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 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

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 basic chart types to analyze crime data and derive insights.

Chart Types:

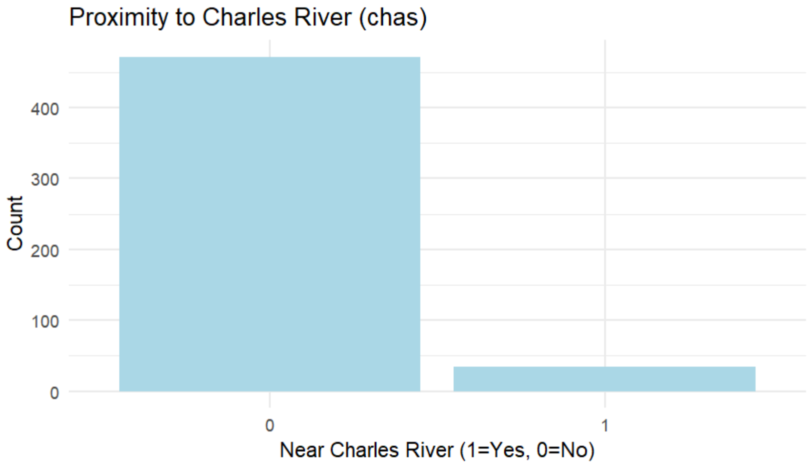
1. **Bar Chart:** A bar chart is used to display categorical data with rectangular bars representing the frequency or count of each category.
2. **Pie Chart:** A pie chart shows the proportion of categories as slices of a pie, useful for comparing parts of a whole.
3. **Histogram:** A histogram is used to represent the distribution of numerical data by grouping it into bins.
4. **Timeline Chart:** A timeline chart visualizes data points in chronological order, often used to show trends over time.
5. **Scatter Plot:** A scatter plot displays the relationship between two numerical variables using points in a Cartesian plane.
6. **Bubble Plot:** A bubble plot is an extension of a scatter plot where the size of the points (bubbles) represents an additional variable.

Dataset:
Link: https://www.kaggle.com/datasets/arunjangir245/boston-housing-dataset

1. crim: Per capita crime rate by town.
2. zn: Proportion of large residential lots (over 25,000 sq. ft.).
3. indus: Proportion of non-retail business acres per town.
4. Chas: Binary variable indicating if the property is near Charles River (1 for yes, 0 for no).
5. nox: Concentration of nitrogen oxides in the air.
6. rm: Average number of rooms per dwelling.
7. age: Proportion of old owner-occupied units built before 1940.
8. dis: Weighted distances to Boston employment centers.
9. rad: Index of accessibility to radial highways.
10. tax: Property tax rate per $10,000.

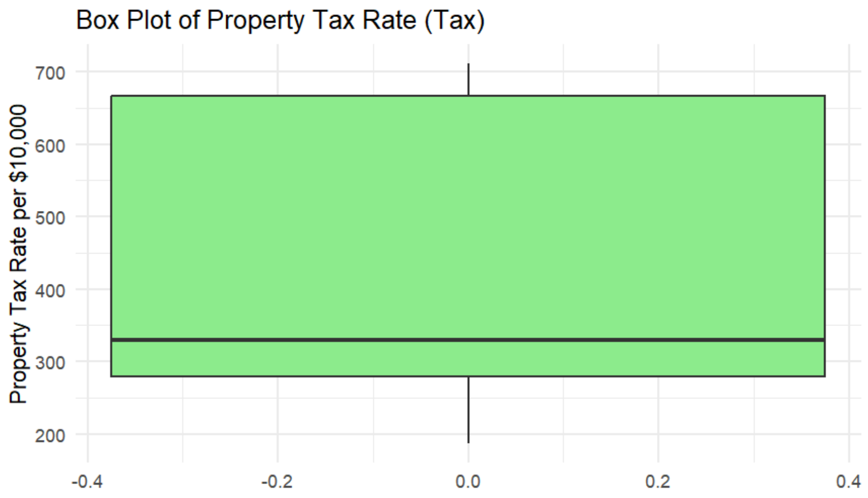
These features provide valuable information about the characteristics of neighborhoods that can influence housing prices.

