Housing Price Prediction Project

Objective of Analysis

The objective is to identify factors significantly affecting house prices through multiple linear regression analysis and goodness of fit tests, considering square footage, number of bedrooms and bathrooms, neighborhood, and construction year. The data encompasses various residential property records, including size, bedroom and bathroom count, location (urban, suburban, rural), construction year, and price. The project undertakes multiple linear regression to quantify relationships between house prices and various features, with a focus on the independence of categories such as bedrooms versus neighborhood type, and the difference between observed and expected housing price frequencies.

Data Description

The dataset consists of residential property records. Key variables include size (Square Feet), Bedroom and Bathroom count, Location (categorized into Urban, Suburban, and Rural neighborhoods), Year of Construction, and House Price. The dataset includes 50,000 observations, with houses varying in size from 1,000 to 2,999 square feet, having 1 to 5 bedrooms and bathrooms, and constructed between 1950 and 2021. House prices range significantly, with the noted presence of outliers and unusual values.

Muhammad Bin Imran. (n.d.). Housing Price Prediction Data. Kaggle. Retrieved [Retrieved, November 29th, 2023], from https://www.kaggle.com/datasets/muhammadbinimran/housing-price-prediction-data

Questions

1. Which categories affect house price?
2. Are the following categories independent of one another?
    a. Bedrooms and Neighborhood type
    b. Year-Built and Neighborhood type
    c. Number of Bedrooms and Bathrooms
    d. Square Feet and Year Built
3. Are observed frequencies of housing prices are significantly different from the expected frequencies based on the distribution of square feet, bedrooms, bathrooms, neighborhood, and year built?

Methodology

1. Multiple linear regression is employed to quantify the relationship between house prices (response variable) and a set of independent variables (predictors). The categorical variable Neighborhood was converted into dummy variables with 'Rural' serving as the reference group.

2. The Chi-Square Test of Independence is employed to assess if the categories are independent.

    a. The calculated Chi-Square statistic is compared to the Chi-Square distribution with degrees of freedom equal to $df = (\#rows - 1) \times (\#columns - 1)$

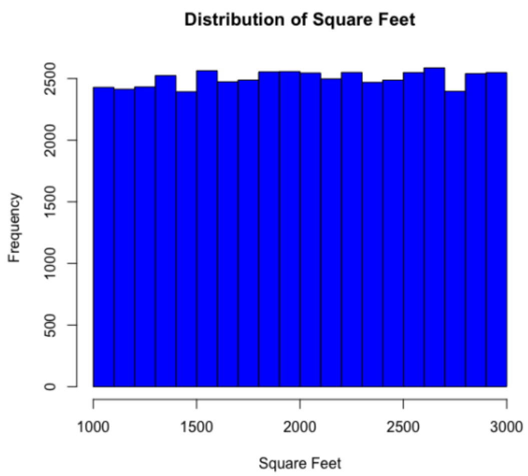
3. A chi-square test of independence was performed for each house characteristic against house price ranges.
    a. **Categorization**: Continuous variables (House Prices, Square Feet, Year Built) were categorized into discrete groups. House prices were divided into 'Low', 'Medium', and

'High'; Square Feet were categorized into 'Small', 'Medium', and 'Large'; Year Built was divided into four periods: '1970 and before', '1971-1990', '1991-2010', and '2011 and after'.

b. **Contingency Tables**: For each characteristic, contingency tables were constructed, comparing the observed frequency of price ranges across the categories of that characteristic.

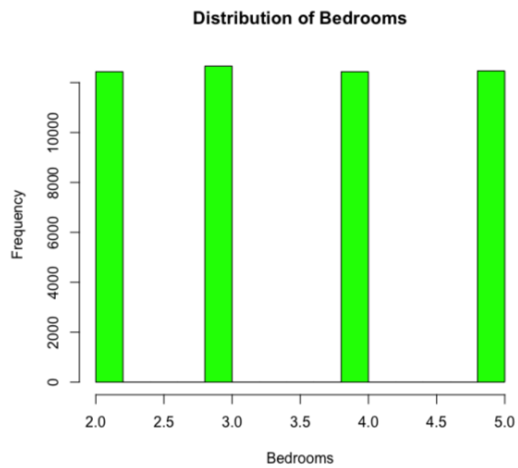
We begin data analysis with simple summary statistics before fitting a statistical model and preforming goodness of fit tests.

Figure 1 Distribution Plot of Square Feet



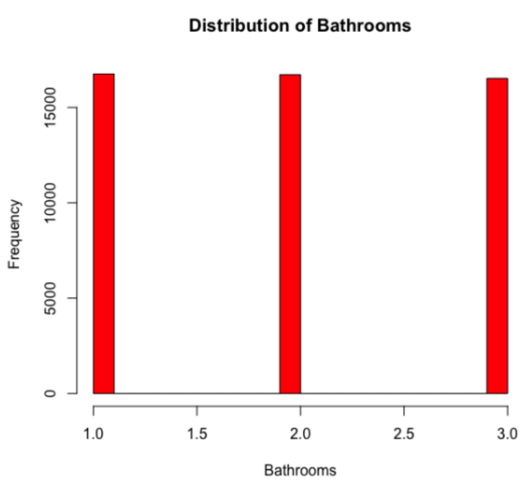
Note. This histogram displays the frequency distribution of the square footage of houses.

Figure 2 Distribution Plot of Bedrooms



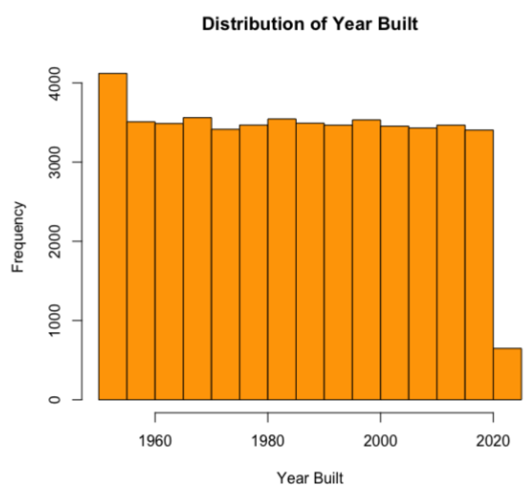
Note. The bar chart illustrates the number of houses distributed by the count of bedrooms.

