House Price Prediction Analysis

Jai Patel

2024-12-10

Contents

Abstract

This project looks at house sales in King County, USA, from May 2014 to May 2015 to understand what affects home prices. The data includes information about house square feet, condition, number of bedrooms, and more. The main finding is that bigger homes, especially those with more living space, usually cost more. The year a house was built doesn't have much effect on its price. The size of the above-ground space is important for price. Lot and basement size differ from one another. As the number of bathrooms increase, the price also goes up. Homes with more bedrooms tend to have better condition ratings. Even though the dataset has some limitations, the analysis gives useful insights into what influences home prices in the area.

Introduction

This project focuses on exploring housing trends using this dataset. Large datasets like this are important for understanding the real estate market, planning cities, and making financial decisions. As housing demand increases, it's important to use data to better understand market patterns and trends.

Researchers have used similar datasets in the past to study pricing trends, identify key property features, and predict how the market would change. However, new uses like machine learning features have improved visualization methods and offer a new way to see the change. This project will analyze the dataset to find patterns that affect housing prices, helping people make smarter decisions.

By examining factors like square feet of the house and condition, this project aims to provide clear, data-driven insights for real estate and property management. The goal is to use models and visualizations to turn this data into practical, useful information.

Data

The dataset contains 21,613 rows and 22 columns, containing many variables. The variables are `X`, `id`, `date`, `price`, `bedrooms`, `bathrooms`, `sqft_living`, `floors`, `waterfront`, `view`, `condition`, `grade`, `sqft_above`, `sqft_basement`, `yr_built`, `yr_renovated`, `zipcode`, `lat`, `long`, `sqft_living15`, and `sqft_lot15` (see Shiv28 [3]). The dataset focuses on house sales in King County, USA, including the city of Seattle. It covers transactions from May 2014 to May 2015, offering a idea of the housing market during that period. However, the dataset has some limits. It only covers one year, so it can't show long-term trends. It also focuses only on King County, so the results may not apply to other areas. Even with these challenges, the dataset is a good starting point for understanding what impacts home prices.

```
data = read.csv("../Documents/kc_house_data_NaN.csv")
summary(data)
```

```
##        X               id                date               price
##  Min.   :    0   Min.   :1.000e+06   Length:21613       Min.   :  75000
##  1st Qu.: 5403   1st Qu.:2.123e+09   Class :character   1st Qu.: 321950
##  Median :10806   Median :3.905e+09   Mode  :character   Median : 450000
##  Mean   :10806   Mean   :4.580e+09                      Mean   : 540088
##  3rd Qu.:16209   3rd Qu.:7.309e+09                      3rd Qu.: 645000
##  Max.   :21612   Max.   :9.900e+09                      Max.   :7700000
##
##     bedrooms        bathrooms      sqft_living       sqft_lot
##  Min.   : 1.000   Min.   :0.500   Min.   :  290   Min.   :    520
##  1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040
##  Median : 3.000   Median :2.250   Median : 1910   Median :   7618
##  Mean   : 3.373   Mean   :2.116   Mean   : 2080   Mean   :  15107
##  3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688
##  Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359
##  NA's   :13       NA's   :10
##      floors        waterfront           view          condition
##  Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000
##  Median :1.500   Median :0.000000   Median :0.0000   Median :3.000
##  Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409
##  3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000
##  Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000
##
```

```
##      grade           sqft_above   sqft_basement       yr_built
##  Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900
##  1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951
##  Median : 7.000   Median :1560   Median :   0.0   Median :1975
##  Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971
##  3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997
##  Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015
##
##   yr_renovated       zipcode           lat             long
##  Min.   :   0.0   Min.   :98001   Min.   :47.16   Min.   :-122.5
##  1st Qu.:   0.0   1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3
##  Median :   0.0   Median :98065   Median :47.57   Median :-122.2
##  Mean   :  84.4   Mean   :98078   Mean   :47.56   Mean   :-122.2
##  3rd Qu.:   0.0   3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1
##  Max.   :2015.0   Max.   :98199   Max.   :47.78   Max.   :-121.3
##
##  sqft_living15    sqft_lot15
##  Min.   : 399   Min.   :   651
##  1st Qu.:1490   1st Qu.:  5100
##  Median :1840   Median :  7620
##  Mean   :1987   Mean   : 12768
##  3rd Qu.:2360   3rd Qu.: 10083
##  Max.   :6210   Max.   :871200
##
```

