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

Name: Sakshi D. Lonare UID: 2021300069

Experiment no 5

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

- 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:

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

Dataset:

https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot?resource=download

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 housing real estate datasets and gain insights.

Chart Types:

- **Regression Plot**: Visualizes the relationship between two variables, often with a fitted regression line.

- **Box and Whisker Plot**: Represents the distribution of data, showing median, quartiles, and potential outliers.
- **Word Cloud Plot**: Displays the frequency of words in a text as the size of the words.
- **Jitter Plot**: Shows the distribution of data points, adding jitter to reduce overlap.
- **3D Scatter Plot**: Visualizes the relationship between three variables in a 3D space.
- **Violin Plot:** Combines elements of a box plot and a kernel density plot, providing a more detailed view of the distribution.

Steps to Perform in R:

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

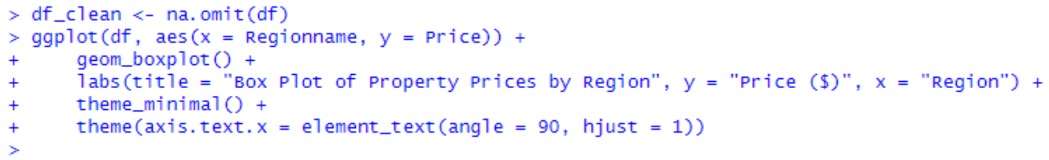
```
install.packages("ggplot2")
install.packages("dplyr")
library(ggplot2)
library(dplyr)
```

2. **Load the Dataset:** 一 Load the housing dataset
3. **Data Preprocessing:** 二 Inspect and clean the data if necessary (handle missing values, filter relevant columns, etc.).

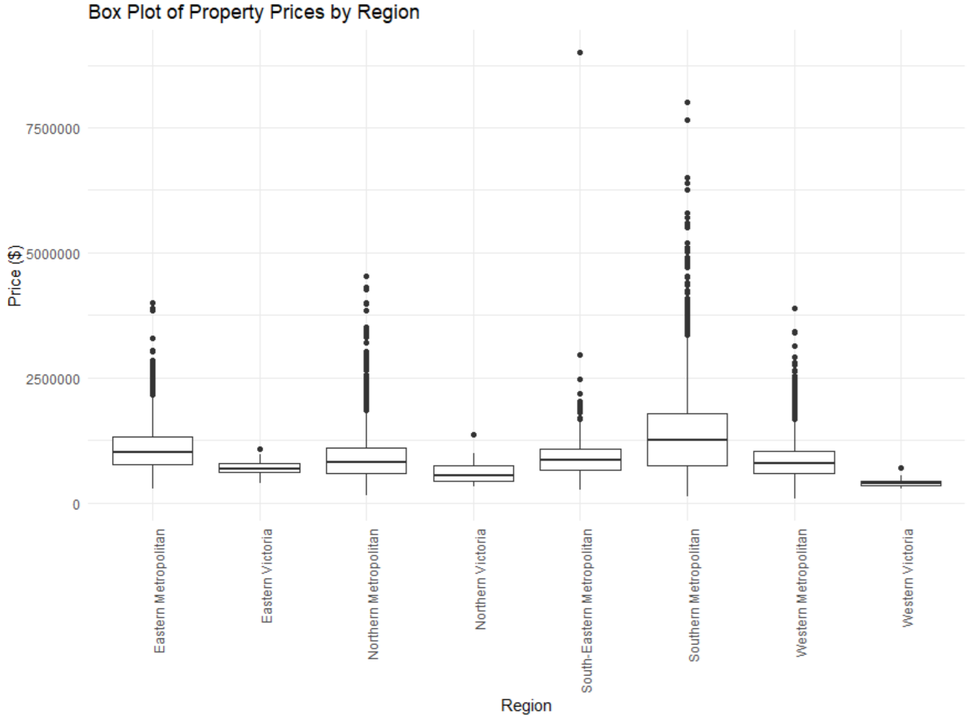
```
df <- df %>% na.omit()
```

4. **Create Visualizations:**

Box and Whisker plot:



