R Notebook

Code

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Hide

```
plot(cars)
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

Hide

```
# Load necessary libraries
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(plotly)
library(grDevices)
library(dplyr)
library(grid)
```

Hide

```
# Load the dataset
data <- read.csv("C:\\Users\\khush\\Downloads\\Housing.csv")

# Check the first few rows of the dataset
head(data)
```

	price <int>	a… <int>	bedroo… <int>	bathroo… <int>	stories <int>	mainro… <chr>	guestro… <chr>	basem… <chr>	hotwaterheating <chr>
1	13300000	7420	4	2	3	yes	no	no	no
2	12250000	8960	4	4	4	yes	no	no	no
3	12250000	9960	3	2	2	yes	no	yes	no
4	12215000	7500	4	2	2	yes	no	yes	no
5	11410000	7420	4	1	2	yes	yes	yes	no
6	10850000	7500	3	3	1	yes	no	yes	no

Hide

```
# Create a data frame for the word cloud
df <- data %>% count(furnishingstatus)

# Create the word cloud
wordcloud(words = df$furnishingstatus, freq = df$n,
          colors = brewer.pal(8, "Dark2"),
          scale = c(3, 0.5))

# Add a title
grid.text("Word Cloud of Furnishing Status",
          x = unit(0.5, "npc"),
          y = unit(1, "npc") - unit(1, "lines"),
          gp = gpar(fontsize = 15, fontface = "bold"))
```

Word Cloud of Furnishing Status

Question Answered:

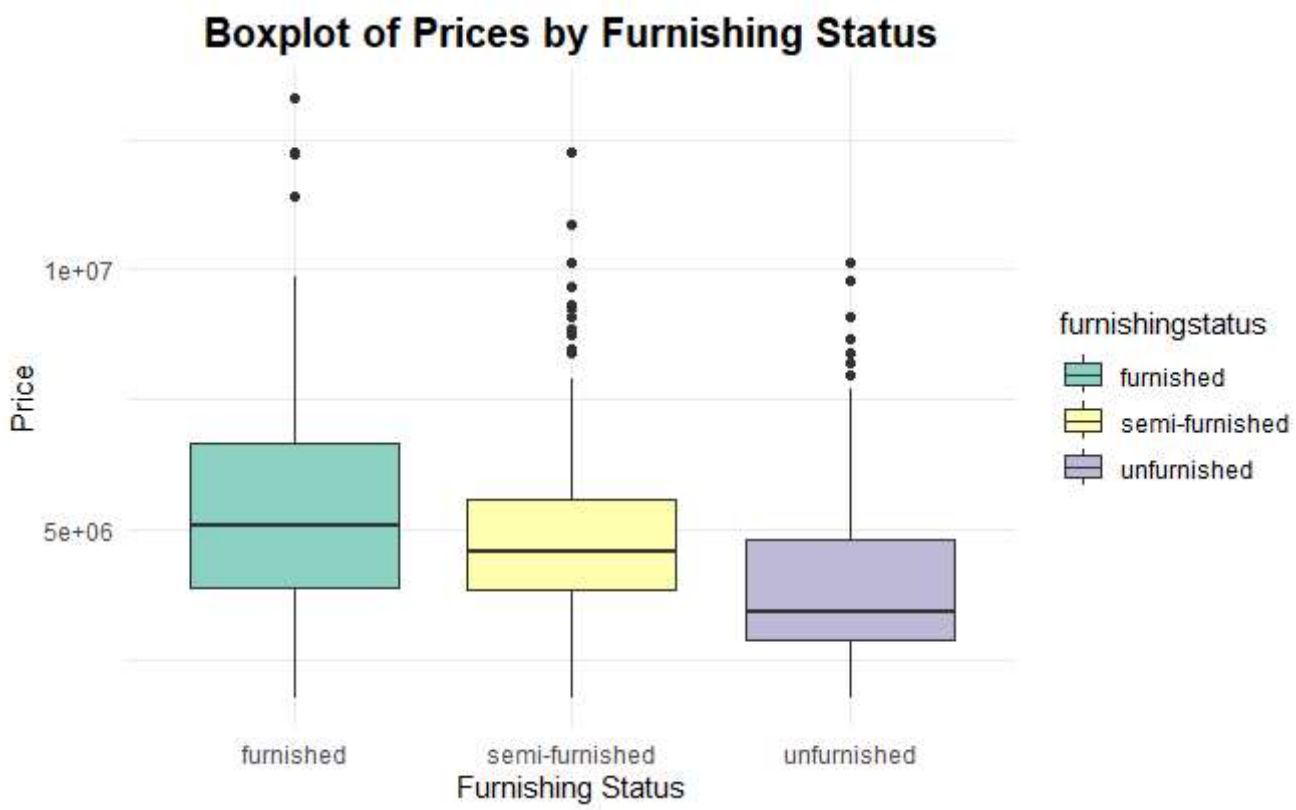
What is the distribution of different furnishing statuses in the dataset?

Observation:

- “Semi-furnished” is the most frequent furnishing status, followed by “Unfurnished” and “Furnished.”
- This indicates that a significant number of houses or apartments in the dataset are semi-furnished, making it the dominant furnishing status.

Hide

