**Project Description**

* What I have done:

1. Downloaded dataset of house prices from Kaggle (data size: 1460 observations with 81 variables, 38 of the variables are numeric and 43 are categorical);
2. Imported dataset using panda;
3. Briefly went through all numeric variables and choose 11 of them (including sales price, which is the object of this analysis) by defining a class to do a further analysis.
4. Built an Ordinary Least Square model and regressed sales price on other 10 variables using multi-variable linear regression.
5. Used the OLS model, calculated the estimated house prices of houses in the dataset, and compared with the real prices provided in the dataset.

* Results:

1. List of variables that being used in the model:

The building class(MSSubClass), Lot size in square feet(LotArea), Overall material and finish quality(OverallQual), Overall condition rating(OverallCond), Original construction date(YearBuilt), Remodel date(YearRemodAdd), Total square feet of basement area(TotalBsmtSF), Above ground living area in square feet(GrLivArea), Size of garage in square feet(GarageArea), Year sold(YrSold).

1. Regression model:

SalePrice = β1MSSubClass + β2LotArea + β3OverallQual + β4OverallCond + β5YearBuilt + β6YearRemodAdd + β7TotalBsmtSF + β8GrLivArea + β9GarageArea + β10YrSold

1. Value of coefficients:

β1 = -141.707 β2 = 0.582634 β3 = 20047.1 β4 = 5022.84

β5 = 416.968 β6 = 174.311 β7 = 21.1518 β8 = 52.3603

β9 = 39.0012 β10 = -625.318

1. Regression results:

SalePrice = -141.707MSSubClass + 0.582634LotArea + 20047.1OverallQual + 5022.84OverallCond + 416.968YearBuilt + 174.311YearRemodAdd + 21.1518TotalBsmtSF + 52.3603GrLivArea + 39.0012GarageArea – 625.318YrSold

1. Among variables that being used in the regression, only the building class and the year sold are negatively related to the house prices, which means that the higher the building class, the cheaper the price is; the later you sell the house, the cheaper it would be.

* Directions on how to run the code

1. Make sure that the original dataset can be seen directly in the file explorer (See below);



1. Run the whole code directly, the coefficients of the regression are shown in the variable explorer under the name of *beta*, and the variables can be seen under the name of *train\_data*. Under the name of *compare*, predicted prices and real prices are listed together to make comparison.