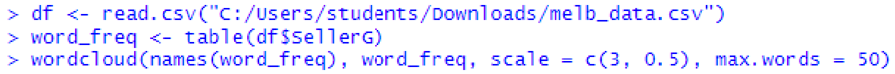
```
> df_clean <- na.omit(df)
> ggplot(df, aes(x = Regionname, y = Price)) +
+     geom_boxplot() +
+     labs(title = "Box Plot of Property Prices by Region", y = "Price ($)", x = "Region") +
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 90, hjust = 1))
>
```



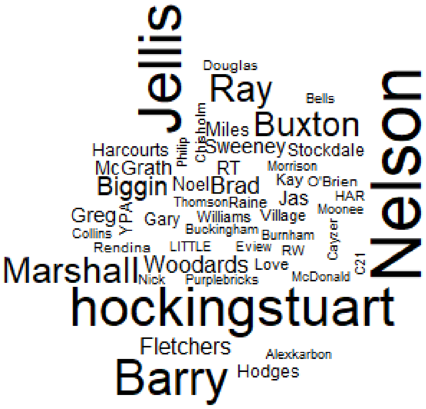
○ **Observations:**

- Eastern Metropolitan and Southern Metropolitan regions generally have higher median prices compared to other regions. Northern Victoria and Western Victoria regions tend to have lower median prices.
- Eastern Metropolitan and Southern Metropolitan regions have wider IQRs, indicating a larger spread of property prices within those regions.

Word Cloud Chart:



```
> df <- read.csv("C:/Users/students/Downloads/melb_data.csv")
> word_freq <- table(df$SellerG)
> wordcloud(names(word_freq), word_freq, scale = c(3, 0.5), max.words = 50)
```

