METCS521 Final Project: Predicting Housing Prices with Linear Regression

Authors: Tomáš Horníček, Wangzhen Li

Introduction

For our project, we took a large dataset, consisting of nearly 3000 house entries, and tried to build a linear regression model, that would help us predict the housing prices in the coming years. Along with creating a linear regression model, we have also explored different variables, which affect our main dependent variable: Sale Price. We have created graph visualizations for these variables, in order to get a good idea of how different houses, with different features, will cost in the future. For our project we have created three python files; *download_housingdataset.ipynb*, *graphs.ipynnb* and *Linear Regression Model.ipynb*.

For the user to run our program and explore how we obtained our data; they need to take the following steps:

1. Run the *download_housingdataset.ipynb* to download the dataset.
2. Run the *graphs.ipynnb* to explore how we manipulated the dataset, to obtain the results we have and to see how certain factors affect the housing prices.
3. Run the *Linear Regression Model.ipynb*, to explore the linear regression model we built, to predict the housing prices

Downloading the Dataset

To analyze the housing prices, the user must first download the dataset. In order to do that they have to run, line by line, the *download_housingdataset.ipynb* file in *Jupyter Notebook*. The script fetches the *URL* of the dataset page and writes the file to the project directory. This file will be then read, by the *graphs.ipynnb* and the *Linear Regression Model.ipynb* files, which then work with the dataset to obtain the information we need.

Manipulating Data and Exploring Variables Affecting Housing Prices

In the file graphs.*ipynnb* file, we have explored different variables that affect the main dependent variable: Sale Price. First we had to manipulate the dataset, making it readable and useful for the project. The *graphs.ipynnb* file first reads the downloaded dataset and then *"dummifies"* the dataset. *"Dummifing"* the dataset is crucial, as we needed the dataset to contain values, which we could then compare against each other. Because the dataset consisted of numbers and strings, we first mapped the dataset to contain only numbers. We

have achieved that through giving the dataset key and value pairs, replacing the text with numbers.

Our next step was to find out what data is useful for our project. We found that out by correlating each category with all the categories and then looked at how closely each category is correlated to the Sale Price. We took the categories that have a high correlation value to be categories that have the absolute value of correlation 0.7 or higher and the categories that have low correlation to have the absolute value of correlation 0.3 or lower. We then have created a dataset without the dropped low correlation values. In our newly created dataset, we still faced a problem of having too many null values. We could've approached it in two way, either drop the houses with null values or fill in the null value with the one closes to it. Our group felt that dropping the null values would keep our findings as accurate as possible, therefore we decided to go with this option. We visualize the correlations between all variables in a heat map. See Figure 1.

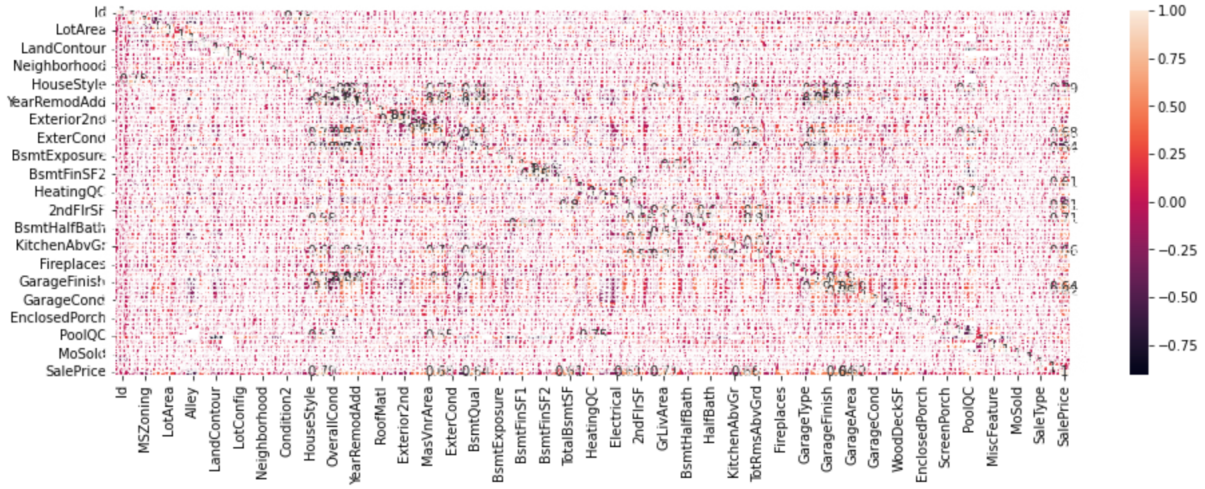

Figure 1: Garage Area(sqft) vs. Sale Price(USD)

