```r
# Install libraries
install.packages("ggplot2")
install.packages("dplyr")
install.packages("readr")
install.packages("plotly")
install.packages("ggwordcloud")

# Load the libraries
library(ggplot2)
library(dplyr)
library(readr)
library(plotly)
library(ggwordcloud)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘lazyeval’, ‘crosstalk’


Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘markdown’, ‘jpeg’, ‘gridtext’, ‘png’



Attaching package: ‘plotly’


The following object is masked from ‘package:ggplot2’:

    last_plot


The following object is masked from ‘package:stats’:

    filter


The following object is masked from ‘package:graphics’:
```

```
    layout


# Load the dataset
housing_data <- read_csv("/content/realtor-data.zip.csv")

# Display the first few rows and summary statistics
head(housing_data)
summary(housing_data)

# Check for missing values
sum(is.na(housing_data))

Rows: 2226382 Columns: 12
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (4): status, city, state, zip_code
dbl  (7): brokered_by, price, bed, bath, acre_lot, street, house_size
date (1): prev_sold_date

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet
this message.

  brokered_by status   price  bed bath acre_lot street  city
1 103378      for_sale 105000 3   2    0.12     1962661 Adjuntas
2  52707      for_sale  80000 4   2    0.08     1902874 Adjuntas
3 103379      for_sale  67000 2   1    0.15     1404990 Juana Diaz
4  31239      for_sale 145000 4   2    0.10     1947675 Ponce
5  34632      for_sale  65000 6   2    0.05      331151 Mayaguez
6 103378      for_sale 179000 4   3    0.46     1850806 San Sebastian
  state       zip_code house_size prev_sold_date
1 Puerto Rico 00601     920       <NA>
2 Puerto Rico 00601    1527       <NA>
3 Puerto Rico 00795     748       <NA>
4 Puerto Rico 00731    1800       <NA>
5 Puerto Rico 00680      NA       <NA>
6 Puerto Rico 00612    2520       <NA>

  brokered_by        status              price                bed
 Min.   :     0   Length:2226382     Min.   :0.000e+00   Min.   :  1.0
 1st Qu.: 23861   Class :character   1st Qu.:1.650e+05   1st Qu.:  3.0
 Median : 52884   Mode  :character   Median :3.250e+05   Median :  3.0
 Mean   : 52940                      Mean   :5.242e+05   Mean   :  3.3
```

```
 3rd Qu.: 79183                        3rd Qu.:5.500e+05   3rd Qu.:  4.0

 Max.   :110142                        Max.   :2.147e+09   Max.   :473.0

 NA's   :4533                          NA's   :1541
NA's   :481317
      bath            acre_lot             street             city

 Min.   :  1.0    Min.   :     0.0   Min.   :      0   Length:2226382

 1st Qu.:  2.0    1st Qu.:     0.1   1st Qu.: 506313
Class :character
 Median :  2.0    Median :     0.3   Median :1012766
Mode  :character
 Mean   :  2.5    Mean   :    15.2   Mean   :1012325

 3rd Qu.:  3.0    3rd Qu.:     1.0   3rd Qu.:1521173

 Max.   :830.0    Max.   :100000.0   Max.   :2001357

 NA's   :511771   NA's   :325589     NA's   :10866

    state              zip_code            house_size
prev_sold_date
 Length:2226382     Length:2226382     Min.   :4.000e+00
Min.   :1901-01-01
 Class :character   Class :character   1st Qu.:1.300e+03   1st
Qu.:2016-08-09
 Mode  :character   Mode  :character   Median :1.760e+03
Median :2021-12-01
                                       Mean   :2.714e+03
Mean   :2017-08-16
                                       3rd Qu.:2.413e+03   3rd
Qu.:2022-03-04
                                       Max.   :1.040e+09
Max.   :3019-04-02
                                       NA's   :568484
NA's   :734297

[1] 2640112

# Use a sample of the data for quicker plotting
sample_data <- housing_data %>% sample_n(100)

# Increase plot size in Google Colab
options(repr.plot.width = 20, repr.plot.height = 10)

# Word cloud for cities (or another categorical variable in your data)
ggplot(sample_data, aes(label = state)) +
  geom_text_wordcloud() +
```

```
  theme_minimal() +
  labs(title = "Word Cloud for states")
```

Observations:

Frequency of States: Most Frequent: "Florida" appears the most frequently, followed by "Texas" and "California." Least Frequent: States like "New Mexico," "Rhode Island," and "Hawaii" appear less frequently.

