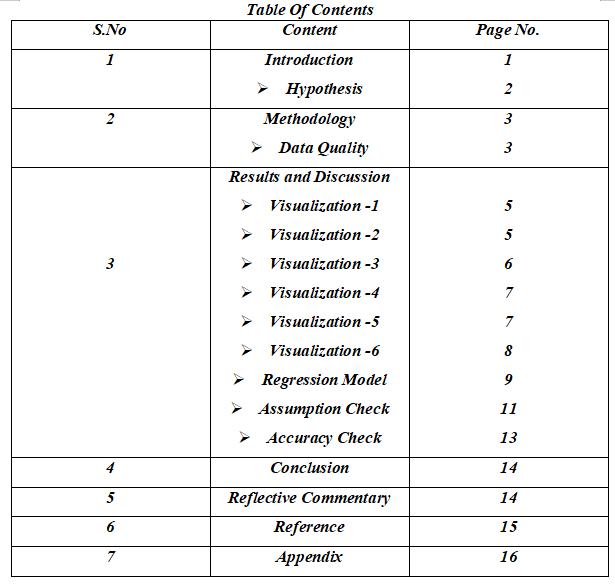
**Unlocking Real Estate Insights: A High-Precision Regression Model for Predicting House Prices**

**Suriya Subbiah Perumal**

****

**Introduction**

In the dynamic and highly competitive realm of real estate, agencies are increasingly leveraging advanced technologies to stay ahead. A critical aspect of this technological adoption is the use of machine learning to enhance accuracy in predicting house prices.This report presents a comprehensive journey into the development of a machine learning model tailored for predicting house prices,utilizing the Ames house price dataset as a foundation.

The report delves into developing a machine learning model for predicting house prices, based on the Ames dataset.We started by selecting key variables believed to influence sale prices. Statistical analyses were performed to explore the variables' impact on pricing,enhanced by visual tools for clearer understanding. The development of predictive models was carried out in R studio,with a focus on achieving high accuracy. Rigorous model testing aimed to balance statistical integrity with precision in price prediction.

Machine learning techniques are increasingly being utilized in the realm of house price prediction, as evidenced by research studies,Truong et al.(2020) explored improved machine learning methods for this purpose, demonstrating advancements in predictive accuracy. Similarly,the use of Random Forest machine learning techniques has been investigated for predicting house prices,focusing on price variance and classification approaches(Zhan,2023). Zaki et,al.(2022) combined hedonic pricing models with machine learning techniques to enhance house price predictions, reflecting a multidisciplinary approach to the problem.Another study presented a hybrid machine learning framework that integrates Hybrid Bayesian Optimization with various techniques like Stacking and Bagging for forecasting house prices, showcasing innovative model combinations for better predictions(Truong,2020). Furthermore, the application of machine learning algorithms in house price prediction has been acknowledged as part of their broader use in real-life applications and research, underscoring the technology's versatility and growing significance(Mora-Garcia,2022).

**Hypothesis**

**H-1:** The hypothesis posits that the sale price of a house will positively correlate with its quality, as quality is a significant consideration for buyers in the property market. This expectation is grounded in the understanding that higher quality houses tend to attract higher prices. According to Jhon Cubbin(1974) buyer tend to judge the quality of the house by the price.

**H-2:**  The assumption is that there is a positive link between the overall constructed area and the property's sales value, which is straightforward to understand: as the number of floors in a building rises, so does the total built-up area,consequently boosting the house's price. "Housing Price Prediction Based on Multiple Linear Regression," Zhang(2021) examines essential determinants of housing prices, including the built-up area.

**H-3:** The construction year of a house is often considered a significant determinant of its selling price, suggesting a positive correlation. This is supported by a study that employed a estimation model to assess the impact of new construction on house sales prices, focusing on houses 1–3 years old in Baton Rouge, Louisiana(Herbert, 2014).

