An Analysis of Home Sales Prices' Relation to Residence Characteristics

Creating a Hedonic Model

Katie Haring

May 6, 2019

Abstract

The selling price for a residential home is based on many factors. Researchers have long tried to wrangle with the importance of each factor using the hedonic pricing model. This analysis sought to use that model to determine which of 11 factors appeared to be relevant to the sales price of 522 homes in a mid-western city in 2002. 4 factors were determined to add useful insight without over-fit, which were: Number of Bedrooms, Year Built, Quality Index, and Lot Size. The dependent variable was transformed using the natural log function to help normalize residuals. The impact of an unusual home with no bedrooms or bathrooms was analyzed and determined to be minimal. Readers are warned against applying this model to locations outside the original city, as hedonic models will vary by location.

Introduction

The hedonic pricing model is a model used to estimate the impact that various internal and external factors have on the price of a home. The model can be used to determine the intrinsic value of each factor or to predict future selling prices. According to Sirmans, MacDonald, Macpherson, and Zietz (2005),the most common factors included in a hedonic pricing model for a single-family home are: square feet, lot size, age, number of bedrooms, number of bathrooms, number of garage spaces, swimming pool indicator, number of fireplaces, and air conditioning indicator. When Sirmans et. al. conducted a meta analysis of over 80 hedonic pricing studies, they found that the impact of many of these factors varied by location. For that reason, it is important that each city or regional location create its own model if it wishes to accurately predict sales prices.

The goal of this paper is to create a hedonic model for a mid-western city, using the methods set forth in Monson's 2009 paper, "Valuation using hedonic pricing models". The factors that will be considered for the model are: Finished Square Feet, Number of Bedrooms, Number of Bathrooms, Air Conditioning Indicator, Garage Size, Pool Indicator, Year Built, Quality Index, Style Indicator, Lot Size, and Highway Adjacency Indicator. The data contains this information from the sales of 522 homes sold in 2002.

Primary Analysis Objectives

1. Determine the best hedonic model for sales price using the 11 property characteristics.

Secondary Analysis Objectives

2. Summarize the homes' property characteristics.

Analytic Methods

Data Source

The data for this analysis were obtained from Appendix C, Data Set 7 from Kutner, Nachtsheim, and Neter (2014). The complete data set is available here. The data set contains: Sales Price (dollars), Finished Area of Residence (square feet), Number of Bedrooms, Number of Bathrooms, Air Conditioning Indicator, Garage Size (cars), Pool Indicator, Year Built, Quality Index (ranking 1-3), Style, Lot Size (square feet), and Highway Adjacency Indicator. The data were collected for a mid-western city by the city tax assessor, looking at the sales of 522 homes during the year 2002. It contains no missing values or obvious mis-keys (e.g., negative Lot Size, Year Built after 2002, etc.). However, there is one data entry that contains 0 bedrooms, 0 bathrooms, a 3 car garage, and air conditioning. This suggests an air-conditioned garage, instead of a residential home. Model fit is examined with and without this data point in the Statistical Analysis section, below.

Model Assumptions

All inferences were conducted using $\alpha = 0.05$ unless stated otherwise. No adjustments for multiplicity were made as this is an exploratory analysis. The error variables (ϵ) were checked for normality and equal variances using the methods described in chapter 5 of *Design and Analysis of Experiments*.

Most discrete variables were summarized with proportions and frequencies. Continuous variables and the Year Built variable were summarized using the following statistics:

- mean
- median
- standard deviation
- quantiles
- minimum
- maximum

Primary Objective Analysis

Preliminary Analysis

To help visual the data, the correlation matrix was produce for each of the variables (Tables 1 and 2):

Table 1: Correlations (Continued Below)

	ID #	**Price**	**Sqr Ft**	**Bed**	**Bath**	**A/C**	**Garage**
ID #	1.00	-0.56	-0.54	-0.27	-0.52	-0.19	-0.39
Price	-0.56	1.00	0.82	0.41	0.68	0.29	0.58
Sqr Ft	-0.54	0.82	1.00	0.56	0.76	0.27	0.53
Bed	-0.27	0.41	0.56	1.00	0.58	0.23	0.32

Bath	-0.52	0.68	0.76	0.58	1.00	0.32	0.49
A/C	-0.19	0.29	0.27	0.23	0.32	1.00	0.32
Garage	-0.39	0.58	0.53	0.32	0.49	0.32	1.00
Pool	-0.10	0.15	0.16	0.13	0.18	0.10	0.11
Year	-0.39	0.56	0.44	0.27	0.51	0.43	0.46
Quality	0.64	-0.76	-0.70	-0.38	-0.68	-0.41	-0.55
Style	-0.30	0.36	0.62	0.38	0.49	0.13	0.23
Size	-0.16	0.22	0.16	0.13	0.15	-0.11	0.15
Highway	-0.19	-0.05	-0.06	-0.03	-0.05	-0.04	0.00

Table 2: Correlations (Continued)

	Pool	**Year**	**Qual.**	**Style**	**Size**	**Highway**
ID #	-0.10	-0.39	0.64	-0.30	-0.16	-0.19
Price	0.15	0.56	-0.76	0.36	0.22	-0.05
Sqr Ft	0.16	0.44	-0.70	0.62	0.16	-0.06
Bed	0.13	0.27	-0.38	0.38	0.13	-0.03
Bath	0.18	0.51	-0.68	0.49	0.15	-0.05
A/C	0.10	0.43	-0.41	0.13	-0.11	-0.04
Garage	0.11	0.46	-0.55	0.23	0.15	0.00
Pool	1.00	0.06	-0.13	0.08	-0.04	-0.04
Year	0.06	1.00	-0.62	0.23	-0.10	0.03
Quality	-0.13	-0.62	1.00	-0.35	-0.12	0.02
Style	0.08	0.23	-0.35	1.00	-0.01	-0.12
Size	-0.04	-0.10	-0.12	-0.01	1.00	0.08
Highway	-0.04	0.03	0.02	-0.12	0.08	1.00

The matrix shows the highest correlation between Price and Finished Square Feet. Quality and Number of Bathrooms were also highly correlated with Price, as well as with each other. This suggests that a model containing both terms might have issues with multicollinearity.

