MGT7177: Statistics for Business

Title : Insightful Predictions: Multiple Linear Regression Approach to House Price Modelling in R

Name: Dhanush Mathighatta Shobhan Babu
Student ID: 40412492
Word Count:2170

Table Of Content

Table of figures

1. Introduction

1.1 Overview and Problem Statement

The task focuses on investigating the dynamics of house pricing using statistical analysis and predictive modelling techniques, employing the Ames Housing dataset. The main objective is to understand the factors that affect the cost of houses, examining different qualities of houses and their influence on pricing. According to (Adair *et al.,* 2000) homebuyers assess the property's proximity to amenities. A prospective house buyer will consider the distance to work, shopping, and school. Transport accessibility assesses how easy it is to get to and from amenities, including journey time, cost, convenience, and transportation options (Adair *et al.,* 2000).

1.2 Literature review

We have conducted a comprehensive analysis of approximately 10 research publications focused on utilising machine learning algorithms for predicting the value of residential properties.
The table below provides an entire summary of the 10 papers.

Title of the paper	Year of publication	Author	Model used, Accuracy or Conclusion
Machine Learning based Predicting House Prices using Regression Techniques	2020	J Mansa,Radha Gupta , N S Narahari	Metric considered is R Square <u>Linear regression :</u> On test data – '0.418' On train data – '-2.12' <u>Ridge regression :</u> On test data – '0.4345' On train data – '0.4358' <u>Lasso regression :</u> On test data – '0.4341' On train data – '0.4430' <u>SVM regression :</u> On test data – '0.799' On train data – '0.6330' XG Boost : On test data – '0.7868' On train data – '0.7584'

Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data	2015	Byeonghwa Park , Jae Kwon Bae	C4.5, RIPPER, Naïve Bayesian, and AdaBoost models are used. Conclusion : The analysis reveals that RIPPER outperforms the C4.5, Naïve Bayesian, and AdaBoost models in terms of performance.
House Price Prediction Using Machine Learning and Neural Networks	2018	Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair	The Models used are : Linear Regression Forest Regression Boosted Regression Neural Networks
Housing Price Prediction via Improved Machine Learning Techniques	2019	Quang Truong, Minh Nguyen, Hy Dang, Bo Mei	Metric considered is RMSLE <u>Random Forest :</u> On test data – '0.12980' On train data – '0.16568' <u>Extreme Gradient Boosting :</u> On test data – '0.16118' On train data – '0.16603' <u>Light Gradient Boosting Machine:</u> On test data – '0.16687' On train data – '0.16944' <u>Hybrid Regression:</u> On test data – '0.14969' On train data – '0.16372' <u>Hybrid Regression:</u> On test data – '0.14969' On train data – '0.16372' <u>Stacked Generalization Regression:</u> On test data – '0.16404' On train data – '0.16350'
House Price Prediction using a	2020	Nor Hamizah Zulkifley , Shuzlina Abdul Rahman,	The models used were :

Machine Learning Model: A Survey of Literature		Nor Hasbiah Ubaidullah ,Ismail Ibrahim	Multiple Linear Regression, Support Vector Regression , Artificial Neural Network , XG Boost
House Price Prediction Using Regression Techniques: A Comparative Study	2019	CH. Raga Madhuri, G.Anuradha ,M. Vani Pujitha	The algorithms and their scores : <u>Multiple Linear Regression :</u> Score - 0. 732072 MSE-39187574448.88446 RMSE - 19795851699 <u>Ridge Regression :</u> Score - 0.732164 MSE- 39174049629.73141 RMSE – 19792435330 <u>Ridge Regression :</u> Score - 0.732164 MSE- 39174049629.73141 RMSE – 19792435330 <u>Lasso Regression :</u> Score - 0.732072 MSE- 39187553734.32263 RMSE – 19795846466 <u>Elastic Net Regression :</u> Score - 0.665228 MSE- 48964293085.00798 RMSE – 22127876781 <u>Ada Boosting Regression :</u> Score - 0.7801099 MSE- 32161481079.94242 RMSE – 17933622355 <u>Gradient Boosting Regression :</u> Score - 0.9177022 MSE- 12037006088.27804 RMSE – 109713 90390

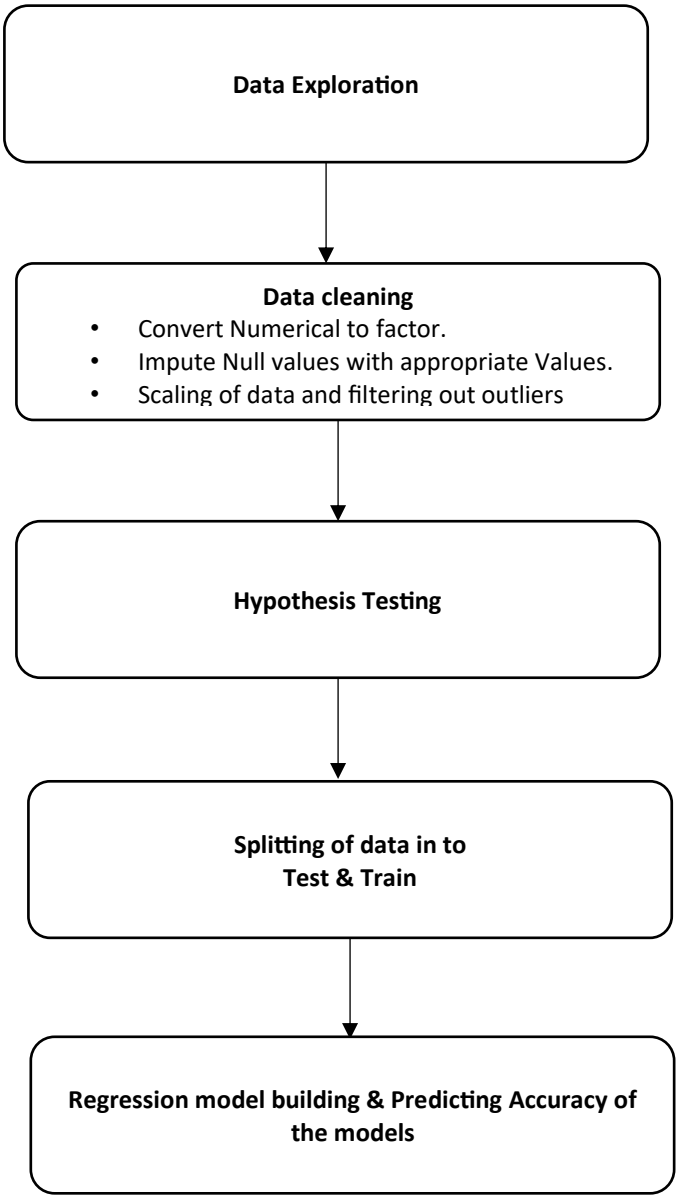
Real Estate Price Prediction with Regression and Classification	2016	Hujia Yu, Jiafu Wu	We have only considered Regression results here : <u>Linear Regression w/o regularization :</u> RMSE w/o PCA : 0.5501 RMSE w/PCA : 0.5473 <u>Lasso :</u> RMSE w/o PCA : 0.5418 RMSE w/PCA : NA <u>Ridge :</u> RMSE w/o PCA : 0.5448 RMSE w/PCA : 0.5447 <u>SVR (linear kernel) :</u> RMSE w/o PCA : 5.503 RMSE w/PCA : NA <u>SVR (Gaussian Kernel) :</u> RMSE w/o PCA : 0.5271 RMSE w/PCA : 0.5269 <u>Random Forest Regression:</u> RMSE w/o PCA : 0.5397 RMSE w/PCA : 0.5323
Factors Influencing Real Estate Property Prices A Survey of Real Estates in Meru Municipality, Kenya	2011	Bernard MESSAH Omboi, Anderson M Kigige	The analysis conducted in Meru municipality demonstrates that the income generated from real estate has a substantial and influential effect on property prices, constituting more than 70% of the overall impact. Conversely, the demand for properties and the involvement of realtors have limited influence on property prices.
House Prices and Accessibility: The Testing of	2010	Alastair Adair,Stanley McGreal,Austin	The research emphasises the significance of sub-market differentiation in

