Bharatiya Vidya Bhavan’s

SARDAR PATEL INSTITUTE OF TECHNOLOGY

Advanced Data Visualization

Experiment no. 5

Submitted To

Prof. Pranav Nerurkar

Submitted By

Name : Mohit Narwaiye

Class : COMPS B

Batch : A

UID : 2021300085

1. Dataset

Can view the dataset here

https://www.kaggle.com/datasets/altavish/boston-housing-dataset

2. Description

This dataset contains information about housing values in suburbs of Boston. It includes various attributes that affect housing prices, such as crime rates, average number of rooms, and property age.

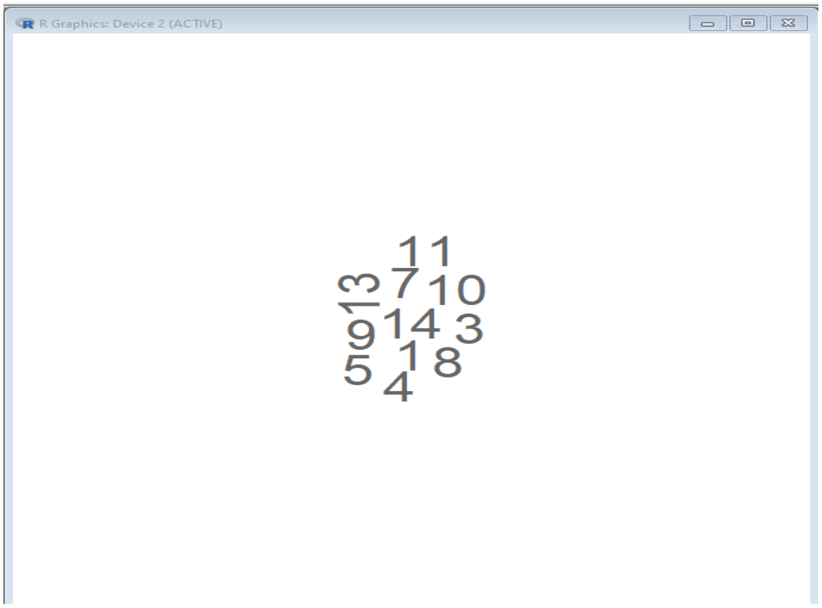
Source: UCI Machine Learning Repository or Kaggle Boston Housing Dataset**.**

3. Metadata

- **CRIM**: Per capita crime rate by town.
- **ZN**: Proportion of residential land zoned for lots over 25,000 sq. ft.
- **INDUS**: Proportion of non-retail business acres per town.
- **CHAS**: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
- **NOX**: Nitric oxides concentration (parts per 10 million).
- **RM**: Average number of rooms per dwelling.
- **AGE**: Proportion of owner-occupied units built prior to 1940.
- **DIS**: Weighted distances to five Boston employment centers.
- **RAD**: Index of accessibility to radial highways.
- **TAX**: Full-value property tax rate per $10,000.
- **PTRATIO**: Pupil-teacher ratio by town.
- **B**: 1000(Bk−0.63)21000(Bk - 0.63)^21000(Bk−0.63)2 where BkBkBk is the proportion of Black residents by town.
- **LSTAT**: Percentage of lower status of the population.
- **MEDV**: Median value of owner-occupied homes in $1000s (target variable).

