AIM: Create advanced charts using R programming language on the dataset - Housing data

DATASET: Housing Price Dataset: https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/data

This dataset represents housing prices, featuring attributes such as area, number of bedrooms, bathrooms, stories, and additional amenities like parking, basement, and air conditioning providing an overview of various factors influencing house prices.

ANALYSIS:

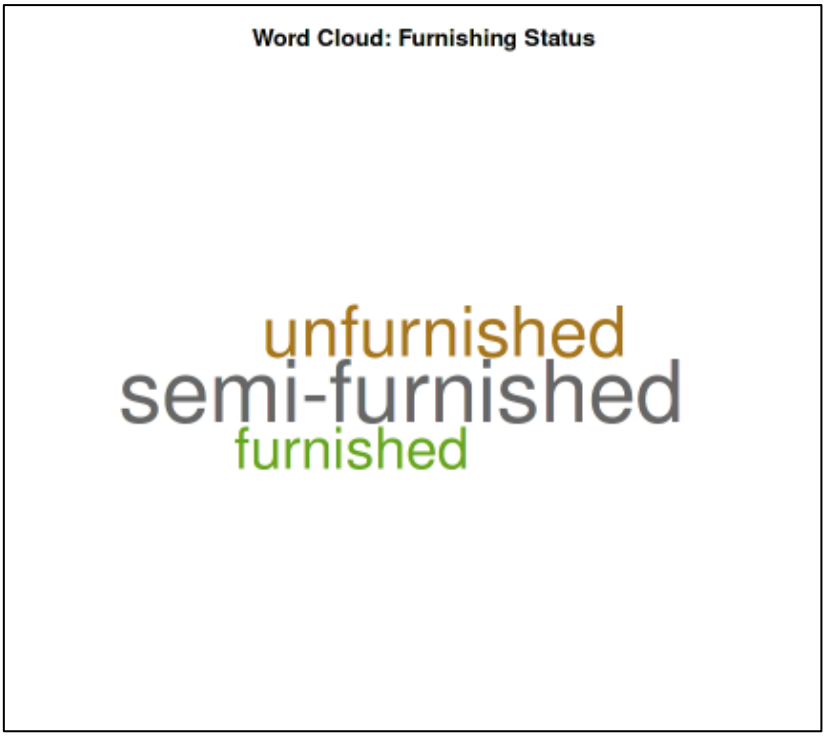
```
library(wordcloud)
library(dplyr)
library(RColorBrewer)
library(ggplot2)
library(plotly)
library(reshape2)

data <- read.csv('/kaggle/input/housing-prices-dataset/Housing.csv')
head(data)
```

A data.frame: 6 × 13

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
	<int>	<int>	<int>	<int>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	<chr>	<chr>
1	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
2	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
3	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
4	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
5	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished
6	10850000	7500	3	3	1	yes	no	yes	no	yes	2	yes	semi-furnished

1] Word Chart/ Word Cloud

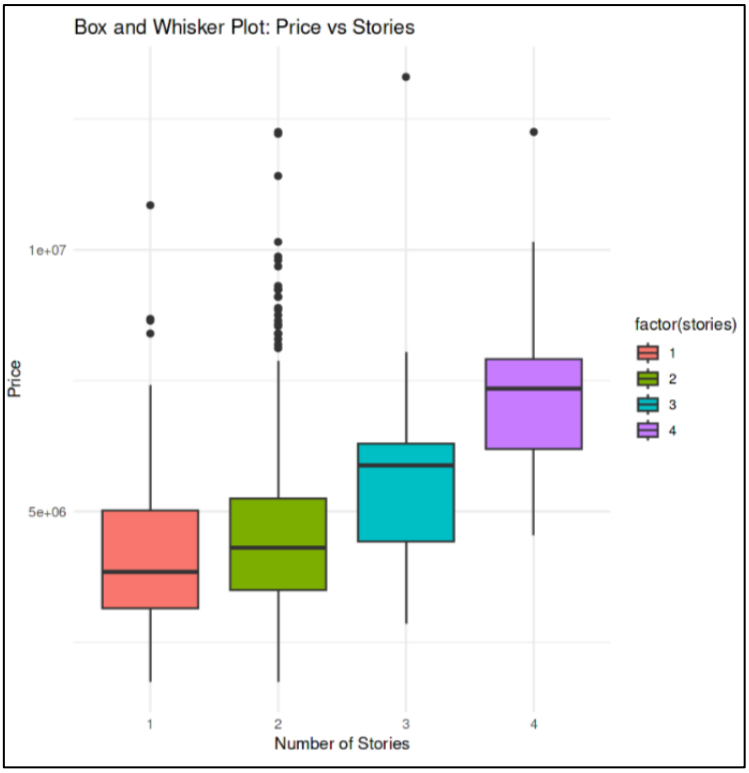
```
furnishing_status_freq <- data %>% count(furnishingstatus)
colors <- brewer.pal(8, "Dark2")

wordcloud(words = furnishing_status_freq$furnishingstatus,
          freq = furnishing_status_freq$n,
          min.freq = 1,colors = colors,
          random.order = FALSE,
          scale = c(4.5, 0.5))

title(main = "Word Cloud: Furnishing Status")
```



