ADV

Code ▾

NAME - OM CHANDRA

UID - 2021700014

BATCH - L

EXPT - 5

Hide

```
housing_data <- read.csv("C:\\Users\\Om Chandra\\Downloads\\archive (22)\\Housing.csv", header = TRUE, stringsAsFactors = FALSE)
```

Hide

```
head(housing_data)
```

Hide

```
summary(housing_data)
```

```
       id                date               price            bedrooms
 Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000
 1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000
 Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000
 Mean   :4.580e+09                      Mean   : 540089   Mean   : 3.371
 3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000
 Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000
   bathrooms      sqft_living       sqft_lot           floors
 Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
 1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000
 Median :2.250   Median : 1910   Median :   7618   Median :1.500
 Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494
 3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000
 Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500
   waterfront            view          condition         grade
 Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
 1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
 Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000
 Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657
 3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
 Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000
   sqft_above   sqft_basement       yr_built     yr_renovated
 Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0
 1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0
 Median :1560   Median :   0.0   Median :1975   Median :   0.0
 Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4
 3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0
 Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0
    zipcode           lat             long         sqft_living15
 Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399
 1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
 Median :98065   Median :47.57   Median :-122.2   Median :1840
 Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987
 3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
 Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210
   sqft_lot15
 Min.   :   651
 1st Qu.:  5100
 Median :  7620
 Mean   : 12768
 3rd Qu.: 10083
 Max.   :871200
```

Hide

```
install.packages("lubridate")
library(lubridate)
```

Hide

```
dataset <- na.omit(housing_data)
```

Hide

```
housing_data$date <- ymd(substr(housing_data$date, 1, 8))
```

Hide

```
head(housing_data)
```

	id	date	price	bedroo...	bathroo...	sqft_living	sqft_lot	floors	wate
	<dbl>	<date>	<dbl>	<int>	<dbl>	<int>	<int>	<dbl>	
1	7229300521	2014-10-13	231300	2	1.00	1180	5650	1	
2	6414100192	2014-12-09	538000	3	2.25	2570	7242	2	
3	5631500400	2015-02-25	180000	2	1.00	770	10000	1	
4	2487200875	2014-12-09	604000	4	3.00	1960	5000	1	
5	1954400510	2015-02-18	510000	3	2.00	1680	8080	1	
6	7237550310	2014-05-12	1225000	4	4.50	5420	101930	1	

6 rows | 1-10 of 21 columns

Hide

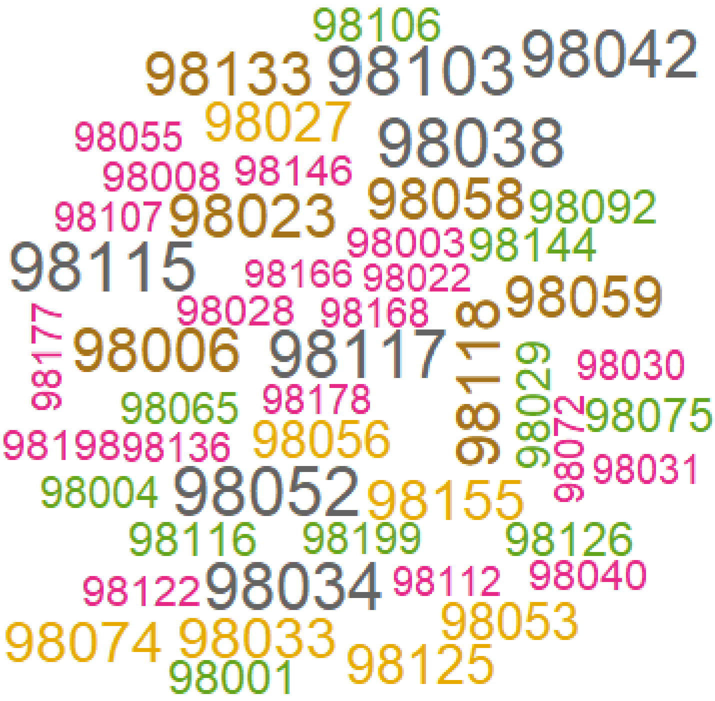
```
install.packages("wordcloud")
```

Hide

```
library(wordcloud)
library(dplyr)

# Generate the Word Cloud for Zipcodes
zipcode_count <- housing_data %>%
  count(zipcode) %>%
  arrange(desc(n))

# Create word cloud
wordcloud(words = zipcode_count$zipcode,
          freq = zipcode_count$n,
          min.freq = 1,
           max.words = 50,
          scale = c(2.2, 0.5),
          colors = brewer.pal(8, "Dark2"))
```



1.)Word Cloud for Zipcodes

Observations:

i.)Zip codes with larger font sizes are more frequently represented in the dataset, indicating higher occurrence.

ii.)Here Zipcodes 98115 ,98103 ,98034 ,etc have large size which means people tend to buy house in this zipcode ,the reason could be good area,affordable prices.

