Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India

Experiment no 5 **Name-Divya Suvarna** **UID-2021200114** **BATCH-A**

Aim :

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

- Advanced - Word chart, Box and whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, Jitter
- Write observations from each chart

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

To explore and visualize housing data using advanced charts in R, including Word chart, Box and Whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, and Jitter plot, in order to uncover patterns and insights in the dataset.

Objectives:

1. To visualize the distribution and relationship between various features in the housing dataset.
2. To identify potential outliers and understand the spread of the data.
3. To explore the relationship between independent variables and the target variable (e.g., house prices).
4. To create informative visualizations that can guide decision-making in the housing market.

Theory:

Data visualization is an essential skill in data analysis that helps in understanding trends, patterns, and relationships within a dataset. R, a powerful statistical programming language, provides a wide range of tools for creating visually appealing and informative charts. In this experiment, we will use advanced chart types to analyze crime data and derive insights.

Chart Types:

- **Word Chart (Word Cloud)**: A visual representation of text data where the size of each word indicates its frequency or importance in the dataset. Larger words represent higher frequency, making it easy to identify key themes or terms.
- **Box and Whisker Plot**: This plot displays the distribution of a dataset through its quartiles, highlighting the median, interquartile range (IQR), and potential outliers. The box represents the IQR, while the whiskers extend to the minimum and maximum values within 1.5 times the IQR.
- **Violin Plot**: Similar to a box plot but adds a rotated density plot on each side, showing the distribution of the data across different categories. It provides more information about the density of the data at different values, making it useful for comparing distributions.

- **Regression Plot**: This type of plot shows the relationship between two variables with a fitted line (linear or nonlinear) representing the trend. It helps visualize how one variable is predicted based on another, aiding in understanding correlations.
- **3D Chart**: A graphical representation that displays data points in three dimensions (x, y, and z axes). It allows for the visualization of complex datasets and relationships that might not be easily interpreted in two dimensions.
- **Jitter Plot**: A scatter plot that adds a small amount of random noise (jitter) to the data points along one or both axes. This technique helps prevent overplotting, making it easier to visualize the distribution of points when many points overlap.

Steps to Perform in R:

1. **Set Up the Environment:**
    - Install and load necessary libraries.

R

```
# Install required packages if not already installed
install.packages("ggplot2")
install.packages("plotly")
install.packages("car")  # For 3D charts
install.packages("wordcloud")
install.packages("dplyr")

# Load the libraries
library(ggplot2)
library(plotly)
library(car)
library(wordcloud)
library(dplyr)
```

2. **Load the Dataset:**
    - Load the crime dataset (replace crime_data.csv with your dataset's file name).

```
# Load dataset
housing_data <- read.csv("/content/Housing.csv")

# Preview the first few rows of the dataset
head(housing_data)
```

3. **Data Preprocessing:**
    - Inspect and clean the data if necessary (handle missing values, filter relevant columns, etc.).

R
Copy code

```
# Check for missing values
missing_values <- colSums(is.na(housing_data))
print(missing_values)

# Fill missing values (example: fill with mean for numerical and mode for categorical)
housing_data <- housing_data %>%
  mutate(across(where(is.numeric), ~ifelse(is.na(.), mean(., na.rm = TRUE), .))) %>%
  mutate(across(where(is.character), ~ifelse(is.na(.), as.character(names(sort(table(.),
decreasing = TRUE)[1])), .)))
```

4. **Create Visualizations:**

Word Cloud:

```
# Word Cloud for Furnishing Status
furnish_table <- table(housing_data$furnishingstatus)
wordcloud(names(furnish_table), furnish_table, colors=brewer.pal(8, "Dark2"))
```

- **Observation:**

Description: The word cloud visually represents the different categories of housing types based on their furnishing status—specifically "furnished," "semi-furnished," and "unfurnished." The size of each word in the cloud indicates its frequency or prominence in the dataset.