The above word cloud states that the majority of the houses that are sold are semi-furnished and furnished homes represent the smallest proportion.

2] Box and whisker plot

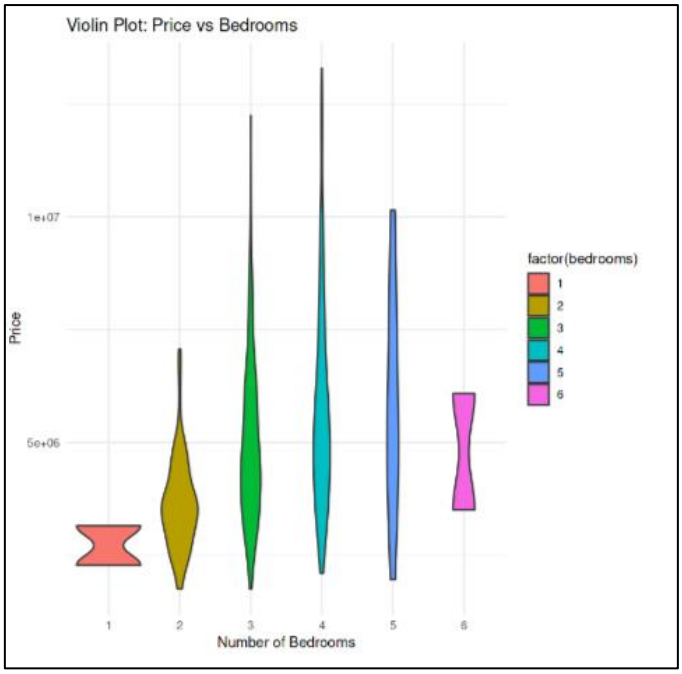
```
box_plot <- ggplot(data, aes(x = factor(stories), y = price)) +
  geom_boxplot(aes(fill = factor(stories))) +
  theme_minimal() +
  labs(title = "Box and Whisker Plot: Price vs Stories", x = "Number of Stories", y = "Price")
print(box_plot)
```



The above Box and Whisker plot states that houses with more stories tend to have a higher price range. There's some price overlap, but outliers are seen particularly in 2-story houses. The highest sold house was of 3-story.

3] Violin plot

```
violin_plot <- ggplot(data, aes(x = factor(bedrooms), y = price)) +
  geom_violin(aes(fill = factor(bedrooms))) +
  theme_minimal() +
  labs(title = "Violin Plot: Price vs Bedrooms", x = "Number of Bedrooms", y = "Price")
print(violin_plot)
```

