Optimization of the Housing Market: Factors Affecting House Sale Prices

by Darling Judah Hsu

I. Abstract:

What factors affect the market value of your house? One of the hottest issues with today's political climate is housing and homelessness. In particular, what can we do to make sure supply meets demand? In order to do so, we need to optimize the housing market not only to optimize profit but to also have more freedom in policymaking, because price is optimized. The data used will be the Ames Housing dataset which includes house prices in Ames Iowa, as well as a multitude of characteristics of the house. Here, we find that fundamental factors of a house like the size and quality seem to have a bigger sway over factors like style. So, this study will intend to build a model determining which factors affect the sales prices of homes most through stepwise processes that include a variety of factors relating to house characteristics and architectural styles.

II. Introduction:

For a flourishing nation like the US, which has continued to grow steadily since creation, housing will be a problem. Immigration tends to follow the business cycle and despite the current inflation issue, the US is still one of the richest countries in the world (*Manuel, 2023*). So naturally, our population has continued to grow, despite declining birth rates and the recent COVID pandemic. To combat increasing demand and limited supply, we have to consider how to optimize the housing market to raise costs. Otherwise, considering low income families, a form of public housing similar to Singapore's Lee Kuan Yew's LAA in the 1960s (albeit less forceful) might be suggested: where the government bought private land below market rates to make room

for clean, decent-quality public housing. If the government could design more-desirable housing, it could encourage people to sell their land at lower costs in return for public housing, while nonetheless keeping private land elsewhere available. Then, it could put a price ceiling on public housing for low-income families. Either way, whether the US decides to raise market cost or take a more centralized stance, market optimization will be needed to make up for loss of limited supply, or a price ceiling.

In particular, this paper will delve into how to optimize the cost of a home and what makes it desirable. When optimizing the cost of a home, a lot of thought goes into the interior design. However, interior design is subjective and malleable; once a house is built, the interior design can be changed. What's harder to change is the build of the house: the style of the roof, the driveway material, the size, etc…

The dataset of interest will be the Ames housing dataset, whose variables range from the configuration of the driveway to the dimensions of the house & floors. Furthermore, because the data is taken from one city, we can generally control for extraneous variables like city conditions or weather (which may affect house preference). With the data, I'll be regressing sale price on HouseStyle, OverallCond (overall condition), OverallQual (overall quality of materials), 1stFlrSF (1st floor square footage), 2ndFlrSF (2nd floor square footage), and Functional (home functionality rating) with control variables like year sold (to filter noise from factors like houses sold between the 2007 Housing Crisis). More variables may be added, based upon a stepwise mode of model building. In particular, I'll likely be using multiple linear regression, as well as logistic regression, to account for categorical features within the dataset.

III. Data

The data used is from the Ames Housing Dataset: a dataset compiled by Dean de Cock, in Ames, Iowa. The dataset monitors characteristics of houses sold in Ames from 2006 to 2010, with 79 variables and 1460 cross sectional units (homes).

Among the variables, the main parameters of interest include SalePrice, LotArea, GarageArea, Foundation, OverallQual, OverallCond, and TotRmsAbvGrd. In addition, other variables were checked beforehand.

The first variable check was price against the area of the first floor in square feet (*figure 3.1*). Ultimately, we decided to use the lot area instead of the first floor but we found that area did indicate a positive correlation with sales price.

Additionally we looked at a histogram of the sales price which was right skewed and centered around $200,000, indicating that the median house price was around $200,000 (*figure 3.2*). This is important because it indicates the wealth level of people buying houses in Ames, which isn't in the extremes of poverty and wealth. Wealth level can impact preferences in a house; nevertheless, since the prices are centered around average house pricing in Nebraska, wealth should not be a concern.

A histogram of the sales prices, stratified by the neighborhoods, also showed that no single neighborhood seemed to be more "wealthy" by house prices than any other, though some appeared slightly more expensive. Thus, the neighborhood was decided not to be a consideration in our model (*figure 3.3*).

A histogram of sales prices stratified by house style also yielded similar results where "house style" included 1-story houses, 2-story houses, etc… Because no significant difference

was noticed between the stratifications, house style will also be excluded from our consideration (*figure 3.4*).

Our last concern is the housing crisis from 2007-2008. The dataset samples data from 2006 all the way until 2010, such that the housing crisis was in the middle. So, we obtained box plots of prices across 2006-2010 and found that, aside from the outliers, there were no significant differences in the quartiles of sales prices at the time (*figure 3.5*)

This point is further exemplified by a stratified histogram of sales prices by year (*figure 3.6*). Here, a difference between years is even less obvious than one between neighborhoods as all the colored sections seem to take on the same shape. Thus, time also won't be a consideration in the model.

Thus, the variables included in the model and their summary statistics will be further specified in the next section.

IV. Empirical Specification

Final Regression Equation:

$$\text{SalePrice} = 156285.6 + 62.06\beta_1 - 0.88\beta_2 + 781.76\beta_3 + 30786.26\beta_4 + 3305.44\beta_5 + 56646.37\beta_6 - 251573.80\beta_7 - 30826.39\beta_8 - 15574.13\beta_9 - 22547.62\beta_{10} - 31445.95\beta_{11} - 39870.59\beta_{12}$$ (*table 4.1*.)

