NAME:	Sakshi Bisen
UID No.	2021600008
BRANCH:	B.E CSE-AIML
BATCH:	I
SUBJECT	Advanced Data Visualization
EXPERIMENT No.	5
DATE:	19/9/2024

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

DATASET: House Pricing Dataset→ The dataset contains information about property listings, including price, area, number of bedrooms and bathrooms, availability of amenities like parking and air conditioning, and the furnishing status.

Link to Dataset: https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/data

ANALYSIS:

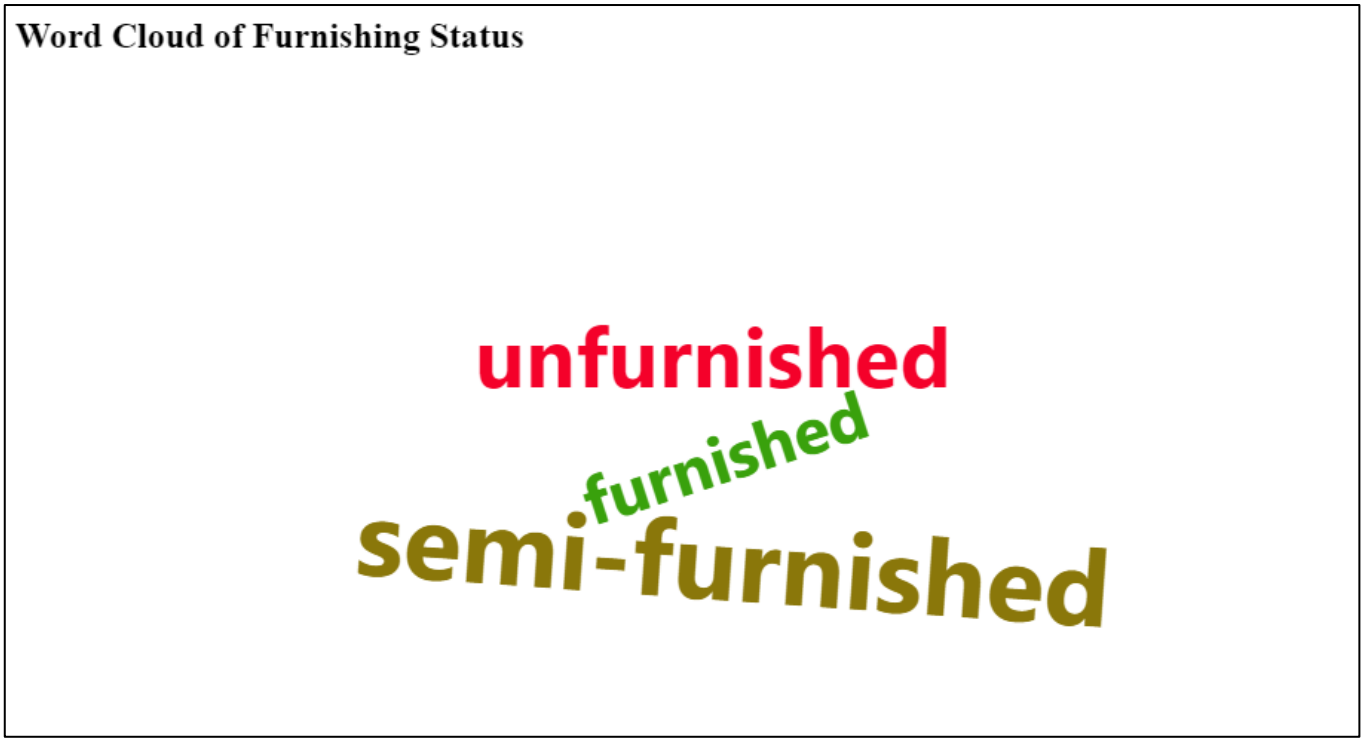
```
library(wordcloud)
library(dplyr)
library(RColorBrewer)
library(ggplot2)
library(plotly)
library(reshape2)
library(wordcloud2)
library(htmlwidgets)

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

1] Word Chart/ Word Cloud

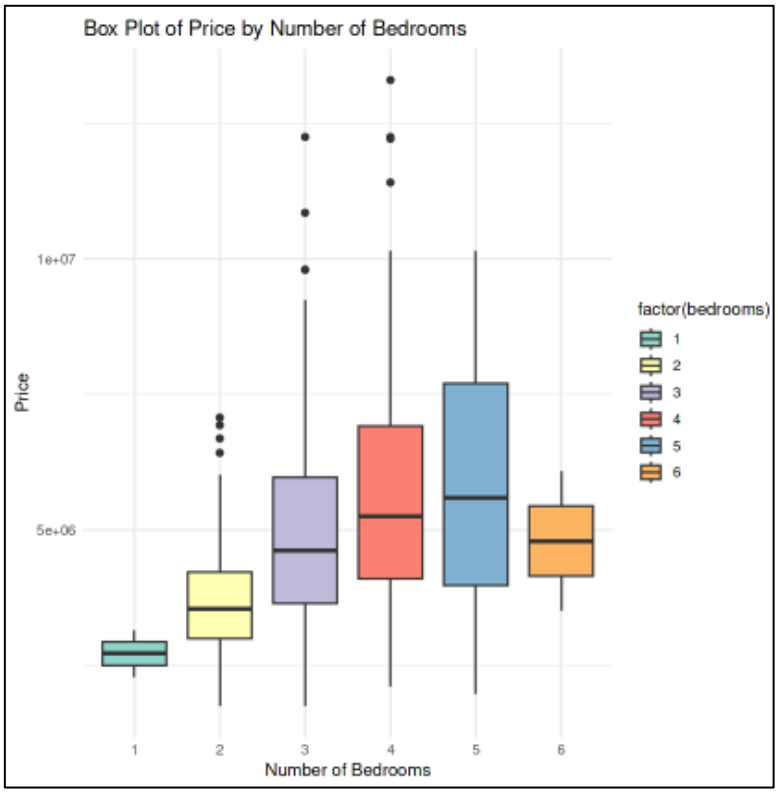
```
wordcloud_data <- as.data.frame(table(data$furnishingstatus))
wordcloud <- wordcloud2(wordcloud_data, size = 0.4)
