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

Aim :
Create advanced charts using R programming language on housing dataset

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

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.

Word Chart
A word chart is a simple visual tool used to represent the frequency of words or terms within a given text. It is often displayed as a bar chart or a word cloud, where the length of each bar or the size of each word corresponds to its frequency.

Box and Whisker Plot
A box and whisker plot is a graphical representation of a dataset that shows the distribution of data, including the median, quartiles, and outliers. It consists of a box, whiskers, and dots.

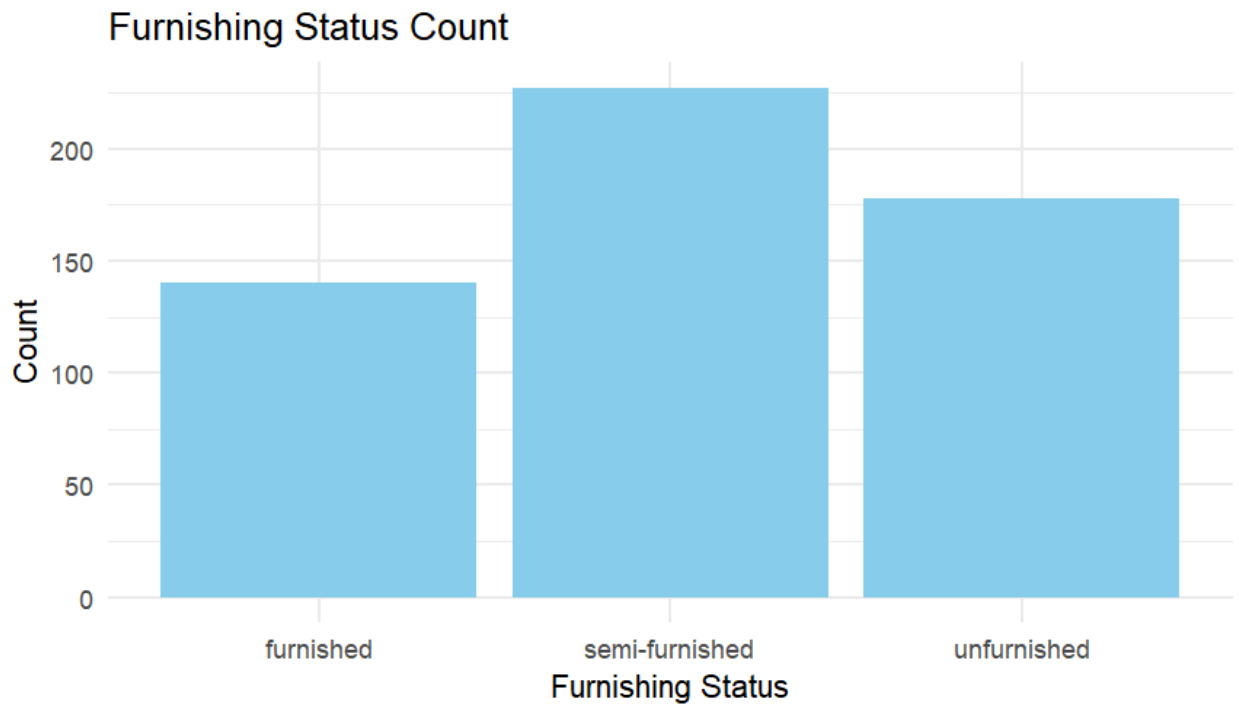
Violin Plot
A violin plot is a combination of a box plot and a kernel density estimate. It provides a more detailed view of the distribution of data, especially for non-normal distributions.

Regression Plot (Linear and Nonlinear)
A regression plot is used to visualize the relationship between two variables. It typically includes a scatter plot of the data points and a regression line.

3D Chart
A 3D chart is a visual representation of data with three dimensions (x, y, and z). It can be used to visualize data that involves three variables or to add depth to a 2D chart.

