Rich Gruss

Reading: Practical Data Science With R, Ch 2 and 3

Chapter 2 describes how to use a relational database in R, but more importantly, it provides a strong justification for why: file-based data sets become unwieldy when they approach a million rows. Another important contribution of this chapter is the detailed description of some useful tools—H2, an in-memory database, SQL Screwdriver, a tool for loading data into relational databases, and SQuirrel, a free GUI database explorer. Finally, this chapter also makes an excellent point about when "big data" doesn't need to be big: unless your data doesn't sample well, such as when you're dealing with rare events, you can find the same information in a sample that you would find in the entire set, and you won't bump up against computational limits.

Chapter 3 deals with data exploration, which consists of summary statistics and visualization. Important points from this chapter: 1) No data set is perfect, and exploration helps to uncover missing values, invalid values, outliers, and unit inconsistencies. 2) visualizations can expose information that isn't easily detected in numerical summaries: the shape of the distribution, existence of sub-populations, and the presence of outliers. 3) Although this is more of a statistics point than an "R" or data science point, lognormal distributions are good for expression data that is heavily skewed, such as income.