Individual Assignment 1

Eran Kan Kohen

15th September 2022

Word Count: 1965

1 Collect and Prepare Data

1.1 Summary

The following paper will construct a theory to illustrate the predictors of house prices in the US. The data made available by Kaggle was initially investigated. Several variables with the type "character" were converted into factor variables. A summary of the variables relevant to the constructed theory is present below:

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
SalePrice	1,460	180,921.200	79,442.500	34,900	755,000
YearBuilt	1,460	1,971.268	30.203	1,872	2,010
GarageCars	1,460	1.767	0.747	0	4
GarageArea	1,460	472.980	213.805	0	1,418
X1stFlrSF	1,460	1,162.627	386.588	334	4,692
X2ndFlrSF	1,460	346.992	436.528	0	2,065

The table gives insight on the main variables which are going to be used in the analyses. Interestingly, both the average 1^{st} floor square foot and the sale price are quite high. Moreover, when looking at the minimum and maximum values for the variables, it shows that the market has all kinds of houses. For example, the age range of the houses are 138, as the oldest and the newest house were built on 1972 and 2010, respectively.

Also, it is quite tragic that the maximum garage area in the data is larger than the mean first floor area, which means that several cars in the US reside in a more spacious place than the average person.

As seen by the table, several houses do not have a second floor. Hence, the square foot in both floors were combined together to create a new variable called TotalSF. The new variable has a mean of 1510, minimum value of 334, and maximum value of 5642.

Moreover, the variable MSZoning is missing from the table as it is a categorical variable. A summary of the types of zoning is present below:

Table 2:

Classification	Frequency
Commercial	10
Floating Village Residential	65
Residential High Density	16
Residential Medium Density	218
Residential Low Density	1151

1.2 Plots

In order to continue the investigation, 3 separate plots were created to illustrate the relationship between possible independent variables and SalePrice.

The first figure shows the relationship between SalePrice and MSZoning. As shown in the graph, Floaring Village Residential and Residential Low Density areas have the highest mean prices. Moreover, the Commercial zone

seems to have the lowest mean price. However, considering that there are only 10 houses which are classified in the zone, the mean outcome might not be very reliable. Also, it is surprising that even though the Low density areas are quite expensive, there seems to be no difference between Medium Density and High Density.

Figure 1: MSZoning x SalePrice

The second figure is focusing on the Total Square Foot and the Sale Price. It is a scatter plot graph which illustrates the relationship between the variables. As seen from the graph below, there seems to be a positive relationship between the variables. However, it needs to be mentioned that, especially as the Year Built increases, the frequency of outliers increase as well. There seems to be a great deal of highly expensive houses built between the years of 1980 and 2010.

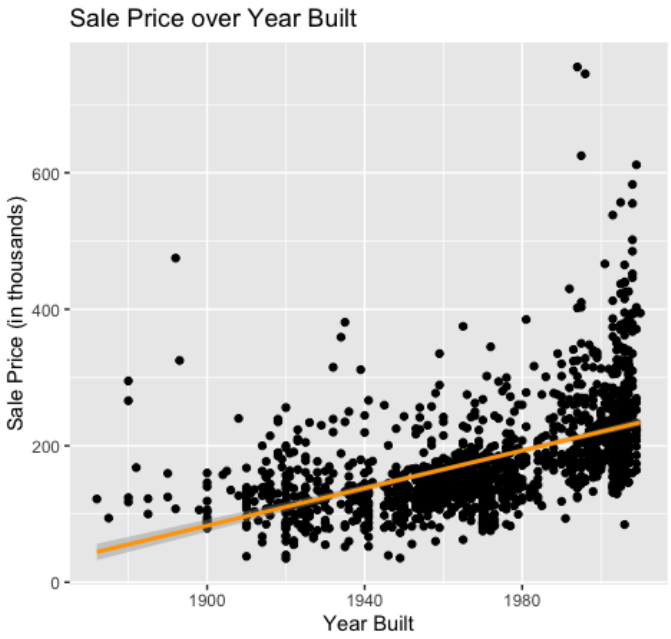

Figure 2: YearBuilt x SalePrice

The third figure is stylistically very similar to the second one as it is a scatter plot graph. It shows the relationship between the year the house was built and its price. Similar to the second graph, there seems to be a positive relationship between Total Square Foot and Sale Price. It should be noted that after a certain point, the variation in the data seems to be quite high. It shows that this relationship can struggle in predicting the sale price for bigger houses. Finally, the two points in the bottom right should be specifically highlighted. The two biggest houses in the data set happens to have close-to-average prices. This seems very weird and is probably due to an unexplored variable.