Questions Answered:

i.)Which zip codes are most prevalent in the dataset?In which area people tend to buy houses more?

ii.)How does the frequency of different zip codes compare visually?

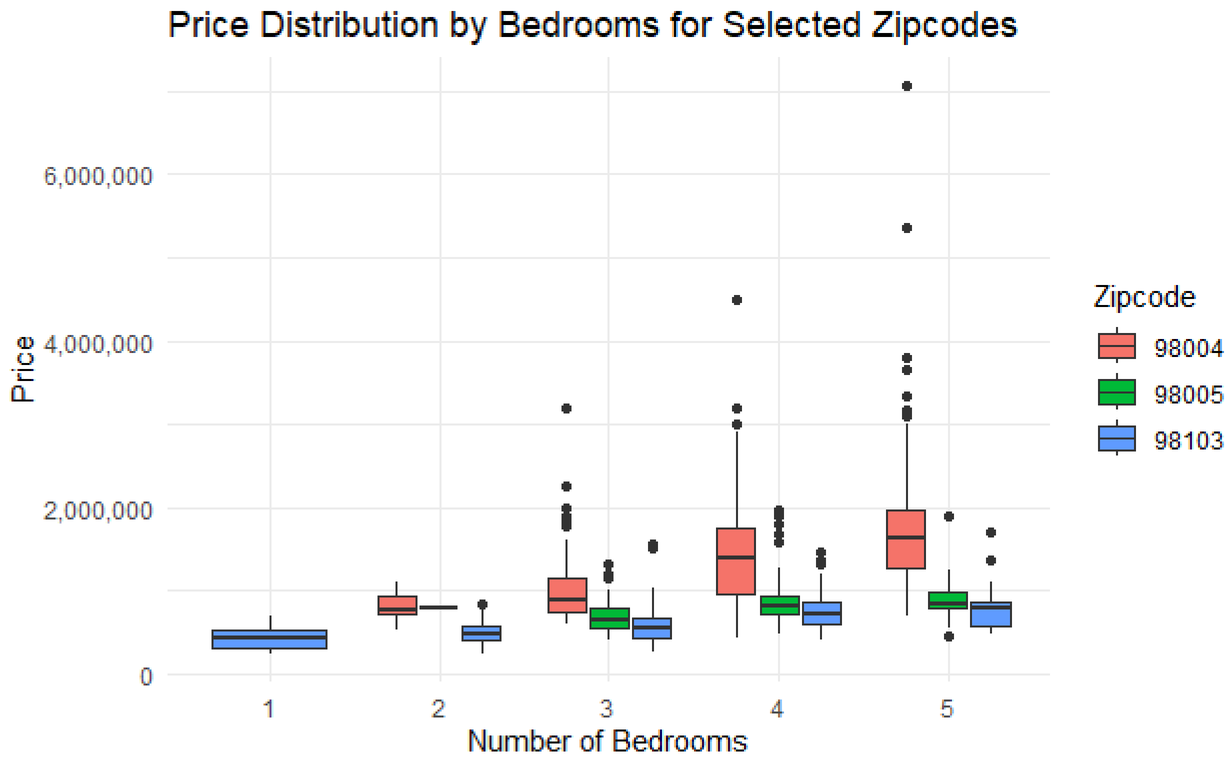
Hide

```
library(ggplot2)
library(scales)  # For formatting y-axis labels

# Define the zip codes to include
zipcodes_to_include <- c(98103, 98004, 98005)

# Filter the data for the specified zip codes and bedrooms up to 10
filtered_data <- housing_data %>%
  filter(zipcode %in% zipcodes_to_include, bedrooms <= 5)

# Create the boxplot
ggplot(filtered_data, aes(x = factor(bedrooms), y = price, fill = factor(zipcode))) +
  geom_boxplot() +
  labs(title = "Price Distribution by Bedrooms for Selected Zipcodes",
       x = "Number of Bedrooms",
       y = "Price",
       fill = "Zipcode") +
  scale_y_continuous(labels = scales::comma) +  # Format y-axis labels with commas
  theme_minimal()
```