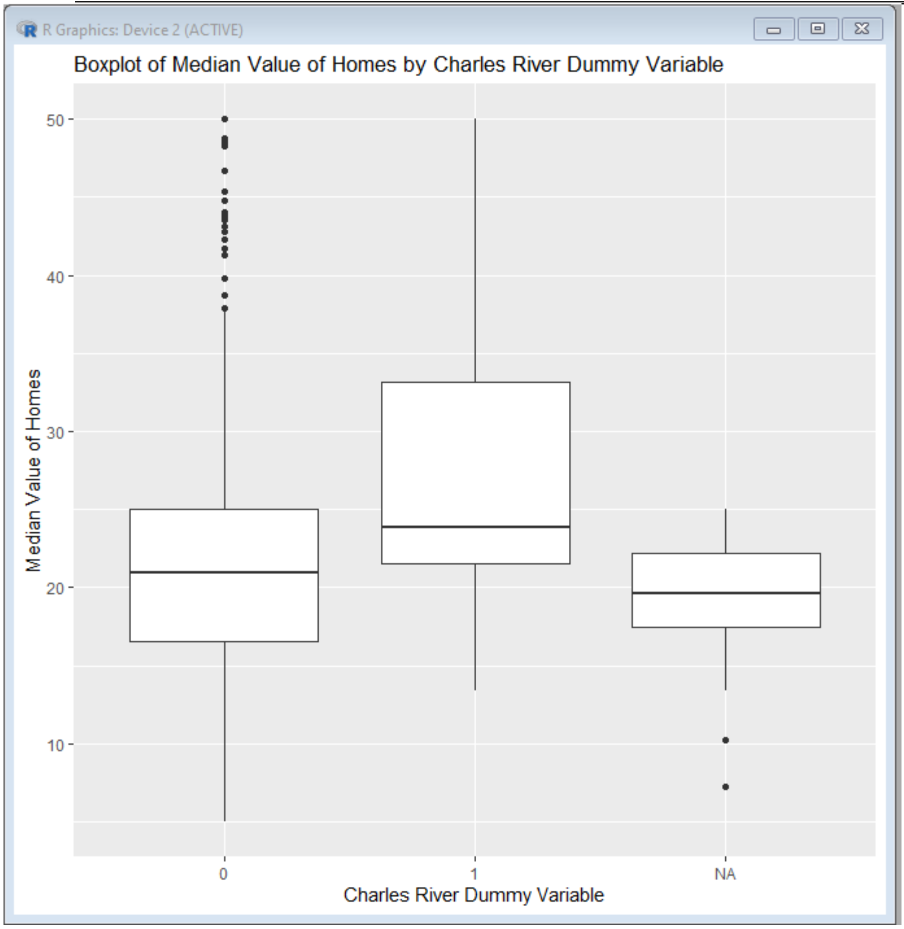
4. Visualizations and Observations

1. WordChart



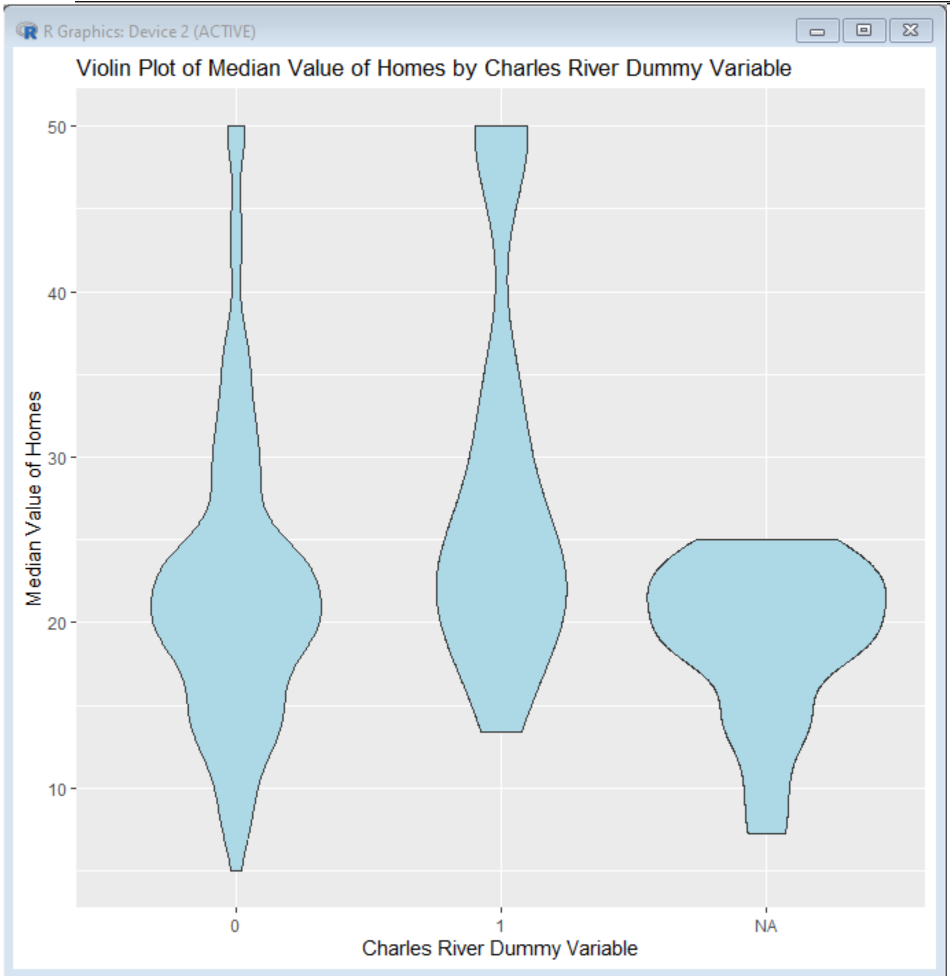
The Word Cloud visualizes the column names from the dataset. It helps identify key features or attributes in the dataset by showing the most frequent terms. Since we are using column names, it highlights the structure of the dataset.