Figure 3 Distribution Plot Bathrooms



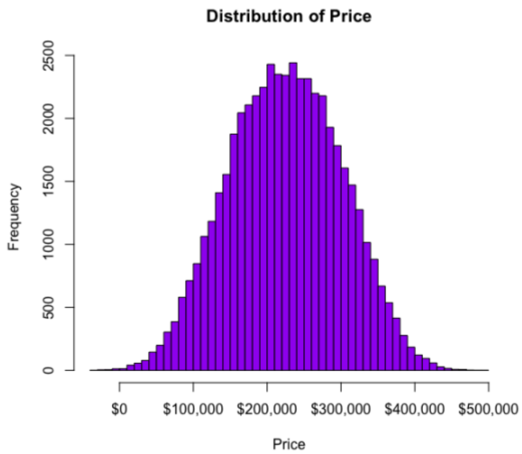
Note. This bar chart shows how many houses fall into each category based on the number of bathrooms.

Figure 4 Distribution Plot of Year Built



Note. The histogram indicates the number of houses built across different years.

Figure 5 Distribution Plot of Price



Note. This histogram represents the distribution of house prices within the dataset.

Table 1

Descriptive Statistics of Housing Data

Variable	N	Min	Q1	Median	Mean	Q3	Max
House	50,000	1	12,501	25,000	25,000	37,500	50,000
Square Feet	50,000	1,000	1,513	2,007	2,006	2,506	2,999
Bedrooms	50,000	2	3	3	3.499	4	5
Bathrooms	50,000	1	1	2	1.995	3	3
Year Built	50,000	1,950	1,967	1,985	1,985	2,003	2,021
Price	50,000	-36,588	169,956	225,052	224,827	279,374	492,195

Note. N = Sample size. Q1 = First quartile. Q3 = Third quartile. Price is reported in USD.

Summary Statistics

- Square Feet: The average size of the houses is approximately 2006 square feet, with a range from 1000 to 2999 square feet.
- Bedrooms: Houses typically have 3 to 4 bedrooms.
- Bathrooms: The number of bathrooms ranges from 1 to 3, with an average close to 2.
- Year Built: The houses were built between 1950 and 2021, with a mean construction year around 1985.
- Price: The average price is about $224,827, though there is significant variation, as indicated by the standard deviation of $76,142. Notably, the minimum value is negative, which might be an error or outlier.

Distribution Plots

- **Figure 1:** The distribution appears uniform across the range, but it is concentrated around the 1,500 to 2,500 square foot range.
- **Figure 2:** Most houses have 3 or 4 bedrooms, with a smaller number having 2 or 5.
- **Figure 3:** The distribution is somewhat evenly split between 1, 2, and 3 bathrooms.

- **Figure 3:** There's a relatively uniform distribution across the years, with no significant peaks or troughs.
- **Figure 4:** The count of houses varies across different neighborhoods, but the exact distribution is not clear.
- **Figure 5:** The price distribution is somewhat bell-shaped but shows some skewness towards the higher values.

Observations

The data seems to cover a wide range of house sizes, ages, and prices. But the presence of a negative house price suggests a need for data cleaning or further investigation. Negative values for house prices are not realistic in real-world scenarios, which implies that these values might be typos or other mistakes during data collection or entry. Since there are only 22 observations with a negative price data cleaning is appropriate as their removal is unlikely to significantly impact the analysis.

After manually cleansing the data, we have new summary of statistics and distribution plots. However, now the lowest price of a house is $154.78 which is quite unusual and raises several questions. Such a low price for a house is highly unlikely under normal market conditions in the U.S. The resource of this dataset has mentioned that it is simulated, and it is from across the world. Therefore, I will keep low priced houses because in some countries it is a possibility for houses to have a very low selling price.

Table 2

Descriptive Statistics of Housing Clean Data

Variable	N	Minimum	Q1	Median	Mean	Q3	Maximum
House	49,978	1	12,495	24,990	24,990	37,484	49,978
Square Feet	49,978	1,000	1,514	2,008	2,007	2,506	2,999
Bedrooms	49,978	2	3	3	3.499	4	5
Bathrooms	49,978	1	1	2	1.995	3	3
Year Built	49,978	1,950	1,967	1,985	1,985	2,003	2,021
Price (USD)	49,978	$154.8	$170,007.5	$225,100.1	$224,931.7	$279,395.8	$492,195.3

Note. N = Sample size. Q1 = First quartile. Q3 = Third quartile. Price is reported in USD.

Square Feet:

- The average size of a house is about 2,006.75 square feet.
- The standard deviation is 575.35, indicating moderate variability in house sizes.
- The smallest house is 1,000 square feet, and the largest is 2,999 square feet.

Bedrooms:

- The average number of bedrooms is approximately 3.5.
- Majority of houses have between 2 to 5 bedrooms.

Bathrooms:

- The average number of bathrooms is close to 2.
- The standard deviation is 0.815, suggesting some variation but generally close to the mean.

Year Built:

- The houses were built between 1950 and 2021.
- The mean year of construction is 1985, indicating a mix of older and newer properties.

Price:

- The average price of a house is approximately $224,931.67.
- The standard deviation is $75,995.68, indicating a wide range in house prices.
- The minimum price is $154.78, which might indicate special cases such as data being collected from several countries.
- The maximum price is $492,195.26.

Houses in each Neighborhood:

- Suburb: 16,716 houses
- Rural: 16,668 houses
- Urban: 16,594 houses
- This shows an even distribution of houses among the three neighborhoods.

