Data scraping, ingestation, and modeling: bringing 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 acknowledgments

- 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` and `ggformula` packages 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**

9,844 mi.

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

Autobahn USA ~ 47 mi. away

888-233-5057 Email Dealer

☐ Save/Compare

☑ Free CARFAX Report

15 Photos/Video

2011 MINI Cooper Base — **$22,500**

13,370 mi.

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

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 to drive a car one 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

- We live in a multivariate world
- If intro stats only addresses bivariate questions (e.g., two-sample t-test) we risk becoming irrelevant
- Straightforward to consider multivariate visualizations and fit multiple regression models (Project MOSAIC "Less Volume, More Creativity")
- *R Journal*: `https://journal.r-project.org/archive/2017/RJ-2017-024`

Cars.com Activity

Groups	of size two
Duration	one or two 50-minute class periods
Requirements	one computer per student
Software	R/RStudio and Excel, Open Office, or Google Spreadsheet
Motivation	Why might we care about car prices?

Cars.com Activity

Group is given a major city in the US (e.g., Atlanta or Los Angeles)

Person 1 searches `cars.com` for used Toyota Prius cars on offer within 50 miles of that city

Person 2 downloads the template `cars.csv` spreadsheet and open it up on their computer

Person 1 reads out car models, year, mileage, and price

Person 2 enters the values into the spreadsheet and reads them out for Person 1 to check

Continue until 40 cars have been entered (note that some groups will be really slow, so may only yield 20-25 cars)

Person 2 emails Person 1 the `cars.csv` spreadsheet then both members upload this into RStudio

Cars.com analysis (part 1)

- The file `student.Rmd` reads `cars.csv`
- Generates descriptive statistics (for data quality assessment, e.g., using $ in price)
- Create visual multivariate displays (e.g., mileage by year)
- Fit regression model and report coefficients
- Scaffolding for additional analyses by the group

Cars.com analysis (part 2)

- Once `student.Rmd` runs without error, the creative step begins
- Students need to find an interesting display and fit a multiple regression model
- Goal is to make an insight
- Publish this on `rpubs.com` using a class-wide login provided by the instructor
- Quickly review several of these to see insights
- Deliverable: full credit for email to instructor (cc-ed to group partner) of modified `student.Rmd` and `cars.csv` files

Examples

```
glimpse(ds)
```

```
## Observations: 40
## Variables: 6
## $ car      <fctr> Toyota Prius, Toyota
## $ model    <fctr> Two, Two, Four, Four,
## $ price    <dbl> 21950, 20887, 19998, 1
## $ year     <int> 2016, 2016, 2013, 2014
## $ mileage  <int> 6255, 9997, 30322, 462
## $ location <fctr> Atlanta, Atlanta, Atl
```