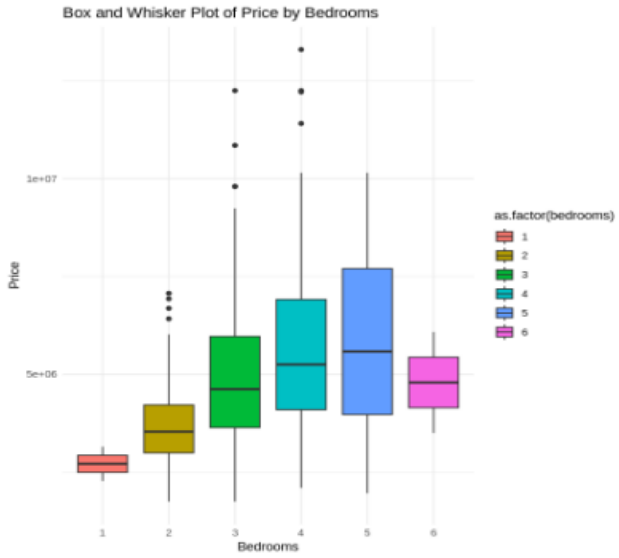
Interpretation:

- **"Semi-furnished"** is likely the most frequently occurring term, indicated by its larger font size, suggesting a higher prevalence of semi-furnished homes in the dataset.
- **"Furnished"** and **"Unfurnished"** might appear in smaller font sizes, indicating their relatively lower frequency compared to "furnished."

- The varying sizes of the words allow viewers to quickly gauge the importance or distribution of different furnishing statuses in the housing market.

Box Plot:

```
# Box and Whisker Plot: Price vs Bedrooms
ggplot(housing_data, aes(x = as.factor(bedrooms), y = price)) +
 geom_boxplot(aes(fill = as.factor(bedrooms))) +
 labs(title = "Box and Whisker Plot of Price by Bedrooms", x = "Bedrooms", y = "Price") +
 theme_minimal()
```



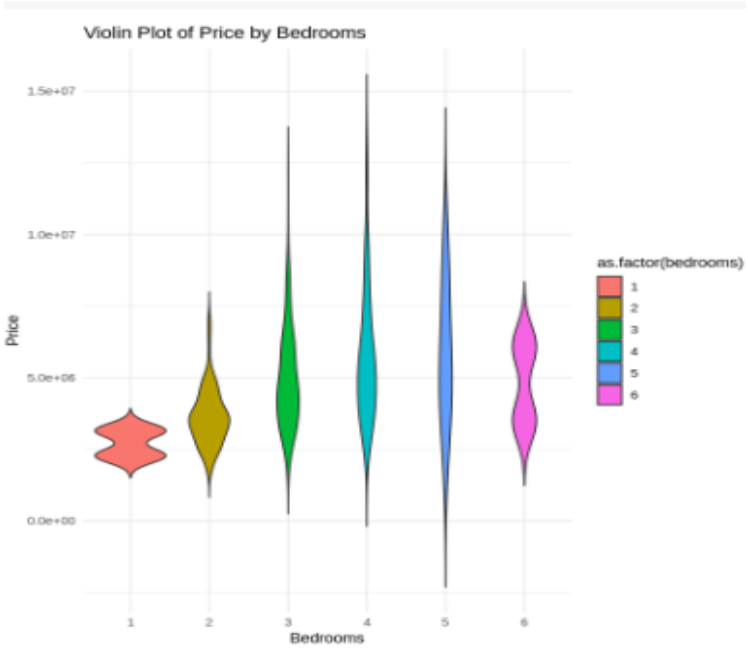
- **Observation:**

Description: This plot displays the distribution of housing prices based on the number of bedrooms. The box plot summarizes the data by showing the median, quartiles, and potential outliers.

Interpretation: Each box represents the interquartile range (IQR), which shows the middle 50% of the data. The line inside each box indicates the median price for that number of bedrooms. The "whiskers" extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartiles, respectively. Outliers are represented as individual points. This graph helps to understand how price varies with the number of bedrooms.

Violin Plot:

```
# Violin Plot: Price Distribution
ggplot(housing_data, aes(x = as.factor(bedrooms), y = price)) +
 geom_violin(aes(fill = as.factor(bedrooms)), trim = FALSE) +
 labs(title = "Violin Plot of Price by Bedrooms", x = "Bedrooms", y = "Price") +
 theme_minimal()
```