Which categories affect house prices?

To determine which factors, affect the price of houses in the dataset, a multiple linear regression analysis is appropriate. This statistical method can help identify the relationship between the price (dependent variable) and other predictors of the houses, such as square footage, number of bedrooms, number of bathrooms, neighborhood, and year built.

Multiple Linear Regression Model:

$Y = 23764.1926 + 99.1681X_1 + 5080.7529X_2 + 2826.1022X_3 - 705.1247X_4 + 1562.2231X_5 - 10.8474X_6$

Table 3

Linear Regression Model Predicting House Prices

Predictor	B (SE)	β	t-value	p	95% CI
Intercept	23764.19 (21389.69)		1.11	.26657	
Square Feet (X_1)	99.17 (0.39)	.25	255.97	< .0001	[98.40, 99.93]
Bedrooms (X_2)	5080.75 (199.66)	.12	25.45	< .0001	[4683.68, 5477.82]
Bathrooms (X_3)	2826.10 (273.20)	.10	10.34	< .0001	[2283.79, 3368.41]
Neighborhood (Suburb) (X_4)	-705.12 (545.43)	-.01	-1.29	.19609	[-1785.42, 375.17]
Neighborhood (Urban) (X_5)	1562.22 (546.45)	.03	2.86	.00425	[488.59, 2635.86]
Year Built (X_6)	-10.85 (10.75)	-.01	-1.01	.31331	[-31.91, 10.22]

Note. *p* < .05. *p* values represent two-tailed tests. The reference category for neighborhood is rural. B = unstandardized regression coefficient; SE = standard error; β = standardized coefficient; CI = confidence interval.

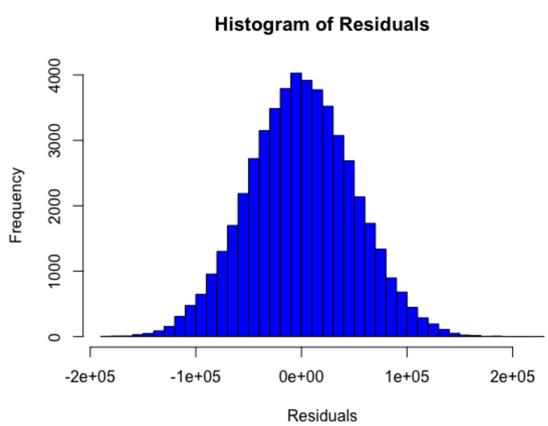
Linear Regression Model Analysis

- Square Feet: Every additional square foot is associated with an increase of approximately $99.16 in the house price. This predictor is highly significant due to p-value < 0.0001.
- Bedrooms: Each additional bedroom increases the house price by about $5,080.75, which is statistically significant (p-value < 0.0001).
- Bathrooms: An additional bathroom is associated with an increase of approximately $2,826.10 in the house price, with high statistical significance (p-value < 0.0001
- Neighborhood:
    - Houses in suburban neighborhoods are priced $705.12 less than those in rural areas, but this is not statistically significant (p-value = 0.19609).
    - Urban houses are priced $1,562.22 higher on average compared to rural houses, with this effect being significant (p-value = 0.00425).
- Year Built: The year a house was built has an insignificant negative association with its price (p-value = 0.31331).

The model explains about 57.02% of the variance in house prices (R-squared = 0.5702). Also, F-statistic is highly significant (p-value < 0.0001), indicating that the overall regression model is statistically significant.

Figure 6

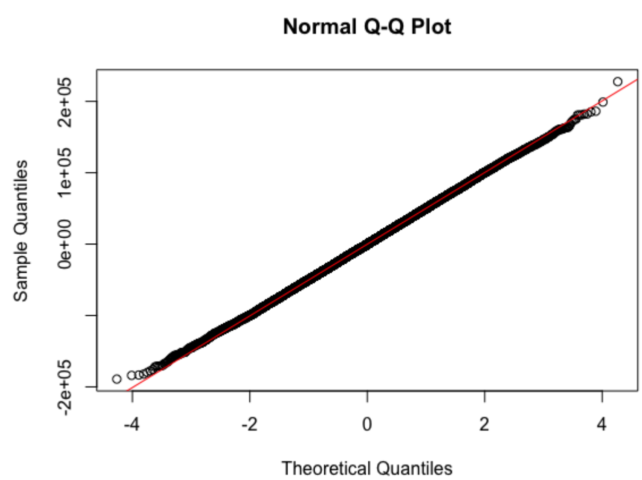
Histogram of Residuals from the Linear Regression Model



Note. The histogram displays the frequency distribution of residuals from the linear regression model, indicating the differences between observed and predicted prices.

Figure 7

Normal Q-Q plot of Residuals from the Linear Regression Model



Note. The Q-Q plot assesses the normality of residuals by comparing their distribution to a theoretical normal distribution. Points closely following the reference line suggest normality.

The histogram in **Figure 6** suggests a bell-shaped distribution, but the tails seem to be a bit heavy, especially on the right side, indicating the presence of outliers or extreme values that the model does not predict well.

Q-Q plot in **Figure 7** shows that the points generally follow the line, but there are deviations at the ends, particularly in the upper right and lower left, which indicates the presence of larger than expected residuals (outliers).

The residuals ranged significantly, suggesting the presence of outliers or that the model does not fully capture all factors influencing house prices. Since the dataset contains housing data from around the world, several factors could contribute to the large range in residuals and the significant variation in house prices. Factors such as market differences, currency fluctuations, and economic conditions.

In conclusion, the size of a house and the number of bedrooms and bathrooms are the primary factors contributing to the price of a house. The location also plays a role, with urban houses generally being more expensive than rural ones. Year of construction does not have a significant impact on price within this dataset. The model appears to have a reasonable fit to the data overall, but the issues highlighted by the residual analysis suggest that there may be room for improvement.