In this regression, there are 12 variables: β_1 (Garage Area), β_2 (Lot Area), β_3 (Square Root of Lot Area), β_4 (Overall Quality), β_5 (Overall Condition), β_6 (Total Rooms Above Grade), β_7 (Square Root of Total Rooms Above Grade), β_8 (Brick & Tile), β_9 (Cinder Blocks), β_{10} (Slab), β_{11} (Stone), β_{12} (Wood)

Garage Area refers to the area of the garage in square feet and lot area refers to the area of the property in general (summary statistics listed at *table 3.2*).

The overall quality refers to a rating of the overall material and finish of the house on a scale of 1-10: 2 being poor, 3 being fair, 5 being average, 7 being good, and 9 being excellent (summary statistics listed at *table 3.3*). The overall condition refers to a rating of the overall condition of the house on the same scale of quality, with 2 being poor and 9 being excellent (summary statistics listed at *table 3.4*). Total rooms above grade simply refers to the number of rooms in the house rated as good quality (summary statistics listed at *table 3.5*).

Lastly, there are 5 dummy variables for the foundation of the house, which is built of six materials: brick & tile, cinder block, poured concrete, slab, stone, and wood. Because poured concrete accounted for the majority of the houses in the sample, we considered dummy variables for the other five foundation materials (summary statistics listed at *table 3.1*).

V. Results

In the steps leading up to the current model, a multitude of things needs to be explained

(**1**) The model began from the sales price as a function of the house's lot area (*table 5.1)*. Lot area was chosen because it's the most consistent variable. Although overall quality should relate with sales price most directly, the ratings themselves would not be as consistent as a measurement. So, lot area was chosen to be observed as a factor contributing to sales price first.

With a p-value of practically 0, lot area is statistically significant; so, it's the first factor attributed to the model. The obtained coefficient was 2.1, meaning that for every unit increase in square footage, the sales price is estimated to increase by $2.10. However, the adjusted

R-squared of the model was only 0.069, meaning that only 6.90% of the variation in the sales prices could be explained by the model

(**2**) Next, three more OLS regressors were added to the model: overall quality, overall condition, and total rooms above grade (*table 5.2)*. Overall quality and overall condition were added to the model due to the fact that, assuming the ratings are accurate representations of quality and condition, quality and condition should relate to sales price most directly. Total rooms above grade was chosen as a regressor since the number of rooms are meant to represent the wealth, covering the possibility of multiple floors.

(**3**)The beta coefficients of overall quality, condition, and total rooms above grade are 39288.90, -134.67, and 10388.18 respectively. These coefficients indicate that for every unit increase in the overall quality score, sales price is expected to increase by approximately $39,300; for every unit increase in overall condition score, sales price is expected to decrease by $134.67; and for every increase in total rooms above grade, sales price is expected to increase by approximately $10,400. Something noticeably odd is that increased overall condition is expected to increase sales price though it isn't in this model; thus, it's possible that there exists multicollinearity. So, this factor will need reconfirmation in future endeavors.

Regardless, with these new variables, the model's adjusted R-squared has jumped to 0.6936, meaning that our model now explains 69.36% of the variance in sales prices.

Additionally, we decided to add in a variable for the foundation material. To do so, we set the standard as poured concrete, creating dummy variables for the remaining five possible foundations (brick & tile, cinder blocks, slabs, stone, and wood). With the new model, we obtained the beta coefficients -36891, -15486, -16699, -30887, and -34128 respectively (*table 5.3*). The first thing to notice is that compared to poured concrete, every other foundation is

expected to decrease sales price. This makes sense considering that poured concrete is used most in this sample.

(**5**) In order to confirm the statistical significance for the dummy variables, a multiple restriction F-test was conducted with the null hypothesis that β_8 (Brick & Tile) = β_9 (Cinder Blocks) = β_{10} (Slab) = β_{11} (Stone) = β_{12} (Wood) = 0 -that the unrestricted model is statistically insignificant- and alternative hypothesis that the unrestricted model accounts for statistically more variance (*test* 5.1). At a significance level of 95%, the F-critical value is 3.0173. An F-statistic of 12.9 was obtained; with an F-statistic greater than our F-critical value, the model is deemed significant despite the individual insignificance of slabs, stone, and wood.

Additionally, the possibility of nonlinear functional forms needed to be considered, for the lot area variable in particular. Because of the concept of diminishing marginal utility, it's very likely that increased lot area is likely to increase in value until a certain point, at which the growth is slowed. So, we considered the possibility of squared lot area having an effect (*table 5.4a*), and the square root of lot area (*table 5.4b*). However, because the squared lot area led to a model with an adjusted r-squared of 0.7193 and the square root of the lot area output an adjusted r-squared of 0.7253, we decided to keep the square root of the lot area. From the square root of the lot area, we obtained a beta coefficient of 950.50, indicating that for every increase in lot area, sales price is expected to decrease by -1.16 but increase by 425.25 times the lot area value.