Data Visualization

From the base dataset, that we have edited to become useful for us, we then took out variables, with high correlation to the Sale Price and that we found interesting. We then created sub datasets with these variables, and our dependent variable; Sale Price, and visualized them using graphs.

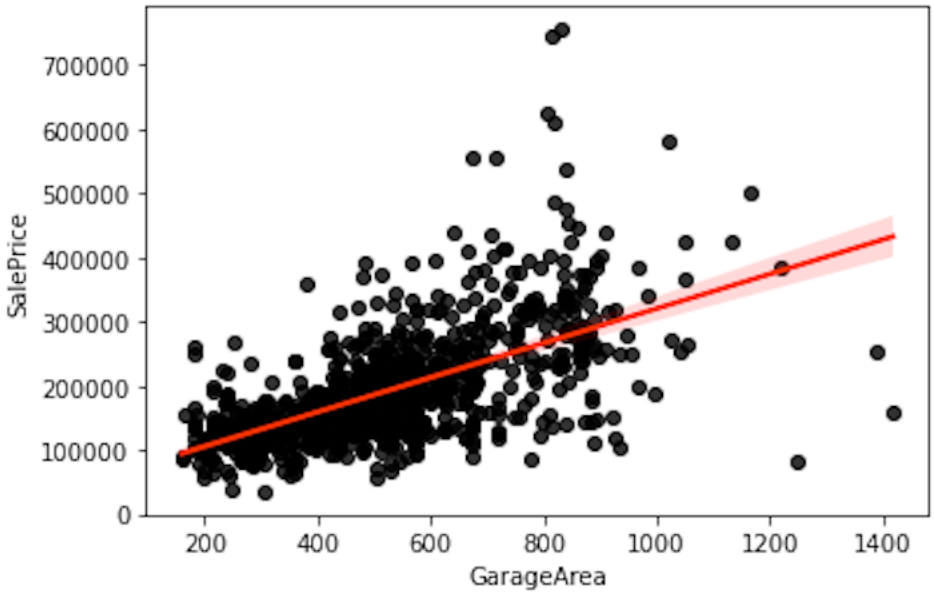

Figure 2: Garage Area(sqft) vs. Sale Price(USD)

The first graph that we visualized was *Figure2*, that showed the correlation between garage area and sale price. Based upon this we can roughly estimate the sale prices of houses based on their garage size. We can do that by obtaining the gradient of the line of best fit and plotting the gradient, with the garage area to the equation of the line, to give us the sale price of the house. There are many other factors that affect the sale price(as we can see with the anomalies that have a higher price than others and garage area ~ 800), however with this graph we can see that the bigger the garage area, the bigger the sale price.