Hide

```
NA
NA
NA
```

2.)Boxplot: Price Distribution by Bedrooms for Selected Zipcodes

Observations:

i.)Houses with more bedrooms generally have higher prices.There is a visible increase in the median price as the number of bedrooms increases.

ii.)The variability in prices is greater for houses in zipcode 98004, as indicated by the wider interquartile ranges.This indicates that the posh area tend to have greater variability in prices.

iii.)We can also see zipcode 98004 and 98005 have no houses with 1 bedroom indicating that the area is a deluxe area.

Questions Answered:

i.)How does the price of houses vary with the number of bedrooms in selected zip codes?

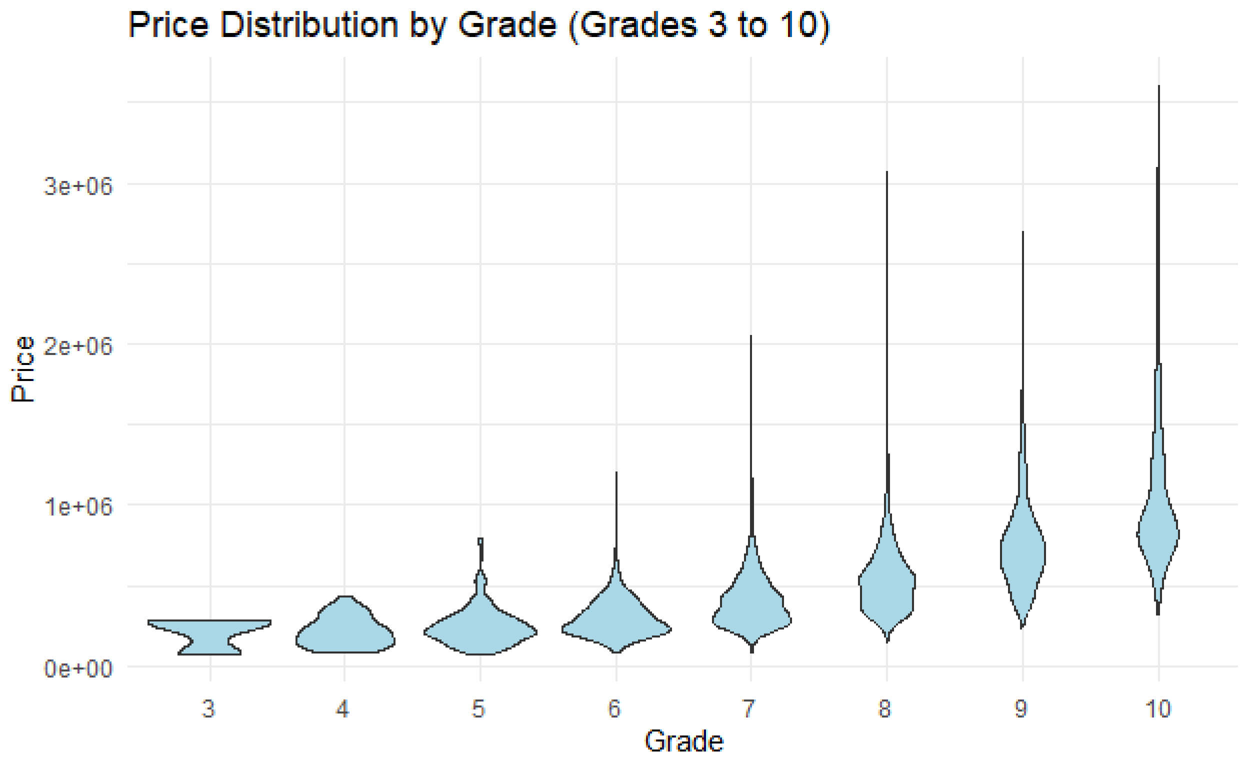
ii.)Are there noticeable trends or patterns in price distribution for different bedroom counts?

