**Module 1 - Week 1: R Practice Assignment**

**(Boston Housing Data Analysis)**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

**ALY 6010.**[**71820**](https://northeastern.instructure.com/courses/196161)**: Probability Theory and Introductory Statistics**

**Professor Antonius Breur**

**September 26th, 2024**

**Introduction:**

For my initial analysis, I selected the "BostonHousing.csv" file from GitHub to conduct statistical analysis and practice my fundamental R skills. This choice was motivated by my recent relocation to Boston and my current interest in the local housing market, making the dataset particularly relevant and engaging for me.

To begin, I loaded the necessary R libraries, including dplyr, ggplot2, reshape2, and GGally. I then imported the "BostonHousing.csv" dataset into RStudio using the read\_csv() function.

The subsequent sections of this report will present these visualizations, and the insights derived from them, focusing on factors that influence housing prices in Boston. This analysis serves not only as a practical exercise in R programming and statistical concepts but also as a valuable resource in my personal house-hunting journey.

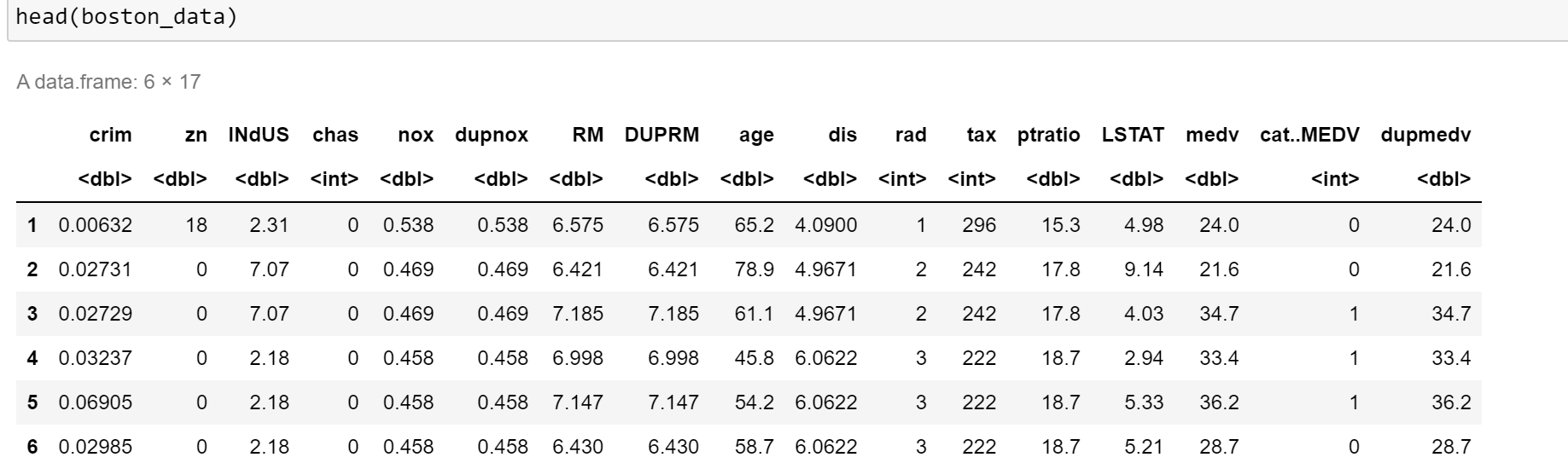
Please note that, I have renamed the actual data set before loading “boston\_housing\_untouched3.csv”.

A white rectangular object with red text

Description automatically generated

**1. Data Overview:**

* **First Few Rows of the Dataset:**



* **Column names and the data structure:**

A screenshot of a computer

Description automatically generated

**2. Data Cleaning and Preparation:**

* Columns were renamed to uppercase to maintain consistency.

