Project: House Price Prediction – Advanced Regression Techniques

Author: **Ngô Đại Phương (Phuong Dai Ngo)**
Programming language: **Python**
Statistics model: **Linear Regression**
Source: **Kaggle**
Dataset:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Table of Content

I. Overview 2
II. Preprocessing 3
1/ MsSubClass: Identifies the type of dwelling involded in the sale 4
2/ MSZoning: Identifies the general zoning classification of the sale 5
3/ LotArea: Lot size in square feet 6
4/ BldgType: Type of dwelling 7
5/ HouseStyle: Style of dwelling 8
6/ YearBuilt: Original construction date 9
7/ Yrsold: Year Sold 10
8/ Sale Type 11
9/ Sale Condition 12
10/ OverallQual: Rates the overall house material and completion 14
III. Model & Evaluation 15

I. Overview: Housing Price Prediction in the Ames, Iowa, United States

79 explanatory variables are provided from the dataset and describe aspects of residential properties in Ames, Iowa. This evaluation model will take 10 features (listed below) as the main influencers to be analyzed and visualized with Python and basic Linear Regression to assess the final Root Mean Squared Error of the US house prices in Ames. With further emphasis on the period of 2006-2010, when the Financial Crisis 2007-2008 happened, this model will discuss which variable becomes the most influential feature affecting the Sale Price in Ames.

10 key features for the analysis: 'SalePrice', 'MSSubClass', 'MSZoning', 'LotArea', 'BldgType', 'HouseStyle', 'YearBuilt', 'YrSold', 'SaleType', 'SaleCondition', ‘OverallQual

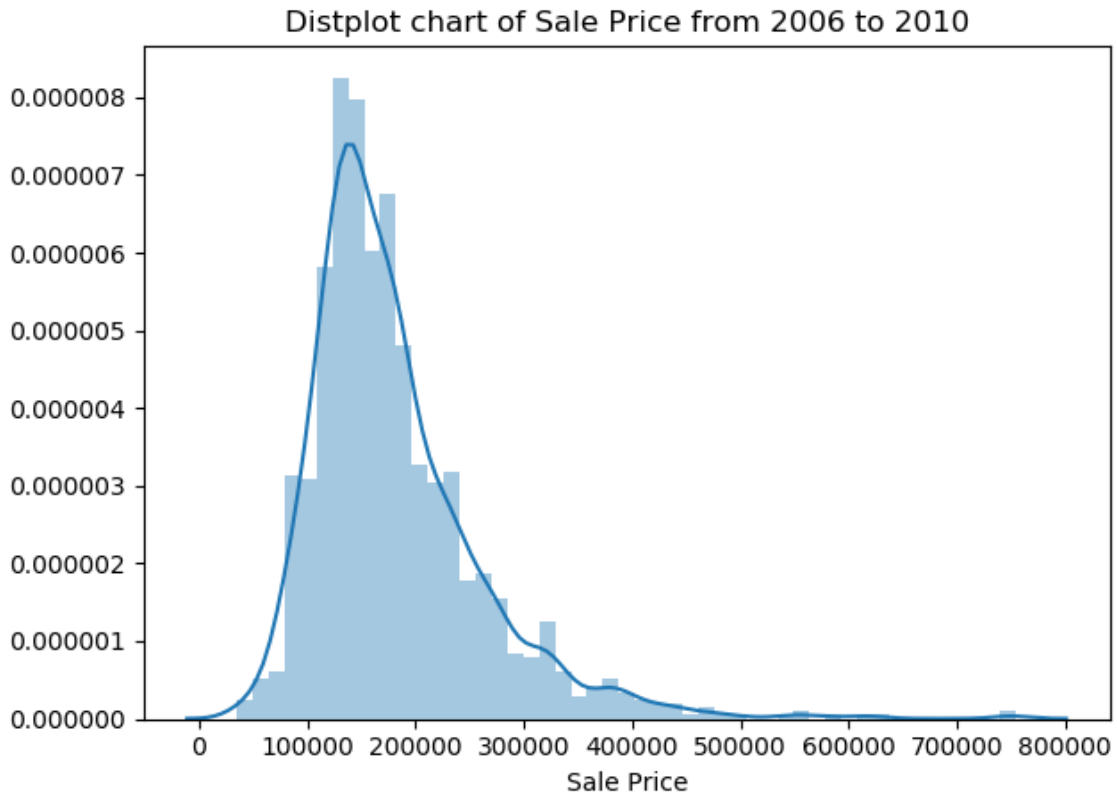
II. Preprocessing:

General data: 1454 x 54

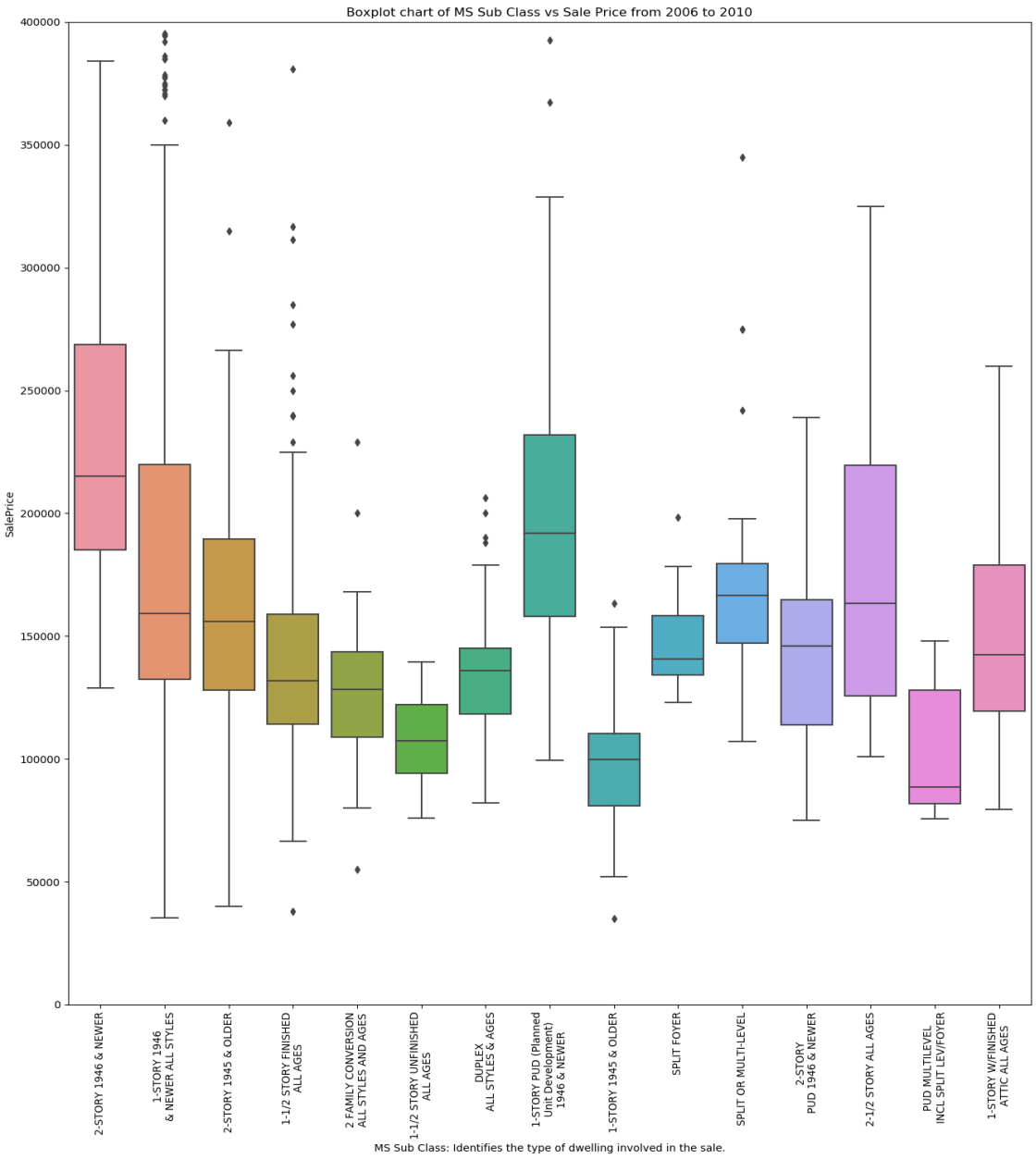
+ Train: 1018 x 54

+ Test: 436 x 54

Sale Prices mainly focus on a range from US$ 100,000 up to 200,000 through the 6 years. Therefore, this chart shows an initial glance of a price range with high spendings in the housing market.



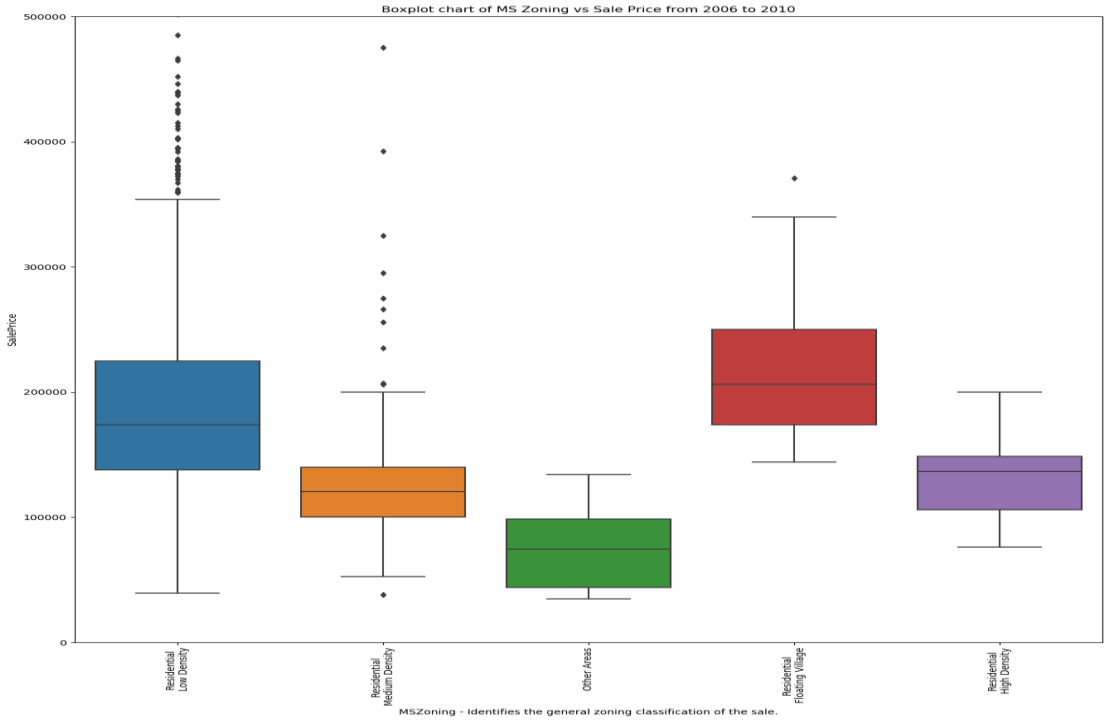
1/ MS Sub Class: Identifies the type of dwelling involved in the sale



Generally, this variable identifies the types of dwellings involved in the sale are assumed to have a strong positive correlation with Sale Prices. However, Sale Prices are randomly distributed in all types of dwellings involved in the sale instead. For most, 1-STORY 1946 & NEWER ALL STYLES and 2-STORY 1946 & NEWER takes the most proportion of identities, mainly ranging from around

