METCS521 Final Project Submission

Predicting Housing Prices with Linear Regression

Author: Ming Cheng, Haoran Ma

Introduction:

In this project, we need to explore the housing price and predict the trend of housing price. We need to download the dataset to Jupyter first, then obtain the data we need by deleting the redundant data, and finally predict the housing price through the prediction model. In order to accomplish this goal, we need to divide the whole project into the following four steps:

1. Downloading the dataset

2. An exploratory correlation analysis to identify ideal variables for future prediction

3. Sorting of variables by distance metric

4. Linear regression of variables and prediction of test dataset

Downloading the Dataset

In order to get the dataset we need, we need to download the CSV file we need to Jupyter and display it in the way of dataset, so that we can display all the data we need about housing price in Jupyter.

Exploration & Correlation (Extension 1):

We started with a dataset of 2919 rows * 81 columns[Fig 1].

```
housing_price_df
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2914	2915	160	RM	21.0	1936	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2006
2915	2916	160	RM	21.0	1894	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	4	2006
2916	2917	20	RL	160.0	20000	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2006
2917	2918	85	RL	62.0	10441	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	Shed	700	7	2006
2918	2919	60	RL	74.0	9627	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	11	2006

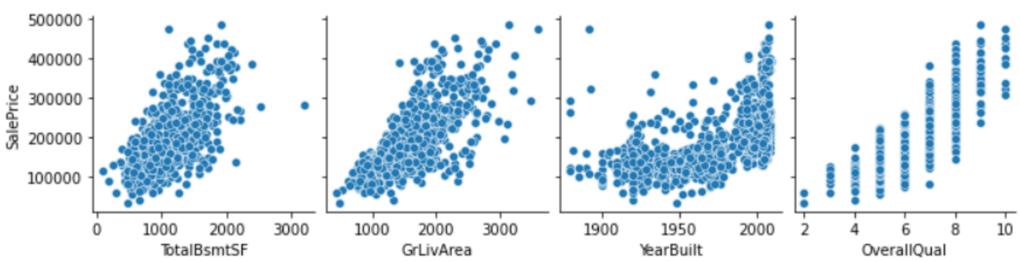
2919 rows × 81 columns

We used df. Columns to get the names of all columns and dropped the columns that were not valid for our analysis. In the following data, all null data is removed, and the resulting data is the data we need to analyze. [Fig 2]

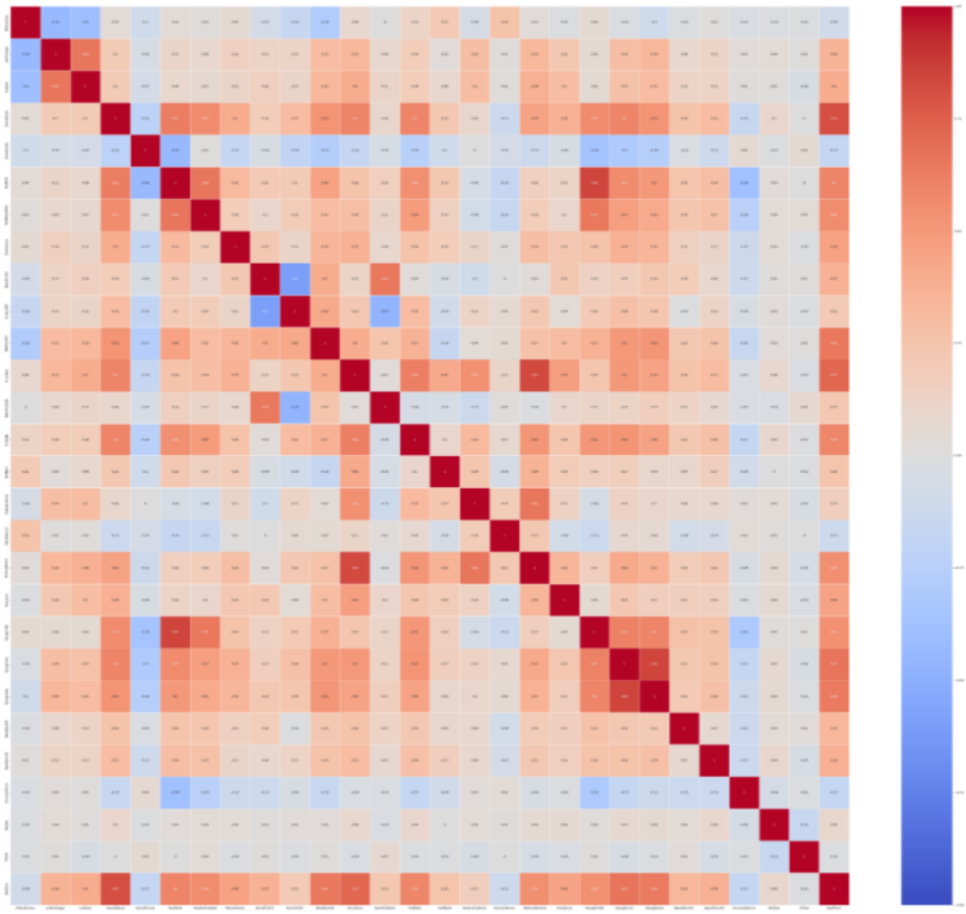
	MSSubClass	MSZoning	LotFrontage	LotArea	LotShape	LotConfig	Neighborhood	Condition1	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd
0	60	RL	65.0	8450	Reg	Inside	CollgCr	Norm	1Fam	2Story	7	5	2003	2003
1	20	RL	80.0	9600	Reg	FR2	Veenker	Feedr	1Fam	1Story	6	8	1976	1976
2	60	RL	68.0	11250	IR1	Inside	CollgCr	Norm	1Fam	2Story	7	5	2001	2002
3	70	RL	60.0	9550	IR1	Corner	Crawfor	Norm	1Fam	2Story	7	5	1915	1970
4	60	RL	84.0	14260	IR1	FR2	NoRidge	Norm	1Fam	2Story	8	5	2000	2000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	60	RL	62.0	7917	Reg	Inside	Gilbert	Norm	1Fam	2Story	6	5	1999	2000