Jitter
Jitter is a technique used to add randomness to data points in a scatter plot to avoid overlapping points. This can make it easier to see the distribution of data and identify patterns.

1. Furnishing Status of Properties

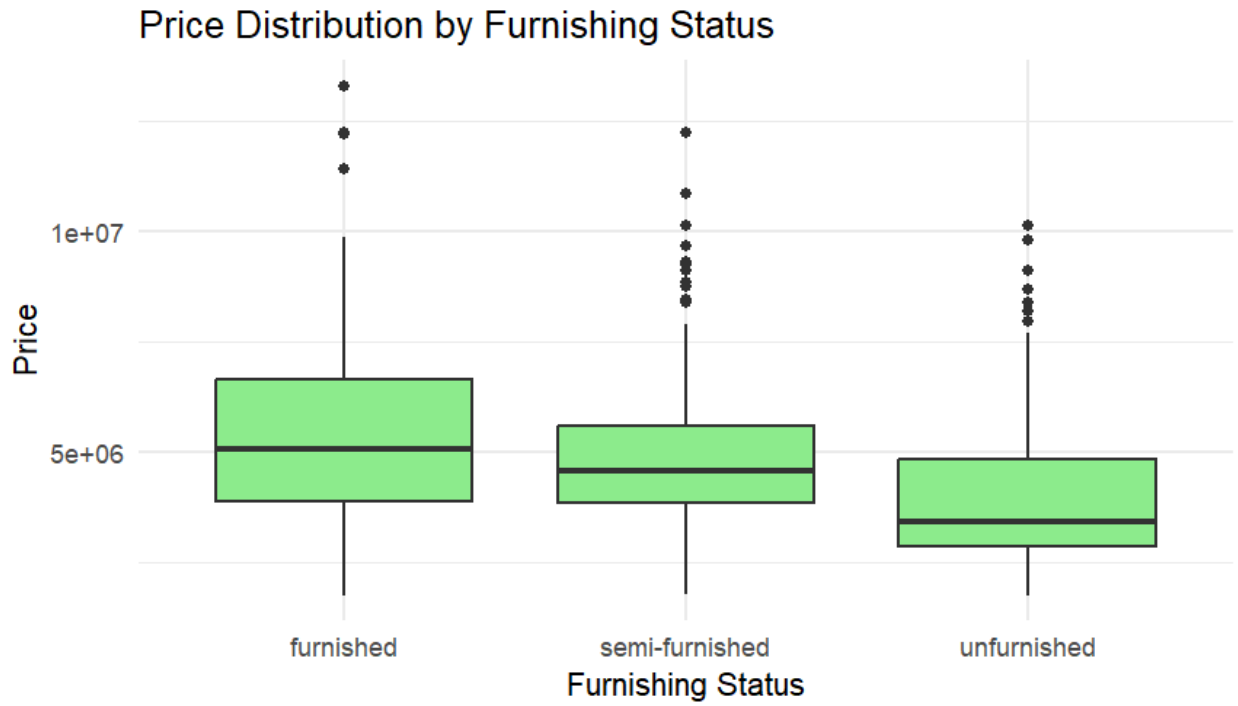


The provided bar graph illustrates the distribution of furnishing statuses among a given dataset. The x-axis represents the different furnishing statuses (furnished, semi-furnished, and unfurnished), while the y-axis indicates the count of properties corresponding to each status.

Analysis of the graph reveals that semi-furnished properties are the most prevalent, followed by unfurnished and furnished properties. This suggests that a significant portion of the properties in the dataset are partially equipped with furniture, while a smaller number are either fully furnished or completely devoid of furniture.

