Michael Kamp

DATA785: Predictive Modeling

Week 1 Assignment : KNN Model

14(a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median.

Answer: See R code included with assignment submission

14(b) The goal of this analysis was to predict whether a car has high or low MPG (mpg01) using the Auto dataset. Based on graphical exploration with boxplots and scatterplots, cylinders, weight, and displacement emerged as the strongest predictors of MPG. Cars with more cylinders, higher weight, and larger engines generally fall into the low-MPG category, while lighter cars with smaller engines and fewer cylinders tend to be high-MPG. Horsepower shows a moderate relationship with MPG, whereas acceleration and year appear less informative. These patterns are illustrated in Figures 1 and 2, which highlight the relationships between each predictor and MPG classification.

Figure 1: "Combined boxplots showing the distribution of each predictor (cylinders, displacement, horsepower, weight, acceleration, year) by high and low MPG (mpg01). Cylinders, weight, and displacement show the clearest separation between groups.

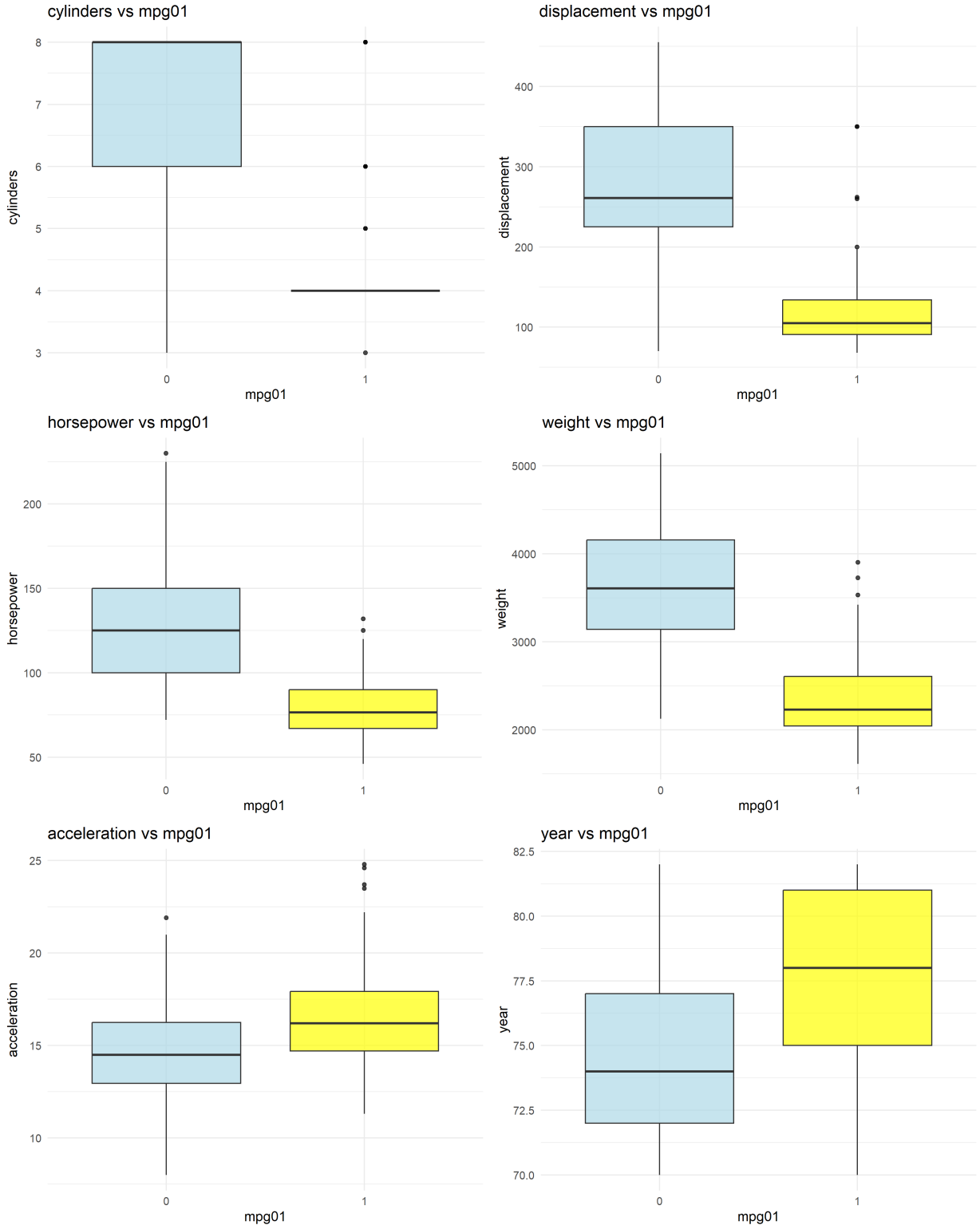
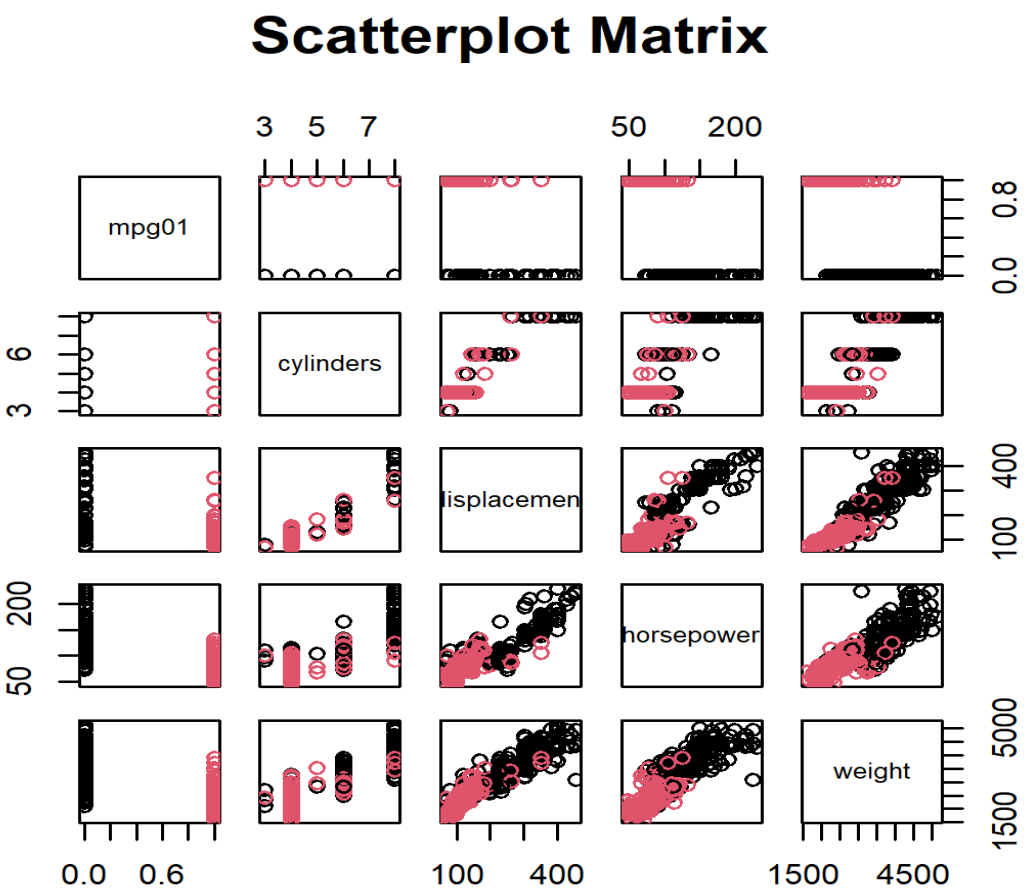


Figure 2: Scatterplot matrix of numeric predictors and mpg01. Positive correlations are visible between weight and displacement, and clear patterns highlight the separation between high- and low-MPG cars



14(c) Split the data into a training set and a test set.

Answer: The data was split into a training set and a test set using a 70/30 split.

14(h)

Answer**:** KNN was performed on the training data using the three variables most strongly associated with mpg01: cylinders, weight, and displacement. Several values of K were tested to evaluate the model's performance on the test set. The test errors for each K were calculated,

showing that smaller K values, such as 1 or 3, tended to have higher variability in accuracy, while moderate values, like K = 5, achieved lower and more stable test errors. Based on this analysis, the K value that produced the lowest test error was selected as the best-performing model. The test error plot and confusion matrices included in the submission illustrate the detailed performance for each K. Overall, this approach demonstrates that KNN can effectively classify cars into high- and low-MPG groups, and that selecting an appropriate K is crucial for balancing bias and variance in the predictions.