Relationships within the Belfast Urban Area		Smyth,James Cooper &Tim Ryley	understanding the variations in house prices. It highlights that accessibility has a limited impact on prices, with income and demand being the main factors influencing prices in the Belfast housing market. Accessibility only explains less than 2% of the differences in prices, except for West Belfast terraced models, where it accounts for 14% of the variance.
Real Estate Value Prediction Using Linear Regression	2018	Nehal N Ghosalkar, Sudhir N Dhage	The present situation of the real estate industry encounters difficulties in efficiently handling and using extensive volumes of data, underscoring the importance of pertinent and meaningful information. By effectively utilising Linear Regression Algorithms, the system improves decision accuracy and minimises investment risks for clients. Future expansions will include the integration of comprehensive home databases from multiple cities, the incorporation of additional influential factors such as economic recessions, and the provision of detailed property information to enable more extensive analysis and operation on a larger scale. These expansions aim to improve the system's usability and accuracy.

2. Methodology

2.1 Analytical Approach and Tasks

The process encompasses several tasks, including as data pre-processing, hypothesis formulation, data visualisation, statistical association measurement, regression analysis, and model evaluation. The approach described is the conventional way typically employed in Machine Learning models. A similar methodology was utilised by (Ghosalkar & Dhage, 2018) and (Manasa et al., 2020). This can also be referred to as an adaption of CRISP-DM (Schröer et al., 2021).
The flowchart below provides a concise overview of the analytical tasks employed.



Flow chart 2.1 : Steps in analytical Tasks

2.2 Data exploration and data quality assessment

After understanding the business problem, the next step is to understand the data. Analysing the given data and characterising it, and assessing its quality are crucial duties in this phase (Schröer et al., 2021). There are a total of 78 variables, with the target variable being sale_price, which is a dependent variable. The remaining variables are independent. There are a 47 of categorical variables and a 31 of numerical variables. We utilised various built-in functions in 'R' for data exploration, including summary(), count(), and is.na().

2.3 Variable Selection

We have selected a total of 17 variables. The selection of these variables has been based on three factors:

- Derived from a hypothesis
- Supported by research papers
- Established by logical reasoning

The variables that are deduced from a hypothesis are:

1. **lot_area**
2. **neighbourhood**
3. **frontage**
4. **year_remod**
5. **room_tot**

The variables that are examined in research papers:

1. **Zone** - (Yu & Wu, 2016)
2. **year_built** - (Truong et al., 2020)
3. **half_bath** - (Park & Bae, 2015)
4. **full_bath** -(Park & Bae, 2015)
5. **bedrooms** -(Park & Bae, 2015)
6. **aircon** -(Yu & Wu, 2016)

The variables that are determined using deductive reasoning:

1. **kitchen** - A modern, fully furnished kitchen often increases house value. Improved appliances, countertops, and layout may attract buyers and raise prices.
2. **Foundations** - Home structural stability depends on its foundation. Buyers want a durable, well-maintained property with a solid foundation and no cracks or structural problems, which boosts the sale price.

3. **stories** - Story count affects a home's value. For practical reasons, some homebuyers choose single-story homes, while others prefer multi-story homes. Pricing depends on square footage, layout, and stories
4. **heat_type** - A homeowner's comfort and utility bills depend on the heating system type. If a home has a modern, more energy-efficient heating system like a gas furnace or geothermal heating, its value may rise.
5. **house_quality** - Quality of structure, finishes, and workmanship affects a home's appeal and price. Quality craftsmanship and materials cost extra.
6. **house_condition** - Inside and exterior, the house's condition matters. Well-kept and well-maintained properties might sell for more than those in need of repairs.

2.4 Data quality issues

We have just focused on resolving the data quality concerns pertaining to the factors that have been indicated previously. The table displays the data quality concerns.

Sl_NO	Variable Name	Data Type	Outlier / Data quality issues
1	lot_area	Numerical	There are multiple outliers that fall below and above the upper and lower bounds.
2	neighboured	Categorical	There are no data quality issues.
3	frontage	Numerical	There are 480 outliers that need to be adjusted using either the mean or median, depending on the distribution of the data. Additionally, there are outliers that fall below the lower bound and above the upper bound.
4	year_remod	Numerical	There are no data quality issues.
5	room_tot	Numerical	When a box plot is generated, certain data points may appear as outliers. However, these points cannot be classified as outliers since it is possible for there to be zero rooms at the minimum and fifteen rooms at the maximum.
6	zone	Categorical	There are no data quality issues.
7	year_built	Numerical	There are outliers that fall below the lower limit.
8	half_bath	Numerical	There are no data quality issues.
9	full_bath	Numerical	There are no data quality issues.
10	bedrooms	Numerical	There are no data quality issues.
11	aircon	Categorical	There are no data quality issues.
12	kitchen	Numerical	There are no data quality issues.

13	foundations	Catergorical	There are no data quality issues.
14	heat_type	Catergorical	There are no data quality issues.
15	house_quality	Catergorical	There is a single data quality concern. The value "11" is an outlier in the house ratings data, as indicated by the data dictionary. The acceptable range for house ratings is from 1 to 11. Therefore, the value of "11" needs to be filtered out.
16	house_condtion	Catergorical	There are no data quality issues.
17	stories	Categorical	There are no data quality issues.

2.5 Addressing data quality issues

Address poor data quality by cleansing it. Create derived characteristics based on the selected model from the first phase. Which strategy is best for these processes depends on the model (Schröer et al., 2021).

