A Linear Regression Model to Predict House Price in the Ames Data Set

Devane Risby

Contents

1.0 Introduction

This report seeks to assess the impact of property characteristics on determining property value within the framework of a hedonistic pricing model using the Ames house dataset. Utilizing a multiple linear regression model, we aim to derive an accurate estimation of property value by examining five proposed hypotheses and seven additional variables with the data set containing 2867 observations. These hypotheses, rooted in prior literature, will guide our analysis with the aim replicate existing findings.

1.1 Literature Review

The pricing models employed in real estate estimates often rely on hedonic pricing frameworks which assess value based on a blend of internal and external attributes that contribute to utility or desirability (Hargrave, M. 2021; Rosen, S. 1974). Fundamental to this assessment is the correlation between value and various underlying factors. Total lot area has been demonstrated as a consistent positive relationship with value (McMillen, D.P. 2008). Furthermore, in Zietz et al (2007) analysis an observation of 52 regression models studied found a positive relationship appears 45 times. Additionally, work by Sirmans, G.S. et al. (2006) found that size was influenced by geographical location but time had a limiting influence. Furthermore the total number of bedrooms has been linked to property value increased of $9,500 per room (Emrath, P. and Taylor, H. 2012) with Sirmans, G.S. et al. (2006) finding showing that bedrooms were not geographically linked, indicating overall appeal.

A property age has been demonstrated to have a notable negative correlation with its value, albeit with diminishing impacts as the property age increases into extremes. Particularly, lower-priced homes exhibit the most pronounced depreciation when newly constructed (Zietz, J., Zietz, E.N. and Sirmans, G.S. 2007). Xu, Y. et al. (2015) discovered that properties tend to experience a substantial decline in value during the their initial years, which tapers off over time.

The location of a property is anticipated to significantly influence its value, as evidenced across various scales of study. Feng, X. et al (2023) observed a correlation between the location of jobs and its impact on property value, with a more local examination by Aziz et al (2020) revealing a direct link between neighborhood services or quality, and a given property value. However, they note that in financially disadvantaged areas this impact is more pronounced, while wealthier areas may even be negative. Both these studies indicated that location plays a role in property value.

Moreover, the quality of a property has been identified as a factor influencing value (Baum, A. (1994). Yet, this factor seems to be greatly influenced by prevailing market conditions. In weakened markets conditions, the overall quality of a property demonstrates limited impact (Miller, N., Sah, V. and Sklarz, M. (2018) suggesting that it is of a lower desirable rank. Additionally, work by Jandásková, T. et al. (2022) backed this up by finding significant correlation between property quality and time on the market, further indicating desirability.

1.2 Hypothesis Formulation

The hypothesis listed below where derived from the existing literature with the aim to find similar results to prove the quality of the regression model created. H1 was selected as a base line to determine the overall reliability of the model.

H1 Larger size is associated with higher sale prices.

H2 Older age correlates with lower sale prices.

H3 Decreasing house quality corresponds to decreasing sale prices.

H4 A higher number of bedrooms is linked to increased sale prices.

H5 The neighborhood significantly influences house prices.

2.0 Methodology

The study leveraged the R programming language in conjunction with packages (reference table 1) to create a comprehensive report. Two multiple linear regression models were formulated; first utilizing the hypotheses variables to determine the overall accuracy of the model (see formula listed below), with the second revised model utilizing 7 additional random variables to demonstrate a more accurate model.

$$sale\ price = \beta_0 + \beta_1 \times Lot\ Area + \beta_2 \times Age + \beta_3 \times Bedroom + \beta_4 \times House\ Quality + \beta_5 \times Neighbourhood$$

The models accuracy was determined following the work outlined by Schneider, A. et al. (2010). By assessing the fitted variables against the residuals, multicollinearity test, adjusted R^2 assessment and the Mean Absolute Error (MSE), these outcomes were utilized to refine the model into a more accurate predictive model.

Pearson's correlation, Spearman's rank correlation, and ANOVA tests (formulas listed below) were performed to assess the correlation of the independent variables with the dependent variable. This analysis was complemented by visualizations generated through ggplot2 to reflect the overall relationships between these variables. The ANOVA test is used only for the neighborhood variable to determine if the mean of sale price is influenced.

Peason's correlation

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2\ (y_i - \bar{y})^2}}$$

Spearman's Ranks Correlation

$$p = 1 - \frac{6\Sigma {d_i}^2}{n(n^2 - 1)}$$

ANOVA Test

$$Total\ Variation = between\ group\ vairation + within\ group\ variation$$

2.1 Data Quality Assessment and Treatment

The Ames dataset underwent a data quality assessment to determine its overall usability. To reduce the overall scope, the removal of variables was determined to be unnecessary due to limited significant value. All task undertaken can been seen in Appendix 1 file 1.

2.1.1 Appropriate Data Type

To allow for analysis to occur, character type variables were factorized. A for loop was used on all variables and automatically converted them to a factor under the condition. This allowed the data to be group for visual and summary analysis.

2.1.1 Identifying and Addressing Missing and Extreme Values

The dataset exhibited several outliers and missing values, prompting the need for corrective actions to improve the analysis. A boxplot (Figure 1) was used to examine 'lot_area,' revealing numerous extreme values exceeding 100,000 which were consequently removed. Furthermore, a data summary revealed instances where the 'lot_area' was as small as 13. Deemed implausible, values below 100 were excluded to ensure data integrity.

Similar analysis was conducted for 'sale_price', identifying outliers beyond $750,000 (See Figure 2). Records linked to a value greater than this were removed. Similar steps were taken for other variables, as outlined in the appendix.