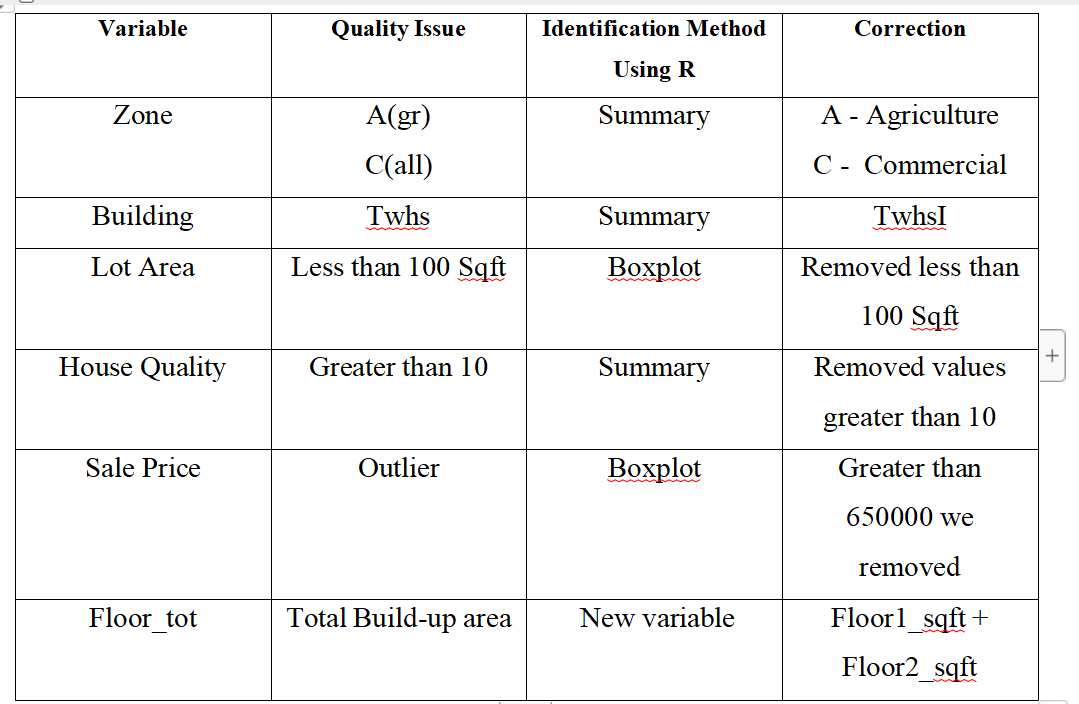
**H-4:**It is hypothesized that the number of full bathrooms in a house may positively correlate with its sale price, a concept supported by general real estate literature which indicates that an increase in amenities, like bathrooms, tends to elevate a house's market value.This is substantiated by a study in The Journal of Real Estate Finance and Economics, which includes the number of bathrooms as a key variable affecting house prices(Jauregui,et,al., 2017).

**H-5:** There will be positive correlation between the year of remodeling and sale price. The positive correlation between a house's year of remodeling and its sale price is further evidenced by a study by Bogin and Doerner(2019), which provides a wide-scale analysis of property renovation bias in repeat-sales house price indices across various U.S. geographies. They found that property improvements often lead to a positive quality drift,in the central districts of large cities. However,this effect tends to diminish outside of downtown areas and is negligible in smaller cities.

**H-6:** The sale price of a house is likely positively correlated with the number of rooms it contains, as more rooms typically enhance a property's functionality and increase its total built-up area, thereby boosting its market value. This relationship is substantiated by research conducted by Zhang(2021) in a study,where a clear upward trend in sale price was observed with an increase in the number of rooms.

**Methodology**

The Ames dataset, which includes approximately 79 variables related to the sales prices of 2,880 houses, was initially imported into R for data quality analysis. Maintaining a high standard of data quality is crucial for achieving optimal results, particularly in production planning and control, as noted by Lindstrom et al.(2023). To assess and rectify data quality issues, various techniques were employed, as detailed in the subsequent table. The identification and resolution of these issues were facilitated using the ggplot2 and tidyverse packages in R.

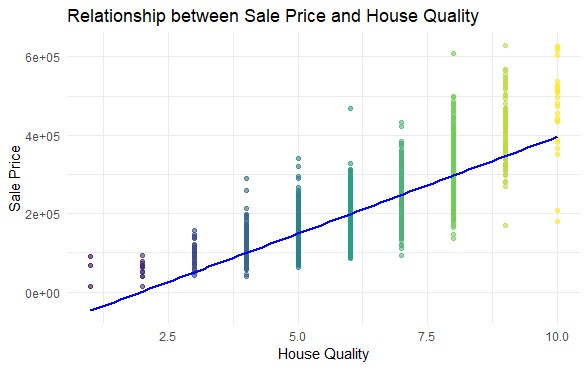


After resolving data quality concerns, the amended dataset was used to produce visualizations with the ggplot2 package in R,focusing on variables that showed a substantial connection with the sale price.Various graphical representations were explored to effectively illustrate the relationship of these variables with the sale price. Additionally,the correlations between these variables and the sale price were calculated using the correlation function in R.

Using the dataset,a regression model was built in R,setting the seed with the student number for reproducibility and splitting the data into training and testing sets. We started with variables that showed a strong positive correlation with sale prices. Through iterative refinement adding and removing factors we enhanced the model's accuracy.The final model, boasting high precision,was validated against the test set and checked against standard modeling assumptions to ensure robustness.