Other things considered in our model include garage area, the log of the lot area, and the square root of the total rooms above grade. The garage area variable yielded a beta coefficient of 62.37, indicating that for each increase in garage area square footage, sales price is expected to increase by $62.37 (*table 5.5*). The log of the lot area yielded a beta coefficient of 24,365.79 meaning that for every percentage point increase of the lot area, sales price is expected to

increase by $24,365.79 (*table 5.6*). However, the resulting adjusted r-squared of 0.7489 was not as large as the adjusted r-squared of our current model, with an R-squared of 0.7504. Lastly, we included the square root of the total rooms above grade in our current mode, which yielded a beta coefficient of -251,573, indicating that for every increase in rooms above grade, sales price is expected to decrease by $251,573 (*table 4.1*). This is another odd finding that should be researched further in the future.

(**4**) Furthermore, in order to test the verifiability of a nonlinear functional form for the total rooms variable, we conducted a single restriction F-test (*test 5.2*) where our null hypothesis is that the unrestricted model, with the square root of rooms, is statistically insignificant ($\beta_6 = 0$)- and the alternative hypothesis that $\beta_6 \neq 0$. In particular, we obtained an F-statistic of 48.51. At a significance level of 95%, the F-critical value is 6.6349. Because we obtained an F-statistic greater than the F-critical value, we conclude that the square root of total rooms is statistically significant.

(**6**) Another important factor to consider is omitted variable bias. Because we didn't collect the data ourselves, there's potential for variables left out that we were unable to check. Many things were controlled for because the houses were all sampled from the same city but one such possibly omitted variable is architectural style of the house. In specific, different styles of house are popular with trends. So, in one season, multi-story houses with lots of rooms might be more preferable. In another season, more modest homes with less rooms but bigger might be more preferred. Even practically, more modest homes are likely to be preferred during economic recessions like during the COVID-19 pandemic. Thus, this would be a possible omitted variable.

In particular, style would likely be correlated with foundation, as well as total number of rooms above grade. Assuming style would always be positively adding to sale price, iff certain

foundations were “in style” or style was positively correlated with more rooms and sales price, then the bias would be positive; otherwise, the bias would be negative.

One way to solve this is to do a web scrape of yelp reviews or zillow prices for houses of different styles across different time periods. By doing so, you could establish a preferred style or trend for the different time periods, or even locations, of each house.

(**7**) Lastly, one model to consider is the probit model. Probit models are far superior to OLS when it comes to binary outcomes primarily for their scaling to output probabilities between 0 and 1. So, in order to further look at the significance of variables, we’re interested in a probit prediction model.

Because we’re trying to predict a continuous variable, sales price, we would need to consider binary outcomes. Unfortunately, due to the specification of the data, we’re unable to try and predict the probability of a house being bought since all houses included have been bought. Instead, we decided to try and measure preferability or “how preferable” a house is, measured by the probability of a house being bought within the first 5 months.

The reason why we decided to use this as an indicator is because generally, (**fact check)**, houses are bought, on average, within the first 3-4 months of being on-sale. So, we looked at a variable indicating how many months it took for a house to be sold, ranging from 1 month to a year (*table 5.8*).

Unfortunately, with our previous model for predicting sales price, our model became completely invalid: with a pseudo R-squared of 0.0055 and every beta coefficient statistically insignificant (*table 5.9*). In fact, it’s likely our measurement of preferability was flawed since

continued failure ensued with many more combinations of variables: each returning statistically insignificant models.

The best probit model we obtained was preference as a function of overall condition: p(preference) = -0.32 - 0.067β_1, which indicates that for every unit increase in overall condition rating, the z-score is expected to decrease by 0.067 units, or 2.66 percentage points. However, even this model has a pseudo r-squared value of 0.0025, and the constant is statistically insignificant because of its p-value, 0.083 which is above an alpha significance level of 95%.

VI. Conclusion

In conclusion, according to the model we've found, the factors that seem to affect housing most are lot area, overall quality, overall condition, and style- which we're attributing to foundation and total rooms. Through a variety of processes and tests, we've determined that styles like total rooms above grade and floor foundation seem to be far more inconsistent in determining sales price. Among the variables that needed the most adjusting and changed most with other variables were those relating to style. Thus, it seems style cannot remedy fundamental differences in houses.

Instead, characteristics like overall quality, condition, and lot area were far more consistent, attributing to the model the most. Thus, we conclude that the most important factors in the sales prices of houses are those fundamental characteristics, more than style.

VII. References

Aalbers, Manuel. *The Great Moderation, the Great Excess and the Global Housing Crisis*, www.tandfonline.com/doi/full/10.1080/14616718.2014.997431. Accessed 25 Nov. 2023.

De Cock, Dean. *Ames, Iowa: Alternative to the Boston Housing Data as an End Of …*, www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627. Accessed 25 Nov. 2023.

Housing and the Economy , OECD, www.oecd.org/housing/topics/housing-economy/. Accessed 2 Nov. 2023.

Mohamad, Muhamad, et al. "Procedia: Social & Behavioral Sciences." *ISSN 1877-0428 (Online) | Procedia: Social & Behavioral Sciences | The ISSN Portal*, portal.issn.org/resource/issn/1877-0428. Accessed 24 Nov. 2023.