1. lot_area

To remove outliers we need first calculate the upper bound and lower bound Through Inter-Quartile Range (IQR), an outlier x can be detected

if: x < Q1 − 1.5 * IQR (IQR) **OR** Q3 + 1.5 (IQR) < x
where: Q1 = 25th percentiles Q3 = 75th percentiles
IQR = Q3 − Q1

(Truong et al., 2020) .

We utilised the filter function from the "dplyr" library to eliminate the outliers.Upon computing the upper and lower limit

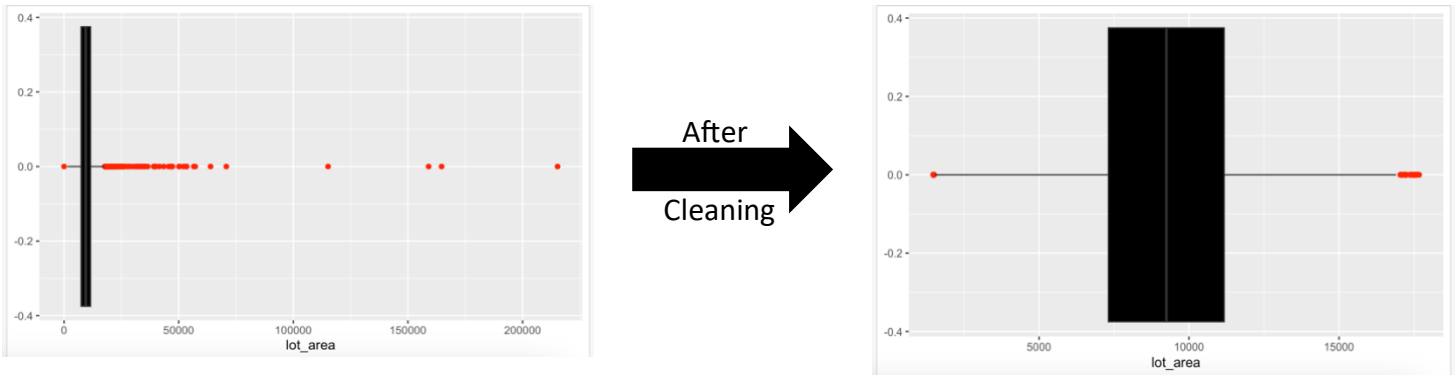


Fig 2.5.1: Boxplot of lot_area before and after filtering outlier

2. frontage

Once again, we have utilised the aforementioned formulas to compute the interquartile range (IQR), upper bound, lower bound, and subsequently eliminated any outliers.

Fig 2.5.2: Boxplot of frontage before and after filtering outlier

3. year_built

There are outliers that have values below the lower limit, but these outliers are removed during the process of correcting the two variables given above.

Fig 2.5.3: Boxplot of year_built before and after filtering outlier

4. house_quality

As previously stated, the value "11" is filtered using the filter function in the 'dplyr' library because it does not exist in the data dictionary.

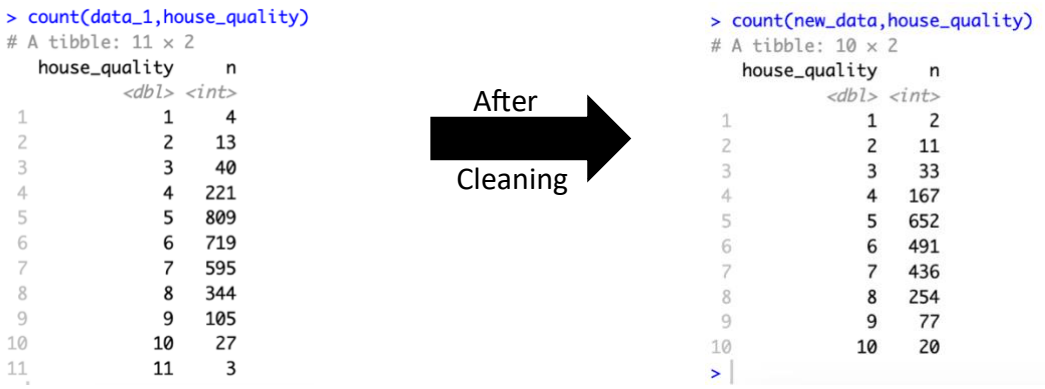


Fig 2.5.4: Before and after cleaning house_quality

2.6 Hypothesis Testing

The most prevalent hypothesis test often entails evaluating the null hypothesis in comparison to the alternative hypothesis.

H0: There is no relationship between X and Y
versus
the alternative hypothesis is
Ha: There is some relationship between X and Y (James et al.).

1. **lot_area vs sale_price**

H0: There is no relationship between lot_area and sale_price
H1: There is a relationship between lot_area and sale_price

The statistical outputs, including the correlation analysis and linear regression model, decisively reject the null hypothesis (H0) in support of the alternative hypothesis (H1). This suggests a significant association between the variables `lot_area` and `sale_price`.The correlation coefficient is roughly 0.377, suggesting a moderate positive linear association between the variables `lot_area` and `sale_price`. Furthermore, the linear regression analysis reveals that the coefficient for `lot_area` is highly statistically significant ($p < 2.2e-16$), suggesting that variations in `lot_area` are linked to variations in `sale_price`.