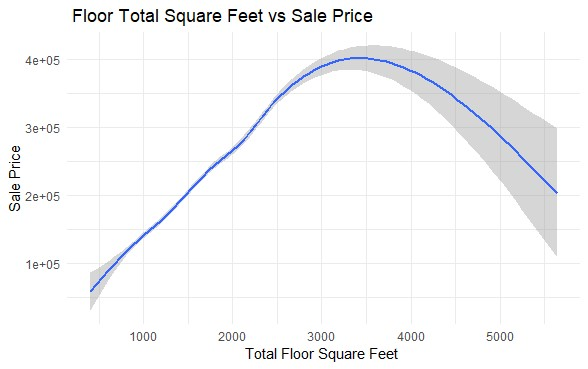
**Results and Discussion**

**Visulaization-1**



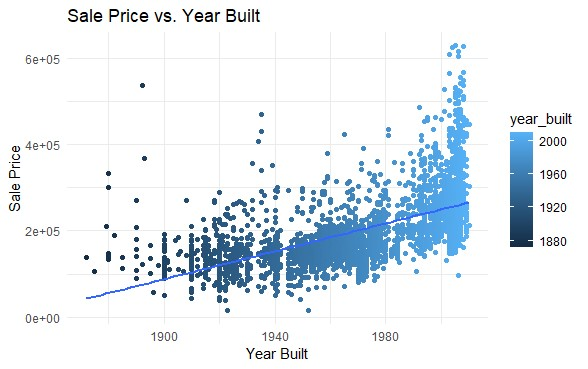
The scatter plot reveals a strong positive correlation between house quality and sale prices, with a Spearman coefficient of 0.8067002, indicating that better quality often results in higher prices. Most data points cluster around the median quality range, with price variability increasing with quality.Outliers, especially in the high-quality segment, suggest some houses command premiums beyond the norm. Hu et al.(2020) also note that quality influences homeowner satisfaction and sale prices.

**Visulaization-2**



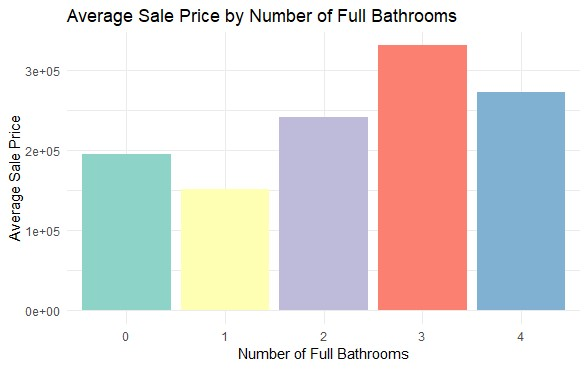
The graph depicts a parabolic trend between building size and sale prices, with prices climbing with size up to around 3000 square feet, then declining for larger homes, suggesting a market preference for mid-sized properties.The peak indicates an optimal size for maximum value,with a denser concentration of sales in smaller homes. A Pearson correlation coefficient of 0.5785147 shows a moderate positive linear relationship between house quality and price, supporting that higher quality typically leads to higher prices. In line with this, Feng(2021) notes that the total price of a larger housing unit will increase at an increasing rate as its size increases, implying a non-linear growth in value.

**Visulaization-3**



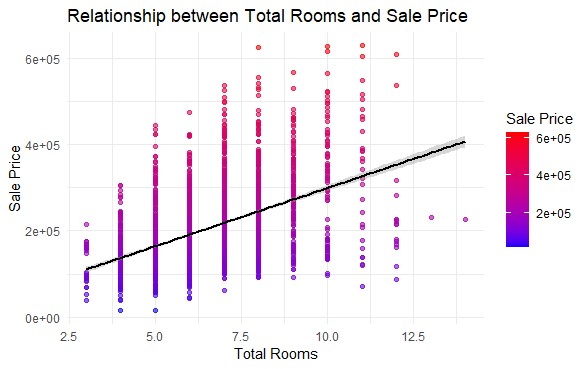
The scatter plot shows a moderately strong positive correlation of 0.679 between the construction year of properties and their sale prices,with newer properties tending to sell for more. The density of data points increases for recent years, indicating a trend of rising sale prices over time. An upward trend line further suggests that modern properties are valued higher,potentially reflecting advancements in building and design. Outliers in later years highlight properties that exceed expected prices, underscoring a market that values modernity. The color gradient may represent different periods or densities, emphasizing changes through time.

**Visulaization-4**



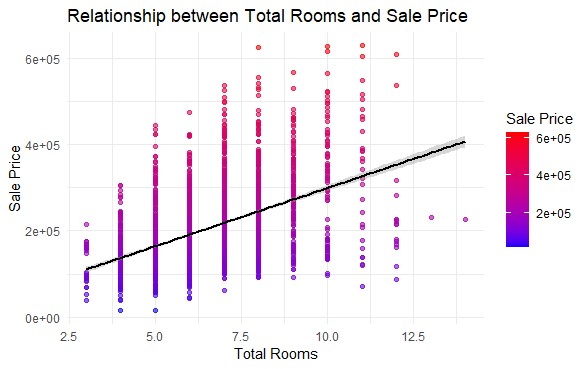
The bar chart demonstrates that as the number of full bathrooms in a property increases from 0 to 3, the average sale price also rises, confirming a moderate positive correlation with a Spearman coefficient of 0.6301274. This suggests additional bathrooms typically add value. However, the average price plateaus and even slightly declines for properties with 4 full bathrooms, indicating a limit to the value added by more bathrooms,possibly due to market saturation or other valuation factors.