Office, U.S. Government Accountability. "The Affordable Housing Crisis Grows While Efforts to Increase Supply Fall Short." *U.S. GAO*, US GAO, 7 July 2020, www.gao.gov/blog/affordable-housing-crisis-grows-while-efforts-increase-supply-fall-sort.

VIII. Appendix

Table 3.1

```
. tab foundation, gen(found_)
```

Foundation	Freq.	Percent	Cum.
BrkTil	146	10.00	10.00
CBlock	634	43.42	53.42
PConc	647	44.32	97.74
Slab	24	1.64	99.38
Stone	6	0.41	99.79
Wood	3	0.21	100.00
Total	1,460	100.00	

Table 3.2

Variable	Obs	Mean	Std. dev.	Min	Max
saleprice	1,460	180921.2	79442.5	34900	755000
lotarea	1,460	10516.83	9981.265	1300	215245
garagearea	1,460	472.9801	213.8048	0	1418

Table 3.3

OverallQual	Freq.	Percent	Cum.
1	2	0.14	0.14
2	3	0.21	0.34
3	20	1.37	1.71
4	116	7.95	9.66
5	397	27.19	36.85
6	374	25.62	62.47
7	319	21.85	84.32
8	168	11.51	95.82
9	43	2.95	98.77
10	18	1.23	100.00
Total	1,460	100.00	

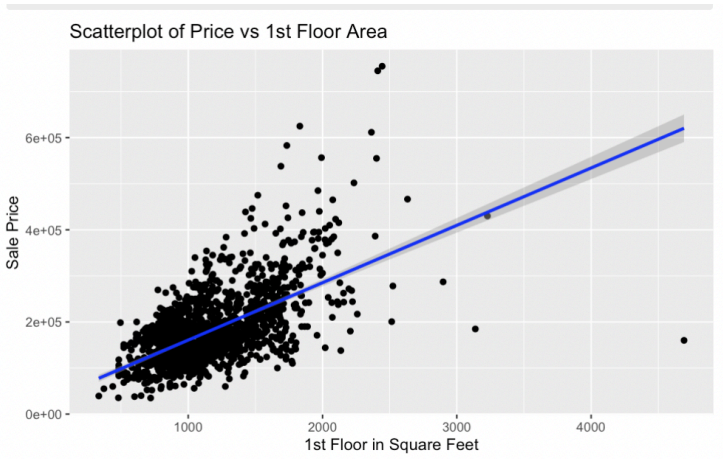
Table 3.4

OverallCond	Freq.	Percent	Cum.
1	1	0.07	0.07
2	5	0.34	0.41
3	25	1.71	2.12
4	57	3.90	6.03
5	821	56.23	62.26
6	252	17.26	79.52
7	205	14.04	93.56
8	72	4.93	98.49
9	22	1.51	100.00
Total	1,460	100.00	

Table 3.5

TotRmsAbvGrd	Freq.	Percent	Cum.
2	1	0.07	0.07
3	17	1.16	1.23
4	97	6.64	7.88
5	275	18.84	26.71
6	402	27.53	54.25
7	329	22.53	76.78
8	187	12.81	89.59
9	75	5.14	94.73
10	47	3.22	97.95
11	18	1.23	99.18
12	11	0.75	99.93
14	1	0.07	100.00
Total	1,460	100.00	

Figure 3.1

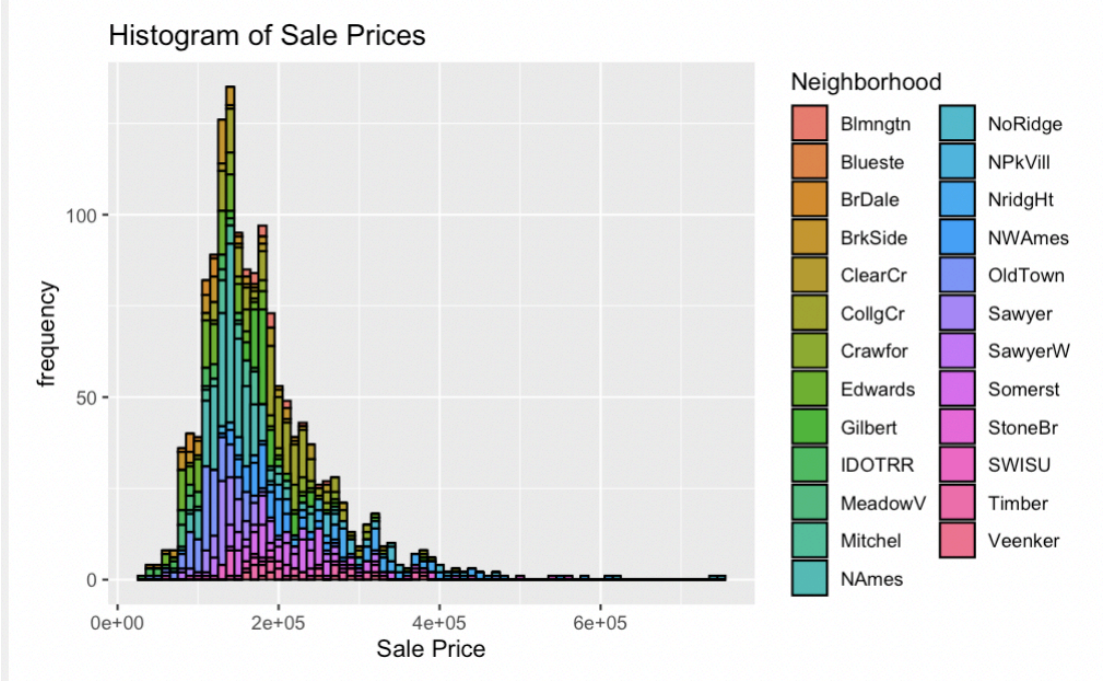


Scatterplot between sale price and the size of the 1st floor in square feet. Here, we see that there's a clear positive correlation though it's important to note, much of the data lies between 1st floor sizes of 0-2000 square feet.

