Solution to find codes of all airports

***Design:***

I have gone through several kinds of design to solve this problem.

Design 1: manage to find a mapping between the postal code and airports on Google and then scrap data sources to extract the table and process it into a csv.

Design 2: Find the whole list of airports, with their IATA codes mapping, which is easy to find online and then find a mapping between IATA and postal code.

Design 3: Find the whole list of airports, with their locations and then go into the locations link to look through their geographic information table on Wikipedia to search for postal code/zip code and capture them by their titles/class names.

Design 4: Find the whole list of airports, then send queries such as “Manchester Airport postal code” to Google and use the API SerpWoW to scrape and parse Google’s search result. Finally, by extracting the “answer” sent back from Google inside the “answer box”, I can get the mapping between airports and postal code.

***Difficulties faced***

Difficulty 1: For Design 1, there is no existing and available mapping between the postal codes and airports name online, so this is a dead end.

Difficulty 2: For Design 2, indeed it is feasible to find the mapping between IATA codes and airports. There is no available mapping between IATA and postal code.

Difficulty 3: For Design 3, it is feasible to find the location(with link to the Wikipedia of that location) and airport. However, in Wikipedia, some places do not have information table; some places do have geographic information table, yet it does not include postal code/zip code, but only area code; other places such as Manchester is too large that it includes a lot of postal codes and it is hard to pin down to the specific area Manchester Airport is in using that link.

Difficulty 4: For Design 4, it is possible to scrape the Google search result. However, Google has limits on the request, around 2.5 requests per day from scraping. However, to get the entire mapping between postal code and airport, it needs at least 10k requests. Also, Google only provides postal code for some large airports. For some other airports such as Allah Valley Airport in Philippines and Churchill Falls Airport in Canada, it does not provide the postal code but only IATA code.

***Implementation:***

Finally, I chose to just do the mapping between IATA and airports. Basically my solution is by scraping several data sources from Wikipedia to get the mapping between the codes and the airport. On Wikipedia, there is a wiki-table that includes information such as the IATA code, ICAO code, Airport name, Location served and so on. So what I did is that I used BeautifulSoup to scrape the whole page of the data source down and then extracts the table that contains the information I needed, followed by extracted the columns that I needed. Then I used pandas to load the columns into a data frame. Because there are 26 alphabets, so I ran a for-loop to loop through all 26 data sources that I have used and then concatenate the data frames as a final data frame and at last, I outputted the data frame out as a csv.