**Visulaization-5**



The scatter plot displays the relationship between the Sale Price of houses and the Year of Remodeling, indicating a positive correlation, as shown by a Spearman correlation coefficient of 0.5995173. This suggests a moderate to strong increasing trend where houses remodeled in more recent years tend to have higher sale prices. The data points become denser and more spread in terms of sale price as time progresses, especially after 2000, with vertical 'stripes' in certain years suggesting clustering around specific sale prices. The variability in sale prices also appears to increase with time, and the presence of outliers, particularly in later years,suggests that there are some sales with prices significantly above the average.

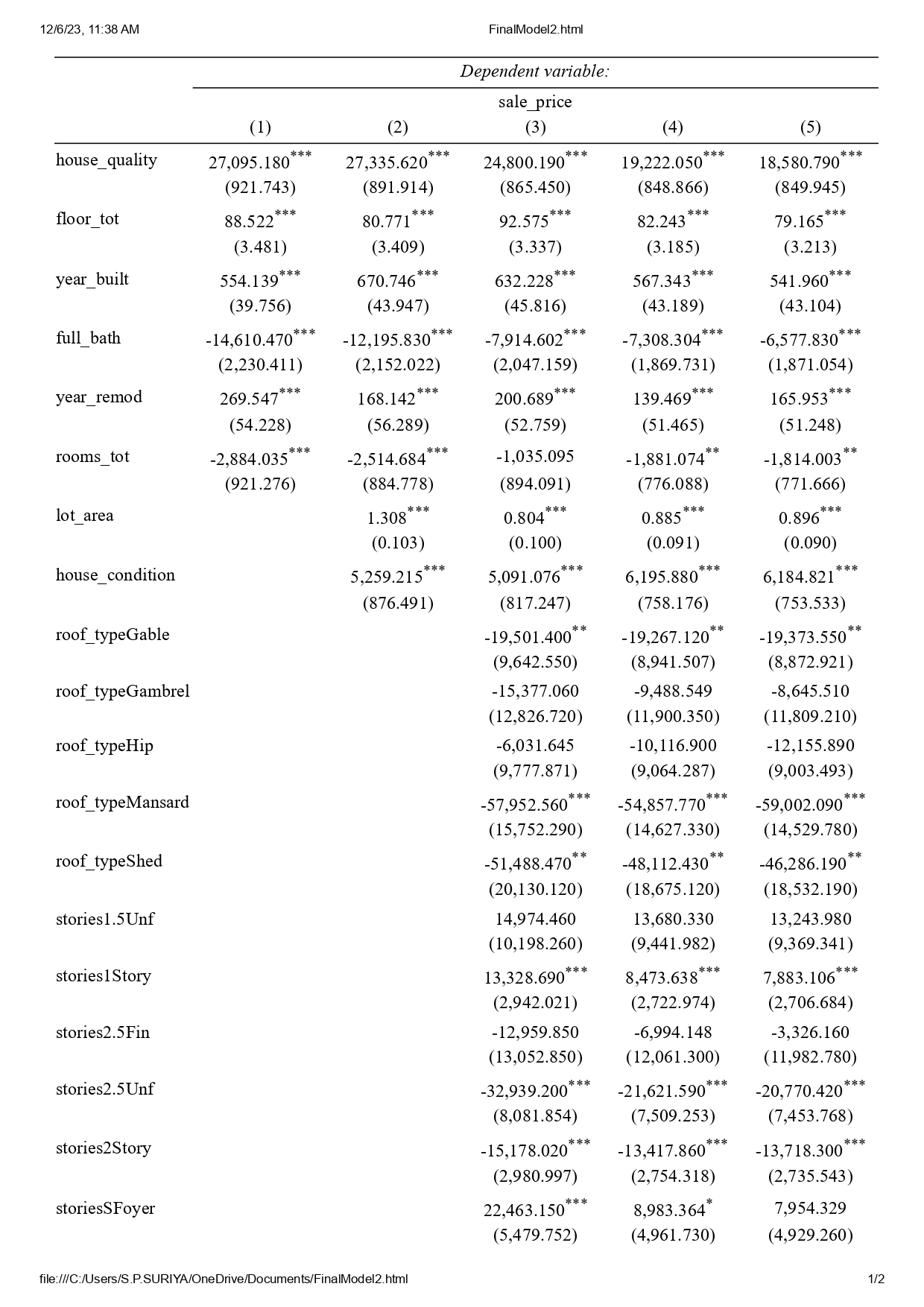
**Visulaization-6**

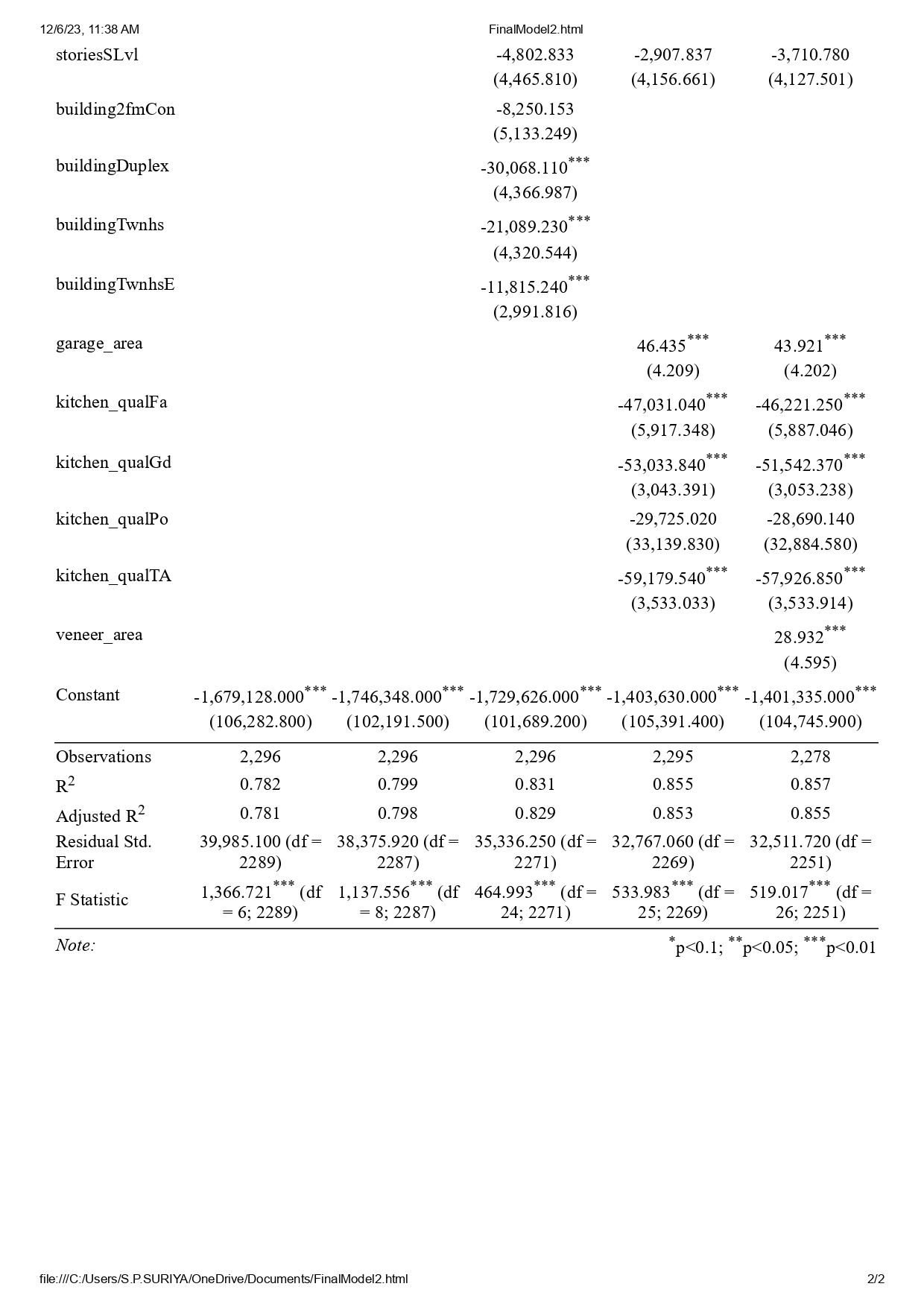


The scatter plot illustrates a moderate positive correlation (Spearman's ρ = 0.493) between the total number of rooms and sale price, indicating that properties with more rooms tend to sell for higher prices,although this relationship is not very strong.The plot also indicates variability in sale prices for properties with similar room counts, suggesting other factors influence price.The data is most dense around 5 to 7 rooms, reflecting the common property size in the market.

