Development of a regression model for predicting house sales prices in Ames.

Jaime Rubio Diaz

Table of Contents

1. Introduction

In the dynamic real estate landscape, predicting home sales prices is a crucial task for homeowners, real estate professionals and investors. In this study, a multilinear regression model is constructed to predict house sale prices in Ames. This approach allows multiple influencing factors to be considered, providing an accurate prediction of sales prices. The objective is not only to create a predictive tool, but also to gain valuable insight into the main factors that determine property values in this locality.

2. Background

Numerous studies have been conducted to predict house prices. The House Price Index (HPI) is a statistical index used by investors to make economic decisions, which creates patterns of house purchase prices based on past sales in a specific location (Liberto, 2023). However, due to the advances in technology more complex statistical methods, such as regression models and more complex machine learning methods are now being employed to achieve greater accuracy in predicting property prices. Through these methodologies, it is possible to consider factors such as location or the number of bedrooms, rather than simply replicating sales patterns from previous years (Truong *et al.*,2020).

Regression models aim to stablish relationships between different variables and the target variable (James, 2022). One of the most world known regression model to predict house pricing is the Hedonic Pricing Model, this model assumes that the price of a house is determined by the combined value of all its individual attributes (Nur et al., 2017).

The variables that affect the price of a house were divided into three groups by Nur, A. et al. (2017): physical condition, which includes characteristics such as the size of the house or the number of bedrooms; the concept, an idea proposed by developers like minimalism or green environment; and location, as it establishes the land price.

H1: There will be a positive relationship between the **garage area** and the **sale price**.

Approximately 88% of individuals aged 15 and older are drivers in the United States, with an average of 1.9 vehicles per household (no author, no date). Hence, the size of the garage seems to be an important feature for a house. Private cars are often perceived as having a value that exceeds their cost, and people exhibit irrational behaviours towards them (Moody et al., 2021). Consequently, the importance of keeping cars in a garage in optimal conditions is likely to influence housing prices.

H2: There will be a positive relationship between the **house quality** and the **sale price**.

Improving or achieving a good quality of housing often requires expenses, especially if the property is older, leading to higher housing prices (Goodman, 2004). Moreover, the better the quality of a house the more expensive the materials or the more advanced the technology in it, hence the selling price is expected to increase with the quality of the house.

H3: There will be a negative relationship between the **house age** and the **sale price**.

The age of a home is a critical factor influencing its market value, age serves as the measure for economic depreciation, and it also affects the house's features (Goodman and Thibodeau, 1995).

H4: There will be a relationship between the **neighbourhood** and the **sale price**.

House prices could be affected by the presence of Iowa State University with 34500 students in Ames. Proximity contributes to changes in housing prices, with greater proximity to the university correlating with higher rent prices (Babalola et al., 2013). Therefore, houses close to the university are expected to be more in demand as companies or individuals will want to buy them to rent to students. Hence, greater demand leads to higher prices.

H5: There will be a positive relationship between the **ground living area** and the **sale price**.

This variable, named 'Total_sf' is calculated as the sum of 'floor1_sf' and 'floor2_sf', as will be explained later.

3. Methodology

A dataset named 'Ames' and its data dictionary have been provided. The Data set is sourced from the Ames, Iowa Assessor's Office, and contains 2881 observations and 78 variables. It stores information on residential houses sold in Ames from 2006 to 2010. The data dictionary includes information across all variables of the data set.

The objective of this project is to develop a multilinear regression model to predict house prices given its features. The steps this process have taken are:

1.	Data understanding
2.	Data preparation
3.	Data analysis
4.	Model deployment
5.	Evaluation of model performance

Table 1. Methods

3.1. Data Preparation

To develop a regression model, a new dataset is created. This new dataset will include the hypotheses derived from the wider literature and 13 additional variables.

VARIABLE	MEANING	TYPE
Bedroom	Number of bedrooms	Ordinal
bsmt_area	Basement size (sqft)	Numerical
d_type	Type of dwelling	Categorical
external_qual	Quality of the exterior of the house	Ordinal
fireplace	Number of fireplaces	Numerical
garage_area	Garage size (sqft)	Numerical
garage_cars	Numbers of cars that fit in the garage	Numerical
heat_qual	Quality of the heater	Ordinal
house_age	Age of the house	Numerical
house_condition	Condition of the house	Ordinal
house_quality	Quality of the house	Ordinal
kitchen	Number of kitchens	Numerical
lot_area	Lot size (sqft)	Numerical
neighborhood	Neighbourhood where the house is located	Categorical
prox_1	Proximity to conditions	Categorical
rooms_tot	Number of rooms	Numerical

total_sf	Ground living area (sqft)	Numerical
full_bath	Number of bathrooms	Numerical
sale_price	Sale price of the house	Numerical (TARGET VARIABLE)

Table 2. Reduced Dataset

All the variables of the Data set are directly selected from 'ames', except two of them.

'House_age': Created by subtracting 'year_sold' from 'year_remod'.

'Total_sf': Created by adding 'floor1_sf' and 'floor2_sf', representing the total square footage of both the first and second floors of the house.

Missing values

There are only three missing values in total, so they are not modified. Due to the aim of the project, the model will be built by omitting missing values. Additionally, McKnight, et al (2007) emphasized considering the implications that the missing data may have in the output as a step to manage them. In the present project missing values will not be include in the model so they will not have any implication.

Outliers

Osborne and Overbay (2019) argued in favour of removing outliers, and demonstrated, through the calculation of correlations before and after outlier removal, that the presence of outliers makes correlation less accurate. In this project, outliers are removed because when splitting the data to build the regression model, missing values are omitted.

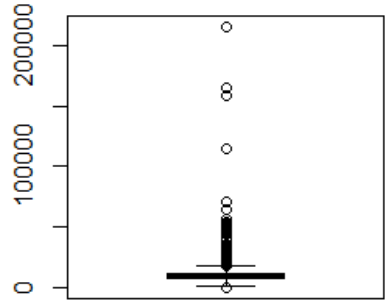

Figure 1. lot_area

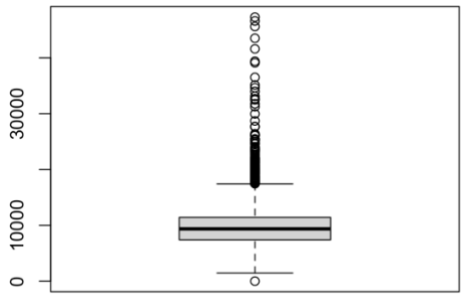

Figure 2. lot_area cleaned

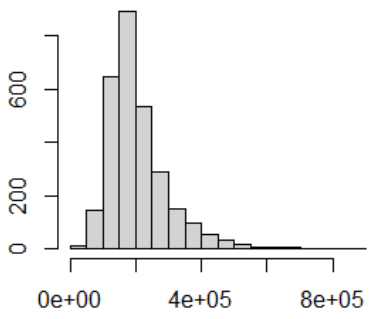

Figure 3. Sale price

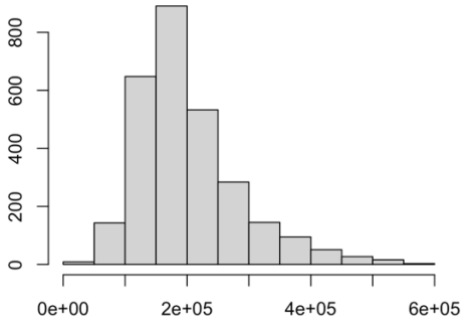

Figure 4. Sale price cleaned

After checking for missing values and removing outliers, categorical and ordinal variables are converted into factors.