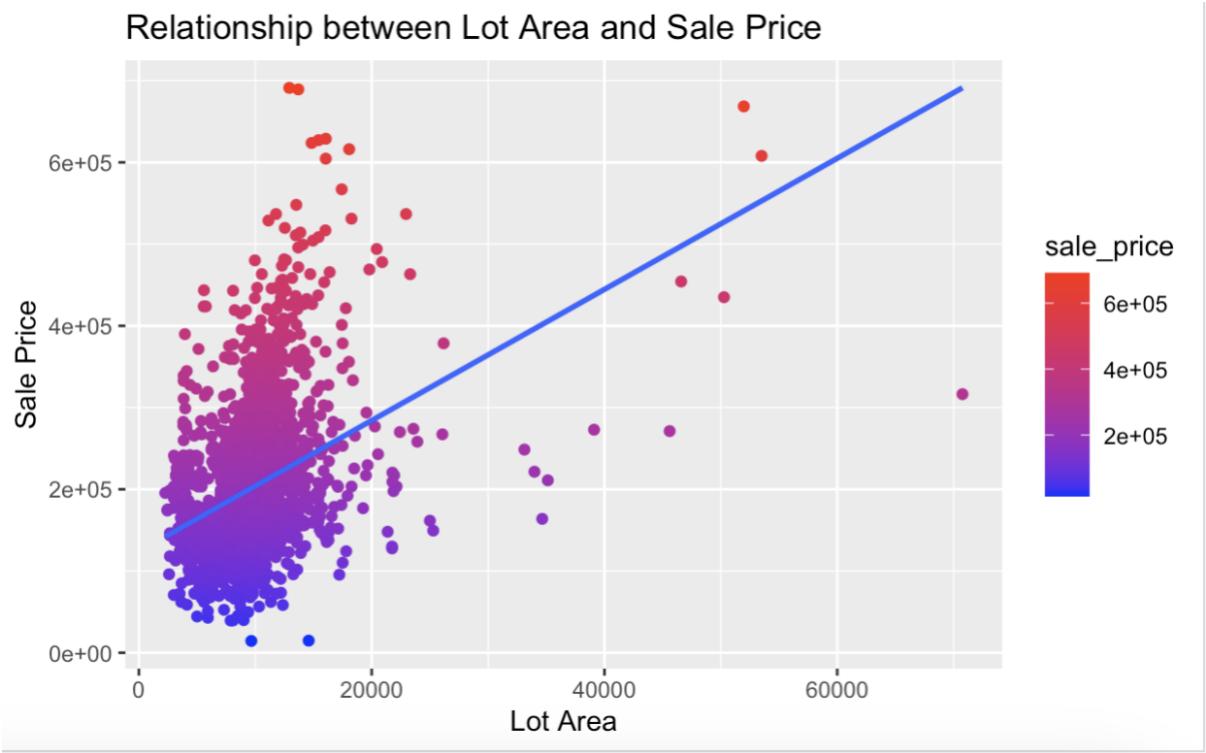


Figure 2.6.1: scatter plot for lot_area vs sale_price

2. neighbourhood vs sale_price

H0: There is no relationship between neighbourhood and sale_price
H1: There is a relationship between neighbourhood and sale_price

An analysis of variance (ANOVA) has been conducted.The ANOVA table reveals a remarkably significant F-statistic (F value = 123.89, p-value < 2.2e-16), suggesting a robust overall association between the `neighbourhood` variable and `sale_price`.Moreover, the coefficients in the linear regression model elucidate the correlation between several levels or categories within the `neighbourhood` variable and `sale_price`.
As an example: The Adjusted R-squared value of 0.5793 indicates that around 57.93% of the variation in `sale_price` can be accounted for by the `neighbourhood` variable in this model. This study presents compelling evidence refuting the null hypothesis (H0: No link exists between `neighbourhood` and `sale_price`) in support of the alternative hypothesis (H1: A relationship exists between `neighbourhood` and `sale_price`).

Figure 2.6.2: bar plot for neighbourhood vs avg_sale_price

3. room_tot vs sale_price

H0: There is no relationship between room_tot and sale_price
H1: There is a relationship between room_tot and sale_price

The statistical analysis strongly supports the alternative hypothesis (H1), rejecting the null hypothesis (H0: No association between `room_tot` and `sale_price`). Correlation Analysis: The Pearson's correlation coefficient between `sale_price` and `rooms_tot` is 0.478, indicating a moderate positive linear relationship between room count and sell prices. Exceptionally low p-value (< 2.2e-16) strongly challenges the null hypothesis of no connection.Linear regression analysis: The linear regression model also shows a link between `rooms_tot` and `sale_price` variables. The coefficient estimate for `rooms_tot` is highly significant (p-value < 2e-16), demonstrating a strong correlation between room count and sale price.

Figure 2.6.3: scatter plot for room_tot vs avg_sale_price

4. year_remod vs sale_price

H0: There is no relationship between year_remod and sale_price
H1: There is a relationship between year_remod and sale_price

The statistical analysis supports the alternative hypothesis (H1), contradicting the null hypothesis (H0: No association between year_remod and sale_price).Correlation Analysis: The Pearson correlation coefficient between `sale_price` and `year_remod` is around 0.555. Sale prices are moderately positively correlated with remodelling year. The low p-value (< 2.2e-16) strongly supports the alternative hypothesis, rejecting the null hypothesis of no connection.Linear regression analysis: The linear regression model confirms the association between `year_remod` and `sale_price`. The coefficient estimate for `year_remod` is of strong statistical significance (p-value < 2e-16). Increased remodelling year leads to a significant increase in sale price.

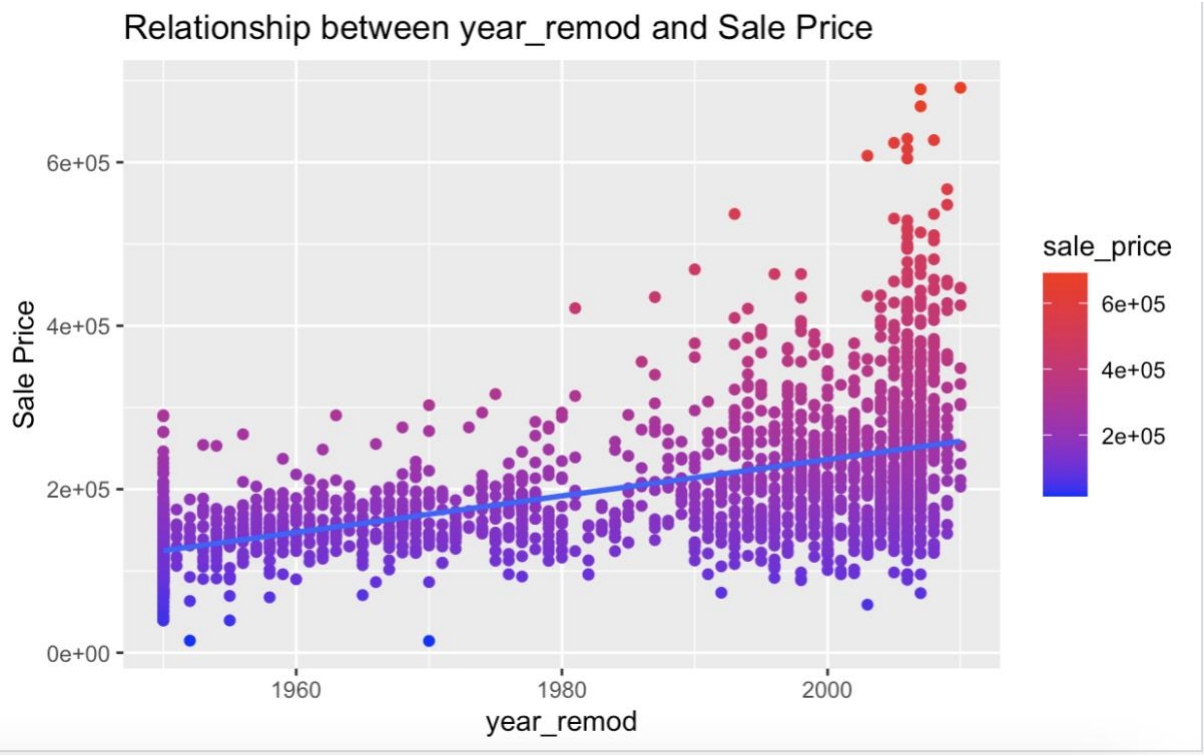


Figure 2.6.4: scatter plot for year_remod vs avg_sale_price

5. **frontage vs sale_price**

H0: There is no relationship between frontage and sale_price
H1: There is a relationship between frontage and sale_price