```
ggplot(data, aes(x = furnishingstatus, y = price, fill = furnishingstatus)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Prices by Furnishing Status",
       x = "Furnishing Status",
       y = "Price") +
  scale_fill_brewer(palette = "Set3") +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5)  # Title styling
  )
```



Boxplot of Prices by Furnishing Status

Question Answered:

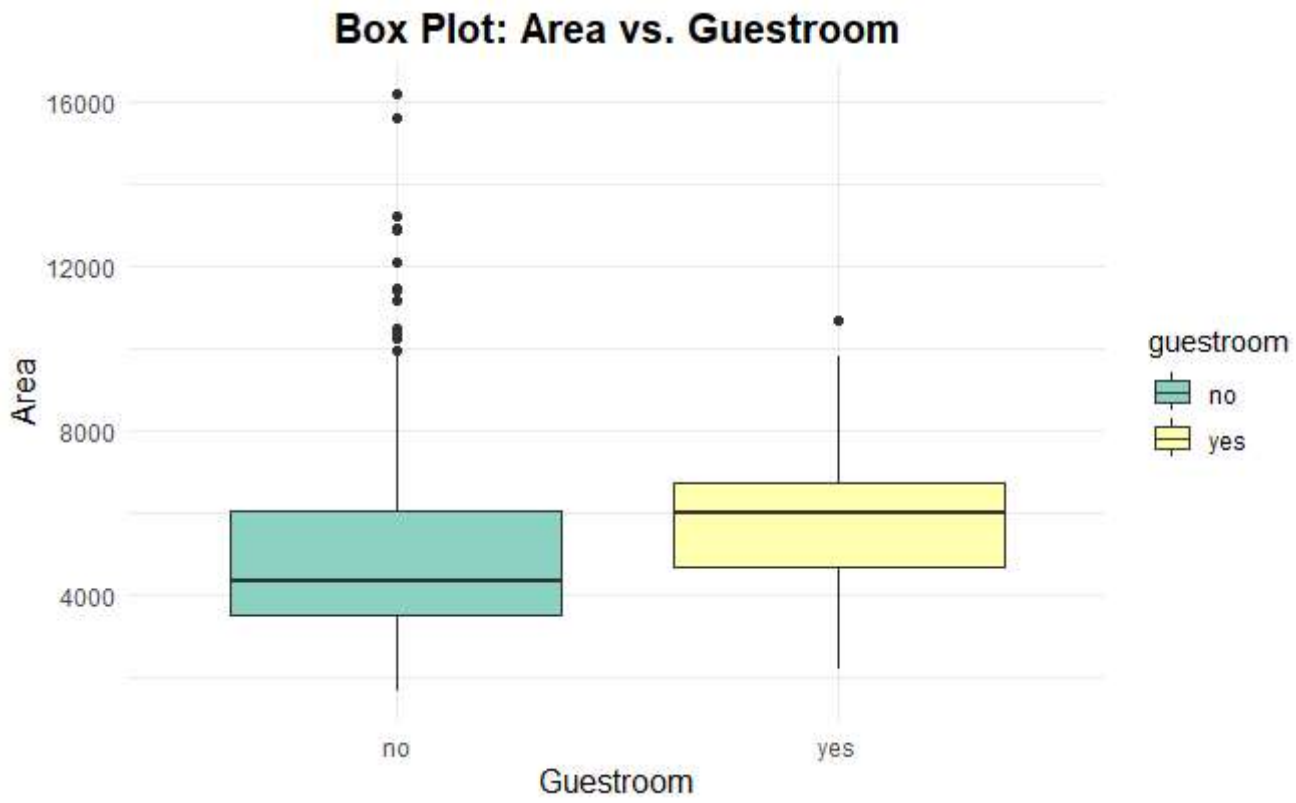
How do prices vary based on the furnishing status?

Observation:

- The median price is highest for “Furnished” properties, followed by “Semi-furnished” and then “Unfurnished.”
- Furnished properties tend to have the highest range of prices, while unfurnished ones have a lower price range, showing a clear relationship between furnishing and property value.

Hide