Hide

```
library(ggplot2)
library(dplyr)  # Ensure dplyr is loaded for data manipulation


filtered_data <- housing_data %>%
  filter(grade >= 3 & grade <= 10)

# Create the violin plot with the filtered data
ggplot(filtered_data, aes(x = factor(grade), y = price)) +
  geom_violin(fill = "lightblue", drop = FALSE) +
  labs(title = "Price Distribution by Grade (Grades 3 to 10)",
       x = "Grade",
       y = "Price") +
  theme_minimal()
```



3.)Violin Plot: Price Distribution by Grade (Grades 3 to 10)

Observations:

i.)Higher-grade properties generally have higher median prices compared to lower-grade properties.

ii.)The range of prices is broader for higher-grade properties, indicating more variability.

iii.)The density of house prices is higher in the low to mid range of grades which indicates more people(as their are more common people) buy house of low-mid grade.

Questions Answered:

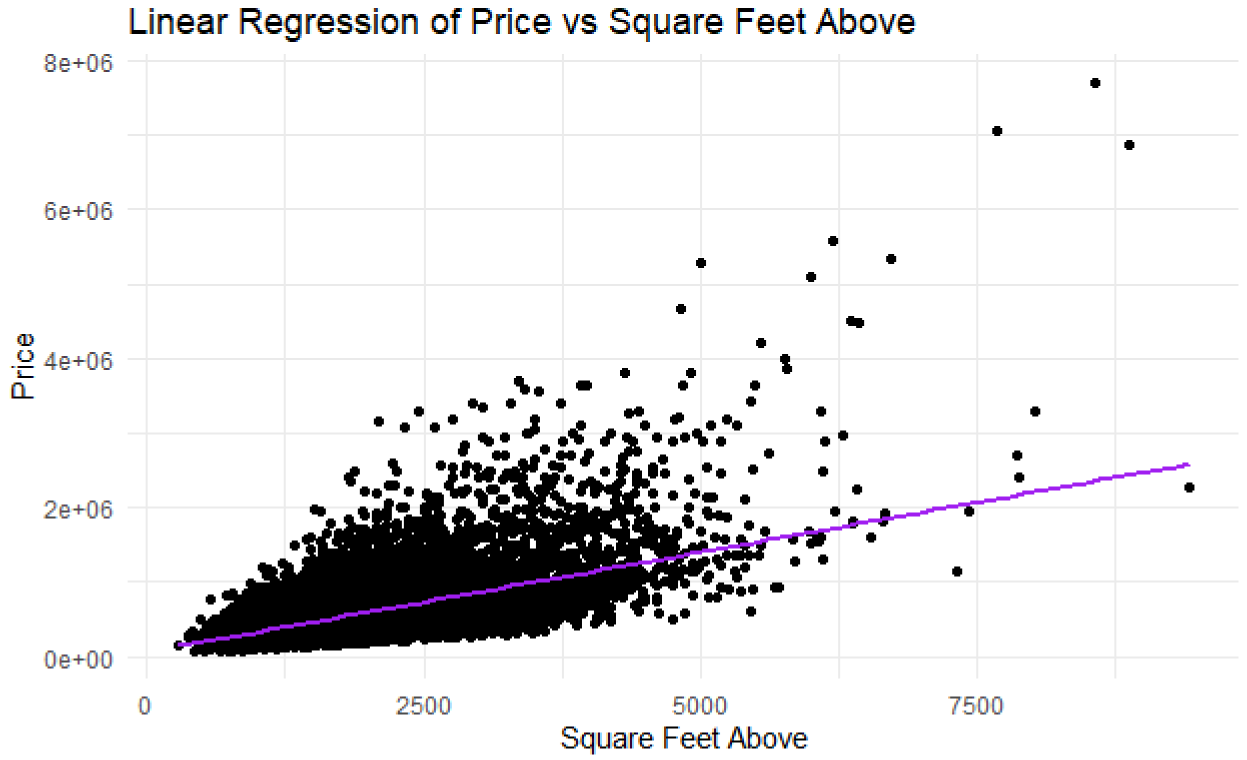
i.)How does the price distribution differ across various property grades?

ii.)What is the spread of house prices for each grade level?