3.2. Data analysis

With the 'summary' function in 'R' a table of descriptive statistics of the data can be generated. The table includes the frequency for factors and a descriptive statistical analysis for numerical data. Additionally, the 'ggplot2' library has been utilized to correlate the target variable 'sale_price' with the rest of the variables. This approach helps identify data trends, distributions, and correlations, all of which are presented in the Results section.

3.3. Model Development

Multilinear regression is employed to build the relationship between a dependent variable and multiple independent variables (Grant et al., 2018). A multilinear regression model will be used to stablish the relation between the sale price of houses in Ames and others dependent variables, including the hypothesis presented.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

Equation 1. Multivariable linear regression model equation

A multilinear regression model should contain variables that are high correlated with the target variable ('sale_price') but no high correlated among themselves (multicollinearity). "Correlation is a term to denote the association or relationship between two (or more) quantitative variables" (Thatte, and Gogtay, 2017).

To build the model the data is split into training and test sets with the training set representing 80% of the original data and the test set representing 20%. The model is trained with the training dataset and its accuracy is tested with the test dataset. Different models are built before stablishing the final one. The quality of the regression model is determined by:

R-squared	Coefficient of determination
RMSE	Root mean square error
MAE	Mean absolute error

Table 3. Linear regression measures

4. Results

```
Call:
lm(formula = sale_price ~ total_sf + neighbourhood + house_quality +
garage_area + house_age + d_type + fireplace + heat_qual +
external_qual, data = train)

Residuals:
Min     1Q  Median     3Q    Max
-225570  -13676    147   13463  150810

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)         54886.857  24278.971   2.261 0.023879 *
total_sf              62.948      2.287  27.527  < 2e-16 ***
neighbourhoodBlueste   3263.343  12265.856   0.266 0.790226
neighbourhoodBrDale    4965.293   9612.292   0.517 0.605519
neighbourhoodBrkSide -12346.662   7440.388  -1.659 0.097179 .
neighbourhoodClearCr  10222.216   8391.959   1.218 0.223321
neighbourhoodCollgCr  -4696.448   6501.321  -0.722 0.470137
neighbourhoodCrawfor  18692.854   7138.840   2.618 0.008895 **
neighbourhoodEdwards -14589.244   6919.825  -2.108 0.035118 *
neighbourhoodGilbert  -8310.866   6819.560  -1.219 0.223099
neighbourhoodGreens   -7140.260  13776.668  -0.518 0.604312
neighbourhoodGrnHill 120507.832  20163.297   5.977 2.66e-09 ***
neighbourhoodIDOTRR  -19986.902   7649.692  -2.613 0.009044 **
neighbourhoodMeadowV  -1726.056   9246.561  -0.187 0.851937
neighbourhoodMitchel  -7205.648   7062.700  -1.020 0.307730
neighbourhoodNAmes   -11934.743   6684.503  -1.785 0.074332 .
neighbourhoodNoRidge  27240.382   7581.092   3.593 0.000334 ***
neighbourhoodNPkVill  10045.056   9112.446   1.102 0.270435
neighbourhoodNridgHt  30640.320   6744.341   4.543 5.85e-06 ***
neighbourhoodNWAmes   -8781.607   6977.549  -1.259 0.208329
neighbourhoodOldTown -27349.453   7064.310  -3.871 0.000111 ***
neighbourhoodSawyer  -12877.848   7022.747  -1.834 0.066832 .
neighbourhoodSawyerW -12801.948   6847.039  -1.870 0.061661 .
neighbourhoodSomerst   8408.524   6581.128   1.278 0.201503
neighbourhoodStoneBr  41491.469   7590.003   5.467 5.12e-08 ***
neighbourhoodSWISU   -18313.096   8300.804  -2.206 0.027477 *
neighbourhoodTimber    8327.726   7298.928   1.141 0.254017
neighbourhoodVeenker   5371.968   8811.603   0.610 0.542160
house_quality2       61550.807  24404.778   2.522 0.011738 *
house_quality3       64190.831  23303.405   2.755 0.005927 **
house_quality4       71928.957  22725.817   3.165 0.001572 **
house_quality5       83025.997  22645.203   3.666 0.000252 ***
house_quality6       92839.721  22698.476   4.090 4.47e-05 ***
house_quality7      102552.320  22748.170   4.508 6.89e-06 ***
house_quality8      132998.881  22852.474   5.820 6.77e-09 ***
house_quality9      176268.301  23215.868   7.593 4.65e-14 ***
house_quality10     218610.215  24774.776   8.824  < 2e-16 ***
garage_area            39.566      4.003   9.884  < 2e-16 ***
house_age            -361.475     41.538  -8.702  < 2e-16 ***
d_type30           -16004.189   3606.472  -4.438 9.56e-06 ***
d_type40           -15679.626  13932.482  -1.125 0.260545
d_type45           -12544.809   7582.190  -1.655 0.098170 .
d_type50           -21019.553   2744.200  -7.660 2.80e-14 ***
d_type60           -14545.990   2187.972  -6.648 3.75e-11 ***
d_type70           -27322.357   3694.002  -7.396 2.00e-13 ***
d_type75           -43005.667   8154.818  -5.274 1.47e-07 ***
d_type80            -9775.227   3103.876  -3.149 0.001659 **
d_type85             7012.132   4721.839   1.485 0.137679
d_type90           -30655.336   3699.557  -8.286  < 2e-16 ***
```

```
d_type120          -23005.815   2988.350  -7.698 2.09e-14 ***
d_type160          -51485.296   4332.914 -11.882  < 2e-16 ***
d_type180          -19897.813   9812.182  -2.028 0.042697 *
d_type190          -24649.939   4585.764  -5.375 8.47e-08 ***
fireplace1          6088.993   1493.442   4.077 4.73e-05 ***
fireplace2         22508.548   2699.805   8.337  < 2e-16 ***
heat_qualFa        -11749.088   3715.258  -3.162 0.001587 **
heat_qualGd         -4780.542   1856.253  -2.575 0.010080 *
heat_qualPo        -11635.071  23213.898  -0.501 0.616274
heat_qualTA         -7481.809   1747.931  -4.280 1.95e-05 ***
external_qualFa     -49660.833   8117.699  -6.118 1.13e-09 ***
external_qualGd     -33817.166   4860.748  -6.957 4.59e-12 ***
external_qualTA     -43591.036   5231.373  -8.333  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27280 on 2146 degrees of freedom
Multiple R-squared:  0.887,             Adjusted R-squared:  0.8838
F-statistic: 276.2 on 61 and 2146 DF,  p-value: < 2.2e-16
```

Table 4. Multilinear model output

Model output explained

The R-squared of the model is 0.88, which means that 88% of the variability of the selling price can be explained by the model. The residual standard error is 27280, so given that the average selling price is 199331 $, the model performs well. It can also be seen that the p-value is less than 2.2e-16 indicating that at least one variable in the model has statistical significance with the sale price. Overall, the model is robust, and its accuracy and quality will be detailed later by comparing it with the test dataset.