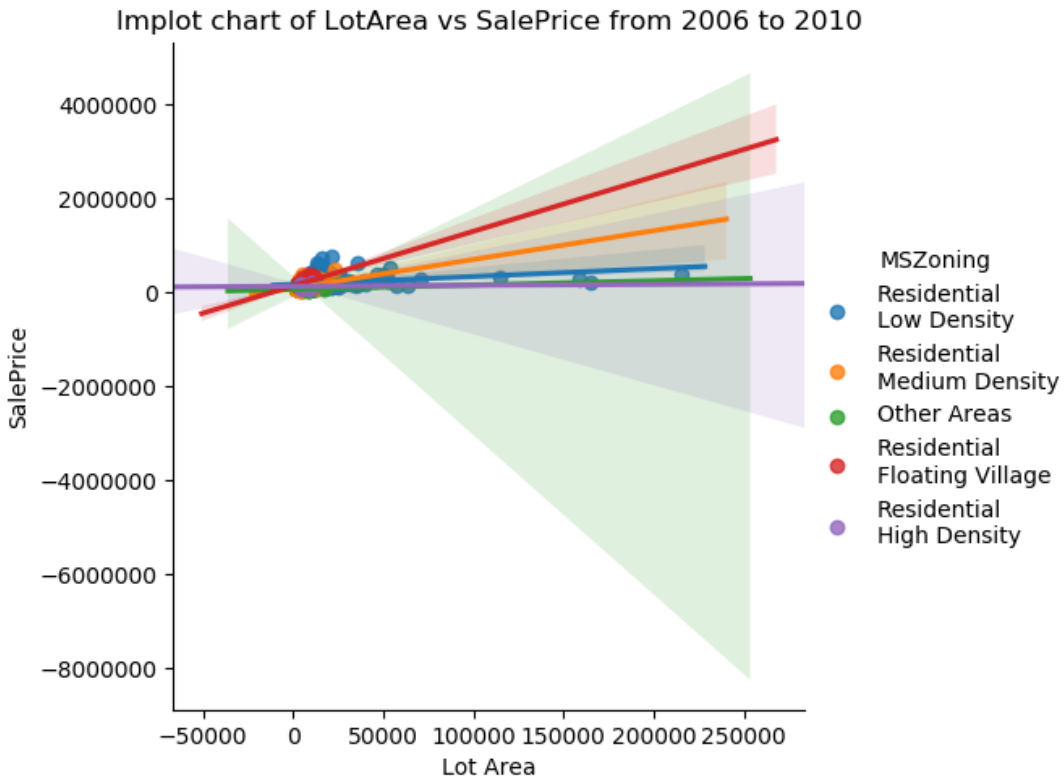
US$20,000 to slightly over 400,000 per property for 1-STORY 1946 & NEWER ALL STYLES. While it takes a span from around US$120,0000 to 460,000 per property for 2-STORY 1946 & NEWER. This proves that homebuyers process home deals with any kind of property, especially 1-story and 2-story houses. We can see that the weak negative correlation between MSSubClass vs SalePrice reaches only at -0.084.

2/ MSZoning - Identifies the general zoning classification of the sale

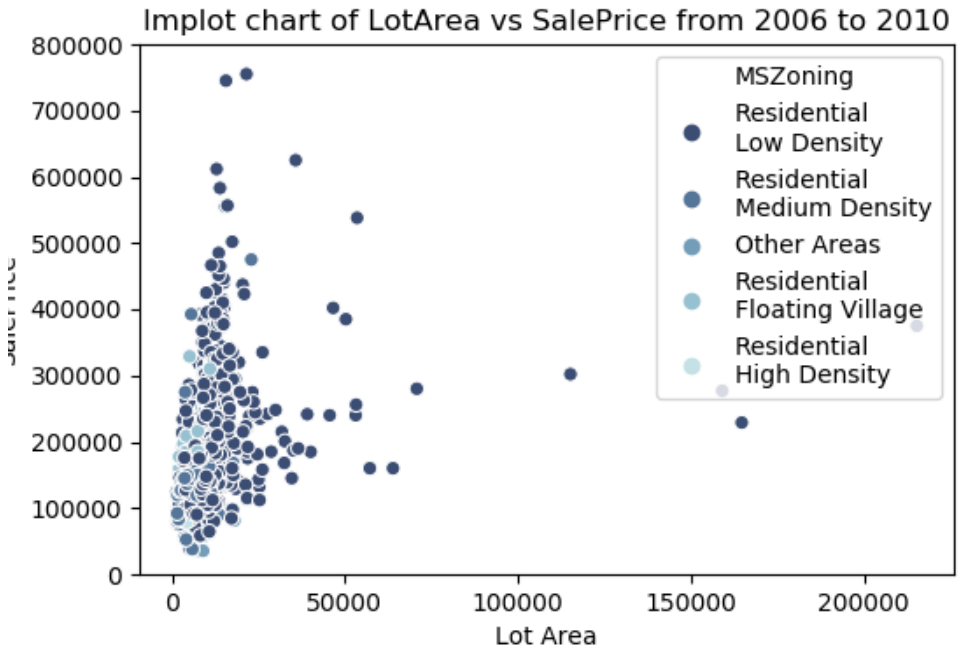


As a whole, homebuyers aimed at a property rate starting from around US$100,000 per property in all zones except some areas such as Agriculture, Commercial, Industrial and Residential Low-Density Park with q25 beginning from US$50,000 while their q75 can reach nearly US$ 100,000. To be more specific, Floating Village Residential takes the greatest Sale Prices with q25 arriving at nearly US175,000 and q75 extending up to US$250,000. Meanwhile, Residential Low Density takes second place with q25 almost extending to US$150,000 and q75 getting as far as approximately US$225,000. All in all, most house deals distribute at and above US$100,000 per property proves that house values are at high rates for a majority of homebuyers.