Box plots were produced for Price and Finished Square Feet (Figure 1):

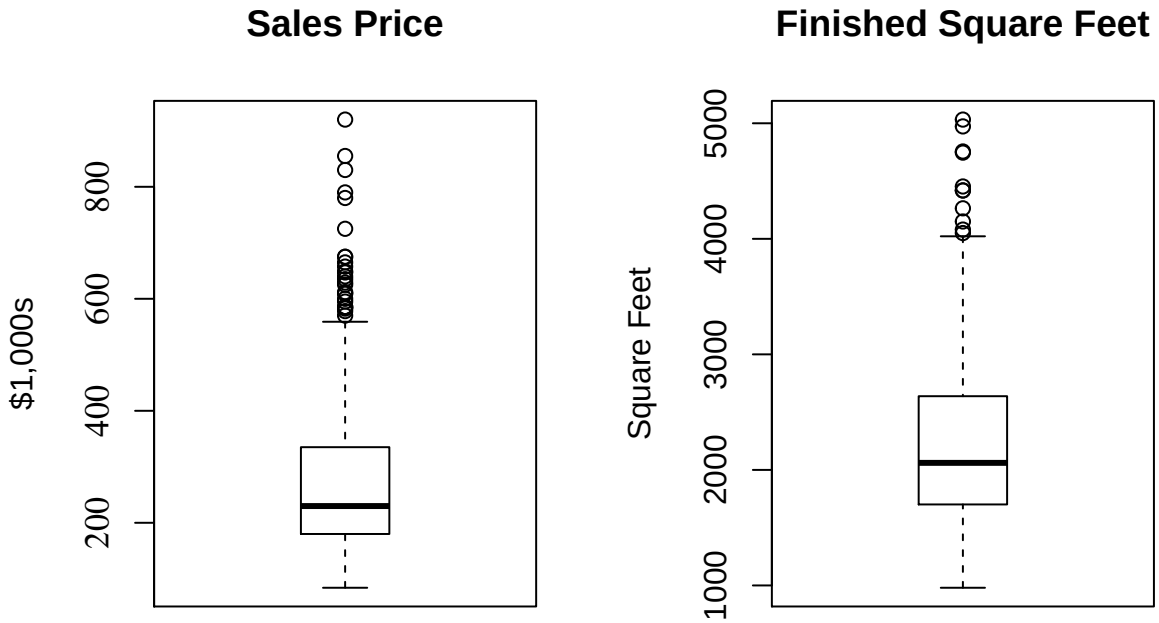

Figure 1: Box Plots of Sales Price and Finished Sqr. Ft.

Both metrics appeared to have somewhat skewed distributions.

Model Fitting

An exhaustive search was performed for the best group of independent variables from the data for predicting the dependent variable. Table 3 shows the adjusted R^2 and Bayesian Information Criterion (BIC) for the 8 models proposed, in order from simplest to most complex, with each table row adding one independent variable to the model.

Table 3: Potential Models

Model	**Adj. R^2**	**BIC**
Finished Sqr Ft	0.6709	-568.63
+ Quality	0.7172	-642.46
+ Style	0.7871	-785.45
+ Year built	0.7994	-811.34
+ Lot Size	0.8155	-849.84
+ Bedroom Ct	0.8207	-859.38
+ Garage Size	0.8222	-858.43
+ Highway Adjacency	0.8236	-857.32

As seen in the table, the most complex model was the most predictive. However, the model with the lowest BIC (an index of model fit, in which the lowest valued model has the best fit) was the model using Finished Square Feet, Quality, Style, Year Built, Lot Size, and Number of Bedrooms. Note that this model eliminates the multicollinearity issue raised in the preliminary analysis. This model was selected for further analysis.

A multiple linear regression model was fit to the data, with the selling price set as the dependent variable and the residential characteristics set as the X independent variables. The model took the form of:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \epsilon_i$$

where:

Y_i = the sales price of the i^{th} randomly selected home
X_{i1} represents Finished Square Feet
X_{i2} represents Quality (1-3)
X_{i3} represents Style
X_{i4} represents Year Built
X_{i5} represents Lot Size in Square Feet
X_{i6} represents Number of Bedrooms
$\epsilon_i \sim idd\text{N}(0,\sigma^2)$
and β_0, β_1, σ^2 were the unknown parameters of interest.

Given that simpler models are generally better than more complex models, this model's R^2 was compared to models that used less parameters to determine if the loss of some predictive power could be accepted for the sake of preventing over-fit. The least impactful predictors determined from Table 4 below were Number of Bedrooms, Style, and Lot Size, in that order. These variables were removed from the model one at a time, and then the impact on model predictive ability was assessed (Table 5).