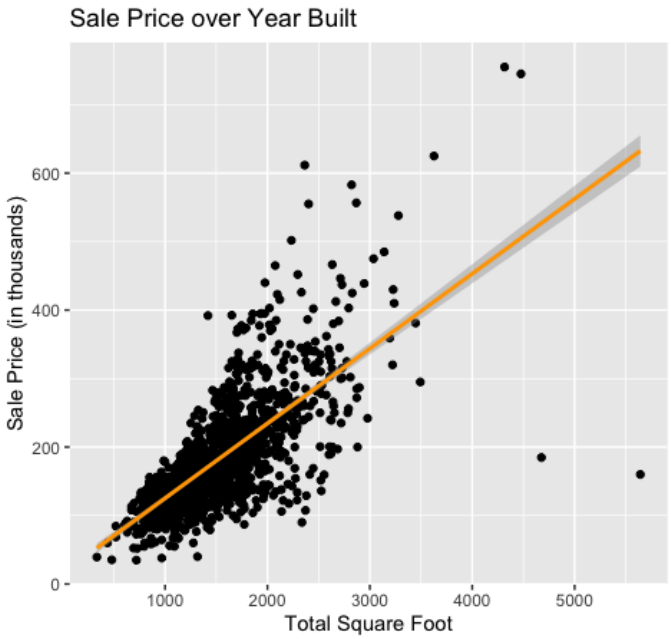

Figure 3: TotalSF x SalePrice

2 Theory and Assumptions

2.1 Theory

A 2001 study which analyzed housing prices in Ohio showed that several variables correlate with housing prices. Several of these indicators are also present in the present data set. These indicators are Structure Age, Building Square Foot, and Garage Capacity. (Bowen et al., 2001) The data set provides two indicators for Garage Capacity with one variable focusing on the area in square foot (GarageArea) and the other focusing on how many cars would fit in the garage (GarageCars).

The study included several characteristics of the neighbourhood's as well. (Bowen et al., 2001) This was added into the model through the variable MSZoning, hence it classifies the area in which the estate is present.

The initial model consisted of the variables TotalSF, YearBuilt, and Gar-

ageArea.

Hence the hypotheses were:

H_1: Bigger houses will have higher sale prices.

H_2: Newly built houses will have higher sale prices

H_3: Houses with bigger garages will have higher sale prices.

$$\begin{aligned}\text{SalePrice} = \alpha &+ \beta_1 \text{TotalSF} \\ &+ \beta_2 \text{YearBuilt} \\ &+ \beta_3 \text{GarageArea} + \epsilon, \epsilon \sim n(0, \sigma)\end{aligned}$$

MSZoning and GarageCars were used in subset analyses.

2.2 Assumptions and Remedies

There are 6 assumptions of to conduct a regression test. There is a chance that several of these will not be satisfied in this data set unless adjustments are made.

First of all, it is highly possible that the data will face the issue of mulicollinearity. The variables TotalSF and GarageArea both concern the space of the house. It is very likely that as one increases the other will increase in linear fashion. If there is a a clear linear relationship, one of the variables could be removed, or GarageArea can be changed with GarageCars.

Secondly, one the assumptions is that the predictor variables will be normally distributed. This might not be the case for TotalSF. As seen through the initial analysis, several houses in the data have only one floor, whereas some houses have two floors. Considering the wealth gap in the US (Hubmer et al., 2021), it is highly likely that there will be a number of much bigger houses than the mean. If this graph does not turn out to be normally distributed, it can be solved in different ways. The variable can be converted with the logarithmic scale.

Also, it is possible that the variable YearBuilt will not meet the homo-

skedaticity assumption. As seen in Figure 2, several houses built between the years of 1980 and 2010 have very high prices. Moreover, there seems to be an inconsistent division of residuals. The standard errors will be checked if there seems to be a violation.

Linearity is about the variables having a linear relationship. This assumption will be violated if the variables end up having a different type of relationship such as quadratic.

Exogeneity concerns the dependence of X on Y. When doing an OLS regression model equation, it is assumed that Y is dependent on X and not the other way around.

Last, data generation assumes that the data can be generated in a fixed or random way. However, since the generation of this data set is unknown, this assumption will not be evaluated.

3 OLS Regression

To test the hypotheses mentioned in section 2, 4 different regression models were run.

