HKUST BUSINESS SCHOOL 香港科大商學院

ISOM 3400 Business Applications Development in Python

Assignment 1

Due date & time: May 3 at 11:59pm HKT

Objective

In this assignment, you will have to scrape information about the **2024 U.S. Presidential Primary Election** from the **CNN website**. The **primary election** is a crucial part of the U.S. electoral process, where political parties select their candidates to run for the presidency. **Super Tuesday** is an important day when the largest number of U.S. states and territories hold primary elections and caucuses. On this day, approximately one-third of all delegates to the presidential nominating conventions can be won. Your task is to navigate the CNN website, identify and extract county-level election data for the U.S. states and territories held on Super Tuesday, and store the results in a CSV file as output. This document will include the requirements and grading criteria for the assignment, offering guidance and helpful hints to support you in successfully completing the task.

Requirements

- Your program should perform automation and web scraping tasks on **Chrome** browser.
- Only use the libraries covered in this course: **Pandas, Selenium and Beautiful Soup**.
- Ensure your submitted program is **error-free** as it will be checked by clicking the **'Run All'** button.
- You must submit a programming file in **ipynb** format and a **CSV** file storing the required outcomes. The data stored in the CSV file must match the output generated by your programming file.

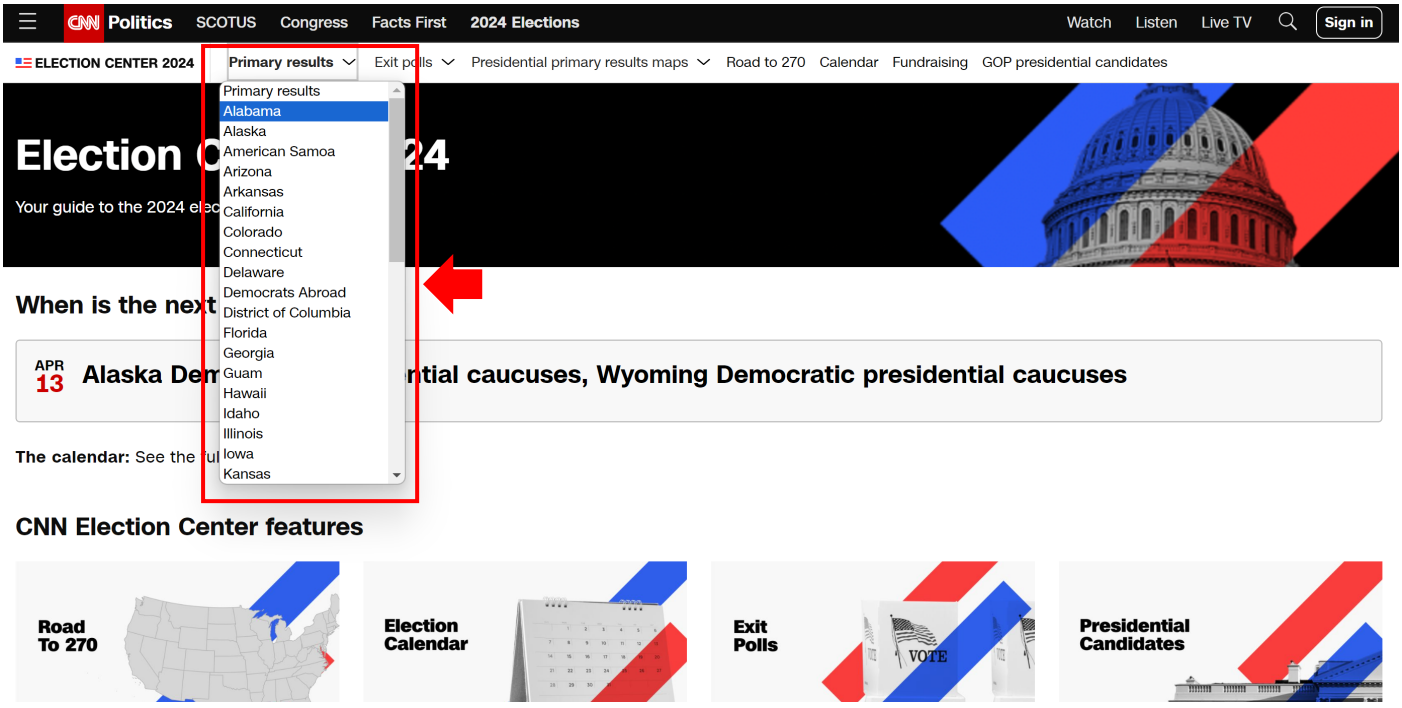
Marking Criteria

- Ensure data correctness and completeness
- Write concise and effective code, avoiding unnecessary repetition
- Make effective use of Python functions
- Write modular code with concise comments
- Use clear and descriptive variable names