wordcloud <- htmlwidgets::prependContent(wordcloud, htmltools::tags$h2("Word Cloud of Furnishing Status"))
wordcloud
```



The above word cloud shows that a large population prefer semi-furnished houses for easy cutomization and furnished homes represent the smallest proportion.

2] Box and whisker plot

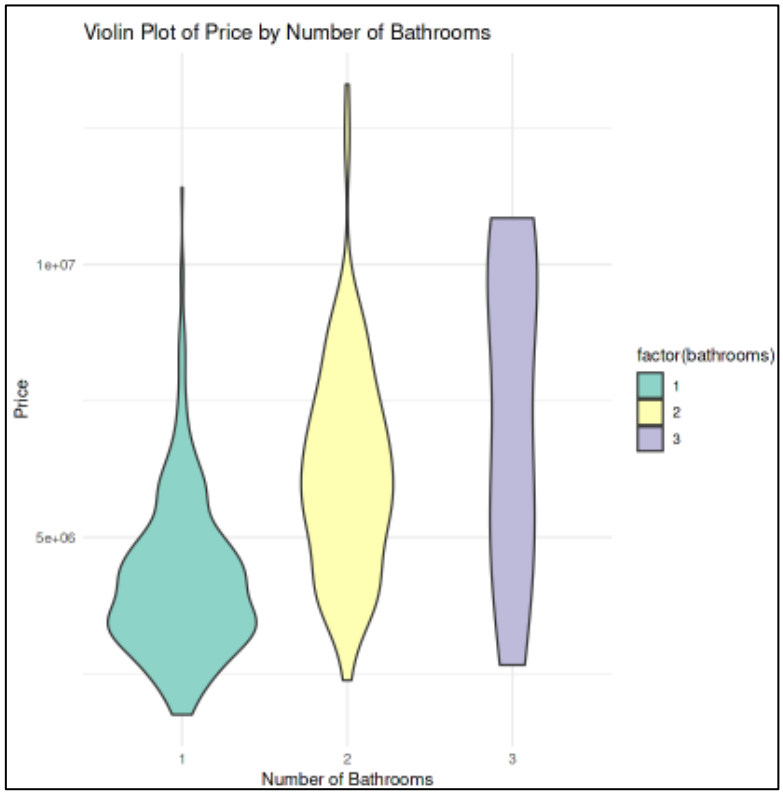
```
boxplot_price <- ggplot(data, aes(x = factor(bedrooms), y = price, fill = factor(bedrooms))) +
  geom_boxplot() + scale_fill_brewer(palette = "Set3") +
  labs(title = "Box Plot of Price by Number of Bedrooms", x = "Number of Bedrooms", y = "Price") +
  theme_minimal()
print(boxplot_price)
```



The above Box and Whisker plot shows that houses with more bedrooms tend to have a higher price. It is observed that house with 3,4,5 bedrooms we sold at higher price than the one with 6 bedrooms. Outliers are seen particularly in 2,3, and 4 bedroom houses. The highest sold house was of 4 bedrooms.

3] Violin plot