Table 4: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Finished.square.feet	1	6.655e+12	6.655e+12	2217	1.15e-172
Number.of.bedrooms	1	2.761e+10	2.761e+10	9.2	0.002566
factor(Year.built)	73	1.354e+12	1.854e+10	6.178	8.146e-35
factor(Quality)	2	3.824e+11	1.912e+11	63.7	5.543e-25
factor(Style)	9	1.092e+11	1.213e+10	4.043	5.444e-05
Lot.size	1	7.993e+10	7.993e+10	26.63	3.754e-07
Residuals	434	1.303e+12	3.001e+09	NA	NA

Table 5: Model Characteristics

Predictors	r.squared	adj.r.squared	sigma	statistic	p.value

All	0.869	0.842	54785.04	32.967	0
-Bedrooms	0.868	0.842	54739.85	33.402	0
-Style	0.860	0.836	55802.42	35.569	0
-Lot Size	0.849	0.824	57940.44	32.990	0

As seen in Table 5, none of the factors had a large impact on R^2, and removing Number of Bedrooms from the model barely affected model predictiveness at all. For this analysis, it was decided to remove Number of Bedrooms and Style from the model, reducing the model to the 4 predictive variables: Finished Square Feet, Year Built, Quality, and Lot Size. Lot Size was left in the model, since it had the biggest impact on the R^2. The model with 4 predictive variables explained 86.0% of variation in Sales Price.

Model Testing

The normality of error terms was tested using a QQ-Plot (Figure 2) and a Shapiro-Wilk normality test:

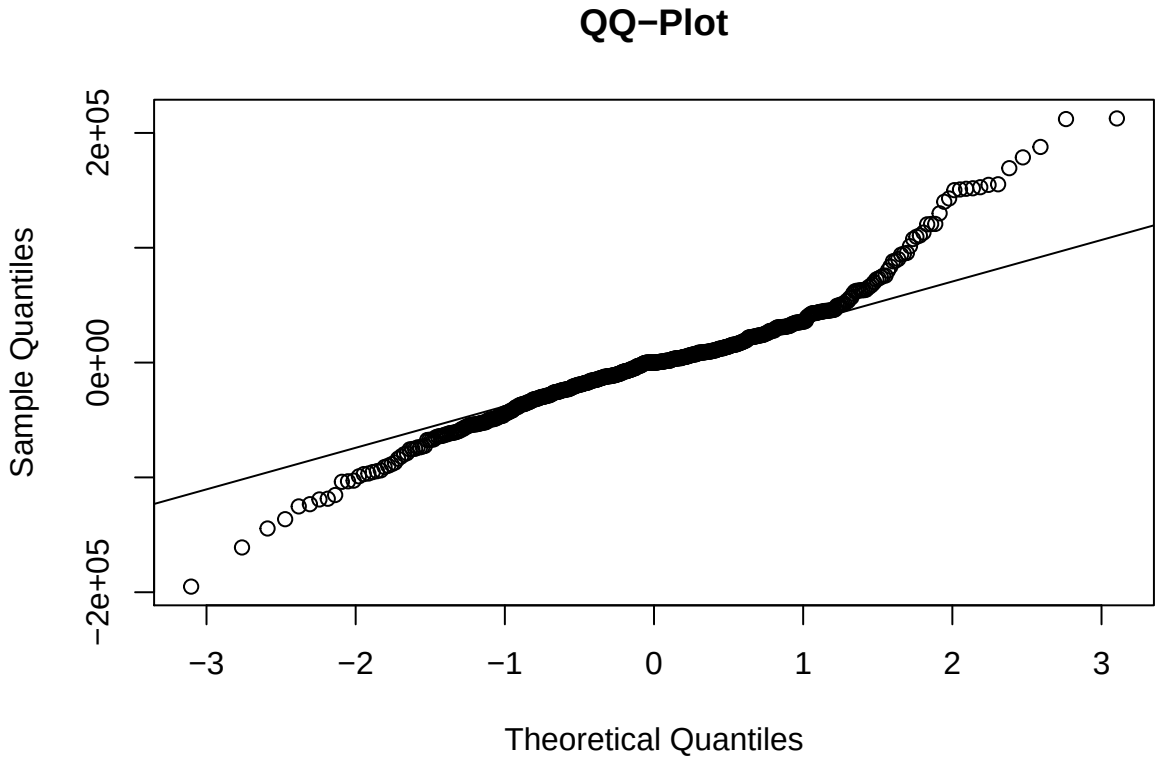

Figure 2: Q-Q Plot

```
Shapiro-Wilk normality test
```

```
data:  residuals(model3)
W = 0.94809, p-value = 1.456e-12
```

As shown in the Shapiro-Wilk results, the error terms failed the normality test at the 0.05 level of significance. To correct the issue, the dependent variable was transformed using the natural log, resulting in the final model:

$$\ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i$$

where:

Y_i = the sales price of the i^{th} randomly selected home
X_{i1} represents Finished Square Feet
X_{i2} represents Quality (1-3)
X_{i3} represents Year Built
X_{i4} represents Lot Size in Square Feet
$\epsilon_i \sim idd\mathrm{N}(0,\sigma^2)$
and β_0, β_1, σ^2 were the unknown parameters of interest.

The new QQ-Plot and normality test results looked like this (Figure 3):