Hide

```
ggplot(housing_data, aes(x = sqft_above, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", color = "purple", se = FALSE) +
  labs(
    title = "Linear Regression of Price vs Square Feet Above",
    x = "Square Feet Above",
    y = "Price"
  ) +
  theme_minimal()
```

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



4.)Linear Regression Plot: Price vs Square Feet Above

Observations:

i.)There is a positive correlation between price and square footage above ground.

ii.)The regression line shows that as the square footage above ground increases, the price also tends to increase except for some outliers. . iii.)The relationship appears linear, with a steady increase in price with more square footage above ground.

Questions Answered:

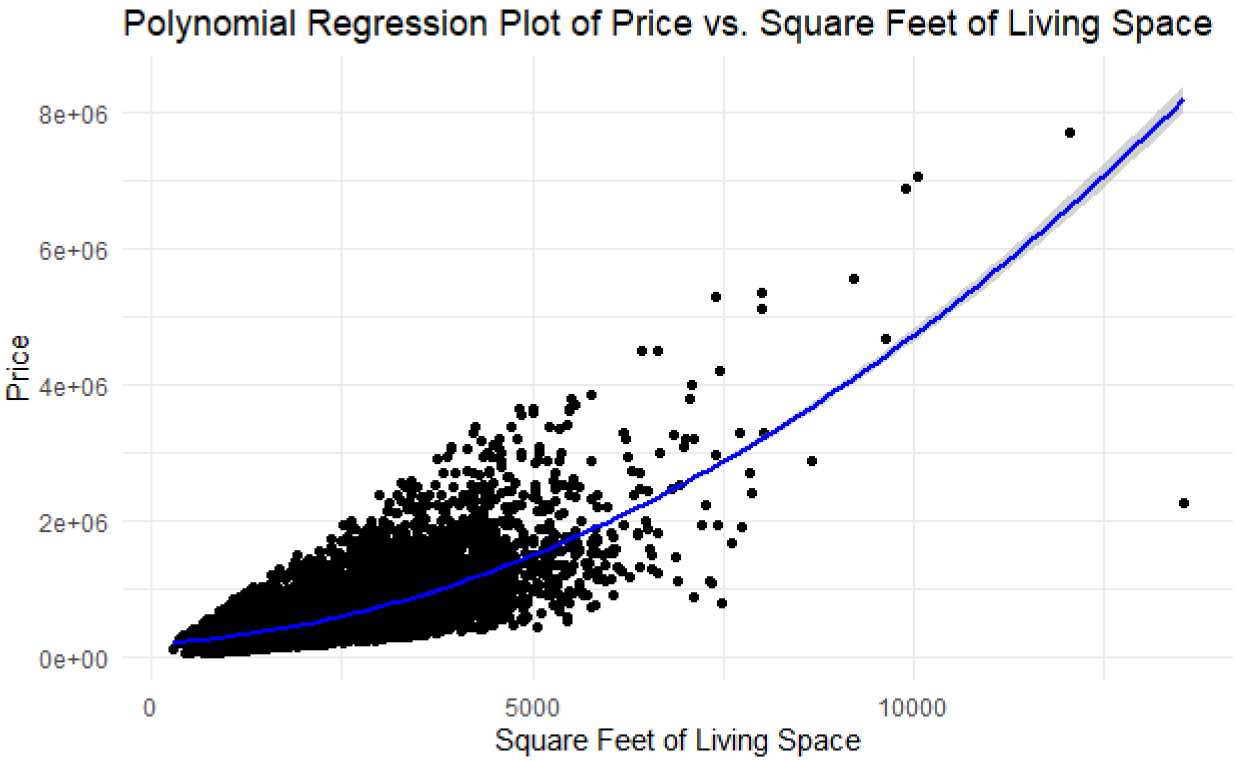
i.)What is the relationship between the price of a property and its square footage above ground?

ii.)Does increasing square footage above ground significantly affect property prices?

Hide

```
poly_model <- lm(price ~ poly(sqft_living, 2), data = housing_data)

# Create the plot
ggplot(housing_data, aes(x = sqft_living, y = price)) +
  geom_point(color = "black") +
  stat_smooth(method = "lm", formula = y ~ poly(x, 2), color = "blue", size = 1) +
  labs(title = "Polynomial Regression Plot of Price vs. Square Feet of Living Space",
       x = "Square Feet of Living Space",
       y = "Price") +
  theme_minimal()
```



5.)Polynomial Regression Plot: Price vs Square Feet of Living Space

Observations:

i.)The polynomial regression curve shows a non-linear relationship between price and square footage of living space.

ii.)We can see as the square foot of living space increases , the prices increase more non-linearly.

iii.)The relationship is more complex than a simple linear trend, suggesting other factors influence price.