```
ggplot(data, aes(x = guestroom, y = area, fill = guestroom)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Box Plot: Area vs. Guestroom",
       x = "Guestroom",
       y = "Area") +
  scale_fill_brewer(palette = "Set3") +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5),  # Title styling
    axis.title.x = element_text(size = 12),  # X-axis title styling
    axis.title.y = element_text(size = 12)   # Y-axis title styling
  )
```



Boxplot of Area vs. Guestroom

Question Answered:

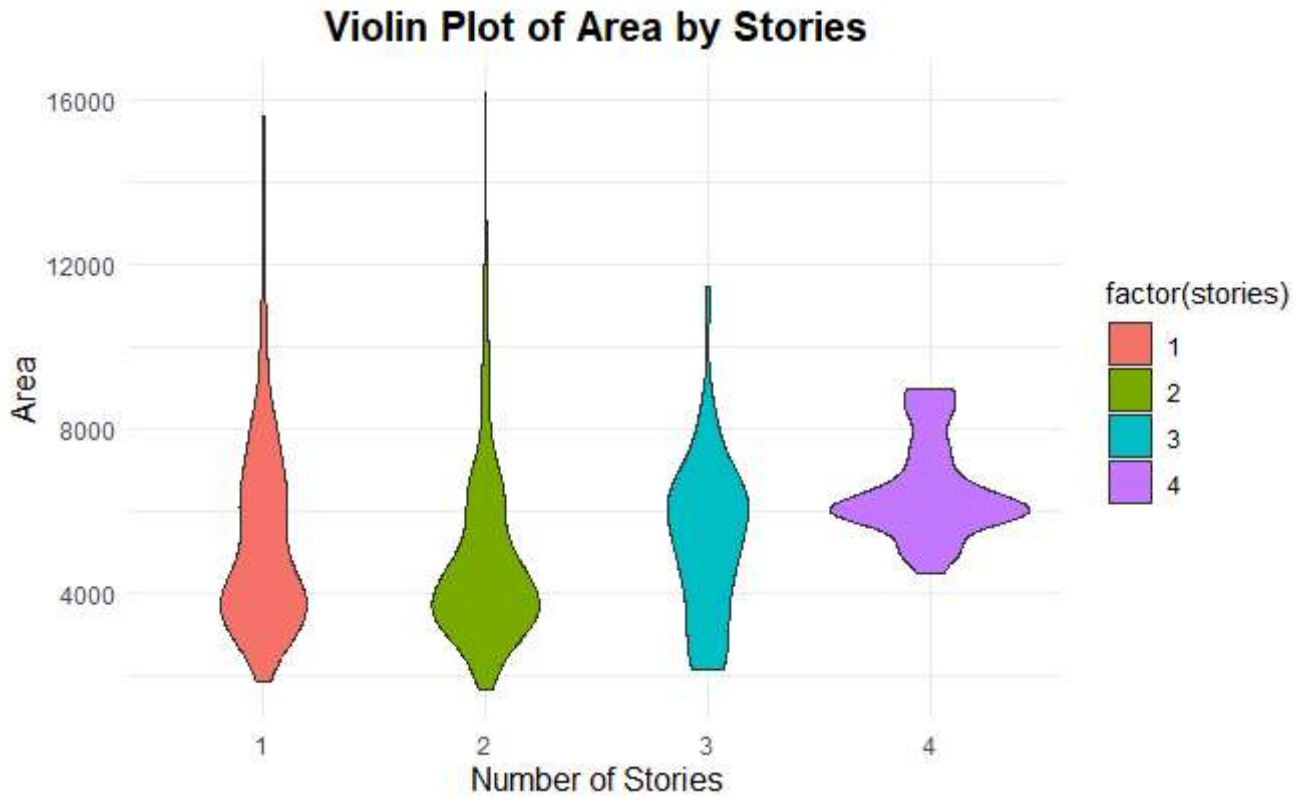
How does having a guestroom impact the area of the house?

Observation:

- Houses with a guestroom tend to have a larger median area compared to those without a guestroom.
- This suggests that homes with guestrooms are typically larger, providing more space overall.

Hide