2. Box and Whisker Plot



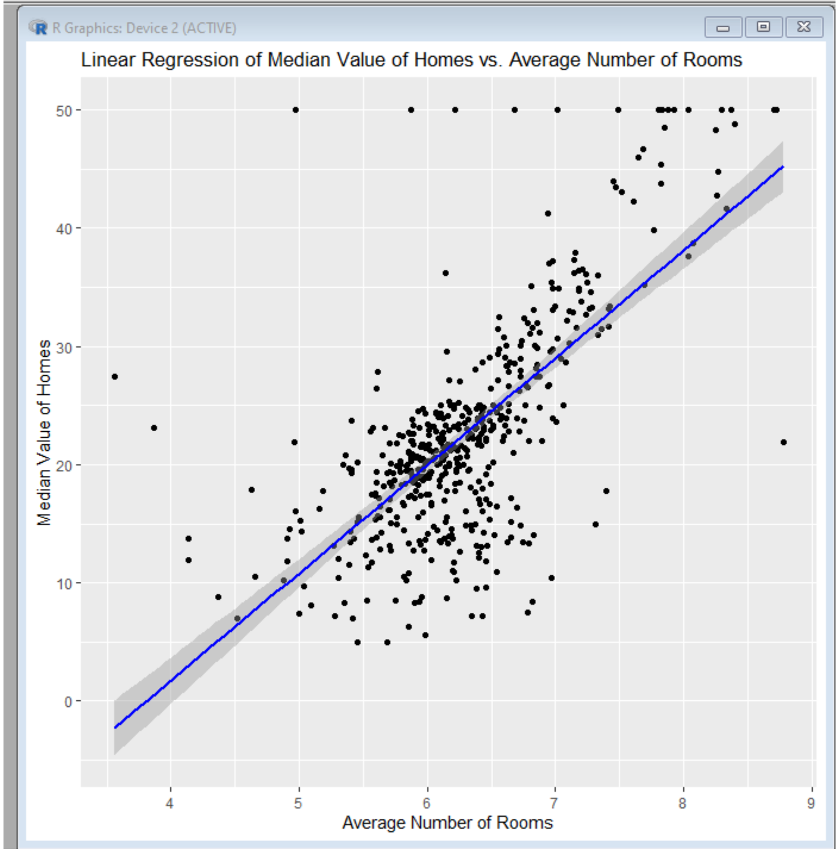
This plot shows the distribution of median home values (MEDV) based on whether the house is near the Charles River (CHAS). The boxplot will reveal the spread, median, and any potential outliers. If the distribution differs significantly between categories, it could indicate a relationship between the proximity to the Charles River and home values.

