Analyze Housing Prices Affect by What Types of Factors by guessing as customers

APPLY MACHINE LEARNING WITH HOUSING DATA

Yingshi Huang, Runjia She

CS 521 Information Structure with Python
2022 Spring A2

Table of Contents

Introduction: *2*

Technology applied: *2*

Method 1: *3*

Steps: 3

LotArea as independent value, SalePrice as dependent value: 4

histogram SalePrice: 4

histogram LotArea: 5

histogram YearUsed: 5

YearUsed as independent value, SalePrice as dependent values: 5

Method 2: *6*

Steps: 7

YearBuilt VS SalePrice: 8

Original: 8

Polynomial_original: 8

Clean_linear: 9

Clean_polynomial: 9

Compare: 9

LotArea VS SalePrice: 10

Original: 10

Clean: 10

Clean_linear: 11

Clean_polynomial: 11

Compare: 11

Conclusion: *13*

Reference: *14*

Introduction:

In the final project, it shows two methods to analyze the housing market. How the factors like area, year built affect the housing prices.

Technology applied:

1. Python
2. Numpy
3. Pandas
4. Matplotlib
5. Seaborn
6. Sklearn.linear_moder
7. Sklearn.pipeline
8. Sklearn.preprocessing
9. Scipy.stats
10. Machine Learning
11. Linear regression
12. Polynomial regression

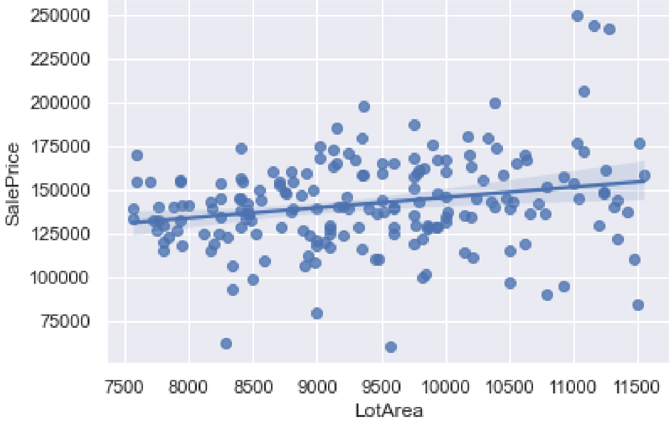
Method 1:

Read the housing data manually and decide range of the data by eyeballing. Reduce mass of data, so that the data is easier to analyze. The new range of housing data is more efficient.

Steps:

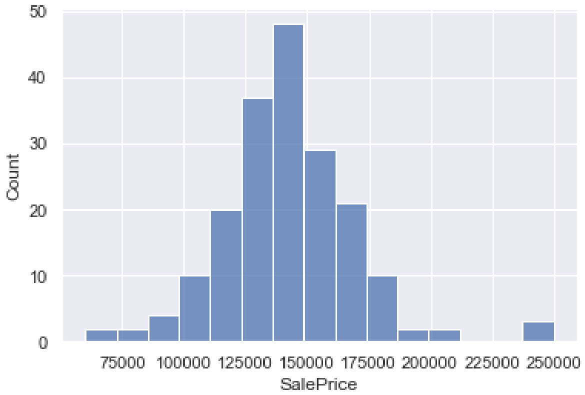
1) Calculate the first factor which might affect the SalePrice of the house: the years of house has been using by minus when the house was built from the year of selling the housing.
2) Pick the second factor LotArea.
3) Drop N/A data from housing.
   a) Use Street to sort the data, there are only two streets—Grvl, Pave.
   b) Ignore Grvl Street because there are only 6 sets of data, which is too less to analyze
4) Only use [Q1, Q3] the middle 50% of values when ordered from lowest to highest from interquartile range (IQR).
   a) Normal test Pave Street median 50% data by SalePrice
5) Plot LotArea VS SalePrice using linear regression
6) Plot histogram SalePrice
7) Plot histogram LotArea
8) Plot histogram YearUsed
9) Plot polynomial curve degree 4 YearUsed VS Price

LotArea as independent value, SalePrice as dependent value:



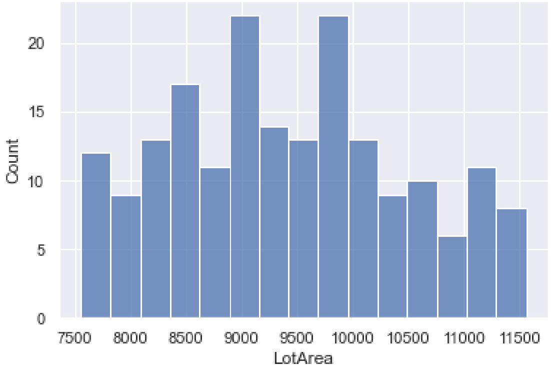
In general thinking, we first thought Lot Area would influence the house pricing, and we build the linear graph to observe the relationship. We can see the trendline follows an upward curve, but not very significant, and we can still observe many points below the line. So, we still need more analysis to find out the most influential factor in house price.

histogram SalePrice:



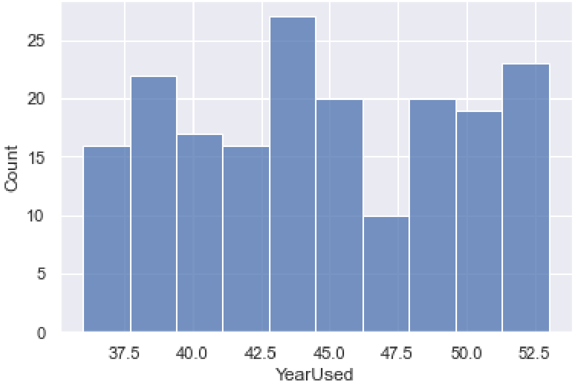
Based on our cleaned dataset, we can observe that most sale prices are around 125000 to 175000.

histogram LotArea:



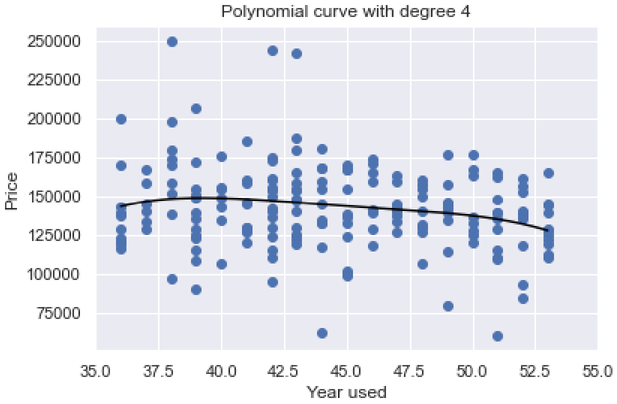
From this graph, we can observe the sales houses lot area distribution, and we can see most houses distributed on the left, and most people preferred houses with lot areas around 10000 square feet and 9000 square feet, and there are not many counts on the right tail.

histogram YearUsed:



In our cleaned dataset, we make the new column to view the year used and make the following histogram to view the distribution of sales houses and their used years. We can see the range for using year are between 30 years to 55 years.

YearUsed as independent value, SalePrice as dependent values:



We also try to analyze the relationship between house used years and their sale price. We build the polynomial curve graph with a degree of 4. Degree four (quartic case) is the highest degree such that every polynomial equation can be solved by radicals, according to the Abel–Ruffini theorem. We can observe the graph and see the curve going down when the year used increases. However, this trend is not enough strong, and we cannot conclude that year used will influence the sale prices a lot.

Method 2:

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.

Steps:

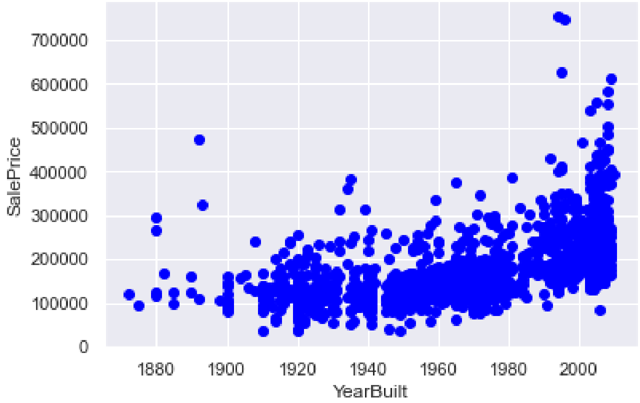
1) Organize housing data (clean invalid data – N/A data)
2) Categorize housing data
   a) Pick the factors that might affect Sale Price by guessing as a buyer:
   b) LotArea, LotShape, OverallCond, YearBuilt, Foundation, Electrical: Electrical system, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, Bedroom, Kitchen, Functional, GarageCars, GarageArea, YrSold, SaleCondition

   c) Use machine learning to evaluate the independent variable if it is in high relative by using correlation coefficient.

   d) In practice if the absolute value of correlation coefficient is larger than |correlation|>=0.4 has a meaningful relationship.
3) Drop outlines
   a) Drop all the data is outside 2 standard deviations.