```
ggplot(data, aes(x = factor(stories), y = area, fill = factor(stories))) +
  geom_violin() +
  labs(title = "Violin Plot of Area by Stories",
       x = "Number of Stories",
       y = "Area") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5),  # Title styling
    axis.title.x = element_text(size = 12),  # X-axis title styling
    axis.title.y = element_text(size = 12)   # Y-axis title styling
  )
```



Violin Plot of Area by Stories

Question Answered:

How does the area of a property vary based on the number of stories?

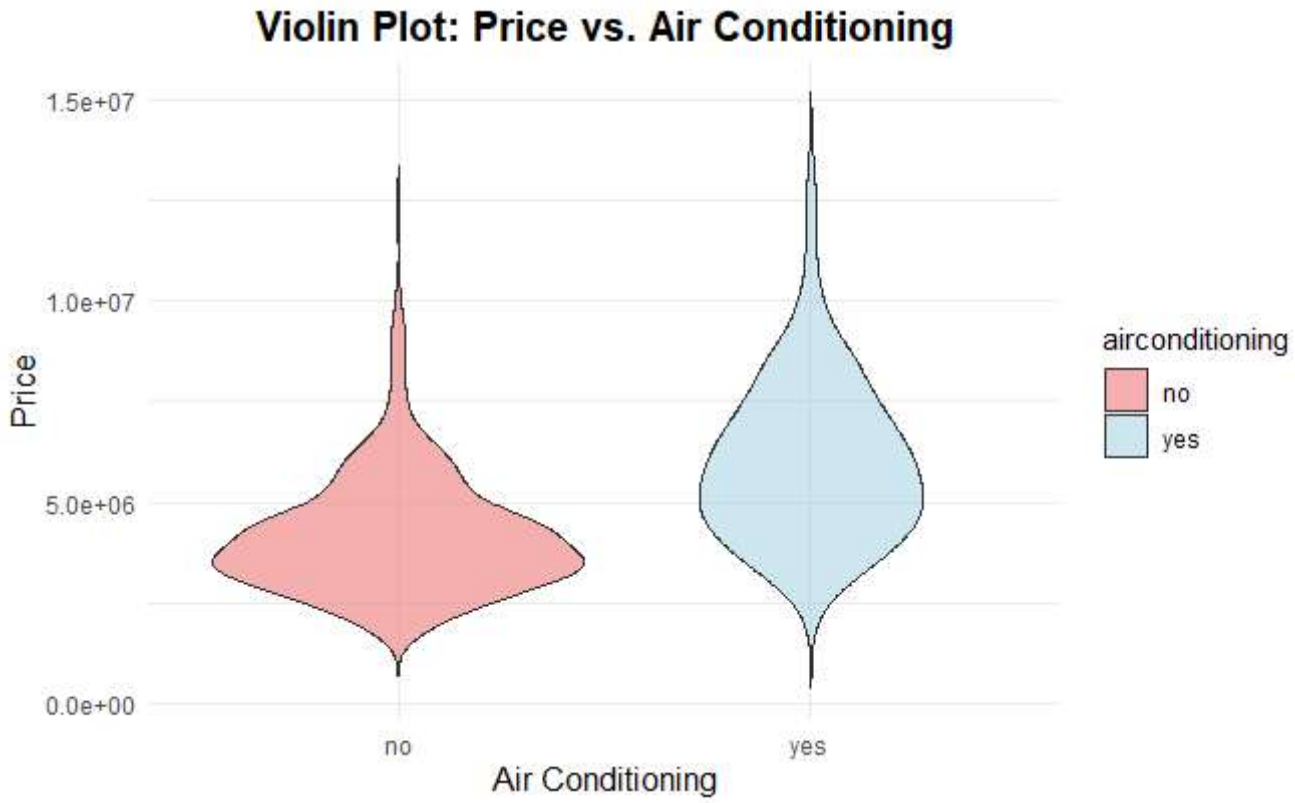
Observations:

- **1-Story Buildings**: These have the widest range of areas, with a high density of smaller properties but also a few large properties.
- **2-Story Buildings**: The area distribution is more concentrated compared to 1-story buildings, with most properties being in the middle range of areas.
- **3-Story Buildings**: The area distribution narrows further, suggesting that properties with 3 stories tend to have a more consistent area size. There are fewer extreme values, and the bulk of the data lies in the middle.

- **4-Story Buildings**:The distribution for 4-story buildings is the narrowest, with most properties having similar areas.

Hide

```
ggplot(data, aes(x = airconditioning, y = price, fill = airconditioning)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  theme_minimal() +
  labs(title = "Violin Plot: Price vs. Air Conditioning",
       x = "Air Conditioning",
       y = "Price") +
  scale_fill_manual(values = c("yes" = "lightblue", "no" = "lightcoral")) +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5),  # Title styling
    axis.title.x = element_text(size = 12),  # X-axis title styling
    axis.title.y = element_text(size = 12)   # Y-axis title styling
  )
```



Violin Plot: Price vs. Air Conditioning

Question Answered:

How does the price of a property vary based on the presence of air conditioning?

Observations:

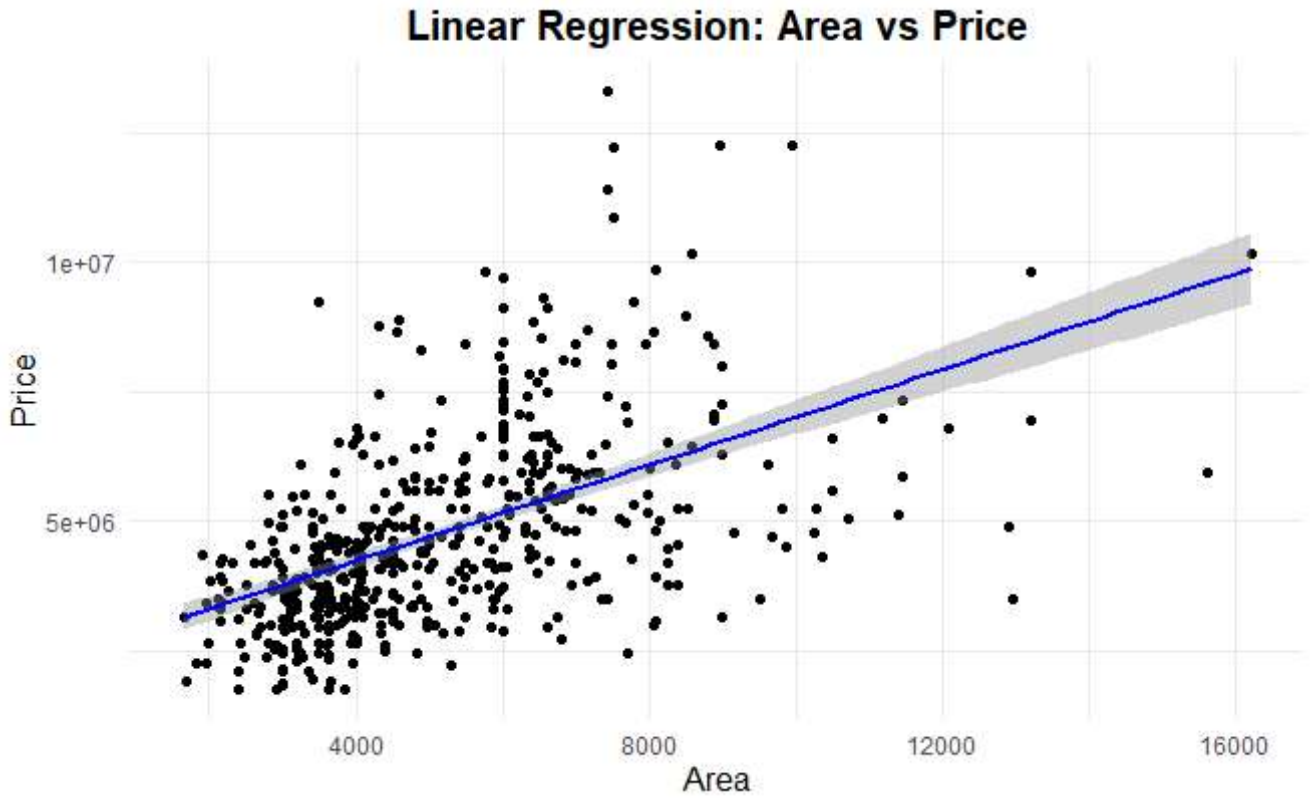
- **Properties without Air Conditioning**: The price distribution shows concentration of properties at lower prices.
- **Properties with Air Conditioning**: The price distribution here is also broad, but there seems to be a greater density around mid- to higher-price ranges.

Properties with air conditioning generally have a higher price range, which could suggest that air conditioning adds value or is a common feature in more expensive properties.

Hide