Table 3: Regression Comparison

	Dependent variable:			
	SalePrice			
	(1)	(2)	(3)	(4)
TotalSF	109.277***			83.430***
	(2.783)			(2.556)
YearBuilt		1,375.373***		787.476***
		(58.717)		(44.100)
GarageArea			231.646***	80.915***
			(7.608)	(6.927)
Constant	15,955.120***	−2,530,308.000***	71,357.420***	−1,535,624.000***
	(4,444.833)	(115,761.300)	(3,949.003)	(85,640.100)
Observations	1,460	1,460	1,460	1,460
R^2	0.514	0.273	0.389	0.685
Adjusted R^2	0.514	0.273	0.388	0.684
Residual Std. Error	55,405.780 (df = 1458)	67,739.660 (df = 1458)	62,135.640 (df = 1458)	44,655.670 (df = 1456)
F Statistic	1,541.515*** (df = 1; 1458)	548.666*** (df = 1; 1458)	926.951*** (df = 1; 1458)	1,053.837*** (df = 3; 1456)
Note:				*p<0.1; **p<0.05; ***p<0.01

As seen through the table above, all of the models have significant relationships, giving us confidence to reject the null hypotheses. Moreover, the R^2 values of the models reveal that the total square foot explains more than 50% of the variance in the relationship. This is followed by garage are and year built by 39% and 27%, respectively. When all 3 of the variables are included in a single model, 69% of the variance in the data is explained. Finally, as predicted, the price of the houses positively correlates with the total space, garage area, and the construction year of the house.

According to this model, a single square feet increase of the house or the garage correlates with a 83 or 80 dollar increase in the house price. It is interesting that a single square feet increase of the garage corresponds to as much value compared to a single square feet increase of the house. Additionally, the year the house was built turned out to have the greatest effect on the price, since a single year increase correlates to a 787 dollar increase in the price.

Furthermore, it is possible that YearBuilt and TotalSF can have an unexplored interaction effect, since a new and larger house can be very expensive. Hence, the regression model was tested with a new interaction variable.

As seen below, the new interaction variable didn't yield a significant relationship between YearBuilt and SalePrice. It also increase the R^2 by a very small margin. Hence, the original regression model was kept.

Finally, as seen through Figure 2, YearBuilt seems to have a quadratic relationship with SalePrice. The price seems to improve exponentially as the houses are built more recently. Hence, square of YearBuilt was added into the model. As seen through the table below, R^2 is improved from the original model, with all of the relationships keeping their significance. Thus, the new quadratic model was kept as the final version.

Table 4: Regression with Quadratic and Interaction Terms

	Dependent variable:		
	SalePrice		
	(1)	(2)	(3)
TotalSF	83.430***	−643.520***	80.212***
	(2.556)	(137.062)	(2.602)
YearBuilt	787.476***	195.865	−25,061.020***
	(44.100)	(119.780)	(4,818.613)
GarageArea	80.915***	80.374***	77.237***
	(6.927)	(6.864)	(6.896)
TotalSF:YearBuilt		0.369***	
		(0.070)	
I(YearBuiltˆ2)			6.593***
			(1.229)
Constant	−1,535,624.000***	−371,040.500	23,797,657.000***
	(85,640.100)	(235,365.600)	(4,723,136.000)
Observations	1,460	1,460	1,460
R^2	0.685	0.691	0.691
Adjusted R^2	0.684	0.690	0.690
Residual Std. Error	44,655.670 (df = 1456)	44,245.200 (df = 1455)	44,235.690 (df = 1455)
F Statistic	1,053.837*** (df = 3; 1456)	812.146*** (df = 4; 1455)	812.652*** (df = 4; 1455)
Note:			*p<0.1; **p<0.05; ***p<0.01

4 Checking Assumptions

As discussed in section 2, the OLS regression model was based on 6 assumptions. However, it is possible that several of these assumptions were not met when the model was created. This section will aim to detect such violations and correct it if possible.

4.1 Normal Distribution

The assumption of normal distribution of the independent variables was checked by creating frequency histogram and conducting a Shapiro-Wilk normality test. The initial checks revealed that neither of the variables met the requirement for normality.

For the variable TotalSF, a log transformation turned out to be fruitful as it resulted in a bell-like frequency graph. Whereas, neither YearBuilt nor GarageArea reacted positively to either log, inverse, or square root transformations.

4.2 Multicollinearity

A vif analysis was conducted among the three independent variables. As seen in the table below, no signs of multicollinearity was detected in the data since none of the scores are close to 5. Hence, no remedies were used.

Table 5: Vif Scores

TotalSF	YearBuilt	GarageArea
1.298	1.298	1.605

4.3 Heteroskedasticity

In order to make sure the model fits in with the assumption of homoskedasticity, the residuals were checked and a Breusch-Pagan test was conducted. It revealed that the variables did not meet the requirements, hence, robust and clustered standard errors were checked. The zoning classification variable "MSZoning" was used for clustering. As expected, both the robust and clustered standard errors widened since the model was heteroskedastic. The same check was done for the quadratic model but the results were indifferent.