Examples

```
favstats(~ price, data=ds)
```

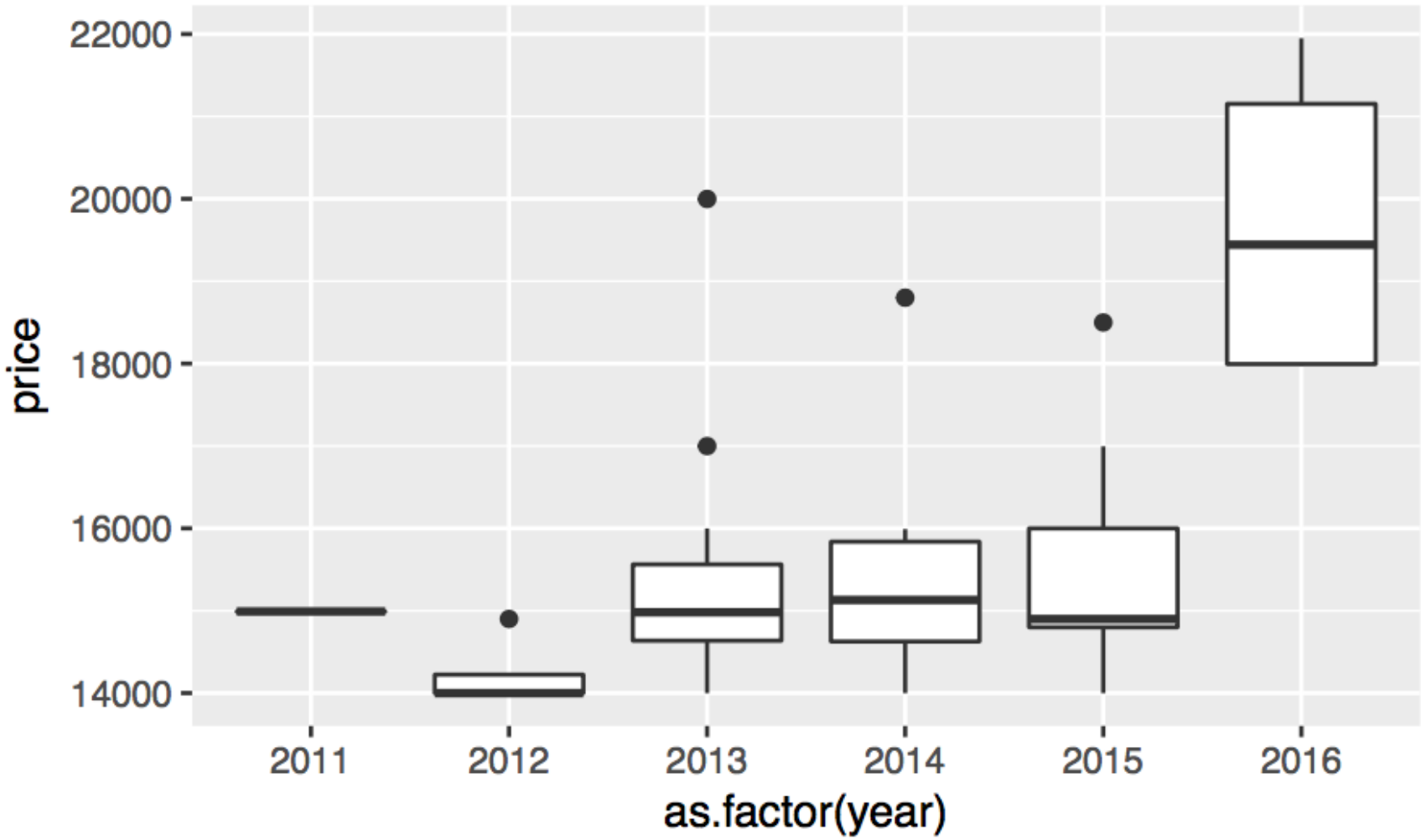
```
##    min    Q1 median    Q3   max     mean       sd  n miss
##  13999 14596  14981 15999 21950 15743.92 1965.711 40
```

```
favstats(~ mileage, data=ds)
```

```
##   min       Q1 median    Q3   max     mean       sd  n mi
##  6255 32688.25  40002 49950 63546 40057.38 14701.67 40
```

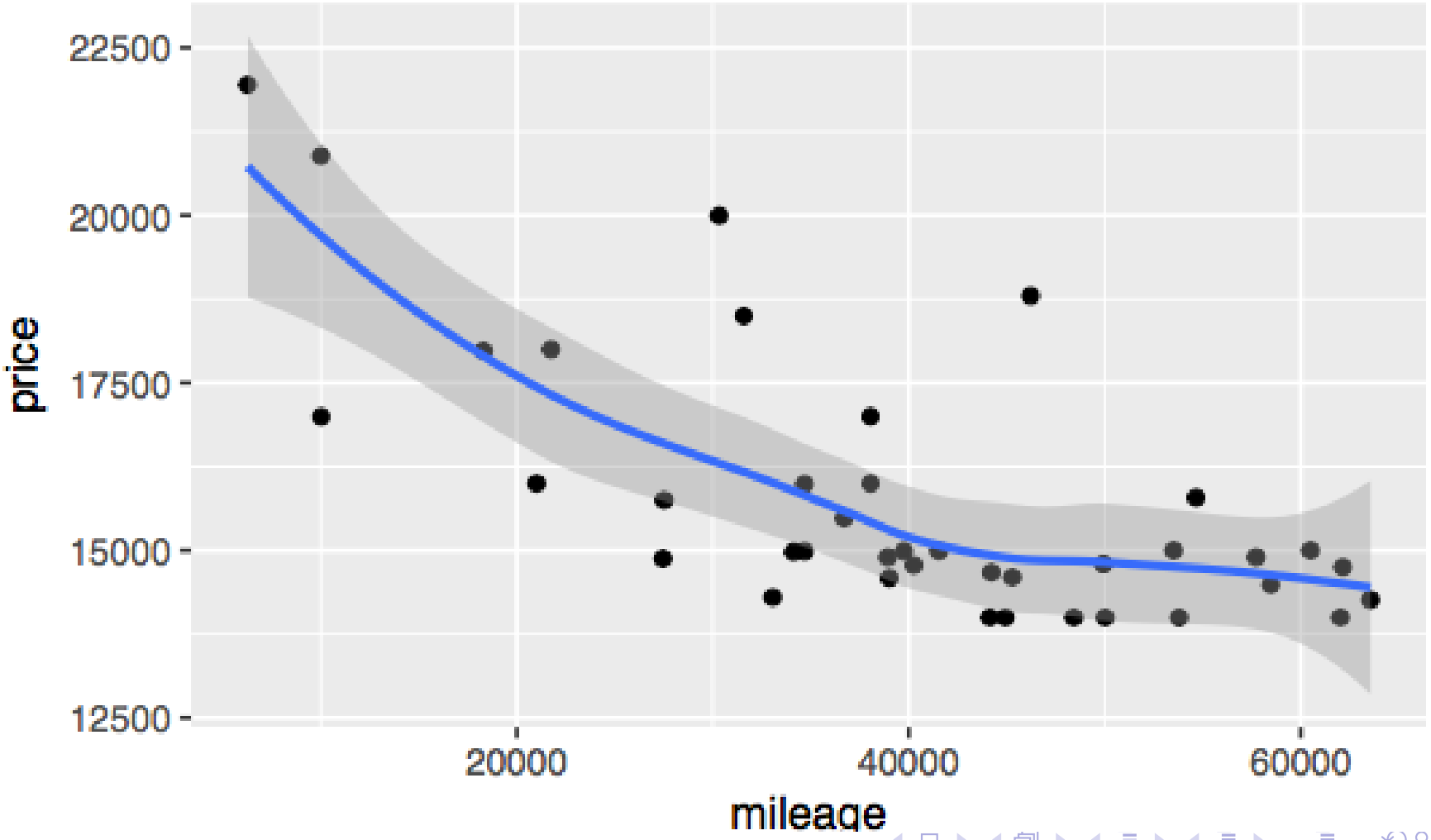
Examples

```
gf_boxplot(price ~ as.factor(year), data=ds)
```