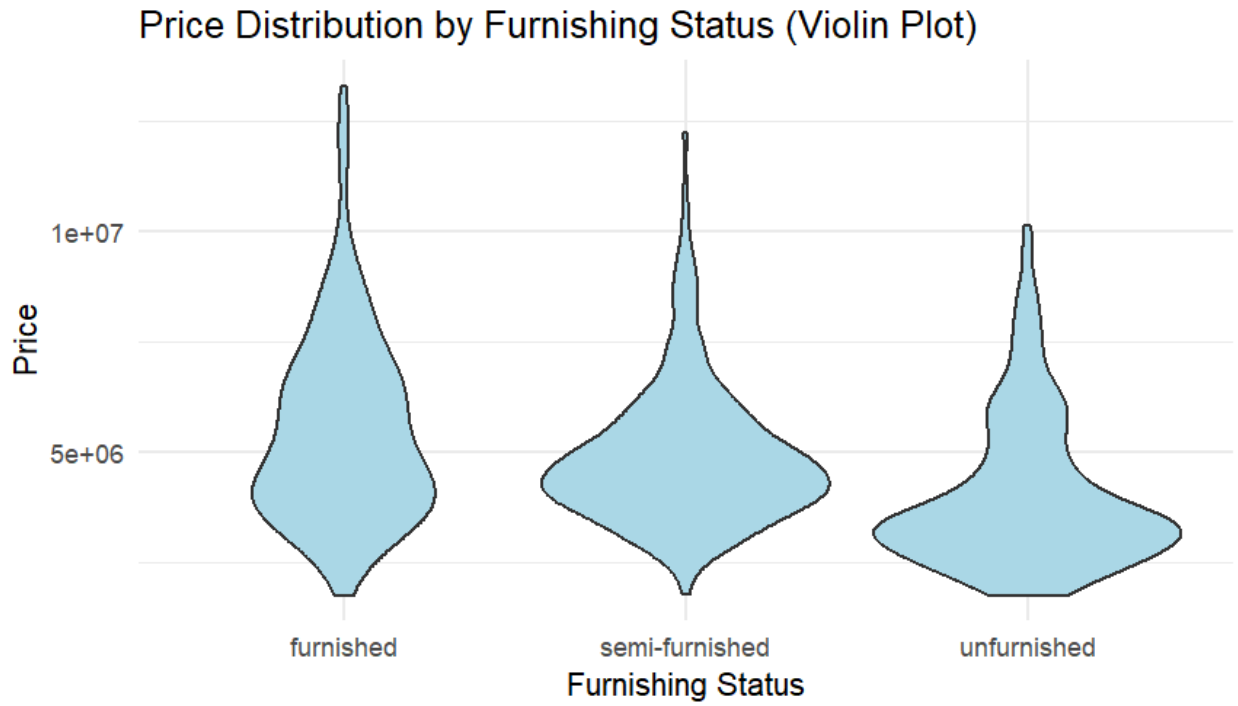
These findings may be valuable for various stakeholders, such as real estate agents, property managers, or interior designers, as they provide insights into the preferences and demands of the market regarding furnishing levels.

2. Box Plot by Furnishing status to detect overpriced or underpriced properties



The box plot illustrates the price distribution of properties based on their furnishing status. It shows that furnished properties generally have a higher median price compared to semi-furnished and unfurnished properties. The interquartile range (IQR), represented by the box, is similar across all three categories, indicating a consistent spread of prices within each group. However, there are outliers (individual data points outside the whiskers) in the furnished and semi-furnished categories, suggesting some properties in these categories have significantly higher or lower prices compared to the majority. Overall, the graph suggests that furnishing status is a factor that influences property prices, with furnished properties tending to be more expensive.