```
ggplot(data, aes(x = area, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Linear Regression: Area vs Price",
       x = "Area",
       y = "Price") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5),  # Title styling
    axis.title.x = element_text(size = 12),  # X-axis title styling
    axis.title.y = element_text(size = 12)   # Y-axis title styling
  )
```

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



Linear Regression: Area vs Price

Question Answered:

How does the area of a property affect its price, and is there a linear relationship between them?

Observations:

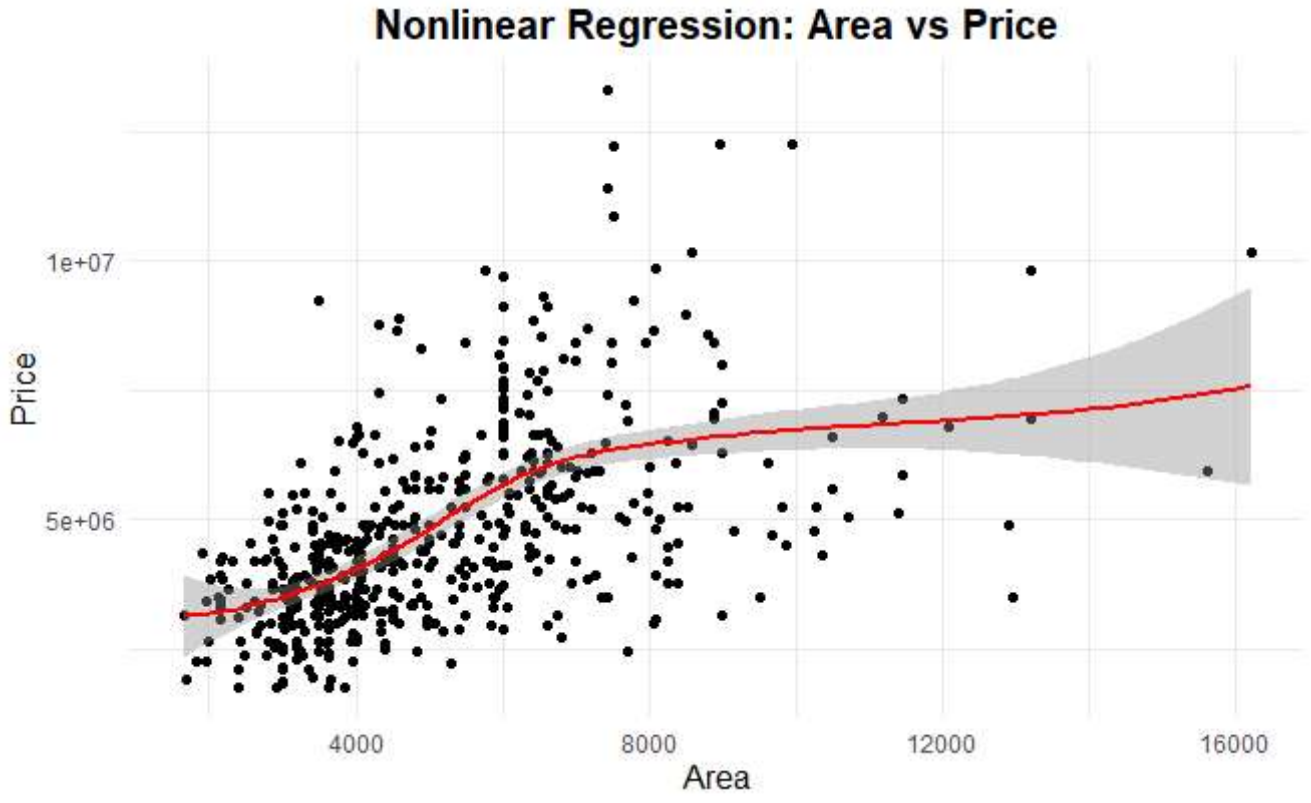
- **Positive Correlation**: The scatter plot shows a clear positive correlation between property area and price. As the area increases, the price also tends to rise.

- **Linear Trend**: The blue regression line indicates the linear trend between area and price, suggesting that for larger properties, the price generally increases at a consistent rate.
- **Variability**: While there is a strong positive correlation, the points are spread around the regression line, particularly at higher areas. This indicates that while area is a strong predictor of price, other factors may also influence property prices, especially for larger properties.

Hide