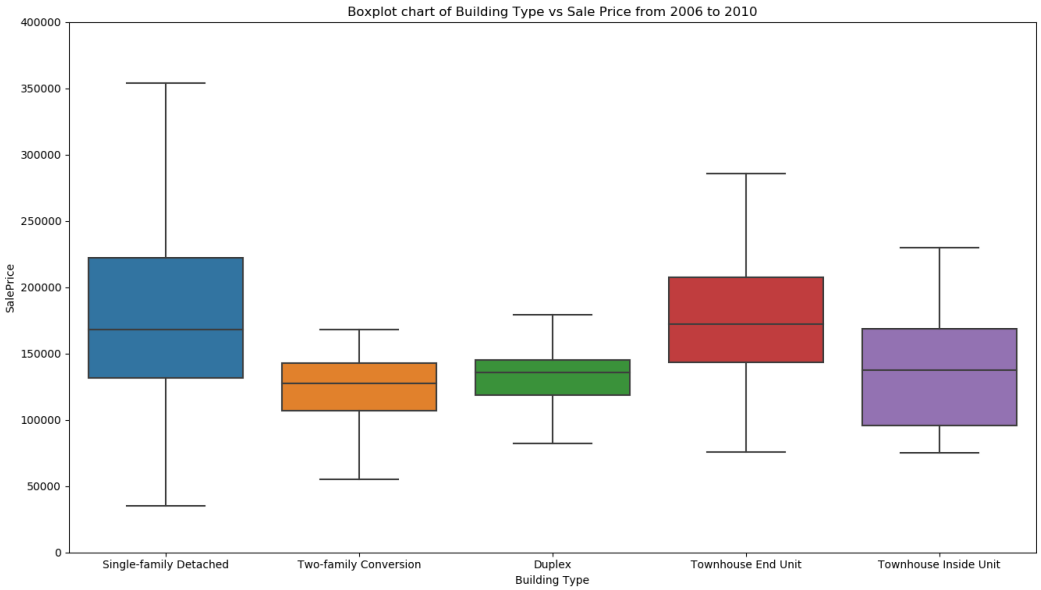
3/ LotArea - Lot size in square feet



As we can see intuitively, a great number of house merchants take place in properties below 50,000 square feet per property. These sizes cover up a range of prices largely below US$400,000. It is supposed that the larger the lot area, the greater the price. However, this is proved uncertain as the Lot Area has a small positive correlation with Sale Price at an index of 0.264. This means that Lot Area would not affect the overall Sale Price even if the lot size could expand further. And this also means that in the housing market, the Lot Area can be larger but it has a minor impact on any house values.



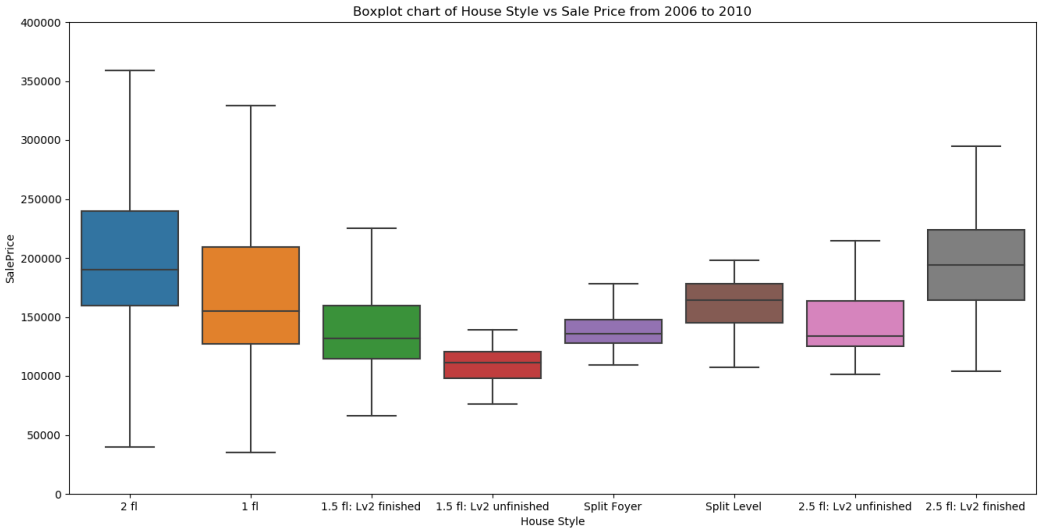
4/ BldgType - Type of dwelling



Another proof to confirm that Sale Prices all start from US$100,000 and above in all building types is the chart of Building Type vs Sale Price. It is highlighted that the Single-family Detached building

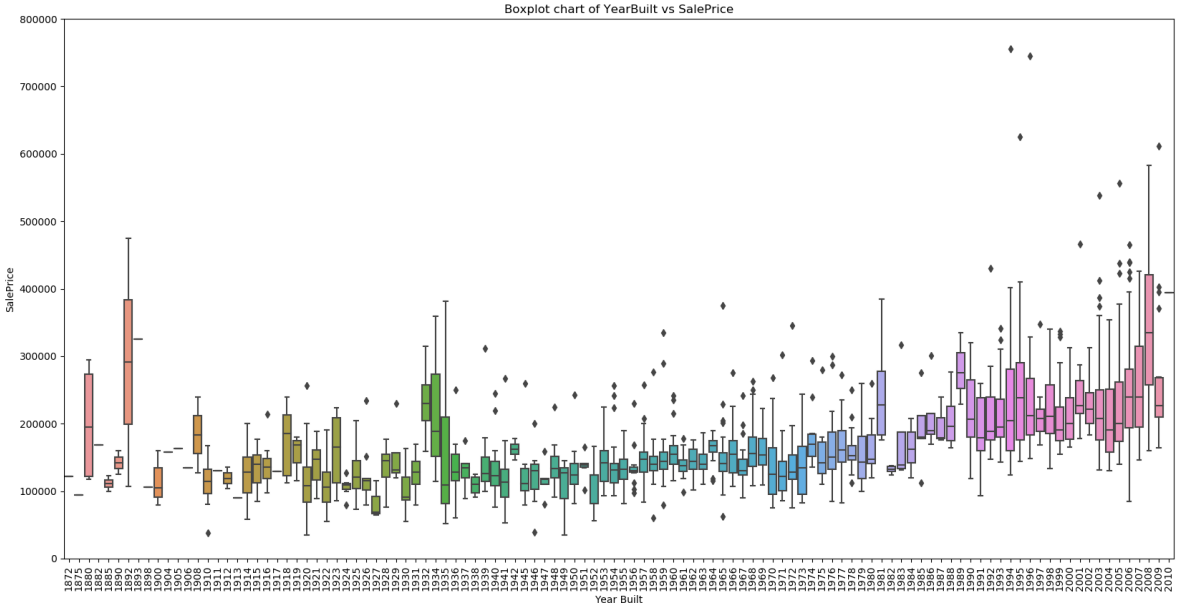
takes the longest range of Sale Price with q25 above US$130,000 and q75 closely to US$ 225,000. Followingly, Townhouse Inside Unit and Townhouse End Unit come at the next places. These are the most common properties for single families with a decent lot area below 50,000 square feet. The more separated between a family and another for a house, the higher the price. Therefore, homebuyers tend to buy more single units than combined or shared properties with other families.

