# Exercise 1

***Goal:***

*A script to acquire a dataset consisting of all records within the specified license type(s). Each record should contain all information associated with a license (name, license#, status, specialty, etc) that is available on the website.*

**Analyze the listed websites and answer the questions below with the goal in mind:**

1. Technology:
   1. What technology does the website utilize?
   2. What does this information tell you about how to approach scraping from this website?
   3. How did you determine the technology utilized?
2. Search Criteria:
   1. What search criteria will allow you to scrape the objective dataset most efficiently?
   2. How would you evaluate the completeness of an individual search?
3. Script Outline:
   1. Select a website below and include a script outline (pseudo-code) which achieves the goal above.

*List of Websites and associated license type*

|  |  |  |  |
| --- | --- | --- | --- |
| Source | URL | License Type | Example record |
| AR Med Board | http://www.armedicalboard.org/public/verify/default.aspx | Physician Assistant (all licenses with ‘PA-‘ prefix) | PA-340 |
| DC Health | https://doh.force.com/ver/s/ | Physical Therapist (all licenses under the license type dropdown value of ‘PHYSICAL THERAPIST’) | PT870060 |
| OR Med Board | https://techmedweb.omb.state.or.us/search | Podiatrist (all licenses under the license type dropdown value of ‘Podiatrist’) | DP00246 |
| PA Licensing System | https://www.pals.pa.gov/#/page/search | Pharmacy (all facility licenses found under the Board dropdown value of ‘State Board of Pharmacy’ and the License Type dropdown of ‘Pharmacy’) | PP481373 |

# Exercise 2

**Provide the css/xpath selector to obtain the following data from the two attached html examples:**

* Name
* Primary Specialty
* Address information
* License Information for each license on page
  + License Number
  + Original Issue Date
  + Expiration Date
  + License Status
* Board History
  + Board Minutes
  + Board Orders