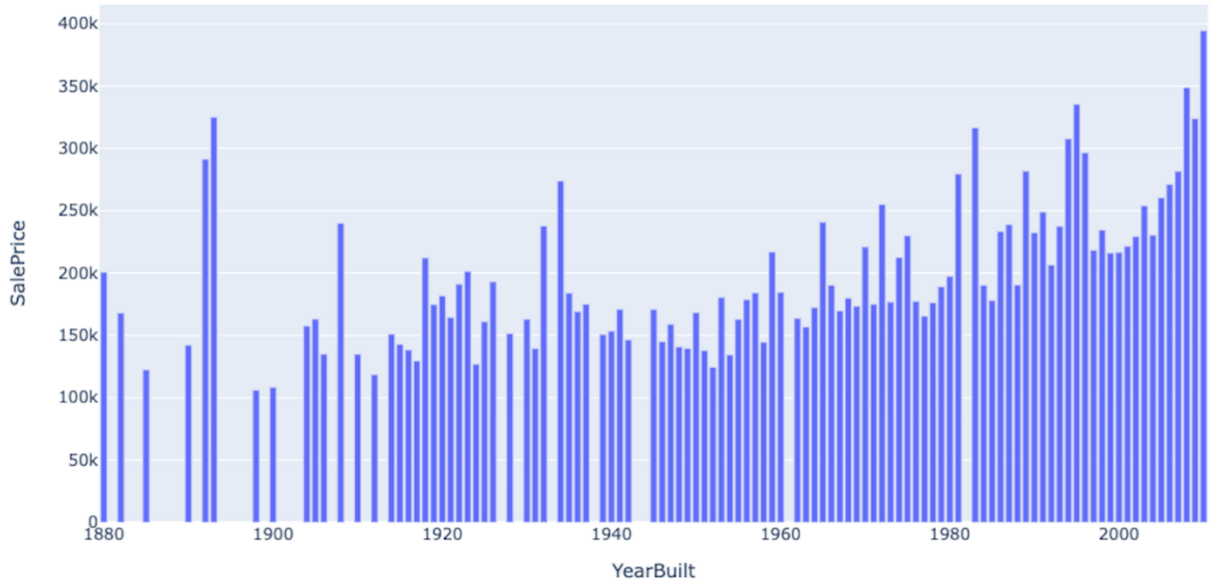

Figure 3: Houses with over 1400 sqft above ground area: Year Built vs Sale Price(USD)

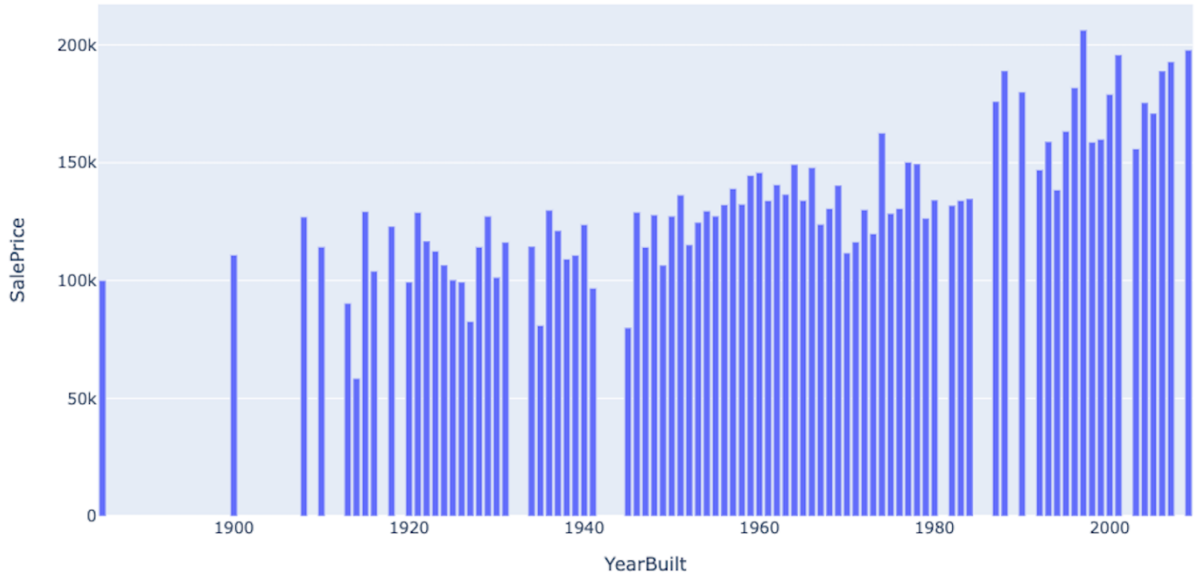

Figure 4: Houses with under 1400 sqft above ground area: Year Built vs Sale Price(USD)