2.1.3 Feature Engineering

To enhance the readability, the 'year_built' variable was converted into a new variable named 'age'. This transformation utilized the formula seen below, resulting a in derivation of an age for each record. This facilitated the analysis focusing on property age, enabling a more comprehensive exploration.

$$Age = \ Current\ Year - Year\ Built$$

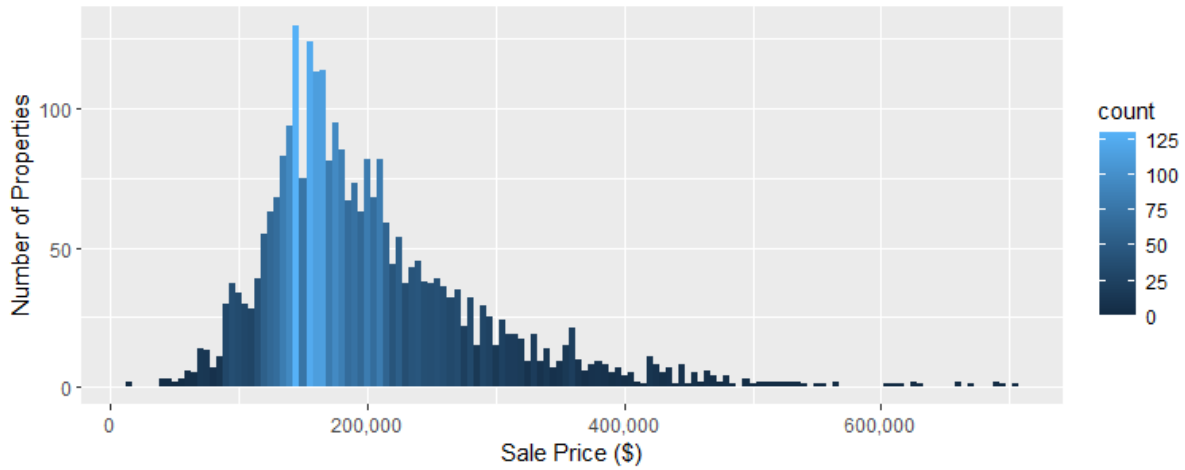
2.1.4 Imputation of missing value for bedroom

The 'bedroom' variable initially included 7 value of 0, which were viewed as miss inputs. The 'mice' package was utilized to predict and substitute for these missing value to enhance the datasets completeness. This approach aimed to preserve the data's integrity by reducing the necessity for removing incomplete records (Buuren, S. van and Groothuis-Oudshoorn, K. (2011).

3.0 Results and Discussion

A review of the sale distribution (Figure 3) illustrates a symmetrical distribution around the median value of $181,083. A skewness of 1.58 indicates a positive lean towards larger values, which is expected due to an absolute 0 value. Notably, a few outliers remain above the upper quartile of $241,109. Moreover, the relatively small SD of $88,580.63 suggests a compact clustering of data points.

Figure 3: Histogram of Sale Price

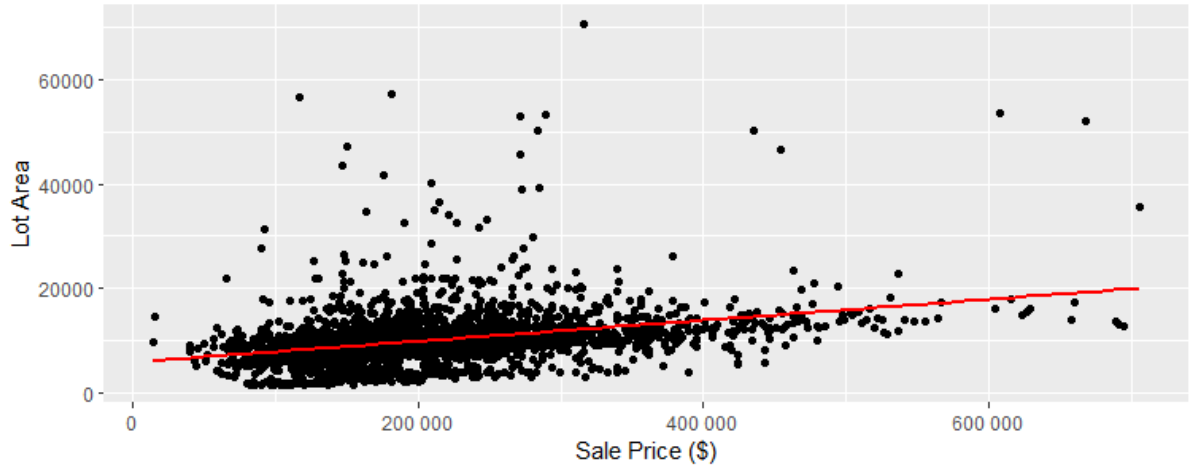


3.1 Hypotheses Testing

3.1.1 Hypothesis 1

A positive association between larger lot areas and sale prices find support in the data distribution. Figure 4 shows a discernible positive trend line indicating that increased lot area aligns with higher sales price. Additionally a Pearsons correlation of 0.3448847 shows a moderate positive correlation. It is essential to acknowledge the presence of numerous outliers, particularly those with a lot area exceeding 40,000. Despite their substantial size, these outliers display a limited correlation with sale price. This observation hints at a potential optimal lot area range, beyond which a non-existent or negative relationship could emerge.