3. Violin Plot

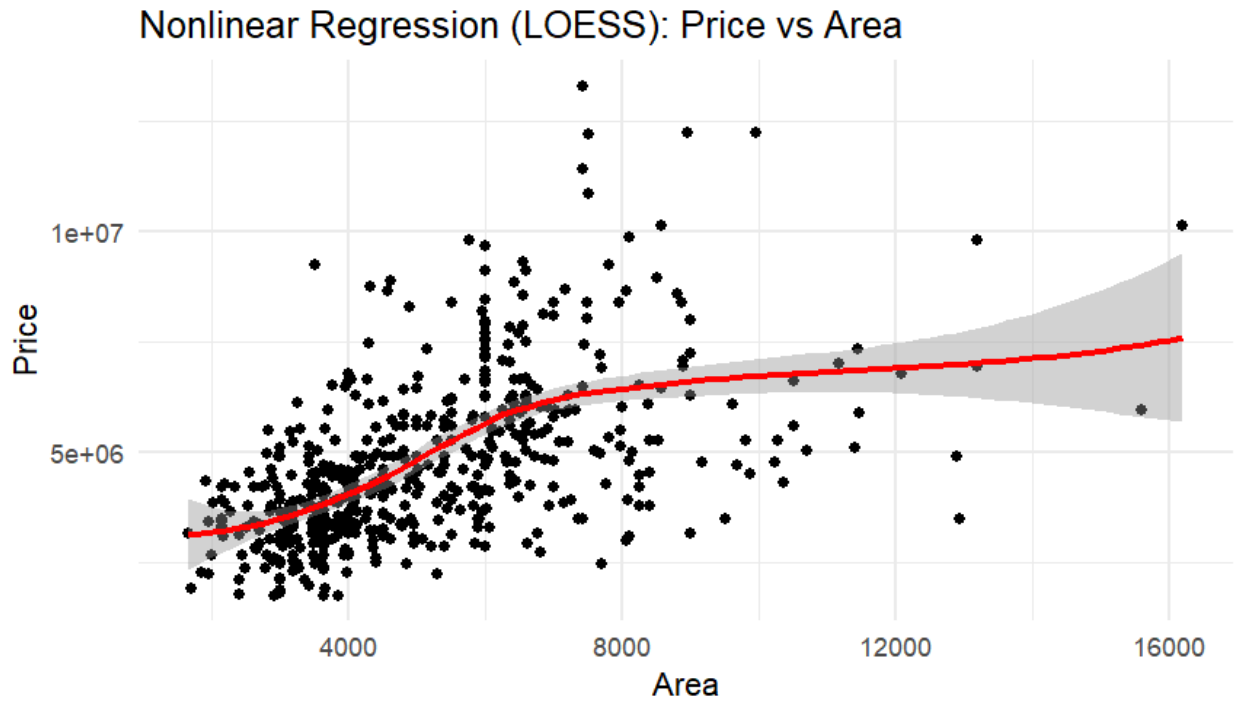


The violin plot visually represents the price distribution of properties based on their furnishing status. It offers a more detailed view of the density of prices within each category compared to a box plot. The violin plots show that furnished properties generally have a higher median price and a wider range of prices compared to semi-furnished and unfurnished properties. The semi-furnished category exhibits a more concentrated distribution of prices, suggesting less variation within this group. The unfurnished category shows a similar pattern to the furnished category, with a higher median price and a wider range of prices. Overall, the violin plot confirms the findings from the box plot, indicating that furnishing status is a significant factor influencing property prices, with furnished properties tending to be more expensive and have a wider range of price variations.