**Regression Model**

The initial regression model was constructed using a training dataset to predict housing prices, selecting variables based on their established significant correlation with the sale price. The base model demonstrated a substantial adjusted R-squared value of 0.781, accounting for 78.1% of the variance in sale prices.To enhance the model's predictive accuracy, additional variables, were incorporated.Such as lot area and house condition,since they have realtion with the house price(Wolverton,1997). Resulting in an improved variance explanation of 79.8%.Buyers are ready to pay more to live in school zones and possesion of garage has small effect on the sale price. (Gibson,et,al,2007,Ball,1973),subsequent adding the variables,ultimately helped achieve an adjusted R-squared value of 85.5%, indicating a high level of predictive accuracy.

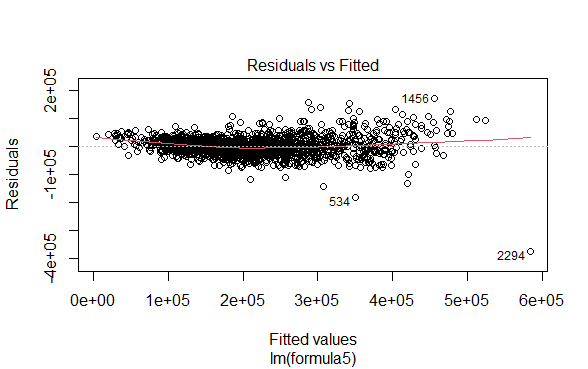
****

****

The increasing R-squared and decreasing residual standard error across models imply an improvement in predictive power and accuracy. Each model's significant F-statistic confirms its collective explanatory variables' predictive utility.The regression analysis uses over 2,278 observations, ensuring robust statistical conclusions.