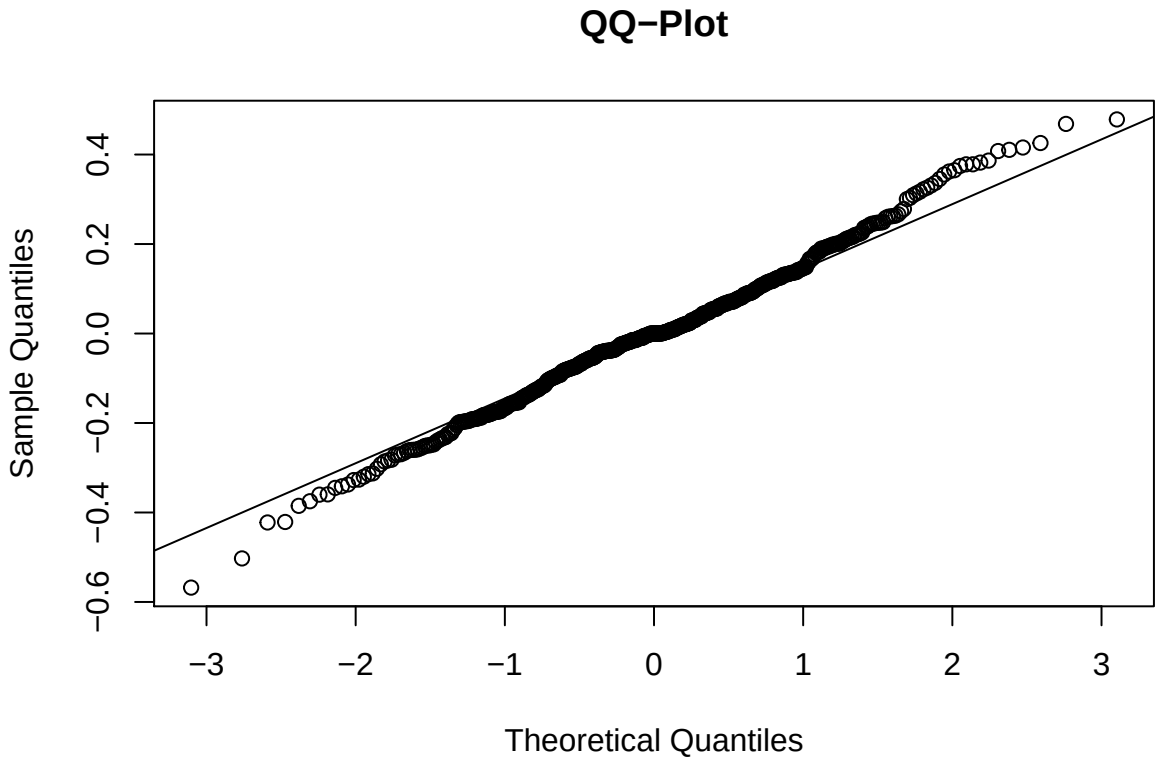

Figure 3: Q-Q Plot

```
    Shapiro-Wilk normality test

data:  residuals(model5)
W = 0.99464, p-value = 0.065
```

Standardized residuals from the new model were plotted against the fitted values produced by it, in Figure 4:

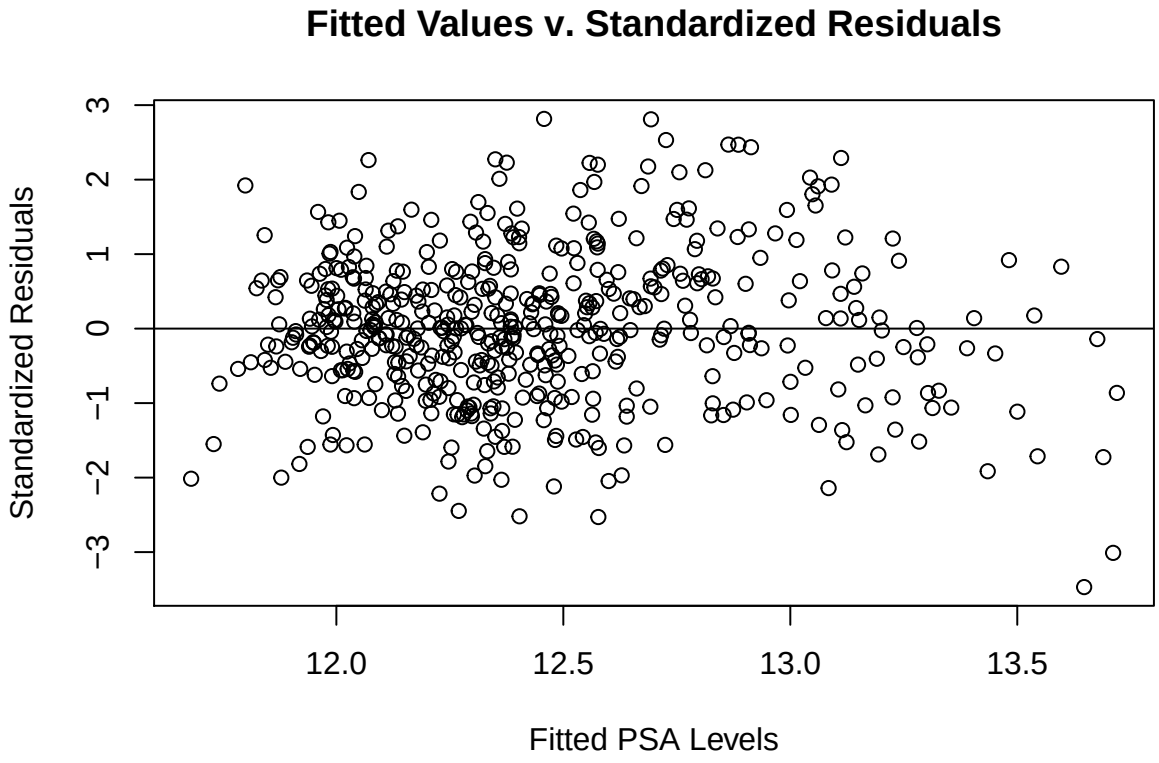

Figure 4: Standardized Residual Plot

The residuals looked approximately random, with one or two outliers. Given the imprecise nature of the model (no housing price is ever going to be predicted exactly right) and that previous review of the data only produced one suspicious data point eligible for potential removal, it was decided to leave the outliers in and instead assess the impact of the one suspicious data point. The entry with a 3-car garage, air conditioning, above average square feet, and no bedrooms or bathrooms was suspected to be either an entry-error or a non-residential property. The model was refit without this entry to assess impact (Table 6).