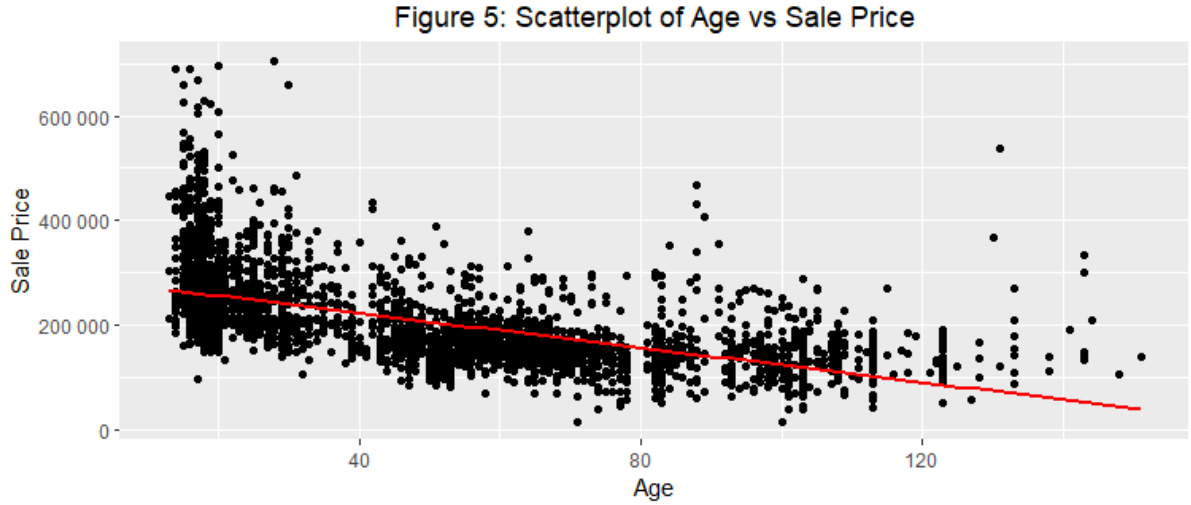
Figure 4: Scatterplot of Lot Area vs Sales Price



3.1.2 Hypothesis 2

Older age correlating with lower sales prices is supported, with figure 5 showing a negative trend line. There is a significant clustering of sale price at lower ages, with the largest sale price being found here. A Pearsons correlation of -0.5663633 indicate a moderate to strong negative correlation.

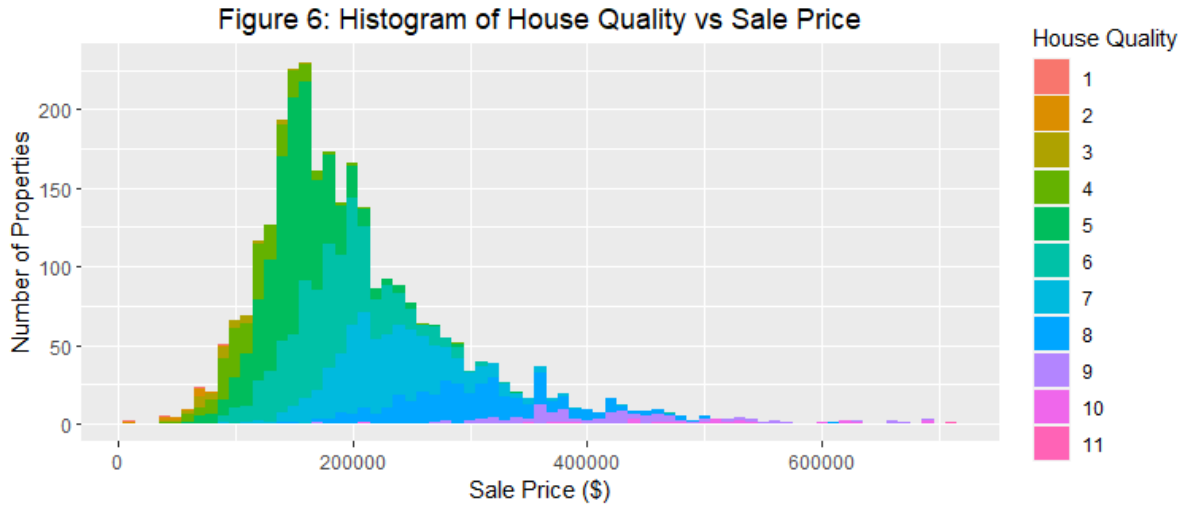
Figure 5: Scatterplot of Age vs Sale Price



