**Slide 1:**

Introduce ourselves.

**Slide 2:**

Explain that our data set came pre-cleaned, give a few examples of the type of information contained and where the data came from.

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the ASA Statistical Graphics Section's 1995 Data Analysis Exposition.

**Slide 3:**

Methods that we used to analyze the data.

Linear Model

Decision Trees

Bagging

Random Forest

**Slide 4:**

Linear Model Results

**Slide 5:**

Tree Results (Side by side)?

**Slide 6:**

Bagging Results

An *ensemble* method is an approach that combines many simple “building block” models to obtain a single and potentially very powerful model. These simple building block models are sometimes known as *weak learners* since they may lead to mediocre predictions on their own.

**Slide 7:**

Random Forest Results

The number of trees B is not a critical parameter with bagging; using a very large value of B will not lead to overfitting.

**Slide 8:**

Boosting

The number of trees B. Unlike bagging and random forests, boosting can overfit if B is too large, although this overfitting tends to occur slowly if at all. We use cross-validation to select B.

**Slide 9:**

Comparison of methods and results

**Slide 10:**

Summary Slide!!

**Decision Trees:** Unpruned and pruned decision trees show similar MSE and error rates.

**Ensemble Methods:** Bagged, Random Forest, and Boosting models generally outperform individual decision trees and linear models, as evidenced by lower MSE and error rates.

**Model Complexity:** Bagged and Random Forest models with larger values of B (number of trees) tend to have lower MSE and error rates, indicating the benefit of increasing model complexity.

**Comparative Performance:** Boosting models show competitive performance, offering lower error rates compared to decision trees and linear models.

**Model Interpretability:** Linear models have higher error rates compared to ensemble methods, indicating potential limitations in capturing complex relationships present in the data.