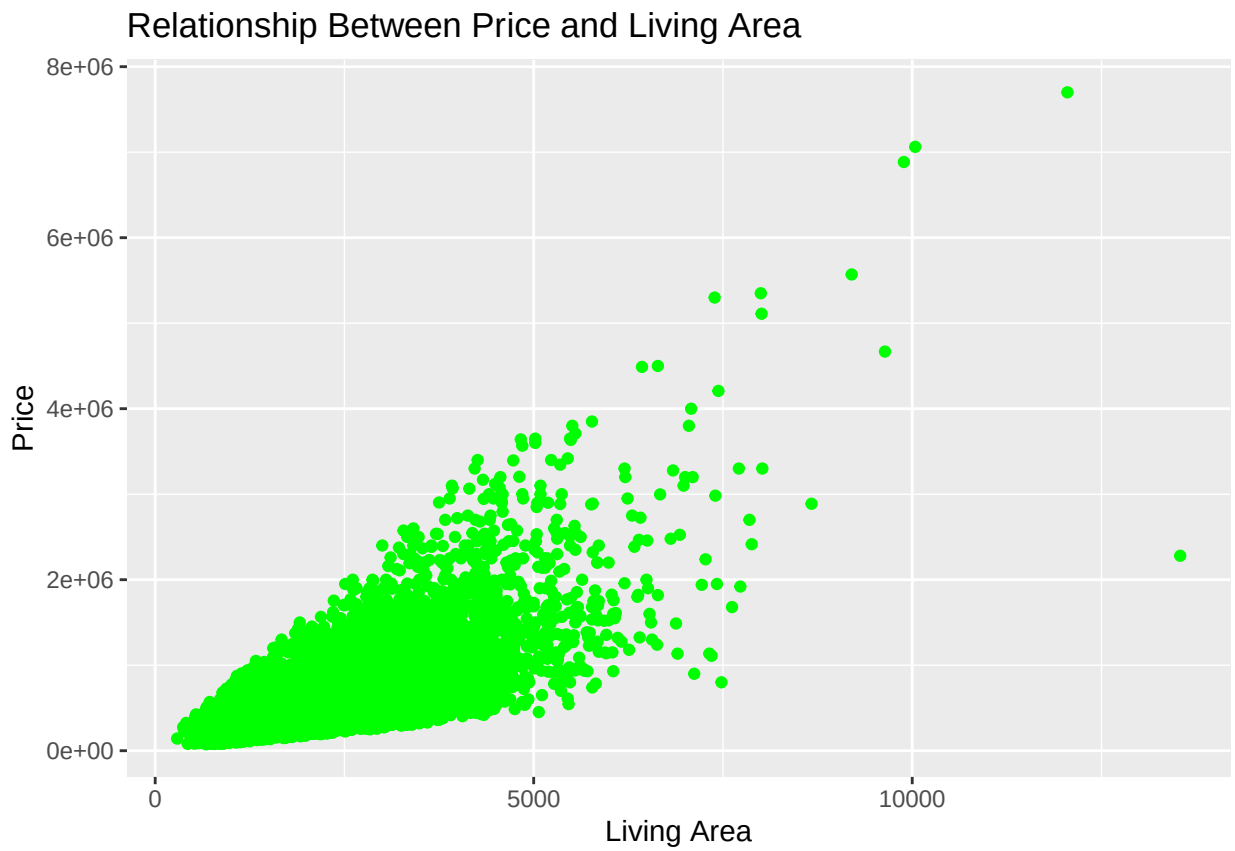
The above data is showing basic information of the categories that we can potentially see a relationship between.