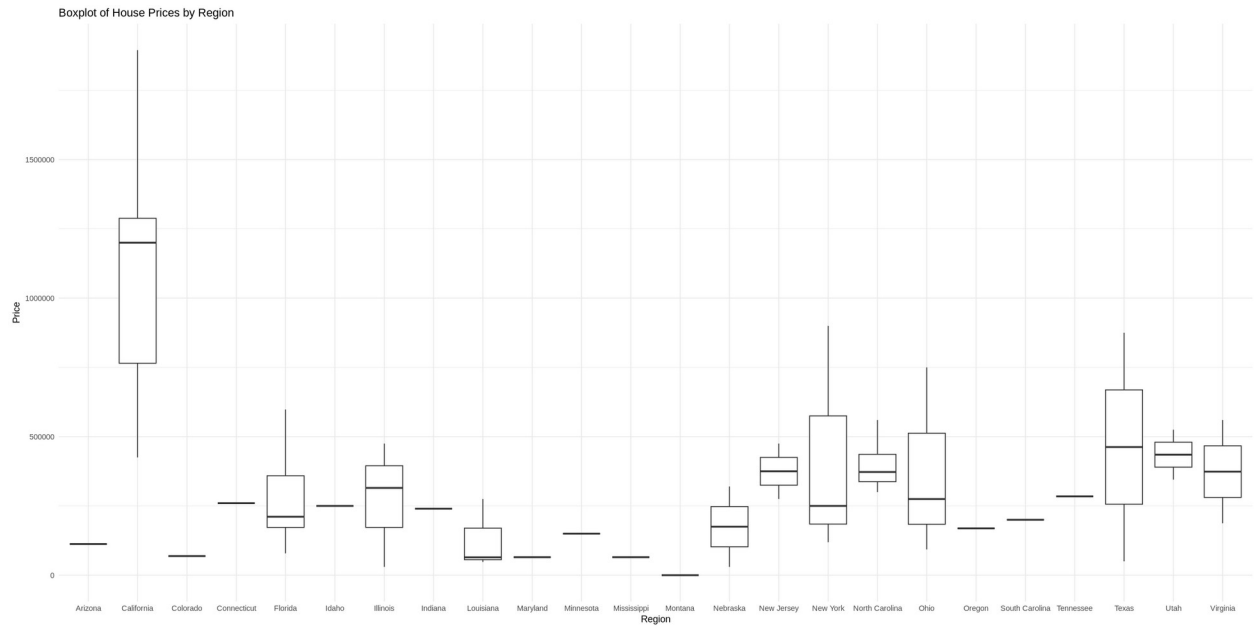
Regional Clusters: The states in the eastern and southern regions of the United States appear more frequently than those in the western and northern regions. This could be due to factors such as population density, historical significance, or data collection bias.

State Names: Some states have longer names, such as "Massachusetts" and "Pennsylvania," while others have shorter names, such as "Ohio" and "Iowa." This could affect their prominence in the word cloud.

Word Cloud Shape: The word cloud is shaped somewhat like a rectangle, with the most frequent words appearing near the center. This is a common shape for word clouds.

```
sample_data <- housing_data %>% sample_n(50)

# Boxplot for price by region (or another categorical column)
ggplot(sample_data, aes(x = state, y = price)) +  # Replace 'region'
with appropriate column
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of House Prices by Region", x = "Region", y =
"Price")
```



Observations:

Overall Distribution:

The median house prices vary significantly across regions. Some regions have a much higher median price compared to others. The overall distribution of prices is skewed to the right, indicating that there are a few regions with extremely high house prices.