Variables

The intercept, represented by Bo in the multilinear regression equation, has an estimate value of 54886.8 in the case of this model. It indicates the sale price when the rest of the dependents variables in the model are equal to '0'.

In the model summary, each variable has 'Estimate', 'Std. Error', 't value' and 'p-value'. Due to the word limitation, the focus will be on the estimate value and the p value for the hypothesis presented earlier.

The estimate value indicates how the target variable value is modified when a unit of a dependent variable is also modified (Hayes, 2023).

A p-value is a statistical measure that determinates the significance of hypothesis tests, p-value lower than 0.05 means that the variables are statistically significant (James et al., 2022).

Garage Area

Estimate	Meaning	Relationship	P vale
39.566	Increasing the garage area size by 1 foot corresponds to a $39.6 increment in the sale price.	Positive	< 2e-16

Table 5. Results garage area

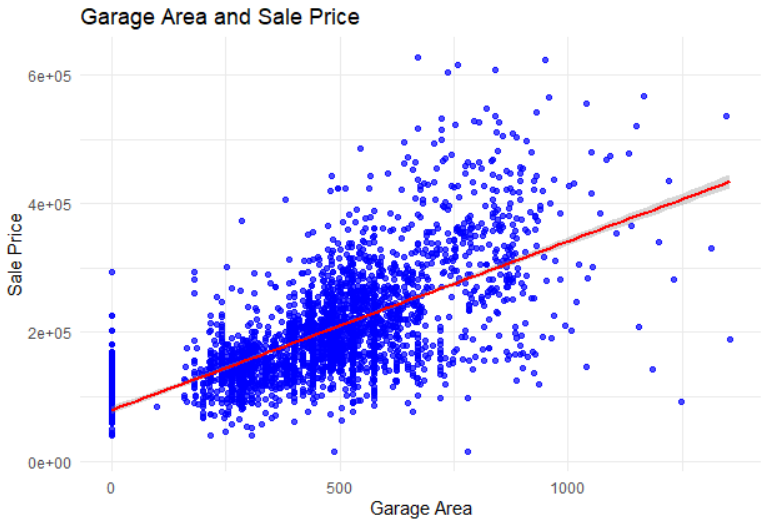

Figure 5. Garage Area and price

House Quality

House quality is an ordinal variable, so in the model output 'R' compares all the levels of the variable with the first level, which will be the reference level. Overall, it is observed that the higher the quality of the house, the higher the price. If a model is built only with sale price and the house quality, a p-value less than 2.2e-16 can be observed, indicating that both variables are statistically significant. The following is the output interpretation for the level 10 of House Quality.

Estimate	Meaning	Relationship	P vale
218610.21	A house with a quality rating of '10' is priced approximately $218,610 higher than houses with a quality rating of '1'	Positive	< 2e-16

Table 6. Results house quality

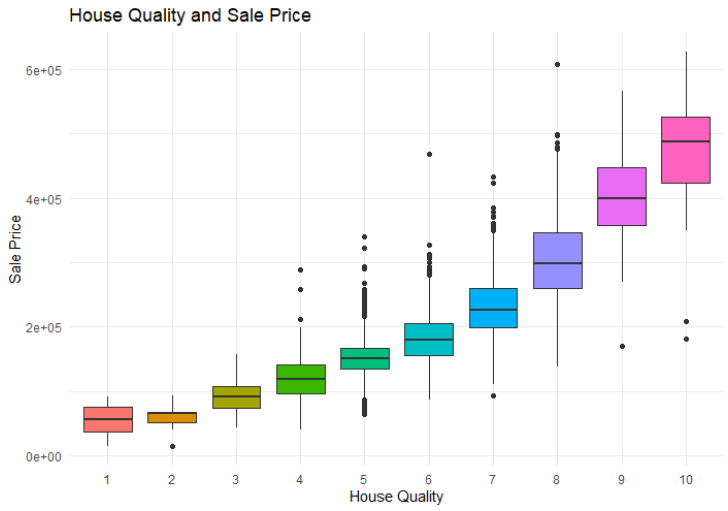

Figure 6. House quality and price

House Age

Estimate	Meaning	Relationship	P vale
-361.5	An increase of 1 year in the age of the house leads to a decrease of $361.5 in the sale price.	Negative	< 2e-16

Table 7. Results House Age

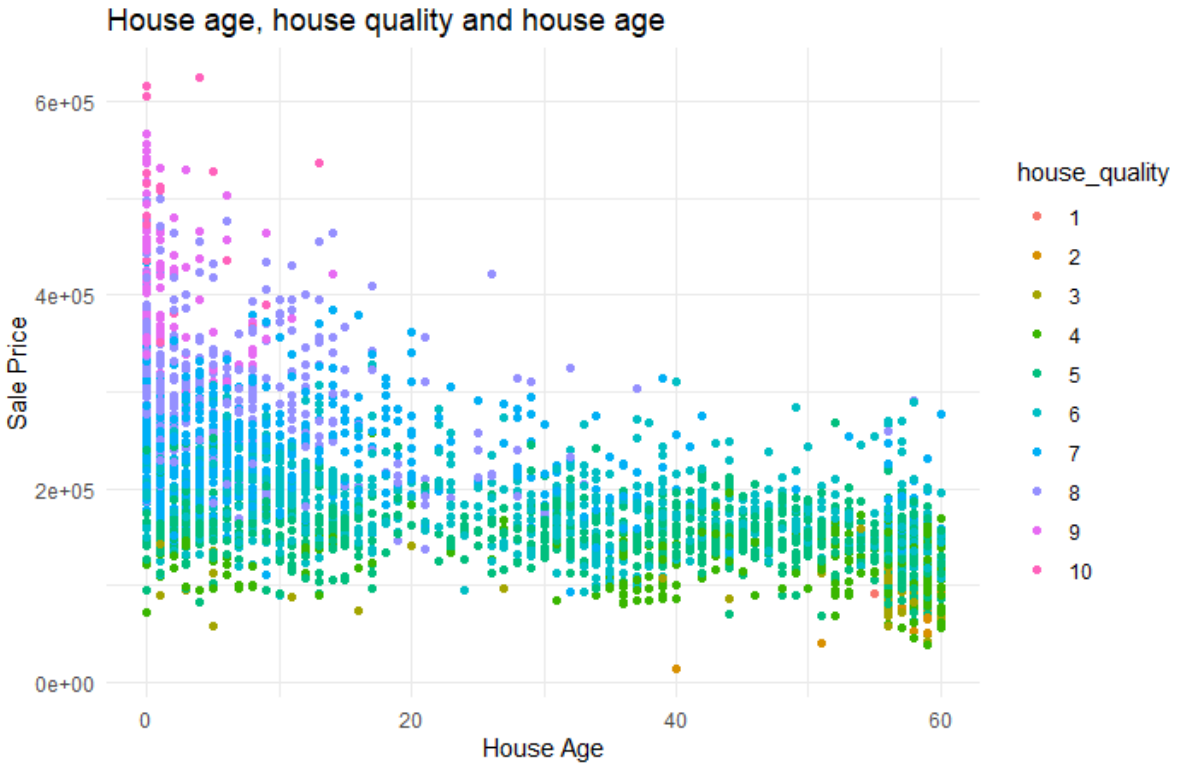

Figure 7. House age and price

Neighbourhood

'R' follows a similar approach with categorical variables as it does with ordinal variables. The Blmngtn neighbourhood is chosen as the reference level, and the rest of the neighbourhoods are compared with it. If a model is built only with sale price and the neighbourhood a p-value less than 2.2e-16 can be observed, indicating that both variables are statistically significant. The following is the output interpretation for StoneBr neighbourhood.