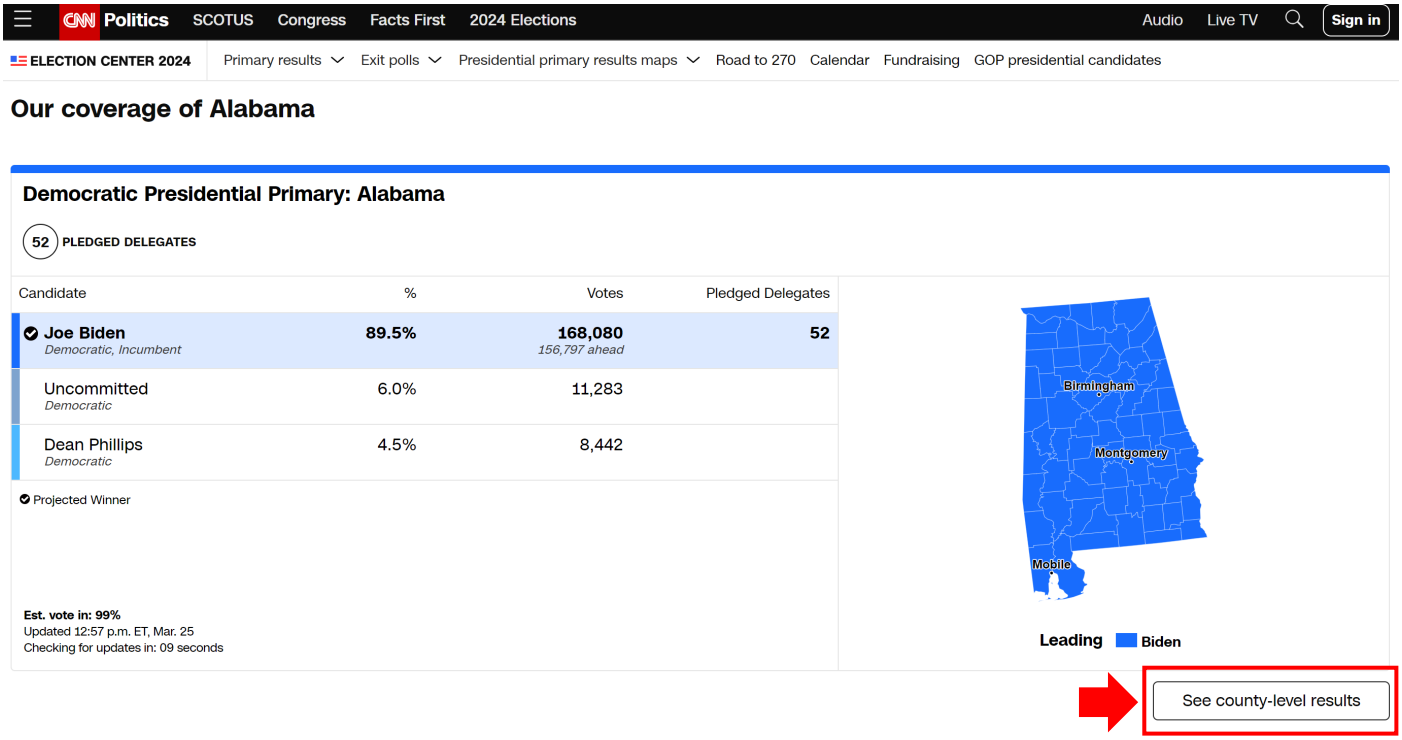
Guidelines for Web Scraping Tasks

Start from **CNN Politics | Election Centre 2024**: https://edition.cnn.com/election/2024