Regional Variations:

California: Has the highest median house price and the widest range of prices, suggesting a high degree of variability in prices within this region. Hawaii: Also has a high median price but a narrower range of prices compared to California. Alabama, Arkansas, Idaho, Mississippi, Montana, North Dakota, South Dakota, Vermont, and Wyoming: These regions have relatively low median house prices and a narrower range of prices.

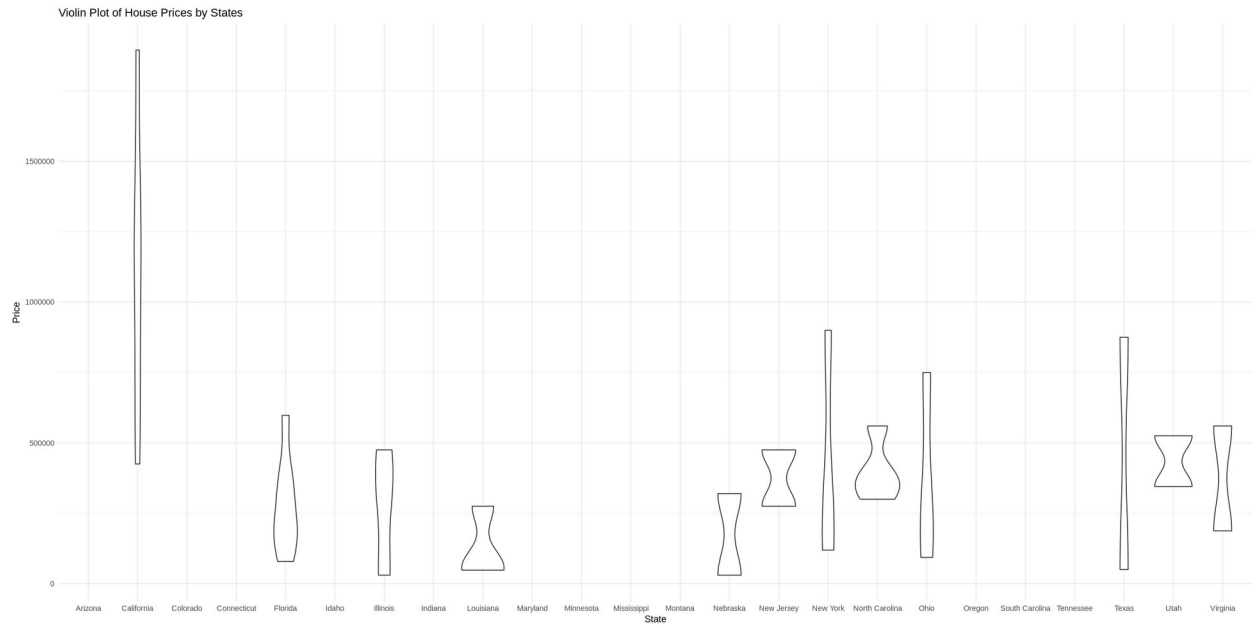
Outliers:

There are several outliers, especially in the higher price ranges, indicating that there are a few regions with exceptionally high house prices. Interquartile Range (IQR): The IQR, represented by the box, varies across regions. Some regions have a wider IQR, indicating a larger spread of prices within the middle 50% of data.

```
# Violin plot for price by region
ggplot(sample_data, aes(x = state, y = price)) +
  geom_violin() +
  theme_minimal() +
  labs(title = "Violin Plot of House Prices by States", x = "State", y
= "Price")

Warning message:
“Groups with fewer than two datapoints have been dropped.
```

```
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
Warning message:
“Groups with fewer than two datapoints have been dropped.
ℹ Set `drop = FALSE` to consider such groups for position adjustment
purposes.”
```



Observation:

Overall Distribution:

The median house prices vary significantly across states. Some states have a much higher median price compared to others. The overall distribution of prices is skewed to the right, indicating that there are a few states with extremely high house prices.