3. Violin Plot



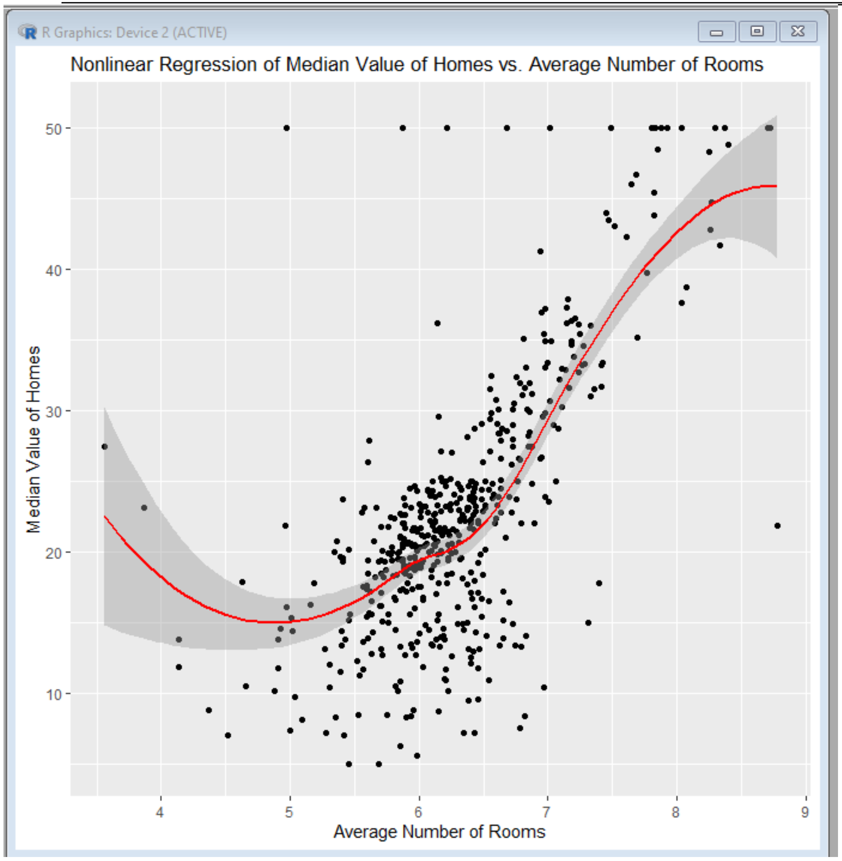
The violin plot combines features of boxplots and density plots. It shows the distribution and density of home values (MEDV) by the Charles River dummy variable (CHAS). The width of the plot indicates the density of the data, helping to visualize how home values are distributed within each category.

4. Linear Regression Plot



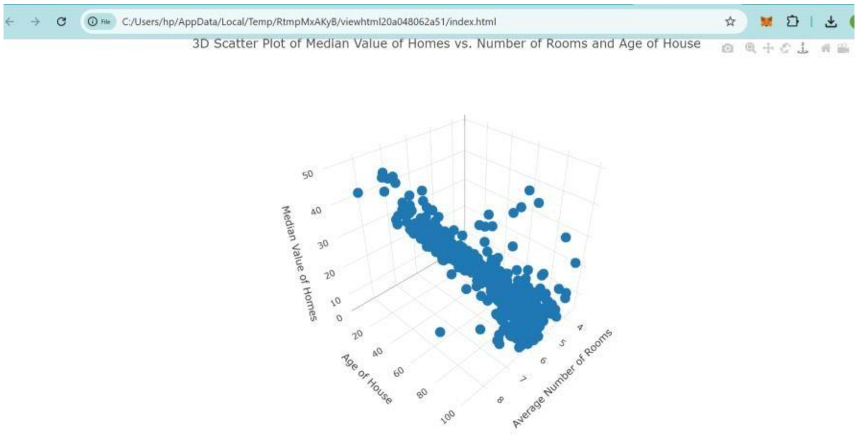
This plot shows the relationship between the average number of rooms per dwelling (RM) and median home values (MEDV). The linear regression line helps identify whether there is a linear trend where an increase in the number of rooms is associated with an increase in home values.