Examples

```
gf_point(price ~ mileage, data=ds) %>%
  gf_smooth()
```



Cars.com followup (next class)

- Collate individual group data into `carscollated2017.csv`
- Perform data cleaning (e.g., numeric zip code rather than city)
- Add location to the multiple regression model
- Practice interpreting regression models with categorical predictors
- Practice interpreting regression models with interactions between mileage and year
- Seek insights: depreciation dramatic for new cars

A used car is best bet

By Tom and Ray Magliozzi
Published: Oct. 26, 2001 12:00 a.m.
Updated: Oct. 26, 2001 10:42 a.m.

+ Leave a commer

Question: My beautiful, normally intelligent wife of 24 years and I disagree mightily about the best timing to buy and sell a vehicle. We're absolutely positive we're each right, and we're absolutely positive the other is wrong. The argument involves economics — how to spend the least amount of money. I say you should buy a car with about 60,000-80,000 miles on it and drive it into the ground. She thinks it's better to buy a 1- or 2-year-old car and keep it only for two or three years. It's time to replace my "driven into the ground" '87 Nissan pickup, and we need your advice. — Kurt

Tom: It's great to get letters from lovebirds like you two, Kurt. If this is all you've got to argue about, things must be pretty good.

Ray: Here's the story. Speaking from a purely economic point of view — how you spend the least amount of money on cars — you're more correct than she is. If you buy an old car, which has already taken the bulk of its depreciation hit, and then drive it into the ground, you will spend the least.

Tom: We actually wrote a pamphlet about this very subject, called "How to Buy a Great Used Car: What Detroit and Tokyo Don't Want You to Know." In it, we lay out several money-saving used-car strategies, and we prove mathematically that the "heap strategy" is the cheapest. If

Cars.com followup

```
tally(~ year, margins=TRUE, data=ds)
```

```
## year
##  2007  2010  2011  2012  2013  2014  2015  2016  2017 Total
##     2     5    37    45   176   237   201   126     2   831
```

```
tally(~ location, margins=TRUE, data=ds)
```

```
## location
##           40202         Atlanta     Bangor, ME   Baton Rouge
##              40              40              40            40
##         Buffalo         Chicago       Cleveland        Dallas
##              40              41              26            41
```

Cars.com followup

```
## locationSeattle                  2136.54194    463.
## locationTampa                   -2152.29736    462.
## mileage                            -0.06065      0.
## as.factor(year)2012              -251.31079   1135.
## as.factor(year)2013              3237.23166    894.
## as.factor(year)2014              3140.19070    888.
## as.factor(year)2015              3252.51391    885.
## as.factor(year)2016              8208.61054    874.
## mileage:as.factor(year)2012         0.01709      0.
## mileage:as.factor(year)2013        -0.01797      0.
## mileage:as.factor(year)2014        -0.00343      0.
## mileage:as.factor(year)2015        -0.00989      0.
## mileage:as.factor(year)2016        -0.18186      0.
##
## Residual standard error: 2040 on 790 degrees o
## Multiple R-squared:  0.736,  Adjusted R-square
```