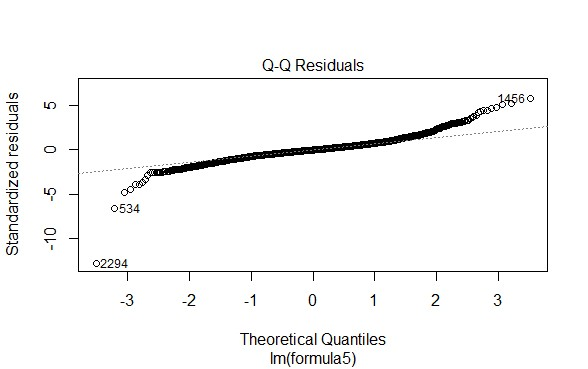
**Assumptions Check**

**Residual Plot**

****

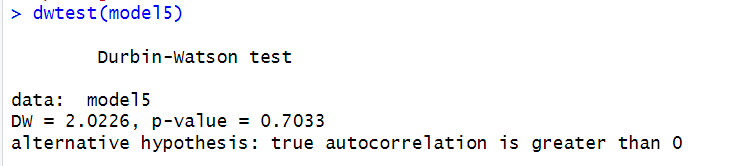
The data points are mostly clustered around the horizontal line at zero,indicating generally accurate predictions,with some spread indicating potential heteroscedasticity where variance increases with the fitted values.Notable outliers are evident, specifically points labeled "14560","534", and "2294",suggesting significant deviations from actual values in those cases. While the plot lacks clear systematic patterns,suggesting no major bias in model predictions,the presence of outliers and changing variance might imply the need for model refinement or a different modeling approach.

**Q-Q Plot**



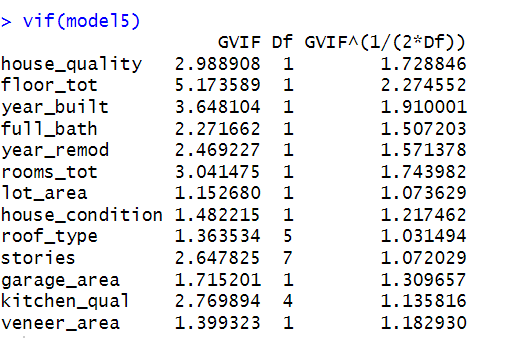
The Q-Q plot indicates that the data closely follows the expected theoretical distribution,which is often the normal distribution.The data points align well with the reference line,suggesting a good match in terms of central tendency and variability.A few outliers are present but they're relatively minimal,suggesting that statistical analyses conducted on this dataset should be valid and reliable.Overall, the data appears to be normally distributed with consistent variance and minimal skewness or heavy tails.

**Durbin-Watson test**

****

The Durbin-Watson test result displayed in the image shows a DW statistic of 2.0226 and a p-value of 0.7033,suggesting no evidence of autocorrelation in the residuals of the regression model.Since the DW statistic is close to 2 and the p-value is well above conventional thresholds for statistical significance.

**Assumption of No Multicollnearity**

****

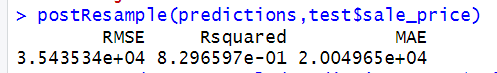
The Variance Inflation Factor analysis indicates that the highest VIF is observed for 'year\_built' at approximately 3.65, suggesting a moderate correlation with other variables, but still within an acceptable range.The remaining variables demonstrate low to moderate multicollinearity, with several variables showing VIF values close to 1,implying that they are relatively independent and do not significantly inflate the variance of the estimated regression coefficients.Overall, the VIF results suggest that multicollinearity is not a major concern for this model.

**Cooks Distance**

**Screenshot 2023-12-06 100728**

A Cook's distance exceeding 1 suggests that the corresponding data point might be disproportionately influencing the model,but in this case,we observe that there are no Cook's distance values surpassing 1.

**Accuracy Prediction**

****

The model was evaluated using a test dataset to ascertain its predictive efficacy due to the presence of high variance in the training data.It yielded a satisfactory RMSE of $35,435,which is relatively minor relative to the peak sale price of $650,000 within the dataset.Additionally,the model demonstrated an R-squared value of approximately 0.830, indicating it accounts for roughly 83% of the variance in house prices.The MAE was reported at $20,049,further underscoring the model's respectable predictive accuracy in the context of the property values concerned.

**Conclusion**

The regression model developed for house price prediction showcases high accuracy while adhering to all necessary assumptions.Its implementation promises to enable precise, automated estimations of property values,enhancing efficiency and potentially boosting profits.This accuracy could also bolster customer trust in real estate agencies.The primary challenge is minor deviations from actual prices, attributed to missing or irrelevant data, slightly impacting accuracy.Despite these issues, the model's high accuracy makes it a valuable tool for estate agencies, aiding in reliable price forecasting.