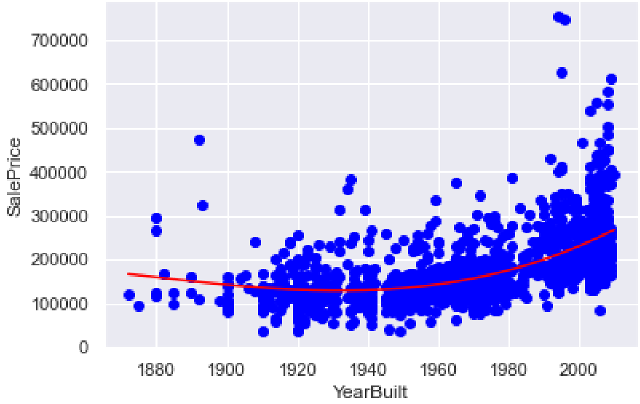
Meaningful Graph YearBuilt VS SalePrice:

Original:



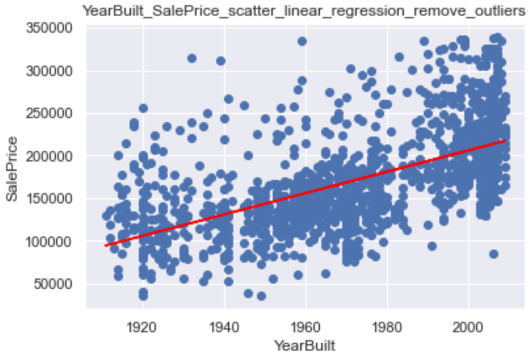
The scatter graph can show the direct relationship between YearBuilt and SalePrice. Most of the data are closer to each other. However, there are still some of the outliers on the left side of the graph and the upper part of the graph. The total price of the houses is mostly steadily increase by years. It is much obvious from 1980 to 2010.

Polynomial_original:



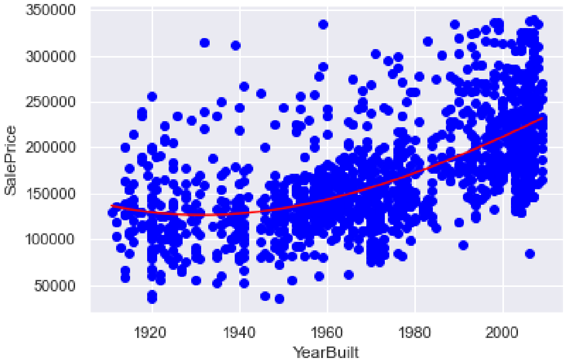
The shape of the original scatter graph is a little curve and going up, so that polynomial degree of 3 might be a good assumption. The red line represents polynomial degree of 3 which seems fit inside the scatter graph.

Clean_linear:



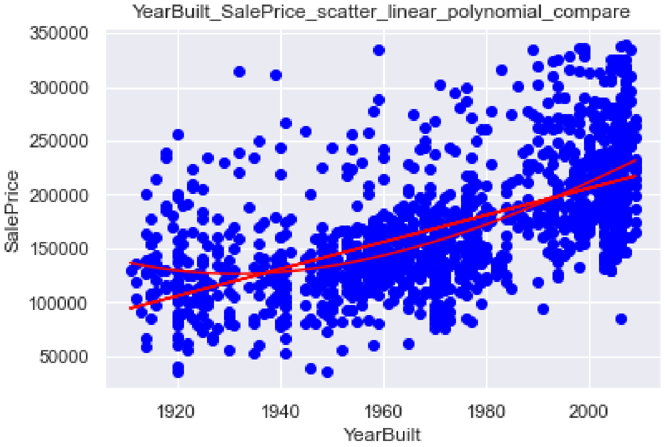
The outliers is far from most of the data, so that clean the outliers is going to help the precise of the linear regression. The outliers are defined as 2 deviations from standard. The normal rule is minus 1.5 standard deviations from both side of the boundaries. However, in this case we decide to reduce 2 standard deviations so that the data are more compact.

