Analyze Housing Prices Affect by What Types of Factors

APPLY MACHINE LEARNING WITH PYTHON

Huang, Yingshi, Runjia She

CS 521 Information Structure with Python
2022 Spring A2

Table of Contents

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

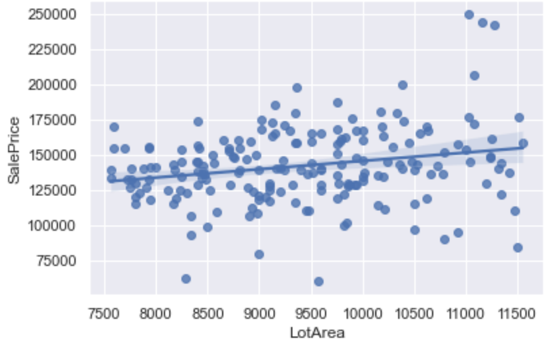
Method 1:

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.

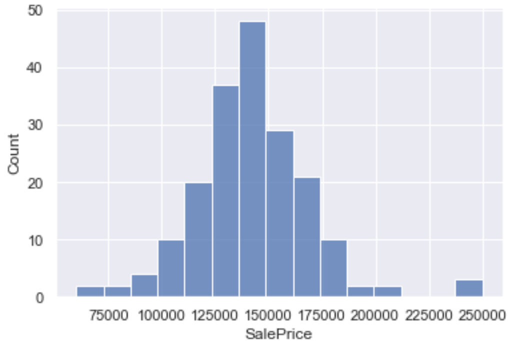
Steps:

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

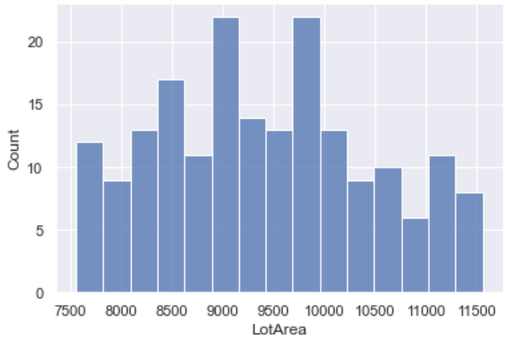
LotArea as independent value, SalePrice as dependent value:



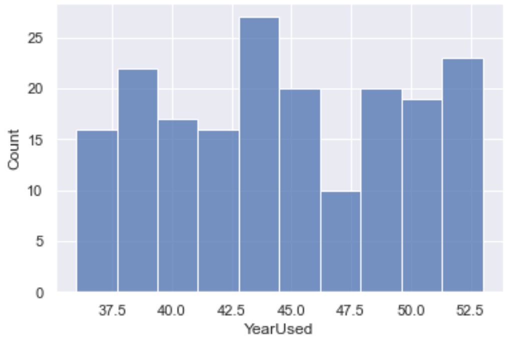
histogram SalePrice:



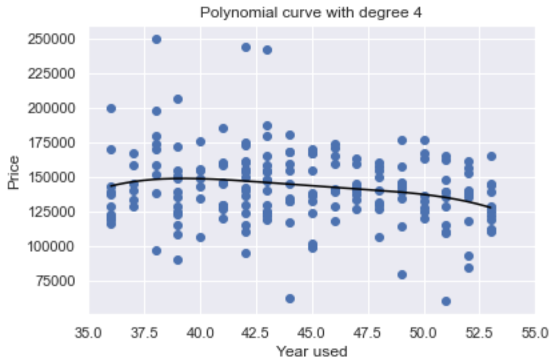
histogram LotArea:



histogram YearUsed:



YearUsed as independent value, SalePrice as dependent values:



Method 2:

Use standard deviation to reduce outliers. Maintain most of the housing data and apply both linear regression and polynomial regression. Compare both of the regressions to see the different.

Steps:

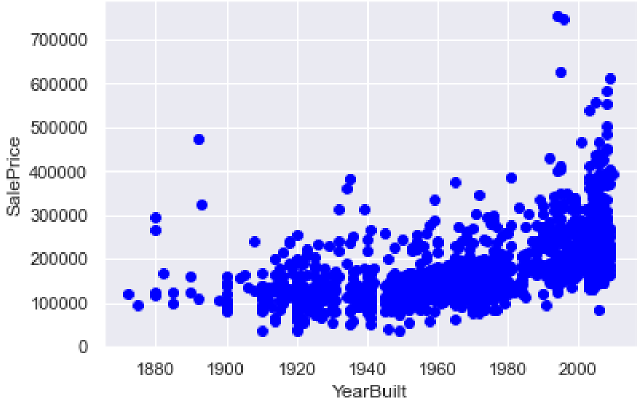
1) Organize housing data(clean invalid data – N/A data)
2) Categorize housing data
   a) Pick the factors that might affect Sale Price by guessing as a buyer:
   b) LotArea, LotShape, W, OverallCond, YearBuilt, Foundation, Electrical: Electrical system, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath,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