Figure 3.2

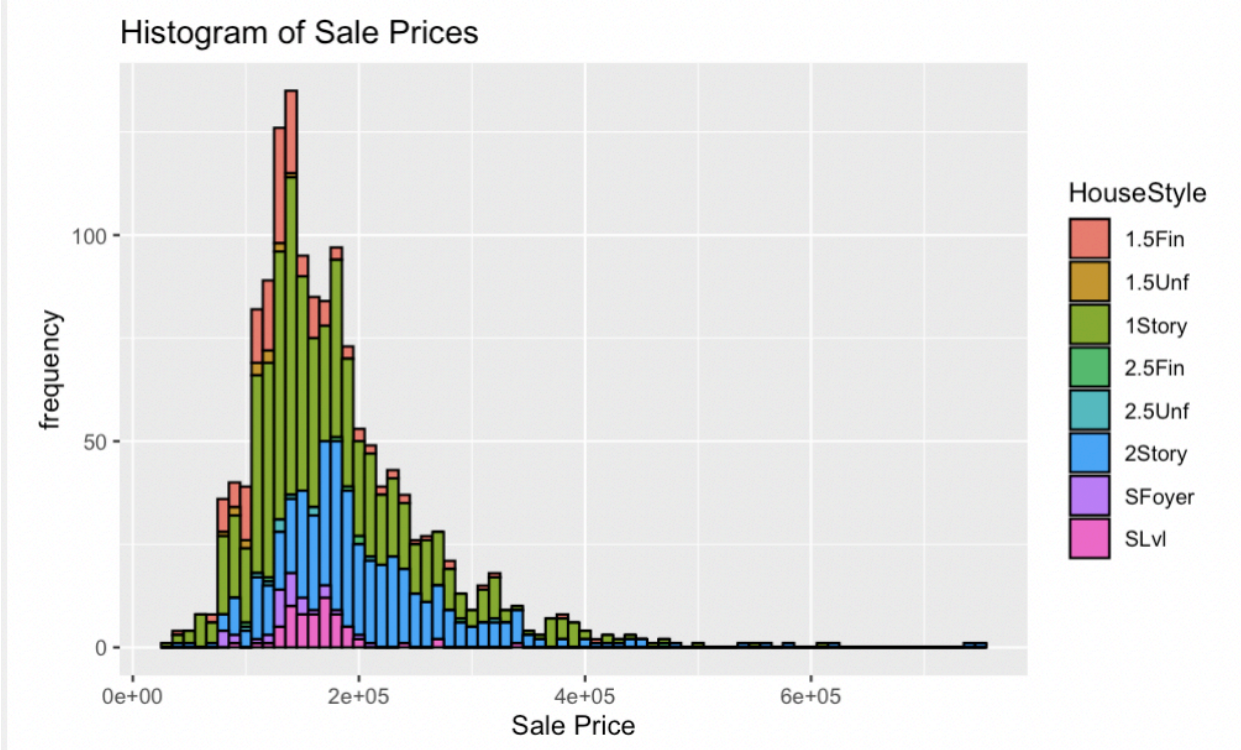
Histogram of sale prices. The histogram is right skewed with most sale prices being around $200,000. With huge values, it might be smart to log the sale price values to make theme easier to interpret. It's approximately normal.