Since the number of bedrooms, bathrooms and neighborhood are the primary factors contributing to the price of a house, then the price of the house is expected to be dependent on those predictors. But year of construction is expected to be independent because the coefficient is so small that it does not have a significant effect on the house price.

The results indicate that there is a real estate market where property values are highly influenced by the number of bedrooms, bathrooms, and type of neighborhood. It is implied that newer homes may not always cost more just because of their age because the construction year has little impact. This might encourage purchasers to place greater emphasis on desired features and location rather than the year of construction. To meet pricing expectations and market demands, this emphasizes for sellers how important it is to prioritize these essential features over modernity.

Are categories independent of one another?

First, let’s begin by testing independence between two predictors: Number of Bedrooms and Neighborhood

Null Hypothesis (H_0): There is no significant relationship between Number of Bedrooms and Neighborhood types.

Alternative Hypothesis (H_a): There is a significant relationship between Number of Bedrooms and Neighborhood types.

Table 4

Observed Frequencies of Houses by Number of Bedrooms and Neighborhood Type

Bedrooms	Rural	Suburb	Urban
2	4117	4137	4177
3	4170	4306	4180
4	4199	4157	4073

5	4182	4116	4164

Note. The table shows the distribution of houses based on the number of bedrooms and the type of neighborhood (Rural, Suburb, Urban).

Table 5

Expected Frequencies of Houses by Number of Bedrooms and Neighborhood Type

Bedrooms	Rural (Expected)	Suburb (Expected)	Urban (Expected)
2	4145.82	4157.76	4127.42
3	4220.86	4233.02	4202.12
4	4145.16	4157.09	4126.75
5	4156.16	4168.13	4137.71

Note. This table presents the expected frequencies of houses in each category, calculated under the assumption that the number of bedrooms and neighborhood type are independent variables.

The Chi-Square test was conducted to examine the independence between the number of bedrooms in houses and their neighborhood types. The test yielded a Chi-Square statistic of $\chi^2(6)$ = 5.27 with 6 degrees of freedom. The calculated Chi-Square statistic is compared to the Chi-Square distribution with degrees of freedom equal to $df\ =\ (4-1)\times(3-1)=6$. The computed critical value is $\chi^2_{0.05,6}=$ 1.635383 which is less then 5.27. Therefore, fail to reject the null hypothesis, suggesting no significant association (i.e. independence)

The associated p-value was 0.51. This result also indicates that there is no significant relationship between the number of bedrooms and the type of neighborhood, suggesting that these two variables are independent.

Next two predictors that will be tested for independence are: Year-Built and Neighborhood

Null Hypothesis (H_0): There is no significant relationship between Year-Built and Neighborhood type.

Alternative Hypothesis (H_a): There is a significant relationship between Year-Built and neighborhood type

Table 6

Observed Frequencies of Houses by Year Built Range and Neighborhood Type

Year Built Range	Rural	Suburb	Urban
Pre-1980	6954	6970	6962
1980-2000	4696	4668	4637
Post-2000	5018	5078	4995

Note. The table shows the distribution of houses based on the year-built range (pre-1980, 1980-2000, post-2000) and the neighborhood type (Rural, Suburb, Urban).

Table 7

Expected Frequencies of Houses by Year Built Range and Neighborhood Type

Year Built Range	Rural (Expected)	Suburb (Expected)	Urban (Expected)
Pre-1980	6965.62	6985.68	6934.70
1980-2000	4669.43	4682.87	4648.70
Post-2000	5032.95	5047.44	5010.61

Note. This table presents the expected frequencies of houses in each category, calculated under the assumption that the year-built range and neighborhood type are independent variables.

The p-value of 0.9552 is significantly higher than the alpha level of 0.05. This indicates that there is no statistically significant relationship between the year the house was built and the neighborhood. The Chi-Square statistic of $\chi^2_{0.05,4} = 0.668$, with 4 degrees of freedom, is greater than the critical value of 0.710723. Pearson's Chi-Square test suggests that the observed frequencies are very close to the expected frequencies under the assumption of independence. Therefore, it can be concluded that the year a house was built, and the neighborhood type are independent variables in this dataset.

Next two predictors that will be tested for independence are: Number of Bedrooms and Bathrooms.

Null Hypothesis (H_0): There is no significant relationship between Number of Bedrooms and Bathrooms.

Alternative Hypothesis (H_a): There is a significant relationship between Number of Bedrooms and Bathrooms.

Table 8

Observed Frequencies of Houses by Number of Bedrooms and Bathrooms

Bedrooms	Bathroom 1	Bathroom 2	Bathroom 3
2	4227	4180	4024
3	4262	4180	4214
4	4113	4181	4135
5	4145	4170	4147

Note. The table shows the observed number of houses with varying numbers of bedrooms and bathrooms.

Table 9

Expected Frequencies of Houses by Number of Bedrooms and Bathrooms

Bedrooms	Bathroom 1	Bathroom 2	Bathroom 3
2	4165.47	4156.52	4109.01
3	4240.87	4231.75	4183.38
4	4164.80	4155.85	4108.35
5	4175.86	4166.88	4119.26

Note. The table presents the expected frequencies of houses with varying numbers of bedrooms and bathrooms.

The p-value (0.5248) is greater than the alpha level of 0.05, which suggests that there is not enough evidence to reject the null hypothesis. The Chi-Square statistic of $\chi^2_{0.05,6} = 5.1491$, with 6 degrees of freedom, is greater than the critical value of 1.635383. Pearson's Chi-Square test also suggests that the two categories are independent.

Next two predictors that will be tested for independence are: Square Feet and Year Built

Null Hypothesis (H_0): There is no significant relationship between Square Feet and Year Built.

Alternative Hypothesis (H_a): There is a significant relationship between Square Feet and Year Built.

Table 10

Observed Frequencies of Houses by Square Feet Category and Year Built Range

Square Feet Category	Pre-1980	1980–2000	Post-2000
Small	7137	4586	4780
Medium	7133	4675	4693
Large	7281	4769	4924

Note. The table shows the observed number of houses with varying square footage categories across different year-built ranges.