4. Price vs Area

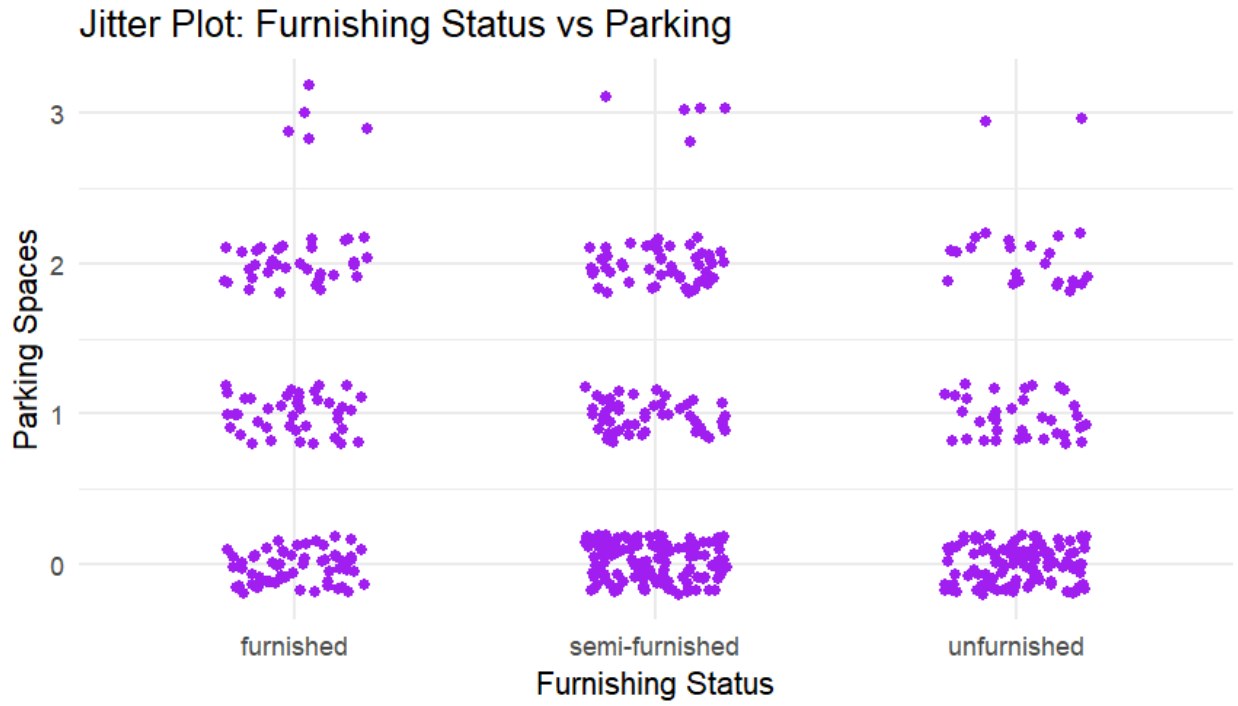
The scatter plot with a linear regression line illustrates the relationship between property price and area. The blue line represents the best-fit linear model, showing a positive correlation between price and area. This suggests that as the area of a property increases, its price tends to increase as well. The shaded area around the line represents the confidence interval, indicating the range of potential price values for a given area. The scatter of the data points around the line suggests some variability in the relationship, indicating that other factors besides area may also influence property prices. Overall, the graph provides evidence that property area is a significant predictor of price, but other variables may also play a role.

5. Non Linear Regression



The scatter plot with a nonlinear regression curve (LOESS) illustrates the relationship between property price and area. The red curve represents the best-fit nonlinear model, suggesting a more complex relationship between the two variables compared to the linear regression model. The curve indicates that while there is a general trend of increasing prices with increasing area, the relationship is not strictly linear. The shaded area around the curve represents the confidence interval, indicating the range of potential price values for a given area. The scatter of the data points around the curve suggests some variability in the relationship, indicating that other factors besides area may also influence property prices. Overall, the graph provides evidence that property area is a significant predictor of price, but a nonlinear relationship may better capture the complexity of the relationship between these two variables.

6. Parking and Furniture



The jitter plot illustrates the relationship between furnishing status and the number of parking spaces associated with properties. It shows that there is a general trend for furnished properties to have more parking spaces compared to semi-furnished and unfurnished properties. However, there is also considerable overlap between the categories, indicating that furnishing status is not a deterministic factor for the number of parking spaces. The jitter plot reveals that while furnished properties may have a higher likelihood of having more parking spaces, there are still many furnished properties with fewer parking spaces and vice versa. This suggests that other factors, such as property size, location, or individual preferences, may also play a role in determining the number of parking spaces available.