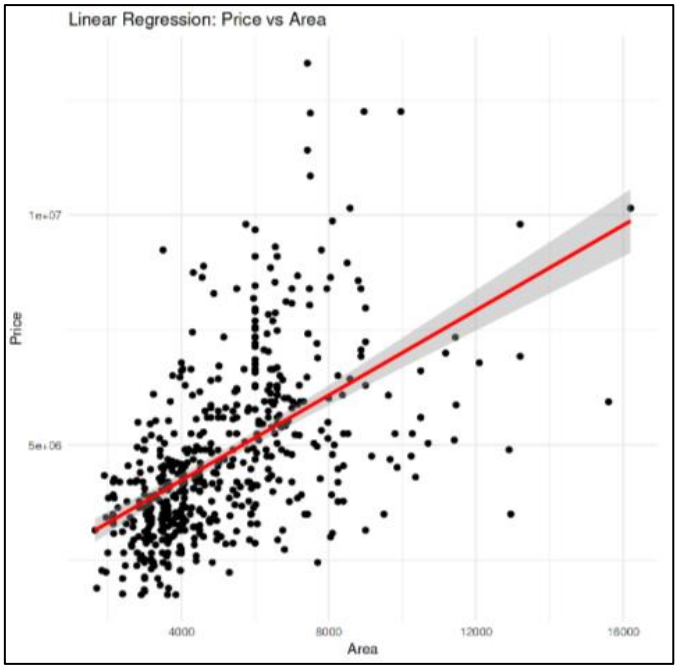


The above Violin plot shows that houses with 3 or 4 bedrooms show a wider price range and houses with fewer or more bedrooms tend to have a narrower range in prices. The highest selling house was of 4 bedrooms.

4] Regression plot (linear and nonlinear)

→ Linear

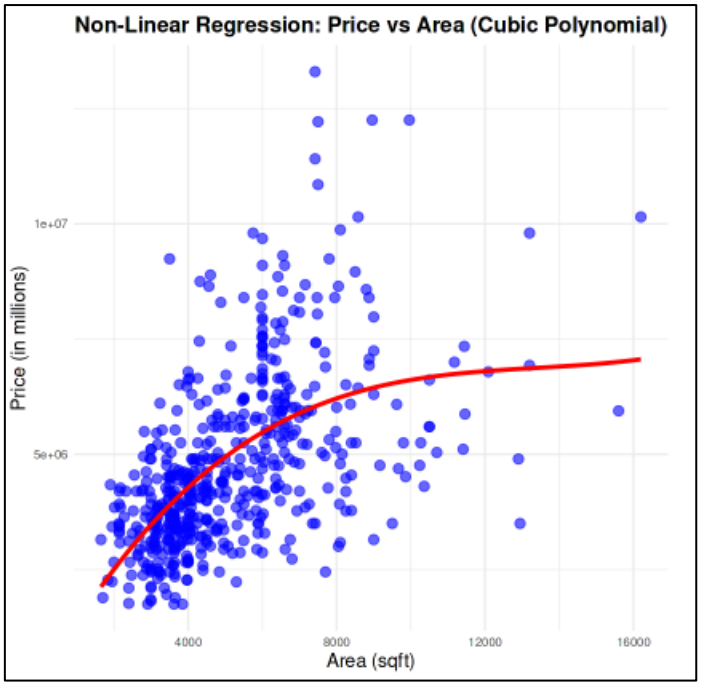
```
linear_reg <- ggplot(data, aes(x = area, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  theme_minimal() +
  labs(title = "Linear Regression: Price vs Area", x = "Area", y = "Price")
print(linear_reg)
```



The above regression plot shows that larger area generally results in a higher price. But area is not the only factor to determine price which can be clearly seen by the above plot.