Table 11

Expected Frequencies of Houses by Square Feet Category and Year Built Range

Square Feet Category	Pre-1980	1980–2000	Post-2000
Small	7116.254	4632.780	4753.966
Medium	7115.392	4632.219	4753.389
Large	7319.354	4765.001	4889.645

Note. The table presents the expected frequencies of houses with varying square footage categories across different year-built ranges, assuming the square footage category and year-built range are independent.

The p-value of (0.6759) is greater than the alpha level of 0.05, which means we fail to reject the null hypothesis. Chi-Square statistic of $\chi^2_{0.05,4} = 2.327$, with 4 degrees of freedom, is greater than the critical value of 0.710723. Pearson's Chi-Square test also suggests that the two categories are independent.

Are observed frequencies of housing prices are significantly different from the expected frequencies based on the distribution of square feet, bedrooms, bathrooms, neighborhood, and year built?

Null Hypothesis (H_0): The observed frequencies of housing prices are consistent with the expected frequencies based on the distribution of square feet, bedrooms, bathrooms, neighborhood, year built.

Alternative Hypothesis (H_a): The observed frequencies differ significantly from the expected frequencies

To determine whether the observed frequencies of housing prices are significantly different from the expected frequencies based on the distribution of square feet, bedrooms, bathrooms, neighborhood, and year built, we will conduct a series of Chi-Square Tests of Independence for each pair of categories (price with each other variable).

However, since housing price is a continuous variable, they would first need to be categorized into discrete groups (e.g., Low, Medium, High) based on some criteria such as quantiles. Then we can create contingency tables for each pair (Price Range with Square Feet Category, Price Range with Bedrooms, etc.) and perform Chi-Square Tests. After categorizing house prices, below we can see results of contingency tables below:

Table 12

Observed Frequencies of House Prices by Square Footage Category

Square Footage	Low	Medium	High
Small	11,703	4,280	510
Medium	4,279	7,960	4,253
Large	521	4,261	12,211

Note. Observed counts of low, medium, and high-priced houses across small, medium, and large square footage categories.

Table 13

Expected Frequencies of House Prices by Square Footage Category

Square Footage	Low	Medium	High
Small	5446.08	5445.42	5601.51
Medium	5445.75	5445.09	5601.17
Large	5611.18	5610.50	5771.32

Note. Expected counts of low, medium, and high-priced houses across small, medium, and large square footage categories, assuming independence between price and square footage.

Table 14

Observed Frequencies of House Prices by Number of Bedrooms

Bedrooms	Low	Medium	High
2	4,537	4,024	3,870
3	4,374	4,261	4,021
4	3,920	4,119	4,390
5	3,662	4,088	4,712

Note. Observed counts of low, medium, and high-priced houses with different bedroom counts.

Table 15

Expected Frequencies of House Prices by Number of Bedrooms

Bedrooms	Low	Medium	High
2	4102.30	4176.55	4237.20
3	4102.05	4176.29	4237.20
4	4226.66	4303.16	4237.20
5	4101.64	4101.39	4127.27

Note. Expected counts of low, medium, and high-priced houses with different bedroom counts, assuming independence between price and number of bedrooms.

Table 16

Observed Frequencies of House Prices by Number of Bathrooms

Price Range	1 Bathroom	2 Bathrooms	3 Bathrooms
Low	5752	5459	5282
Medium	5512	5529	5451
High	5483	5723	5787

Note. The table shows the observed number of low, medium, and high-priced houses with one, two, and three bathrooms.

Table 17

Expected Frequencies of House Prices by Number of Bathrooms

Price Range	1 Bathroom	2 Bathrooms	3 Bathrooms
Low	5526.60	5514.72	5451.69
Medium	5526.26	5514.38	5451.36
High	5694.14	5681.90	5616.96

Note. The table presents the expected number of low, medium, and high-priced houses with one, two, and three bathrooms, assuming independence between price range and number of bathrooms.

Due to the extensive range of years covered in the dataset, the observed and expected frequencies for house prices by year built produce a very large contingency table. Including this table in its entirety within the report would hinder readability and clarity. Therefore, I have summarized the findings here and have provided a comprehensive table in an appendix.

<u>Square Footage vs. Price Range</u>: A highly significant Chi-Square statistic of 25,929 with 4 degrees of freedom and a p-value less than 0.0001 indicates a strong relationship between the size of the house and its price. (fail to reject the null hypothesis)

<u>Bathrooms vs. Price Range:</u> A Chi-Square statistic of 28.387 with a p-value of 0.00001 suggests a significant relationship between the number of bathrooms and house price.

Bedrooms vs. Price Range: With a Chi-Square statistic of 224.39 and a p-value less than 0.0001, there is a significant relationship between the number of bedrooms and house price. (fail to reject the null hypothesis)

Neighborhood vs. Price Range: The Chi-Square statistic of 22.349 and a p-value of 0.0001708 demonstrate a significant association between neighborhood and house price. (fail to reject the null hypothesis)

Year Built vs. Price Range: A Chi-Square statistic of 121.13 with 142 degrees of freedom and a p-value of 0.897 indicates no significant relationship between the year a house was built and its price range, suggesting these variables are independent. (reject the null hypothesis)

The analysis indicates that the observed frequencies of housing prices are significantly different from the expected frequencies when considering the distributions of square footage, number of bedrooms, number of bathrooms, and neighborhood type. However, the year a house was built does not appear to be associated with its price range. Multiple linear regression model has also confirmed these findings, since its coefficients indicated which predictors had the highest impact on house prices. Predictors that had the most significant impact, resulted in being dependent and vice versa.

These conclusions imply that in the housing market, certain features like square footage, bedroom and bathroom count, and neighborhood type are crucial in influencing house prices. The lack of a significant relationship between the year built and price suggests that buyers may prioritize current condition or location over age. Real estate stakeholders might consider emphasizing and investing in these key features to improve property value.