Analysis

Price & Square Foot in Living Relationship

```
library(tidyverse)
library(ggplot2)

ggplot(data, aes(x = sqft_living, y = price)) + geom_point(color = "green") +
    labs(title = "Relationship Between Price and Living Area",
        x = "Living Area", y = "Price")
```




```
correlation = cor(data$sqft_living, data$price, use = "complete.obs")
paste0("Correlation is: ", correlation)
```

```
## [1] "Correlation is: 0.7020350546118"
```

Result

The scatter plot (see R Core Team [2]) shows that bigger houses with more square feet of living space usually cost more. There's a upward trend, but there are a few exceptions where smaller houses have high prices, probably because of things like location or special features like have a view with water.

The correlation (see Posit team [1]) between the size of the living area and price is 0.7020350546118, which means there's a strong positive relationship. This shows that as the size of the living area increases, the price also tends to go up.

This means bigger homes are generally more expensive, which makes sense in the housing market. This information is useful for understanding how size affects home prices.

Price & Year Built Relationship

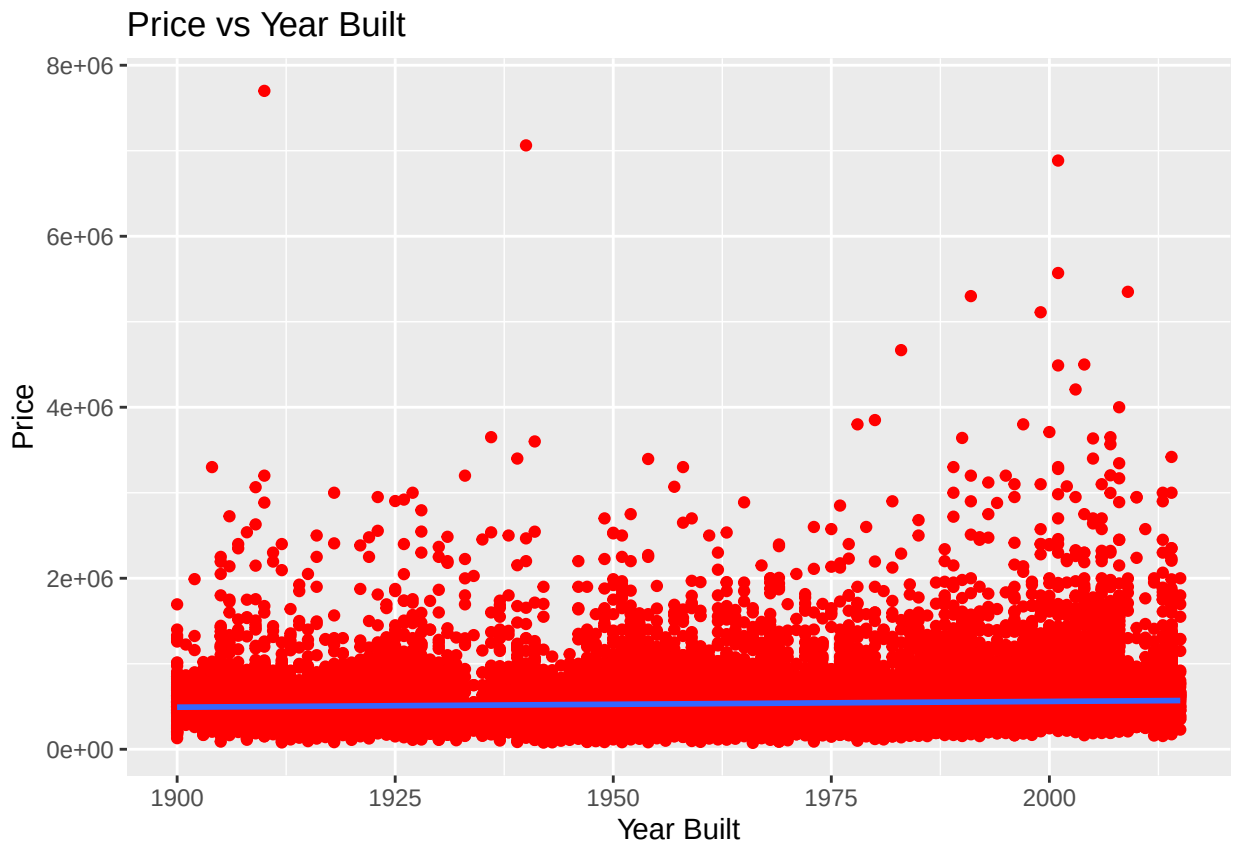
```
model = lm(price ~ yr_built, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = price ~ yr_built, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -461709 -221337  -87006  104064 7201095