→ Non-Linear

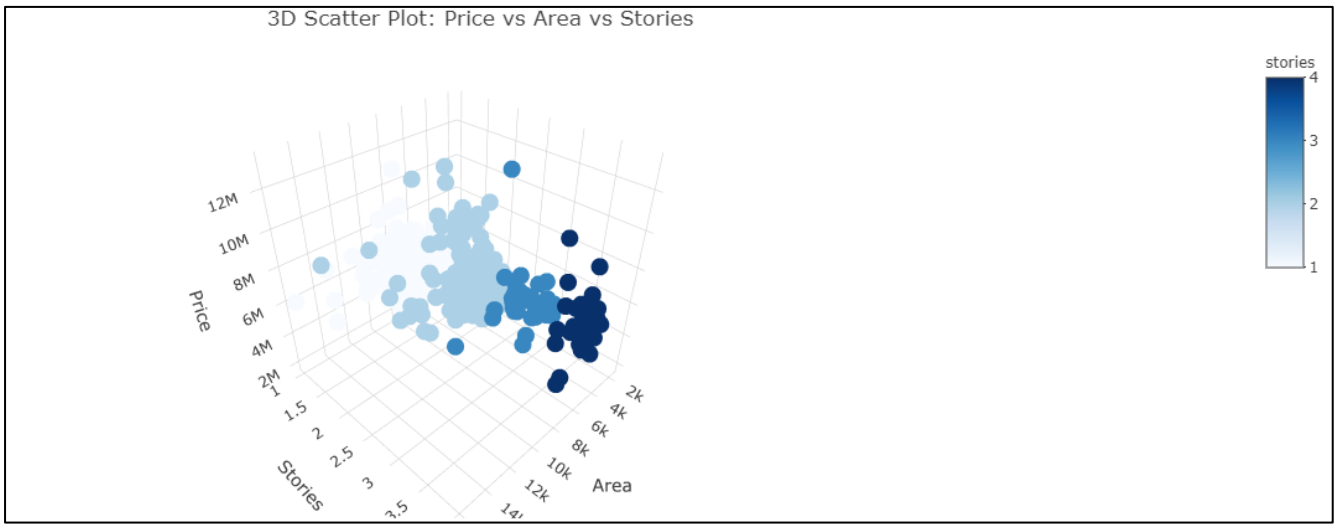
```
model <- lm(price ~ poly(area, 3), data = data)
data$predicted_price <- predict(model)

non_linear_plot <- ggplot(data, aes(x = area, y = price)) + geom_point(color = "blue", alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 3), color = "red", se = FALSE, size = 1.2) +
  theme_minimal() +
  labs(title = "Non-Linear Regression: Price vs Area (Cubic Polynomial)",
       x = "Area (sqft)",
       y = "Price (in millions)") +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14))

print(non_linear_plot)
```



The above regression plot shows that the relationship between Area and Price is non-linear. The price appears to rise quickly for moderate increases in area but flattens as area continues to increase.

5] 3D chart

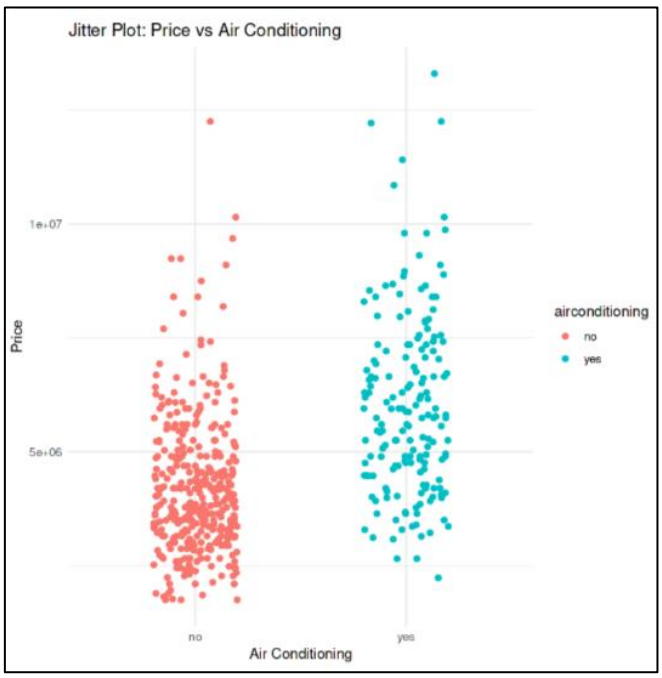
```
plot_3d <- plot_ly(data, x = ~area, y = ~stories, z = ~price,
                   color = ~stories, colors = brewer.pal(n = 9, name = "Blues"),
                   type = "scatter3d",mode = "markers") %>%
  layout(scene = list(xaxis = list(title = 'Area'),
                      yaxis = list(title = 'Stories'),
                      zaxis = list(title = 'Price')),
         title = "3D Scatter Plot: Price vs Area vs Stories")
plot_3d
```



The above 3D plot shows that houses with more stories tend to have higher prices and larger areas also generally lead to higher prices, but there are exceptions.

6] Jitter plot

```
jitter_plot <- ggplot(data, aes(x = airconditioning, y = price)) +
  geom_jitter(aes(color = airconditioning), width = 0.2, height = 0) + theme_minimal() +
  labs(title = "Jitter Plot: Price vs Air Conditioning", x = "Air Conditioning", y = "Price")
print(jitter_plot)
```



The above jitter plot shows that Price and air conditioning has no definite relationship. Most of the houses sold were with no air-conditioning. The highest sold house was with air conditioning.