Clean_polynomial:



The Linear regression seems fit in the graph but use another method to test can help developers found the results better. Therefore, we plot one more polynomial to.

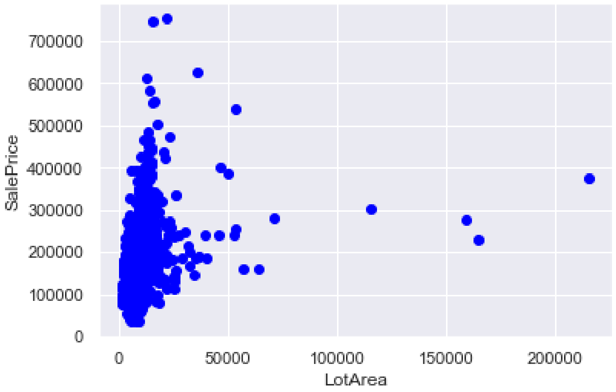
Compare:



Plot both regression together, they are very close to each other, and both fit the scatter graph.

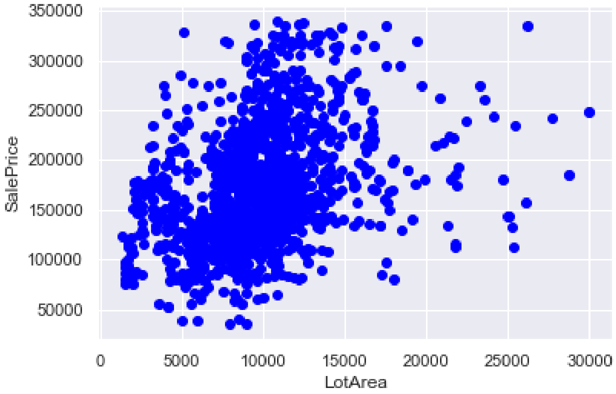
LotArea VS SalePrice:

Original:



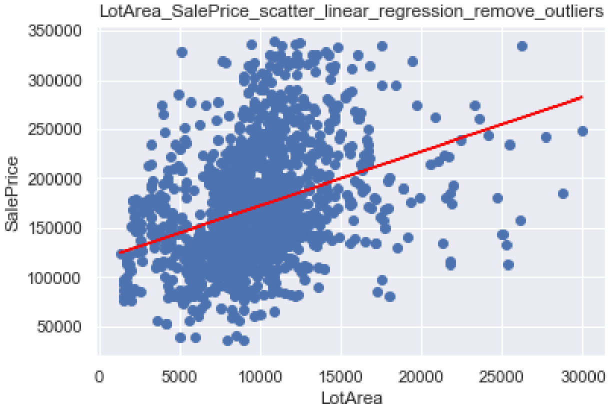
This the scatter graph of LotArea VS SalePrice, which seems unrelative. Additionally, the outliers are extremely far from most of the data. Therefore, clean outliers can have better affection.

Clean:



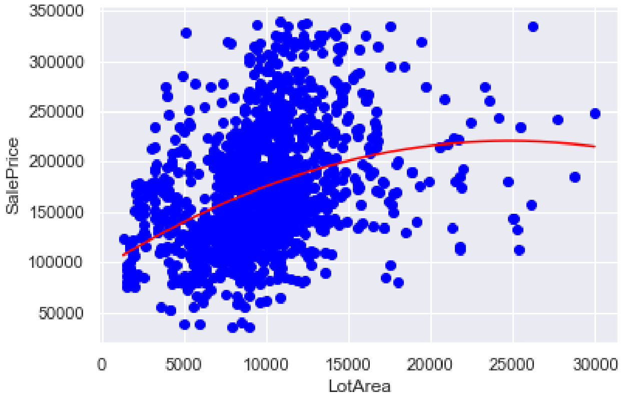
Data without outliers by minus 2 standard deviations from each side of boundaries. There are some of them if seem far from central. However, these data are in the range, so it is fair data.