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -790477.9   167350.2  -4.723 2.33e-06 ***
## yr_built        675.1       84.9   7.952 1.93e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 366600 on 21611 degrees of freedom
## Multiple R-squared:  0.002917,   Adjusted R-squared:  0.002871
## F-statistic: 63.23 on 1 and 21611 DF,  p-value: 1.93e-15
```

```
ggplot(data, aes(x = yr_built, y = price)) + geom_point(color = "red") +
    geom_smooth(method = "lm") + labs(title = "Price vs Year Built",
    x = "Year Built", y = "Price")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Result

The model shows a very weak link between the year a house was built and its price. On average, each additional year increases the price by $675.10. However, this number is small compared to other factors like the house size.

The R-squared value is 0.002917, meaning only about 0.2917% of the variation in house prices can be explained by the year the house was built. This suggests that other factors like size and condition have a much greater impact on the price.

The scatter plot shows that newer houses are a bit more expensive, but the trend is very weak. The line in the plot is almost flat, and there's a lot of scatter around it, saying that the year built doesn't have a strong influence on price by itself.

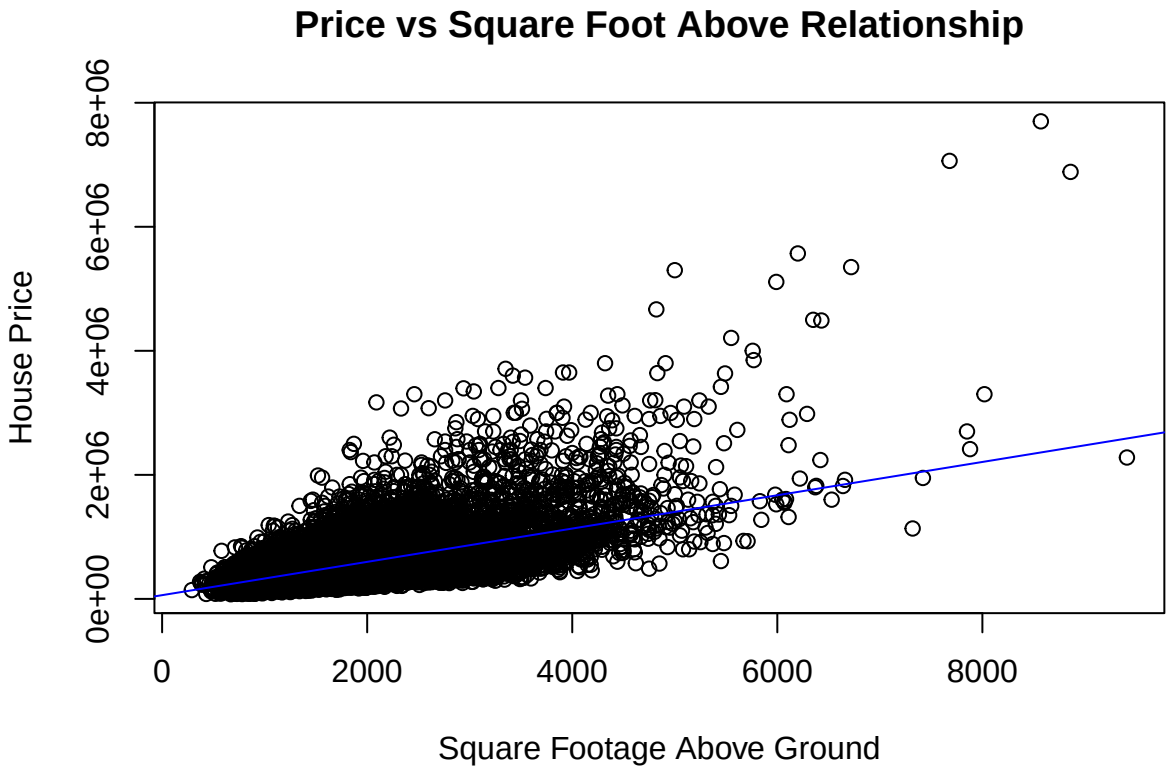
Square Foot Above and Price Relationship