Estimate	Meaning	Relationship	P vale
41491.5	The houses in the StoneBr neighborhood are priced approximately $41,491.5 higher than those in Blmngtn.	Positive	5.12e-08

Table 8. Results neighbourhood

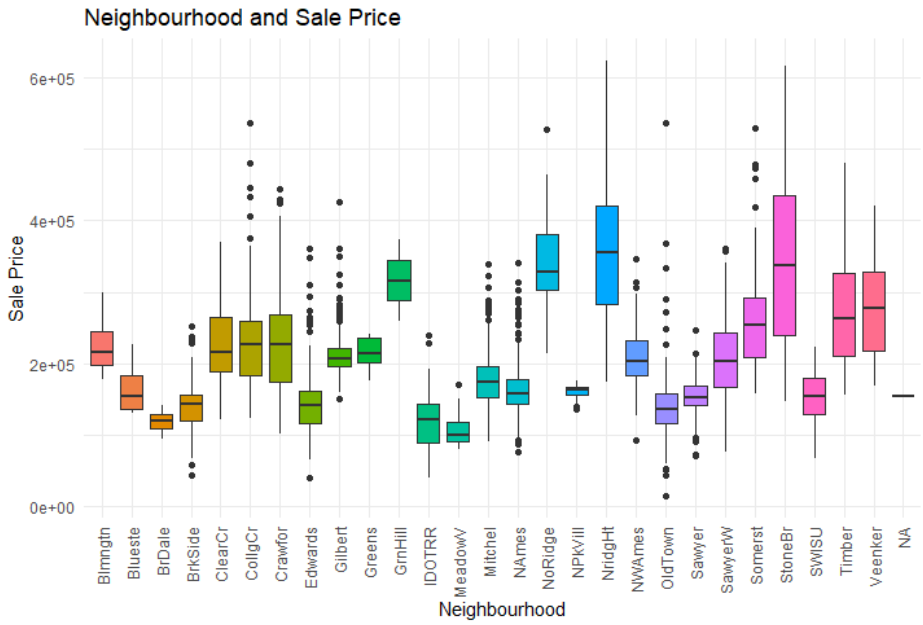

Figure 8. Neighbourhood and price

Total ground living

Estimate	Meaning	Relationship	P vale
62.9	An increase of 1 foot of the total ground living area will cause an increase of $62.9 in the sale price.	Positive	< 2e-16

Table 9. Results total ground living

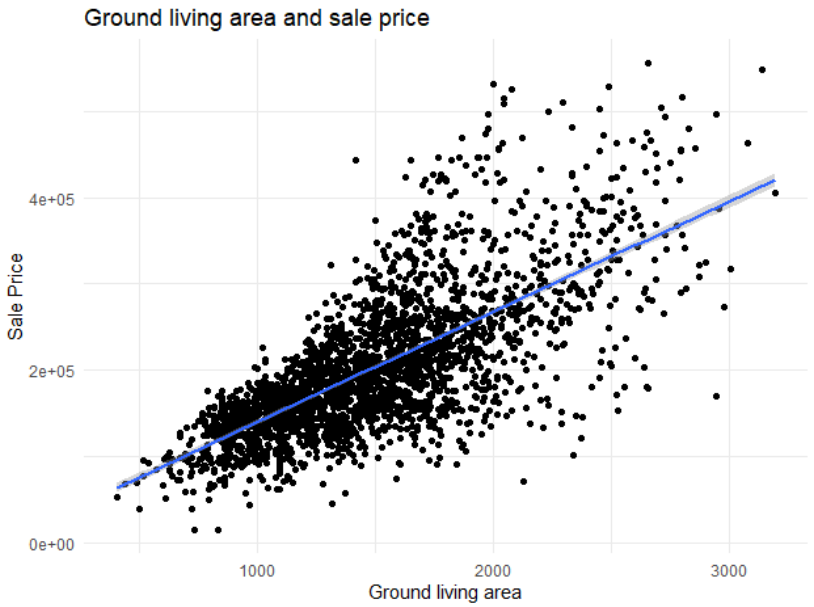

Figure 9. Living area and sale price

Residuals of the model

```
Residuals:
    Min      1Q  Median      3Q     Max
-210550  -13245     379   13295  143984
```

Figure 10. Model residuals

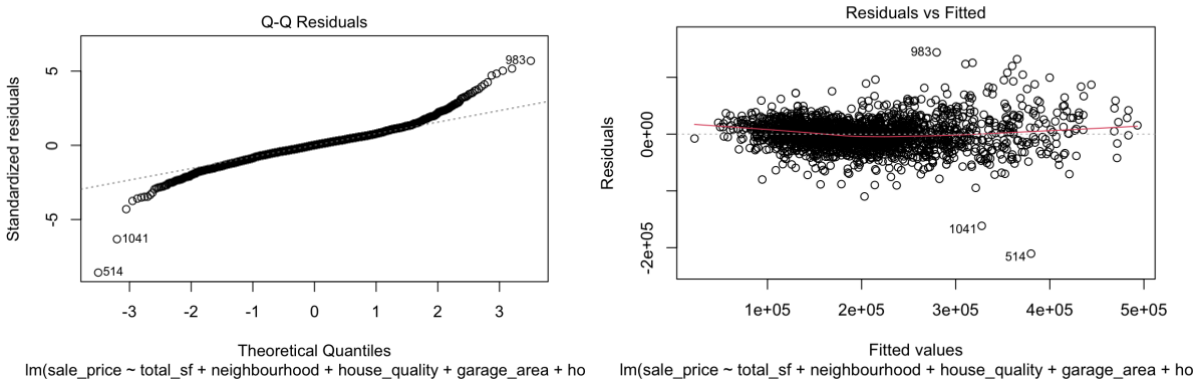

Figure 11. Residual plots

It can be appreciated that the model has a good alignment showing good results for intermediate prices. However, the model's performance is not optimal for both high sale prices and low sale prices.

Model performance with test dataset

The most accurate model is 'model4' (Table 4). When the predictions of the model are compared to the 'test' dataset the following results are achieved:

```
        RMSE      Rsquared           MAE
2.570842e+04 9.028937e-01 1.850486e+04
```

Figure 12. Model measures

- A R-squared of 0.903 indicates that approximately 90.3% of the variability in the sale price can be explained by the independent variables used to build the model, indicating a strong predictive power (Caffo, no date).
- The Mean Absolute Error indicates that the average of the absolute difference between the predicted values and actual values of the 'test' data is $18504.
- The RMSE for the same 'test' data is $25708. This means that, on average, the squared differences between the predicted and actual values, when averaged and then squared, result in $25658.

- RMSE is similar to calculate than MAE, however the squaring of errors in RMSE derives in a higher value, indicating that the model errors have large magnitudes, which may refer to outliers (Willmott and Matsuura, 2005).