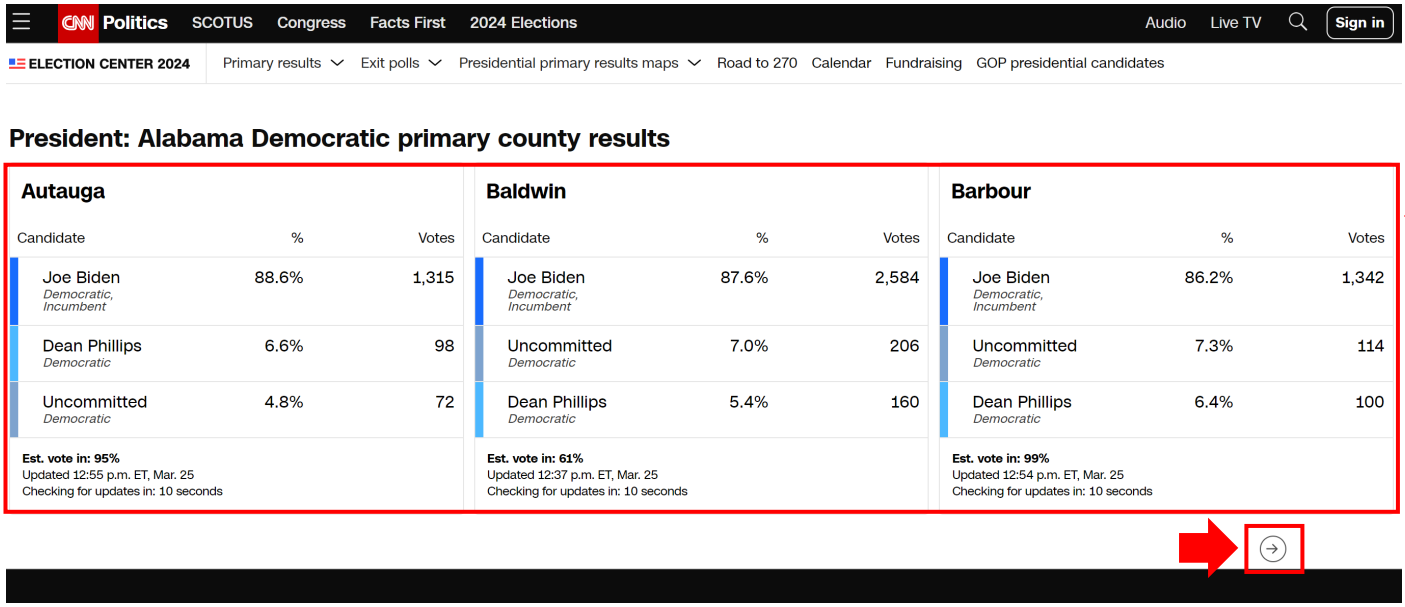
This website displays all the results of the 2024 U.S. elections. Your task is to scrape data for the **presidential primary election**. Find the menu bar at the top of the site, you can choose U.S. states from the first dropdown menu. By locating to this dropdown menu, you can scrape the names of U.S. states and territories which have completed primaries.