```
model = glm(price ~ sqft_above, data = data, family = gaussian())

summary(model)
```

```
##
## Call:
## glm(formula = price ~ sqft_above, family = gaussian(), data = data)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59953.2     4729.8   12.68   <2e-16 ***
## sqft_above     268.5        2.4  111.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 85360045823)
##
##     Null deviance: 2.9129e+15  on 21612  degrees of freedom
## Residual deviance: 1.8447e+15  on 21611  degrees of freedom
## AIC: 605341
##
## Number of Fisher Scoring iterations: 2
```

```
plot(data$sqft_above, data$price, xlab = "Square Footage Above Ground",
    ylab = "House Price", main = "Price vs Square Foot Above Relationship")
abline(model, col = "blue")
```



Result

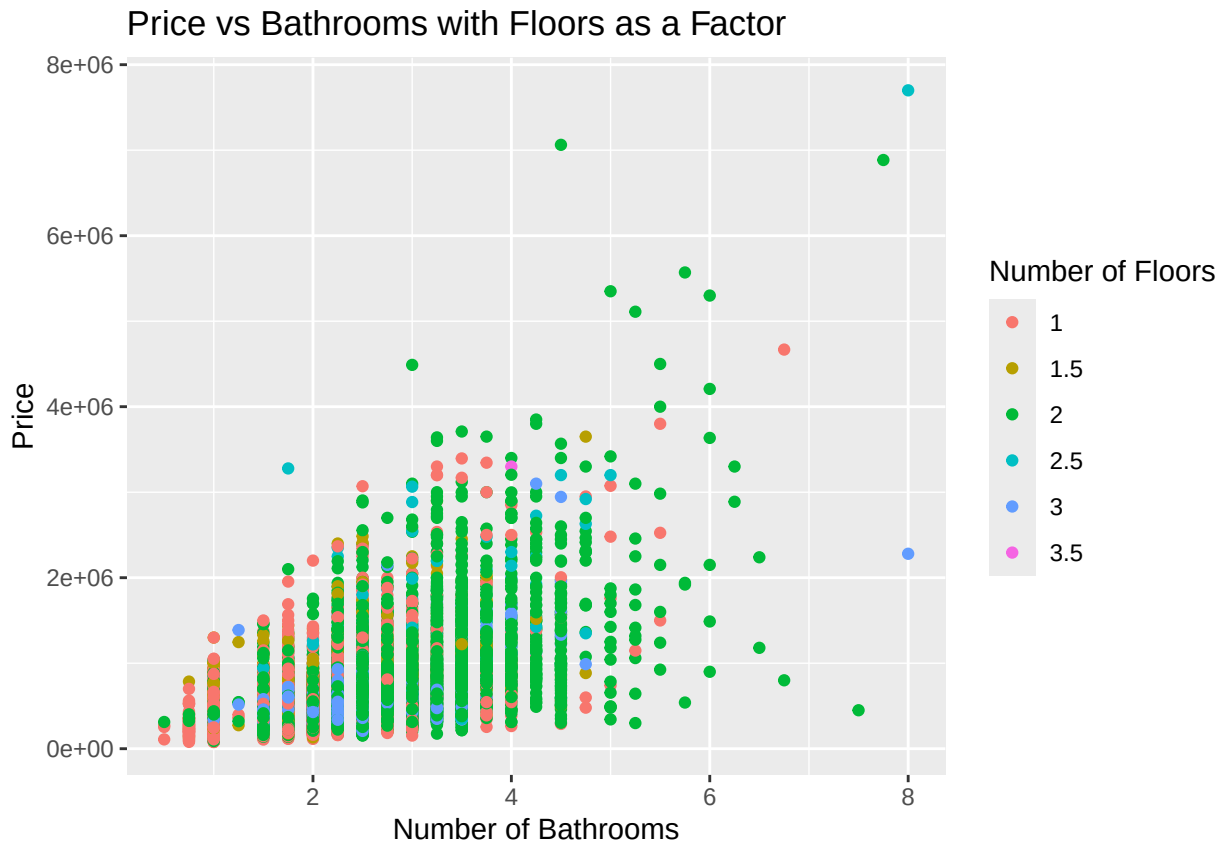
In this analysis, the relationship between the square footage of above-ground space and house price. The results show a strong positive correlation, where each additional square foot of above-ground space is associated with an increase in house price by $268.50. The model's statistical significance, shown by the very low p-value (< 2e-16), shows that square footage is a meaningful predictor of house prices.

Floors + Bathrooms and Price Relationship