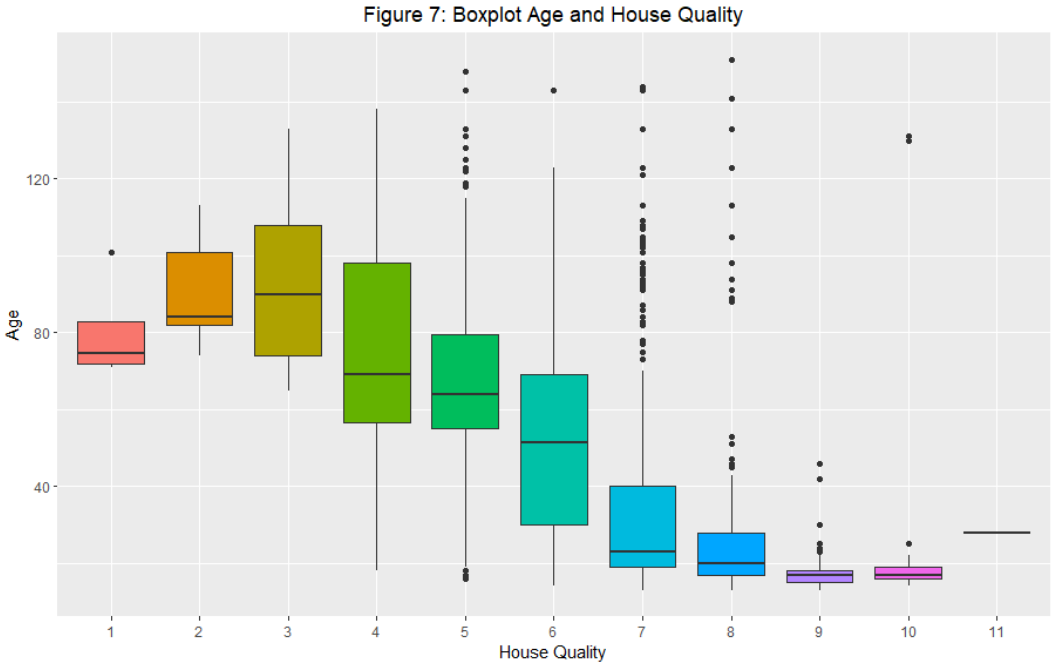
3.1.3 Hypothesis 3

Decreasing house quality corresponding with decreasing sale price is supported, with figure 6 showing a clear correlation. Properties with a house quality of less than 4 shows to be distributed at the lower end with an increase to 7 showing a greater distribution on the right. The effects of quality decreases as price increases, with the final house qualities of 9+ covering a larger area of the graph

compared to lower qualities. Additionally, a Spearman's rank of 0.8103997 shows an extremely strong positive relationship.



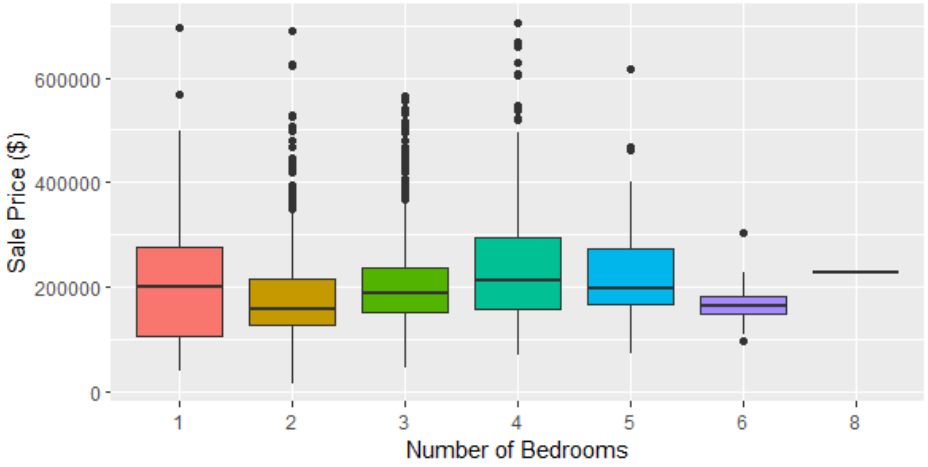
An intriguing relationship emerges with age, displaying a negative correlation. Figure 7 suggests that increases in property age coincides with a decline in quality. The anomaly within 1 and 2 is noteworthy, exhibiting a younger average age. This deviation from the trend might relate to properties initially constructed to a lower standard which are then removed from the market leaving a survivorship bias.



3.1.4 Hypothesis 4

A higher number of bedrooms being linked to increased sale price was not conclusively determined. Figure 8 shows that there are micro trends in the data with an increase in prices between 2 and 4 bedrooms and a drop there after following a bell curve. However, there are significant outliers to this trend. 1 bedroom being more valuable than 2 suggests demand for single bed houses being greater due to macro market trends. Furthermore, a spearman's rank of 0.1914932 shows a limited positive relationship, although this isn't significant enough to determine correlation.

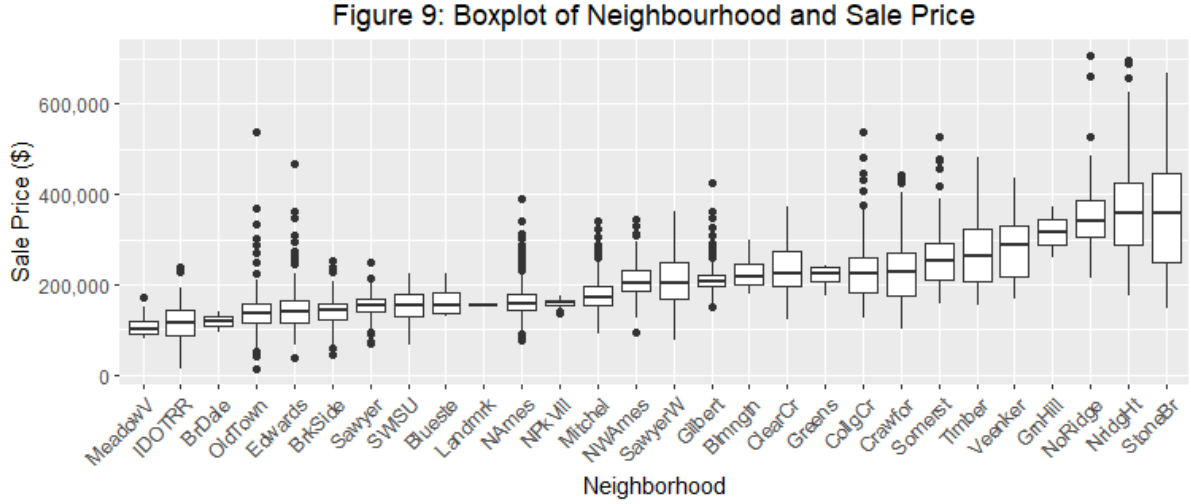
Figure 8: Boxplot of Bedrooms and Sale Price



3.1.5 Hypothesis 5

Neighbourhood's significant influence on sale price was proven to be correct with a significant association. Figure 9 shows that there is a trend based on the neighbourhood selected, with MeadowV containing the least valuable properties compared to NoRidgHt, showing that the variable is playing a significant role. Additionally, an ANOVA test shows that the relationship isn't random with a p value of 2.2 × 10^-16.