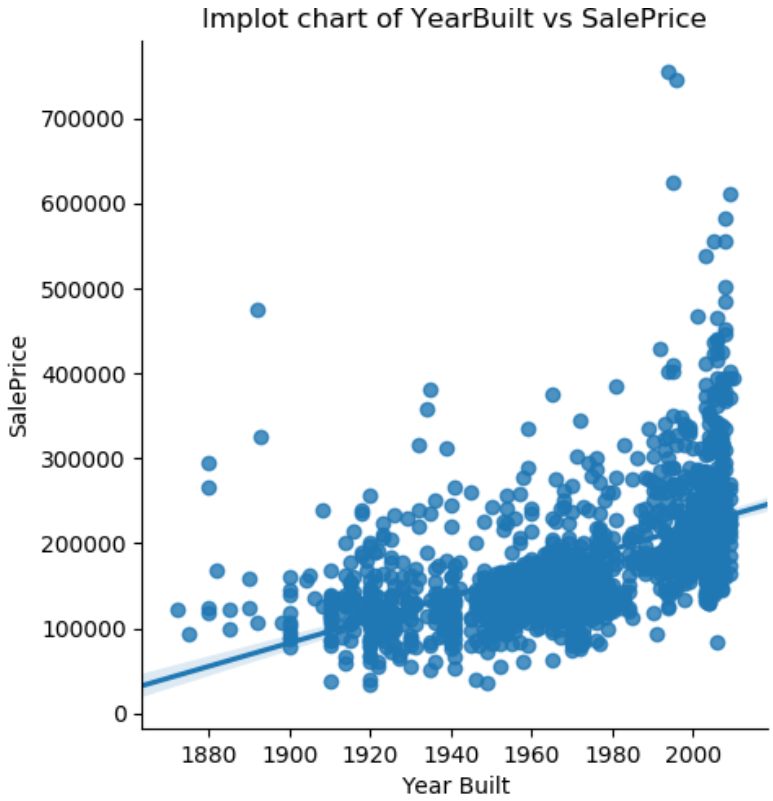
5/ HouseStyle - Style of dwelling



This chart also illustrates that all house styles have a range of prices mostly starting from US$100,000 and above no matter what statuses they are or how many stories they have.
We can easily see that more stories in a single unit and/or finished status of the house can lead to higher prices for homebuyers. Specifically, a 2-story house takes the longest range and it is also the most expensive style when its q25 begins above US$150,000 and q75 ends up nearly at US$ 250,000. Another thing to mention is that two and one-half stories: 2nd level (finished) and One story (finished) take high prices as well with their q75 in the mid US$210,000-230,000. In the meantime, unfinished houses such as the one and one-half story: 2nd level unfinished is the least preferred with both q275 and q75 below US$ 120,000.

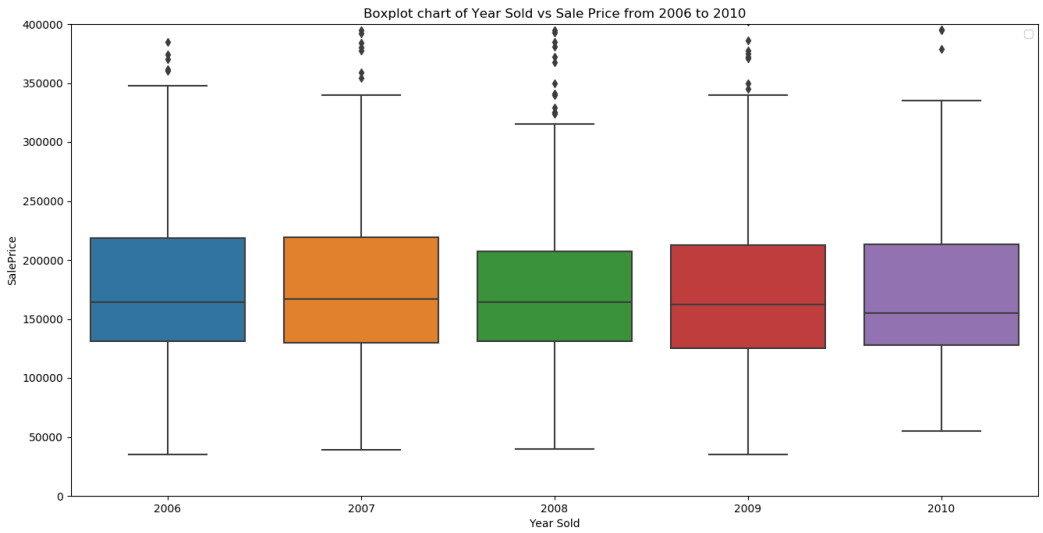
6/ YearBuilt - Original construction date