```
library(ggplot2)

ggplot(data, aes(x = bathrooms, y = price, color = factor(floors))) +
    geom_point() + labs(title = "Price vs Bathrooms with Floors as a Factor",
    x = "Number of Bathrooms", y = "Price", color = "Number of Floors")
```

```
## Warning: Removed 10 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Result

The analysis shows a positive relationship between the number of bathrooms and house price—meaning that as the number of bathrooms increases, the price tends to go up as well. The number of floors appears to be spread out across the graph, with no clear trend, but it seems to increase in the upward direction.

Square Foot Basement and Square Foot Lot Relationship

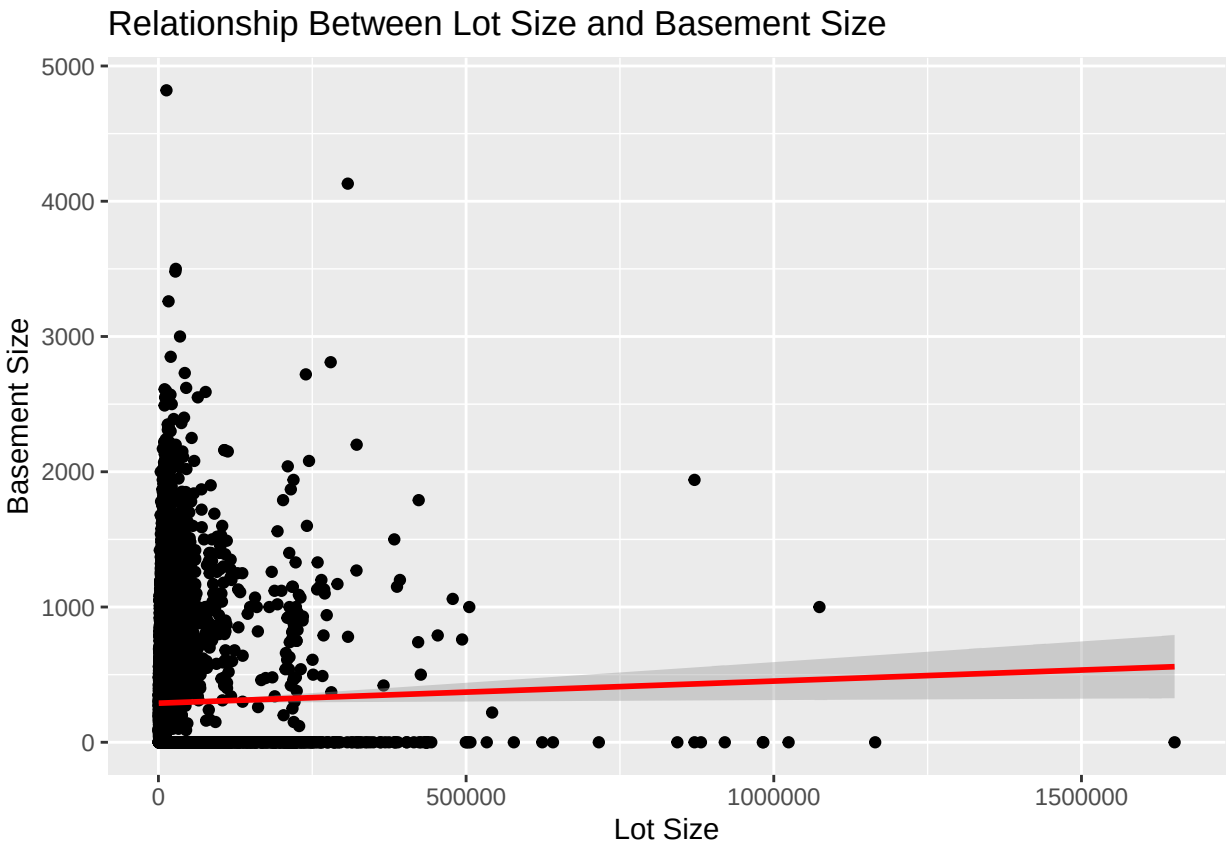
```
model = lm(sqft_basement ~ sqft_lot, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = sqft_basement ~ sqft_lot, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -558.8 -290.4 -289.8  270.2 4528.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.890e+02  3.204e+00  90.209   <2e-16 ***