Table 6: Model Characteristics w/ and w/o Unusual Entry

Predictors	r.squared	adj.r.squared	sigma	statistic	p.value

All Data	0.857	0.833	0.177	34.674	0
-Entry 108	0.857	0.832	0.176	34.514	0

As seen above, the entry was found to have minimal impact on the model and was kept in the data, since the context of the point was unknown.

The final model, $\ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i$, coefficents are summarized in Table 7 below, and the coefficient confidence intervals are given in Table 8:

Table 7: Model Coefficients Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.329	0.195	58.002	0.000
Finished.square.feet	0.000	0.000	18.155	0.000
factor(Year.built)1900	0.353	0.250	1.411	0.159
factor(Year.built)1908	0.366	0.251	1.455	0.146
factor(Year.built)1914	0.551	0.252	2.188	0.029
factor(Year.built)1918	0.262	0.216	1.212	0.226
factor(Year.built)1920	0.413	0.217	1.898	0.058
factor(Year.built)1921	0.441	0.251	1.755	0.080
factor(Year.built)1922	0.345	0.216	1.593	0.112
factor(Year.built)1923	0.401	0.250	1.607	0.109
factor(Year.built)1925	0.567	0.195	2.912	0.004
factor(Year.built)1927	0.331	0.250	1.323	0.187
factor(Year.built)1928	0.046	0.251	0.182	0.856
factor(Year.built)1934	0.344	0.218	1.582	0.114
factor(Year.built)1935	0.287	0.217	1.324	0.186
factor(Year.built)1936	0.351	0.250	1.401	0.162
factor(Year.built)1937	0.850	0.250	3.402	0.001
factor(Year.built)1938	0.362	0.219	1.650	0.100
factor(Year.built)1939	0.483	0.250	1.930	0.054
factor(Year.built)1940	0.384	0.198	1.936	0.054
factor(Year.built)1941	0.397	0.198	2.009	0.045
factor(Year.built)1942	0.539	0.251	2.151	0.032
factor(Year.built)1944	0.385	0.218	1.763	0.079
factor(Year.built)1946	0.448	0.205	2.189	0.029
factor(Year.built)1947	0.431	0.190	2.264	0.024
factor(Year.built)1948	0.352	0.191	1.843	0.066
factor(Year.built)1949	0.408	0.199	2.056	0.040
factor(Year.built)1950	0.440	0.186	2.362	0.019
factor(Year.built)1951	0.339	0.185	1.836	0.067
factor(Year.built)1952	0.596	0.192	3.109	0.002
factor(Year.built)1953	0.436	0.187	2.337	0.020

factor(Year.built)1954	0.517	0.188	2.743	0.006
factor(Year.built)1955	0.408	0.186	2.189	0.029
factor(Year.built)1956	0.511	0.182	2.816	0.005
factor(Year.built)1957	0.441	0.183	2.408	0.016
factor(Year.built)1958	0.497	0.187	2.660	0.008
factor(Year.built)1959	0.507	0.183	2.772	0.006
factor(Year.built)1960	0.457	0.184	2.479	0.014
factor(Year.built)1961	0.542	0.186	2.918	0.004
factor(Year.built)1962	0.543	0.185	2.928	0.004
factor(Year.built)1963	0.444	0.186	2.389	0.017
factor(Year.built)1964	0.409	0.191	2.143	0.033
factor(Year.built)1965	0.495	0.187	2.649	0.008
factor(Year.built)1966	0.469	0.184	2.549	0.011
factor(Year.built)1967	0.653	0.196	3.327	0.001
factor(Year.built)1968	0.569	0.188	3.021	0.003
factor(Year.built)1969	0.613	0.187	3.279	0.001
factor(Year.built)1971	0.493	0.206	2.389	0.017
factor(Year.built)1972	0.558	0.186	2.997	0.003
factor(Year.built)1973	0.563	0.200	2.817	0.005
factor(Year.built)1974	0.482	0.195	2.471	0.014
factor(Year.built)1975	0.584	0.205	2.843	0.005
factor(Year.built)1976	0.510	0.185	2.750	0.006
factor(Year.built)1977	0.628	0.185	3.400	0.001
factor(Year.built)1978	0.564	0.184	3.067	0.002
factor(Year.built)1979	0.582	0.189	3.081	0.002
factor(Year.built)1980	0.628	0.187	3.348	0.001
factor(Year.built)1981	0.641	0.196	3.270	0.001
factor(Year.built)1982	0.664	0.189	3.506	0.001
factor(Year.built)1983	0.514	0.196	2.623	0.009
factor(Year.built)1984	0.612	0.185	3.310	0.001
factor(Year.built)1985	0.439	0.187	2.355	0.019
factor(Year.built)1986	0.672	0.189	3.558	0.000
factor(Year.built)1987	0.546	0.185	2.957	0.003
factor(Year.built)1988	0.642	0.192	3.349	0.001
factor(Year.built)1989	0.633	0.188	3.359	0.001
factor(Year.built)1990	0.747	0.221	3.389	0.001
factor(Year.built)1991	0.794	0.194	4.093	0.000
factor(Year.built)1992	0.786	0.190	4.145	0.000
factor(Year.built)1993	0.626	0.207	3.018	0.003
factor(Year.built)1994	0.805	0.208	3.860	0.000
factor(Year.built)1995	0.756	0.198	3.827	0.000
factor(Year.built)1996	0.758	0.190	3.986	0.000

factor(Year.built)1997	0.897	0.192	4.680	0.000
factor(Year.built)1998	0.911	0.252	3.617	0.000
factor(Quality)2	-0.246	0.035	-7.107	0.000
factor(Quality)3	-0.382	0.045	-8.571	0.000
Lot.size	0.000	0.000	6.646	0.000

