The Data Science Design Manual

Solutions to Selected Exercises

Phillip Price

August 4, 2021

Introduction

As part of my ongoing effort to further educate myself, I will tackle the realm of Data Science. "The Data Science Design Manual," by Steven Skiena, is the first stepping stone. I am beholden to none but myself, so to that end I will engage only the exercises that appeal most to me. What follows is my attempt to answer those questions. No one else is checking my answers, so stay sharp!

Chapter 1) What is Data Science?

Exercise 1-1

a) Project Gutenburg (gutenburg.org) has, according to their website, an XML/RDF file containing metadata for the collection of books on the site. There are programs online that can convert this to a CSV file, or this can be done with a word processor program as a middleman to transform the XML file into a more useful filetype.

While a site like Amazon appears to be a gold mine, it would likely require either the use of their API (for which they charge a not-insignificant monthly fee in addition to a cumbersome registration process) or the use of web-scraping tools, which have the potential to get your IP address blocked by Amazon.

Open Library (openlibrary.org) has Data Dump files intended for bulk downloads. These files contain data relevant to its collection of books. See the LibrariesHacked/openlibrary-search GitHub page for more info and for guidance on turning the download into a more readable/queriable format.

The Library of Congress (loc.gov) has an API to search through its collections, but imposes rate limits to prevent denial of service attacks.

The Seattle Public Library has a dataset on its inventory as a result of City of Seattle's Open Data Program.

b) The Jockey Club (jockeyclub.com) offers some statistics that are available for download in CSV format.

Equibase (equibase.com) has PDF summaries of race statistics. Turning this into usable data would require either manual entry or the use of a natural language processing (NLP) / image recognition algorithm to parse the document for relevant info. Of note, these PDF's contain a small, more in-depth description of how the race played out in real time.

Horse Racing Datasets (horseracingdatasets.com, who would've guessed that?) contains community uploaded data sets in a variety of formats, most notably in CSV, XSLX, and Google Sheets.

c) Nasdaq (nasdaq.com) offers datasets on the historical price of individual stocks. For looking at one company, the data would require little to no alteration. Using the data of more than one company would involve manually downloading the dataset for each company and then designing a custom algorithm to merge the dataset to reflect whatever question you're interested in.

   Marketstack (marketstack.com) offers an API, limited to 1000 requests/month, that retrieves End-of-Day stock data in a JSON format.

   eoddata (eoddata.com) offers datasets for the closing prices of a wide number of stocks. This is limited by being only for a specific day. To use this data requires registering with the website first, thereupon you would have to download a file for each day of data you wanted. The files are available for download in a variety of formats, most notably CSV. There are also additional options to purchase in bulk the historical records for groups of stocks going back up to 20 years.

d) The Center for Disease Control and Prevention hosts some datasets primarily associated with the risks/indicators of chronic disease (chronicdata.cdc.gov). These datasets can be exported in CSV format. The CDC also has other datasets, some of which require an application process to view (most health-related data is on the borderline of personally-identifiable, and thus greater care is required when managing these datasets).

e) The National Center for Education Statistics (nces.ed.gov) hosts datasets pertaining to colleges and universities. These datasets are available for download in a variety of formats for use with popular programming software (R, Sas) as well as CSV.

   The Department of Education (ed.gov) also hosts links to various datsets related to colleges and universities. The datasets are in a variety of formats.

f) The FBI Crime Data Explorer provides select datasets for download, available in CSV format.

g) North American Breeding Birds Survey hosts a dataset on bird populations in the North American region. It is available in an XML format.

   data.gov also contains datasets about bird populations in a variety of formats.

   The Migratory Bird Data Center contains links to various datasets related to bird populations, although as of 2015 the links on the website are no longer being updated.

Exercise 1-8

The phone book is a limited subset of all available numbers. Not only that, but it is likely an inherently biased database (do you think Jeff Bezos or Mark Zuckerberg has their phone number in the phone book?). By dialing random strings of digits it is possible to reach those numbers not in the phone book as well as those in it, and since such polls are trying to reach a representative population rather than a specific person the random digit dialing method prevails.

Exercise 1-10

See accompanying Jupyter notebook.

Of note, one of the most interesting takeaways for me was how much time was spent processing the data. Even with the complete dataset I had access to, the majority of my time was spent working on getting the data into a form useful for analysis, although this may have had more to do with my relative inexperience in working with the Pandas library.

Exercise 1-12

At a high level, I would say regression is continuous while classification is discrete. Regression predicts the value of a numerical quantity, whereas classification determines which label among a finite set of labels is appropriate for a given input.

Exercise 1-16

See accompanying Jupyter notebook.