Cars.com followup

```
carfun <- makeFun(mod)
carfun(year=2016, mileage=1000, location='Tampa')
```

```
##     1
## 22875
```

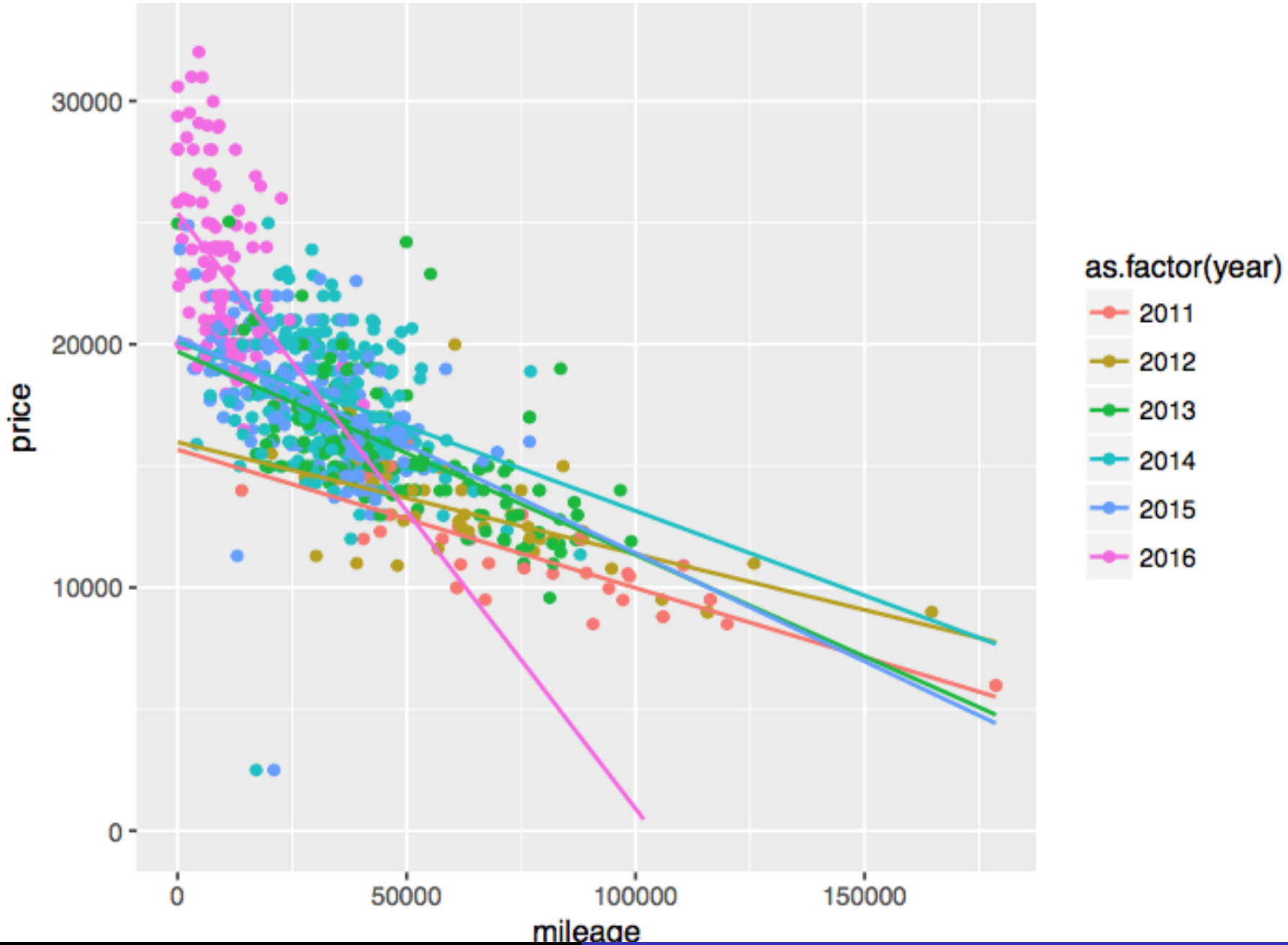
```
carfun(year=2016, mileage=1000, location='Tampa') -
carfun(year=2016, mileage=1001, location='Tampa')
```

```
##      1
## 0.2425
```

```
carfun(year=2012, mileage=1000, location='Tampa') -
carfun(year=2012, mileage=1001, location='Tampa')
```

```
##       1
## 0.04355
```

Cars.com followup



Extension

- Interpreting this model as an exam question
- Residual diagnostics
- Functional form of relationship (non-linear?)
- Outlier detection
- Account for different car models (sparsity and inconsistent coding)
- Automate data scraping

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.

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 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