Future questions/ways to analyze the data:

How does the age of a house impact its price, considering renovations or historic value?

Is there a premium on prices in certain neighborhoods, and is it justified by the amenities available?

What factors contribute to the existence of high-priced homes in the lower square footage category?

Time Series Analysis: analyze the pricing trends over time and incorporate macroeconomic indicators.

Interaction Effects: study how combinations of features, like size and location, interact to predict prices.

Appendix (R codes and Outputs)

```
> hist(data$Price, main="Distribution of Price", xlab="Price", col="purple",
+      breaks=50,xaxt='n');
> axis(side=1, at=axTicks(side=1), labels=scales::dollar(axTicks(side=1)));
> print(summary_stats);
     House         SquareFeet      Bedrooms       Bathrooms     Neighborhood         YearBuilt
 Min.   :    1   Min.   :1000   Min.   :2.000   Min.   :1.000   Length:50000       Min.   :1950
 1st Qu.:12501   1st Qu.:1513   1st Qu.:3.000   1st Qu.:1.000   Class :character   1st Qu.:1967
 Median :25000   Median :2007   Median :3.000   Median :2.000   Mode  :character   Median :1985
 Mean   :25000   Mean   :2006   Mean   :3.499   Mean   :1.995                      Mean   :1985
 3rd Qu.:37500   3rd Qu.:2506   3rd Qu.:4.000   3rd Qu.:3.000                      3rd Qu.:2003
 Max.   :50000   Max.   :2999   Max.   :5.000   Max.   :3.000                      Max.   :2021
     Price
 Min.   :-36588
 1st Qu.:169956
 Median :225052
 Mean   :224827
 3rd Qu.:279374
 Max.   :492195
> library(readxl)
> library(ggplot2)
> data <- read_excel("/Users/kamala/Desktop/houses.xlsx")

> summary_stats <- summary(data)
> hist(data$SquareFeet, main="Distribution of Square Feet", xlab="Square Feet", col="blue")
> hist(data$Bedrooms, main="Distribution of Bedrooms", xlab="Bedrooms", col="green")
> hist(data$Bathrooms, main="Distribution of Bathrooms", xlab="Bathrooms", col="red")
> hist(data$YearBuilt, main="Distribution of Year Built", xlab="Year Built", col="orange")
> cleandata <- read_excel("/Users/kamala/Desktop/houses.xlsx")

> summary_stats <- summary(cleandata)
> print(summary_stats)
     House         SquareFeet      Bedrooms       Bathrooms     Neighborhood         YearBuilt
 Min.   :    1   Min.   :1000   Min.   :2.000   Min.   :1.000   Length:49978       Min.   :1950
 1st Qu.:12495   1st Qu.:1514   1st Qu.:3.000   1st Qu.:1.000   Class :character   1st Qu.:1967
 Median :24990   Median :2008   Median :3.000   Median :2.000   Mode  :character   Median :1985
 Mean   :24990   Mean   :2007   Mean   :3.499   Mean   :1.995                      Mean   :1985
 3rd Qu.:37484   3rd Qu.:2506   3rd Qu.:4.000   3rd Qu.:3.000                      3rd Qu.:2003
 Max.   :49978   Max.   :2999   Max.   :5.000   Max.   :3.000                      Max.   :2021
     Price
 Min.   :   154.8
 1st Qu.:170007.5
 Median :225100.1
 Mean   :224931.7
 3rd Qu.:279395.8
 Max.   :492195.3
> houses_per_neighborhood <- table(cleandata$Neighborhood)
> print(houses_per_neighborhood)

 Rural Suburb  Urban
 16668  16716  16594
```

```
> cleandata$Neighborhood <- as.factor(cleandata$Neighborhood)
> model <- lm(Price ~ SquareFeet + Bedrooms + Bathrooms + Neighborhood + YearBuilt, data = cleandata)
>
> summary(model)

Call:
lm(formula = Price ~ SquareFeet + Bedrooms + Bathrooms + Neighborhood +
    YearBuilt, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-189092  -34031    -230   33695  227603

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         23764.1926 21389.6938   1.111  0.26657
SquareFeet             99.1681     0.3874 255.965  < 2e-16 ***
Bedrooms             5080.7529   199.6688  25.446  < 2e-16 ***
Bathrooms            2826.1022   273.2041  10.344  < 2e-16 ***
NeighborhoodSuburb   -705.1247   545.4287  -1.293  0.19609
NeighborhoodUrban    1562.2231   546.4529   2.859  0.00425 **
YearBuilt             -10.8474    10.7580  -1.008  0.31331
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49830 on 49971 degrees of freedom
Multiple R-squared:  0.5702,	Adjusted R-squared:  0.5701
F-statistic: 1.105e+04 on 6 and 49971 DF,  p-value: < 2.2e-16
> residuals <- residuals(model)
> library(ggplot2)
> ggplot(data.frame(residuals), aes(x = residuals)) +
+     geom_density(fill = "blue", alpha = 0.5) +
+     labs(title = "Density Plot of Residuals from the Linear Regression Model",
+          x = "Residuals",
+          y = "Density") +
+     theme_minimal()
> residuals <- residuals(model)
> hist(residuals, main="Histogram of Residuals", xlab="Residuals", col="blue", breaks=50)
> qqnorm(residuals)
> qqline(residuals, col="red")
> contingency_table <- table(cleandata$Bedrooms, cleandata$Neighborhood)
> chi_square_test <- chisq.test(contingency_table)
> print(chi_square_test)

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 5.2666, df = 6, p-value = 0.5101

> print(contingency_table)

    Rural Suburb Urban
  2  4117   4137  4177
  3  4170   4306  4180
  4  4199   4157  4073
  5  4182   4116  4164


> critical_value <- qchisq(1 - 0.05, 6, lower.tail = FALSE)
> print(critical_value)
[1] 1.635383
```

```
> cleandata$PriceRange <- cut(cleandata$Price, breaks=quantile(cleandata$Price, probs=0:3/3),