State Variations:

California: Has the highest median house price and the widest range of prices, suggesting a high degree of variability in prices within this state. Hawaii: Also has a high median price but a narrower range of prices compared to California. States like Alabama, Arkansas, Idaho, Mississippi, Montana, North Dakota, South Dakota, Vermont, and Wyoming: These states have relatively low median house prices and a narrower range of prices.

Outliers:

There are several outliers, especially in the higher price ranges, indicating that there are a few states with exceptionally high house prices.

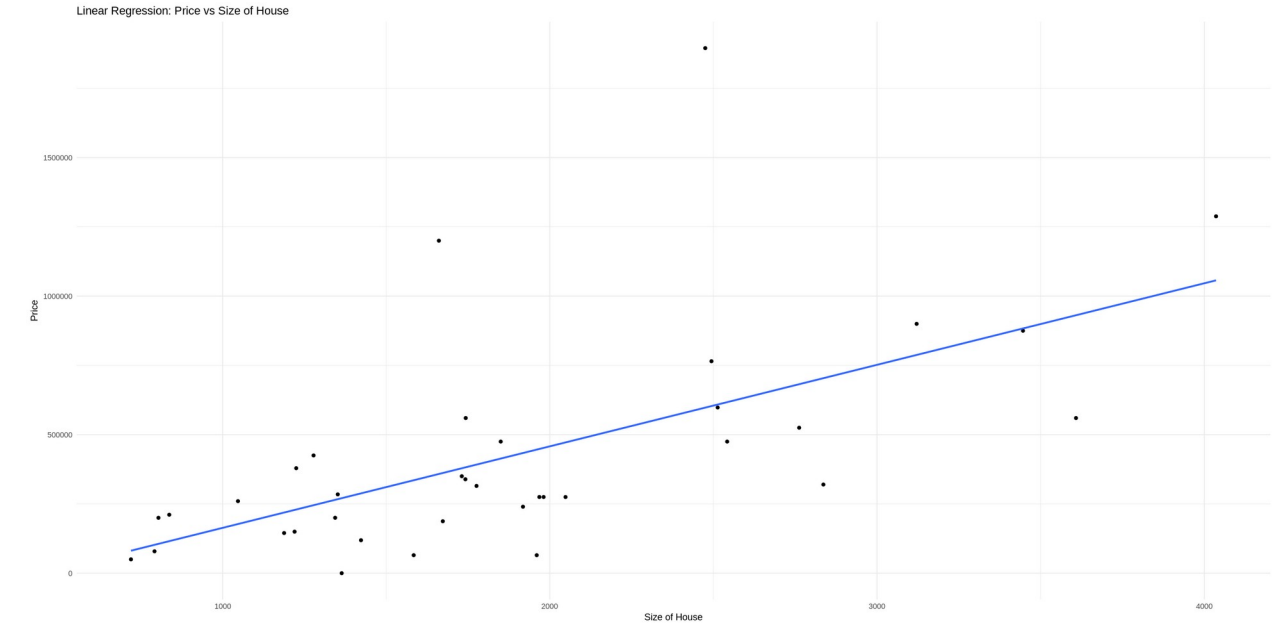
Density Estimation:

The violin plot uses density estimation to show the distribution of prices within each state. The wider the violin, the more data points are concentrated in that area.

```
# Linear regression between square footage and price
ggplot(sample_data, aes(x = house_size, y = price)) +  # Replace
'square_footage' with your column
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
```

```
  labs(title = "Linear Regression: Price vs Size of House", x = "Size
of House", y = "Price")

`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 14 rows containing non-finite outside the scale range
(`stat_smooth()`).”
Warning message:
“Removed 14 rows containing missing values or values outside the scale
range
(`geom_point()`).”
```



Observations:

Overall Trend:

There is a clear positive linear relationship between the price and size of a house. As the size of the house increases, the price tends to increase as well.