Questions Answered:

i.)How does the price of a property relate to its square footage of living space in a non-linear fashion?

ii.)Is square footage of living space is the only factor affecting price of a house?

Hide

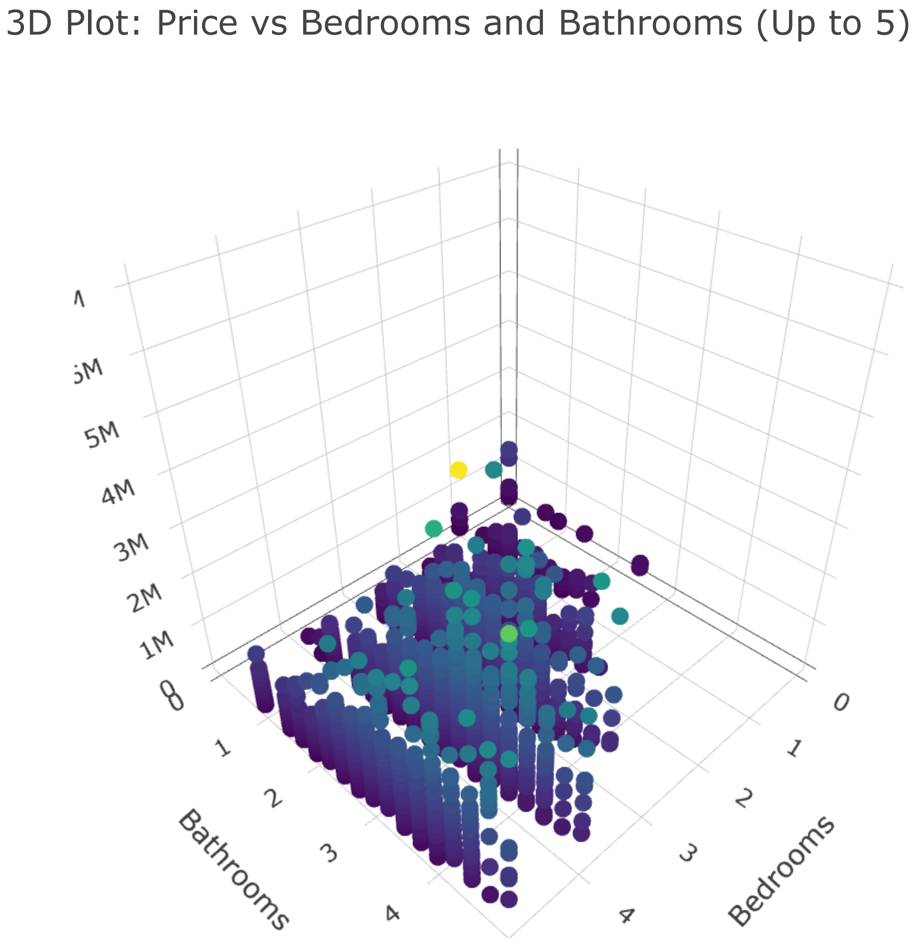
```
install.packages("plotly")
```

Hide

```
library(plotly)
library(dplyr)

# Filter the data for bedrooms and bathrooms up to 5
filtered_data_3d <- housing_data %>%
  filter(bedrooms <= 5, bathrooms <= 5)

# Create the 3D scatter plot
plot_ly(filtered_data_3d, x = ~bedrooms, y = ~bathrooms, z = ~price,
        type = "scatter3d", mode = "markers",
        marker = list(size = 5, color = ~price, colorscale = "Viridis")) %>%
  layout(title = "3D Plot: Price vs Bedrooms and Bathrooms (Up to 5)",
         scene = list(xaxis = list(title = 'Bedrooms'),
                      yaxis = list(title = 'Bathrooms'),
                      zaxis = list(title = 'Price')))
```



Hide

```
NA
NA
```

6.)3D Scatter Plot: Price vs Bedrooms and Bathrooms

Observations:

i.)Properties with more bedrooms and bathrooms tend to have higher prices.

ii.)The 3D plot shows a clear spread of data points, indicating how price varies with both variables.

iii.)There is a noticeable upward trend in price with increasing numbers of bedrooms and bathrooms except for some outliers.

iv.)We can also see a linear relationship between no. of bedrooms and no.of bathrooms except for some cases.