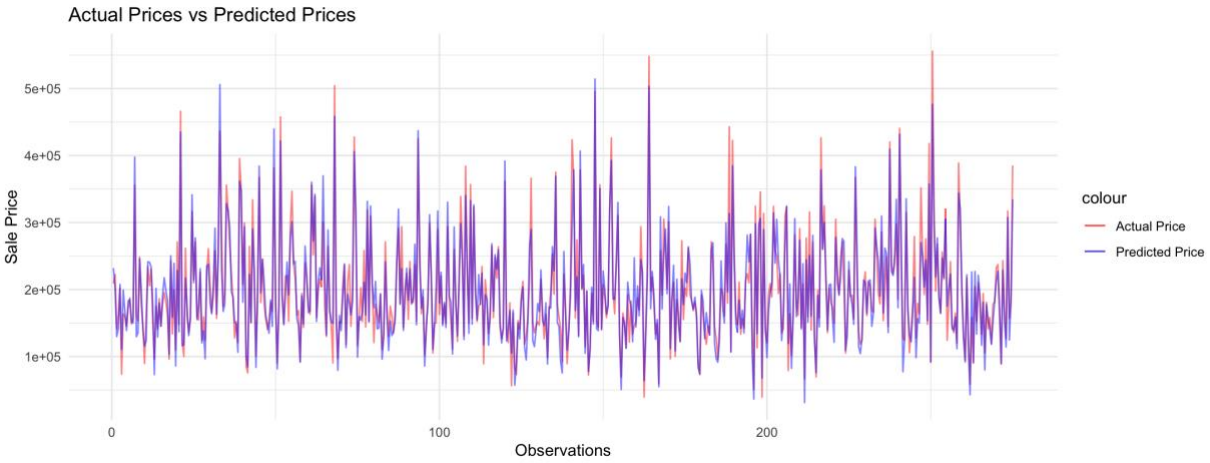

Figure 13. Actual prices vs predicted prices.

In addition to the model interpretation and hypothesis plots, here are some interesting graphs created from the data.

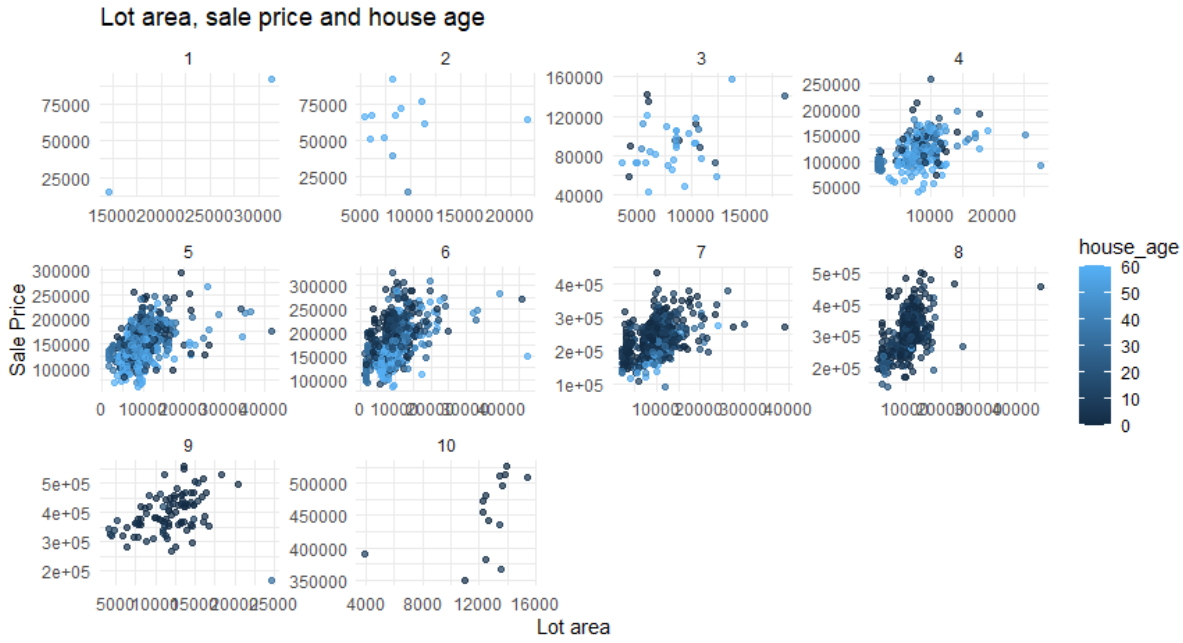

Figure 14. Lot area sale price and house age

Figure 15. Total rooms and sale price

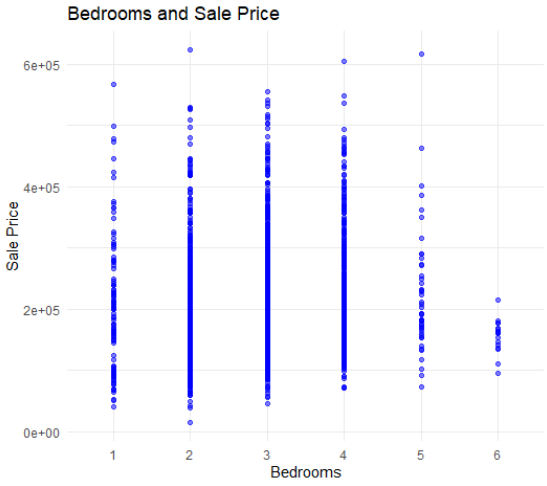

Figure 16. Bedrooms and sale price

5. Discussion

All the hypothesis based on wider literature presented have been confirmed, except the relationship between the neighbourhood and the sale price. Although, the neighbourhood is statistically significant for the model, the hypothesis about the proximity to university is not confirmed. South & West of Iowa State University and Edwards are the closest neighbourhoods to university, and they are determined as the most expensive of Ames (no author, no date). Both neighbourhoods can be observed in Figure 8 as ones of the cheapest neighbourhoods.

It was concluded by Wu (2020) with a multilinear regression approach for house prices in California that houses with more rooms require more space and therefore the price is higher. This is confirmed for 'Ames' data set as it can be seen in Figure 14.

The sale price presents a negative relationship with the number of bedrooms (Figure 15). Finding alginate with (Ozgur et al., 2016) that found with a USA data set a correlation of -0.05 between bedrooms and sale price with a multivariable regression approach as well.

In terms of the accuracy of the model. On the one hand, Madhuri, Anuradha, and Pujitha (2019) realized a comparison between different algorithms to predict house prices with 'kingcounty' dataset, they found that Gradient Boosting Regression was the most accurate, and their multiple linear regression algorithms achieved a r-squared of 0.73. Being the presented model more accurate. On the other hand, Lu, et al., (2017) with the same objective developed an approach by combining regression models, resulting in a 0.112 of MSE with LASSO and gradient boosting combination. Being the presented model less accurate.

6. Conclusions

The present project has focused on constructing a multilinear regression model to predict house sale prices in Ames. The model has demonstrated a high predictive accuracy, as evidenced by the R-squared value of 90%.

It could be beneficial to consider using LASSO regression to improve the model. Lasso regression not only aims to predict the dependent variable, but also perform variable selection by reducing the coefficients of some variables exactly to zero, leading to a reduction in error (Ranstam and Cook, 2018). Additionally, if greater accuracy is desired, others machine learning models should be tried, such as random forest or decision trees.