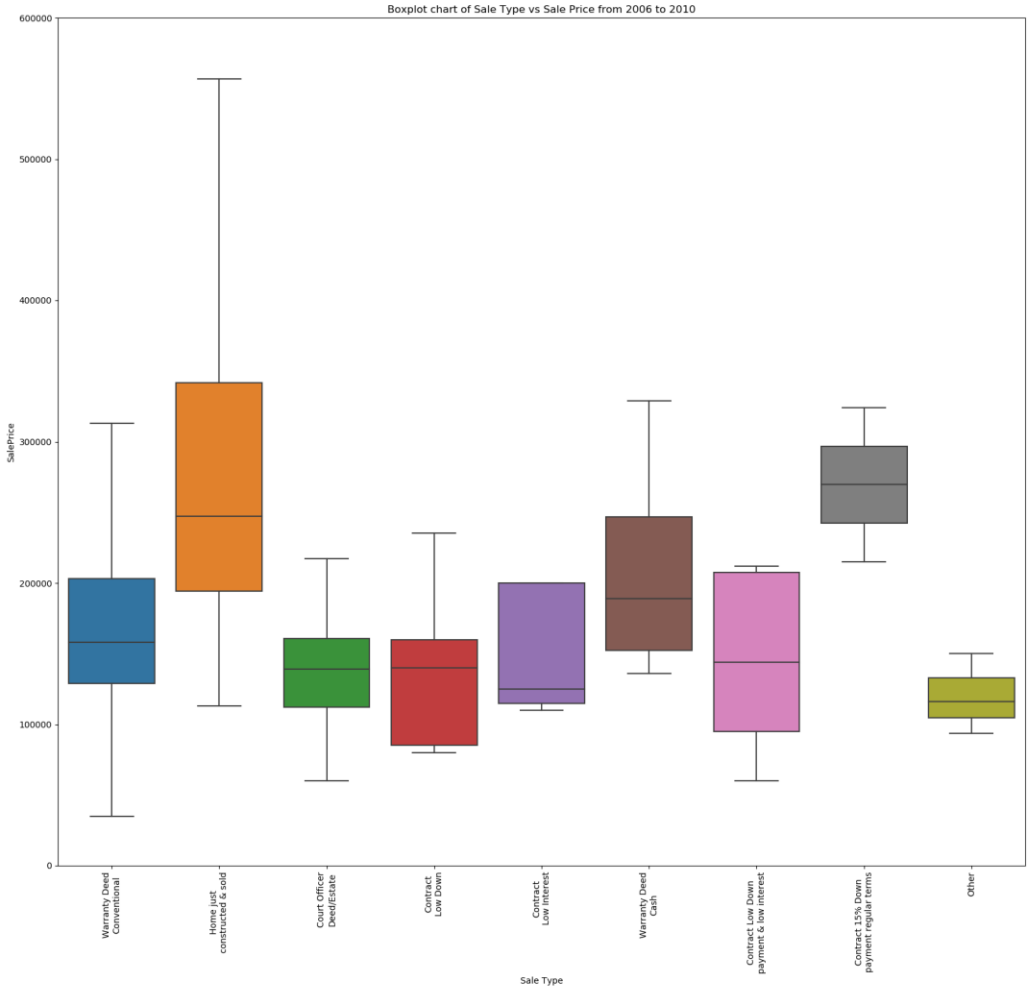
According to the lm chart, there is a moderate positive correlation between Year Built and Sale Price from 1872 to 2010 at 0.523. Some highlights can be mentioned that there were several long-term downturns in newly built houses throughout World War I (1914-1918), the Great Depression (1929-1933), and World War 2 (1939-1945). In contrast, during the Economic Crisis (2007-2008), houses built in these years and the previous period have a surge in Sale Price as demand rises with more lax mortgages in the previous crisis.

7/ YrSold - Year Sold



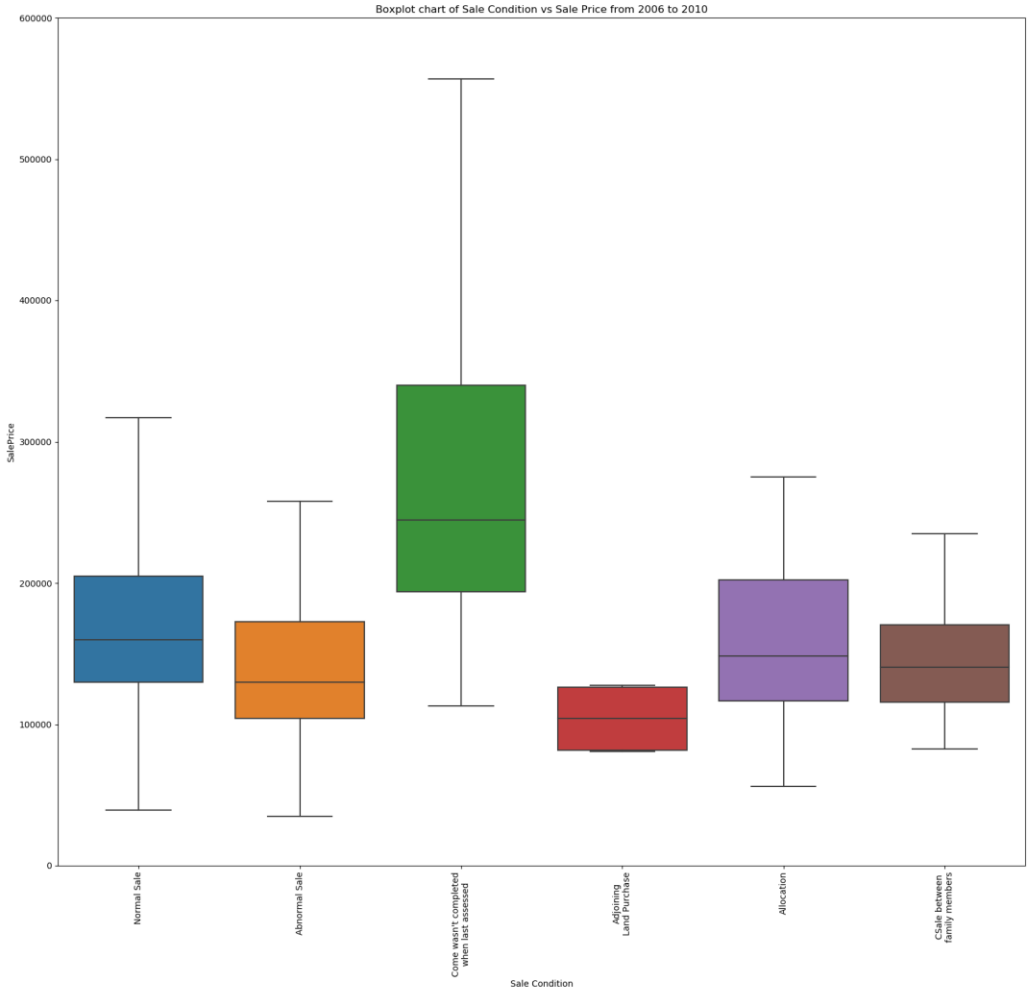
Logically, SalePrice would increase due to inflation annually. However, from 2006 to 2008, housing prices started to decrease, showing in q50, q75, q99, and range from q25 to q75 when comparing with 2007. The reason was that the 2007-2008 economic crisis had affected the housing market. There was also a slightly negative correlation between Year Sold and Sale Price at -0.029. This minor index was probably thanks to the Sale Price then started to recover at a slow velocity in 2009 and 2010 after United States President Barack Obama and US Congress's multiple regulatory and long-term responses. Therefore, there would be high prices generally for homebuyers to buy with newly revised mortgages.