Next we wanted to compare the square footage of the above ground level of the houses and how that compared to the Sale price over the years. For this graph visualization, we split the dataset into two datasets; one containing the houses with under 1400 sqft above level and one containing houses with over 1400 sqft. Next we took the average price of the houses for the given year for the two categories and compared it with the sale price. This has resulted in

graphs *Figure 3*: showing the houses with over 1400 sqft above ground area and how their sale price progressed throughout the years, and it also shows us *Figure 4:* showing the houses with under 1400 sqft above ground area and how their sale price progressed throughout the years. We then compared the two graphs and come to the obvious conclusion that the bigger the square footage above ground, the higher the sale price. However, comparing the two graphs shows us much more. From the two graphs we can read information such as how did the square footage affect the price of the houses for the given year. We can achieve that by comparing th percentage increase for the houses with over 1400 sqft and houses under 1400 sqft. For example, comparing the houses in the early 2000's we can see that the difference between houses under/over 1400 sqft is close to 100%. However, in the 1950's - 1960 's, the difference between houses under/over 1400 sqft is between 10-25%. Based upon these findings we can conclude that the square footage played a bigger role in the sale price, in the early 2000's than it did in the 1950's - 1960 's.

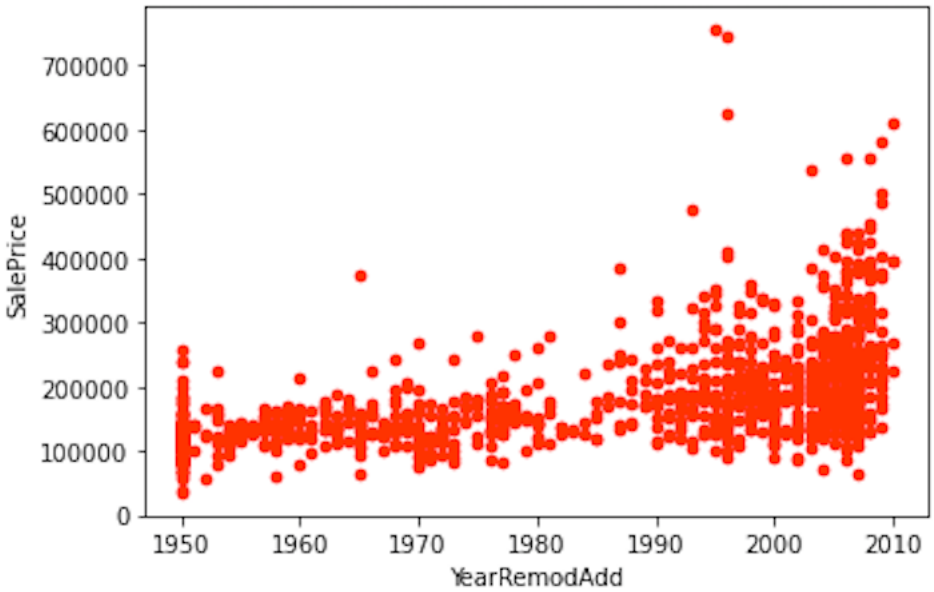

Figure 5: Year Remodel added vs. Sale Price

Next, we found it interesting to explore how the year a remodel was added to the house, affected the house price. The product of our findings can be seen in *Figure 5.* We can observe an interesting trend in *Figure 5*. We can see that, if the remodeling of the house was done before the year 1995, it didn't really influence the sale price. If the remodeling was done after the year 1995, we can see that the newer the remodeling, the higher the sale price. This could be for several reasons; we think that it is because of technology advancements in the 1990's throughout to 2010. House technology, such as fingerprint entries, security cameras, etc. was advancing at that time and logically the newer and more advanced technology, the higher the sale price.