The analysis and model provide valuable information on the factors that influence property values in Ames. This allows companies to develop marketing campaigns and highlight features that match local market preferences, such as the garage area. Additionally, companies can set house prices based on expected prices and individuals can estimate the correct prices.

REFERENCES

Ames, IA Real Estate Data; Demographic Data (no date) NeighborhoodScout. Available at: https://www.neighborhoodscout.com/ia/ames (Accessed: 27 November 2023).

Babalola, S.J., Umar, A.I., y Sulaiman, L.A. (2013). "An economic analysis of determinants of house rents in the university environment." European Scientific Journal, 9(19).

Caffo, B. (no date) Regression Models for Data Science in R, RPubs. Available at: https://rpubs.com/iabrady/residual-analysis (Accessed: 27 November 2023).

Goodman, A.C. and Thibodeau, T.G. (1995) 'Age-Related Heteroskedasticity in Hedonic House Price Equations', Journal of Housing Research, 6(1), pp. 25–42. doi:10.1111/1540-6229.00742. Available at: https://www.jstor.org/stable/24825889?seq=18 Accessed: (24/11/2023)

Goodman, J. (2004) 'Determinants of operating costs of multifamily rental housing', Journal of Housing Economics, 13(3), pp. 226–244. doi:10.1016/j.jhe.2004.07.003. Available at: https://www.sciencedirect.com/science/article/pii/S1051137704000348

Grant, S.W., Hickey, G.L. and Head, S.J. (2018) 'Statistical primer: Multivariable regression considerations and pitfalls', European Journal of Cardio-Thoracic Surgery, 55(2), pp. 179–185. doi:10.1093/ejcts/ezy403. Available at: https://academic.oup.com/ejcts/article/55/2/179/5265263 Accessed: 25/11/2023

Hayes, A. (2023) Multiple linear regression (MLR) definition, formula, and example, Investopedia. Available at: https://www.investopedia.com/terms/m/mlr.asp (Accessed: 04 December 2023).

Household, individual, and vehicle characteristics (no date a) Bureau of Transportation Statistics. Available at: https://www.bts.gov/archive/publications/highlights_of_the_2001_national_household_travel_survey/section_01 (Accessed: 24 November 2023).

James, G. et al. (2022) An introduction to statistical learning: With applications in R. 2nd edn. Boston: Springer.

Liberto, D. (2023) Understanding the House price index (HPI) and how it is used, Investopedia. Available at: https://www.investopedia.com/terms/h/house-price-index-hpi.asp#:~:text=How%20the%20House%20Price%20Index,big%20implications%20for%20the%20economy . (Accessed: 28 October 2023).

Lu, S. et al. (2017) 'A hybrid regression technique for house prices prediction', 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) [Preprint]. doi:10.1109/ieem.2017.8289904. Available at: https://ieeexplore.ieee.org/abstract/document/8289904 Accessed: 26/11/2023

Madhuri, CH.R., Anuradha, G., and Pujitha, M.V. (2019) 'House price prediction using regression techniques: A comparative study', 2019 International Conference on Smart Structures and Systems (ICSSS) [Preprint]. doi:10.1109/icsss.2019.8882834. Available at: https://ieeexplore.ieee.org/document/8882834 Accessed: 26/11/2023

McKnight, P.E. et al. (2007) Missing data: A gentle introduction [electronic]. New York: Guilford Press. pag. 14 (Accessed: 24 Nov. 2023).

Moody, J., Farr, E., Papagelis, M. and Keith, D.R. (2021). The value of car ownership and use in the United States. *Nature Sustainability*, 4. doi:https://doi.org/10.1038/s41893-021-00731-5. Accessed: 25/11/2023.

Nur, A. et al. (2017) 'Modelling house price prediction using regression analysis and particle swarm optimization case study : Malang, East Java, Indonesia', International Journal of Advanced Computer Science and Applications, 8(10). doi:10.14569/ijacsa.2017.081042. Accessed 28/10/2023.

Osborne, J.W. and Overbay, A. (2019) 'The power of outliers (and why researchers should ALWAYS check for them)', Practical Assessment, Research, and Evaluation, 9(6). doi:https://doi.org/10.7275/qf69-7k43. Available at: https://scholarworks.umass.edu/pare/vol9/iss1/6 (Accessed: 17 Nov. 2023).

Ozgur, C., Hughes, Z., Rogers, G., and Parveen, S. (2016). "Multiple Linear Regression Applications in Real Estate Pricing." Business Faculty Publications, College of Business, 10. Valparaiso University. Available at: https://core.ac.uk/download/pdf/303864121.pdf . Accessed: 27/11/2023

Thatte, U. and Gogtay, N. (2017) 'Principles of Correlation Analysis', Journal of The Association of Physicians of India, 65, pp. 79–81. Available at: https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&q=Principles+of+Correlation+Analysis&btnG= Accessed: 25/11/2023

Truong, Q. et al. (2020) 'Housing price prediction via Improved Machine Learning Techniques', Procedia Computer Science, 174, pp. 433–442. doi:10.1016/j.procs.2020.06.111. Accessed: 28/10/2023.

Willmott, C. and Matsuura, K. (2005) 'Advantages of the mean absolute error (mae) over the root mean square error (RMSE) in assessing average model performance', Climate Research, 30, pp. 79–82. doi:10.3354/cr030079. Available at: https://int-res.com/abstracts/cr/v30/n1/p79-82/. Accessed: 27 November 2023

Wu, Z. (2020) 'Prediction of California House Price Based on Multiple Linear Regression', Academic Journal of Engineering and Technology Science, 3(7), pp. 11–15. doi:10.25236/AJETS.2020.030702. Available at: https://francis-press.com/papers/2868 Accessed: 27/11/2023

Ranstam, J. and Cook, J.A. (2018) 'Lasso regression', British Journal of Surgery, 105(10), pp. 1348–1348. doi:10.1002/bjs.10895.