```
ggplot(data, aes(x = area, y = price)) +
  geom_point() +
  geom_smooth(method = "loess", col = "red") +
  labs(title = "Nonlinear Regression: Area vs Price",
       x = "Area",
       y = "Price") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5),  # Title styling
    axis.title.x = element_text(size = 12),  # X-axis title styling
    axis.title.y = element_text(size = 12)   # Y-axis title styling
  )
```

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



Nonlinear Regression: Area vs Price

Question Answered:

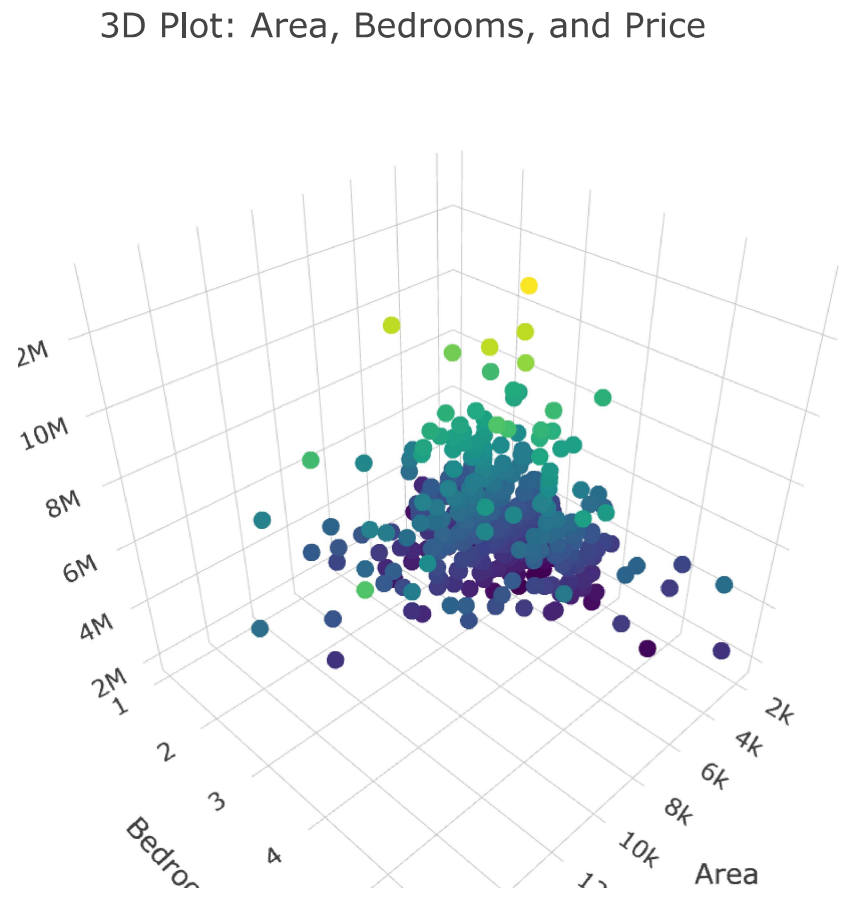
Is there a nonlinear relationship between the area of a property and its price, and how does the price vary across different property sizes?

Observations:

- **Nonlinear Trend**: The red LOESS (Locally Estimated Scatterplot Smoothing) curve shows a nonlinear relationship between the area and price, suggesting that the relationship is not purely linear. The price increase varies across different ranges of the area.
- **Lower Areas**: In the lower area range, the slope of the curve is steeper, suggesting that as the area increases in this range, the price rises more sharply.
- **Middle Areas**: For medium-sized properties, the price increases at a relatively steady pace.
- **Higher Areas**: In the higher area range, the curve starts to flatten, showing a diminishing rate of price increase. This suggests that after a certain point, adding more area to a property may result in only marginal increases in price, reflecting a saturation point in value for very large properties.

Hide