1456	20	RL	85.0	13175	Reg	Inside	NWAmes	Norm	1Fam	1Story	6	6	1978	1988
1457	70	RL	66.0	9042	Reg	Inside	Crawfor	Norm	1Fam	2Story	7	9	1941	2006
1458	20	RL	68.0	9717	Reg	Inside	NAmes	Norm	1Fam	1Story	5	6	1950	1996
1459	20	RL	75.0	9937	Reg	Inside	Edwards	Norm	1Fam	1Story	5	6	1965	1965

1075 rows × 56 columns

Next, we visualized the data, taking SalePrice as y value and 'TotalBsmtSF', 'GrLivArea', 'YearBuilt' and 'OverallQual' as X value, we obtained four scatterplots. Drop outliers greater than 2700 in totalBusmtSF and less than 1900 in Year Build. [Fig 3]



Before using linear Regression Model to predict the holiday, we need to do a correlation analysis to find out which variables have the strongest correlation with housing price. In the figure below, we can see that some variables have strong correlation with sales price (0.5).[Fig 4]



Linear Regression (Extension 2):

For the prediction part, we need to handle all categorical variables firstly and filter them by comparing the levels of each categorical variables with SalePrice through histograms. As a results, we narrow down the categorical variables and map encode ‘ExterQual’ ‘BsmtQual’ ‘ KitchenQual’’ SaleCondition’ into numerical variables. As for the features selection, we narrow the scope of features by applying RFE Regression and put the selected features into sm.OLS model. After that, we use the train_test_split method to make linear regression prediction model(data frame is not so large and we set test_size=0.3) and calculate mean_abs_error which is 19783.92. From the sm.OLS model, we can see that there are some variables with p-value larger than 0.05 which means that these variables are not likely related to y-variable(SalePrice) and we drop these variables. After applying new linear regression prediction model with same train_test_split method and calculating new mean_abs_error, we can see that value of mean_abs_error is decreasing to 19363.9.

Conclusion:

The overall logic behind our prediction code is to select most helpful features. With applying histograms, VarianceThreshold, RFE, we actually create a smaller subset of all variables. After building every linear regression prediction model of every different feature selection model, we compare the mean_abs_value and select the subset with smallest value. The last result of our code is showing that there are about 10-15 variables highly related to SalePrice prediction model. It turns out the importance of features selection.