APPENDIX

```
setwd('Q:/')
.libPaths(c(.libPaths(),"C:/Rpackages"))

library(ggplot2)
library(readxl)
library(tidyverse)
library(caret)
library(corrplot)
library(lmtest)
library(car)

ames <- read_excel('ames.xlsx')
colnames(ames)
#house age and gr living sf are calculated
ames$house_age <- ames$year_sold - ames$year_remod
ames$total_sf <- ames$floor1_sf + ames$floor2_sf
#create a new table with the selected variables
data <- ames %>% select(d_type, lot_area, `exter l_qual`,
                        neighbourhood, prox_1, garage_cars,
                        house_quality, house_condition,
                        bsmt_area,heat_qual, total_sf,
                        bedroom, kitchen, rooms_tot, full_bath,
                        fireplace, garage_area,house_age,
                        sale_price)
summary (data)

##data cleaning -----

#Convert the categorical variables into factors
```

```
data <- data %>% mutate_if(is.character,as.factor)
#LOT_AREA
hist(data$lot_area)
table(data$lot_area)
boxplot(data$lot_area)
data %>% count(lot_area>50000)
data$lot_area[data$lot_area > 50000] <- NA
data$lot_area[data$lot_area < 20] <- NA
summary(data$lot_area)
#BSMT_AREA
hist(data$bsmt_area)
boxplot(data$bsmt_area)
data %>% count(bsmt_area>2470)
data$bsmt_area[data$bsmt_area > 2470] <- NA
summary(data$bsmt_area)
table(data$bsmt_area)
#total_sf
boxplot(data$total_sf)
data %>% count(total_sf>3200)
data$total_sf[data$total_sf > 3200] <- NA
#garage area
boxplot(data$garage_area)
data %>% count(garage_area>1100)
data$garage_area[data$garage_area > 1100] <- NA
#house age
boxplot(data$house_age)
data$house_age[data$house_age < 0] <- NA
summary(data$house_age)
#sale_price
boxplot(data$sale_price)
data %>% count(sale_price>601000)
```

```
data$sale_price[data$sale_price > 601000] <- NA
summary(data$sale_price)
hist(data$sale_price)
#bedrooms
boxplot(data$bedroom)
table(data$bedroom)
data$bedroom[data$bedroom > 6] <- NA
data$bedroom[data$bedroom < 1] <- NA
#kitchen
boxplot(data$kitchen)
table(data$kitchen)
data$kitchen[data$kitchen > 2] <- NA
data$kitchen[data$kitchen < 1] <- NA
#full_bath
boxplot(data$full_bath)
table(data$full_bath)
data$full_bath[data$full_bath < 1] <- NA
data$full_bath[data$full_bath > 3] <- NA
#fireplace
boxplot(data$fireplace)
table(data$fireplace)
data$fireplace[data$fireplace > 3] <- NA
#neighborhood
table(data$neighbourhood)
data$neighbourhood <- droplevels(data$neighbourhood, exclude = 'Landmrk')
#garage_cars
boxplot(data$garage_cars)
table(data$garage_cars)
data$garage_cars[data$garage_cars > 3] <- NA
#extenal quality
data$'exter l_qual' <- droplevels(data$'exter l_qual', exclude = "Good")
```

```
data %>%
  rename(external_qual  = 'exter l_qual') -> data
#house quality
table(data$house_quality)
data$house_quality[data$house_quality == '11'] <- NA
#house_condition
table(data$house_condition)
#rooms_totals
table(data$rooms_tot)
data$rooms_tot[data$rooms_tot == '0'] <- NA
data$rooms_tot[data$rooms_tot > 11] <- NA
#fireplace
table(data$fireplace)
data$fireplace[data$fireplace == '3'] <- NA
#convert the variables which levels affect the price and the ordinal variables in factors
data$d_type <- as.factor(data$d_type)
data$house_quality <- as.factor(data$house_quality)
data$house_condition <- as.factor(data$house_condition)
data$fireplace <- as.factor(data$fireplace)
data$garage_cars <- as.factor(data$garage_cars)
data$rooms_tot <- as.factor(data$rooms_tot)
summary(data)
data <- na.omit(data)

##graphs ------

#garage area
ggplot(data, aes(x = garage_area, y = sale_price)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_smooth(method = 'lm', color = "red") +
  labs(title = "Garage Area and Sale Price",
```

```
    x = "Garage Area",
    y = "Sale Price") +
  theme_minimal()

#house_quality
ggplot(data, aes(x = house_quality, y = sale_price, fill = house_quality)) +
  geom_boxplot() +
  labs(title = "House Quality and Sale Price",
    x = "House Quality",
    y = "Sale Price") +
  theme_minimal() +
  guides(fill = FALSE)

#house age, sale price and house condition
ggplot(data, aes(x = house_age, y = sale_price, color = house_quality)) +
  geom_point() +
  labs(title = "House age, house quality and house age",
    x = "House Age",
    y = "Sale Price") +
  theme_minimal()

#Neighborhood and sale price
ggplot(data, aes(x = neighbourhood, y = sale_price, fill = neighbourhood)) +
  geom_boxplot() +
  labs(title = "Neighbourhood and Sale Price",
    x = "Neighbourhood",
    y = "Sale Price") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  guides(fill = FALSE)
```

```
#total_sf and sale price
ggplot(data, aes(x = total_sf, y = sale_price)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Ground living area and sale price",
       x = "Ground living area",
       y = "Sale Price") +
  theme_minimal()

#lot area, sale price and house age
ggplot(data, aes(x = lot_area, y = sale_price, color = house_age)) +
  geom_point(alpha = 0.7) +
  facet_wrap(~ house_quality, scales = "free") +
  labs(title = "Lot area, sale price and house age",
       x = "Lot area",
       y = "Sale Price") +
  theme_minimal()

#Total rooms and sale price
ggplot(data, aes(x = as.factor(rooms_tot), y = sale_price)) +
  geom_point(color = "blue", alpha = 0.5) +
  labs(title = "Total rooms and Sale Price",
       x = "Total rooms",
       y = "Sale Price") +
  theme_minimal()

#bedroom and sale price
ggplot(data, aes(x = bedroom, y = sale_price)) +
  geom_point(color = "blue", alpha = 0.5) +
  labs(title = "Bedrooms and Sale Price",
       x = "Bedrooms",
```

```
      y = "Sale Price") +
  theme_minimal()
##Correlation---------
#Pearson correlation numeric variables
data_numeric <- data[sapply(data, is.numeric)]
correlation_matrix <- cor(data_numeric, use = "complete.obs")
corrplot(correlation_matrix, method = "circle")

#numeric hypothesis correlation
cor.test(data$total_sf, data$sale_price)
cor.test(data$house_age, data$sale_price)
cor.test(data$garage_area, data$sale_price)
pairs(data[, c("sale_price", "house_age", "garage_area", "total_sf")])

#categorical hypothesis
anova_result <- aov(sale_price ~ house_quality + neighbourhood, data = data)
cor.test(data$sale_price, as.numeric(data$house_quality), method = "spearman", exact = FALSE)
cor.test(data$sale_price, as.numeric(data$neighbourhood), method = "spearman", exact = FALSE)

###Model --------

#split the data 80 TRAIN Y 20 TEST
set.seed(40425150)
index <- createDataPartition(data$sale_price, p = 0.8, list = FALSE)
train <- data[index,]
test <- data[-index, ]

#build the models
model0 <- lm(sale_price ~ neighbourhood, data = train)
model01 <- lm(sale_price ~ house_quality, data= train)
```

```
model1 <- lm(sale_price ~ total_sf + neighbourhood + house_quality + garage_area + house_age, data = train)


summary(model1)

test$predictions1 <- predict(model1, test)

print(test$predictions1)

postResample(test$predictions1, test$sale_price)

model2 <- lm(sale_price ~ total_sf + neighbourhood + house_quality + garage_area + house_age + d_type + bedroom + bsmt_area, data = train)

summary(model2)

test$predictions2 <- predict(model2, test)

print(test$predictions2)

postResample(test$predictions2, test$sale_price)

model3 <- lm(sale_price ~ total_sf + neighbourhood + house_quality + garage_area + house_age + kitchen + bsmt_area + fireplace + external_qual, data = train)

summary(model3)

test$predictions3 <- predict(model3, test)

print(test$predictions3)

postResample(test$predictions3, test$sale_price)

model4 <- lm(sale_price ~ total_sf  + neighbourhood + house_quality + garage_area + house_age + d_type + bsmt_area + bedroom + fireplace + heat_qual + external_qual, data=train)

summary(model4)

test$predictions4 <- predict(model4, test)

print(test$predictions4)

postResample(test$predictions4, test$sale_price)

plot(model4)


#line chart actual values vs predicted values with half of the test dataset for increased clarity

ggplot(test, aes(x = seq_along(sale_price)/2)) +

  geom_line(aes(y = sale_price, color = "Actual Price"), alpha = 0.5, size = 0.5) +

  geom_line(aes(y = predictions4, color = "Predicted Price"), alpha = 0.5, size = 0.5) +

  labs(title = "Actual Prices vs Predicted Prices",
```

```
    x = "Observations",
    y = "Sale Price") +
  scale_color_manual(values = c("Actual Price" = "red", "Predicted Price" = "blue")) +
  theme_minimal()
#residuals examination. 105 standardised residuals
diag <- train
diag$residuals <- resid(model4)
diag$standar_residuals <- rstandard(model4)
diag$large_residuals <- diag$standar_residuals > 2 | diag$standar_residuals < -2
sum(diag$large_residuals)
diag_investigate <- diag[diag$large_residuals, c("sale_price", "total_sf", "neighbourhood", "house_quality", "garage_area", "house_age", "d_type", "bsmt_area", "bedroom", "fireplace", "heat_qual", "external_qual")]
diag_investigate
hist(diag$residuals)
#assumption independent erros
dwtest(model4)
#assumption of no multicollinearity
vif(model4)
mean(vif(model4))
```