Clean_linear:



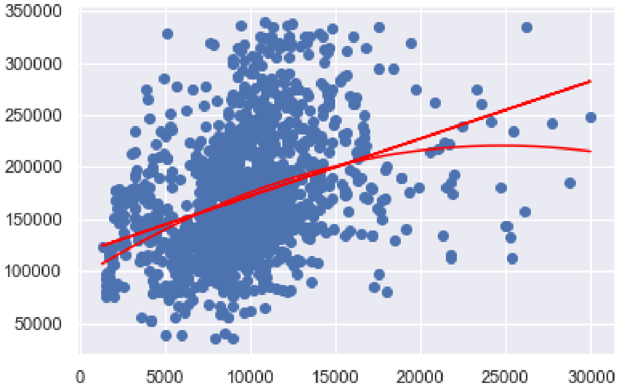
Plot Linear regression. It is affect by the edge data.

Clean_polynomial:



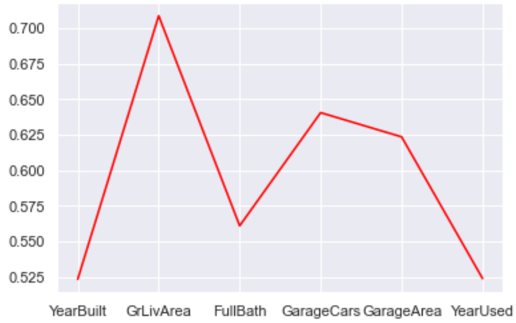
Plot polynomial by 2 degree, it is the curve is larger than YearBuilt vs SalePrice.

Compare:



It is very different from two regressions, so that people should be careful to apply any one of the regressions.

High correlation from factors from guessing:



We elevate all the factors that from guessing. In practical only absolute value larger than 0.4 is meaningful in correlation. Therefore, this graph only applies the large influence factors as X value, and the correlation values as Y value.

Conclusion:

Compare two methods, they both provide information of how to factors might affecting the SalePrice. Method 1 clears the data and analyzes with eyeballing and computer calculation. Method 2 use the standard deviation to clear outliers and evaluate by correlation larger than 0.4.

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.

Reference:

House Prices | Stacked regression | Top 8%
https://www.kaggle.com/code/gabedossantos/house-prices-stacked-regression-top-8
House Price - Multi Linear Regression with EDA
https://www.kaggle.com/code/harshghadiya/house-price-multi-linear-regression-with-eda
House Price(step-by-step modeling)
https://www.kaggle.com/code/adibouayjan/house-price-step-by-step-modeling
House_price_AI 기초_guide(220626)
https://www.kaggle.com/code/laplace8/house-price-ai-guide-220626
HousePrices Simple ML Workflow | Top 11%
https://www.kaggle.com/code/uyeanil/houseprices-simple-ml-workflow-top-11
House Prices. Feature engineering, EDA
https://www.kaggle.com/code/georgyzubkov/house-prices-feature-engeneering-eda
HousePrice (step-by-step)🔥🏘
https://www.kaggle.com/code/abdelrahmantarek13/houseprice-step-by-step
Machine Learning - Polynomial Regression
https://www.w3schools.com/python/python_ml_polynomial_regression.asp
Np.polyfit: How To Use Numpy Polyfit() Method In Python
https://appdividend.com/2022/01/28/numpy-polyfit-method-in-python/
Is there a numpy builtin to reject outliers from a list
https://stackoverflow.com/questions/11686720/is-there-a-numpy-builtin-to-reject-outliers-from-a-list
“numpy remove nan from 2d array” Code Answers
https://www.codegrepper.com/code-examples/python/frameworks/file-path-in-python/numpy+remove+nan+from+2d+array
Linear Regression in 6 lines of Python
https://towardsdatascience.com/linear-regression-in-6-lines-of-python-5e1d0cd05b8d
Convert the NumPy Array to Pandas DataFrame
https://datatofish.com/numpy-array-to-pandas-dataframe/
Correlation and Linear Regression
https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_correlation-regression/bs704_correlation-regression_print.html
Introduction to Correlation and Regression Analysis
https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_multivariable/mobile_pages/BS704_Multivariable5.html

How to detect constant, quasi-constant features in your dataset
https://towardsdatascience.com/how-to-detect-constant-quasi-constant-features-in-your-dataset-a1ab7aea34b4
Dropping Constant Features using VarianceThreshold: Feature Selection -1
https://medium.com/nerd-for-tech/removing-constant-variables-feature-selection-463e2d6a30d9