A screenshot of a computer code

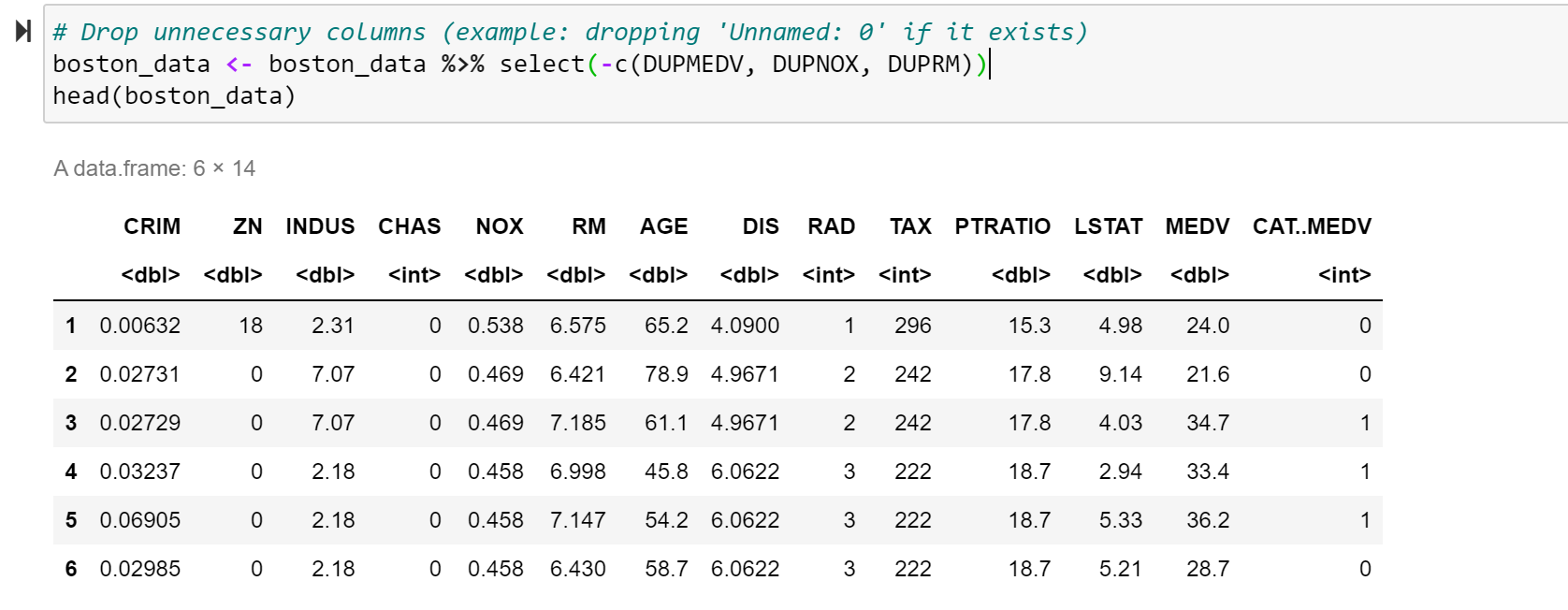
Description automatically generated

* Checked and removed missing values from the dataset.

A screenshot of a computer

Description automatically generated

* Dropping Unnecessary Columns DUPMEDV, DUPNOX, DUPRM as they are duplicates and are of no use.