5. Nonlinear Regression Plot



The nonlinear regression plot uses a loess smoother to capture more complex relationships between the average number of rooms (RM) and median home values (MEDV). This plot helps identify whether there is a nonlinear trend that a linear model might miss.

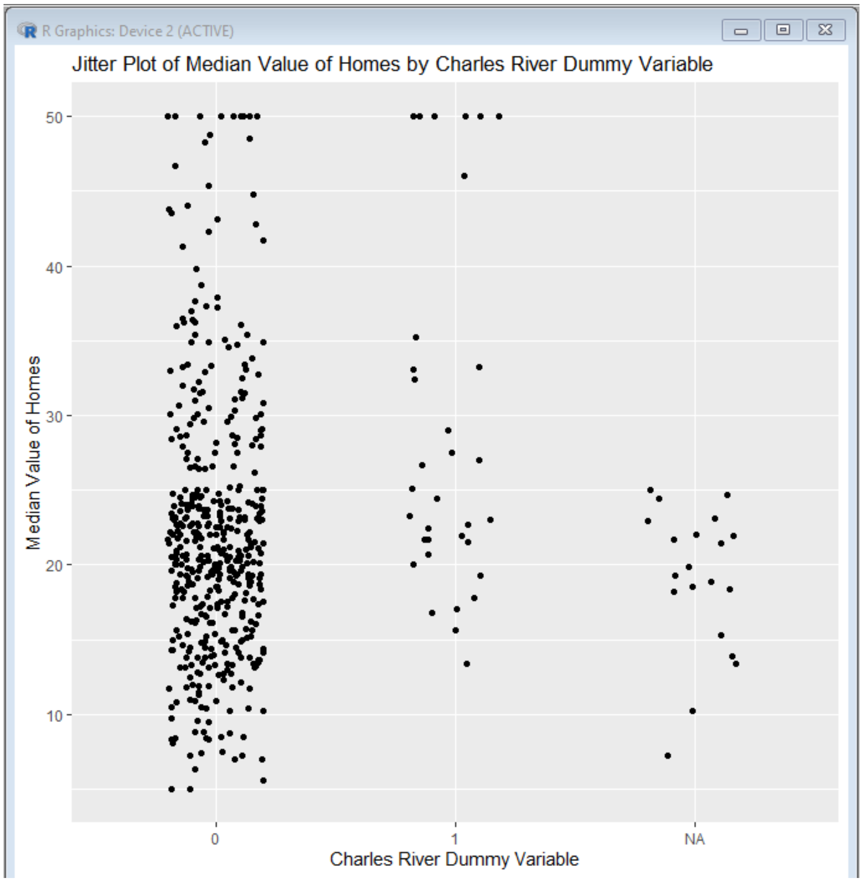
6.3D Scatter Plot



The 3D scatter plot shows the relationship between three variables: the average number of rooms

(RM), the age of the house (AGE), and median home values (MEDV). This visualization helps in understanding how these three variables interact with each other and how they influence home prices.

7.Jitter Plot



The jitter plot shows the distribution of median home values (MEDV) with minimal overlap in relation to the Charles River dummy variable (CHAS). This plot helps to visualize how home values are spread across the categories and can highlight any clustering or patterns.

5.Source Code

https://github.com/dmo-27/ADV_experiments/blob/main/exp5.R