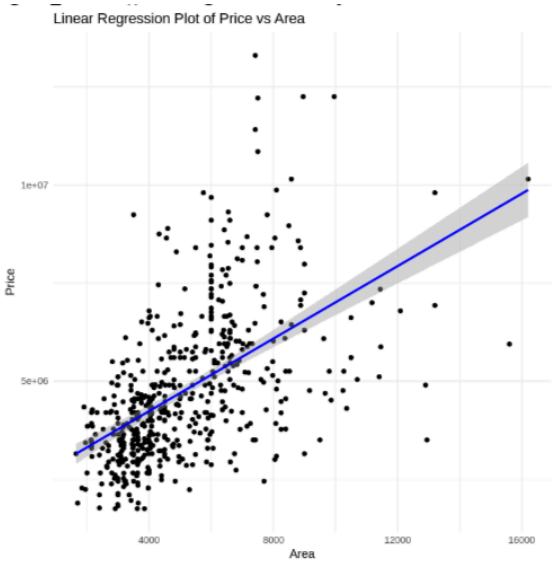
- **Observation:**

Description: A violin plot combines a box plot and a kernel density plot. It shows the distribution of housing prices for different categories of bedrooms.

Interpretation: The width of each "violin" represents the density of prices at different values, giving insight into the distribution shape. Similar to the box plot, the median and quartiles are shown within each violin. This plot allows for comparison of price distributions across different bedroom categories, highlighting areas where prices cluster or are sparse.

Linear Regression Plot:

```
# Linear Regression Plot
ggplot(housing_data, aes(x = area, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Linear Regression Plot of Price vs Area", x = "Area", y = "Price") +
  theme_minimal()
```



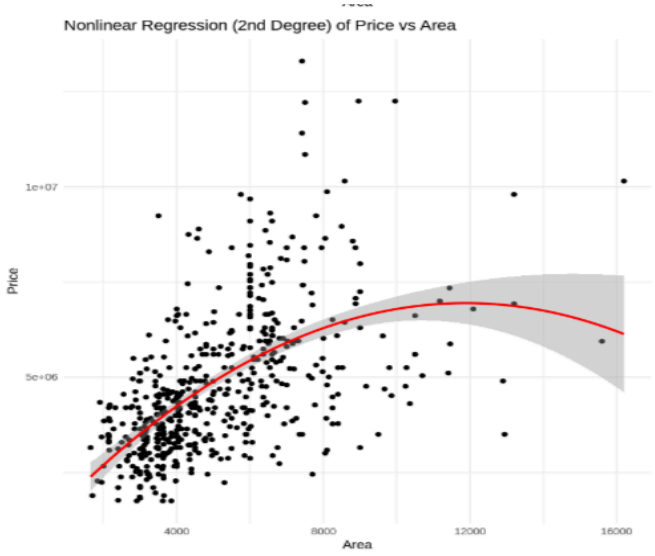
- **Observation:**

Description: This scatter plot visualizes the relationship between housing prices and the area of the houses. A linear regression line is overlaid to indicate the general trend.

Interpretation: Each point represents a house, with the x-axis showing the area and the y-axis showing the price. The linear regression line suggests that as the area increases, the price tends to increase, indicating a positive correlation. The shaded area around the line indicates the confidence interval, showing the uncertainty of the predictions.

Non-Linear Regression Plot:

```
# Polynomial (Nonlinear) Regression Plot
ggplot(housing_data, aes(x = area, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), col = "red") +
  labs(title = "Nonlinear Regression (2nd Degree) of Price vs Area", x = "Area", y =
         "Price") +
  theme_minimal()
```



Observation:

Description: This scatter plot also shows the relationship between housing prices and area but fits a second-degree polynomial regression line instead of a linear regression.

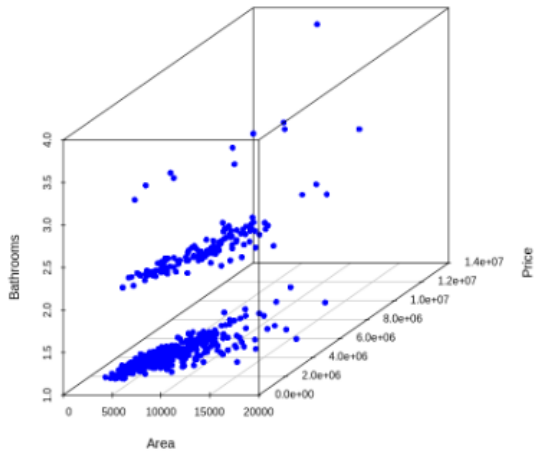
Interpretation: This indicates a more complex relationship between area and price, where the trend may curve. The regression line shows how price changes with area in a nonlinear manner. The shaded area indicates the confidence interval, similar to the linear regression plot. This graph suggests that the relationship between area and price may not be constant and may change at different levels of area.