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

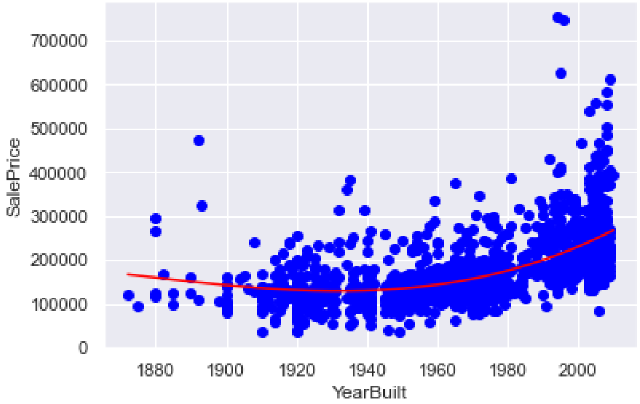
   d) In practice if the absolute value of correlation coefficient is larger than |cc|>=0.4 has a meaningful relationship.
3) Drop outlines
   a) Drop all the data is outside 2 standard deviation

YearBuilt VS SalePrice:

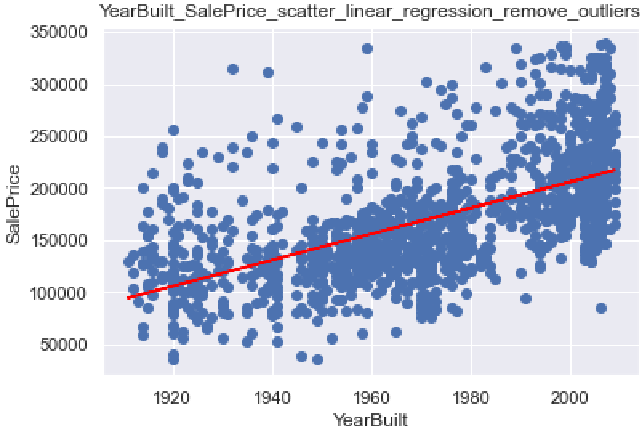
Original:



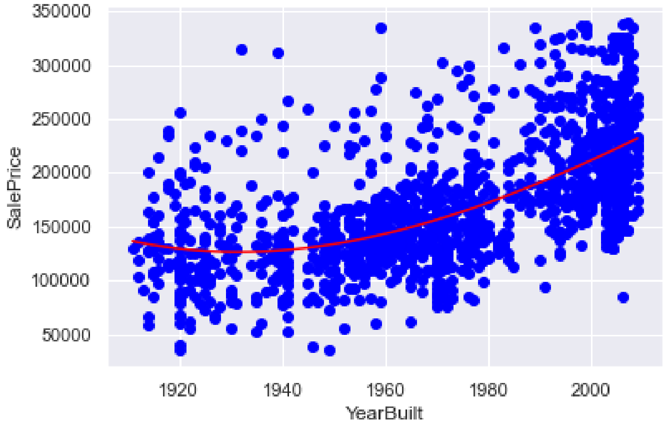
Polynomial_original:



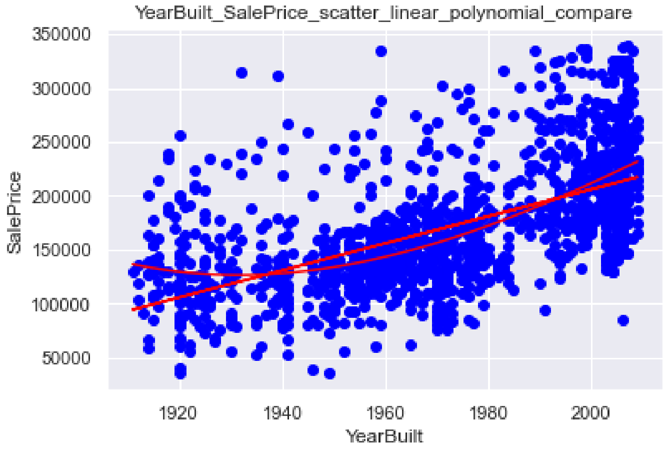
Clean_linear:



Clean_polynomial:

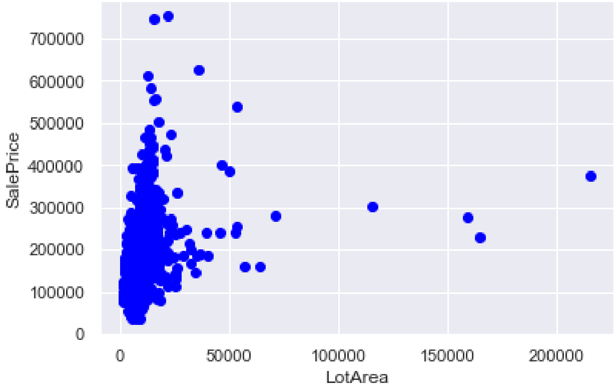


Compare:

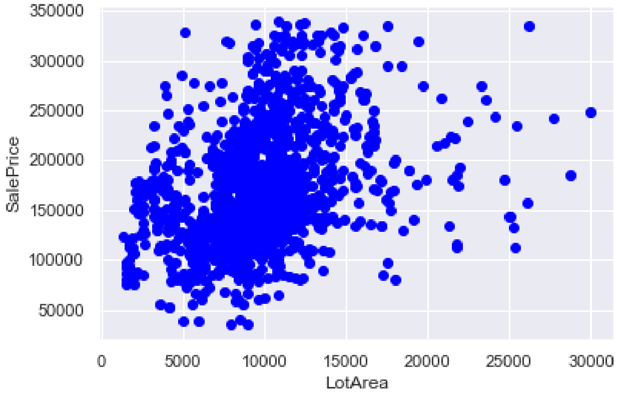


LotArea VS SalePrice:

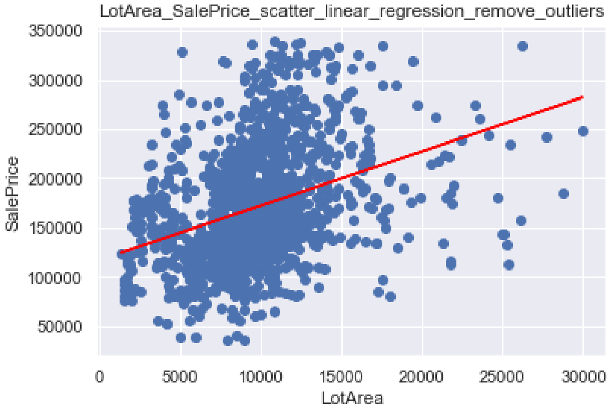
Original:



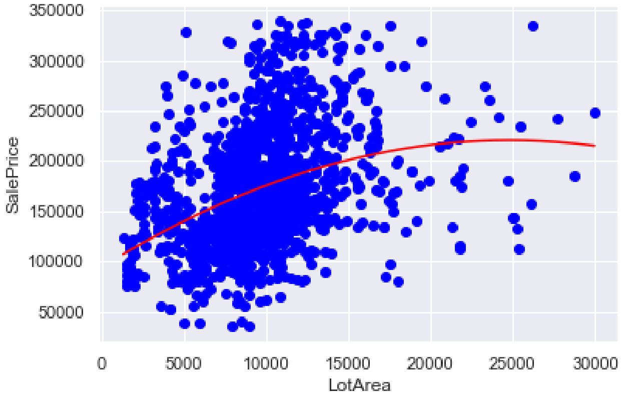
Clean:



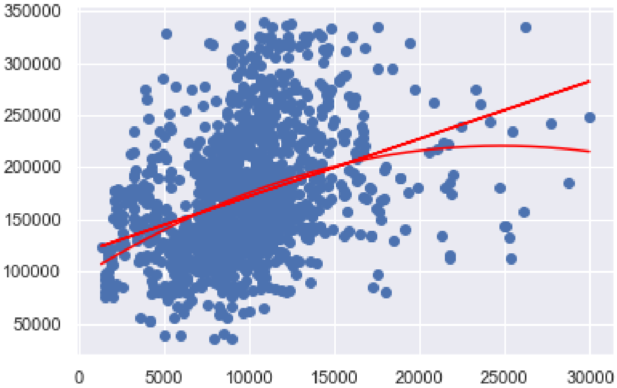
Clean_linear:



Clean_polynomial:



Compare:



Conclusion:

Compare two methods, they both provide information of how to clear the data and analyze with eyeballing and computer calculation.

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