A statistical research shows a strong association between `frontage` and `sale_price` variables.Correlation Analysis: The Pearson correlation coefficient between `sale_price` and `frontage` is approximately 0.339. Property sale values are slightly positively correlated with frontage size. The low p-value (< 2.2e-16) strongly rejects the null hypothesis of no correlation. This validates the alternative theory that 'frontage' and'sale_price' are related.Linear regression analysis: Additionally, the linear regression model validates the association between `frontage` and `sale_price` variables. The coefficient estimate for `frontage` is highly significant (p-value < 2e-16), indicating a strong correlation between frontage and sale price.

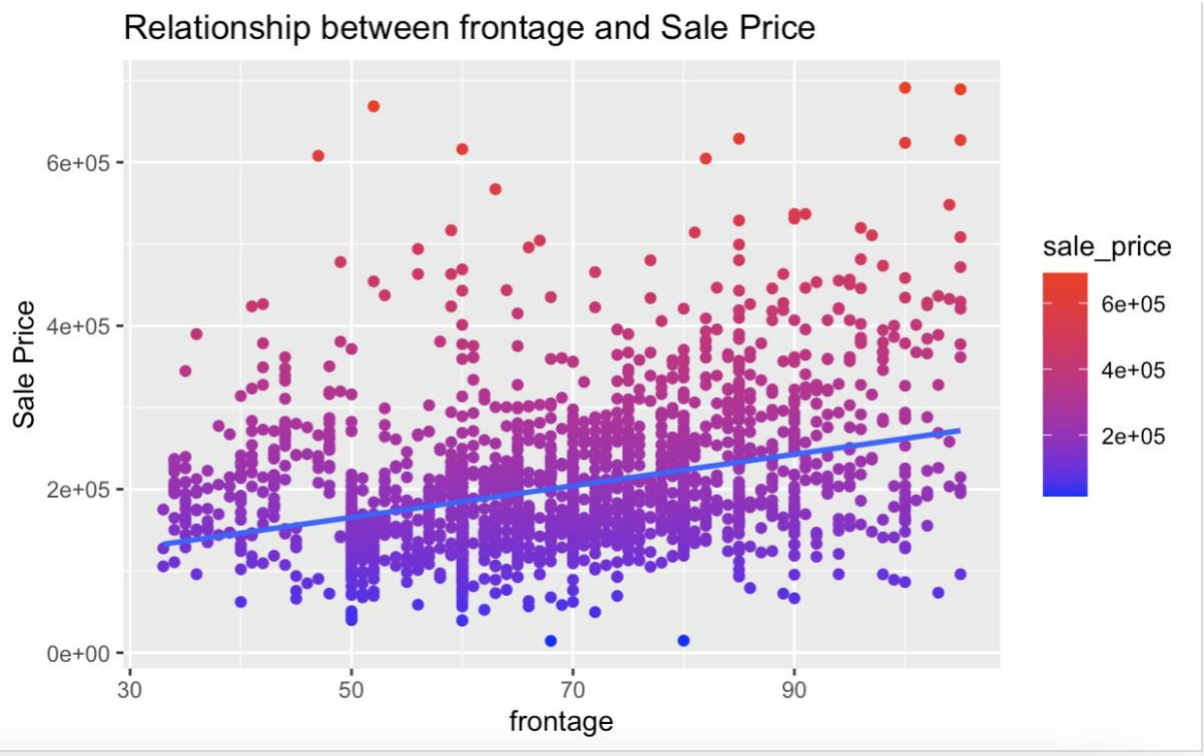


Figure 2.6.5: scatter plot for frontage vs avg_sale_price

2.7 Regression Model Techniques

Here we are making use of multiple linear regression technique to build our model. Regression analysis is conducted to ascertain the correlations between two or more variables that have cause-and-effect relationships, and to create predictions for the subject matter based on these relationships (Uyanık & Güler, 2013). Multivariate regression analysis aims to simultaneously consider the influence of independent variables on the dependent variable, taking into account their variation (Unver & Gamgam, 1999).

Multivariate regression analysis model is formulated as in the following;

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

y = dependent variable
X_i = independent variable
β_i = parameter
ε = error

Figure 2.7: formula to calculate multiple linear regression.

2.8 Model building

We have employed the forward approach to construct our model. A forward-selection rule commences with an absence of explanatory variables and subsequently incorporates variables, one at a time, based on their statistical significance, until there are no remaining variables that exhibit statistical significance (Smith, 2018) .

In total we have built 4 models:
Model 1 : based on variables derived from hypothesis
Model 2: based on variables derived from hypothesis +Literature review
Model 3: Adding all the variables (hypothesis +Literature review Logical thinking)

3. Results and Discussion

3.1 Presentation of Key Outputs

Model 1: The model has an adjusted R-squared of 0.7236, which indicates that 72.36% of the variance in sale_price can be explained by the model. The p-value for the F-statistic is less than 2.2e-16, which means that the model is statistically significant.

Model 2: Overall, Model 2 is a better fit for the data than Model 1. This is because Model 2 has a higher adjusted R-squared (0.7627 vs. 0.7236), which indicates that it explains more of the variance in the sale_price data. Additionally, Model 2 has a higher F-statistic (127.9 vs. 141.2), which indicates that it is a more significant model

Model 3: Model 3 has a higher adjusted R-squared (0.8651 vs. 0.7627) and a higher F-statistic (134.6 vs. 127.9) than Model 2, which suggests that Model 3 is a better fit for the data. Model 3 also has a lower residual standard error (32430 vs. 43020), which suggests that it is more accurate at predicting sale_price.

Overall, Model 3 is a more powerful and accurate predictor of sale_price than Model2.