Table 8: 95% Coefficients Confidence Intervals

	2.5 %	97.5 %
(Intercept)	10.945	11.713
Finished.square.feet	0.000	0.000
factor(Year.built)1900	-0.139	0.845
factor(Year.built)1908	-0.128	0.860
factor(Year.built)1914	0.056	1.045
factor(Year.built)1918	-0.163	0.688
factor(Year.built)1920	-0.015	0.840
factor(Year.built)1921	-0.053	0.935
factor(Year.built)1922	-0.081	0.770
factor(Year.built)1923	-0.089	0.892
factor(Year.built)1925	0.184	0.950
factor(Year.built)1927	-0.161	0.822
factor(Year.built)1928	-0.447	0.539
factor(Year.built)1934	-0.083	0.772
factor(Year.built)1935	-0.139	0.713
factor(Year.built)1936	-0.141	0.843
factor(Year.built)1937	0.359	1.341
factor(Year.built)1938	-0.069	0.792
factor(Year.built)1939	-0.009	0.976
factor(Year.built)1940	-0.006	0.773
factor(Year.built)1941	0.009	0.786
factor(Year.built)1942	0.047	1.032
factor(Year.built)1944	-0.044	0.813
factor(Year.built)1946	0.046	0.851
factor(Year.built)1947	0.057	0.804
factor(Year.built)1948	-0.023	0.728
factor(Year.built)1949	0.018	0.798
factor(Year.built)1950	0.074	0.807
factor(Year.built)1951	-0.024	0.702
factor(Year.built)1952	0.219	0.973
factor(Year.built)1953	0.069	0.804
factor(Year.built)1954	0.147	0.887
factor(Year.built)1955	0.042	0.775

	2.5 %	97.5 %
factor(Year.built)1956	0.154	0.868
factor(Year.built)1957	0.081	0.801
factor(Year.built)1958	0.130	0.864
factor(Year.built)1959	0.147	0.866
factor(Year.built)1960	0.095	0.820
factor(Year.built)1961	0.177	0.907
factor(Year.built)1962	0.179	0.907
factor(Year.built)1963	0.079	0.810
factor(Year.built)1964	0.034	0.785
factor(Year.built)1965	0.128	0.862
factor(Year.built)1966	0.107	0.831
factor(Year.built)1967	0.267	1.039
factor(Year.built)1968	0.199	0.939
factor(Year.built)1969	0.245	0.980
factor(Year.built)1971	0.087	0.898
factor(Year.built)1972	0.192	0.924
factor(Year.built)1973	0.170	0.956
factor(Year.built)1974	0.099	0.866
factor(Year.built)1975	0.180	0.987
factor(Year.built)1976	0.145	0.874
factor(Year.built)1977	0.265	0.992
factor(Year.built)1978	0.203	0.925
factor(Year.built)1979	0.211	0.953
factor(Year.built)1980	0.259	0.996
factor(Year.built)1981	0.256	1.027
factor(Year.built)1982	0.292	1.037
factor(Year.built)1983	0.129	0.900
factor(Year.built)1984	0.248	0.975
factor(Year.built)1985	0.073	0.806
factor(Year.built)1986	0.301	1.043
factor(Year.built)1987	0.183	0.909
factor(Year.built)1988	0.265	1.019
factor(Year.built)1989	0.263	1.003
factor(Year.built)1990	0.314	1.181
factor(Year.built)1991	0.413	1.175
factor(Year.built)1992	0.413	1.158
factor(Year.built)1993	0.218	1.034
factor(Year.built)1994	0.395	1.214
factor(Year.built)1995	0.368	1.145
factor(Year.built)1996	0.384	1.132
factor(Year.built)1997	0.520	1.274
factor(Year.built)1998	0.416	1.405
factor(Quality)2	-0.315	-0.178

	2.5 %	97.5 %
factor(Quality)3	-0.470	-0.295
Lot.size	0.000	0.000

Secondary Objective Analysis

Summary statistics were produced for baseline characteristics (see Tables 9 - 17, below).

Table 9: Number of Bedrooms Proportion Table

Number.of.bedrooms	n	Proportion
0	1	0.002
1	9	0.017
2	64	0.123
3	202	0.387
4	179	0.343
5	52	0.100
6	12	0.023
7	3	0.006

Table 10: Number of Bathrooms Proportion Table

Number.of.bathrooms	n	Proportion
0	1	0.002
1	71	0.136
2	171	0.328
3	175	0.335
4	84	0.161
5	17	0.033
6	1	0.002
7	2	0.004

Table 11: Style Proportion Table

Style	n	Proportion
1	214	0.410
2	58	0.111
3	64	0.123
4	11	0.021
5	18	0.034
6	18	0.034
7	136	0.261

9	1	0.002
10	1	0.002
11	1	0.002

Table 12: Garage Size (in Cars) Proportion Table

Garage.size	n	Proportion
0	7	0.013
1	52	0.100
2	353	0.676
3	106	0.203
4	2	0.004
5	1	0.002
7	1	0.002

Table 13: Pool Indicator Proportion Table

Pool	n	Proportion
0	486	0.931
1	36	0.069

Table 14: Quality Proportion Table

Quality	n	Proportion
1	68	0.130
2	290	0.556
3	164	0.314

Table 15: Air Conditioning Indicator Proportion Table

Air.conditioning	n	Proportion
0	88	0.169
1	434	0.831

Table 16: Highway Adjacency Indicator Proportion Table

Adjacent.to.highway	n	Proportion
0	511	0.979
1	11	0.021

Table 17: Summary Statistics for Continuous Variables

Metric	Sales Price ($)	Finished Square Feet	Year Built	Lot Size (square feet)
Minimum	84000.0	980.00	1885	4560.00
1st Quartile	180000.0	1701.25	1956	17204.75
Median	229900.0	2061.00	1966	22200.00
Mean	277894.2	2260.63	1967	24369.70
3rd Quartile	335000.0	2636.25	1981	26786.75
Max	920000.0	5032.00	1998	86830.00
Std Dev	137923.4	711.07	18	11684.08

Conclusions

A hedonic multiple linear regression model was created to predict residential Sales Price using Number of Bedrooms, Year Built, Quality, and Lot Size. All of these factors were listed as common for hedonic models by Sirmans, MacDonald, Macpherson, and Zietz (2005). The final model transformed Sales Price using the natural log function, in order to normalize the residuals. A suspicious data point without any bedrooms or bathrooms was explored, but ultimately left in the data due to lack of impact on the final model. Finally, summary statistics were shown for all data. It is not recommended that this model be used to predict residential home prices in areas outside of the data source, as hedonic models will vary by location. Instead, this analysis may be used as a starting point when considering data and methods for model development.