## sqft_lot    1.633e-04  7.267e-05   2.247   0.0246 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 442.5 on 21611 degrees of freedom
## Multiple R-squared:  0.0002337,  Adjusted R-squared:  0.0001874
## F-statistic: 5.051 on 1 and 21611 DF,  p-value: 0.02462
```

```
ggplot(model, aes(x = sqft_lot, y = sqft_basement)) + geom_point() +
    geom_smooth(method = "lm", color = "red") + labs(title = "Relationship Between Lot Size and Basemen
    x = "Lot Size", y = "Basement Size")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Result

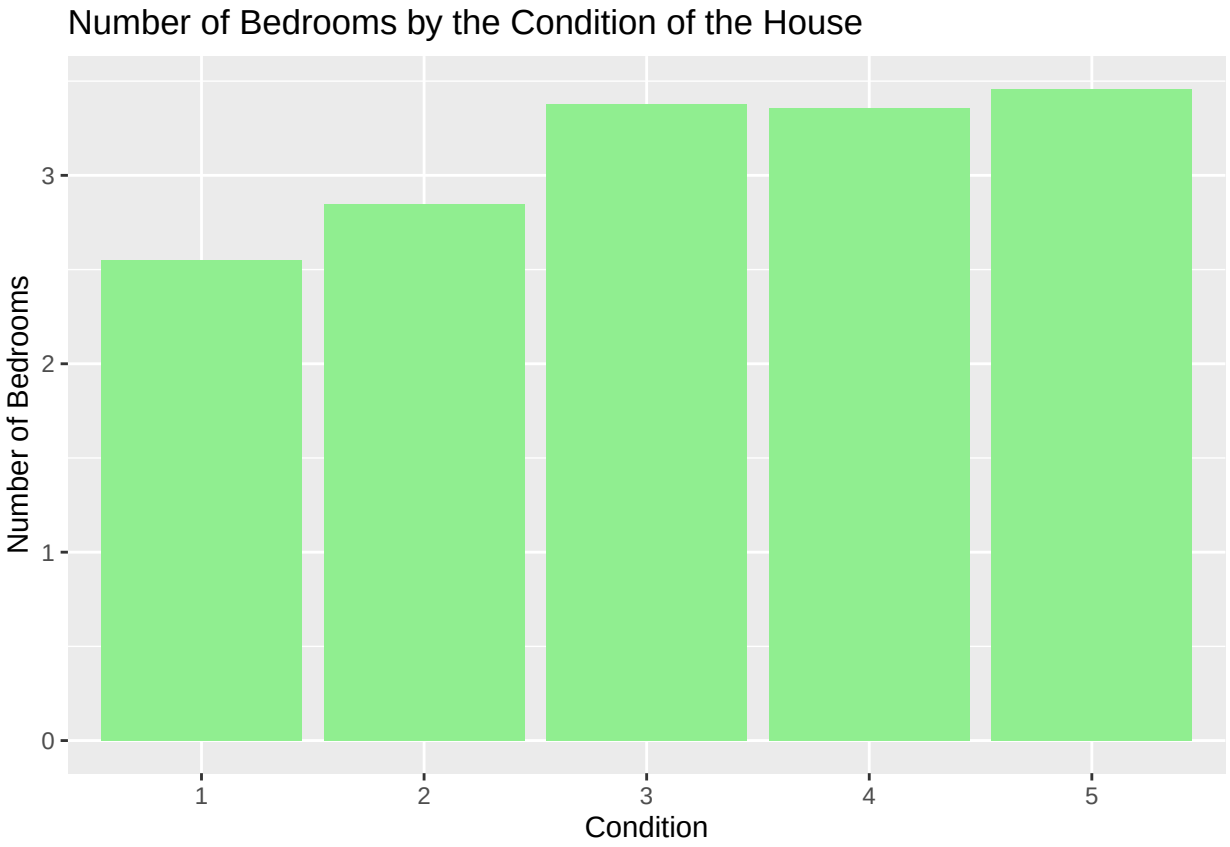
The analysis looks at the connection between lot size and basement size. It shows a very small positive trend, where basement size increases by about 0.00016 sqft for each additional square foot of lot size. While the p-value of 0.0246 means the relationship is statistically significant, the R-squared value is only 0.023%, so the model explains very little of the changes in basement size. The graph is mostly showing that the smaller the lot size, the bigger the basement size is.

Condition and Bedroom Relationship

