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: Siddhesh Kirdat
UID: 2021300064
Batch: BE CSE(DS)

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 understand and apply basic data visualization techniques in R.
- To create various types of charts (Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot) using a crime-related dataset.
- To interpret and analyze the data through visual representations.

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 basic chart types to analyze crime data and derive insights.

Steps to Perform in R:

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

R
Copy code
```
install.packages("ggplot2")
install.packages("dplyr")
library(ggplot2)
library(dplyr)
```

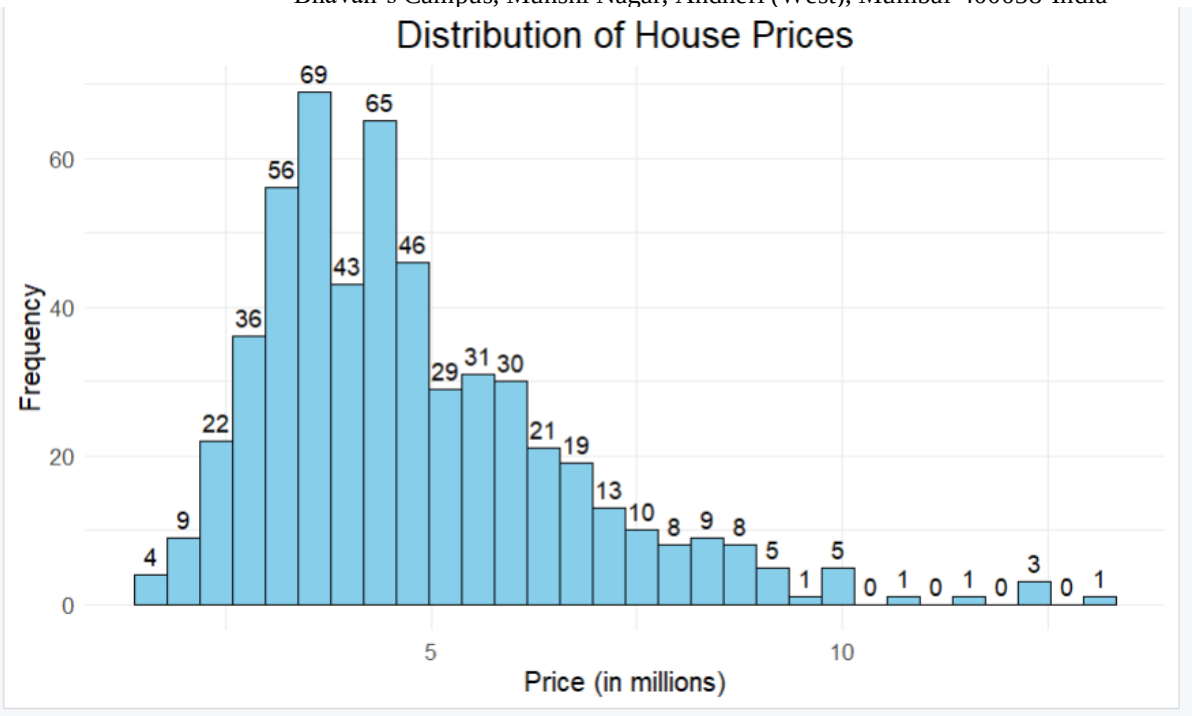
2. **Load the Dataset:**

```
> housing_data<-read.csv("D:/ADV_EXPS/ADV-LAB-EXPS/Exp5/Housing.csv",header=TRUE)
> 
```

3. **Create Visualizations:**

Histogram:
R

```
> ggplot(data, aes(x = Area_Name)) +
+     geom_bar(aes(y = Cases_Property_Stolen, fill = "Stolen"), stat = "identity", position = "dodg
e") +
+     geom_bar(aes(y = Cases_Property_Recovered, fill = "Recovered"), stat = "identity", position =
"dodge") +
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
+     labs(title = "Burglary Cases by Region (2001)",
+          x = "Region",
+          y = "Number of Cases",
+          fill = "Case Type") +
+     scale_fill_manual(values = c("Stolen" = "red", "Recovered" = "green")) +
+     coord_flip()
```



Key observations from the house price distribution:

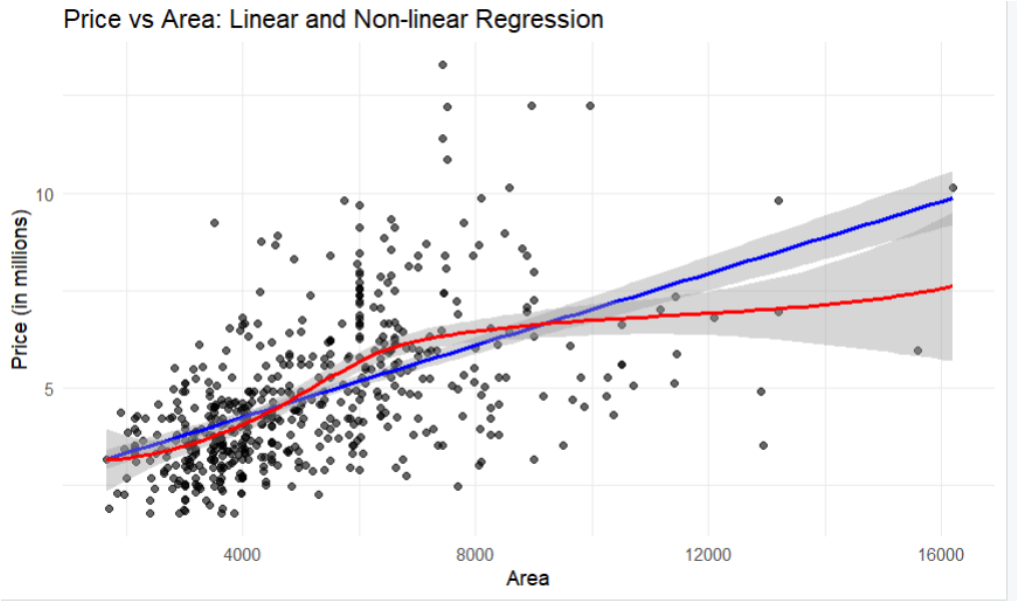
1. Right-skewed distribution (more lower-priced homes)
2. Peak frequency around 4-5 million (65-69 houses)
3. Most houses priced between 3-7 million
4. Very few houses above 10 million (rare luxury properties)
5. Gradual decline in frequency after the peak
6. Minimum frequency at both extremes (below 3M and above 10M)

2.
Violin Plot

Key observations from the violin plot:

1. House prices vary significantly by number of bedrooms
2. 3 and 4 bedroom houses show the widest price distribution and highest density
3. Most houses have 2-4 bedrooms
4. 1 and 6 bedroom houses are rare with narrow price ranges
5. 3 bedroom houses show multiple price clusters (visible by the wider sections in the violin shape)
6. 5 bedroom houses tend to be more expensive with less price variation
7. There appear to be some outlier prices, particularly in the 3 and 4 bedroom categories (visible as individual dots above the main distribution)
8. Price range increases with number of bedrooms up to 4 bedrooms, then starts to narrow

4. **Regression Plot**

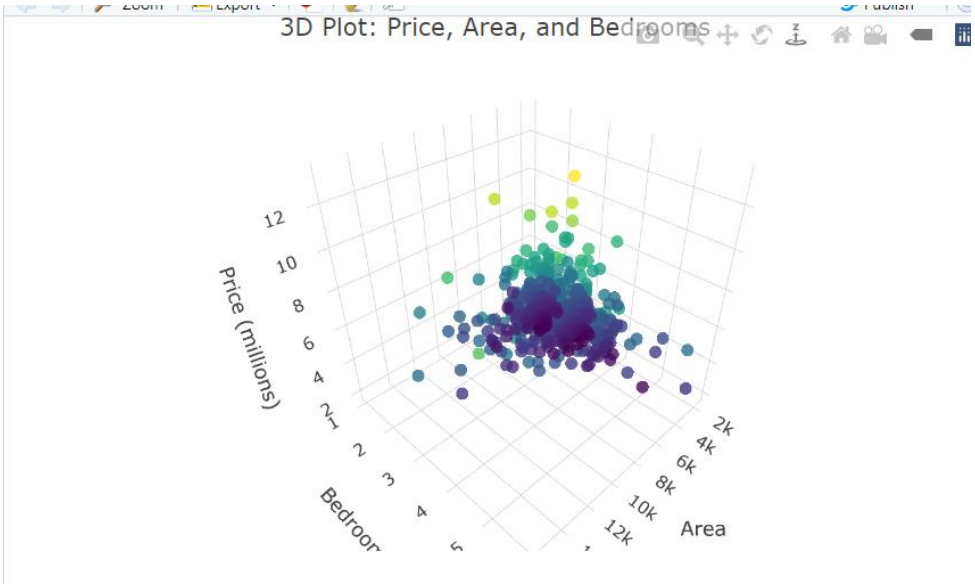


Key observations from the Price vs Area regression plot:

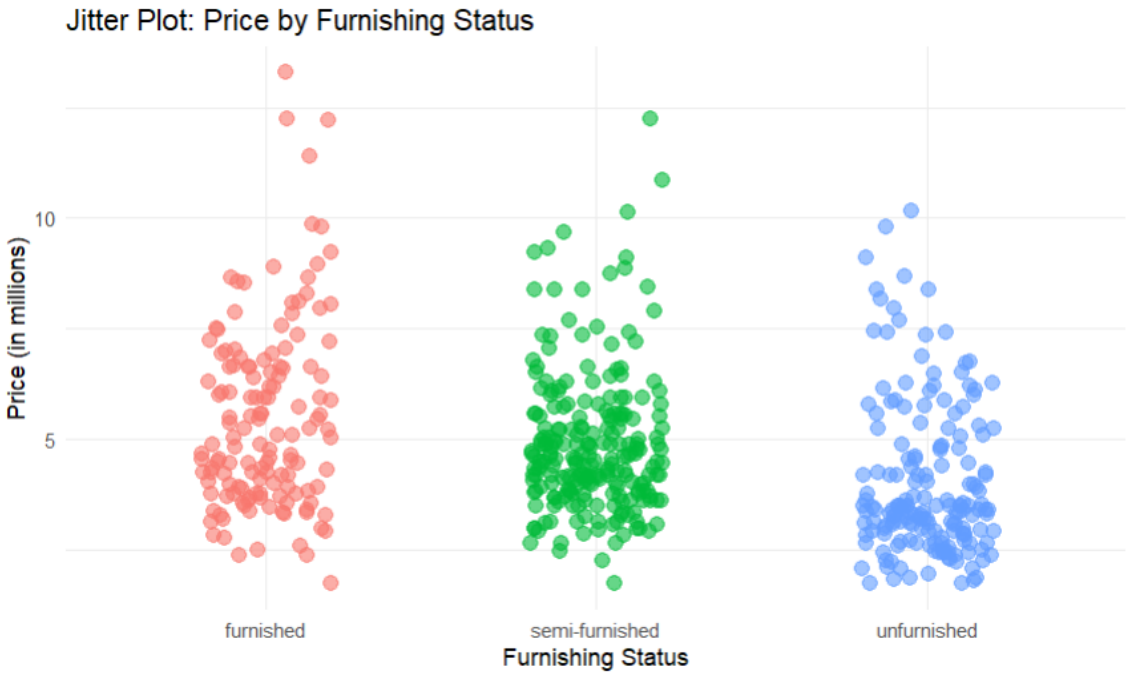
1. **Positive correlation: As area increases, price tends to increase**
2. **Non-linear relationship (red line) shows stronger price growth in smaller areas, then levels off for larger areas**
3. **Linear regression (blue line) suggests a consistent price increase with area**
4. **High scatter/dispersion of points indicates that area alone doesn't determine price**
5. **Wider confidence intervals (grey bands) at extreme areas due to fewer data points**
6. **Most properties are clustered in the 4000-8000 area range**
7. **Some outlier properties with high prices relative to their area**
8. **The relationship appears stronger for smaller properties than larger ones**
9. **The non-linear model suggests diminishing returns on price as area increases**

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

3D Chart



Jitter Plot



Key observations from the Jitter Plot of Price by Furnishing Status:

1. **Similar price ranges across all furnishing statuses (approximately 2-12 million)**
2. **Furnished homes show slightly higher concentration in the mid-price range (5-8 million)**
3. **Semi-furnished homes appear to have a more even distribution across price ranges**
4. **Unfurnished homes show higher density in the lower price range (2-5 million)**
5. **All categories have some outliers in the high price range (above 10 million)**
6. **The spread of points suggests furnishing status isn't a major price determinant**
7. **More homes appear to be furnished or semi-furnished than unfurnished**
8. **Each category shows a core concentration around 3-7 million price range**
9. **The jittering reveals clear clustering patterns within each furnishing status**

Outcomes:

- Successfully created multiple types of charts using R to visualize crime data.
- Gained insights into the distribution, frequency, and relationships within the crime dataset.
- Developed an understanding of how different chart types can be used to analyze and present data effectively.

Conclusion:

This experiment demonstrated the power of data visualization in uncovering patterns and trends in a crime dataset. By using R, we efficiently created visual representations that allowed us to explore the data from different perspectives, leading to better-informed conclusions.