It is important to note that the plot of the residuals showed that the issues are more persistent as the price increases. It is possible that the model will be worse at predicting higher house prices compared to lower prices.

5 Subset Analyses

5.1 Garage Space in Cars

The regression model was compared within three subsets: houses with no garages, houses with garage space for one car, houses with garage space for multiple cars.

As seen in the table below, for the set with no garages, the variable GarageArea was removed from the model, whereas for the set with garage space for one car, the relationship of GarageArea was not significant and the R^2 was much lower. The highest R^2 was for the set with garage space for multiple cars.

Moreover, the assumption of homoskedasticity was checked again for the subsets. Interestingly, the set with no garage met the assumption. This supports the idea that the model is better at predicting prices for cheaper houses as houses with no garages are much less expensive.

As seen in the table below,

Table 6:

	Dependent variable:		
	SalePrice		
	(1)	(2)	(3)
TotalSF	53.871***	61.193***	85.733***
	(5.073)	(3.684)	(3.279)
YearBuilt	387.804***	488.421***	938.262***
	(80.291)	(63.406)	(59.422)
GarageArea		−3.876	124.661***
		(13.843)	(11.078)
Constant	−712,108.500***	−896,691.500***	−1,865,805.000***
	(157,282.500)	(124,573.400)	(116,840.400)
Observations	81	369	1,010
R^2	0.606	0.434	0.622
Adjusted R^2	0.596	0.429	0.620
Residual Std. Error	20,852.430 (df = 78)	22,973.140 (df = 365)	49,853.290 (df = 1006)
F Statistic	60.059*** (df = 2; 78)	93.308*** (df = 3; 365)	550.811*** (df = 3; 1006)
Note:			*p<0.1; **p<0.05; ***p<0.01

5.2 Zoning Classification

For the next subset analysis, the zoning classification of houses were checked. As shown in Figure 1, Floating Village Residential and Residential Low Density houses were more expensive than houses in high and low density residential areas. The Commercial zone was neglected as there were only 10 houses classified to be in the zone.

The table below revealed very interesting details about the data. Even though the average price was the same of medium and high density areas, the model had a very high R^2 value for high density residential areas, it is even higher than the finalized complete model. Whereas the medium density subset had the lowest R^2 value, indicating that the model is worse at explaining the variance in the subset.

Moreover, it is interesting to see that for Floating Village Residential and high density residential areas, the garage area variable was not significant in the model. Hence, it can be stated that for analyzing in those specific subsets, the variable could be removed. Overall, it is reassuring to see that the model holds up in multiple zones.

Table 7:

	Dependent variable:			
	SalePrice			
	(FV)	(RL)	(RM)	(RH)
TotalSF	122.506***	83.452***	58.258***	56.599***
	(13.618)	(2.988)	(5.659)	(9.285)
YearBuilt	4,330.019***	794.512***	375.345***	714.112**
	(1,183.855)	(61.456)	(76.456)	(244.097)
GarageArea	10.380	85.829***	73.983***	−34.448
	(31.123)	(8.448)	(12.548)	(26.144)
Constant	−8,664,070.000***	−1,549,929.000***	−705,316.400***	−1,330,357.000**
	(2,370,850.000)	(119,382.400)	(150,890.500)	(476,070.300)
Observations	65	1,151	218	16
R^2	0.684	0.671	0.509	0.757
Adjusted R^2	0.668	0.670	0.502	0.696
Residual Std. Error	30,164.460 (df = 61)	46,382.210 (df = 1147)	34,232.720 (df = 214)	19,682.370 (df = 12)
F Statistic	43.969*** (df = 3; 61)	780.010*** (df = 3; 1147)	73.987*** (df = 3; 214)	12.462*** (df = 3; 12)
Note:				*p<0.1; **p<0.05; ***p<0.01

References

Bowen, W. M., Mikelbank, B. A., & Prestegaard, D. M. (2001). Theoretical and Empirical Considerations Regarding Space in Hedonic Housing Price Model Applications [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0017-4815.00171]. *Growth and Change*, *32*(4), 466–490. https://doi.org/10.1111/0017-4815.00171

Hubmer, J., Krusell, P., & Smith Jr, A. A. (2021). Sources of US wealth inequality: Past, present, and future [Publisher: The University of Chicago Press Chicago, IL]. *NBER Macroeconomics Annual*, *35*(1), 391–455.