3.2 Presentation of Key Outputs of all models

Model_Name	RMSE	Rsquared	MAE
Model 1	4.380253e+04	7.390951e-01	3.050533e+04
Model 2	4.066174e+04	7.749076e-01	2.753121e+04
Model 3	2.938687e+04	8.823757e-01	2.084334e+04

3.3 Plot of Key Outputs of Model3

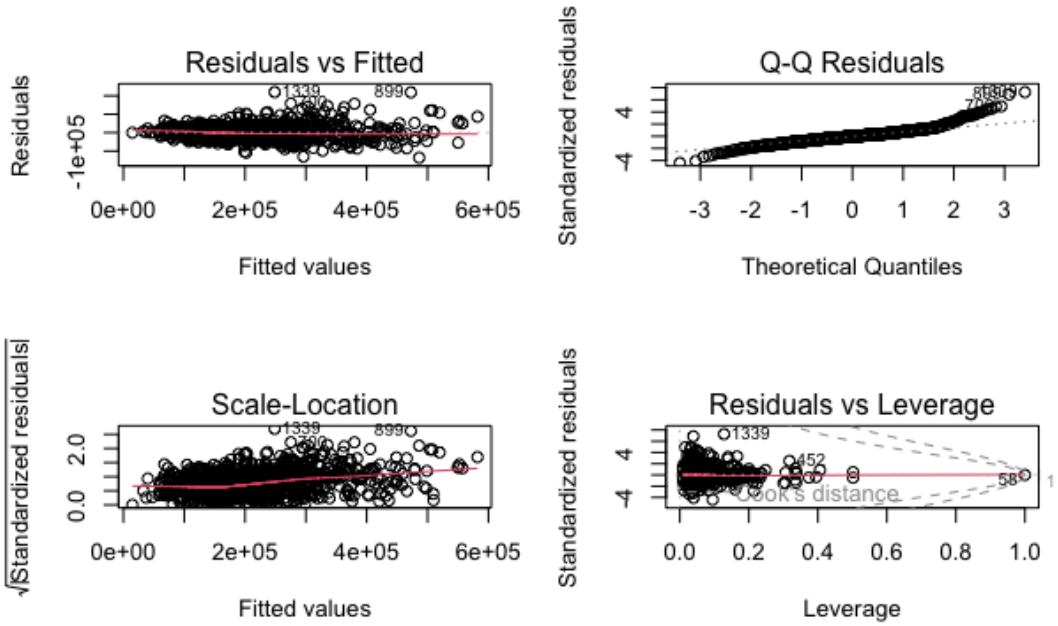


Figure 3.3 : Plots of best model (model 3)

3.4 Model Assumptions

1. **Independent Assumption** – The Durbin-Watson test does not reject the null hypothesis of no autocorrelation at the 0.05 significance level (Field et al., 2012). In other words, there is not enough evidence to suggest that there is positive serial correlation in the residuals.

2. **Multicollinearity Assumption** - In this model, all of the VIFs are below 10, which suggests that there is no severe multicollinearity present. However, there are a few VIFs that are slightly elevated (7.98 for foundations, 7.22 for stories, and 5.01 for heat_type). This suggests that these variables may be slightly correlated with each other, but not enough to cause a significant problem. Overall, the VIFs in this model suggest that there is no severe multicollinearity present, and the model is relatively robust to multicollinearity.

3. **Residuals Assumptions** – The Q-Q plot at figure (), shows that at the end there is a bit deviation from the line stating that the residuals might not be normally distributed.

4.0 Reflective Commentary

4.1 Further Steps

The primary objective of developing these models is to predict values for new observations. To facilitate widespread utilization, we deploy various hosting platforms. R-Shiny, an integrated library in R, proves instrumental in creating user interfaces (UI) for these models. These predictive models are not restricted to a single department but are intended for deployment across diverse organizational segments, including executive management, sales, and marketing. By leveraging R-Shiny's capabilities, we construct intuitive and accessible interfaces, ensuring that the models cater to a broad audience within the company. The precision of these models is geared towards providing actionable insights for strategic decision-making at different organizational levels.

4.2 Learnings and Future Aspiration

This module helped me master CARET, LM, TIDYVERSE, GGPLOT, and other powerful libraries to create complex linear regression models. I want to be a business analyst who contributes to machine learning, especially supervised learning algorithms. As I prepared a report, I discovered a variety of advanced ML algorithms that excited me and strengthened my desire to contribute to this dynamic field. I love using machine learning to gain actionable insights, make informed decisions, and guide business strategies with data. My passion for this field is discovering and applying advanced algorithms to improve predictive modelling, classification, and outcomes.

5. References

Adair, A. et al. (2000) 'House prices and accessibility: The testing of relationships within the Belfast Urban Area', Housing Studies, 15(5), pp. 699–716. DOI Available at: [https://www.tandfonline.com/doi/epdf/10.1080/02673030050134565?needAccess=true [Accessed on 7th November]].

Field, A., Miles, J., & Field, Z. (2012, March 7). Discovering Statistics Using R. SAGE.

Ghosalkar, N.N. and Dhage, S.N. (2018) 'Real estate value prediction using linear regression', 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) [Preprint]. DOI Available at: [https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8697639 [Accessed on 15th November 2023]].

Hamizah Zulkifley, N. et al. (2020) 'House price prediction using a machine learning model: A survey of literature', International Journal of Modern Education and Computer Science, 12(6), pp. 46–54. doi:Available at: [https://www.mecs-press.org/ijmecs/ijmecs-v12-n6/v12n6-4.html [Accessed on 2nd November 2023]].

James, G. et al. (no date) An introduction to statistical learning: With applications in R. Springer.

Kigige, A.M. (2011) Factors Influencing Real Estate Property Prices A Survey of Real Estates in Meru Municipality, Kenya [Preprint]. DOI Available at: [https://www.researchgate.net/profile/Bernard-Omboi/publication/268262225_Factors_Influencing_Real_Estate_Property_Prices_A_Survey_of_Real_Estates_in_Meru_Municipality_Kenya/links/54cf107d0cf24601c092c091/Factors-Influencing-Real-Estate-Property-Prices-A-Survey-of-Real-Estates-in-Meru-Municipality-Kenya.pdf [Accessed on 10th November 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. DOI available at: [https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8882834&tag=1 Accessed on 9th November]].

Manasa, J., Gupta, R. and Narahari, N.S. (2020) 'Machine learning based predicting house prices using regression techniques', 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) [Preprint]. doi:Available at: [https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9074952 [Accessed on 30th September 2023]].

Park, B. and Bae, J.K. (2015) 'Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia Housing Data', Expert Systems with Applications, 42(6), pp. 2928–2934. doi:Available at: [https://www.sciencedirect.com/science/article/pii/S0957417414007325?via%3Dihub [Accessed on 30th September 2023]].

Schröer, C., Kruse, F. and Gómez, J.M. (2021) 'A systematic literature review on applying CRISP-DM process model', Procedia Computer Science, 181, pp. 526–534. doi:Available at:[

https://www.sciencedirect.com/science/article/pii/S1877050921002416 [Accessed on 9th November]].

Smith, G. (2018) 'Step away from stepwise', Journal of Big Data, 5(1). doi:Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6 Accessed on 2nd December 2023]].

Truong, Q. et al. (2020) 'Housing price prediction via Improved Machine Learning Techniques', Procedia Computer Science, 174, pp. 433–442. doi:Available at: [https://www.sciencedirect.com/science/article/pii/S1877050920316318?via%3Dihub [Accessed on 5th November 2023]].