Graphs:



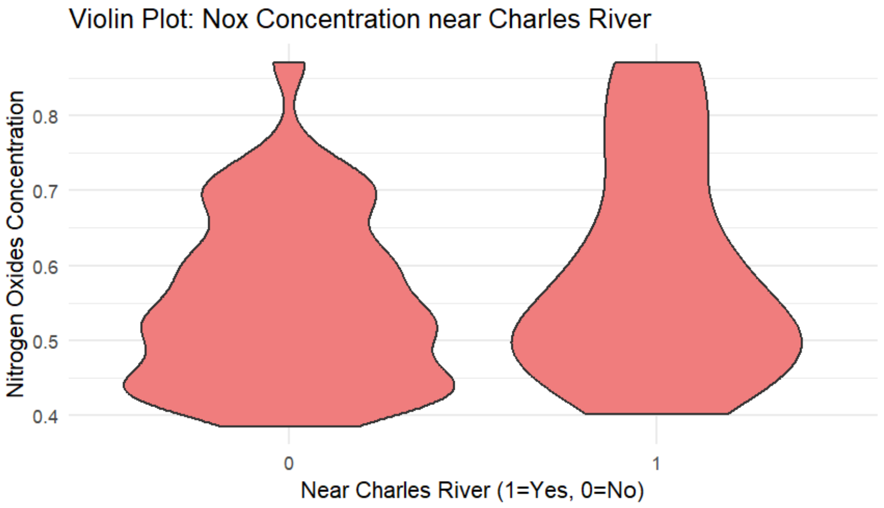
Key Observations:

1. **Skewed Distribution:** The distribution is highly skewed to the right, with a much larger number of data points in the "0" category (not near the Charles River) compared to the "1" category (near the Charles River).
2. **Dominance of "0" Category:** The vast majority of data points fall into the "0" category, indicating that most of the houses in the dataset are not located near the Charles River.
3. **Small Proportion Near the River:** The "1" category, representing houses near the Charles River, has a relatively small number of data points.



Key Observations:

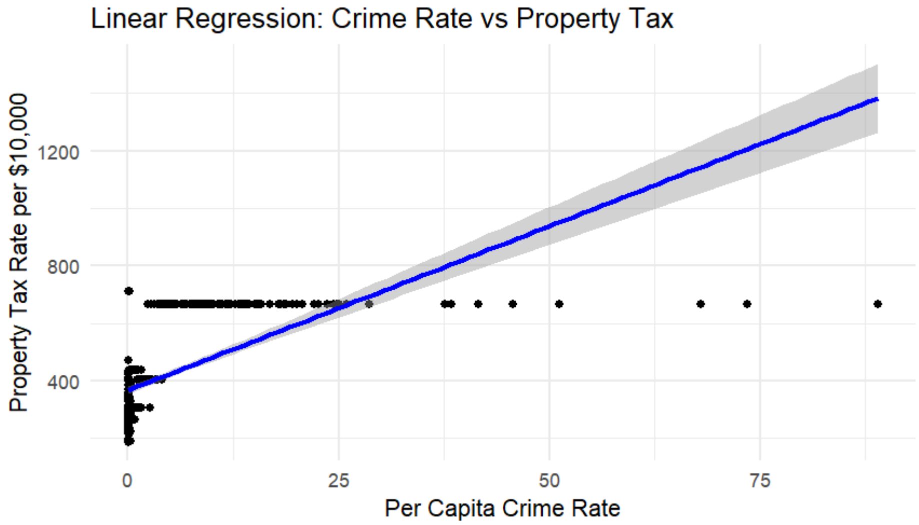
1. **High Median Tax Rate:** The median property tax rate (represented by the horizontal line within the box) is relatively high, indicating that a significant portion of the data points fall within the upper range of the distribution.
2. **Limited Variation:** The box, which represents the interquartile range (IQR), is relatively narrow, suggesting a limited range of property tax rates among the majority of the data points.
3. **Outliers:** There are no visible outliers (data points outside the whiskers), indicating that the distribution is relatively symmetrical and free from extreme values.
4. **Skewness:** The distribution appears to be slightly skewed to the left, with a longer tail on the lower end. This suggests that there are a few properties with significantly lower property tax rates compared to the majority.