Figure 9: Boxplot of Neighbourhood and Sale Price



3.2 Regression Model

Table 1 shows that the outcome of hypotheses variables all contained a significance of below 0.01. However, Models 2 and 3 show that the H5 variable experiencing extreme fluctuations is significant when compared against the vacuum model. A multicollinearity test (Table 3) shows that the H5 variable had significant correlation within the model at 11.362, and was removed to form the revised model.

Table 2: Multicollinearity test for Hmodel

Variable	GVIF	Df	GVIF^(1/(2*Df))
lot_area	1.473	1	1.213
Age	5.138	1	2.267
Bedroom	1.190	1	1.091

House_qaility	2.430	1	1.559
Neighbourhood	11.362	27	1.046

The refined regression model (Table 3) reveals an adjusted R^2 of 0.871, showcasing a robust accuracy for the model (equation depicted below). The model exhibits a median error of -491, suggesting a relatively accurate prediction, with a MEA of 23,186 further backing this accuracy. The model is minimally affected by outlier data, as indicated by a maximum Cook's distance of 0.09. However a standard error of 31,630.090 signifies some, albeit not substantial, variance.

Revised Regression Model Formula

$$sale\ price = \beta_0 + \beta_1 \times Lot\ Area + \beta_2 \times Age + \beta_3 \times Bedroom + \beta_4 \times House\ Quality + \beta_5 \times Heat\ Quality + \beta_6 \times Year\ Remod + \beta_7 \times Basement\ Quality + \beta_8 \times Basement\ Area + \beta_{10} \times Aircon + \beta_{11} \times Floor1\ Sq + \beta_{12} \times Floor2\ Sq + \beta_{13} \times Sale\ Conditon$$

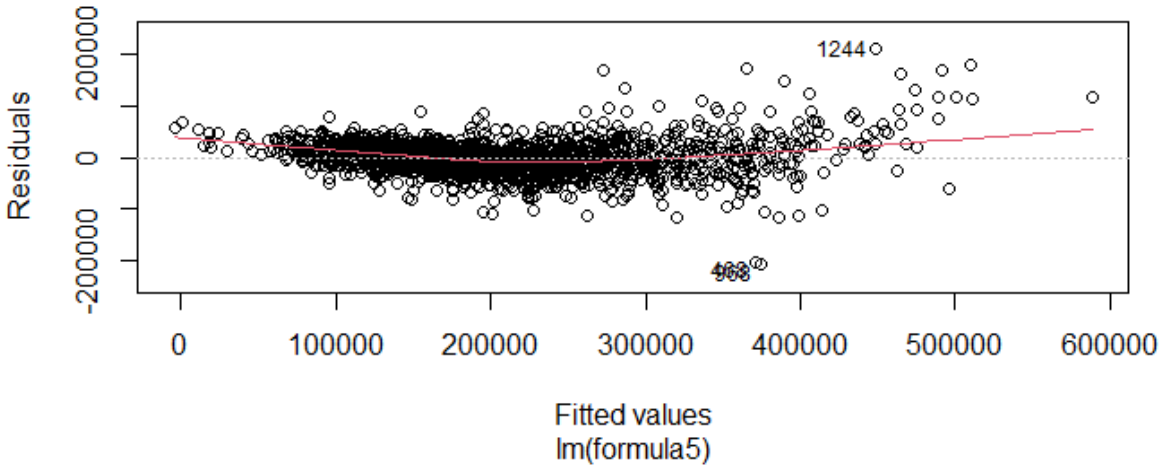
Table 3: Revised Regression Models

	Dependent variable:		
	sale_price		
	(1)	(2)	(3)
lot_area	4.699***	4.440***	2.286***
	(0.219)	(0.205)	(0.170)
age	-464.610***	-200.691***	-317.793***
	(42.077)	(53.845)	(43.555)
bedroom	6,255.836***	8,935.441***	-12,980.470***
	(1,289.996)	(1,226.754)	(1,235.808)
house_quality	42,058.700***	32,876.200***	17,369.910***
	(915.059)	(1,007.387)	(878.026)
heat_qualFa		-18,397.010***	-15,707.590***
		(6,089.965)	(4,661.600)
heat_qualGd		-3,780.572	-2,461.175
		(2,855.893)	(2,184.526)
heat_qualPo		-2,355.440	12,750.910
		(29,619.350)	(22,609.380)
heat_qualTA		-10,808.550***	-5,867.428***
		(2,676.937)	(2,059.376)
year_remod		282.236***	280.266***
		(66.016)	(51.466)
basement_qualFa		-80,383.460***	-45,413.320***
		(7,839.047)	(6,074.203)
basement_qualGd		-73,325.180***	-52,230.210***
		(3,841.553)	(3,051.983)
basement_qualTA		-80,951.900***	-48,324.850***
		(5,029.054)	(3,979.197)
bsmt_area			39.961***
			(4.182)
airconY		6,235.553	11,580.250***
		(4,622.088)	(3,584.045)
floor1_sf			79.999***
			(4.508)
floor2_sf			82.934***
			(2.857)
full_bath			-10,422.490***
			(1,987.856)
sale_condAdjLand			21,640.210*
			(12,490.860)
sale_condAlloca			31,101.590***
			(10,036.590)
sale_condFamily			-4,444.628
			(6,118.815)
sale_condNormal			12,688.130***
			(2,892.947)
sale_condPartial			24,069.000***

			(4,013.206)
Constant	-93,540.670***	-546,690.400***	-548,785.500***
	(7,568.753)	(131,322.100)	(102,215.900)
Observations	2,009	1,952	1,952
R^2	0.732	0.780	0.873
Adjusted R^2	0.732	0.778	0.871
Residual Std. Error	45,833.860 (df = 2004)	41,594.370 (df = 1938)	31,683.090 (df = 1929)
F Statistic	1,370.970*** (df = 4; 2004)	527.087*** (df = 13; 1938)	600.949*** (df = 22; 1929)
Note:			*p**p***p<0.01

Figure 10 shows the residuals vs fitted, displaying a somewhat random distribution but slightly skewed towards the higher values, indicating a reduction in accuracy as sale price increases into the extremes. This is further backed up by Figure 11, indicating a bias towards both extremes, potentially influencing the accuracy of the model at these extremes.