8/ Sale Type

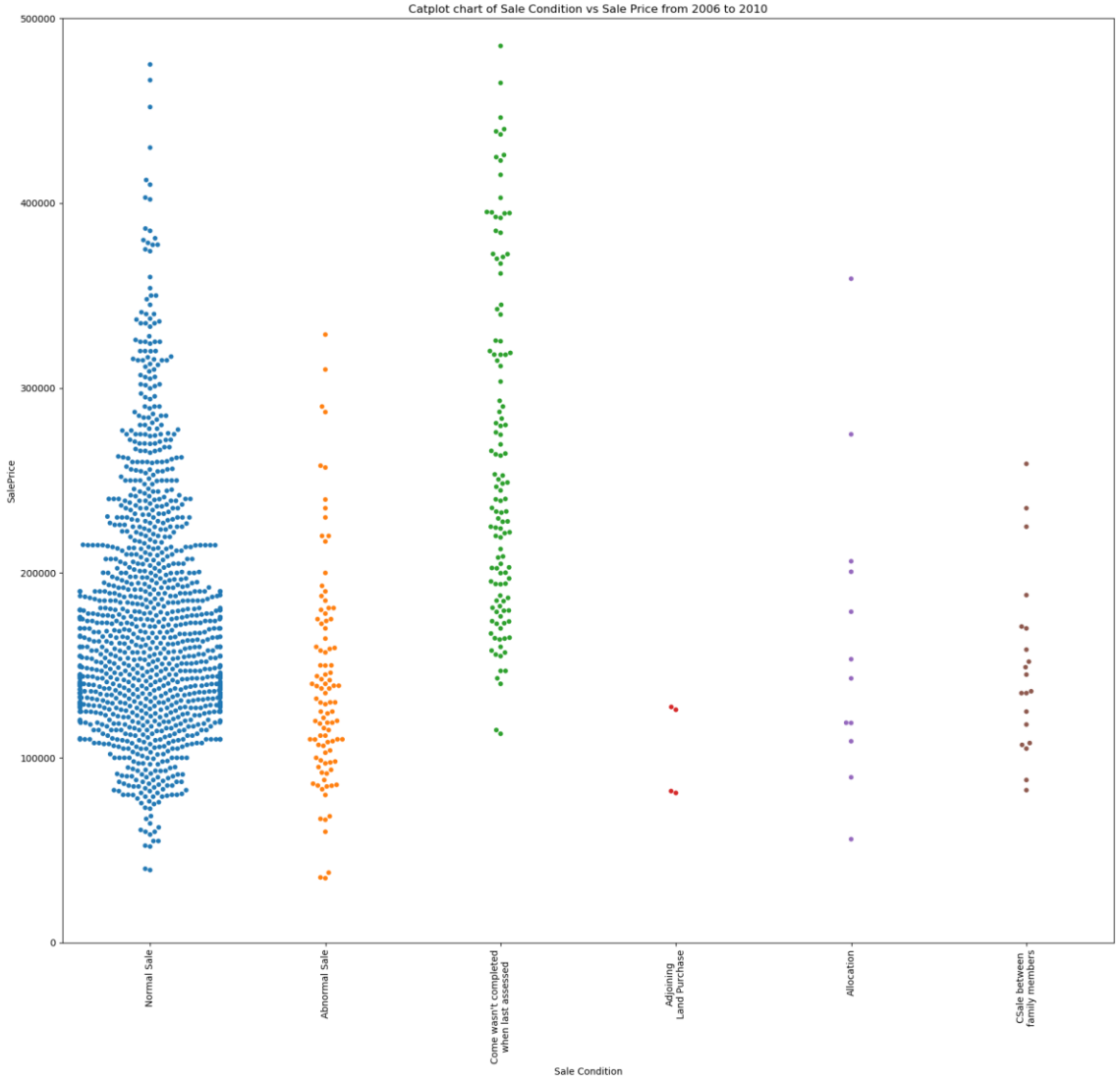


According to the boxplot chart, overall, all types of sales are over US$100,000 per property. This is at a high rate for the housing market. It seems that homebuyers tended not to buy properties at a moderate rate but spend a large number of banking loans thanks to mortgages on housing matters. Furthermore, they seemed to buy new ones at the highest range of prices. When looking at the swarm plot chart, Warranty Deed (Conventional)'s payment method took the greatest amount of housing sales, and Home just constructed and sold took the second place. This is because there were too many homeowners with questionable credit and that banks had allowed people to take out loans for 100% or more of the value of their new homes.