Uyanık, G.K. and Güler, N. (2013) 'A study on multiple linear regression analysis', Procedia - Social and Behavioral Sciences, 106, pp. 234–240. doi:Available at:[https://www.sciencedirect.com/science/article/pii/S1877042813046429 [Accessed on 2nd December 2023]].

Varma, A. et al. (2018) 'House price prediction using Machine Learning and Neural Networks', 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) [Preprint]. doi:Available at: [https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8473231&tag=1 [Accessed on: 2nd December 2023]].

Yu, H. and Wu, J. (2016) 'Real Estate Price Prediction with Regression and Classification', CS 229 Autumn 2016 Project Final Report [Preprint]. DOI available at:[https://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf [Accessed on 7th November]].

Ünver, Ö., Gamgam, H. (1999) *Uygulamalt Istatistik Yöntemleri*. Ankara: Siyasal Kitabevi.

6.Appendix

6.1 R code

```
library(tidyverse)
library(dplyr)

#to know the working directory
getwd()

#set the working directory
setwd("/Users/dhanush/Desktop/Bussiness analytics /Stastics for bussiness ")

#import the data set
library(readxl)
data <- read_excel("ames.xlsx")

colnames(data)
summary(data$ID)
unique(data)

class(data$ID)
summary(data$d_type)
summary(data$zone)
summary(data$lot_area)
summary(data$road)
summary(data$house_quality)
summary(data$house_condition)
summary(data$year_built)
summary(data$year_remod)
summary(data$veneer_area)
summary(data$bsmt_sf1)
summary(data$bsmt_sf2)
summary(data$bsmt_unf)
summary(data$bsmt_area)
summary(data$aircon)
summary(data$floor1_sf)
summary(data$floor2_sf)
summary(data$low_qual_sf)
summary(data$bsmt_full_bath)
summary(data$bsmt_half_bath)
summary(data$full_bath)
summary(data$half_bath)
summary(data$bedroom)
summary(data$kitchen)
summary(data$rooms_tot)
```

```
summary(data$fireplace)
summary(data$garage_cars)
summary(data$garage_area)
summary(data$deck_sf)
summary(data$open_porch_sf)
summary(data$encl_porch_sf)
summary(data$season_porch)
summary(data$screen_porch)
summary(data$pool_sf)
summary(data$features_val)
summary(data$month_sold)
summary(data$year_sold)
summary(data$sale_price)

#Data cleaning

#1)lot_area
#cleaning of lot area
summary(lot_area)

ggplot(data , aes(x = lot_area)) +
  geom_boxplot(fill = "black", outlier.color = "red")  # Color outliers in red

# IOQ = Q3-Q1
#lower bound = Q1 - 1.5 * IQR
#upper boound = Q3 + 1.5 * IQR

new_data <- data %>% filter(lot_area >= 1271.5 & lot_area <=17753.5 )

#ggplot after cleaning the data
ggplot(new_data, aes(x = lot_area)) +
  geom_boxplot(fill = "black", outlier.color = "red")  # Color outliers in red


#2)zone
#cleaning of zone

summary(data$zone)
count(data,zone)

sum(is.na(data$zone))

#convert nominal data into factor
new_data$zone <- as.factor(new_data$zone)
summary(new_data$zone)

ggplot(new_data)+
```

```
  geom_bar(aes(x=zone,fill="red"))

# 3) Neighbourd
summary(data$neighbourhood)
count(data,neighbourhood)
sum(is.na(data$neighbourhood))

#convert nominal data into factor
new_data$neighbourhood <- as.factor(new_data$neighbourhood)
summary(new_data$neighbourhood)

# 4) frontage

summary(data$frontage)
# again summarise data without including NA
summary(data$frontage, na.rm = T)
mean(data$frontage, na.rm = T)
median(data$frontage, na.rm= T)

ggplot(data, aes(x = frontage)) +
  geom_boxplot(fill = "black", outlier.color = "red")  # Color outliers in red

ggplot(data, aes(x = lot_area)) +
  geom_histogram(fill = "black", outlier.color = "red")  # Color outliers in red

#replace NA with mean
new_data <- data %>%
  mutate(frontage = ifelse(is.na(frontage), 68, frontage))

summary(new_data$frontage)
ggplot(new_data, aes(x = frontage)) +
  geom_boxplot(fill = "black", outlier.color = "red")  # Color outliers in red

#remove outliers
new_data <- data %>% filter(frontage >= 32.625 & frontage <= 105.625 )

#ggplot after rermoving outliers
ggplot(new_data, aes(x = frontage)) +
  geom_boxplot(fill = "black", outlier.color = "red")  # Color outliers in red

# 5)year_built
summary(data$year_built)
sum(is.na(data$year_built))

ggplot(new_data, aes(x = year_built)) +
  geom_boxplot(fill = "black", outlier.color = "red")  # Color outliers in red
# as we have already cleaned some data in above it has already filtered some outliers
```

```
# 6) cleaning of year_remod
summary(year_remod)
sum(is.na(year_remod))

#ggplot using clean data of shown above
ggplot(new_data, aes(x = year_remod)) +
  geom_boxplot(fill = "black", outlier.color = "red")

#7) half_bath
summary(data$half_bath)
count(data,half_bath)
sum(is.na(half_bath))

#ggplot using clean data of shown above
ggplot(new_data, aes(x = half_bath)) +
  geom_boxplot(fill = "black", outlier.color = "red")

#8) full_bath
summary(data$full_bath)
count(data,full_bath)
sum(is.na(full_bath))
class(full_bath)

#ggplot using clean data of shown above
ggplot(new_data, aes(x = full_bath)) +
  geom_boxplot(fill = "black", outlier.color = "red")

#9)bedroom
summary(data$bedroom)
count(data,bedroom)
sum(is.na(data$bedroom))

#ggplot using clean data of shown above
ggplot(new_data, aes(x = bedroom)) +
  geom_boxplot(fill = "black", outlier.color = "red")

#10) kitchen
summary(data$kitchen)
count(data,kitchen)
sum(is.na(data$kitchen))

#ggplot using clean data of shown above
ggplot(new_data, aes(x = kitchen)) +
  geom_boxplot(fill = "black", outlier.color = "red")

#11) foundations
```

```
summary(data$foundations)
count(data,foundations)
sum(is.na(data$foundations))

#convert data in to factor
new_data$foundations <- as.factor(new_data$foundations)

summary(new_data$foundations)
count(new_data,foundations)
sum(is.na(new_data$foundations))

barchart(new_data$foundations)

# 12) stories

summary(data$stories)
count(data,stories)
sum(is.na(data$foundations))

#convert data in to factor
new_data$stories <- as.factor(new_data$stories)

summary(new_data$stories)
count(new_data,stories)
sum(is.na(new_data$stories))

#plotting to check if converted to factor
barchart(new_data$stories)

#13) room_ tot

summary(data$rooms_tot)