include.lowest=TRUE, labels=c("Low", "Medium", "High"))
> cleandata$SquareFeetRange <- cut(cleandata$SquareFeet, breaks=quantile(cleandata$SquareFeet,
probs=0:3/3), include.lowest=TRUE, labels=c("Small", "Medium", "Large"))
> cleandata$YearBuiltRange <- cut(cleandata$YearBuilt, breaks=c(min(cleandata$YearBuilt), 197
0, 1990, 2010, max(cleandata$YearBuilt)), include.lowest=TRUE, labels=c("1970 and before", "19
71-1990", "1991-2010", "2011 and after"))
>
> chi_square_results <- lapply(c("SquareFeetRange", "Bedrooms", "Bathrooms", "Neighborhood",
"YearBuiltRange"), function(x) {
+     contingency_table <- table(cleandata[[x]], cleandata$PriceRange)
+     chisq.test(contingency_table)
+ })
> print(chi_square_results)
[[1]]

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 26065, df = 4, p-value < 2.2e-16


[[2]]

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 222.5, df = 6, p-value < 2.2e-16


[[3]]

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 29.653, df = 4, p-value = 5.759e-06


[[4]]

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 21.274, df = 4, p-value = 0.0002795


[[5]]

	Pearson's Chi-squared test

data:  contingency_table
X-squared = 1.5715, df = 6, p-value = 0.9546
```

```
> square_feet_bins <- quantile(data$SquareFeet, probs = c(0, 0.33, 0.66, 1))
> data$SquareFeetCategory <- cut(data$SquareFeet, breaks = square_feet_bins, include.lowest = TRUE,
+                                labels = c("Small", "Medium", "Large"))
>
> year_built_bins <- c(0, 1980, 2000, Inf)
> data$YearBuiltCategory <- cut(data$YearBuilt, breaks = year_built_bins, include.lowest = TRUE,
+                               labels = c("Pre-1980", "1980-2000", "Post-2000"))
>
> contingency_table_sf_yb <- table(data$SquareFeetCategory, data$YearBuiltCategory)
>
> chi_square_test_sf_yb <- chisq.test(contingency_table_sf_yb)
> print(chi_square_test_sf_yb)

	Pearson's Chi-squared test

data:  contingency_table_sf_yb
X-squared = 2.327, df = 4, p-value = 0.6759

> observed_frequencies <- chi_square_test_sf_yb$observed
> expected_frequencies <- chi_square_test_sf_yb$expected
> print(observed_frequencies)

         Pre-1980 1980-2000 Post-2000
  Small      7137      4586      4780
  Medium     7133      4675      4693
  Large      7281      4769      4924
> print(expected_frequencies)

         Pre-1980 1980-2000 Post-2000
  Small  7116.254  4632.780  4753.966
  Medium 7115.392  4632.219  4753.389
  Large  7319.354  4765.001  4889.645
```

```
> price_bins <- quantile(data$Price, probs = c(0, 0.33, 0.66, 1))
> data$PriceCategory <- cut(data$Price, breaks = price_bins, include.lowest = TRUE,
+                           labels = c("Low", "Medium", "High"))
> contingency_table_price_sf <- table(data$PriceCategory, data$SquareFeetCategory)
> contingency_table_price_bed <- table(data$PriceCategory, data$Bedrooms)
> contingency_table_price_bath <- table(data$PriceCategory, data$Bathrooms)
> contingency_table_price_yearbuilt <- table(data$PriceCategory, data$YearBuilt)
> contingency_table_price_neighb <- table(data$PriceCategory, data$Neighborhood)
> chi_square_test_price_sf <- chisq.test(contingency_table_price_sf)
> chi_square_test_price_bed <- chisq.test(contingency_table_price_bed)
> chi_square_test_price_bath <- chisq.test(contingency_table_price_bath)
> chi_square_test_price_yearbuilt <- chisq.test(contingency_table_price_yearbuilt)
> chi_square_test_price_neighb <- chisq.test(contingency_table_price_neighb)
> print(chi_square_test_price_sf)

        Pearson's Chi-squared test

data:  contingency_table_price_sf
X-squared = 25929, df = 4, p-value < 2.2e-16

> print(chi_square_test_price_bath)

        Pearson's Chi-squared test

data:  contingency_table_price_bath
X-squared = 28.387, df = 4, p-value = 1.041e-05

> print(chi_square_test_price_bed)

        Pearson's Chi-squared test

data:  contingency_table_price_bed
X-squared = 224.39, df = 6, p-value < 2.2e-16

> print(chi_square_test_price_neighb)

        Pearson's Chi-squared test

data:  contingency_table_price_neighb
X-squared = 22.349, df = 4, p-value = 0.0001708
```

```
> print(chi_square_test_price_yearbuilt)

	Pearson's Chi-squared test

data:  contingency_table_price_yearbuilt
X-squared = 121.13, df = 142, p-value = 0.897

> observed_frequencies <- chi_square_test_price_sf$observed
> print(observed_frequencies)

         Small Medium Large
  Low    11703   4280   510
  Medium  4279   7960  4253
  High     521   4261 12211
> expected_frequencies <- chi_square_test_price_sf$expected
> print(expected_frequencies)

            Small   Medium    Large
  Low    5446.076 5445.416 5601.508
  Medium 5445.746 5445.086 5601.169
  High   5611.178 5610.498 5771.323
> observed_frequencies <- chi_square_test_price_bed$observed
> expected_frequencies <- chi_square_test_price_bed$expected
> print( observed_frequencies)

            2    3    4    5
  Low    4537 4374 3920 3662
  Medium 4024 4261 4119 4088
  High   3870 4021 4390 4712
> print(expected_frequencies)

                2        3        4        5
  Low    4102.295 4176.546 4101.635 4112.525
  Medium 4102.046 4176.293 4101.386 4112.275
  High   4226.659 4303.162 4225.979 4237.200
> observed_frequencies <- chi_square_test_price_bath $observed