Figure 3: Test error plotted against different K values for the KNN model using randomly selected K values. Moderate K values, such as K = 5, produce lower and more stable test errors.

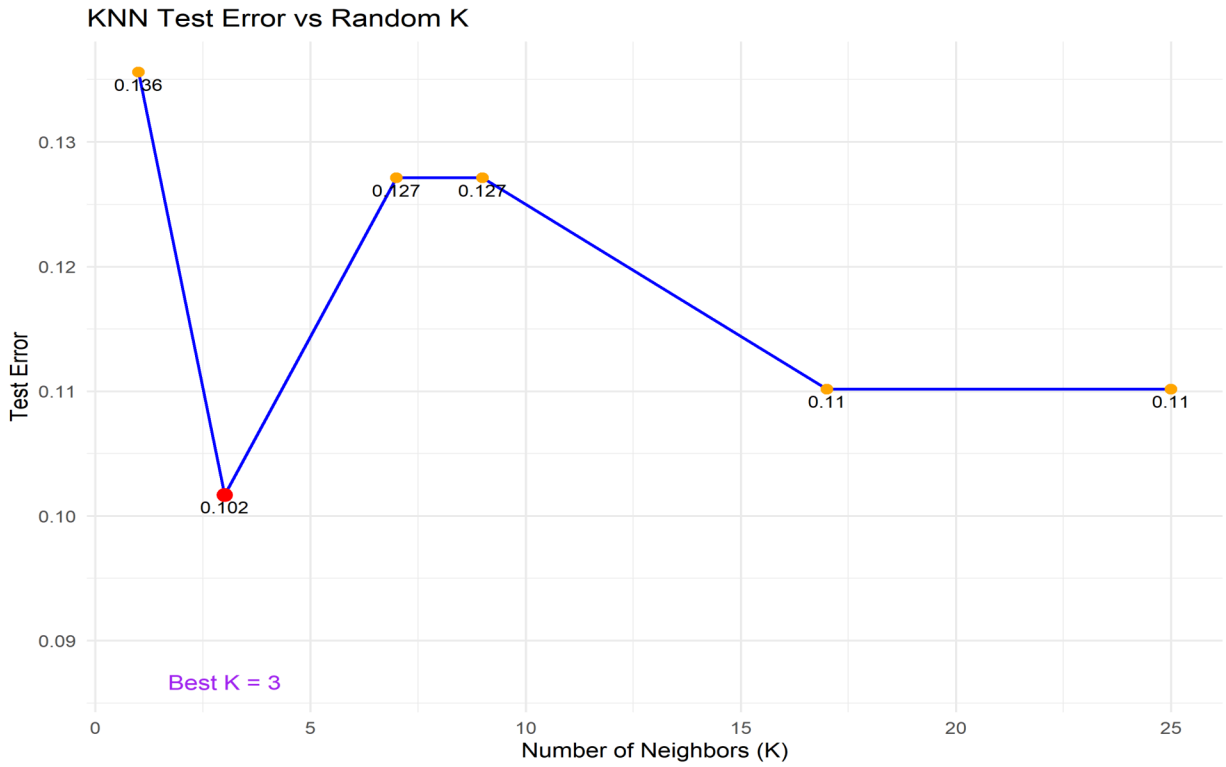


Figure 4: Confusion matrices for each K value tested in the KNN model. These show how well the model correctly classifies high- and low-MPG cars for each K

