they prefect latest built and larger garage more than more bathrooms in the house.

### Train, and test data with different percentage:

We pick the highest correlation to train and test our model. Therefore, the accuracy will be higher. First, we use only 80% to train the model, then 90% of the data. Since we are not using the same 80% or 90% percent, so the accuracy will change each time of new complier. We observe that the higher percentage can lead to higher accuracy.

It is also one of the reasons why real-life data set is extremely large to train and test. With larger amount of the housing data, it is easier to find the correct SalePrice.

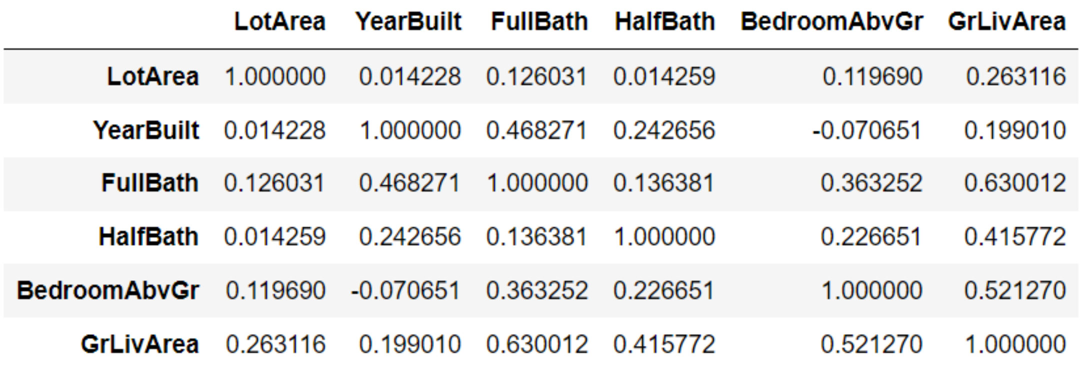
In this data, we only take GivLivArea as independent data, but we can also switch to different factors. In theory, the lower correlation will lead to lower accuracy.

### Method of Zillow: Variables they provide visitor to select. Extension compared to Zillow

Our first method didn't give very significant results, and we thought maybe we should consider more influenced factors. We go to zillow.com, one of the most famous house rentals and sell websites, to see what choices they provide to people, who are looking to buy or rent a house. On zillow.com, we observe the first line on its homepage is a list of factors visitors maybe care about most are "Location, for rent or sell, Price, how many bedrooms and bathrooms, and Home type", and when you click more, you can see more factors to help you lock on the specific house, such as "Square feet, Lot size, and Year built, etc."
We try to use Zillow's method to choose factors, and view the correlation in the following table (table 1). Based on the following table, we can see there is no highly correlated variable in the new data frame for the "Lot area", "YearBuilt", "FullBath", "HalfBath", "BedroomAbvGr", and "GrLivArea" variables, because all correlation values from the above table between the two variables are smaller than 0.8. So, I believe these variables have an acceptable correlation with each other, and I can use them together in a linear model without incurring multicollinearity-related risk.