Figure 10: Residuals vs Fitted



4.0 Conclusion

Hypotheses 1-4 were proven correct, with H1 conditions having a significant impact on property value and good indication of the overall value of a property. Lot area has an impact on sale price at ration of 1|2.826, indicating that larger houses received a direct benefit in value. The model falls in line with Zahirovich-Herbert, V. and Gibler, K.M. (2014) research indicating that size and age have a direct impact in new builds. However, that paper indicates that hypothesis 5 may also be true. The model's assessment of preferred bedroom ranges lacks accuracy with a constant -12,980.470 per unit increases, primarily due to an overgeneralization of the constant negative association with bedrooms excluding desirable ranges.

The H5 variable may wield considerable influence on property value prediction. Nonetheless, the existing model grapples with a significant challenge of multicollinearity, thereby hampering the model's accuracy. A deeper exploration and manipulation of data are imperative to establish a more accurate correlation involving this variable. Gao, G. et al. (2022) reinforce the importance of location in determining property value, emphasizing its intricate nature within the model.

5.0 Reflective Commentary

The dataset faced limitations in accurately predicting sale prices, particularly at the extremes. To address this, leveraging synthetic data might fill the gaps, enhancing the comprehensiveness of the dataset for more accurate modeling (Noble, A. (2022). However, adopting this method necessitates extensive data creation, expanding the project's scope. Despite the additional workload, using synthetic data in regression modeling stands as an increasingly valuable approach (Yue, Y. *et al.* 2018).

Constructing the regression model presented an intriguing challenge that significantly enhanced my proficiency in the R programming language. While I previously worked with Python and Kotlin, transitioning to R proved relatively seamless. Engaging with various R package libraries allowed me to refine my workflow significantly. Unlike my prior undergraduate experience, I relied less on platforms like Stack Overflow to troubleshoot errors. Instead, I sought direct assistance from ChatGPT (OpenAI 2023), expediting the debugging process. This shift notably accelerated my coding pace, eliminating the time-consuming search for solutions in forums or intricate documentation breakdowns. However, it's essential to acknowledge that this approach might limit personal learning opportunities by not actively seeking solutions independently.

6.0 References

Aziz, A., Anwar, M.M. and Dawood, M. (2020) 'The impact of neighborhood services on land values: An estimation through the hedonic pricing model', GeoJournal, 86(4), pp. 1915–1925. doi:10.1007/s10708-019-10127-w.

Baum, A. (1994) 'Quality and property performance', Journal of Property Valuation and Investment, 12(1), pp. 31–46. doi:10.1108/14635789410050494.

Buuren, S. van and Groothuis-Oudshoorn, K. (2011) 'mice: Multivariate Imputation by Chained Equations in R', Journal of Statistical Software, 45(3). doi:10.18637/jss.v045.i03

Emrath, P. and Taylor, H. (2012) 'Housing Value, Costs, and Measures of Physical Adequacy', Cityscape, 14(1), pp. 99–125.

Feng, X. et al. (2023) 'Location, location, Location: Manufacturing and House price growth', The Economic Journal, 133(653), pp. 2055–2067. doi:10.1093/ej/uead008.

Gao, G. et al. (2022) 'Location-centered house price prediction: A multi-task learning approach', ACM Transactions on Intelligent Systems and Technology, 13(2), pp. 1–25. doi:10.1145/3501806.

Hargrave, M. (2021) Hedonic pricing: Definition, how the model is used, and example, Investopedia. Available at: https://www.investopedia.com/terms/h/hedonicpricing.asp (Accessed: 19 October 2023).

Jandásková, T. et al. (2022) 'Technical condition of houses: A framework for the Czech market', International Journal of Housing Markets and Analysis, 16(7), pp. 58–79. doi:10.1108/ijhma-07-2022-0106.

McMillen, D.P. (2008) 'Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients?', Journal of Urban Economics, 64(3), pp. 573–589. doi:10.1016/j.jue.2008.06.002.

Miller, N., Sah, V. and Sklarz, M. (2018) 'Estimating property condition effect on residential property value: Evidence from U.S. Home Sales Data', Journal of Real Estate Research, 40(2), pp. 179–198. doi:10.1080/10835547.2018.12091497.

Noble, A. (2022) What is synthetic data and how can it advance research and development ..., The Alan Turing Institute. Available at: https://www.turing.ac.uk/blog/what-synthetic-data-and-how-can-it-advance-research-and-development (Accessed: 23 November 2023).

OpenAI (2023) Chatgpt, ChatGPT. Available at: https://openai.com/chatgpt (Accessed: 04 December 2023).

Rosen, S. (1974) 'Hedonic prices and implicit markets: Product differentiation in pure competition', Journal of Political Economy, 82(1), pp. 34–55. doi:10.1086/260169.

Schneider, A., Hommel, G. and Blettner, M. (2010) 'Linear regression analysis', Deutsches Ärzteblatt international [Preprint]. doi:10.3238/arztebl.2010.0776.