Figure 3.3



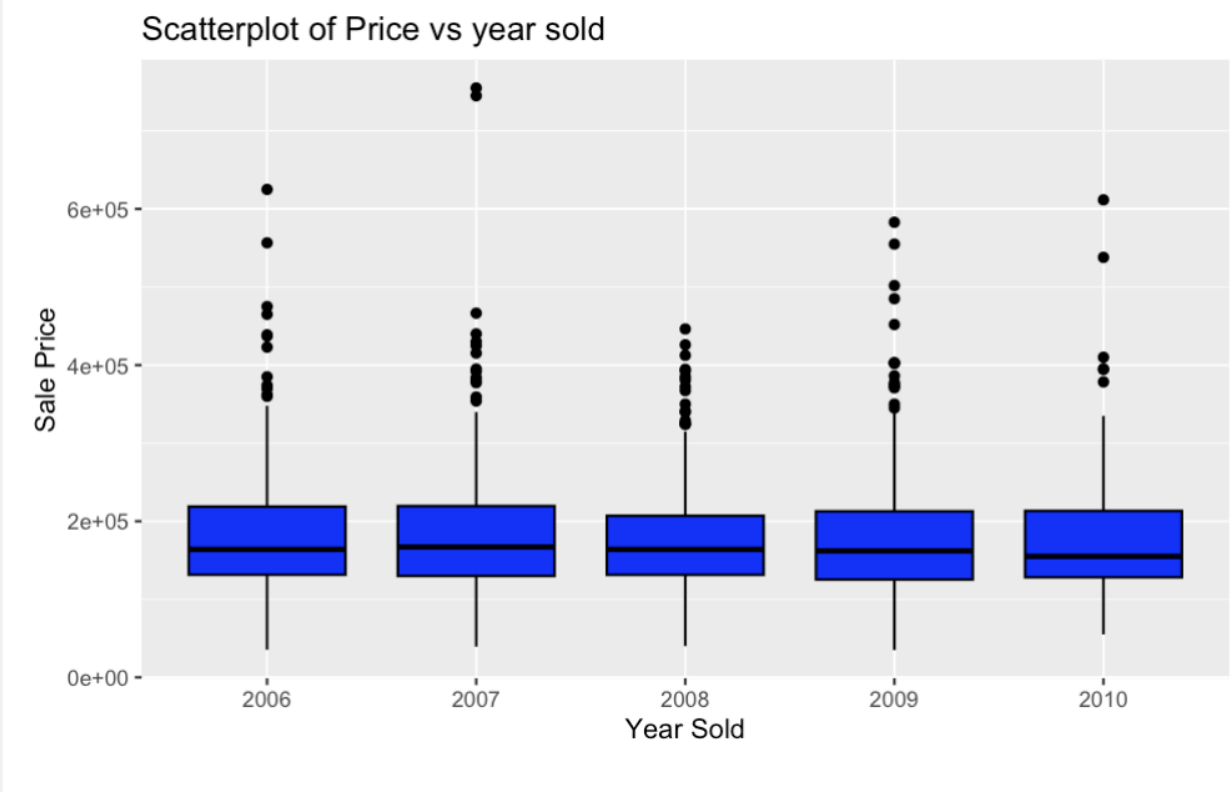
Histogram of sale prices, stratified by neighborhoods. Luckily, neighborhoods don't seem to have a huge effect on the sale prices though further analysis may be needed as Veenker does seem a bit more expensive than NAmes on average. Overall though, bias is minimal.

Figure 3.4



Histogram of sale prices, stratified by the style. Here, there doesn't seem to be a huge preference among house styles with the only noticeable difference being between 2-story houses versus others. Thus, dummy variables may need to be taken into account.

Figure 3.5



Though initially concerned about the year sold, the scatterplots of sale prices indicate that overall, there shouldn't be too much difference. Extra analysis will be needed to account for the 2007-2008 housing crisis but initial visualization appears positive.

Figure 3.6

Histogram further showing distribution of sale prices. Sale prices noticeably even across years; there doesn't seem to be any skew among the years.

Table 4.1

Source	SS	df	MS	Number of obs	=	1,460
				F(12, 1447)	=	366.44
Model	6.9281e+12	12	5.7734e+11	Prob > F	=	0.0000
Residual	2.2798e+12	1,447	1.5755e+09	R-squared	=	0.7524
				Adj R-squared	=	0.7504
Total	9.2079e+12	1,459	6.3111e+09	Root MSE	=	39693

saleprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
garagearea	62.06274	6.259427	9.92	0.000	49.78422	74.34126
lotarea	-.8828441	.2611472	-3.38	0.001	-1.395112	-.3705764
lotarea_sqrt	781.7644	95.50439	8.19	0.000	594.4226	969.1063
overallqual	30786.26	1097.355	28.05	0.000	28633.68	32938.84
overallcond	3305.439	1003.041	3.30	0.001	1337.868	5273.009
totrmsabvgrd	56646.37	6962.32	8.14	0.000	42989.05	70303.69
totrms_sqrt	-251573.8	36120.78	-6.96	0.000	-322428.5	-180719.1
found_1	-30826.39	4128.126	-7.47	0.000	-38924.14	-22728.64
found_2	-15574.13	2792.633	-5.58	0.000	-21052.17	-10096.08
found_4	-22547.62	8689.747	-2.59	0.010	-39593.47	-5501.771
found_5	-31445.95	16514.59	-1.90	0.057	-63841.05	949.1415
found_6	-39870.59	22991.21	-1.73	0.083	-84970.26	5229.085
_cons	156285.6	46764.13	3.34	0.001	64552.83	248018.3

Current linear regression model

Table 5.1

```
Residuals:
    Min      1Q  Median      3Q     Max
-275668  -48169  -17725   31248  553356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.588e+05  2.915e+03   54.49   <2e-16 ***
LotArea     2.100e+00  2.011e-01   10.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76650 on 1458 degrees of freedom
Multiple R-squared:  0.06961,   Adjusted R-squared:  0.06898
F-statistic: 109.1 on 1 and 1458 DF,  p-value: < 2.2e-16
```

Model 1: House price as a function of lot area

Model1: SalePrice = 158,800 + 2.1LotArea

Because the LotArea variable has a p-value of essentially 0, evidence suggests that it's a statistically significant variable in regards to predicting house SalePrice.