count(data,rooms_tot)
sum(is.na(data$rooms_tot))

#plot box plot to check outliers
ggplot(new_data, aes(x = rooms_tot)) +
  geom_boxplot(fill = "black", outlier.color = "red")
#keep outliers as u can back those

#14) Aircon
summary(data$aircon)
count(data,aircon)
sum(is.na(data$aircon))

#convert this into factor
new_data$aircon <- as.factor(new_data$aircon)
```

```
#to check if converetd into factor
summary(new_data$aircon)
count(new_data,aircon)
sum(is.na(new_data$aircon))

barchart(new_data$aircon)

#15) Heat_type
summary(data$heat_type)
count(data,heat_type)
sum(is.na(data$heat_type))

#convert to factor
new_data$heat_type <- as.factor(new_data$heat_qual)

summary(new_data$heat_type)
count(new_data,heat_type)
sum(is.na(new_data$heat_type))

barchart(new_data$heat_type)

#16) house_quality

summary(data$house_quality)
count(new_data,house_quality)
sum(is.na(data$house_quality))

new_data <- new_data %>%
  filter(house_quality <11)

count(data_1,house_quality)

hist(data_1$house_quality)

#convert to nomial data as mentioned in the data dictonary
new_data$house_quality <- as.factor(new_data$house_quality)

barchart(new_data$house_quality)

#17) House condtion

summary(data$house_condition)
count(new_data,house_condition)
sum(is.na(data$house_condition))

#convert to ordinal data
```

```
new_data$house_condition <- as.factor(new_data$house_condition)

barchart(new_data$house_condition)

#Convert all character into factor
new_data<-new_data %>% mutate_if(is.character,as.factor)

#Hypothesis testing

#H1 - lot_area vs sale_price

cor.test(new_data$sale_price, new_data$lot_area, method = "pearson")
m1_lot_area <- lm(new_data$sale_price ~ new_data$lot_area)
summary(m1_lot_area)

#plotting relationship between lot_area vs sale_price
ggplot(new_data, aes(x = lot_area, y = sale_price, color = sale_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Lot Area", y = "Sale Price") +
  ggtitle("Relationship between Lot Area and Sale Price") +
  scale_color_gradient(low = "blue", high = "red")

#H2 - Neighbourhood vs sale_price
m2_neighbourhood <- lm(new_data$sale_price ~ new_data$neighbourhood)
summary(m2_neighbourhood)
anova_neighbourhood <- anova(m2_neighbourhood)
summary(anova_neighbourhood)
print(anova_neighbourhood)

#plotting relationship between Neighbourhood vs sale_price
ggplot(hq1, aes(x = neighbourhood, y = avg_sale_price, fill = avg_sale_price)) +
  geom_bar(stat = "identity") +  # Creating bars
  labs(x = "Neighbourhood", y = "Average Sale Price", title = " Average Sale Price by
Neighbourhood") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +  # Rotating x-axis labels
  scale_fill_gradient(low = "blue", high = "red")

#H3 room_tot vs sale_price
cor.test(new_data$sale_price, new_data$rooms_tot, method = "pearson")
m3_room_tot <- lm(new_data$sale_price ~ new_data$rooms_tot)
summary(m3_room_tot)

#create a scatter plot for the same
ggplot(new_data, aes(x = rooms_tot, y = sale_price, color = sale_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
```

```
  labs(x = "rooms_tot", y = "Sale Price") +
  ggtitle("Relationship between rooms_tot and Sale Price") +
  scale_color_gradient(low = "blue", high = "red")

#H4 year of remod
cor.test(new_data$sale_price, new_data$year_remod, method = "pearson")
m4_year_remod <- lm(new_data$sale_price ~ new_data$year_remod)
summary(m4_year_remod)

#create a scatter plot for the same
ggplot(new_data, aes(x = year_remod, y = sale_price, color = sale_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "year_remod", y = "Sale Price") +
  ggtitle("Relationship between year_remod and Sale Price") +
  scale_color_gradient(low = "blue", high = "red")


#H5 frontage vs sale_price
cor.test(new_data$sale_price, new_data$frontage, method = "pearson")
m5_frontage <- lm(new_data$sale_price ~ new_data$frontage)
summary(m5_frontage)

#plottig sactter plot
ggplot(new_data, aes(x = frontage, y = sale_price, color = sale_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "frontage", y = "Sale Price") +
  ggtitle("Relationship between frontage and Sale Price") +
  scale_color_gradient(low = "blue", high = "red")

class(new_data$house_quality)

hq <- new_data %>% group_by(house_quality) %>% summarise(avg_sale_price =
mean(sale_price))
hq

ggplot(hq)+
  geom_bar(aes(x = house_quality, y = avg_sale_price, fill = house_quality), stat = "identity")

hq1 <- new_data %>% group_by(neighbourhood) %>% summarise(avg_sale_price =
mean(sale_price))
hq1

#divide the clean data into train and test
library(caret)
set.seed(40412492)
```

```
index <-createDataPartition(new_data$sale_price,times = 1 , p=0.7 , list = F)
train_data <- new_data[index,]
test_data <- new_data[-index,]

summary(test_data$neighbourhood)

#model testing

#model building using forward approcah

#model1 using hpyothesis vairables

model_1 <-lm(sale_price ~ lot_area + neighbourhood + rooms_tot + year_remod +
frontage,data = train_data)
summary(hypothesis_based_model)

predictions_hypo_model <- predict(hypothesis_based_model,newdata = test_data)
postResample(predictions_hypo_model,test_data$sale_price)

#model2 using literature backed variables+hypo
model_2 <-lm(sale_price ~lot_area + neighbourhood + rooms_tot + year_remod + frontage+
year_built + zone + half_bath + full_bath + bedroom + aircon,data = train_data)
summary(lit_based_model)

predictions_lit_model <- predict(lit_based_model,newdata = test_data)
postResample(predictions_lit_model,test_data$sale_price)


#model 3 including all variables
model_3 <- lm(sale_price ~ lot_area+ zone + neighbourhood + frontage + year_built +
year_remod+half_bath+full_bath+bedroom+kitchen+foundations+stories+rooms_tot+airco
n+heat_type+house_quality+house_condition,data = train_data)
summary(all_var_model)

predictions_all_var_model <- predict(all_var_model,newdata = test_data)
postResample(predictions_all_var_model,test_data$sale_price)

plot(all_var_model)


plot(model_3, which = 2)

install.packages("lmtest")
library(lmtest)
install.packages('car')
library(car)
```

```
vif(all_var_model)
plot(all_var_model)
dwtest(all_var_model)
```