**Upwork – Web Scraper for Retailers’ Websites (Posted 5/29/19)**

More Specific Details:

* NOTES: we have our own proxy subscription through proxymesh, so we can either use that or switch to it as we hand off the project
* We are still finalizing the last retailers we want to initially scrape, but here are the likely ones:
  + Target.com (see overview below)
  + Staples.com (will create same Target-like write-up)
  + Homedepot.com (will create same Target-like write-up)
  + Bedbathandbeyond.com (will create same Target-like write-up)
  + BarnesandNoble.com (will create same Target-like write-up)
  + Wholefoods.com (will create same Target-like write-up)
  + Oreillyauto.com (will create same Target-like write-up)
* For any retailer we are scraping:
  + Want to scroll through all categories / menus / submenus for every product potentially sold in stores, and get as much information as we generally can (excluding review text, etc):
    - Data we definitely need:
      * Product Name
      * Product URL
      * SKU ID (if there)
      * UPC (if there, might also be called GTIN)
      * Brand
      * Thumbnail url (image thumbnail)
    - Data that we want to capture if possible:
      * Price (might differ by location though)
      * Menu / category path for product
      * Same day pickup available
      * If only sold in stores / online
      * Note: we have a “unique attributes” column in our data that can store a hash (hstore with key value pairs) of different attributes we find useful that don’t fit in our standard set of columns
  + We want to have the data be pulled into a pandas dataframe and be pulled in Postrgres
  + We want to be able to run this automatically to update and get new product data regularly (we don’t want to delete old records, just update old ones and find new products)
* Target.com
  + How we are thinking to get every product sold in stores:
    - OPTION 1) get every product, so long as it doesn’t say “not sold in stores” (e.g. on product page: <https://www.target.com/p/az-patio-heaters-174-recycled-fire-pit-fire-glass-royal-blue/-/A-51258989>, e.g. on menu page: <https://www.target.com/c/outdoor-heating-accessories-fire-pits-patio-heaters-garden/-/N-5xtpj?Nao=0>)
    - OPTION 2) if Option 1 is wasting a ton of calls, we could potentially choose a certain home store that is large (Set location to “king of prussia mall”) and select “available in store” products.
    - OPTION 3) same as option 2, but do for multiple stores in different places to make sure we cover every product
  + Issues to be aware of:
    - Clothing: sizes and colors might make things harder
    - Pagination stops at 50 pages (1200 products I think)
    - Products might change quickly for clothing because products are seasonal