References

Sirmans, G.S., MacDonald, L., Macpherson, D.A., & Zietz, E.M. (2005). The Value of Housing Characteristics: A Meta Analysis. *The Journal of Real Estate Finance and Economics, 33*(3), 215-240.

Monson, M. (2009). Valuation using hedonic pricing models. *Cornell Real Estate Review, 7,* 62-73.

Kutner, M.H., Nachtsheim, C.J., & Neter, J. (2014). *Applied Linear Regression Models* (4th ed.). New York, NY; McGraw-Hill/Irwin.

Dean, A., Voss, D., & Draguljic, D. (2017). *Design and Analysis of Experiments* (2nd ed.). New York, NY; Springer International Publishing AG.

Document Information

All of the statistical analyses in this document were performed using R version 3.5.2 (2018-12-20). R packages were maintained using the packrat dependency management system. However, the Packrat-enabled files were incredibly large and difficult to submit via Blackboard. Packrat-version available upon request.

```
sessionInfo()
```

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_0.8.0.1     pander_0.6.3      broom_0.5.1       leaps_3.0
## [5] kableExtra_1.0.1  data.table_1.12.0 knitr_1.22
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.0        highr_0.7         pillar_1.3.1
##  [4] compiler_3.5.2    tools_3.5.2       digest_0.6.18
```

```
##  [7] lattice_0.20-38   nlme_3.1-137      evaluate_0.12
## [10] tibble_2.0.1      viridisLite_0.3.0 pkgconfig_2.0.2
## [13] rlang_0.3.1       rstudioapi_0.9.0  yaml_2.2.0
## [16] xfun_0.4          httr_1.4.0        stringr_1.3.1
## [19] xml2_1.2.0        generics_0.0.2    hms_0.4.2
## [22] grid_3.5.2        webshot_0.5.1     tidyselect_0.2.5
## [25] glue_1.3.0        R6_2.3.0          rmarkdown_1.11
## [28] tidyr_0.8.2       readr_1.3.1       purrr_0.3.0
## [31] magrittr_1.5      codetools_0.2-15  backports_1.1.3
## [34] scales_1.0.0      htmltools_0.3.6   assertthat_0.2.0
## [37] rvest_0.3.2       colorspace_1.4-0  stringi_1.3.1
## [40] munsell_0.5.0     crayon_1.3.4
```

Appendix

R code

```
# load packages
library(knitr)

# declare global chunk options
# knitr::opts_chunk$set(echo = FALSE)

# determine output format dynamically
out_type <- knitr::opts_knit$get("rmarkdown.pandoc.to")

# load data sets
library(data.table)
data<-fread('http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/
data/textdatasets/KutnerData/Appendix%20C%20Data%20Sets/APPENC07.txt')

colnames(data) <- c("Id.number", "Sales.price", "Finished.square.feet",
"Number.of.bedrooms", "Number.of.bathrooms", "Air.conditioning",
"Garage.size", "Pool", "Year.built", "Quality", "Style", "Lot.size",
"Adjacent.to.highway")

attach(data)

corrdata<-round(cor(data),2)
rownames(corrdata) <- c("ID #","Price","Sqr Ft", "Bed","Bath","A/C",
"Garage","Pool","Year","Quality","Style","Size","Highway")

corrdata1<-corrdata[,c(1:7)]
corrdata2<-corrdata[,c(8:13)]


library(kableExtra)
kable(corrdata1, col.names=
        c("ID #","Price","Sqr Ft", "Bed","Bath","A/C","Garage"),
      caption="Correlations (Continued Below)",
      longtable = TRUE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
        row_spec(0,bold=TRUE)



kable(corrdata2, col.names=
        c("Pool","Year","Qual.","Style","Size","Highway"),
```

```
      caption="Correlations (Continued)",
      longtable = TRUE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
        row_spec(0,bold=TRUE)

data("data")
par(mfrow=c(1,2))


# Define the position of tick marks
v1 <- c(0,200000,400000,600000,800000)

# Define the labels of tick marks
v2 <- c("0","200","400","600","800")

# Plot the data
boxplot(data[,c(2)],
          boxwex = 0.5, main="Sales Price", ylab="$1,000s",
     yaxt = "n")

# Add axis to the plot
axis(side = 2,
     at = v1,
     labels = v2,
     tck=-.1,
     tcl = -0.5,
     cex.axis=1.05,
     font.axis=5)


boxplot(data[,c(3)],
          boxwex = 0.5, main="Finished Square Feet", ylab="Square Feet")

library("leaps")
rs<-regsubsets(Sales.price ~ Finished.square.feet + Number.of.bedrooms
+ Number.of.bathrooms + factor(Air.conditioning) +  Garage.size
+   factor(Pool) +  Year.built +    factor(Quality) + factor(Style)
+   Lot.size +  factor(Adjacent.to.highway), data= data,
method="exhaustive")


mdl_tbl <- data.frame("Model" = c("Finished Sqr Ft","+ Quality",
"+ Style","+ Year built","+ Lot Size", "+ Bedroom Ct",
"+ Garage Size", "+ Highway Adjacency"),
  "Adj. R^2" = round(summary(rs)$adjr2,4),
```