**Reflective Commentary**

Throughout the module, I've significantly improved my understanding and application of statistics and R programming. Initially challenging concepts like probability and hypothesis testing have become more familiar with practice.Although mastering R, particularly the ggplot2 package, was initially difficult,persistent practice and community support helped me overcome these hurdles.I still find interpreting complex statistical models to be challenging, but I'm tackling this through deeper study and practical application.The most intriguing aspect has been machine learning, which has vast applications across various sectors.Moving forward, I plan to incorporate the statistical and programming skills I've developed into my career, utilizing them to make informed,data-driven decisions and to address real-world problems.

**References**

Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F. and Oluwadara, G., 2022. “House price prediction using random forest machine learning technique.” *Procedia Computer Science*, *199*, pp.806-813. doi: https://doi.org/10.1016/j.procs.2022.01.100

Ball, M.J., 1973. “Recent empirical work on the determinants of relative house prices”.*Urban studies*, *10*(2), pp.213-233. doi:<https://doi.org/10.1080/00420987320080311>

Bogin, A.N. and Doerner, W.M. ,2019 “Property Renovations and Their Impact on House Price Index Construction”, *Journal of Real Estate Research*, 41(2), pp. 249-283.Available at: https://www.tandfonline.com/doi/abs/10.1080/10835547.2019.12091526

Feng, S.-T., Peng, C.-W., Yang, C.-H., & Chen, P.-W.,2021. “Non-linear relationships between house size and price”. *International Journal of Strategic Property Management*, *25*(3), 240-253. <https://doi.org/10.3846/ijspm.2021.14607>

Gibson, J., Sabel, C., Boe-Gibson, G. and Kim, B., 2007, July. “House prices and school zones: does geography matter?”. In *Spatial Econometrics Association conference, Cambridge*.

Hu, M., Yang, Y. and Yu, X., 2020. “Living better and feeling happier: An investigation into the association between housing quality and happiness”. *Growth and Change*, *51*(3), pp.1224-1238.doi: <https://doi.org/10.1111/grow.12392>

Jauregui, A., Tidwell, A. and Hite, D., 2017. “Sample selection approaches to estimating house price cash differentials.” *The Journal of Real Estate Finance and Economics*, *54*, pp.117-137. doi:https://doi.org/10.1007/s11146-015-9529-9

John Cubbin (1974),“Price, quality, and selling time in the housing market”, *Applied Economics*, 6:3, 171-187, DOI: 10.1080/00036847400000017

Mora-Garcia, R.T., Cespedes-Lopez, M.F. and Perez-Sanchez, V.R., 2022. “Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times.” *Land*, *11*(11), p.2100.doi: https://doi.org/10.3390/land11112100

Truong, Q., Nguyen, M., Dang, & Mei, B. (2020). “Housing Price Prediction via Improved Machine Learning Techniques.” *Procedia Computer Science*, 174, 433-442. <https://doi.org/10.1016/j.procs.2020.06.111>

[Wolverton, Marvin L.](https://marketplace.copyright.com/rs-ui-web/mp/search/author/Wolverton, Marvin L.),”Empirical study of the relationship between residential lot price, size and view”, *Journal of property valuation and investment*, 01 Mar 1997, Vol. 15, Issue 1, pages 48 - 57.**DOI:** [10.1108/14635789710163801](https://doi.org/10.1108/14635789710163801" \t "https://marketplace.copyright.com/rs-ui-web/mp/search/all/_blank)

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: https://doi.org/10.1016/j.jhe.2014.06.003

Zhan, C., Liu, Y., Wu, Z., Zhao, M. and Chow, T.W., 2023. “A hybrid machine learning framework for forecasting house price.” *Expert Systems with Applications*, *233*, p.120981.doi: https://doi.org/10.1016/j.eswa.2023.120981

- Zhang, Q. (2021). “Housing Price Prediction Based on Multiple Linear Regression.” *Hindawi*. [https://doi.org/10.1155/2021/7678931.](https://doi.org/10.1155/2021/7678931【15†source】.)

Zaki, J., Nayyar, A., Dalal, S. and Ali, Z.H., 2022. “House price prediction using hedonic pricing model and machine learning techniques.” *Concurrency and Computation: Practice and Experience*, *34*(27), p.e7342. doi:https://doi.org/10.1002/cpe.7342

**Appendix**

1. **Addressing Data Quality**

ames\_processed <- ames %>%

mutate(

zone = str\_replace(zone, "A\\(agr\\)", "A"),

zone = str\_replace(zone, "C\\(all\\)", "C"),

building = str\_replace(building, "Twhs", "TwhsI")

) %>%

filter(

lot\_area >= 100,

house\_quality <= 10,

rooms\_tot != 0,

sale\_price <= 650000

) %>%

mutate(floor\_tot = floor1\_sf + floor2\_sf)

library(openxlsx)

write.xlsx(ames\_processed,"C:/Users/S.P.SURIYA/OneDrive/Desktop/Statics For Business/ames\_processed.xlsx")

summary(ames$zone)

1. **Correlation**

**House Quality and Sale Price**

cor(ames\_processed$sale\_price, ames\_processed$house\_quality, method = "spearman") ,