3D Scatter Plot:

```
# Install scatterplot3d if not already installed
install.packages("scatterplot3d")

# Load scatterplot3d library
library(scatterplot3d)

# Create a 3D scatter plot
scatterplot3d(housing_data$area, housing_data$price,
housing_data$bathrooms,
              pch = 16, color = "blue", xlab = "Area", ylab = "Price", zlab =
"Bathrooms")
```



○ **Observation:**

Description: This 3D scatter plot visualizes the relationship between housing price, area, and the number of bathrooms. Each point represents a house with its corresponding values for area, price, and bathrooms.

Interpretation: The x-axis represents the area, the y-axis represents the price, and the z-axis represents the number of bathrooms. The points are plotted in 3D space, allowing viewers to see how the price varies with both area and the number of bathrooms. Generally, it appears that houses with larger areas and more bathrooms tend to have higher prices, indicating a positive correlation among these variables. This visualization helps to understand multidimensional relationships in the housing market.

Jitter Plot:

```
# Jitter Plot: Parking vs Price
ggplot(housing_data, aes(x = as.factor(parking), y = price)) +
 geom_jitter(aes(color = as.factor(parking)), width = 0.2) +
 labs(title = "Jitter Plot of Price by Parking Availability", x = "Parking Spaces",
y = "Price") +
 theme_minimal()
```

○ **Observation:**

Description: The jitter plot displays the distribution of housing prices categorized by the number of parking spaces available. The jitter effect adds random noise to the data points to avoid overlap and enhance visibility.

Interpretation: The x-axis represents the number of parking spaces (0 to 3), while the y-axis shows the price of the houses. Each color represents a different category of parking availability (0, 1, 2, or 3 parking spaces). The spread of points within each category indicates the variation in prices for houses with that specific number of parking spaces. The jittered effect helps to visualize the distribution and identify potential clusters or trends, suggesting that houses with more parking spaces might have different pricing trends compared to those with fewer or no parking spaces.

Outcomes:

- **Created Multiple Chart Types**: Successfully generated various visualizations such as bar charts, box and whisker plots, violin plots, regression plots (both linear and nonlinear), 3D scatter plots, and jitter plots using R. Each chart helped to display different aspects of the housing data, allowing for a deeper understanding of the dataset.
- **Analyzed Price Distribution**: The box and whisker plot revealed insights into the distribution of housing prices based on the number of bedrooms, highlighting price ranges and potential outliers. The violin plot provided a visual representation of price distributions, showing the density of prices across different bedroom counts.
- **Explored Relationships Between Variables**: The linear and nonlinear regression plots effectively illustrated the relationships between area and price, indicating trends such as the positive correlation between larger areas and higher prices. The 3D scatter plot depicted the relationship between area, price, and the number of bathrooms, enriching the analysis by adding a third dimension.

- **Comparison of Furnishing Types**: The word cloud visualization highlighted the prevalence of different furnishing types—furnished, semi-furnished, and unfurnished—providing insights into the market trends related to the types of housing available.
- **Developed Visualization Skills**: This experiment enhanced the understanding of how various chart types can be applied to analyze specific elements of a dataset. Each chart type provided a unique perspective, demonstrating the power of visual representations in summarizing large datasets and making the information more accessible.

Conclusion:

The exploration of various chart types provided a comprehensive understanding of the housing market dynamics. By employing visualizations such as box and whisker plots, violin plots, and regression analyses, we uncovered significant patterns in housing prices, particularly in relation to the number of bedrooms and the size of the properties. The integration of 3D scatter plots allowed for a nuanced view of how multiple variables interact, revealing intricate relationships between area, price, and the number of bathrooms.Additionally, the word cloud visualization emphasized the market's focus on furnishing types, illustrating the demand for furnished, semi-furnished, and unfurnished properties. This multifaceted approach to data visualization not only enhanced our analytical capabilities but also underscored the importance of visual tools in transforming complex datasets into clear, actionable insights.Ultimately, the findings from this analysis can inform stakeholders—whether they are real estate agents, buyers, or policymakers—about market trends, helping them make informed decisions based on data-driven evidence.