Questions Answered:

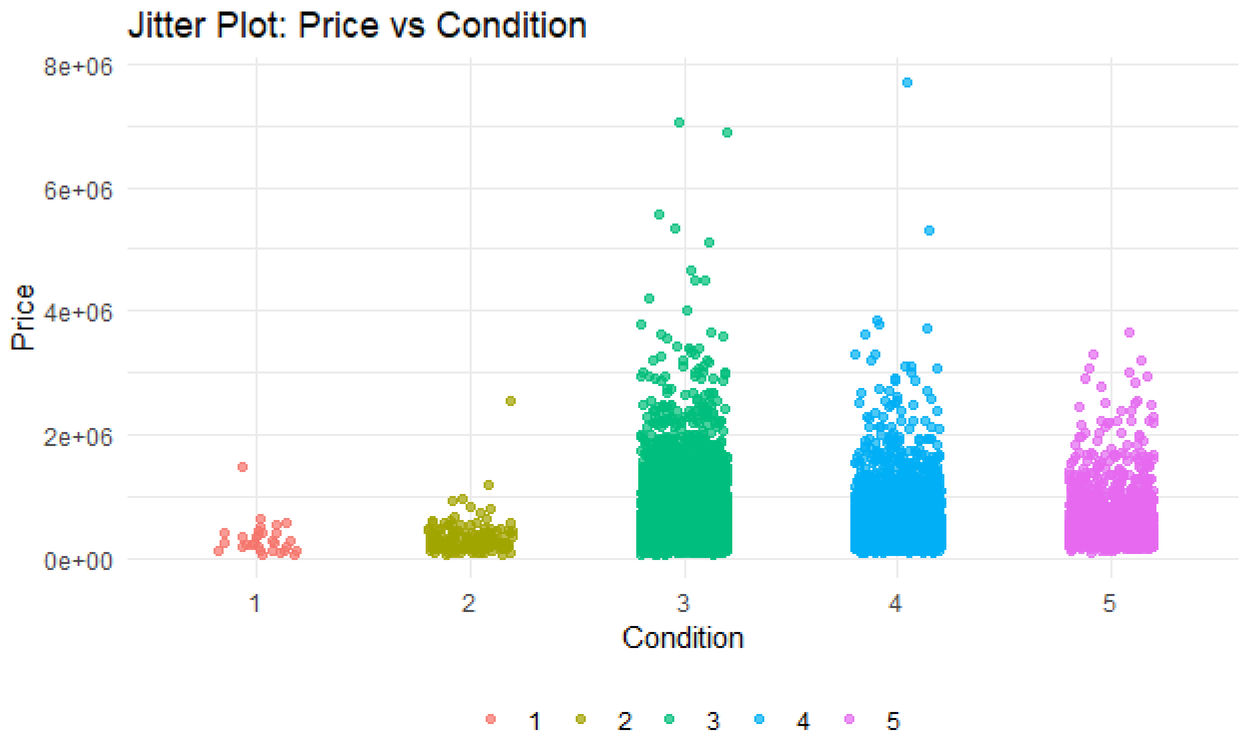
i.)How does the price of a property vary with the number of bedrooms and bathrooms?

ii.)What is the combined effect of bedrooms and bathrooms on property prices?

Hide

```
library(ggplot2)

# Jitter plot for price vs condition
ggplot(housing_data, aes(x = as.factor(condition), y = price, color = as.factor(condition)))
+
  geom_jitter(width = 0.2, height = 0, alpha = 0.7) +
  labs(x = "Condition", y = "Price", title = "Jitter Plot: Price vs Condition") +
  theme_minimal() +
  theme(legend.title = element_blank(), legend.position = "bottom")
```



7.)Jitter Plot: Price vs Condition

Observations:

i.)Prices are spread across different property conditions, showing variability within each condition.

ii.)Higher property conditions(3 to 5) tend to have a broader range of prices.

iii.)The spread of points suggests that price can vary significantly even within the same condition category for example condition 3 and 4.

Questions Answered:

i.)How does property condition impact pricing?

ii.)Are there significant differences in prices among different property conditions?