Lab 1: Wrangling and Exploring Data

In this lab exercise, you will build on HW 1 to practice some more data wrangling for exploratory data analysis. You should work in a group of 4-5 students of your choice; you will only need to submit once as a group (see submission instructions below). Remember that labs are graded for sincere effort: if you do not finish the entire lab assignment before the end of class, turn in what you have so far.

You are welcome to use whatever tools you deem helpful. In particular, you have practiced using regular expressions to search and extract text, OpenRefine for cleaning and refining datasets, Pandas for working with structured data in Python, and programming in Python itself. You are encouraged to use documentation and other references to learn new features and functionalities of these powerful tools as they become necessary for the task at hand.

We will work with the same dataset you saw in HW 1. govtrack.us (https://www.google.com/url?q=http://www.govtrack.us/&sa=D&ust=1516309781268000&usg=AFQjCNHmEaudoYkhtS2Du47GafoIeXdJYw) contains a wealth of data on the U.S. Congress (that is, both the house and senate). Along with partners, they curate a publicly available dataset containing, among other things, information on all current members of the U.S. Congress: https://github.com/unitedstates/congress-legislators (https://github.com/unitedstates/congress-legislators). Pull the legislators-current file from the repository. It is available in different formats: the master branch maintains it as a .yaml file, but if you scroll down it is also available as a .json or .csv; the choice is up to you.

Name and NetID

Write the names and netIDs of all of your group members. Eugene Kim (yk171), Nathan Kim (njk24)

Part 0: Clean and Parse the Dataset

Nothing to turn in for this part, just a reminder that you may want to start by cleaning or pre-processing the dataset to get it into an easier shape for the exploration/analysis questions below.

Part 1: Gender by Party

This dataset identifies members of the U.S. Congress as Republicans, Democratas, or Independents and also as male or female. For each combination of party affiliation and gender, how many members of the current U.S. Congress are there with that combination? Provide the code you use to answer the questions. If you use OpenRefine, write the sequence of steps you took in OpenRefine explicitly.

Using OpenRefine, we simply filtered the Republicans in the "party" column, filtered the males(M) for the "gender" column and counted the matches. We did the same procedure for the rest of the combinations and counted the matches. There are 228 Republican males, 24 Republican females, 174 Democrat males, 107 Democrat females, 3 Independent males and no indepedent females.

In [2]:
```
import pandas as pd
f = pd.read_csv("C:/Users/ek99k/Downloads/legislators-current (1).csv")
f['party']
```

Out[2]:
```
0         Democrat
1         Democrat
2         Democrat
3         Democrat
4         Democrat
           ...
531       Democrat
532     Republican
533     Republican
534     Republican
535     Republican
Name: party, Length: 536, dtype: object
```

In [11]:
```
a = f[f['gender'] == 'F']
a
```

Out[11]:

	last_name	first_name	middle_name	suffix	nickname	full_name	birthday	gender	type
1	Cantwell	Maria	NaN	NaN	NaN	Maria Cantwell	1958-10-13	F	sen
5	Feinstein	Dianne	NaN	NaN	NaN	Dianne Feinstein	1933-06-22	F	sen
6	Klobuchar	Amy	Jean	NaN	NaN	Amy Klobuchar	1960-05-25	F	sen
9	Stabenow	Debbie	Ann	NaN	NaN	Debbie Stabenow	1950-04-29	F	sen
15	Collins	Susan	M.	NaN	NaN	Susan M. Collins	1952-12-07	F	sen
...	...	...	...	...	...	...	...	...	...
522	Wexton	Jennifer	NaN	NaN	NaN	Jennifer Wexton	1968-05-27	F	rep
523	Schrier	Kim	NaN	NaN	NaN	Kim Schrier	1968-08-23	F	rep
525	Miller	Carol	D.	NaN	NaN	Carol D. Miller	1950-11-04	F	rep
530	McSally	Martha	NaN	NaN	NaN	Martha McSally	1966-03-22	F	sen
535	Loeffler	Kelly	NaN	NaN	NaN	Kelly Loeffler	1970-11-27	F	sen

131 rows × 34 columns

Part 2: Party Identity by State

In addition to party and gender, this dataset identifies the state that each legislator represents. For each state, count the number of Republicans, Democrats, and Independents who represent that state. Use that information to identify the most Republican-leaning, most Democrat-leaning, most Independent-leaning, and most Bipartisan-leaning state. The question is intentionally vague about how to quantify these things, and there are multiple reasonable ways of defining each. In your answer, discuss at least a couple of interpretations, and explain or justify why you chose your particular answers.

In [ ]:
```
We first tried solving/organizing this through openrefine. However, we were no
t able to find the solution to this part.
We initially used multiple filters for state and party, but that proved to be
quite clunky and disorganized. We then tried
using the text facet feature in openrefine, which organized and grouped by sta
te, but that also did not work, as we realized
that we'd have to go through each individual state to record the number of dem
ocrats and republicans, which defeats the purpose
of cleaning data. Then we attempted to use pandas by looping through the state
to create a dictionary. The code is below.
However we were not able to finish and have since discovered the groupby funct
ion.
```

In [9]:
```
f = pd.read_csv("C:/Users/ek99k/Downloads/legislators-current (1).csv")
fdict={}
for x in 50:
    if f['state'] not in f:
        fdict
```

Out[9]:
```
0      OH
1      WA
2      MD
3      DE
4      PA
       ..
531    ME
532    PA
533    NC
534    NC
535    GA
Name: state, Length: 536, dtype: object
```

Part 3: The Longest Serving Members of Congress

So far, we have only looked at current terms for our analysis. In this part, we want to look at more of the historical data to search for the longest serving (current) members of congress. We want to break this down by party and by the chamber in which the member serves: either the house or the senate. So for each combination of party affiliation (Republicans, Democrats, and Independents) and chamber of the congress (house and senate), find the longest serving current member (more precisely, the member of that party affiliation and chamber who has served in that chamber for the longest). Be careful in your analysis: note that members can (and do) sometimes change from one chamber to another (especially from the house to the senate).

In [ ]:
```
We were not able to figure out how to do this one with pandas but our intended
procedure was to use openrefine and filter
through the dates and then sort to find the oldest starting term.
```

Submitting Lab 1

1. Double check that you have written all of your answers along with your supporting work in this notebook. Make sure you save the complete notebook.
2. Double check that your entire notebook runs correctly and generates the expected output. To do so, you can simply select Kernel -> Restart and Run All.
3. You will download two versions of your notebook to submit, a .pdf and a .py. To create a PDF, we reccomend that you select File --> Download as --> HTML (.html). Open the downloaded .html file; it should open in your web broser. Double check that it looks like your notebook, then print a .pdf using your web browser (you should be able to select to print to a pdf on most major web browsers and operating systems). Check your .pdf for readability: If some long cells are being cut off, go back to your notebook and split them into multiple smaller cells. To get the .py file from your notebook, simply select File -> Download as -> Python (.py).
4. Upload the .pdf to gradescope under lab1 report and the .py to gradescope under lab1 code. Only submit once per group, but be sure to add your partner using the group feature on gradescope (https://www.gradescope.com/help#help-center-item-student-group-members).

In [ ]:
```
```