Data Mining II: Individual Boston Housing Case Study

BANA 7047-001

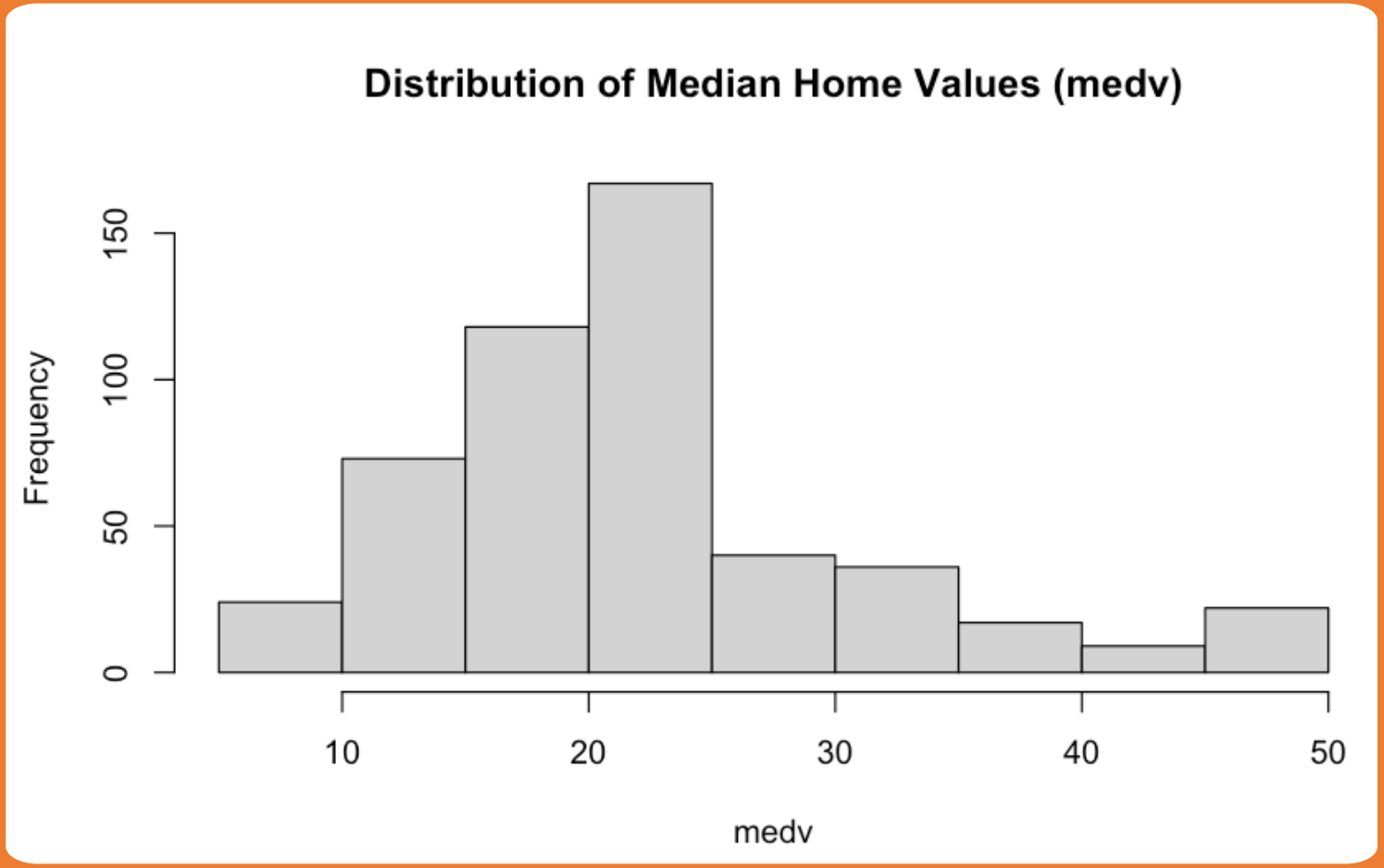
Habel, Meme

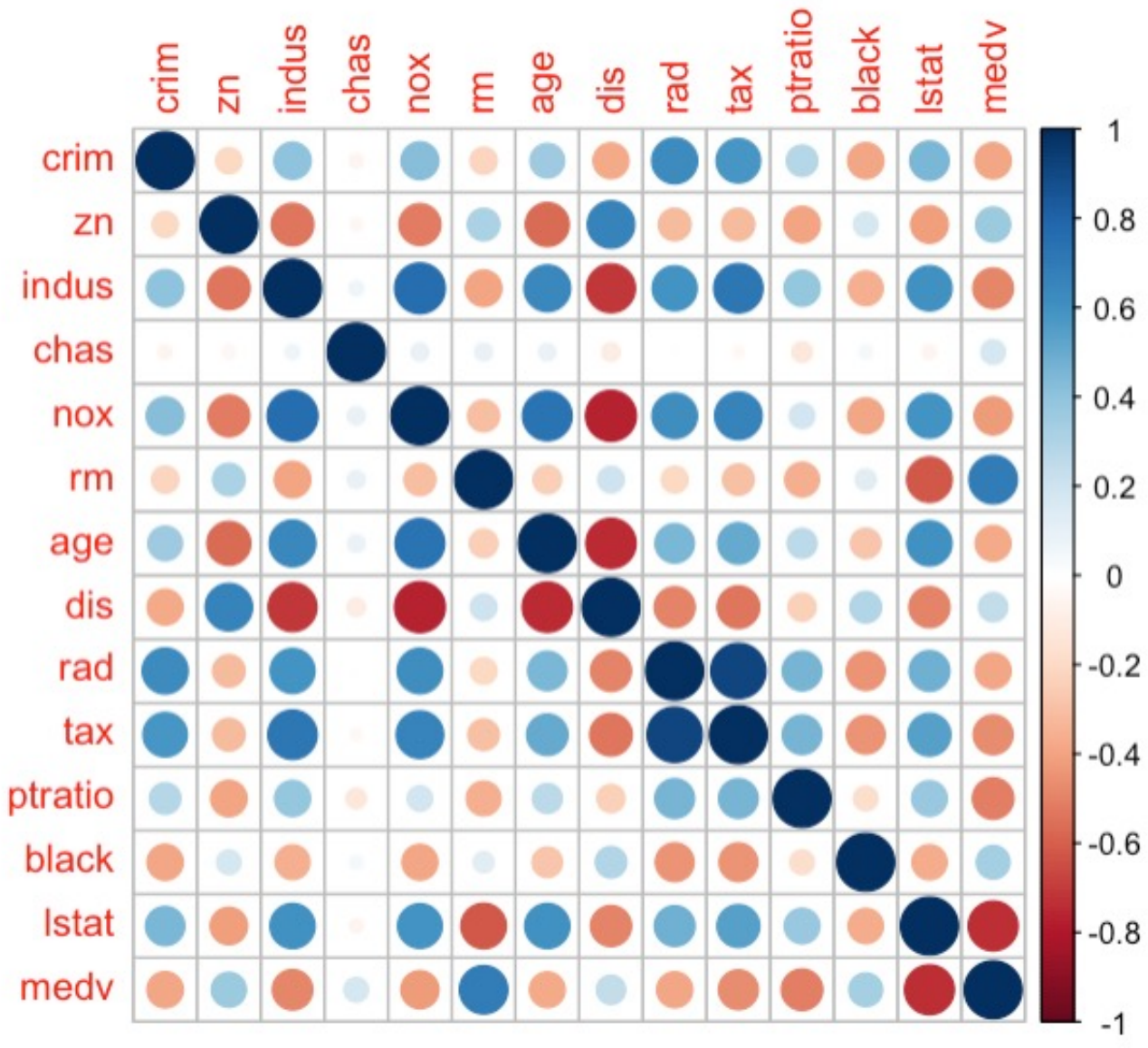
Introduction

- Boston Housing Data
  - Target variable: ‘medv’ (continuous)
- 80%/20% training and testing data split

Distribution of Median Home Values (‘medv’)

Histogram showing the distribution within the medv data values.





Correlation Plot

This correlation plot shows a visual representation of the correlation between each pair of variables.

Model Evaluation Summary

Random Seed (80%/20%) M12470675	Method "best"	Average Sum Squared Error (ASE) (in-sample)	Mean Squared Prediction Error (MSPE) (out-of-sample)
	Linear Model *(Stepwise Variable Selection with AIC)*	22.001	22.731
	Regression Tree *(Pruned)*	14.511	19.513
	k-NN with optimal k= *(scaled X)*	9.469	14.469
	Random Forests	1.999	9.811
	Boosting	7.287	14.365
	GAM	7.754	16.215
	Neural networks *(scaled X&Y)*	4.608	11.929