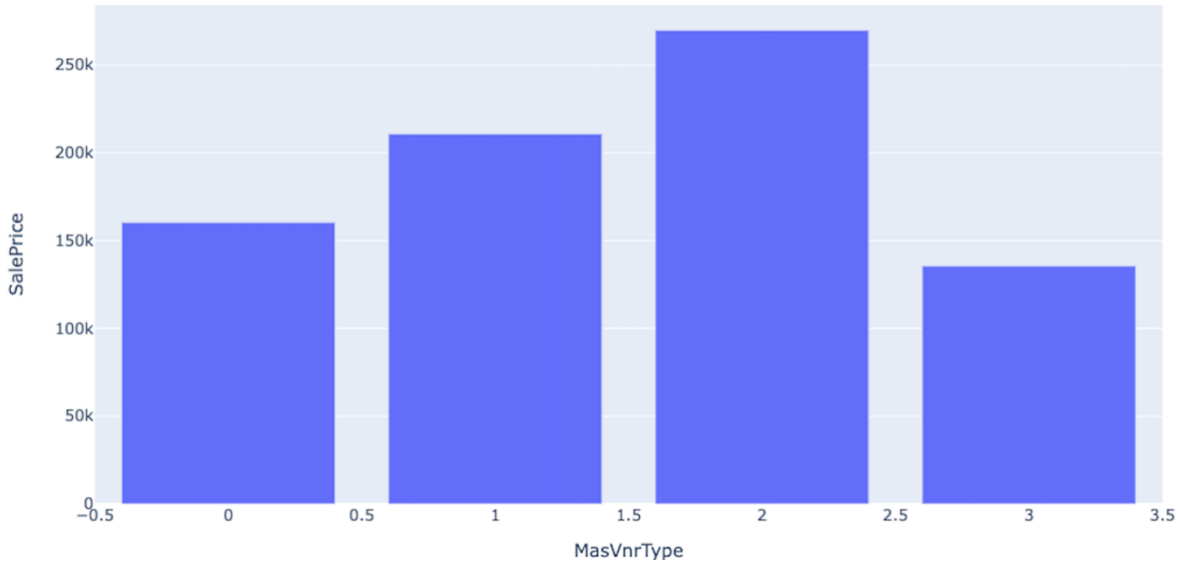

Figure 6: Masonry Veneer Type("None": 0, "Brick Face": 1, "Stone": 2, "Cinder Block": 3) vs. Sale Price(USD)

Finally in *Figure 6*, is a visualization of each masonry veneer type against the sale price. For this graph visualization, we have taken the average of all the houses having the same veneer type and compared it to the sale price. This gives us an idea of which masonry veneer type is most costly and will give us a very rough idea in predicting the price of a house based on the veneer type.

Linear Regression Model

At the beginning, we decided to start with one variable with high absolute value of coefficient among all attributes, which is "GrLivArea" with coefficient 0.708624. We built a simple linear regression model based on overall living area and sales price, as shown in *Figure 7* .

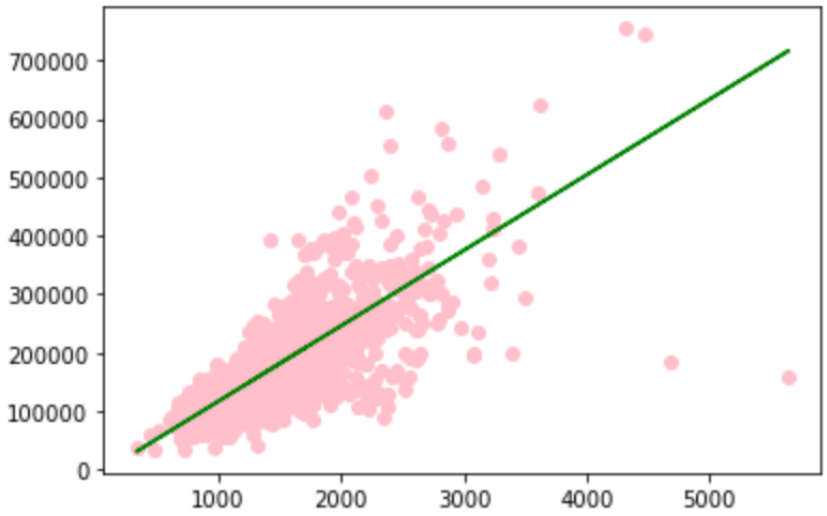

Figure 7

Now, in order to predict the housing price more accurately, we decided to try more attributes. First, we created a new data set, we used the SKLearn library to generate a linear regression model.
First, we created a new data frame with 27 attributes including "Sales_price", whose absolute value of coefficient is greater than 0.3, and then we used train_test_split from SKlearn to divide the entire data set into two parts: test and train. Due to the small amount of overall data, we decided to train as much data as possible, so we set the test size to 0.2 and the train size to 0.8. Next, we built a linear regression model with the train data and Linear Regression in SKlearn. Since both the training data and the test data are just numerical values instead of vectors at this time, we need to reshape both the test data and the training data into two-dimensional vectors. The final model tested R square value is about 0.71, which shows that the goodness of fit of the linear regression model is high, and the accuracy of the prediction is also relatively high.