> expected_frequencies <- chi_square_test_price_bath $expected
> print(observed_frequencies); print(expected_frequencies)
```

```
         1    2    3
Low    5752 5459 5282
Medium 5512 5529 5451
High   5483 5723 5787

              1        2        3
Low    5526.597 5514.717 5451.686
Medium 5526.262 5514.383 5451.355
High   5694.141 5681.900 5616.959
> observed_frequencies <- chi_square_test_price_yearbuilt $observed
> expected_frequencies <- chi_square_test_price_yearbuilt $expected
> print(observed_frequencies); print(expected_frequencies)

       1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970
Low     218  228  249  222  271  207  245  218  220  219  212  218  226  243  232  231  216  220  238  236  236
Medium  233  221  248  232  198  229  214  232  239  223  242  247  210  236  235  232  230  234  251  214  235
High    192  249  225  230  235  230  263  240  257  239  243  228  218  233  257  240  236  275  288  241  211

       1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991
Low     221  229  215  212  239  230  263  253  241  199  226  212  239  246  236  220  236  234  233  218  230
Medium  244  223  236  230  221  219  244  216  236  234  234  213  235  225  227  250  218  223  229  246  242
High    239  223  225  215  242  209  234  226  231  232  233  261  266  232  257  242  253  220  240  228  227

       1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Low     242  216  224  229  234  233  252  249  227  207  247  223  239  240  206  245  213  233  224  233  215
Medium  223  222  238  229  240  234  224  245  227  201  226  237  226  251  211  244  216  235  219  210  205
High    231  243  230  241  229  242  221  234  240  213  226  258  224  235  222  229  239  237  256  260  217

       2013 2014 2015 2016 2017 2018 2019 2020 2021
Low     225  218  226  223  254  230  220  227  202
Medium  244  237  221  208  221  241  212  254  211
High    242  271  241  211  217  233  230  221  235

           1950     1951     1952     1953     1954     1955     1956     1957     1958     1959     1960
Low    212.1933 230.3436 238.2638 225.7236 232.3237 219.7835 238.2638 227.7036 236.2837 224.7335 230.0136
Medium 212.1805 230.3297 238.2493 225.7099 232.3096 219.7701 238.2493 227.6898 236.2694 224.7199 229.9997
High   218.6262 237.3267 245.4869 232.5666 239.3668 226.4464 245.4869 234.6066 243.4469 231.5465 236.9867

           1961     1962     1963     1964     1965     1966     1967     1968     1969     1970     1971
Low    228.6936 215.8234 234.9637 238.9238 231.9937 225.0635 240.5738 256.4140 228.0336 225.0635 232.3237
Medium 228.6797 215.8103 234.9495 238.9093 231.9796 225.0499 240.5592 256.3985 228.0198 225.0499 232.3096
High   235.6267 222.3663 242.0868 246.1670 239.0268 231.8866 247.8670 264.1875 234.9466 231.8866 239.3668

           1972     1973     1974     1975     1976     1977     1978     1979     1980     1981     1982
Low    222.7535 223.0835 216.8134 231.6637 217.1434 244.5339 229.3536 233.6437 219.4535 228.6936 226.3836
Medium 222.7400 223.0700 216.8003 231.6496 217.1303 244.5190 229.3397 233.6295 219.4402 228.6797 226.3698
High   229.5065 229.8465 223.3863 238.6867 223.7263 251.9471 236.3067 240.7268 226.1064 235.6267 233.2466

           1983     1984     1985     1986     1987     1988     1989     1990     1991     1992     1993
Low    244.2038 231.9937 237.6037 234.9637 233.3137 223.4135 231.6637 228.3636 230.6736 229.6836 224.7335
Medium 244.1890 231.9796 237.5893 234.9495 233.2995 223.4000 231.6496 228.3498 230.6597 229.6697 224.7199
High   251.6071 239.0268 244.8069 242.0868 240.3868 230.1865 238.6867 235.2866 237.6667 236.6467 231.5465

           1994     1995     1996     1997     1998     1999     2000     2001     2002     2003     2004
Low    228.3636 230.6736 231.9937 233.9737 230.0136 240.2438 229.0236 204.9332 230.6736 236.9437 227.3736
Medium 228.3498 230.6597 231.9796 233.9595 229.9997 240.2292 229.0097 204.9208 230.6597 236.9294 227.3598
High   235.2866 237.6667 239.0268 241.0668 236.9867 247.5270 235.9667 211.1460 237.6667 244.1269 234.2666

           2005     2006     2007     2008     2009     2010     2011     2012     2013     2014     2015
Low    239.5838 210.8733 236.9437 220.4435 232.6537 230.6736 231.9937 210.2133 234.6337 239.5838 227.0436
Medium 239.5693 210.8605 236.9294 220.4301 232.6396 230.6597 231.9796 210.2006 234.6195 239.5693 227.0298
High   246.8470 217.2661 244.1269 227.1264 239.7068 237.6667 239.0268 216.5861 241.7468 246.8470 233.9266

           2016     2017     2018     2019     2020     2021
Low    211.8633 228.3636 232.3237 218.4634 231.6637 213.8434
Medium 211.8505 228.3498 232.3096 218.4502 231.6496 213.8304
High   218.2862 235.2866 239.3668 225.0864 238.6867 220.3262
> observed_frequencies <- chi_square_test_price_neighb $observed
> expected_frequencies <- chi_square_test_price_neighb $expected
> print(observed_frequencies); print(expected_frequencies)

       Rural Suburb Urban
Low     5555   5659  5279
Medium  5479   5530  5483
High    5634   5527  5832

          Rural   Suburb    Urban
Low    5500.527 5516.367 5476.106
Medium 5500.193 5516.032 5475.774
High   5667.280 5683.601 5642.119
```