Sirmans, G.S. et al. (2006) 'The value of housing characteristics: A meta analysis', The Journal of Real Estate Finance and Economics, 33(3), pp. 215–240. doi:10.1007/s11146-006-9983-5.

Xu, Y. et al. (2015) 'House age, Price and rent: Implications from land-structure decomposition', SSRN Electronic Journal. doi:10.2139/ssrn.2664928.

Yue, Y. et al. (2018) 'Synthetic Data Approach for classification and regression', 2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP) [Preprint]. doi:10.1109/asap.2018.8445094.

Zahirovich-Herbert, V. and Gibler, K.M. (2014) 'The effect of new residential construction on housing prices', Journal of Housing Economics, 26, pp. 1–18. doi:10.1016/j.jhe.2014.06.003.

Zietz, J., Zietz, E.N. and Sirmans, G.S. (2007) 'Determinants of house prices: A quantile regression approach', The Journal of Real Estate Finance and Economics, 37(4), pp. 317–333. doi:10.1007/s11146-007-9053-7.

Package Name	**Reference**	**License**
Tidyverse	Wickham, H. et al. (2019) 'Welcome to the Tidyverse', Journal of Open Source Software, 4(43), p. 1686. doi:10.21105/joss.01686.	MIT
ggplot2	Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.	MIT
Psych	William Revelle (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.3.9, https://CRAN.R-project.org/package=psych.	GPL-2, GPL-3
Scales	Hadley , W., Pedersen, T.L. and Seidel, D. (2023) *Scale functions for visualization*, *Scale Functions for Visualization* •. Available at: https://scales.r-lib.org/ (Accessed: 30 November 2023).	MIT
Stargazer	Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer	

Mice	Buuren, S. van and Groothuis-Oudshoorn, K. (2011) 'mice: Multivariate Imputation by Chained Equations in R', *Journal of Statistical Software*, 45(3). doi:10.18637/jss.v045.i03.	GPL-2, GPL-3
Caret	Kuhn, M. (2008) 'Building Predictive Models in R Using the caret Package ', *Journal of Statistical Software*, 28(5). doi:10.18637/jss.v028.i05.	GPL-2, GPL-3
Car	Fox J, Weisberg S (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.	GPL-2, GPL-3
Summarytools	Comtois, D. (2022) *Dcomtois/summarytools: R package to quickly and neatly summarize data*, *GitHub*. Available at: https://github.com/dcomtois/summarytools (Accessed: 05 December 2023).	GPL-2