Extension 1: Use the model to predict house price in real life

After building the model, we wanted to test the accuracy of the model in real-time on Zillow. Since the housing price information of our data set is from Iowa, we found a house in Iowa on sale with the following features;

Address: 2632 Pinto Ln, Iowa City, IA 52240
6 rooms above ground(3 bedrooms 3 bathrooms)
Built in 2012
Total number of fireplaces: 1
2 Attached garage spaces
Total interior livable area: 1,591 sqft
The full detail is as below.
https://www.zillow.com/homedetails/2632-Pinto-Ln-Iowa-City-IA-52240/123317216_zpid/?

As you can see in *Figure 8*, the real sale price on Zillow is $180,000. After we enter the listed information of the house, the predicted price, based on our linear regression model, is $182,442.07768641, which is very close to the real price. This proves that our model is highly accurate, as our model was off by about $2,500.

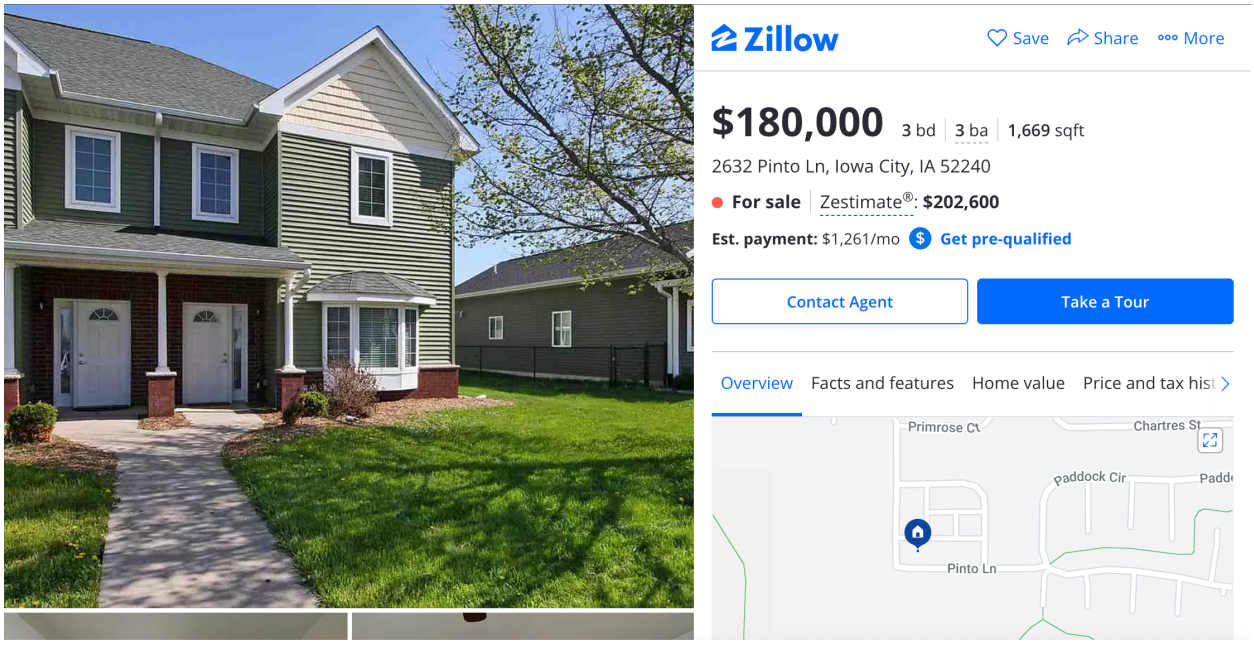

Figure 8

Extension 2: Use Osl to build the model

In the process of modeling with SKlearn, we learned that OLS is also a very useful tool for building linear regression models. Therefore we decided to model with OLS, in order to explore how different variables of the same categorical type , such as “Area/Square feet” and “Quality”, impact the final sales price.