Key Observations:

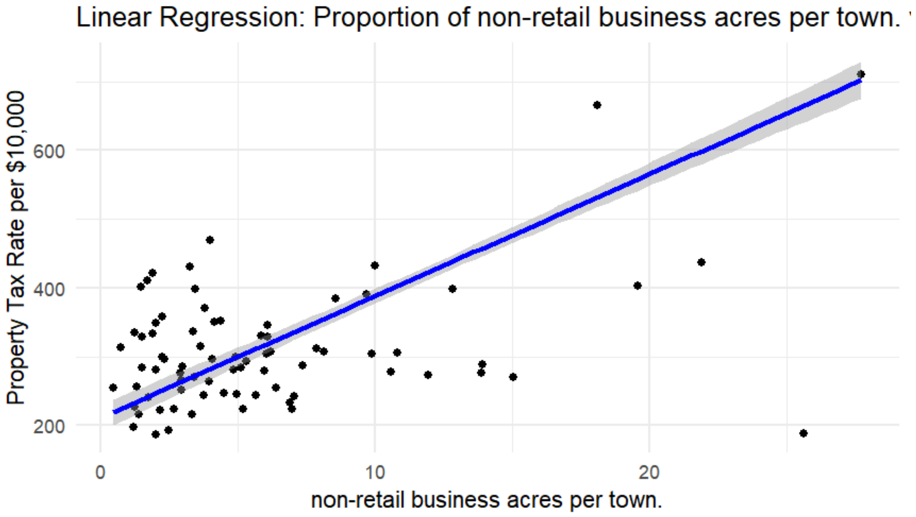
1. Distribution Comparison: The violin plot compares the distribution of nitrogen oxide (NOX) concentrations for houses near and far from the Charles River.

2. Overlapping Distributions: The two violin plots overlap to a significant extent, indicating that there is a considerable amount of overlap in the NOX concentrations between houses near and far from the river.
3. Median Comparison: The median NOX concentration (represented by the white dot within each violin) appears slightly lower for houses near the Charles River compared to those farther away.
4. Density Comparison: The density of the violin plot for houses near the Charles River is slightly more spread out, suggesting a wider range of NOX concentrations in that group.



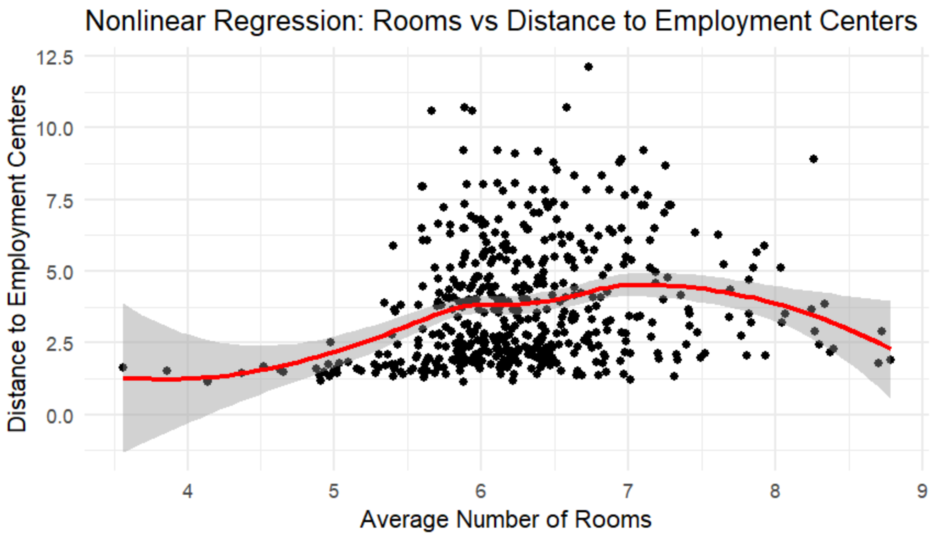
Key Observations:

1. **Negative Correlation:** The scatter plot shows a clear negative correlation between per capita crime rate and property tax rate. This suggests that, generally, areas with higher crime rates tend to have lower property tax rates.
2. **Linear Relationship:** The regression line, a straight line fitted to the data points, indicates a linear relationship between the two variables. This means that as the per capita crime rate increases, the property tax rate tends to decrease in a linear fashion.
3. **Outliers:** There are a few outliers, particularly on the left side of the plot, representing areas with relatively high crime rates but low property tax rates. These might be due to various factors, such as government policies, economic conditions, or specific neighborhood characteristics.
4. **Confidence Interval:** The shaded area around the regression line represents the confidence interval, indicating the range of potential values for the property tax rate at a given crime rate. A wider confidence interval suggests more uncertainty in the prediction.