```
filtered_data <- data %>% group_by(bathrooms) %>% filter(n() > 1)
violin_plot <- ggplot(filtered_data, aes(x = factor(bathrooms), y = price, fill = factor(bathrooms))) +
  geom_violin() + scale_fill_brewer(palette = "Set3") +
  labs(title = "Violin Plot of Price by Number of Bathrooms", x = "Number of Bathrooms", y = "Price") +
  theme_minimal()
print(violin_plot)
```

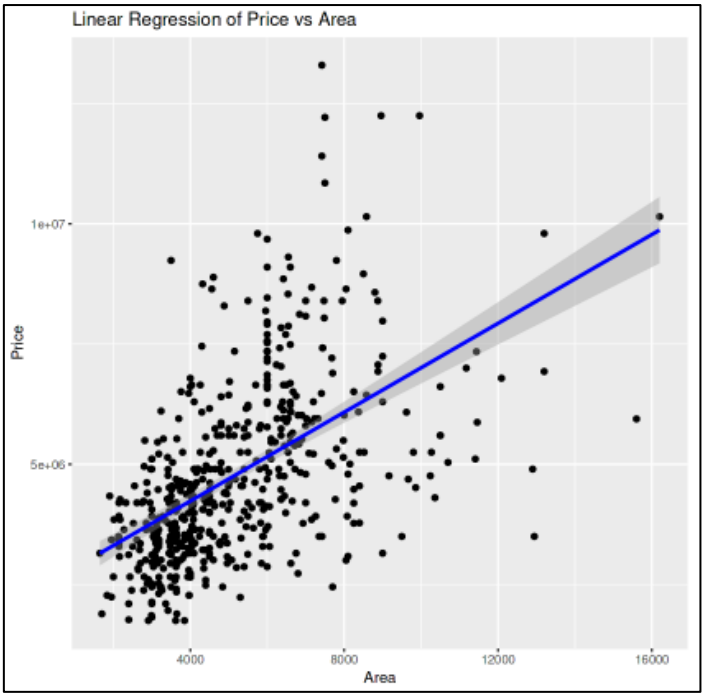


The above Violin plot states that houses with 1 and 2 bathrooms show a wider price range and houses with fewer bathrooms tend to have a narrower range in prices. The highest selling house was the one with 2 bathrooms.

4] Regression plot (linear and nonlinear)

→ Linear

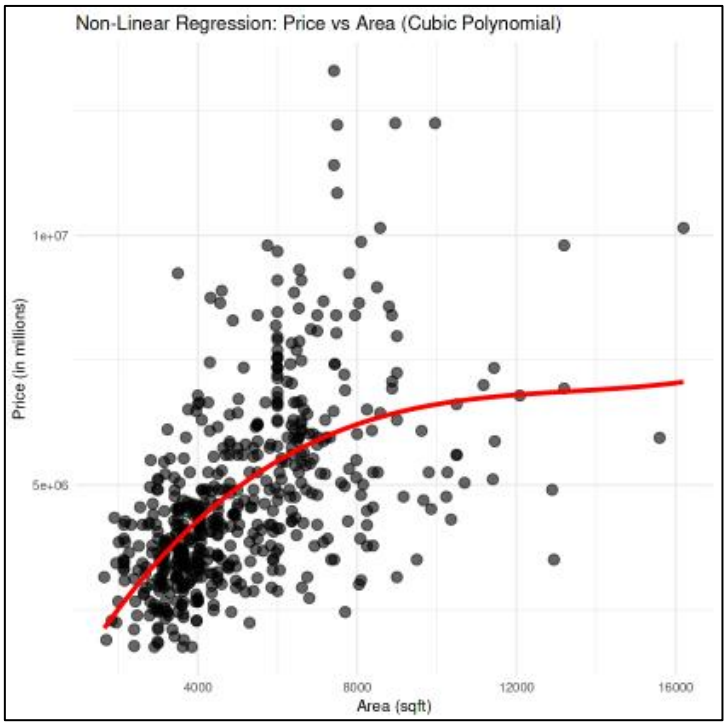
```
linear_plot <- ggplot(data, aes(x = area, y = price)) +
  geom_point() + geom_smooth(method = "lm", color = "blue") +
  labs(title = "Linear Regression of Price vs Area", x = "Area", y = "Price")
print(linear_plot)
```



The linear regression plot suggests a positive correlation between area and price, indicating that larger properties tend to be priced higher.

→ Non-Linear

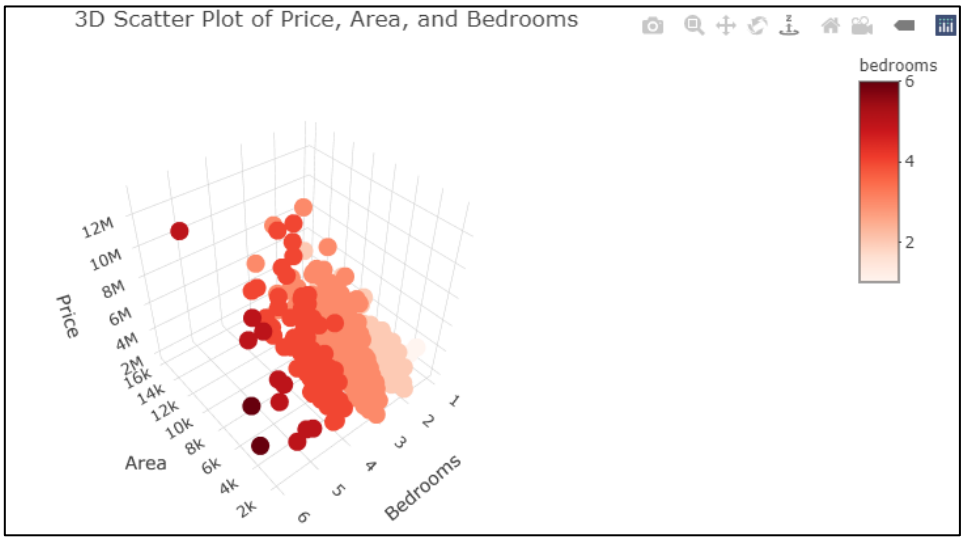
```
model <- lm(price ~ poly(area, 3), data = data)
data$predicted_price <- predict(model)
non_linear_plot <- ggplot(data, aes(x = area, y = price)) + geom_point( alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 3), color = "red", se = FALSE, size = 1.2) +
  theme_minimal() + labs(title = "Non-Linear Regression: Price vs Area (Cubic Polynomial)",
       x = "Area (sqft)", y = "Price (in millions)")

