**Intro to ML (Summer 2022) – Final Group Project:**

* Purpose: To predict the weekly sales of Walmart giving a Kaggle data set
* Cleaning of the data: In order to get the most accurate results when predicting the weekly sales of Walmart, we decided to clean the data through the removal of any extreme outliers pertaining to each of our predictors
  + In order to do this, we made box plots in Python of each of our weekly sales predictors, such as CPI, Unemployment, Fuel Price, and Temperature
  + When plotted, we can clearly see unusual observations in the data, and decided to remove any severe outliers in order to get a clean dataset and ensure the accuracy of our prediction
* Implementation: Use three different models taught in class to display our results and analyze how results from each model vary
  + Multiple Linear Regression (parametric)
  + Boosting (non-parametric)
  + Random Forest (non-parametric)
  + Split the data into a training and test set and use the training set to fit the data to each of our selected models
  + Use the test set to measure effectiveness of each model through accuracy, RMSE, and the relevance of each predictor on weekly sales
* Results:
  + Multiple Linear Regression: was least accurate (13%), it also used all predictors and we noticed an accuracy tradeoff for more interpretability
  + Boosting: Since the parametric model did not work, we decided to test the boosting model on our data and found that we had 91% accuracy
  + Random Forest: Our accuracy was 92% in this model and the ideal number of variables was seen to be 5 (around half of all predictors in the set)
  + The most relevant predictors were CPI and Unemployment (macro-trends) rather than more small-scale factors such as temperature and month