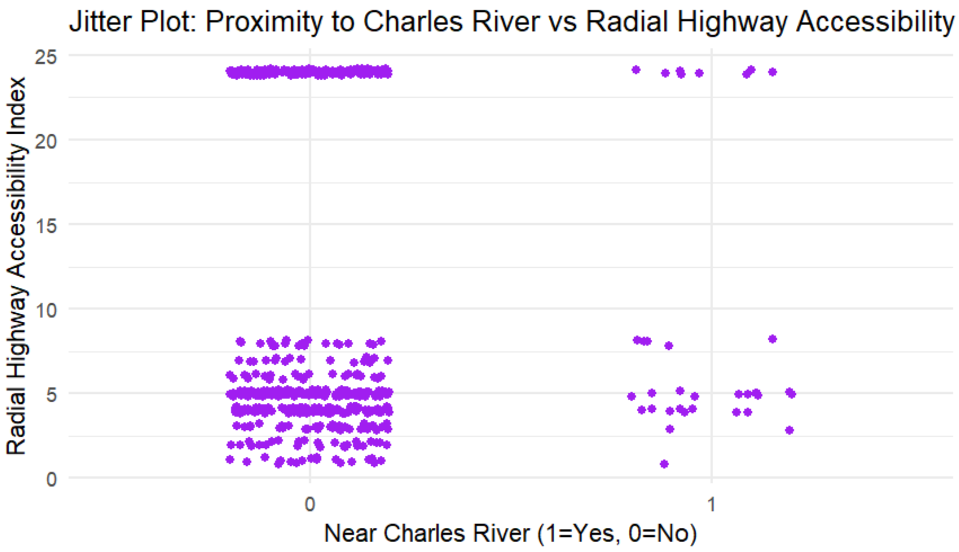
Key Observations:

1. **Positive Correlation:** The scatter plot shows a clear positive correlation between the proportion of non-retail business acres per town and the property tax rate. This suggests that areas with a higher concentration of non-retail businesses tend to have higher property tax rates.
2. **Linear Relationship:** The regression line, a straight line fitted to the data points, indicates a linear relationship between the two variables. This means that as the proportion of non-retail business acres increases, the property tax rate tends to increase in a linear fashion.
3. **Outliers:** There are a few outliers visible, representing areas with relatively high or low property tax rates compared to the general trend. These might be due to various factors, such as specific zoning regulations, economic conditions, or historical development.
4. **Confidence Interval:** The shaded area around the regression line represents the confidence interval, indicating the range of potential values for the property tax rate at a given proportion of non-retail business acres. A wider confidence interval suggests more uncertainty in the prediction.



Key Observations:

1. **Nonlinear Relationship:** The scatter plot shows a clear nonlinear relationship between the average number of rooms per dwelling and the distance to employment centers. This indicates that the relationship is not a simple linear one, and a more complex model is needed to capture the underlying pattern.
2. **Curvilinear Trend:** The red curve, representing the nonlinear regression model, suggests a curvilinear trend. The curve initially slopes downwards, indicating that as the average number of rooms increases, the distance to employment centers tends to decrease. However, after reaching a certain point, the curve starts to slope upwards, suggesting that further increases in the number of rooms might be associated with longer distances to employment centers.
3. **Outliers:** There are a few outliers visible, representing data points that deviate significantly from the general trend. These could be due to various factors, such as specific neighborhood characteristics, historical development, or data errors.
4. **Confidence Interval:** The shaded area around the regression curve represents the confidence interval, indicating the range of potential values for the distance to employment centers at a given average number of rooms. A wider confidence interval suggests more uncertainty in the prediction.



Key Observations:

1. **Distinct Clusters:** The jitter plot shows two distinct clusters of data points, one centered around the "0" category for proximity to the Charles River and the other around the "1" category.
2. **Higher Accessibility Near the River:** The cluster of data points representing houses near the Charles River (category "1") generally has higher values for the radial highway accessibility index. This suggests that houses near the river tend to have better access to radial highways.
3. **Overlapping Distribution:** There is some overlap between the two clusters, indicating that a few houses near the Charles River have lower accessibility to radial highways, and some houses farther from the river have higher accessibility.
4. **Limited Variation:** Within each cluster, there is a limited amount of variation in the radial highway accessibility index. This suggests that proximity to the Charles River is a strong predictor of highway accessibility, with less variation within each category.

Outcomes:

● Successfully created multiple types of charts using R to visualize crime data. ● Gained insights into the distribution, frequency, and relationships within the housing 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 housing dataset. By using R, we efficiently created visual representations that allowed us to explore the data from different perspectives, leading to better-informed conclusions.