print(non_linear_plot)
```



The non-linear regression plot provides a better fit for the relationship between area and price, highlighting potential price increases at a non-linear rate as area increases and saturates after a point.

5] 3D chart

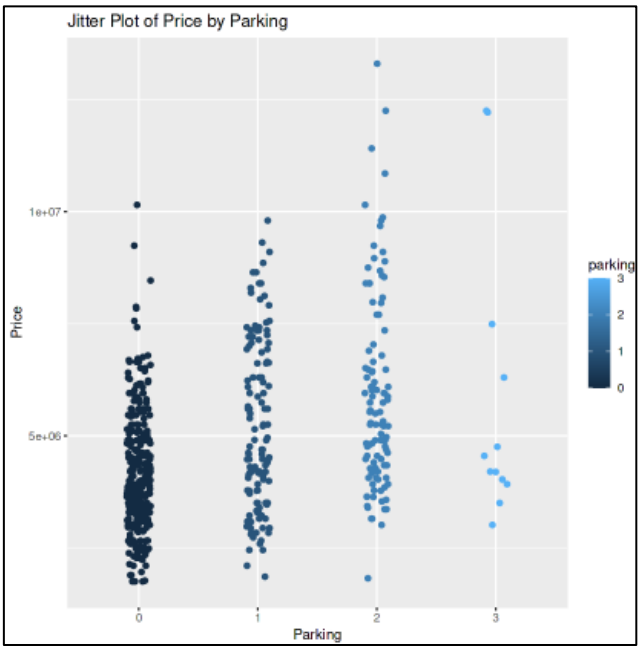
```
fig <- plot_ly(data, x = ~area, y = ~bedrooms, z = ~price, type = "scatter3d", mode = "markers",
               color = ~bedrooms, colors = brewer.pal(n = 9, name = "Reds") )
fig <- fig %>% layout(title = "3D Scatter Plot of Price, Area, and Bedrooms",
                      scene = list(xaxis = list(title = "Area"),
                                   yaxis = list(title = "Bedrooms"),
                                   zaxis = list(title = "Price")))
fig
```



The 3D scatter plot shows a general trend of higher prices for properties with larger areas and more bedrooms. However, there are also many outliers, indicating that other factors besides area and bedrooms influence the price.

6] Jitter plot

```
jitter_plot <- ggplot(data, aes(x = factor(parking), y = price)) +
  geom_jitter(width = 0.1, height = 0, aes(color = parking)) +
  labs(title = "Jitter Plot of Price by Parking", x = "Parking", y = "Price")
print(jitter_plot)
```



The above jitter plot shows that Price and parking conditioning has some relation. Most of the houses sold were with no parking. The highest sold house was with 2 parking availability conditioning. People do not prefer to have a place with 3 parking.