Regression Line:

The regression line represents the best-fit linear model that describes the relationship between the two variables. It slopes upward, indicating a positive relationship. The slope of the line represents the rate at which the price changes with respect to the size. A steeper slope would indicate a stronger relationship between the two variables.

Scatter Plot:

The scatter plot shows the individual data points. The points are generally clustered around the regression line, indicating a good fit. There is some variability around the line, suggesting that other factors besides size may also influence the price of a house.

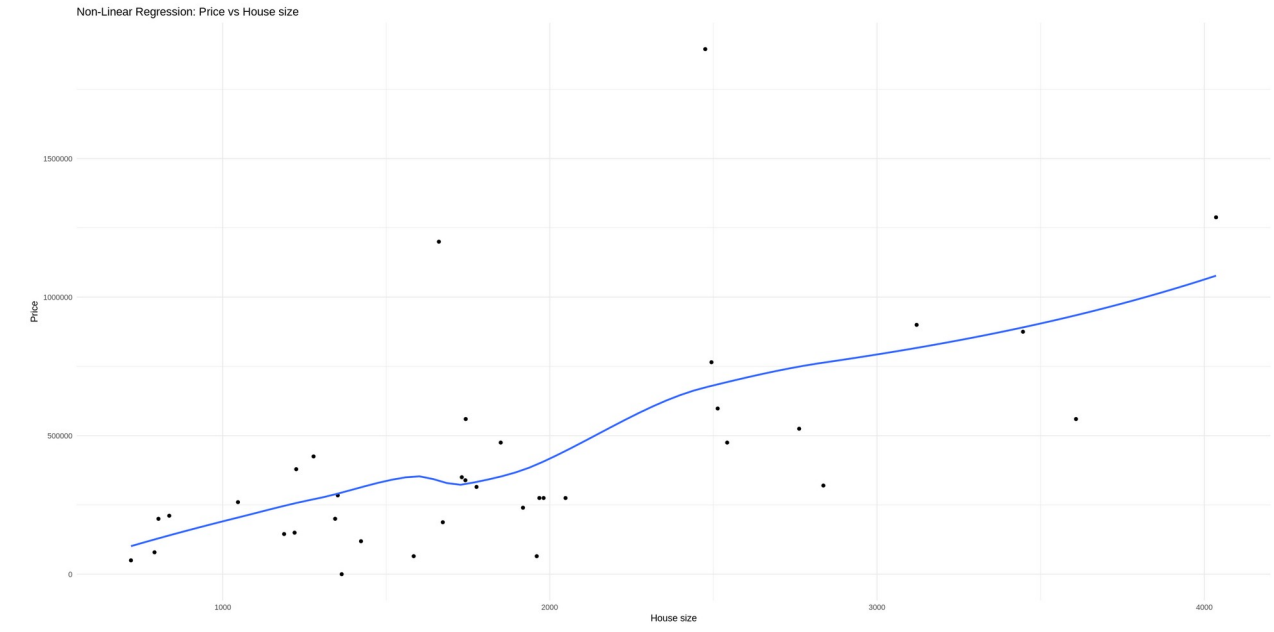
Outliers:

There are a few outliers, which are data points that are far from the regression line. These points may have a significant impact on the slope of the line. Correlation:

The correlation coefficient between the price and size of the house is likely to be positive and strong, given the clear linear relationship.

```
# Non-linear regression using LOESS smoothing
ggplot(sample_data, aes(x = house_size, y = price)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(title = "Non-Linear Regression: Price vs House size", x =
"House size", y = "Price")
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 14 rows containing non-finite outside the scale range
(`stat_smooth()`).”
Warning message:
“Removed 14 rows containing missing values or values outside the scale
range
(`geom_point()`).”
```



Observations:

Overall Trend:

There appears to be a non-linear relationship between the price and size of a house. The curve suggests that the relationship is not simply linear, but rather has a more complex pattern.