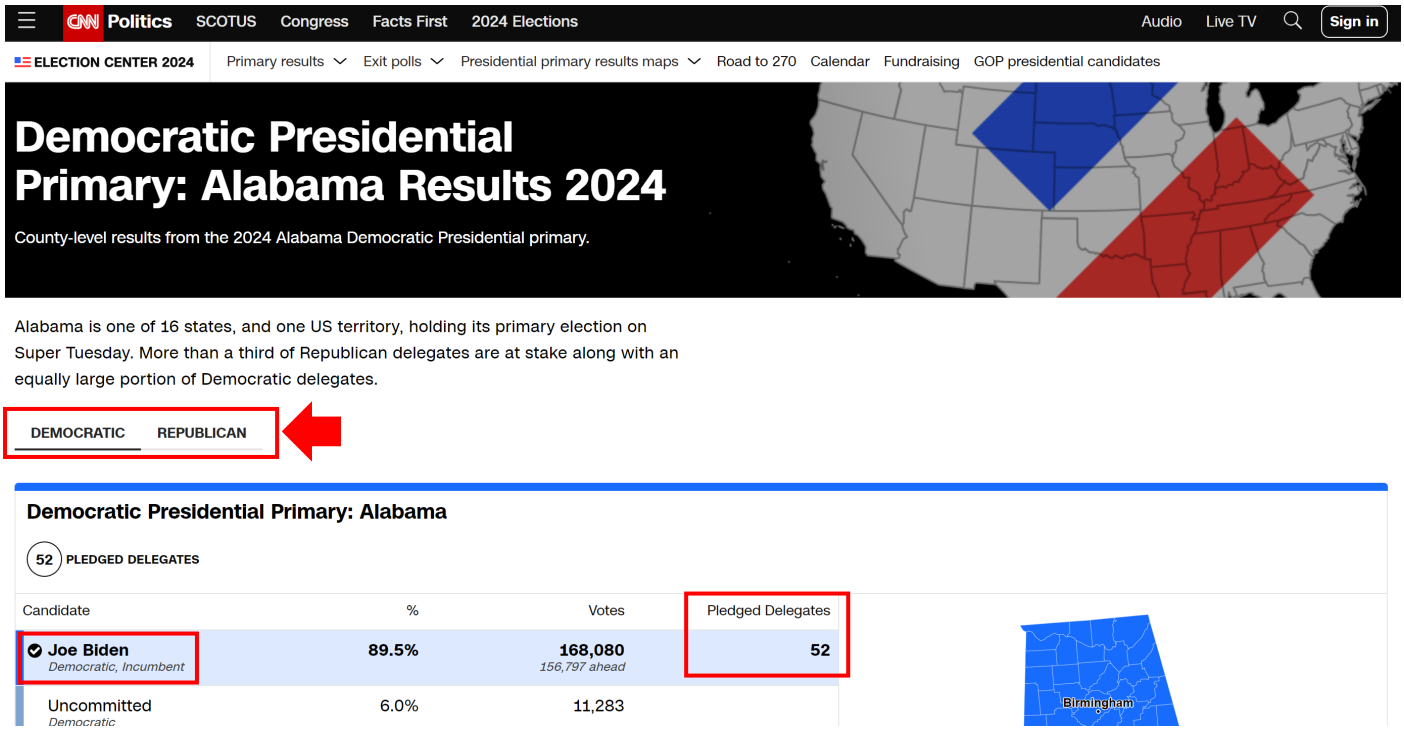
Taking one of the states, Alabama as example, the complete primary results for both the Democratic and Republican parties are displayed in tables on the pages. Since your task is to scrape the data in county level, click the button "**See county-level results**" to access another page.



From the primary county results tables, scrape the election data of ALL counties for both parties. Use Selenium methods to locate and click the right button. This will allow you to access more counties until the button is no longer visible on the page.



You should scrape data for both Democratic and Republican parties.