Table 5.2

Source	SS	df	MS	Number of obs	=	1,460
				F(4, 1455)	=	826.75
Model	6.3945e+12	4	1.5986e+12	Prob > F	=	0.0000
Residual	2.8134e+12	1,455	1.9336e+09	R-squared	=	0.6945
				Adj R-squared	=	0.6936
Total	9.2079e+12	1,459	6.3111e+09	Root MSE	=	43973

saleprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lotarea	1.202458	.1175269	10.23	0.000	.9719176	1.432998
overallqual	39288.9	923.7084	42.53	0.000	37476.96	41100.84
overallcond	-134.6709	1039.17	-0.13	0.897	-2173.103	1903.761
totrmsabvgrd	10388.18	793.9972	13.08	0.000	8830.679	11945.68
_cons	-138317.6	8623.931	-16.04	0.000	-155234.2	-121400.9

Linear regression model including overall quality, overall condition, and total rooms above grade included on top of lot area

Table 5.3

saleprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lotarea	1.238244	.1159195	10.68	0.000	1.010857	1.465632
overallqual	35188.84	1097.316	32.07	0.000	33036.34	37341.34
overallcond	2667.998	1082.809	2.46	0.014	543.9587	4792.037
totrmsabvgrd	11112.57	785.9802	14.14	0.000	9570.788	12654.35
found_1	-36891.27	4420.909	-8.34	0.000	-45563.33	-28219.21
found_2	-15486.52	3000.021	-5.16	0.000	-21371.37	-9601.679
found_4	-16699.73	9374.364	-1.78	0.075	-35088.5	1689.03
found_5	-30887.3	17874.31	-1.73	0.084	-65949.57	4174.965
found_6	-34128.91	24882.22	-1.37	0.170	-82937.9	14680.08
_cons	-123147.9	9122.759	-13.50	0.000	-141043.2	-105252.7

Linear regression model including dummy variables for foundation

Test 5.1

```
. test(found_1==0)(found_2==0)(found_4==0)(found_5==0)(found_6==0)

 ( 1)  found_1 = 0
 ( 2)  found_2 = 0
 ( 3)  found_4 = 0
 ( 4)  found_5 = 0
 ( 5)  found_6 = 0

       F(  5,  1447) =   12.90
            Prob > F =    0.0000
```

Multiple Restriction F-test for foundation variables

Table 5.4a

```
. gen lotarea_sq = lotarea^2

. regress saleprice lotarea lotarea_sq overallqual overallcond totrmsabvgrd  fou
> nd_1 found_2 found_4 found_5 found_6
```

Source	SS	df	MS	Number of obs	=	1,460
				F(10, 1449)	=	374.82
Model	6.6407e+12	10	6.6407e+11	Prob > F	=	0.0000
Residual	2.5672e+12	1,449	1.7717e+09	R-squared	=	0.7212
				Adj R-squared	=	0.7193
Total	9.2079e+12	1,459	6.3111e+09	Root MSE	=	42091

saleprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lotarea	3.039224	.254018	11.96	0.000	2.540942	3.537506
lotarea_sq	-.0000122	1.54e-06	-7.93	0.000	-.0000152	-9.19e-06
overallqual	34623.39	1077.014	32.15	0.000	32510.72	36736.07
overallcond	2741.707	1060.482	2.59	0.010	661.4627	4821.951
totrmsabvgrd	9714.466	789.6994	12.30	0.000	8165.389	11263.54
found_1	-36840.74	4329.591	-8.51	0.000	-45333.68	-28347.8
found_2	-17237.96	2946.349	-5.85	0.000	-23017.52	-11458.39
found_4	-17334.28	9181.067	-1.89	0.059	-35343.88	675.3256
found_5	-28538.49	17507.59	-1.63	0.103	-62881.42	5804.442
found_6	-38076.54	24373.32	-1.56	0.118	-85887.3	9734.215
_cons	-126606.4	8944.961	-14.15	0.000	-144152.8	-109059.9

Linear regression model with lot area squared as a factor

Table 5.4B

Source	SS	df	MS	Number of obs	=	1,460
				F(10, 1449)	=	386.15
Model	6.6955e+12	10	6.6955e+11	Prob > F	=	0.0000
Residual	2.5124e+12	1,449	1.7339e+09	R-squared	=	0.7271
				Adj R-squared	=	0.7253
Total	9.2079e+12	1,459	6.3111e+09	Root MSE	=	41640

saleprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lotarea	-1.162657	.26984	-4.31	0.000	-1.691976	-.6333386
lotarea_sqrt	950.4779	97.1332	9.79	0.000	759.9412	1141.015
overallqual	34580.19	1064.948	32.47	0.000	32491.19	36669.2
overallcond	2693.606	1049.078	2.57	0.010	635.7325	4751.48
totrmsabvgrd	8803.098	797.2296	11.04	0.000	7239.25	10366.95
found_1	-36829.83	4283.183	-8.60	0.000	-45231.73	-28427.93
found_2	-17799.47	2916.153	-6.10	0.000	-23519.81	-12079.14
found_4	-18079.46	9083.406	-1.99	0.047	-35897.49	-261.429
found_5	-28381.7	17319.34	-1.64	0.101	-62355.36	5591.958
found_6	-41459.06	24118.66	-1.72	0.086	-88770.29	5852.169
_cons	-171818.1	10141.92	-16.94	0.000	-191712.5	-151923.7