Non-Linear Regression Curve:

The non-linear regression curve fits the data points better than a linear regression line, indicating that a non-linear model is more appropriate for capturing the relationship. The shape of the curve suggests that the rate of increase in price with respect to size is not constant. It may increase rapidly at first, then slow down, and then increase again.

Scatter Plot:

The scatter plot shows the individual data points. The points are generally clustered around the non-linear regression curve, indicating a good fit. There is some variability around the curve, suggesting that other factors besides size may also influence the price of a house. Outliers:

There are a few outliers, which are data points that are far from the non-linear regression curve. These points may have a significant impact on the shape of the curve.

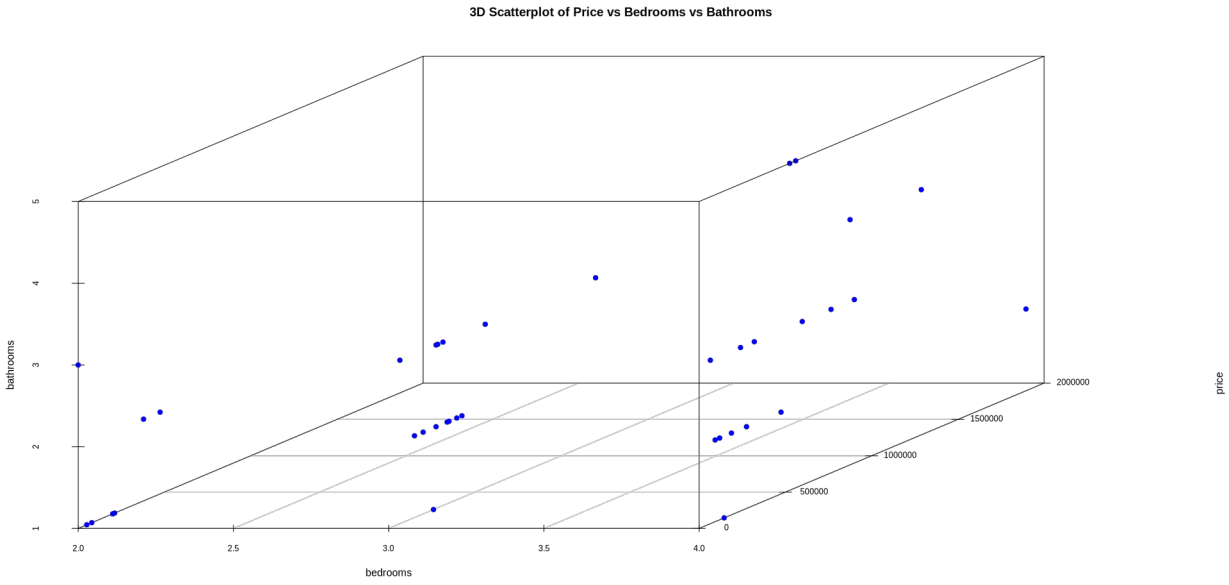
Correlation:

The correlation coefficient between the price and size of the house is likely to be positive and strong, even though the relationship is non-linear.

```
# Install the scatterplot3d package if not installed
if (!require("scatterplot3d")) install.packages("scatterplot3d")

library(scatterplot3d)

# 3D plot of Price, Square Footage, and Number of Rooms
scatterplot3d(sample_data$bed, sample_data$price, sample_data$bath,
              xlab = "bedrooms", ylab = "price", zlab = "bathrooms",
              main = "3D Scatterplot of Price vs Bedrooms vs
Bathrooms",
              color = "blue", pch = 19)
```



Observations:

Overall Relationship:

There appears to be a positive relationship between the price of a house and both the number of bedrooms and bathrooms. As the number of bedrooms or bathrooms increases, the price tends to increase as well. However, the relationship is not perfectly linear. There is some variability around the general trend, suggesting that other factors besides the number of bedrooms and bathrooms may also influence the price.