```
plot_ly(data, x = ~area, y = ~bedrooms, z = ~price, type = 'scatter3d', mode = 'markers',
        marker = list(size = 5, color = ~price, colorscale = 'Viridis')) %>%
  layout(
    title = '3D Plot: Area, Bedrooms, and Price',
    scene = list(
      xaxis = list(title = 'Area'),
      yaxis = list(title = 'Bedrooms'),
      zaxis = list(title = 'Price')
    )
  )
```



3D Plot: Area, Bedrooms, and Price:

Question Answered:

How do the number of bedrooms and the area of a property together affect its price?

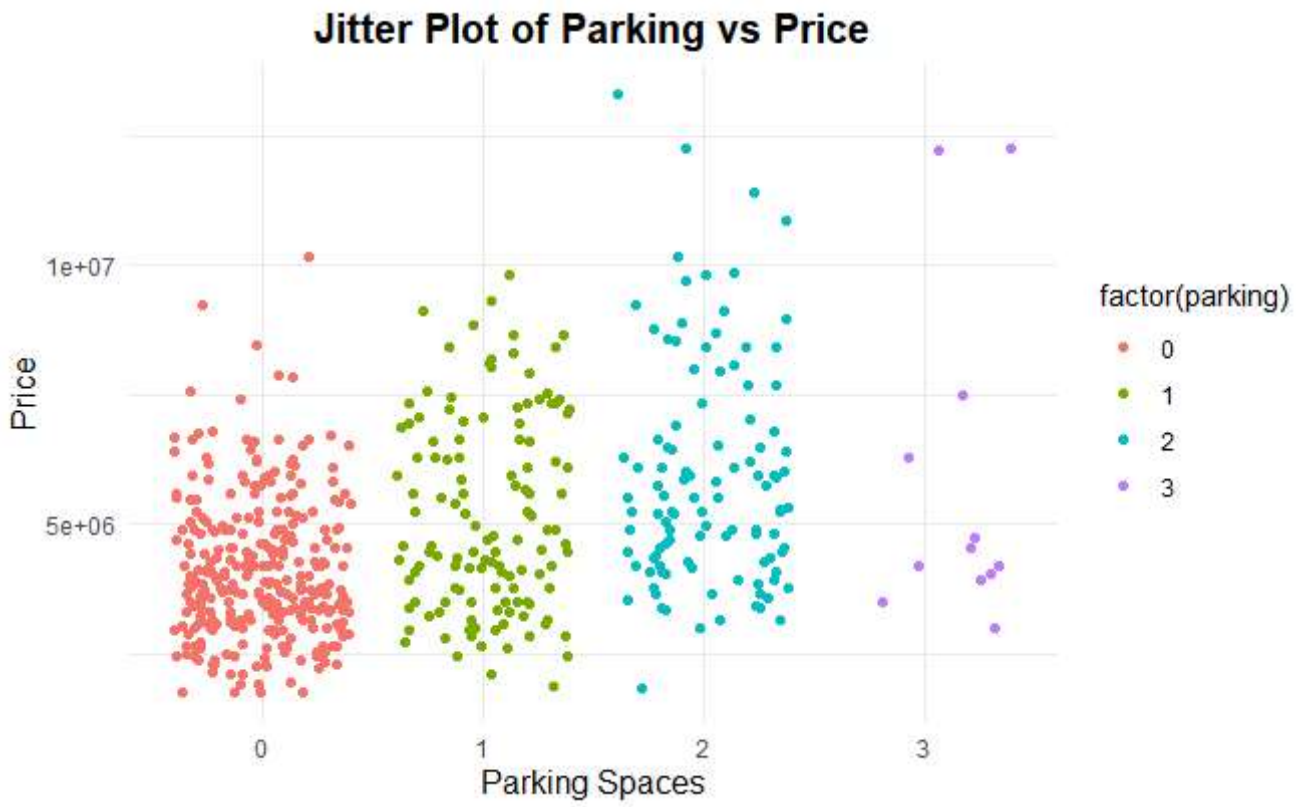
What is the 3-dimensional relationship between these three variables?

Observations:

- **Price as the Color Gradient**: The color gradient (Viridis scale) represents price, with darker colors indicating lower prices and brighter colors representing higher prices.
- **Positive Correlation**: Generally, properties with a higher number of bedrooms and larger areas tend to have higher prices, visible through the clustering of brighter points in the upper-right region of the plot.
- **Outliers**: There may be some outliers where the price is either unusually high for a smaller property or relatively low for a larger one, potentially due to other factors not captured in this 3D plot (such as location, amenities, or condition of the property).

Hide

