DSC 200 – Data Wrangling

Final Project

This final project is a group project. Groups cannot be larger than three students. Information about the due date, etc. is provided in the Canvas assignment from which this file is linked.

Instructions:

This is a two-part project. Each part is required. The sections below describe the requirement for each part of the assignment.

Part 1

Assume that you are the data scientist for a research organization that investigates current health issues. At this point the organization is interested in understanding the effect of the different variants of COVID. They want to analyze the number of persons being diagnosed, hospitalized and/or dying each period (day, week, or month) for each variant of the virus from March 2020. However, it turns out that there is no single source of data to solve the problem. You have been tasked to look for and extract data from at least five distinct reputable sources. The sources of data should include CSV files, PDFs, API and Excel files. The following are possible API sources you could use. Though provided for your guidance, you should not limit yourself to these:

- https://covidtracking.com/data/api
- https://www.mongodb.com/developer/article/johns-hopkins-university-covid-19-restapi/
- https://www.mulesoft.com/exchange/68ef9520-24e9-4cf2-b2f5-620025690913/covid19-data-tracking-api/

You are required to write a python script that downloads the datasets or extracts the data from all the sources and processes the data at a level of granularity such that all the data from the different sources could be merged into a single CSV file. (By level of granularity, we mean daily, weekly, or monthly) Your program should then save the resulting CSV file as **group_[group_number]_covid.csv**.

Note that for each dataset, write a function that performs the extraction, processing, cleaning, etc. for that dataset and should return a **pandas** DataFrame that will be merged with data from the other sources. There should be a main function that calls each of the functions in a sequential manner and reports its progress to the console. In the merged dataset, include a field (of type string and named ds_source) that contains the name of the organization from which the data was sourced. The function should download the files. In other words, you do not have to manually download and submit these files.

Part 2

A university is interested in collecting information about the currently available jobs that have something to do with the data science field. They would like to get the following information extracted from job advertising sites. Some of the job advertising sites include ziprecuriter, indeed, craigslist, and others. You are required to write a python script that scrapes data with the following headers: job title, company name, salary, city, state, job type, minimum qualifications, and desired qualifications. The minimum and desired qualifications section could be textual data. (*For 10 extra credits, you may extract the distinct job qualifications into separate columns and for each job that lists that qualification, use a Boolean value to indicate presence or otherwise of that qualification*.) If any of the websites provides an API, you are allowed to use that API as well. You are required to include a minimum of three of the job advertising sites for this part of the project. Ensure that the chosen website legally allows web scraping. The downloaded file should be a merger of data from all (at least) three websites. Name the downloaded file using the following format: ***group_[group number]_dsc_jobs.csv***.

Like part A, for each company/website, you should write a separate function for scraping the website.

Menu

Putting all together, implement a menu-based system to call the different functions from parts 1 and 2. You may write a function each for part 1 and part 2. Each of these functions may call all 3 or 5 sub main functions as appropriate to the task. Your menu should have two options, one for each part.

Rubric:

1.	Part 1		
	a	At least 5 sources with their respective functions included	5 points
	b	Functions extract the required data from listed source	60 points
	c	Data Merged and Saved as a CSV file	10 points
	d	Code appropriately commented	5 points
2.	Part 2		
	a	At least 3 sources included	5 points
	b	A function correctly implemented for each source	40 points
	c	Data multiple sources appropriately merged	10 points
	d	Code appropriately commented	5 points
3.	Appropriate Menu Implemented		10 points
	Total		150 points

What to submit:

1. Submit a single Python script named ***group_[your_group_number]_project.py*** that contains the implementation of parts 1 and 2 above via the linked Canvas assignment.