Jericho McLeod
G00986513

Chapter 2 is, at its core, a jargon introduction chapter in the first half, with a bit of data preparation in the latter half. I'm not going to summarize the jargon introduction, but relating it to my current dataset, the relations component is something I've been dealing with.

My dataset has tournaments, each with up to eight contestants, and each contestant two lists of cards called a deck and a sideboard. In one form, I can have each card listed with each contestant and tournament, and that is indeed how the data was collected. I have converted out of the long-form to a JSON format already, so it is a confirmation of my decision to see that expounded in a textbook on data mining.

More relevant to my current research is the section on inaccurate values. One of the things I would like to consider is variations within named decklists; however, this field is created by the contestant, and is very inconsistently applied. For instance, out of 75 possible cards, a contestant may change out 30% of cards and use the same name, while another exchanges 5% of the cards and gives the list a new name altogether. Additionally, minor changes in the list may lead to suffixes and prefixes that reflect specific attributes of the list, such as 'UR Delver' compared to 'UB Delver'. While the book provides no guidance on handling inaccurate values, my current thought process is to match identical lists and provide them with the most prevalent title, in the hopes that this preserves integrity while mitigating the most significant inconsistency of the dataset.

Imbalance in my data may also prove to be somewhat problematic, and in ways that I previously had not considered. As an example, if a deck list is completely unique, and is the only observation of any of its cards, and competes in a tournament with other relatively unique decks, I will have limited capacity to relate it to common competition. Oversampling could even introduce additional issues, as there are known factors that are not contained in my dataset, and relationships matter quantitatively. Performance is more accurately described as a directed network, so broad representation of a deck list in the sample data it is difficult or impossible to overcome unbalanced data by oversampling.

The 'Getting to Know Your Data' component of this chapter is probably the most frequent thing I have been told lately, and need to continue being told. I have been advised to, at every step of my research, find some ways to examine the state of my data. Subsequently, in class, Ed Tufte was mentioned, and triggered some independent reading, and now the textbook is expanding on the point. Indeed, most of my identified data issues thus far have come from examining my data, though primarily in summary statistics, histograms, and the like.
-End assignment-

Dr. Kennedy,
My apologies if this is a bit more rambling than philosophical discussion; I'm primarily using the assignment as a sounding board to relate the reading material to my research. If I was able to consider data mining information without contextualizing it obsessively to my own experiences and data problems, though, I would have to wonder if I was in the wrong field.