Choose one of the following data sets. You will use the methods we have selected for that data set.

2004 New Cars and Trucks Data: 04cars.csv

* Data dictionary: <http://jse.amstat.org/datasets/04cars.txt>
* Response variable: Retailprice (quantitative)
* Methods: Robust regression and Elastic net

Australian Athletes data set: Athletes.csv

* Data dictionary: <https://www.rdocumentation.org/packages/alr4/versions/1.0.5/topics/ais>
  + To simplify the prediction problem, we have replaced the Sport variable (which had 10 categories) with a Sport\_group variable (which has 3 categories).
* Response variable: Sport\_group (categorical with 3 levels)
* Methods: KNN and Random forest

Ozone data set: Ozone\_LA.csv

* Data dictionary: <https://rdrr.io/rforge/bfp/man/ozone.html>
  + This data set is a good choice if you want to work with missing data.
* Response variable: hour\_average\_max (quantitative)
* Methods: Linear regression and Boosted trees

Pima Indians Diabetes data set: diabetes.csv

* Data dictionary: See p. 2 of <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf>
* Response variable: Outcome (categorical with 2 levels)
* Methods: Single decision tree and QDA
  + For QDA, investigate whether the assumptions are satisfied.  How does (or doesn’t) this affect the results?