Appendix 1

Figure 1: Boxplot of Lot Area

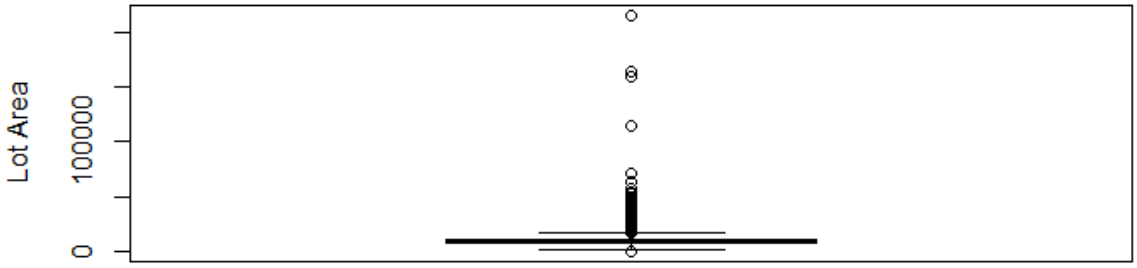


Figure 2: Boxplot of Sale Price

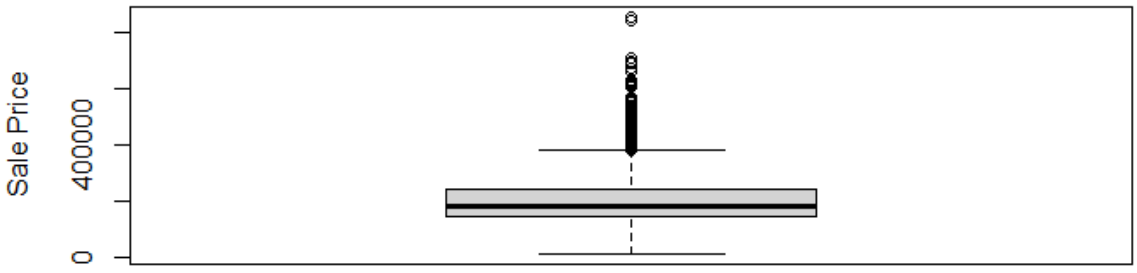


Figure 11: Q-Q Residual

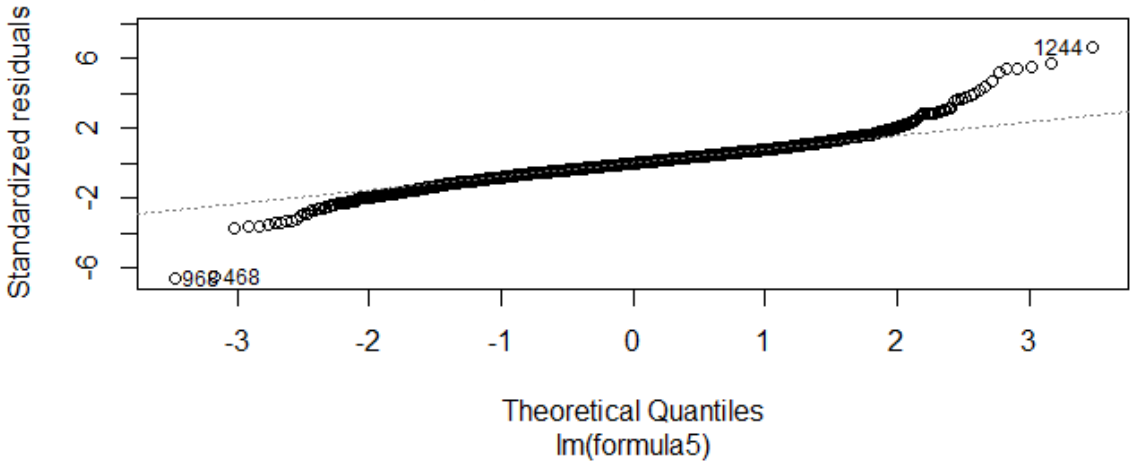


Table 1: Hypothesis Regression Models

	Dependent variable:		
	sale_price		
	(1)	(2)	(3)
lot_area	4.586***		4.699***
	(0.229)		(0.219)
age	-648.685***		-464.610***
	(68.455)		(42.077)
bedroom	8,257.052***		6,255.836***
	(1,218.499)		(1,289.996)
house_quality	32,671.010***		42,058.700***
	(1,013.261)		(915.059)
neighbourhoodBlueste	-16,157.410	-45,430.640*	
	(17,321.270)	(24,453.630)	
neighbourhoodBrDale	-24,728.220*	-93,107.630***	
	(13,059.540)	(18,171.340)	
neighbourhoodBrkSide	17,232.160	-75,466.020***	
	(11,638.550)	(15,009.020)	
neighbourhoodClearCr	3,827.234	25,600.800	
	(12,532.660)	(16,709.030)	

neighbourhoodCollgCr	-4,713.284	15,033.350
	(10,001.780)	(13,953.950)
neighbourhoodCrawfor	37,069.170***	19,616.550
	(11,446.850)	(15,185.990)
neighbourhoodEdwards	-4,614.308	-63,915.090***
	(10,717.700)	(14,294.330)
neighbourhoodGilbert	-21,492.610**	2,902.662
	(10,348.380)	(14,317.850)
neighbourhoodGreens	-10,684.650	-2,617.763
	(19,358.610)	(27,170.700)
neighbourhoodGrnHill	91,782.020***	102,754.700**
	(30,479.110)	(43,132.150)
neighbourhoodIDOTRR	-1,810.261	-99,000.650***
	(11,910.650)	(15,158.480)
neighbourhoodLandmrk	-25,455.590	-58,835.260
	(42,038.190)	(59,528.020)
neighbourhoodMeadowV	6,855.183	-108,212.600***
	(12,594.970)	(17,245.520)
neighbourhoodMitchel	-12,154.890	-29,213.410**
	(10,822.220)	(14,789.860)
neighbourhoodNAmes	-1,397.992	-49,197.950***
	(10,280.340)	(13,714.000)
neighbourhoodNoRidge	74,323.460***	151,334.600***
	(11,404.040)	(15,726.180)
neighbourhoodNPkVill	-13,007.720	-55,279.260***
	(14,036.670)	(19,687.040)
neighbourhoodNridgHt	63,935.540***	146,170.600***
	(10,386.650)	(14,414.680)
neighbourhoodNWAmes	-9,582.447	-1,395.791
	(10,695.690)	(14,634.640)
neighbourhoodOldTown	11,159.370	-76,124.170***
	(11,363.190)	(14,047.680)

neighbourhoodSawyer	-5,256.117	-59,429.900***	
	(10,876.970)	(14,594.250)	
neighbourhoodSawyerW	-10,321.980	-5,691.930	
	(10,519.800)	(14,607.440)	
neighbourhoodSomerst	7,648.785	42,797.710***	
	(10,145.920)	(14,294.330)	
neighbourhoodStoneBr	78,687.930***	159,215.900***	
	(11,979.600)	(16,709.030)	
neighbourhoodSWISU	10,025.700	-54,492.900***	
	(12,990.150)	(17,011.550)	
neighbourhoodTimber	19,404.220*	73,380.590***	
	(11,500.600)	(16,041.570)	
neighbourhoodVeenker	15,425.390	69,667.740***	
	(13,828.430)	(19,084.070)	
Constant	-37,636.820***	213,645.300***	-93,540.670***
	(11,982.680)	(13,310.870)	(7,568.753)
Observations	2,009	2,009	2,009
R^2	0.789	0.576	0.732
Adjusted R^2	0.786	0.570	0.732
Residual Std. Error	40,938.320 (df = 1977)	58,020.740 (df = 1981)	45,833.860 (df = 2004)
F Statistic	238.994*** (df = 31; 1977)	99.692*** (df = 27; 1981)	1,370.970*** (df = 4; 2004)
Note:			*p**p***p<0.01

Table 4: Summary Statistics of Hypothesis variables

	age	bedroom	house_quality	lot_area	sale_price
Mean	51.635	2.858	6.092	9,903.180	203,839.400
Std.Dev	30.294	0.817	1.406	5,147.021	88,590.630
Min	13	1	1	1,470	14,452
Q1	22	2	5	7,440	146,335
Median	50	3	6	9,428	181,083
Q3	69	3	7	11,500	241,244
Max	151	8	11	70,761	706,250
MAD	37.065	0	1.483	3,006.713	62,407.080
IQR	47	1	2	4,059.500	94,773.500
CV	0.587	0.286	0.231	0.520	0.435
Skewness	0.607	0.400	0.171	3.595	1.589
SE.Skewness	0.046	0.046	0.046	0.046	0.046

Kurtosis	-0.504	1.796	0.054	26.325	3.861
N.Valid	2,867	2,867	2,867	2,867	2,867
Pct.Valid	100	100	100	100	100