Correlation for variables

**Table 1**

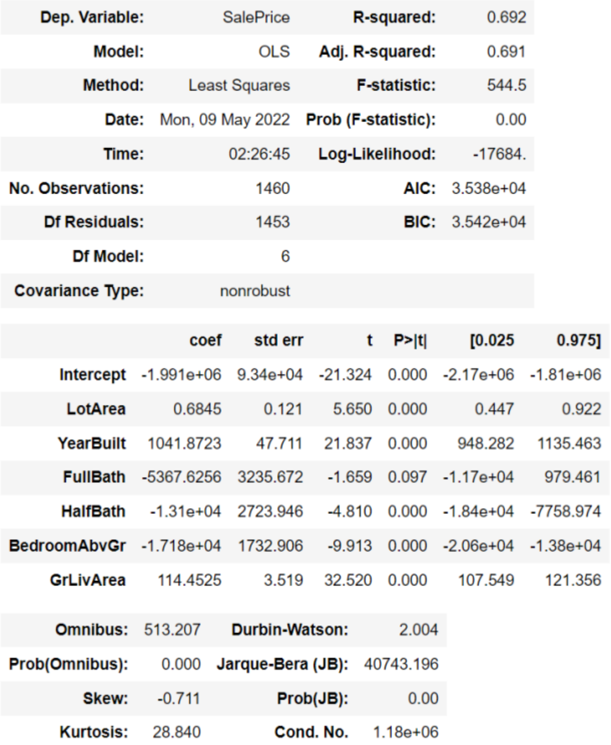


| | LotArea | YearBuilt | FullBath | HalfBath | BedroomAbvGr | GrLivArea |
|---|---|---|---|---|---|---|
| **LotArea** | 1.000000 | 0.014228 | 0.126031 | 0.014259 | 0.119690 | 0.263116 |
| **YearBuilt** | 0.014228 | 1.000000 | 0.468271 | 0.242656 | -0.070651 | 0.199010 |
| **FullBath** | 0.126031 | 0.468271 | 1.000000 | 0.136381 | 0.363252 | 0.630012 |
| **HalfBath** | 0.014259 | 0.242656 | 0.136381 | 1.000000 | 0.226651 | 0.415772 |
| **BedroomAbvGr** | 0.119690 | -0.070651 | 0.363252 | 0.226651 | 1.000000 | 0.521270 |
| **GrLivArea** | 0.263116 | 0.199010 | 0.630012 | 0.415772 | 0.521270 | 1.000000 |

Based on the correlation result, we build a model (Table 2) that uses SalePrice as the outcome variable, with LotArea, YearBuilt, FullBath, HalfBath, BedroomAbvGr, and GrLivArea as the input variables. From the following table, the prob(F-statistic) is our p-value, which is 0.00. This means the probability of the null hypothesis being true is very low. As per the above results, the probability is close to zero. This implies that the regression is meaningful. Moreover, in the table below, we can also observe the predictor variables for all input variables are significant because all their p-values are closed to 0.00. These values are smaller than the common alpha level of 0.1, so we can conclude that they are statistically significant, and we can use these coefficients to predict the house sale price.

## OLS Regression

**Table 2**



| | | | |
|---|---|---|---|
| **Dep. Variable:** | SalePrice | **R-squared:** | 0.692 |
| **Model:** | OLS | **Adj. R-squared:** | 0.691 |
| **Method:** | Least Squares | **F-statistic:** | 544.5 |
| **Date:** | Mon, 09 May 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 02:26:45 | **Log-Likelihood:** | -17684. |
| **No. Observations:** | 1460 | **AIC:** | 3.538e+04 |
| **Df Residuals:** | 1453 | **BIC:** | 3.542e+04 |
| **Df Model:** | 6 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **Intercept** | -1.991e+06 | 9.34e+04 | -21.324 | 0.000 | -2.17e+06 | -1.81e+06 |
| **LotArea** | 0.6845 | 0.121 | 5.650 | 0.000 | 0.447 | 0.922 |
| **YearBuilt** | 1041.8723 | 47.711 | 21.837 | 0.000 | 948.282 | 1135.463 |
| **FullBath** | -5367.6256 | 3235.672 | -1.659 | 0.097 | -1.17e+04 | 979.461 |
| **HalfBath** | -1.31e+04 | 2723.946 | -4.810 | 0.000 | -1.84e+04 | -7758.974 |
| **BedroomAbvGr** | -1.718e+04 | 1732.906 | -9.913 | 0.000 | -2.06e+04 | -1.38e+04 |
| **GrLivArea** | 114.4525 | 3.519 | 32.520 | 0.000 | 107.549 | 121.356 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 513.207 | **Durbin-Watson:** | 2.004 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 40743.196 |
| **Skew:** | -0.711 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 28.840 | **Cond. No.** | 1.18e+06 |

**Conclusion:**

Compare all the methods above, they both provide information of how to factors might affecting the SalePrice. If the data has very strong relationship, and the data is concentric, use IQR1-IQR3 will be fast and less data to deal with. Clear N/A data and reshape the data set is one of the most important ways to computer the data.

At the beginning, when we are not familiar with data set, analyzes with eyeballing is fasting (plot the entire scatter graph is also helpful). Remove outliers by using the standard deviation is affective. It also able to reduce bias when we predict Regressions.

Every type of regressions might have their own benefit like linear regression is direct and easy to get the result. Polynomial regression is easier to curve the shape of the scatter, it is more flexible. We can decide to use which one after compare both or even use one in certain range, and another in another range. (Use Linear Regression is it is concentric, and steady. Use Polynomial if the data is in different spots all over the graph.)