```
mod = aggregate(bedrooms ~ condition, data, mean)

barplot = ggplot(mod, aes(x = factor(condition), y = bedrooms)) +
    geom_bar(stat = "identity", fill = "lightgreen") + labs(title = "Number of Bedrooms by the Condition
    x = "Condition", y = "Number of Bedrooms")

barplot
```



Result

The barplot reveals that as the number of bedrooms increases, the average condition rating (on a scale of 1 to 5) also improves. This could indicate that houses with more bedrooms are often better constructed or maintained. However, it's important to interpret this trend cautiously, as other factors, such as location or view might also influence the condition ratings.

Conclusion

In conclusion, this project helps us understand housing prices by looking at a dataset of house sales in King County, USA, from May 2014 to May 2015. The analysis shows that larger homes, with more square footage, generally have higher prices, with a clear connection between size and cost. On the other hand,

the year a house was built has very little effect on its price. The size of the above-ground space, however, does impact the price, with each additional square foot increasing the price. The relationship between lot size and basement size is that the smaller the lot size, the bigger the basement size. Also, houses with more bedrooms tend to have better condition ratings. While the dataset is limited in terms of time and location, the findings provide useful insights into the factors that influence housing prices.

References

[1] Posit team. *RStudio: Integrated Development Environment for R*. Posit Software, PBC. Boston, MA, 2024. URL: http://www.posit.co/.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[3] Shiv28. *House price prediction in King County, USA*. Feb. 2022. URL: https://www.kaggle.com/code/shiv28/house-price-prediction-in-king-county-usa/notebook.