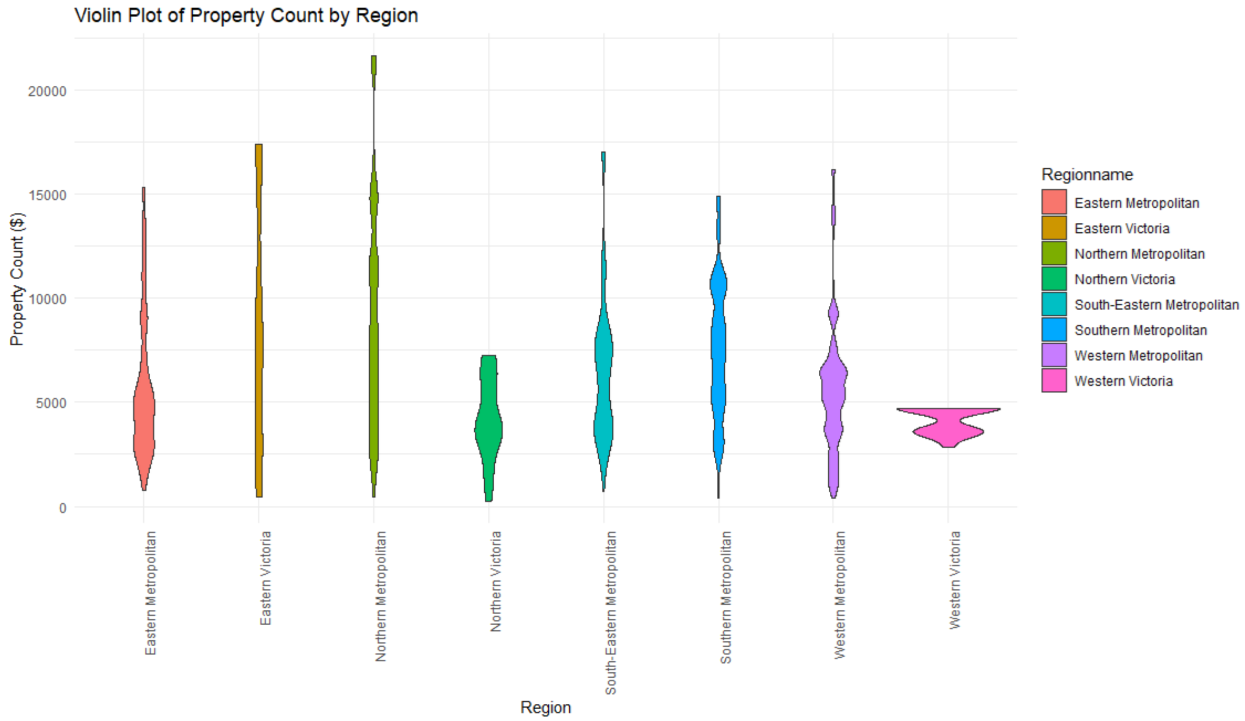


- **Observation:**

- The word cloud provides a visual representation of the frequency of occurrence of different real estate agent names in the dataset. The larger the size of a word, the more frequently it appears.
- Marshall, hockingstuart, and Barry are the most prominent real estate agencies, suggesting that they have a higher market share or are more frequently mentioned in the dataset.
- McGrath and Biggin are close together, possibly indicating a group or network of agencies.

Violin Plot:

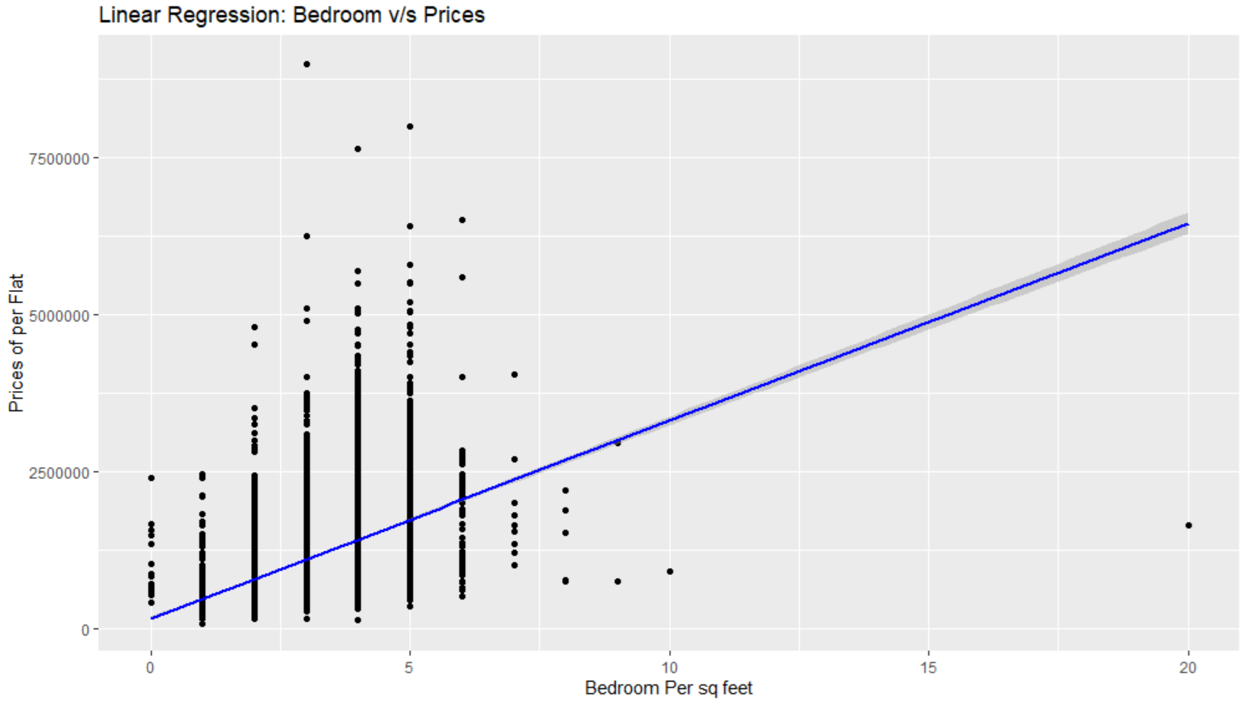
```
> ggplot(df, aes(x = Regionname, y = Propertycount, fill = Regionname)) +
+    geom_violin() +
+    labs(title = "Violin Plot of Property Count by Region", y = "Property
Count ($)", x = "Region") +
+    theme_minimal() +
+    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



- **Observation:**
  - Eastern Metropolitan and Southern Metropolitan regions appear to have the highest median property counts while Northern Victoria and Western Victoria regions seem to have the lowest median property counts.
  - The shape of the violins indicates the distribution of property counts within each region. Eastern Metropolitan and Southern Metropolitan regions have wider violin plots, suggesting a larger spread of property counts. Northern Victoria and Western Victoria regions have narrower violin plots, indicating a more concentrated distribution of property counts.
  - The violin plots for some regions overlap, indicating that there might be some overlap in the property count

Regression Plot:

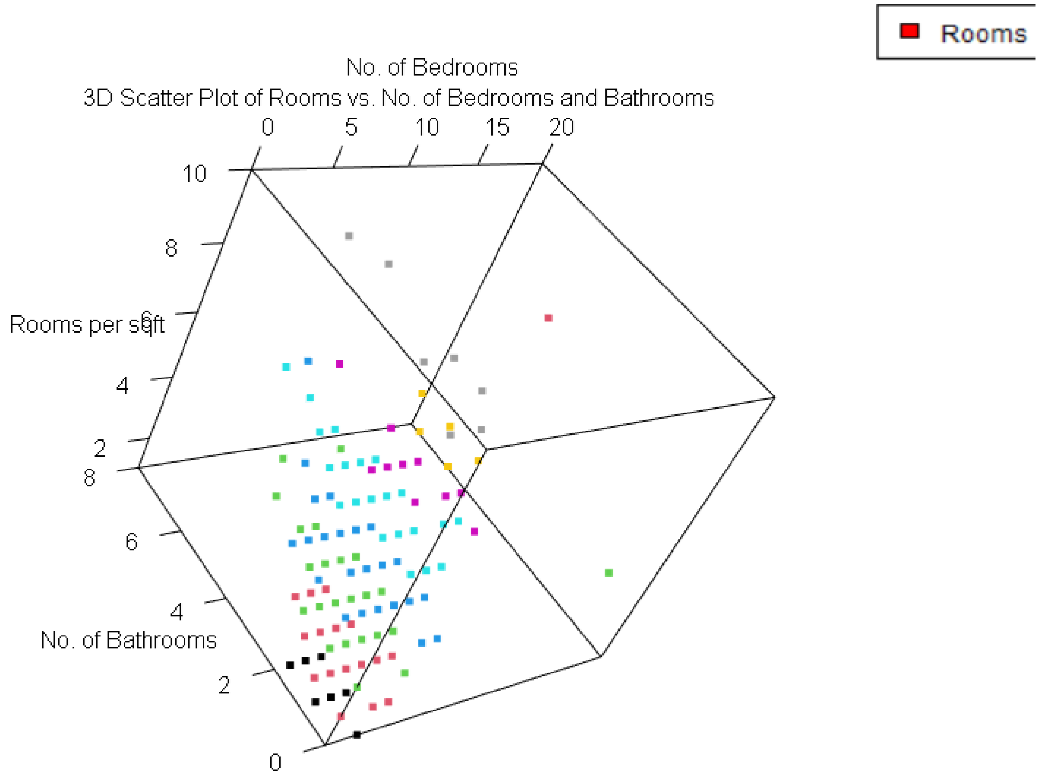
```
ggplot(df, aes(x = Bedroom2, y = Price)) +
+    geom_point() +
+    geom_smooth(method = "lm", col = "blue") +
+    ggtitle("Linear Regression: Bedroom v/s Prices") +
+    xlab("Bedroom Per sq feet") +
+    ylab("Prices of per Flat")
```



- **Observations:**
- There appears to be a positive linear relationship between the number of bedrooms per square foot and the prices of flats. As the number of bedrooms per square foot increases, the prices tend to increase as well
- The regression line slopes upward, indicating the positive correlation between the two variables.
- There are a few outliers, especially on the higher end of bedroom per square foot, which might be influencing the regression line.

3D Chart:

```
plot3d(df$Bedroom2, df$Bathroom, df$Rooms,
+        col = df$Rooms, size = 5,
+        xlab = "No. of Bedrooms", ylab = "No. of Bathrooms", zlab = "Rooms
per sqft",
+        main = "3D Scatter Plot of Rooms vs. No. of Bedrooms and Bathrooms")
>
> # Add color legend
> legend3d("topright", legend = "Rooms", fill = rainbow(10))
```



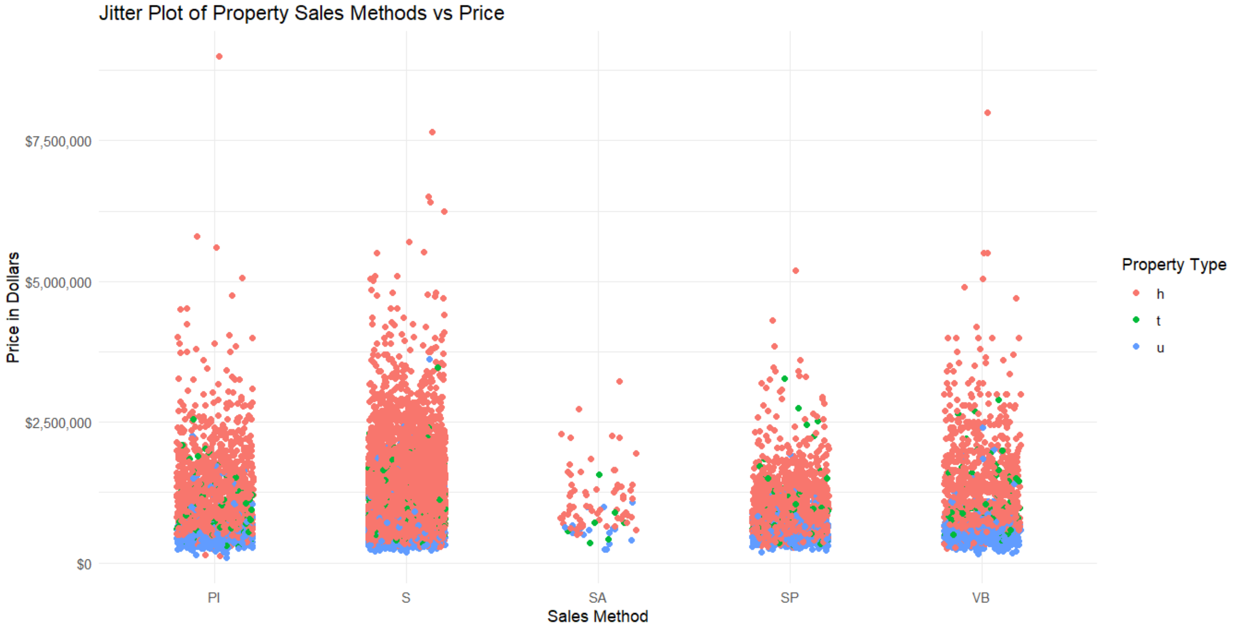
- **Observation:**

The 3D scatter plot provides valuable insights into the relationships between rooms, bedrooms, and bathrooms. It helps to visualize the interplay of these variables and identify potential patterns or anomalies in the data.

- There seems to be a strong positive correlation between the number of rooms and the number of bedrooms
- The data points seem to cluster in certain areas of the plot, suggesting that there might be combinations of rooms, bedrooms, and bathrooms that are more common or preferred

Jitter Plot:

```
ggplot(df, aes(x = Method, y = Price, color = Type)) +
+     geom_jitter(width = 0.2, height = 0) +
+     scale_y_continuous(labels = scales::dollar) +
+     labs(title = "Jitter Plot of Property Sales Methods vs Price",
+          x = "Sales Method",
+          y = "Price in Dollars",
+          color = "Property Type") +
+     theme_minimal()
```



Observations:

The jitter plot indicates that property sales methods and property types might have some influence on the price distribution in the Melbourne real estate market.

- PI (Private Treaty) and VB (Vendor Bid) methods seem to have a wider range of prices, with some properties sold at significantly higher or lower prices compared to other methods. SA (Sale by Auction) and SP (Set Price) methods appear to have a more concentrated distribution of prices, with fewer outliers. S (Sale) method seems to have a slightly higher median price compared to the other methods.

Outcomes:

- Created various advanced charts in R to effectively visualize and analyze housing data of Melbourne.
- Derived key insights into the distribution, frequency, and correlations within the dataset.
- Developed proficiency in using different chart types to explore and present data comprehensively.

Conclusion:

From this experiment Ive showcased the effectiveness of data visualization in revealing patterns and trends within a real estate dataset of Melbourne. Using R, we efficiently generated visual representations that enabled us to examine the data from multiple angles, leading to more informed insights and conclusions.