Notes on MOH Scraping

* Scrapes for cases, measles cases, confirmed cases; doesn’t look into other languages so will either need to translate tweets/posts or find what they are in all languages
* Need to add a filter for generating the final caseload data from cumulative data that filters out for when the posts are from separate outbreaks (ex: there may be caseload data from 2014 and some from 2017 but the way it does it now will calculated cases for the beginning of 2017 based off the end of 2014)
* Facebook scraping keeps timing out, need to keep rerunning
  + Facebook scraping also takes a long time, either find a new scraping package or get an API?
  + \*\*\*\*\*\*\*this is a big necessity if this is going to be continuously used to update the site; takes about a full day to run with all of its timeouts/reruns
* Don’t think this can be fixed in the scraping itself but need to manually check a lot of the results from the scraping
  + Sometimes the numbers are for other countries or aren’t actually case numbers
  + No way to distinguish between caseload for the month and other data like cases for the entire year
  + May want to consider exporting the specific tweets with caseload data to manually check the numbers (I had to go through the feeds and find the posts myself)
* At the end of the script, will probably want to transpose the sheet so countries aren’t on the top, I did it manually in excel but can probably do it in pandas too
* New Zealand is currently having an outbreak in Auckland, so no results are coming up in scraping from New Zealand twitter because they’re posting about it in the Auckland one; need to figure out how to include retweets in the twitter scraping to get retweets from cities

Debugging stuff

* I think some of the months and years are still being treated differently as ints and strs, so they’re not merging in the pd.merge in to the same column; pretty easy manual fix but you can also try to fix in the code

If being more utilized as a means to find when new outbreaks occur, can simplify the code a lot (ex: run Facebook on fewer pages, no need to calculate caseload data exactly from the script)