Confusion Matrix for K=1

0	1
54	9
7	48

Confusion Matrix for K=9

0	1
50	4
11	53

Confusion Matrix for K=25

0	1
52	4
9	53

Confusion Matrix for K=17

0	1
52	4
9	53

Confusion Matrix for K=7

0	1
51	5
10	52

Confusion Matrix for K=3

0	1
52	3
9	54

Confusion Matrix for K=5

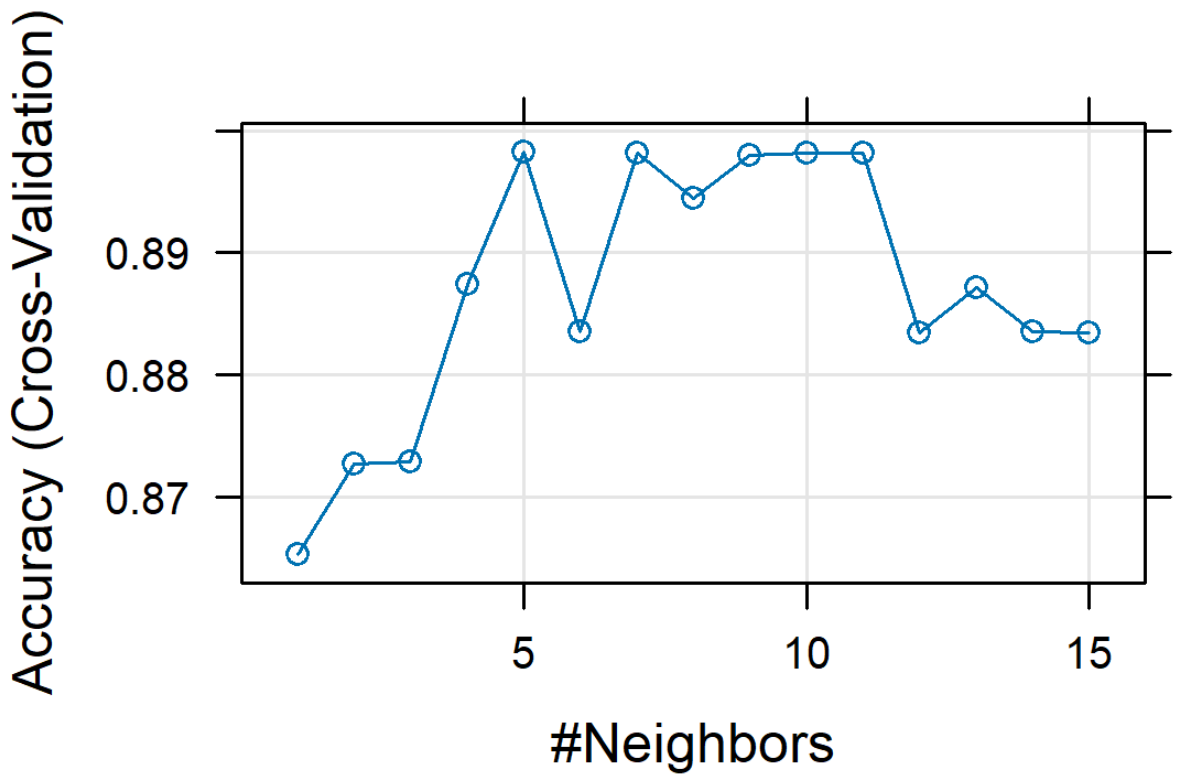
0	1
51	4
10	53

Cross-Validation Analysis:

In addition, cross-validation (10-fold CV) was used to systematically select the optimal K. The CV-selected K, K = produced a confusion matrix showing accurate classification and test

error comparable to or slightly better than the randomly tested K values. Comparing both approaches demonstrates that while the choice of K affects model stability and accuracy, both methods confirm that cylinders, weight, and displacement are the most useful features. Overall, the combination of graphical exploration, KNN modeling, and cross-validation provides a thorough understanding of the data and produces a model that effectively predicts high- and low-MPG cars while balancing bias and variance.

Figure 5: 10-fold cross-validation results for K = 1 to 15, showing test error for each K. The optimal K selected by cross-validation is indicated.



Final Summary / Conclusion:

This analysis aimed to classify cars as high- or low-MPG (mpg01) using the Auto dataset. After identifying cylinders, weight, and displacement as the strongest predictors, KNN models were evaluated across several K values. Test error plots and confusion matrices showed that moderate K values (around 5) achieved lower and more stable test errors. Cross-validation confirmed the optimal K, balancing bias, and variance effectively. Overall, the KNN analysis demonstrates that using the most informative variables allows accurate classification of MPG categories, supporting the insights gained from the initial graphical exploration.

References / Acknowledgements

Mathews, J. (2021). *R programming for dummies* (3rd ed.). Wiley.

OpenAI. (2025). *ChatGPT (GPT-5 mini) [Large language model]*. Retrieved September 8, 2025, from https://chat.openai.com

The above resources were used to guide the R coding, KNN analysis, and report preparation for this assignment.