**Floor Total and Sale Price**

cor(ames\_processed$sale\_price, ames\_processed$floor\_tot, method = "pearson")

**Year Built and Sale Price**

cor(ames\_processed$sale\_price, ames\_processed$year\_built, method = "spearman")

**Full Bath and Sale Price**

cor(ames\_processed$sale\_price, ames\_processed$full\_bath, method = "spearman")

**Year Remodeled and Sale Price**

cor(ames\_processed$sale\_price, ames\_processed$year\_remod, method = "spearman")

**Rooms Total and Sale Price**

cor(ames\_processed$sale\_price, ames\_processed$rooms\_tot, method = "spearman")

1. **Visualization - 1**

ggplot(ames\_processed, aes(x = house\_quality, y = sale\_price)) +

geom\_point(aes(color = house\_quality), alpha = 0.6) +

geom\_smooth(method = "lm", se = FALSE, color = "blue") +

scale\_color\_viridis\_c() +

labs(title = "Relationship between Sale Price and House Quality",

x = "House Quality",

y = "Sale Price") +

theme\_minimal() +

theme(legend.position = "none")

1. **Visualization - 2**

ggplot(ames\_processed, aes(x = floor\_tot, y = sale\_price)) +

geom\_smooth() +

theme\_minimal()+

labs(title = " Floor Total Square Feet vs Sale Price",

x = "Total Floor Square Feet ",

y = "Sale Price")

**5)Visualization - 3**

ggplot(ames\_processed, aes(x = year\_built, y = sale\_price)) +

geom\_point(aes(color = year\_built)) +

geom\_smooth(method = "lm", se = FALSE) +

labs(title = "Sale Price vs. Year Built",

x = "Year Built",

y = "Sale Price") +

theme\_minimal()

**6)Visualization -4**

ggplot(ames\_processed, aes(x = factor(full\_bath), y = sale\_price, fill = factor(full\_bath))) +

geom\_bar(stat = "summary", fun = "mean") +

scale\_fill\_brewer(palette = "Set3") +

labs(title = "Average Sale Price by Number of Full Bathrooms",

x = "Number of Full Bathrooms",

y = "Average Sale Price") +

theme\_minimal() +

theme(legend.position = "none")

1. **Visualization -5**

ggplot(ames\_processed, aes(x = year\_remod, y = sale\_price)) +

geom\_point(color = "green") +

labs(title = 'Sale Price vs. Year of Remodeling',

x = 'Year of Remodeling',

y = 'Sale Price') +

theme\_minimal()

1. **Visualization - 6**

ggplot(ames\_processed, aes(x = rooms\_tot, y = sale\_price, color = sale\_price)) +

geom\_point(alpha = 0.6) +

theme\_minimal() +

labs(title = "Relationship between Total Rooms and Sale Price",

x = "Total Rooms",

y = "Sale Price",

color = "Sale Price") +

scale\_color\_gradient(low = "blue", high = "red") +

geom\_smooth(method = "lm", color = "black")

1. **Data Seperation**

data<- ames\_processed

install.packages("caret")

library(caret)

set.seed(40423910)

index <- createDataPartition(data$sale\_price, times=1, p=0.8,list = FALSE)

train <- data[index,]

test <- data[-index,]

1. **Regression Model**

**Model-1**

formula <- (sale\_price ~ house\_quality + floor\_tot + year\_built + full\_bath + year\_remod + rooms\_tot)

model1 <- lm(formula = formula, data = train)

summary(model1)

**Model-2**

formula2 <- (sale\_price ~ house\_quality + floor\_tot + year\_built + full\_bath + year\_remod + rooms\_tot + lot\_area + house\_condition)

model2 <- lm(formula = formula2, data = train)

summary(model2)

**Model-3**

formula3<- (sale\_price ~ house\_quality + floor\_tot + year\_built + full\_bath + year\_remod + rooms\_tot + lot\_area + house\_condition +roof\_type + stories)

model3 <- lm(formula = formula3, data = train)

summary(model3)

**Model-4**

formula4 <- (sale\_price ~ house\_quality + floor\_tot + year\_built + full\_bath + year\_remod + rooms\_tot + lot\_area + house\_condition +roof\_type + stories + garage\_area + kitchen\_qual)

model4 <- lm(formula = formula4, data = train)

summary(model4)

**Model-5**

formula5 <- (sale\_price ~ house\_quality + floor\_tot + year\_built + full\_bath + year\_remod + rooms\_tot + lot\_area + house\_condition +roof\_type + stories + garage\_area + kitchen\_qual + veneer\_area )

model5 <- lm(formula = formula5, data = train)

summary(model5)

1. **Assumptions Checks**

**VIF**

install.packages("car")

library(car)

vif(model5)

**Cooks Distance**

sum(cooks.distance(model5)>1)

**Durbin-Watson test**

install.packages("lmtest")

library(lmtest)

dwtest(model5)

**Residual Plot**

plot(model5)

1. **Accuracy Check**

library(caret)

predictions <- predict(model5, newdata = test)

test$predictions <- predictions

postResample(predictions,test$sale\_price)