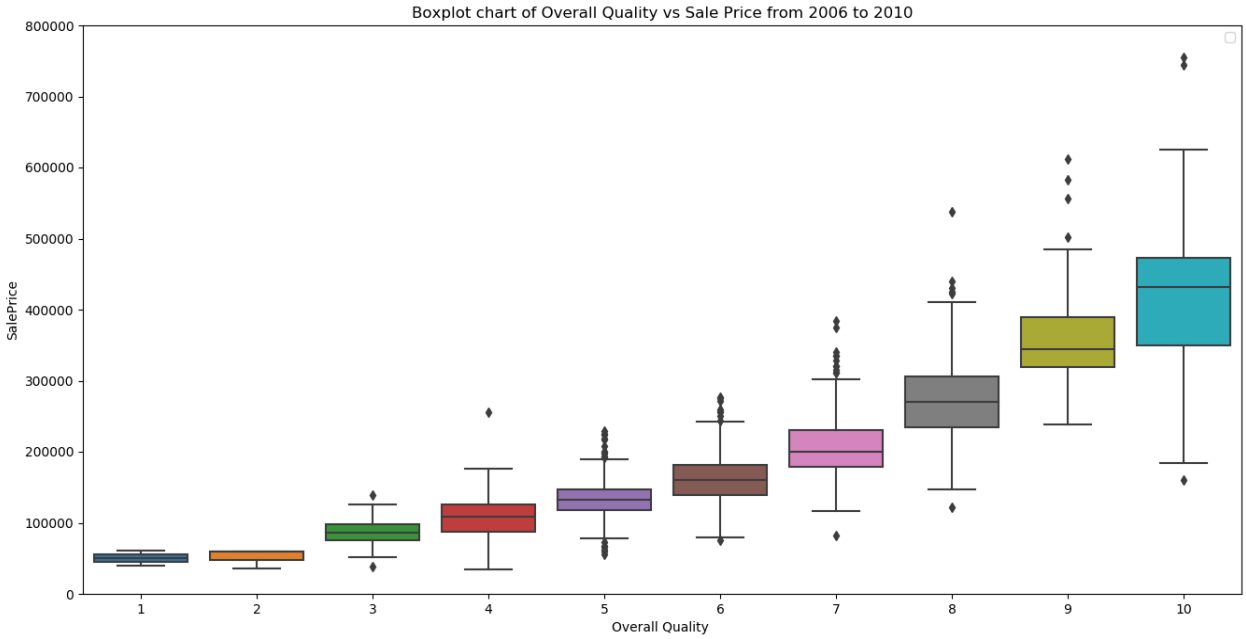
9/ Sale Condition



According to these 2 charts, housing sale cases mainly form in groups of Normal Sale, Abnormal Sale - trade, foreclosure, short sale, and Partial Sale - Home was not completed when last assessed (associated with New Homes). Specifically, Normal Sales has the most cases ranging strongly from US$ 100,000 up to 300,000 while the range of the Partial Sales fluctuates shorter from US$ 150,000 to 250,000 and the one of Abnormal Sale spans the shortest from US$100,000 to 150,000. In general, these 3 conditions all started from around US$100,000 and above which illustrates house deals are completed at a high-rate level. This also draws attention to an overview of different conditions for the majority of housing deals with normal and short processes.


Catplot chart of Sale Condition vs Sale Price from 2006 to 2010
SalePrice
500000
400000
300000
200000
100000
0
Normal Sale
Abnormal Sale
Come wasn't completed when last assessed
Adjoining Land Purchase
Allocation
CSale between family members
Sale Condition

10/ OverallQual: Rates the overall house material and completion



The Overall Quality has a very strong correlation with Sale Price as its correlation index is roughly 0.79. It seems that they have a linear relationship. In general, the higher point a property could get, the better prices its value could reach. This trend was not changed even though the assessed period when the Financial Crisis 2007-2008 took place.

III. Modeling & Evaluation:

The model evaluated on Root Mean Square Error (rsme) receives a result at roughly $38,700 which shows that this error in predicting house prices is satisfactory. The variable having the most impact on Sale Price is Overall Quality which indicates the highest correlation index at 0.79 among all features.

Homebuyers in the US should take care of the Overall Quality of their property as their main feature to consider before selling and buying. What comes next is when to do it as before the Financial Crisis 2007-2008, house prices might be higher with more flexible sale types and conditions. As the Crisis began in 2007, a drop in prices could be a great loss for sellers but a once-in-a-lifetime deal for buyers if they afforded to purchase as previous bank loans became tightened until the US President Obama pushed the US Congress for further actions aiding the US economy.

```
# METHOD 1

def featuring():
    col = ['Id', 'SalePrice', 'MSSubClass', 'MSZoning', 'LotArea', 'BldgType',
           'HouseStyle', 'YearBuilt', 'YrSold', 'SaleType', 'SaleCondition', 'Overall
    df = df[col]

#    dum_mssub = pd.get_dummies(df["MSSubClass"], prefix="mssub_")
#    df = pd.merge(df, dum_mssub[:-1], left_index=True, right_index=True) #merge data
#
    dum_mszoning = pd.get_dummies(df["MSZoning"], prefix="mszoning_")
    df = pd.merge(df, dum_mszoning[:-1], left_index=True, right_index=True)

    dum_buildingtype = pd.get_dummies(df["BldgType"], prefix="buildingtype_")
    df = pd.merge(df, dum_buildingtype[:-1], left_index=True, right_index=True)

    dum_housestyle = pd.get_dummies(df["HouseStyle"], prefix="housestyle_")
    df = pd.merge(df, dum_housestyle[:-1], left_index=True, right_index=True)

    dum_saletype = pd.get_dummies(df["SaleType"], prefix="saletype_")
    df = pd.merge(df, dum_saletype[:-1], left_index=True, right_index=True)

    dum_salecondition = pd.get_dummies(df["SaleCondition"], prefix="salecondition_")
    df = pd.merge(df, dum_salecondition[:-1], left_index=True, right_index=True)

#    dum_overallquality = pd.get_dummies(df["OverallQual"], prefix="overallquality_"
#    df = pd.merge(data, dum_type[:-1], left_index=True, right_index=True)

    df.drop(["MSSubClass", "MSZoning", "BldgType", "HouseStyle",
             "SaleType", "SaleCondition"],
             axis=1, inplace=True)

    return df
```

Name	Type	Size	Value
etype	DataFrame	(1457, 9)	Column na…
	float64	1	28790.970…
	float64	1	149780350…
	float	1	38701.466…
	DataFrame	(437, 39)	Column na…
	DataFrame	(1018, 39)	Column na…
	DataFrame	(437, 37)	Column na…
	DataFrame	(1018, 37)	Column na…
ct	float64	(437,)	[181091.2…

Help | Variable explorer | File explorer | Profiler | Static code analysis

IPython console

Console 1/A

```
In [27]: model.fit(x_train, y_train)
Out[27]: LinearRegression(copy_X=True,
fit_intercept=True, n_jobs=None,
normalize=False)

In [28]: y_predict = model.predict(x_test)

In [29]: y_test = y_test.reset_index(drop=True)

In [30]: mae = mean_absolute_error(y_test,
y_predict)
```