```
  "BIC" = round(summary(rs)$bic,2))

library(kableExtra)
kable(mdl_tbl, col.names=c("Model","Adj. R^2","BIC"),
caption="Potential Models",longtable = TRUE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
        row_spec(0,bold=TRUE)

library(kableExtra)
library(broom)
library(pander)

model1<-lm(Sales.price ~ Finished.square.feet + Number.of.bedrooms
+   factor(Year.built) +    factor(Quality) + factor(Style) +   Lot.size,
data= data)
#Number of Bedrooms is least significant, so remove
model2<-lm(Sales.price ~ Finished.square.feet + factor(Year.built)
+   factor(Quality) + factor(Style) +   Lot.size, data= data)
#Style is least significant, so remove
model3<-lm(Sales.price ~ Finished.square.feet + factor(Year.built)
+   factor(Quality) +  Lot.size, data= data)
#Lot Size is least significant, so remove
model4<-lm(Sales.price ~ Finished.square.feet + factor(Year.built)
+   factor(Quality) , data= data)

pander(anova(model1))

Predictors<-matrix(c("All","-Bedrooms","-Style","-Lot Size"),
nrow=4,ncol=1)

summtable<-rbind(glance(model1)[c(1,2,3,4,5)],
  glance(model2)[c(1,2,3,4,5)],
  glance(model3)[c(1,2,3,4,5)],
  glance(model4)[c(1,2,3,4,5)])
summtable2<-cbind(Predictors,summtable)

summtable2 %>% kable(caption = "Model Characteristics",
longtable = TRUE, digits = 3)


qqnorm(residuals(model3),main="QQ-Plot")
qqline(residuals(model3))

shapiro.test(residuals(model3))
```

```
model5<-lm(log(Sales.price) ~ Finished.square.feet + factor(Year.built)
+    factor(Quality) +    Lot.size, data=data)

qqnorm(residuals(model5),main="QQ-Plot")
qqline(residuals(model5))

shapiro.test(residuals(model5))


plot(fitted(model5),rstandard(model5), main="Fitted Values v. Residuals",
ylab="Standardized Residuals",xlab="Fitted PSA Levels")
abline(h=0)


newdata<-data[-108,]

model6<-lm(log(Sales.price) ~ Finished.square.feet + factor(Year.built)
+    factor(Quality) +    Lot.size, data=newdata)


Predictors<-matrix(c("All Data","-Entry 108"),nrow=2,ncol=1)

summtable<-rbind(glance(model5)[c(1,2,3,4,5)],
  glance(model6)[c(1,2,3,4,5)])
summtable2<-cbind(Predictors,summtable)

summtable2 %>% kable(caption =
"Model Characteristics w/ and w/o Unusual Entry",
longtable = TRUE, digits = 3)

kable(summary(model5)$coef,caption = "Model Coefficients Summary",
longtable = TRUE, digits=3)

library(knitr)
library(kableExtra)
confint(model5) %>% kable(format="pandoc",
caption = "95% Coefficients Confidence Intervals",
longtable = TRUE, digits = 3)

library(dplyr)
library(knitr)

par(mfrow=c(2,4))
```

```
data%>%
  group_by(Number.of.bedrooms)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Number of Bedrooms Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Number.of.bathrooms)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Number of Bathrooms Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Style)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Style Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Garage.size)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Garage Size (in Cars) Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Pool)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Pool Indicator Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Quality)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Quality Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Air.conditioning)%>%
  summarize(n=n())%>%
```

```
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Air Conditioning Indicator Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Adjacent.to.highway)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Highway Adjacency Indicator Proportion Table",
      longtable = TRUE)


attach(data)

library(knitr)
library(kableExtra)

text_tbl <- data.frame("Metric" = c("Minimum", "1st Quartile",
"Median","Mean","3rd Quartile","Max","Std Dev"),
  "Sales Price ($)" = c(round(min(Sales.price),2),
    round(quantile(Sales.price,.25),2),
    round(quantile(Sales.price,.5),2),
    round(mean(Sales.price),2),
    round(quantile(Sales.price,.75),2),
    round(max(Sales.price),2),
    round(sqrt(var(Sales.price)),2)),
  "Finished Square Feet" = c(round(min(Finished.square.feet),2),
    round(quantile(Finished.square.feet,.25),2),
    round(quantile(Finished.square.feet,.5),2),
    round(mean(Finished.square.feet),2),
    round(quantile(Finished.square.feet,.75),2),
    round(max(Finished.square.feet),2),
    round(sqrt(var(Finished.square.feet)),2)),
  "Year Built" = c(round(min(Year.built),0),
    round(quantile(Year.built,.25),0),
    round(quantile(Year.built,.5),0),
    round(mean(Year.built),0),
    round(quantile(Year.built,.75),0),
    round(max(Year.built),0),
    round(sqrt(var(Year.built)),0)),
  "Lot Size (square feet)" = c(round(min(Lot.size),2),
    round(quantile(Lot.size,.25),2),
    round(quantile(Lot.size,.5),2),
    round(mean(Lot.size),2),
    round(quantile(Lot.size,.75),2),
```

```
    round(max(Lot.size),2),
    round(sqrt(var(Lot.size)),2))
)

kable(text_tbl, col.names=c("Metric", "Sales Price ($)",
"Finished Square Feet", "Year Built","Lot Size (square feet)"),
caption = "Summary Statistics for Continuous Variables",
longtable = TRUE
      ) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
        row_spec(0,bold=TRUE)
```