First, we created a new data set that included all the data about "Quality" and mapped categorical variables to numbers according to a specific rule (the better the quality, the higher the value). Poor is 0, Fair is 1, Typical is 2, Good is 3, and Excellent is 4. The final data set includes "ExterQual", "BsmtQual'", "HeatingQC", "KitchenQual", "FireplaceQu", "GarageQual". Then we added these numbers together to calculate the overall quality score "Total_quality" for each house ". Then we calculated the correlation coefficient between "Total_quality" and "Sales_price" which ended up to be 0.70. This number is considered to be a relatively high coefficient, therefore, we assumed that the quality of External, basement, heating, kitchen, fireplace, garage are very likely higly correlated with the sales price and therefore have a high impact on the sales price .

Next, we built a linear regression model3 with OLS. The model profile is shown in *Figure 9*. We can see that the R-square is 0.559. This is not very ideal value. When we check the p-value, we can see that, the p-value of HeatingQC(0.793), FireplaceQu(0.181), GarageQual(0.246) is too

high(>0.05). According to knowledge we had in Business Analytic class, we assumed that the p-value value will be improved as the variables with too high p-values are removed. But when we remove these variables, the R-square decreases by 0.003 to 0.556. We are disappointed with this change, but the result also shows that if there is too few variables available, we cannot build an accurate model with them.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.559
Model:	OLS	**Adj. R-squared:**	0.554
Method:	Least Squares	**F-statistic:**	121.9
Date:	Tue, 10 May 2022	**Prob (F-statistic):**	2.99e-99
Time:	00:13:41	**Log-Likelihood:**	-7278.2
No. Observations:	585	**AIC:**	1.457e+04
Df Residuals:	578	**BIC:**	1.460e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.142e+05	2.53e+04	-4.518	0.000	-1.64e+05	-6.45e+04
ExterQual	3.826e+04	7103.758	5.385	0.000	2.43e+04	5.22e+04
BsmtQual	3.773e+04	4949.137	7.624	0.000	2.8e+04	4.75e+04
HeatingQC	-912.1069	3473.008	-0.263	0.793	-7733.360	5909.147
KitchenQual	3.513e+04	5708.054	6.154	0.000	2.39e+04	4.63e+04
FireplaceQu	5112.7898	3814.774	1.340	0.181	-2379.720	1.26e+04
GarageQual	1.257e+04	1.08e+04	1.161	0.246	-8683.384	3.38e+04

Omnibus:	258.133	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2054.601
Skew:	1.760	**Prob(JB):**	0.00
Kurtosis:	11.480	**Cond. No.**	72.1

Figure 9

The result is not as good as expected. Therefore we guessed that it could be the interaction between variables that impacts the goodness of fit. So we decided to try adding interaction effect test (ExterQual:GarageQual and KitchenQual:FireplaceQu). Because we found that, using our common sense, the External Quality and Garage Quality is highly correlated. The better the External Quality, the better the Garage Quality. The same goes for the correlation between Kitchen Quality and Fireplace Quality.

After we added these 2 interaction effects, we surprisingly found that the R-square improved(See *Figure 10*).The p-value of ExterQual:GarageQual and KitchenQual:FireplaceQu tells us that the interaction effect test are is statistically significant. Consequently, we know that the quality of GarageQual depends on the quality of ExterQual. And the quality of fireplace depends on the quality of Kitchen. That's the unpredictable nature of an interaction effect.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.560
Model:	OLS	**Adj. R-squared:**	0.556
Method:	Least Squares	**F-statistic:**	147.2
Date:	Tue, 10 May 2022	**Prob (F-statistic):**	1.12e-100
Time:	07:48:09	**Log-Likelihood:**	-7277.5
No. Observations:	585	**AIC:**	1.457e+04
Df Residuals:	579	**BIC:**	1.459e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.592e+04	1.27e+04	-5.996	0.000	-1.01e+05	-5.1e+04
ExterQual	2.221e+04	1.15e+04	1.935	0.053	-334.098	4.48e+04
BsmtQual	3.775e+04	4940.494	7.641	0.000	2.8e+04	4.75e+04
KitchenQual	2.964e+04	7189.725	4.122	0.000	1.55e+04	4.38e+04
ExterQual:GarageQual	7597.8231	4405.609	1.725	0.085	-1055.100	1.63e+04
KitchenQual:FireplaceQu	1955.5592	1423.896	1.373	0.170	-841.072	4752.191