```
ggplot(data, aes(x = factor(parking), y = price)) +
  geom_jitter(aes(color = factor(parking))) +
  labs(title = "Jitter Plot of Parking vs Price",
       x = "Parking Spaces",
       y = "Price") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5),  # Title styling
    axis.title.x = element_text(size = 12),  # X-axis title styling
    axis.title.y = element_text(size = 12)   # Y-axis title styling
  )
```



Jitter Plot of Parking vs. Price

Question Answered:

How does the number of parking spaces relate to property prices?

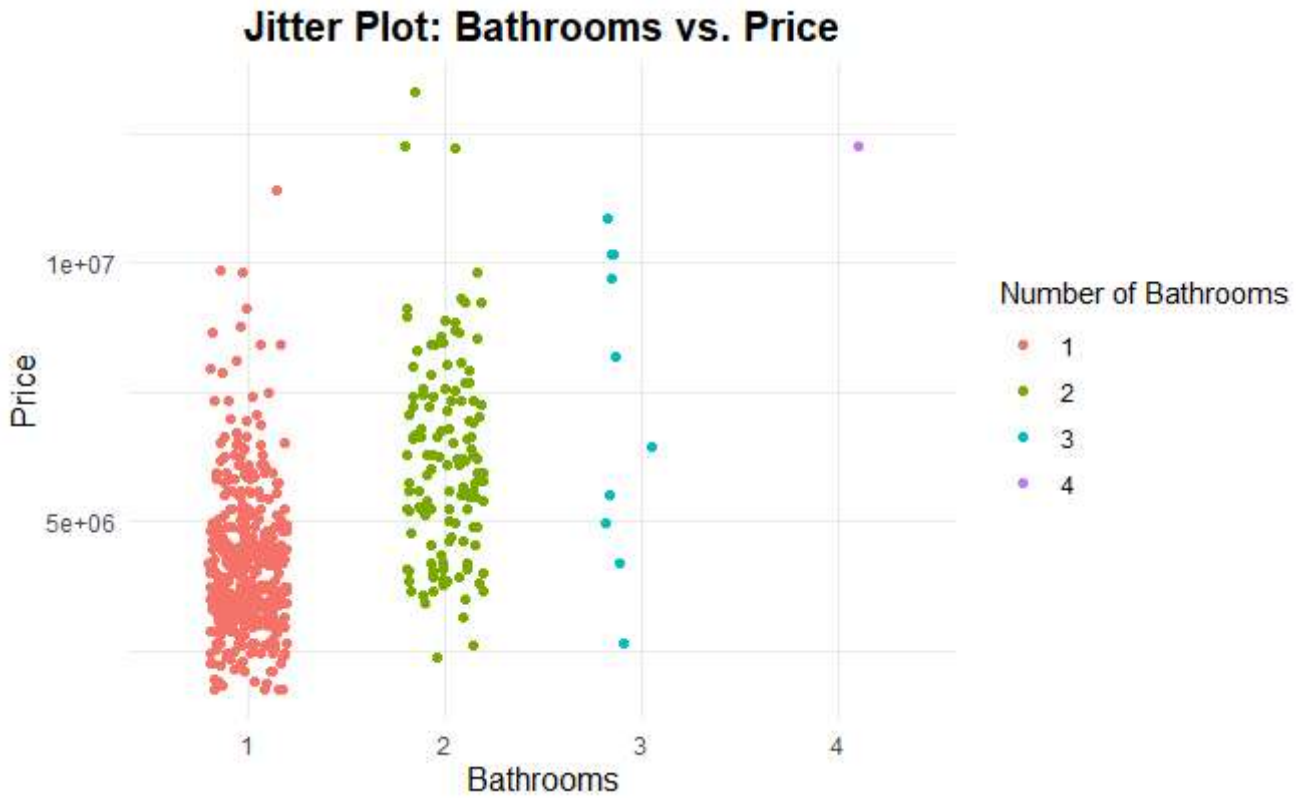
Observations:

- Properties with more parking spaces (e.g., 2 or more) generally tend to have higher prices, as observed by the concentration of points at higher price ranges.
- Most properties seem to have 1 or 2 parking spaces, and the prices for these properties are more densely packed.
- There are some outliers where properties with fewer parking spaces have significantly higher prices, possibly due to other factors like location or property size.

Hide

```
# Ensure bathrooms is treated as a factor
data$bathrooms <- as.factor(data$bathrooms)

ggplot(data, aes(x = bathrooms, y = price)) +
  geom_jitter(width = 0.2, height = 0, aes(color = bathrooms)) +
  labs(title = "Jitter Plot: Bathrooms vs. Price",
       x = "Bathrooms",
       y = "Price") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 15, face = "bold", hjust = 0.5),  # Title styling
    axis.title.x = element_text(size = 12),  # X-axis title styling
    axis.title.y = element_text(size = 12)   # Y-axis title styling
  ) +
  scale_color_discrete(name = "Number of Bathrooms")  # Use discrete color scale
```



Jitter Plot: Bathrooms vs. Price

Question Answered:

How does the number of bathrooms influence the price of a property?

Observations:

- Properties with more bathrooms (e.g., 2 or more) generally show a trend toward higher prices. The concentration of higher-priced properties increases as the number of bathrooms rises.
- For properties with fewer bathrooms (1-2), there is a wider range of prices, meaning that properties with the same number of bathrooms can have very different prices depending on other factors.

- This indicates that while the number of bathrooms influences price, other features like area, location, and amenities also play a significant role.