1 · Data exploration and simple models

Due Monday, February 1, 2016

(1) Warm-ups

(A) Complete the R walkthrough on the Titanic survival data (nothing to turn in for this). Then work through the case study on the Austin City Limits music festival, available here and linked through the class website. Turn in a short write-up of your answers.

(B) Complete the R walkthrough on city temperatures (nothing to turn in for this). Then work through the case study on Chipotle calorie consumption, available here and linked through the class website. Turn in a short write-up of your answers.

(2) Exploring multivariate data

The data in `sclass.csv` contains data on over 29,000 Mercedes S Class vehicles—essentially every such car in this class that was advertised on the secondary automobile market during 2014. For websites like Cars.com or Truecar that aim to provide market-based pricing information to consumers, the Mercedes S class is a notoriously difficult case. There is a huge range of sub-models that are all labeled "S Class," from large luxury sedans to high-performance sports cars; one sub-category of S class even serves as the official pace car in Formula 1 Races. Moreover, individual submodels involve cars with many different features. This extreme diversity—unusual for a single model of car—makes it difficult to provide accurate information to consumers.

The variables in this data set are:

id: a numerical id for this car in a larger data set on Mercedes vehicles
trim: categorical variable for car's trim level, e.g. 550, 63 AMG, etc. The trim is like a sub-model designation.
subTrim: only options are either hybrid or "unspecified"
condition: new, used, or Certified Pre-Owned
isOneOwner: true or false (one owner = true)
mileage: mileage on the car
year: the model year
color: exterior paint color of the car
displacement: engine size in liters; coded as a categorical variable
fuel: gas, diesel, hybrid
state: state in which the car was advertised

region: region of the US in which the car was advertised.
soundSystem: brand of sound system
wheelType: alloy, chrome, etc.
wheelSize: wheel size in inches, e.g. 18 inches, 19 inches, etc.
featureCount: how many features listed on the sticker does the car have.
price: the price in dollars

Your task is to explore the data set using the tools you have learned. The goal is to understand what predicts price. Find an interesting relationship involving three variables, one of which is price.[1] Present appropriate visual/numerical evidence to summarize that relationship, and explain briefly what you have found. You definitely need a picture or two, but you shouldn't need more than a few paragraphs of text.

The idea here is for you to spend a decent amount of time exploring the data and looking for relationships, but ultimately to zoom in on a single good story and tell it well. Later in the semester, we'll revisit this data set and build some fancy regression models to predict price.

[1] A natural thing to try here here is to focus on price versus something, and then independently focus on price versus something else. That would be two bivariate relationships involving price. It's good to start your data exploration by looking at bivariate relationships, but ideally that's not where it should end. The best responses will tell a story about a genuinely multivariate relationship among price and two other variables at once.

(3) Austin food critics

The data in "afc.csv" contains the following information about 104 restaurants scattered around central Austin:
Name: the name of the restaurant
Neighborhood: where the restaurant is located
Type: the type of food served at the restaurant
FoodScore: a numerical rating of the food (0–10) assigned by food critics from the Austin newspaper, with 10 being best.
FeelScore: a numerical rating of the atmosphere (0–10) assigned assigned by those same food critics, with 10 being best.
Price: average price of a meal at the restaurant, including tax, tips, and drinks

(A) Which are the two most expensive neighorhoods, on average? Which are the two cheapest? How did you judge this?

(B) Which seems to predict the price of a meal better: the food quality, or the atmosphere? How did you judge this?

(C) Now, using tools you've learned, compute a "food-adjusted value" measure for each restaurant: that is, a measure of the price of a meal at a restaurant that adjusts for the quality of the food one finds there, as judged by Austin food critics. Which are the two "best-value" neighborhoods, on average? Which are the two worst?

Describe your procedure concisely, and present visual and/or quantitative evidence in support of your judgments.