Omnibus:	258.954	**Durbin-Watson:**	1.989
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2074.813
Skew:	1.764	**Prob(JB):**	0.00
Kurtosis:	11.525	**Cond. No.**	57.5

Figure 10

Next, we tried to model the "Area" variable in the same way. The variables used in Model 4 are "MasVnrArea", "GrLivArea", "TotalBsmtSF", "GarageArea", "SalePrice". The correlation coefficient of "TotalSQFT" and "SalesPrice" (0.7612) shows that total square feet are highly related to sales price, which is higher than the correlation coefficient of quality to price. When we built the model with OLS, the resulting profile is shown in *Figure 10*.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.605
Model:	OLS	**Adj. R-squared:**	0.602
Method:	Least Squares	**F-statistic:**	222.3
Date:	Wed, 27 Apr 2022	**Prob (F-statistic):**	1.58e-115
Time:	23:14:58	**Log-Likelihood:**	-7245.6
No. Observations:	585	**AIC:**	1.450e+04
Df Residuals:	580	**BIC:**	1.452e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.685e+04	9419.963	-1.788	0.074	-3.53e+04	1655.253
MasVnrArea	51.5051	12.201	4.221	0.000	27.541	75.469
GrLivArea	66.0107	5.285	12.489	0.000	55.630	76.392
TotalBsmtSF	37.9567	6.468	5.868	0.000	25.252	50.661
GarageArea	126.4957	15.595	8.112	0.000	95.867	157.124

Omnibus:	373.940	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23616.060
Skew:	-2.057	**Prob(JB):**	0.00
Kurtosis:	33.853	**Cond. No.**	9.02e+03

Figure 10

The R-square for Model 4 is 0.605. Goodness of fit is greater than model 3. So we can conclude that the correlation between house area and sales price is stronger than the correlation between house quality and sales price, and it is more in line with the law of linear regression. When building a linear regression model, more consideration should be given to the Square feet of building area of the house.

Conclusion

All in all, in the housing price predicting project, we firstly manipulated the data, including cleaning the N/A values in the data, dummifying the categorical data, and visualizing some of the variables in the data set. Next we explored the variables that affect sales price. We calculated the correlation coefficient between each variables and the sales price and selected the data whose absolute value are greater than 0.3 to create a new data frame. Then, we divide the data into two parts, test and train. Using SKlearn, OLS to build different linear regression models on the data set and learned some very interesting findings. And we even tested the housing price prediction model with house in real world on Zillow. In the end, we came to the following conclusions.

1. When the amount of data is not very large, the test size should be set as small as possible.
2. Both SKlearn and OLS are useful machine learning tools that can well build linear regression models on data sets. But in contrast, OSL can generate a summary profile, which can clearly view the full range of data of the model, including the R-square and the p-value and coefficient of each variable, which can help us fine-tune the model.
3. In the process of building a linear regression model, the number of variables affects the accuracy of fit. The more variables there are, the better the fit is overall. Because the influencing factors are multivariate relative to the sales price, the improvement of the overall accuracy of fit by a single type of variable is limited.
4. The correlation coefficient is a very effective indicator for selecting variables. Data with an absolute value greater than 0.3 can be regarded as a potential factor affecting the final selling price.
5. When the R-square is not as expected, we could consider adding an interaction effect test to the variables with high p-values to adjust the accuracy of fit. Because there might be correlations between variables. When considering how to add interactions, we need to think about the meaning of variables and common sense.
6. Analyzing our different variables, that affect the sales price, we also came to interesting conclusions. We found out that the newer the year the house was remodeled, the higher the price, but that only goes for renovations done after 1995(*Figure 6*). We also found that the most expensive masonry veneer type is stone (*Figure 5)*. Next, we found out, when comparing *Figure 3* and *Figure 4,* that the square footage started to play a bigger role in the present, than it did in the past. Finally, we saw that the bigger the garage square footage, the bigger the sale price, which was predictable (*Figure 2*). These findings give us a rough estimate, when we want to quickly predict the price of a house, based on its properties.