Linear Regression model with the square root of lot area

Table 5.5

Source	SS	df	MS	Number of obs	=	1,460
				F(9, 1450)	=	465.97
Model	6.8422e+12	9	7.6025e+11	Prob > F	=	0.0000
Residual	2.3657e+12	1,450	1.6315e+09	R-squared	=	0.7431
				Adj R-squared	=	0.7415
Total	9.2079e+12	1,459	6.3111e+09	Root MSE	=	40392

saleprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
garagearea	62.36669	6.369442	9.79	0.000	49.87238	74.86099
lotarea	-.7470811	.2649819	-2.82	0.005	-1.26687	-.2272922
lotarea_sqrt	726.8378	96.82847	7.51	0.000	536.8989	916.7766
overallqual	30683.52	1109.513	27.65	0.000	28507.1	32859.94
overallcond	3201.366	1013.082	3.16	0.002	1214.103	5188.628
totrmsabvgrd	8324.224	770.7724	10.80	0.000	6812.276	9836.172
found_1	-29773.23	4176.201	-7.13	0.000	-37965.27	-21581.18
found_2	-14334.78	2815.359	-5.09	0.000	-19857.4	-8812.174
found_4	-16556.81	8801.949	-1.88	0.060	-33822.73	709.1006
_cons	-162050.8	9885.153	-16.39	0.000	-181441.5	-142660

Linear regression model with garage area added as a factor

Table 5.6

Source	SS	df	MS		
Model	6.9120e+12	10	6.9120e+11	Number of obs	= 1,460
Residual	2.2959e+12	1,449	1.5845e+09	F(10, 1449)	= 436.23
				Prob > F	= 0.0000
Total	9.2079e+12	1,459	6.3111e+09	R-squared	= 0.7507
				Adj R-squared	= 0.7489
				Root MSE	= 39806

saleprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
garagearea	62.79228	6.266787	10.02	0.000	50.49933	75.08522
lotarea	.273914	.1472814	1.86	0.063	-.0149935	.5628216
lotarea_log	24365.79	3080.102	7.91	0.000	18323.85	30407.73
overallqual	31163.38	1095.393	28.45	0.000	29014.65	33312.1
overallcond	2974.8	998.823	2.98	0.003	1015.506	4934.094
totrmsabvgrd	58202.04	6995.195	8.32	0.000	44480.25	71923.84
totrms_sqrt	-260023.9	36342.58	-7.15	0.000	-331313.5	-188734.2
found_1	-29823	4115.958	-7.25	0.000	-37896.88	-21749.13
found_2	-14190.55	2770.131	-5.12	0.000	-19624.45	-8756.658
found_4	-21923.51	8705.757	-2.52	0.012	-39000.74	-4846.275
_cons	8631.475	50841.95	0.17	0.865	-91100.21	108363.2

Linear regression model containing the log of the lot area

Test 5.2

```
. test totrms_sqrt==0

 ( 1)  totrms_sqrt = 0

       F(  1,  1447) =   48.51
            Prob > F =    0.0000
```

Single restriction F-test to determine whether the square root of total rooms is significant

Table 5.7

Probit regression — Number of obs = 1,460
LR chi2(1) = 4.10
Prob > chi2 = 0.0430
Log likelihood = -810.05447 — Pseudo R2 = 0.0025

Preference2	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
overallcond	-.066557	.0330608	-2.01	0.044	-.131355	-.001759
_cons	-.32271	.1863292	-1.73	0.083	-.6879086	.0424885

Best probit regression model found with only the overall condition variable

Table 5.8

Probit regression — Number of obs = 1,460
LR chi2(13) = 10.77
Prob > chi2 = 0.6305
Log likelihood = -967.1305 — Pseudo R2 = 0.0055

Preference	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
saleprice	-2.81e-07	8.44e-07	-0.33	0.739	-1.93e-06	1.37e-06
garagearea	-7.83e-06	.0002073	-0.04	0.970	-.0004142	.0003985
lotarea	-.0000134	.0000123	-1.08	0.278	-.0000376	.0000108
lotarea_sqrt	.0028436	.00371	0.77	0.443	-.0044279	.0101151
overallqual	-.0261044	.0438103	-0.60	0.551	-.1119709	.0597621
overallcond	-.0311647	.0324358	-0.96	0.337	-.0947378	.0324083
totrmsabvgrd	.167308	.2284385	0.73	0.464	-.2804233	.6150393
totrms_sqrt	-.9663202	1.177613	-0.82	0.412	-3.274399	1.341759
found_1	-.0371678	.1356205	-0.27	0.784	-.302979	.2286434
found_2	.0058697	.0907777	0.06	0.948	-.1720513	.1837907
found_4	-.1459368	.2812084	-0.52	0.604	-.6970952	.4052216
found_5	.7539872	.5326371	1.42	0.157	-.2899624	1.797937
found_6	-.1112112	.7488177	-0.15	0.882	-1.578867	1.356445
_cons	1.309906	1.499268	0.87	0.382	-1.628606	4.248417

Regression table for probit regression of previous model