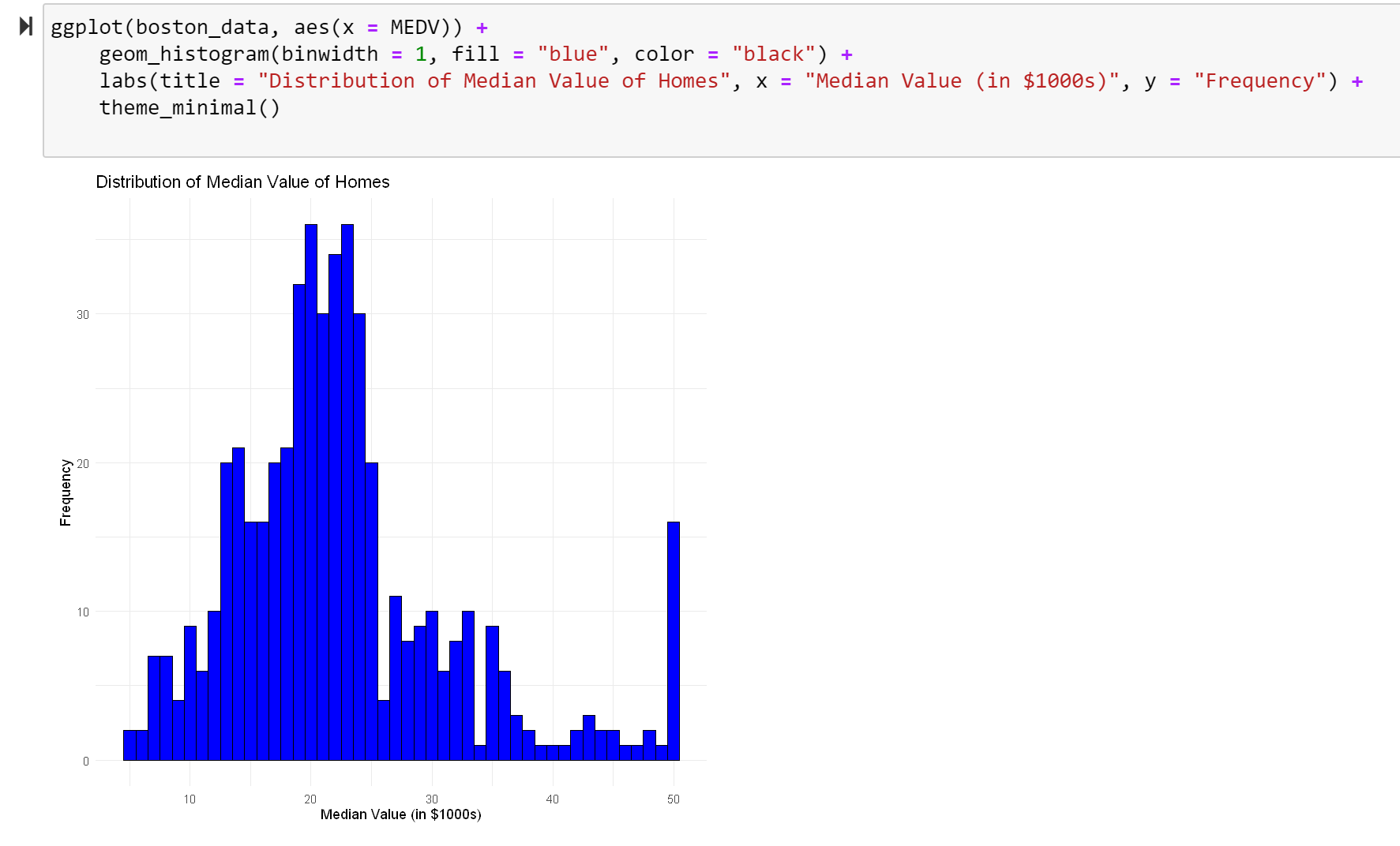
**3. Exploratory Data Analysis:**

* **Frequency table for 'CHAS':** Created a frequency table for the CHAS variable to understand how many properties are located near the Charles River.

A white rectangular object with text

Description automatically generated with medium confidence

* **Histogram for MEDV:** This histogram illustrates the distribution of median home values.



* **Bar Plot for CHAS:** This bar plot shows the count of properties near the Charles River.