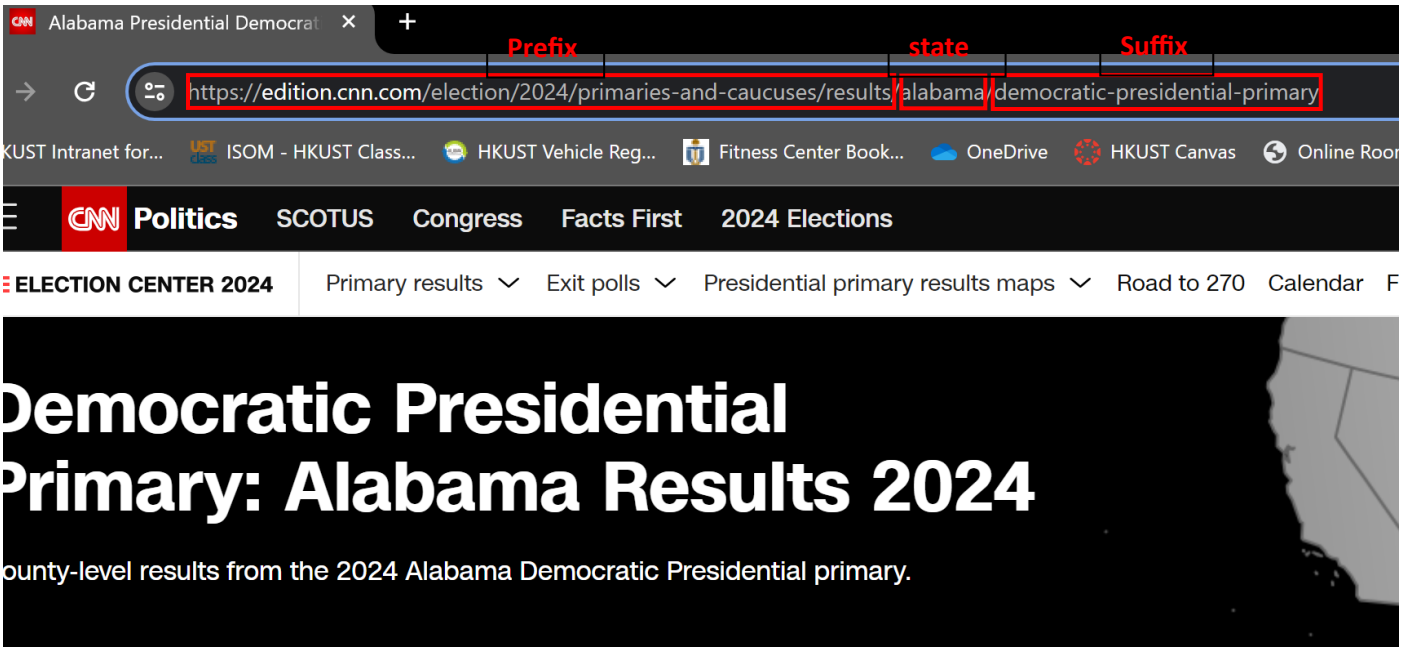


Hint: The winners and delegates data

Our goal is to extract election results for the states and territories that held their primaries on Super Tuesday. To identify these states and territories, we can refer to the following table excerpted from Wikipedia, which lists the Super Tuesday primaries for both parties.

16 states / territories held Democratic primaries on Super Tuesday	15 states / territories held Republican primaries on Super Tuesday
Alabama American Samoa Arkansas California Colorado Iowa Maine Massachusetts Minnesota North Carolina Oklahoma Tennessee Texas Utah Vermont Virginia	Alabama Alaska Arkansas California Colorado Maine Massachusetts Minnesota North Carolina Oklahoma Tennessee Texas Utah Vermont Virginia

To navigate the required pages, you can combine the link prefix and suffix with the names of the states and territories. Then, use the ***Selenium*** method "*.get()*" to navigate to the page. Only the pages of the states and territories that held primaries on Super Tuesday should be visited.



The following data should be scraped:

State, County, Candidate, Party, Incumbent, Vote, Pct, Winner, Delegates

After scraping all the data, store it in a Pandas DataFrame and use "*.to_csv()*" method to create a CSV file. Make sure to set *index=False* to avoid creating an index column. The table below shows the required data and its data format. Note that this does not include all rows of data, and you should try to scrape as many as possible.

	A	B	C	D	E	F	G	H	I
1	State	County	Candidate	Party	Incumbent	Vote	Pct	Winner	Delegates
2	alabama	Autauga	Joe Biden	Dem	Yes	1,315	88.60%	Yes	52
3	alabama	Autauga	Dean Phillips	Dem	No	98	6.60%	No	0
4	alabama	Autauga	Uncommitted	Dem	No	72	4.80%	No	0
5	alabama	Baldwin	Joe Biden	Dem	Yes	2,584	87.60%	Yes	52
6	alabama	Baldwin	Uncommitted	Dem	No	206	7.00%	No	0
7	alabama	Baldwin	Dean Phillips	Dem	No	160	5.40%	No	0
8	alabama	Barbour	Joe Biden	Dem	Yes	1,342	86.20%	Yes	52
9	alabama	Barbour	Uncommitted	Dem	No	114	7.30%	No	0
10	alabama	Barbour	Dean Phillips	Dem	No	100	6.40%	No	0
11	alabama	Bibb	Joe Biden	Dem	Yes	330	87.50%	Yes	52
12	alabama	Bibb	Dean Phillips	Dem	No	28	7.40%	No	0
13	alabama	Bibb	Uncommitted	Dem	No	19	5.00%	No	0
14	alabama	Blount	Joe Biden	Dem	Yes	228	80.60%	Yes	52
15	alabama	Blount	Uncommitted	Dem	No	34	12.00%	No	0
16	alabama	Blount	Dean Phillips	Dem	No	21	7.40%	No	0

Important note: The non-contiguous state and territory (Alaska and American Samoa) have a different page arrangement compared to the other pages. Therefore, a different logic should be implemented for scraping these pages. Failure to scrape data from these state and territory will result in a deduction of total marks.

It is important to ensure the completeness and accuracy of all scraped data, as both factors will affect your final score for the assignment. Please note that there should be more than 10,900 rows of data scraped. All the data in the required states / territories and their counties should be included. Please refer to the previous table for the list of the states / territories that held primaries on Super Tuesday.

Submission Guidelines

You should submit a program file in ipynb format and a CSV file storing the required outcomes. Name both file by your group number, e.g. **group_07.ipynb** & **group_07.csv**. Upload the files to the assignment module on Canvas for submission. The data contained in the CSV file should match exactly with the data generated by your program file.