Data scraping, ingestation, and modeling: bringing data from cars.com into the intro stats class

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

CAUSE webinar, November 21, 2017

nhorton@amherst.edu
http://nhorton.people.amherst.edu,
https://github.com/Amherst-Statistics/Cars-Scraping-Webinar

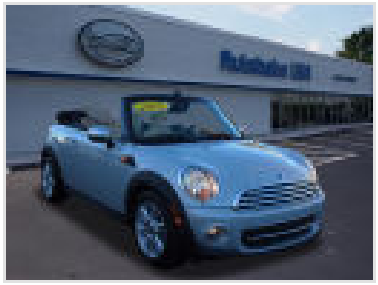
Thanks and acknowledgements

- Danny Kaplan (for the original idea)
- Project MOSAIC: Danny Kaplan (Macalester College), Randy Pruim (Calvin College), Ben Baumer (Smith College), and Johanna Hardin (Pomona College)
- NSF # 0920350

Goal

- I will describe a classroom activity where pairs of students hand scrape data from cars.com, ingest these data into R, then carry out analyses of the relationships between price, mileage, and model year for a selected type of car.
- This early in the semester activity can help illustrate the statistical problem solving process.
- The "Less Volume, More Creativity" approach utilized by the mosaic package facilitates the analysis with a minimal amount of syntax.
- Key concepts that are introduced and reinforced including data ingestion, multivariate thinking through graphical visualizations, and regression modeling.
- Extensions and additional use of the dataset will be discussed along with potential pitfalls.

Cars, cars, and more cars



24 Photos/Video

2012 MINI Cooper Base — **$22,500**

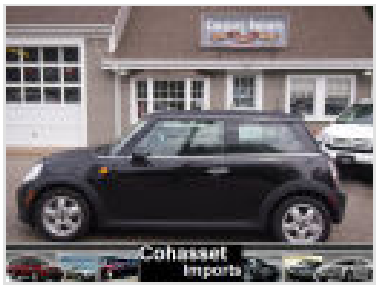
Highclass Gray Metallic, 2 door, FWD, Convertible, 6-Speed Manual, 1.6L I4 16V MPFI DOHC, Stock# MI265375. — 9,844 mi.

Autobahn USA ~ 47 mi. away

888-233-5057 Email Dealer

☐ Save/Compare

☑ Free CARFAX Report



15 Photos/Video

2011 MINI Cooper Base — **$22,500**

Midnight Black Metallic, 2 door, FWD, Hatchback, Automatic, 1.6L I4 16V MPFI DOHC, Stock# 093365. — 13,370 mi.

Cohasset Imports ~ 87 mi. away

888-586-6530 Email Dealer

☐ Save/Compare

☑ Free CARFAX Report

2012 MINI Cooper Base — **$22,165**

16,737 mi.

Questions?

- How much do cars cost?
- How much do car prices vary?
- How are car prices associated with mileage?
- How are car prices associated with age?
- How quickly do new cars depreciate?
- How much does it cost for a car to drive a mile?

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

revised GAISE College Report (2016)

1. Teach statistical thinking.
   - Teach statistics as an investigative process of problem-solving and decision-making.
   - Give students experience with *multivariable thinking*.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

Motivation for multivariate thinking

- How much do cars cost?
- How much do car prices vary?
- How are car prices associated with mileage?
- How are car prices associated with age?
- How quickly do new cars depreciate?
- How much does it cost for a car to drive a mile?

Closing thoughts

- Ensure that students see multivariate examples early and often
- Ensure that students use real tools
- Once they have some experience with "tame data", have them ingest their own
- Motivate automated data scraping procedures
- Practice composing and answering questions with data

Data scraping, ingestation, and modeling: bringing data from cars.com into the intro stats class

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

CAUSE webinar, November 21, 2017

nhorton@amherst.edu
http://nhorton.people.amherst.edu,
https://github.com/Amherst-Statistics/Cars-Scraping-Webinar