APPENDIX 2

```
    d_type        lot_area      external_qual neighbourhood      prox_1     garage_cars house_quality house_condition
20     :1020   Min.   : 1470   Ex:  82       NAmes  : 424   Norm   :2386   0: 145      5      :788   5      :1550
60     : 541   1st Qu.: 7400   Fa:  31       CollgCr: 259   Feedr  : 144   1: 753      6      :701   6      : 511
50     : 276   Median : 9350   Gd: 937       OldTown: 228   Artery :  88   2:1535      7      :570   7      : 369
120    : 183   Mean   : 9640   TA:1708       Edwards: 182   RRAn   :  47   3: 325      8      :326   8      : 138
30     : 133   3rd Qu.:11355                 Somerst: 173   PosN   :  31               4      :215   4      :  95
70     : 124   Max.   :47280                 Gilbert: 163   RRAe   :  28               9      : 92   3      :  46
(Other): 481                                  (Other):1329   (Other):  34               (Other): 66   (Other):  49
  bsmt_area      heat_qual    total_sf       bedroom         kitchen        rooms_tot     full_bath     fireplace
Min.   :   0.0   Ex:1405   Min.   : 407   Min.   :1.000   Min.   :1.000   6      :821   Min.   :1.000   0:1356
1st Qu.: 784.0   Fa:  88   1st Qu.:1114   1st Qu.:2.000   1st Qu.:1.000   7      :621   1st Qu.:1.000   1:1223
Median : 981.5   Gd: 450   Median :1428   Median :3.000   Median :1.000   5      :567   Median :2.000   2: 179
Mean   :1031.7   Po:   3   Mean   :1465   Mean   :2.847   Mean   :1.038   8      :320   Mean   :1.556
3rd Qu.:1275.5   TA: 812   3rd Qu.:1720   3rd Qu.:3.000   3rd Qu.:1.000   4      :187   3rd Qu.:2.000
Max.   :2461.0             Max.   :3194   Max.   :6.000   Max.   :2.000   9      :130   Max.   :3.000
                                                                           (Other):112

 garage_area       house_age       sale_price
Min.   :   0.0   Min.   : 0.00   Min.   : 14452
1st Qu.: 315.0   1st Qu.: 4.00   1st Qu.:145770
Median : 474.0   Median :15.00   Median :180684
Mean   : 463.4   Mean   :23.64   Mean   :199331
3rd Qu.: 576.0   3rd Qu.:43.00   3rd Qu.:237300
Max.   :1092.0   Max.   :60.00   Max.   :555960
```

Figure 17. Data summary

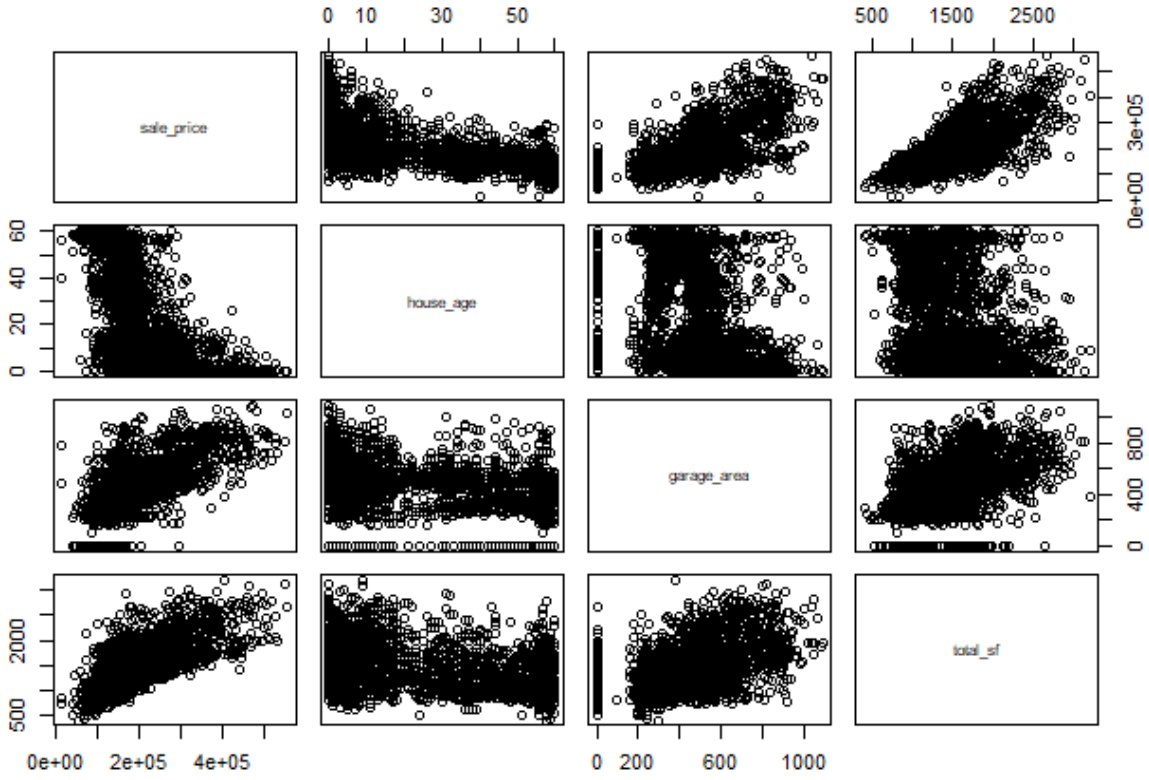

Figure 18 Numeric hypothesis correlations.

```
        Durbin-Watson test

data:  model4
DW = 2.0295, p-value = 0.7556
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 19. Durbin-Watson test. Model 4

	GVIF	Df	GVIF^(1/(2*Df))
total_sf	3.159059	1	1.777374
neighbourhood	147.613931	26	1.100814
house_quality	20.191644	9	1.181706
garage_area	1.922261	1	1.386456
house_age	2.219690	1	1.489862
d_type	51.228930	14	1.150944
fireplace	1.739001	2	1.148352
heat_qual	2.943575	4	1.144483
external_qual	7.145308	3	1.387832

Figure 20. Checking for multicollinearity. Model 4