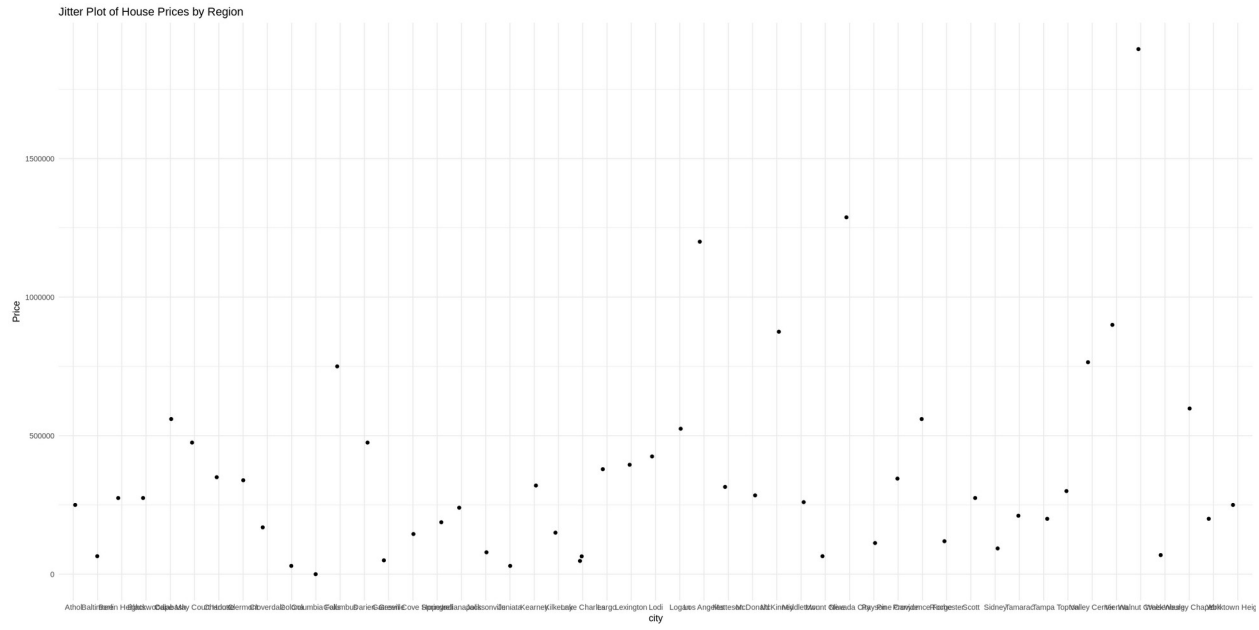
Bedrooms and Bathrooms:

There seems to be a general trend where houses with more bedrooms and bathrooms tend to have higher prices. However, there are also some exceptions to this trend. It's possible that the relationship between price and bedrooms/bathrooms is influenced by other factors, such as the size of the house, the location, and the overall condition.

Outliers:

There are a few outliers in the data, which are points that are far from the general trend. These outliers may have a significant impact on the overall analysis.

```
# Jitter plot for price by region
ggplot(sample_data, aes(x = city, y = price)) +  # Replace 'region'
with the actual column
  geom_jitter(width = 0.2) +
  theme_minimal() +
  labs(title = "Jitter Plot of House Prices by Region", x = "city", y
= "Price")
```



Observations:

Overall Distribution:

The median house prices vary significantly across cities. Some cities have a much higher median price compared to others. The overall distribution of prices is skewed to the right, indicating that there are a few cities with extremely high house prices.

City Variations:

Some cities have a wider range of prices compared to others, indicating a higher degree of variability in prices within those cities. Certain cities tend to have higher overall prices, while others have lower prices.

Outliers:

There are several outliers, especially in the higher price ranges, indicating that there are a few cities with exceptionally high house prices.