A screenshot of a computer

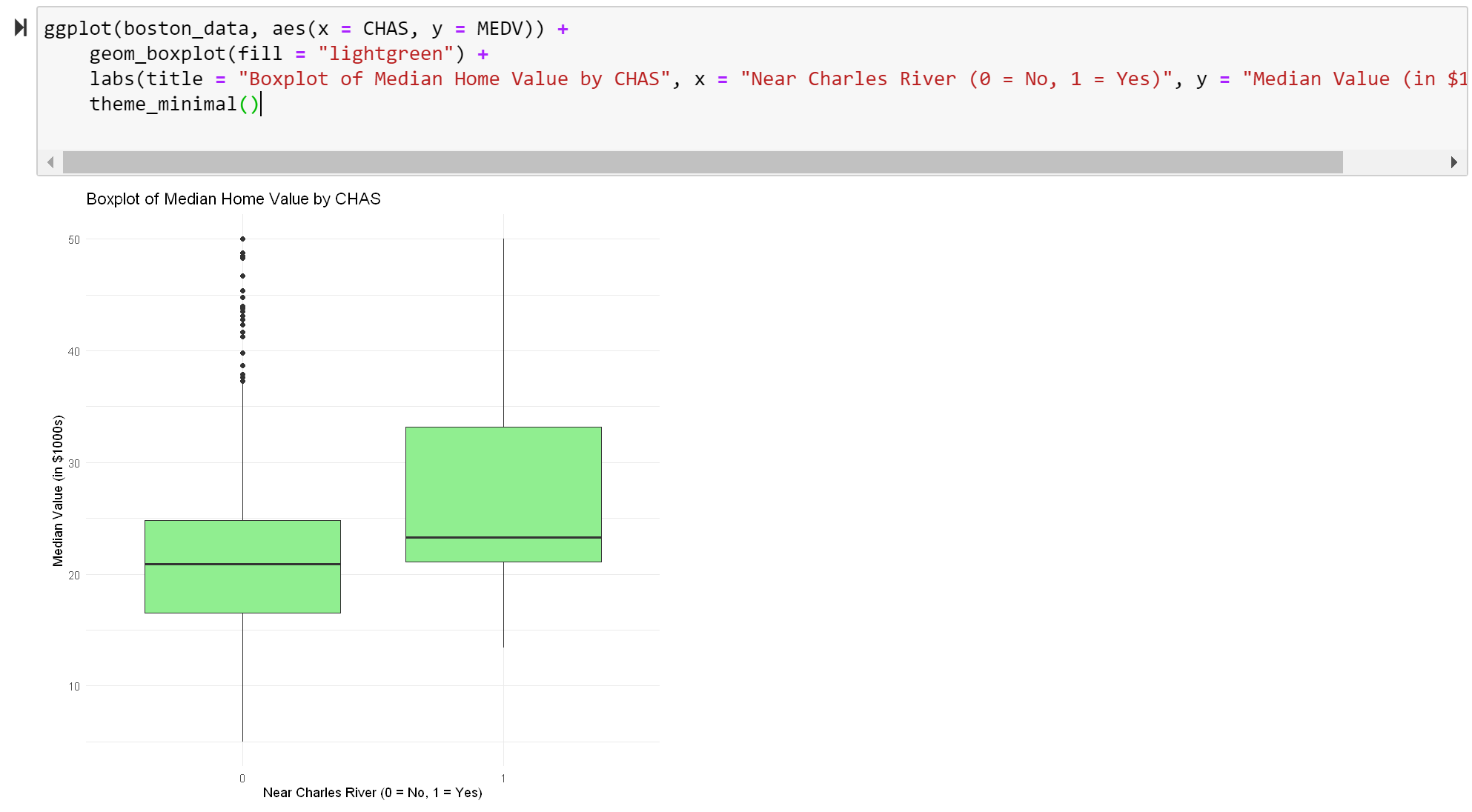
Description automatically generated

* **Scatter Plot of CRIM vs. MEDV:** This scatter plot visualizes the relationship between crime rate and median home value.

A screenshot of a computer

Description automatically generated

* **Boxplot of Median home value by CHAS:** This box plot visualizes the relationship between home value near and away from Charles River.



* **Pair plots of selected features:** "CRIM", "RM", "AGE", "MEDV".

A screenshot of a computer screen

Description automatically generated

* **Density plot of median home values:**

A screen shot of a graph

Description automatically generated

**4. Summary of Findings:**

* **Frequency Table for CHAS**:
  + This table indicates the number of properties located near the Charles River.
* **Histogram of MEDV**:
  + This histogram illustrates the distribution of median home values.
  + The data is heavily clustered between $15,000 and $35,000, with a noticeable spike at $50,000 (likely a cap in the dataset).
  + Most homes fall between $20,000 to $30,000
* **Bar Plot for CHAS**: **(Proximity to Charles River)**
  + The bar chart indicates how many properties are near the Charles River.
  + Only a small number of properties are close to the river (coded as 1).
* **Scatter Plot of CRIM vs. MEDV**: This scatter plot visualizes the relationship between crime rate and median home value.
  + Negative correlation: Higher crime rates are generally associated with lower home values.
  + The data points are clustered in the lower left corner, indicating that most neighborhoods have relatively low crime rates and median home values.
  + Outliers: A few neighborhoods have high crime rates but relatively high home values.
* **Boxplot of median home value by CHAS:**
  + Higher median home values near Charles River.
  + More variability in home values near river.
* **Pair plots of selected features:** "CRIM", "RM", "AGE", "MEDV".
  + Higher average room size (RM) is generally associated with higher median home values (MEDV).
  + Older houses (AGE) tend to have lower median home values (MEDV).
  + There is a weak negative correlation between crime rate (CRIM) and median home value (MEDV), suggesting that higher crime rates might be associated with lower home values.
* **Density plot of median home values:**
  + The distribution is skewed to the right, indicating that there are a few homes with very high values, while most homes have values in the relative lower range.
  + Most homes have median values between 20 and 25 thousand dollars.
  + The range of median home values is relatively wide.

**5. Conclusion:**

* + Proximity to the Charles River is a factor that influences home values, with homes near the river generally having higher values.
  + The distribution of median home values is skewed to the right, with a concentration between $20,000 and $30,000.
  + Crime rate and average number of rooms are also factors that influence home values. Higher crime rates are associated with lower values, while more rooms generally lead to higher values.
  + Age of the house might also be a factor, with older houses potentially having lower values.

**6. References:**

Wickham, H. and Grolemund, G. (2017) R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 1st edition. Sebastopol, CA: O'Reilly Media. Available at: <https://r4ds.had.co.nz/> (Accessed: September 26, 2024).

Grolemund, G. (2014) Hands-On Programming with R: Write Your Own Functions and Simulations. 1st edition. Sebastopol, CA: O'Reilly Media. Available at: <https://rstudio-education.github.io/hopr/> (Accessed: September 26, 2024).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning: with Applications in R. New York: Springer. Available at: <https://www.statlearning.com/> (Accessed: September 26, 2024).

selva86 (no date) Boston Housing Dataset. GitHub. Available at: <https://github.com/selva86/datasets/blob/master/BostonHousing.csv> (Accessed: September 25, 2024).